ABSTRACT

DYNAMICS OF ONLINE CHAT

by
Mihai Moldovan

Millions of people use online synchronous chat networks on a daily basis for work, play and education. Despite their widespread use, little is known about their user dynamics. For example, one does not know how many users are typically co-present and actively engaged in public interaction in the individual chat rooms of any of the numerous public Internet Relay Chat (IRC) networks found on the Internet; or what are the factors that constrain the boundaries of user activity inside those chat rooms. Failure to collect and present such data means there is a lack of a good understanding of the range of user interaction dynamics that large-scale chat technologies support.

This dissertation addresses this gap in the research literature through a year-long field study of the user-dynamics of Austnet, a medium-sized IRC network (103 million messages sent to 7,180 publicly active chat-channels by 489,562 unique nicknames over a one-year period). Key results include: 1) the first rich quantitative description of a medium-sized chat network; 2) empirical evidence for user information-processing constraints to patterns of chat-channel engagement (maximum 40 posters and 600 public messages per chat-channel per 20-minute interval); 3) a short-term channel engagement model which highlights the extent to which immediate channel activity can be reliably predicted, and identifies the best predictor variables; 4) a model for the identification of factors that can be used to distinguish highly predictable channels from unpredictable channels; and 5) the first empirical study of how the Critical Mass theory can help in

predicting the channels' long-term chances of survival by looking at their initial starting conditions.

Collectively, the results highlight how the knowledge of chat network dynamics can be used in making accurate predictions about the chat-channels' levels of short-term activity, and long-term survivability. This is important because it can lead to improved designs of future synchronous chat technologies. Such designs would benefit both the users of the systems, by providing them real-time recommendations about where to find successful group discourse, and the managers of the systems, by providing them vital information about the health of their communities.

DYNAMICS OF ONLINE CHAT

by
Mihai Moldovan

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Information Systems

Department of Information Systems

May 2008

APPROVAL PAGE

DYNAMICS OF ONLINE CHAT

Mihai Moldovan

Dr. Quentin Jones, Dissertation Advisor — Date
Assistant Professor of Information Systems, NJIT

Dr. Starr Roxanne Hiltz, Committee Member — Date
Professor Emeritus of Information Systems, NJIT

Dr. Michael Bieber, Committee Member — Date
Professor of Information Systems, NJIT

Dr. Brian Amento, Committee Member — Date
Research Professor of Information Systems, NJIT

Dr. Brian Butler, Committee Member — Date
~~Assistant~~ Associate Professor of Business Administration, University of Pittsburgh

BIOGRAPHICAL SKETCH

Author: Mihai Moldovan

Degree: Doctor of Philosophy

Date: May 2008

Undergraduate and Graduate Education:

- Doctor of Philosophy in Information Systems
  New Jersey Institute of Technology, Newark, NJ, 2008

- Bachelor of Science in Computer Science,
  Technical University of Cluj Napoca, Cluj Napoca, Romania, 1999

Major: Information Systems

Presentations and Publications:

Mihai Moldovan and Quentin Jones,
"Predicting Group Interaction in Synchronous Chat Systems,"
Proceedings of the 2006 Conference on Computer Supported Cooperative Work, Alberta, Canada, November 2006.

Mihai Moldovan and Quentin Jones,
"Chaos and Group Interactions on Internet Relay Chat: a large scale field study of chat channels dynamics,"
Proceedings of the 2006 Conference of the Association of Internet Researchers, Brisbane, Australia, September 2006.

Bartel Van de Walle and Mihai Moldovan,
"An Information Market for Multi-Agent Decision Making: Observations from a Human Experiment,"
Proceedings of the 2003 Conference on Knowledge-Based Intelligent Information and Engineering Systems, Oxford, UK, pp. 66-72, September 2003.

Mihai Moldovan, Bartel Van de Walle and Aabhas Paliwal,
"Trading in the market: an experiment in group decision dynamics,"
Proceedings of the 2003 Conference on Information Technology: Research and Education, Newark, NJ, pp. 370-374, August 2003.

Soției mele, Elena, pentru toată dragostea, răbdarea şi înțelegerea cu care m-a susținut de-a lungul acestor ani.

To my wife, Elena, for all the love, patience and understanding she supported me with during the last years.

Părinților mei, Victoria şi Florin Moldovan, pentru toată dragostea, încrederea, efortul şi susținerea depusă de-a lungul întregii mele vieți.

To my parents, Victoria and Florin Moldovan, for all the love, trust, effort and support they offered me during my entire life.

ACKNOWLEDGMENT

I would like to express my deepest appreciation to Dr. Quentin Jones, who not only served as my research supervisor, providing valuable and countless resources, insight, and intuition, but also constantly gave me support, encouragement and reassurance, and was my trusted volleyball partner for many seasons.

Special thanks are given to Dr. Starr Roxanne Hiltz, Dr. Michael Bieber, Dr. Brian Amento, and Dr. Brian Butler for actively participating in my committee.

I would like to express my appreciation for Dr. Starr Roxanne Hiltz and Dr. Murray Turoff for their guidance and help throughout my first years at NJIT.

My fellow graduate students in the Information Systems Department also deserve recognition for the support and valuable feedback they provided.

Last, but not least, my wife Elena, and my parents, Victoria and Florin deserve a great amount of recognition for their thorough love, support, encouragement, and patience.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

Table **Page**

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

Table **Page**

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

Millions of people use synchronous online chat systems, such as Internet Relay Chat (IRC), on a daily basis. Such systems often contain thousands of chat-channels (group interaction spaces), populated by even larger numbers of users, providing individuals with the opportunity to discuss a wide variety of topics. Despite their widespread use, little is known about their user dynamics. For example, one does not know how many users are typically co-present and actively engaged in public interaction in the individual chat spaces of any of the numerous IRC networks found on the Internet; or what are the factors that constrain the boundaries of user activity inside those chat rooms. Unfortunately, the navigation of large-scale chat systems is difficult. In part, this is because only relatively impoverished mechanisms exist for user navigation of such large-scale, highly dynamic synchronous chat systems. These navigation mechanisms are through (1) users selecting channels based on the presentation of the names, topics, and number of users (people or software agents) in either the full or a subset of the thousands of channels available; (2) users searching for known conversational partners and joining channels in which they are present; and (3) users learning through personal experience the online places and times where a population of potential interaction partners is likely to be present. Each of these mechanisms presents users with as many irrelevant as relevant channels, or relies on extensive user knowledge of the system in question.

In theory, the navigational difficulties associated with large-scale chat systems could be alleviated by the development of tools that identify and recommend in near real time a smaller subset of channels that are likely to be of value. The value of a channel to a user can be assessed in multiple ways. Three basic assessment questions are the following: (1) Is a sustained group of users likely to be present in a chat-channel for engaged interaction? (2) Is the topic of discussion likely to be of interest? and (3) Are suitable interaction partners likely to be present? However, at present it is not known how to effectively identify any of these features in large-scale chat systems, each of which could form the basis of a synchronous chat recommendation system.

To research this gap in the literature, this dissertation will explore the first of these factors, namely, when is a sustained group of users likely to be present in a chat-channel for engaged interaction? This question can be divided into two topics that complement each other. The first one addresses the short-term group activity of chat-channels, i.e., the potential of a group to engage in lively public interactions for short periods of time, and the ability to predict such intervals of high activity or low activity. The second one addresses the long-term group sustainability, i.e., the likelihood of a group to survive over longer periods of time.

Regarding the short-term activity of chat-channels, currently one does not know the extent to which reliable short-term predictions about this activity can be made. First, while it is known that generally large-scale synchronous systems do not limit the number of users or postings in a channel, little is understood about the boundaries imposed by the users' capabilities. The Information-processing constraints theory (Jones 1997; Jones and Rafaeli 1999) argues that one of the main influences on a user's participation in

computer mediated communication (CMC) is the level of information overload to which the user is exposed when using the system. Prior research on asynchronous CMC systems has shown that the level of activity within such a system can only rise up to a certain level. After this level is reached, due to the effects of information overload, the activity either remains constant or decreases. This theory can guide empirical research aimed at identifying the boundaries to short-term activity imposed by user information processing constraints in synchronous chat systems. Second, to date no empirical work has investigated the extent to which short-term measures of chat-channel activity can be reliably predicted; or the factors that differentiate highly predictable chat-spaces from unpredictable chat-spaces. In theory, linear, nonlinear, and logistic regression modeling can address these shortfalls. Provided enough information about the activity of groups can be collected and analyzed, the regression models may predict the short-term level of activity of chat-channels as well as their overall degree of predictability, based on sets of various independent variables computed using the collected group-dynamics data.

Regarding the long-term sustainability of chat-channels, there are also empirical and theoretical gaps in the literature. The Critical Mass theory of sociologists Oliver, Marwell, and Teixeira (1985) provides the only well-known theoretical model of group interaction trajectories. It postulates that various production functions, or the relationship between resources contributed and collective output over time, together with a group's level of heterogeneity can be used to distinguish between the likelihood of longer-term group success. Although this theory was adapted to electronic media (Markus 1987), its modeling aspect has not been fully applied to CMC systems. Only a small number of its hypotheses were tested using asynchronous CMC systems (Thorn and Connolly 1987;

Rafaeli and LaRose 1993) and the researchers failed to observe the theory's predictions. In theory, the Critical Mass theory could help predict the long-term sustainability of groups in synchronous spaces.

The aim of this research is to explore, empirically and theoretically, whether it is possible to predict the likelihood of sustained short-term and long-term group interaction inside a large-scale synchronous CMC system. This is important because such prediction algorithms may be used to design systems to benefit both individual users, by providing them real-time recommendations about where to find successful group discourse, and managers of group spaces, by providing them vital information about the health of their communities. Understanding the short-term channel activity predictability would help in providing instant recommendations. Understanding long-term sustainability would help in recommending channels that have a good chance of survival, and in determining when recommendations would lead to a more functional group in the long run.

In what follows, the content of the remaining chapters of this document will be described briefly.

First, a basic decision needs to be made about the type of synchronous chat system to analyze. Therefore, Chapter 2 is dedicated to an overview of the literature on the most common classes of synchronous CMC systems. This will lead to an understanding of the main characteristics of such systems, how are they used, and which type would be the best match for this research, in terms of relevance to the other existing systems, popularity among users, and data-collection capabilities.

Second, it is necessary to understand the current methods that are used for providing social recommendations, as well as their limitations. Such methods include

social visualizations, which allow users of CMC systems to be aware of the activities inside the spaces they inhabit, and social matching systems. The relevant literature shall be reviewed in Chapter 3.

Third, one needs to be aware of the work that attempted to identify the rhythms of various synchronous and asynchronous CMC systems. Accordingly, Chapter 4 will be dedicated to a review of the research in this area.

Fourth, the work that has been done on identifying group interaction trajectories in CMC systems must be considered. Chapter 5 will start with a review of the literature on CMC interaction dynamics analysis. It will continue with theoretical considerations about the Critical Mass theory and conclude with an overview of previous works that used this theory to research interactive CMC media.

The research questions and hypotheses, and the associated research methods will be presented in Chapter 6.

The results of the research will be presented in Chapters 7 through 11. Chapter 7 will report various descriptive statistics of the analyzed IRC network. Chapter 8 will identify the information-processing limits that constrain the community interaction dynamics seen in IRC channels. Chapter 9 will investigate the extent to which short term measures of activity can be reliability predicted for IRC channels. Chapter 10 will identify the factors that can be used to distinguish highly predictable channels from unpredictable channels. Chapter 11 will describe the factors that contribute to the long term survivability of IRC channels. Chapter 12 will discuss the contributions of this work and will present several topics of interest for future research.

CHAPTER 2

TYPES OF TEXT-BASED SYNCHRONOUS CMC SYSTEMS

This chapter provides an overview of the most common types of synchronous computer mediated communication (CMC) systems and reviews various alternative candidate systems for the empirical research intended to be conducted. The overview of the literature on the most common classes of synchronous CMC systems will allow a better understanding of the main characteristics of such systems. These characteristics include how they are typically used and which type would be the best match for this research, in terms of relevance to the other existing systems, popularity among users, and data-collection capabilities. In this research the focus is strictly on text-based synchronous computer mediated communication systems, because they are the most common type. Technologies such as Push-To-Talk, a method of conversing over half-duplex communication lines, and VoIP (Voice over Internet Protocol) applications such as Skype are omitted from this review. The work that has been done on Short Message Services (SMS), which is asynchronous in its nature, was excluded from the literature review.

Hiltz and Turoff (1993) and Kerr and Hiltz (1982) defined CMC as the "use of computer to structure, store and process written communication among a group of persons." This definition could be enhanced by noting that a computer mediated communication system is any software or hardware system that allows people and/or groups of people to interact using written, video, or audio communication methods.

Traditionally, the starting point for categorizing a CMC system has been in terms of whether it supports synchronous or asynchronous communication (Newhagen and Rafaeli 1996).

While this categorization can be applied to many systems, it should be noted that such categorical distinctions are not clear-cut. Communications created using synchronous technologies can be stored and made persistent and searchable, thus enabling asynchronous use of the medium. Further, synchronous communication is often "near" real-time, meaning that users must hit the carriage return key before the information they have typed is shared. On the other hand, asynchronous communication tools such as email can be used for quick message exchanges that make the interactions near synchronous. These issues highlight both the elasticity of synchronicity and the importance of understanding the significance of making interactions persistent (Erickson and Laff 2001). In this research, systems that support either near real-time or true real-time communication shall be referred to as synchronous systems. The evolution of such systems started with operating system commands, which initially allowed synchronous text communication between two users of a system. These were later enhanced to accommodate more than two users. The next step was the emergence of smaller synchronous group chat systems organized around spatial metaphors (early 1980s). Those systems were followed by large-scale mass interaction chat systems (late 1980s), after which instant messaging systems arrived (mid-1990s). The research associated with various systems will be addressed in five sections:

1. Early dyadic (in pairs)/group synchronous communication systems (*talk*, *ytalk*) – the precursors to synchronous group CMC
2. Multi-User Domains/Multi-User Domains Object Oriented (MUDs/MOOs)
3. Internet Relay Chat (IRC)
4. Instant Messaging (IM)
5. Other chat systems and emerging trends

2.1 Early Dyadic/Group Synchronous CMC

Operating system commands, such as UNIX's *talk* and *ytalk* (and similar commands on other operating systems) – are among the oldest forms of synchronous CMC (Hiltz and Turoff 1993). The *talk* command was originally used for synchronous text communication between users of a multi-user computer, running the UNIX operating system; but it eventually started to accommodate users from different computers as well. It allowed two people using different machines (or different terminals of a single computer) to establish a real-time text communication channel. The dyadic nature of the *talk* command user interface made it impossible for more than two users to interact synchronously.

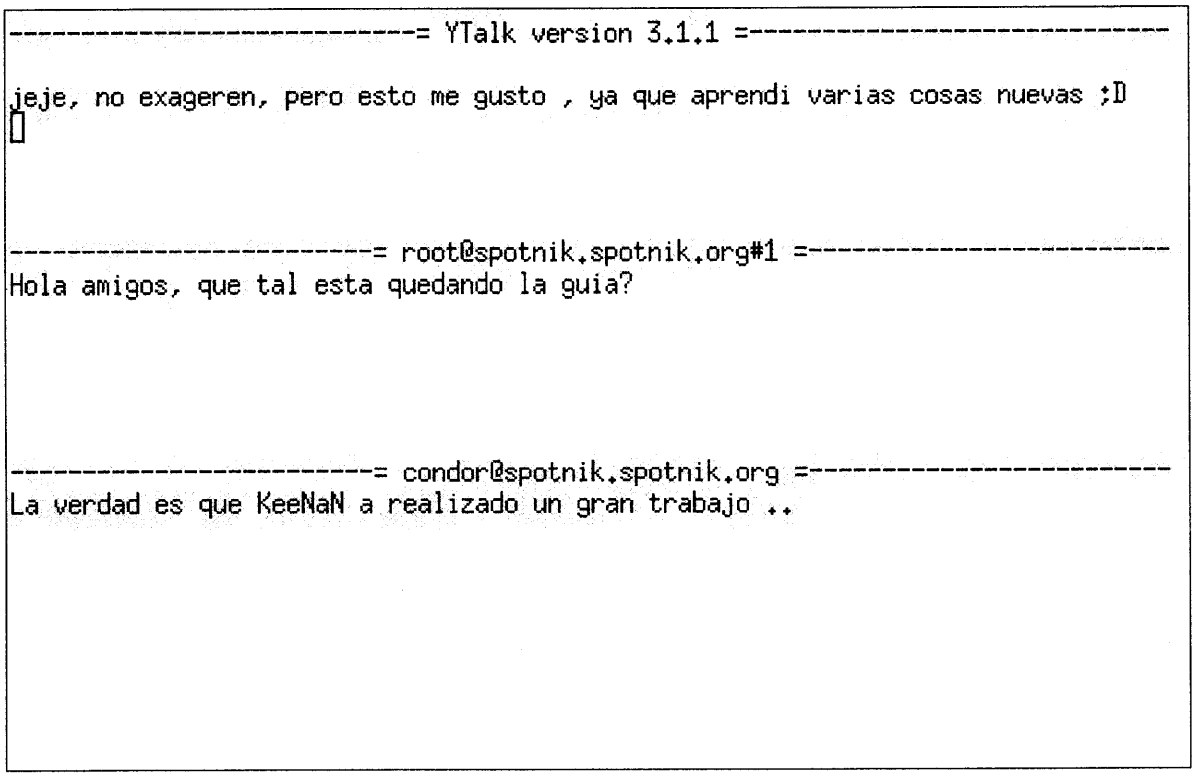

Figure 2.1 The UNIX *ytalk* command interface.

This inconvenience led to improved commands such as *ntalk* or *ytalk* (Figure 2.1), which allowed several participants to exchange text messages in real-time. In all these cases the communication happened in true real-time. Each character typed by one of the users was immediately displayed on the screens of the other users. The user interfaces were simple and did not convey information other than the text typed by the users.

The "true" synchronous nature of these early computer mediated communication tools distinguishes them from most of today's systems. The current synchronous chat software interfaces are often referred to as being near-synchronous or quasi-synchronous (Garcia and Jacobs 1999). The argument in favor of this term was that "although posted messages are available synchronously to participants, the message production process is available only to the person composing the message." In other words, the messages are

sent to the interaction partners only after they are fully composed, giving the sender a certain degree of control in the creation and distribution of messages.

Although systems such as *talk* and *ytalk* are interesting technologies, they are not relevant for this research for several reasons: They are rarely used for large group interaction; they are not very common anymore; they cannot offer multiple interaction spaces; and the features they provide can be found in today's large-scale synchronous chat systems.

2.2 MUD/MOO Research

2.2.1 MUDs and MOOs Overview and History

MUDs/MOOs were the first generation of applications dedicated to synchronous group interaction over computer networks. The distinctive feature of the MUDs/MOOs was their representation of computer mediated interaction spaces through a spatial metaphor (Figure 2.2). The interaction spaces inside these systems were organized as rooms, often mimicking the geographical structure of some real-world buildings or various other places. To understand the structure of the system and to meet new people, the users had to navigate through these rooms sequentially, without the possibility of being in multiple places at the same time. This geographical metaphor made the MUDs/MOOs fundamentally different from other synchronous CMC systems that will be reviewed later in this chapter.

In August 1978, Roy Trubshaw, a student at University of Essex, England, started to write a text-based adventure game on a DECsystem-10 machine. His initial aim was twofold: to create a true multi-player adventure game and to write an interpreter for a

database definition language. Soon he was joined in his efforts by Richard Bartle, also an Essex student at that time. Together, they created and wrote Multi-User Dungeon (MUD) in 1979. Initially, only students at the University of Essex used the game, but players from other universities soon joined, and the game became very popular in England. With the authors' permission, the source code was used for other projects such as Frog, BLUD, UNI, and MIST. Additionally, some copies of the code were sent to USA, Norway, Sweden, and Australia; after that it continued to spread and evolve.

Initially, the term MUD referred to one particular game – Multi-User Dungeon – the first game of this type, created by Roy Trubshaw and Richard Bartle. However, the term became widely spread, and soon it started to be used to refer to this entire class of games, not only the initial Multi-User Dungeon game, a practice that is still common today. However, there are some variations to what the letter "D" represents. Some authors (Mehlenbacher et al. 1994; Sempsey 1995; Garbis and Waern 1997) use the term Multi-User Domains, while others (Curtis 1996) prefer the term Multi-User Dimension. However, the majority of the authors continue to use the initial meaning of MUD (Multi-User Dungeon). As MUDs became more and more popular, some of the users were not satisfied with the features and attributes of the available MUD applications and enhanced the software. One of the most important MUD spin-offs, created to diversify the realm of interactive text-based gaming, was MUD Object Oriented (MOO). MOO was created by Stephen White in 1990 at University of Waterloo, and then adopted and extended by Pavel Curtis at Xerox PARC (Dourish 1998). It differed from the traditional MUD in that it allowed the users to extend the existing environment using the object-oriented programming paradigm. LambdaMOO, (Curtis 1996) and MediaMOO (Bruckman 1993)

were perhaps the largest environments implemented on the MOO platform. Other variations of MUD systems include Multi-User Simulation Environments (MUSEs) and Multi-User Shared Hallucinations (MUSHs) (Bruckman 1992).

All these terms however, refer to one single type of computer application, "a software program that accepts connections from multiple users across some kind of network (e.g., telephone lines or the Internet) and provides to each user access to a shared database of 'rooms,' 'exits,' and other objects" (Curtis and Nichols 1993).

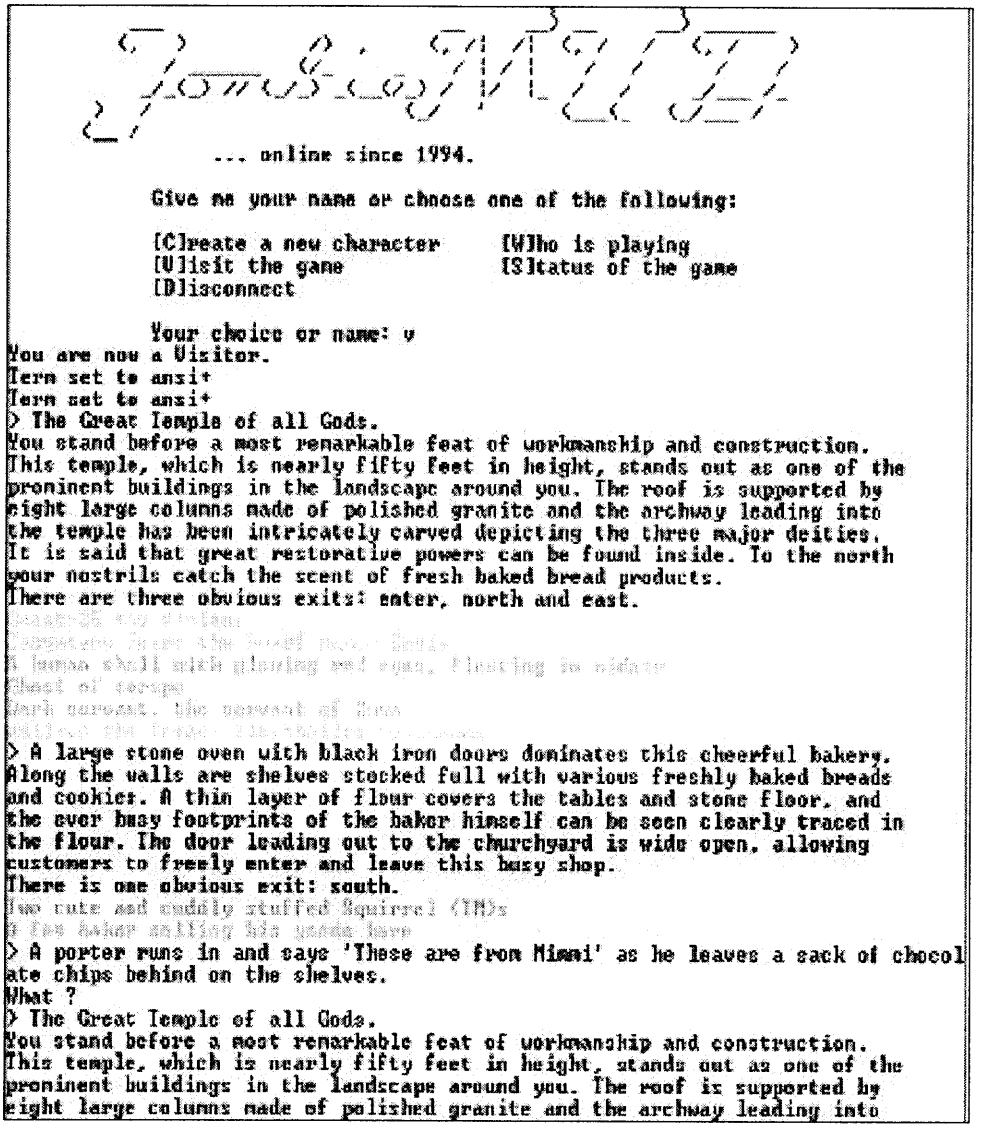

Figure 2.2 An example of MUD user interface.

A more pertinent definition of the term MUD would be "a text-based virtual environment" (Churchill and Bly 1999) or a "text-based virtual reality" (Cherny 1994) that may support interpersonal communications, community building, game-playing, or other social or educational activities. For the remainder of this chapter, the term MUD will be used to refer to any of the above-mentioned variations of this type of applications. MUDs are of interest because they were among the first, if not the first, online synchronous group communication spaces. Chatting (group or in pairs) is typically either the most or second most important activity in a MUD (Curtis 1996).

2.2.2 MUDs Classification

MUDs can be categorized according to the main ways in which they are used. Bartle (1992) was the first one to observe two different patterns of MUD usage: (a) the predominant, traditional, adventure-game style, where users compete for achieving higher status within the game; and (b) a more social-oriented style, where interacting and socializing with other people tends to be the primary purpose of the entire MUD experience. Another author who mentioned two different styles of MUD usage was Bruckman (1992). She described two types of MUDs, "those which are like adventure games and those which are not." Garbis and Waern (1997) conducted a workshop in November 1996 that had the "design and use of MUDs for serious purposes" as the main theme. Within the workshop, three main areas of use were identified: education and teaching, crisis action management, and general communication environment. Finally, another potential area of application for MUDs is in the general workplace. Evard (1993) had the idea of using MUDs as system tools, to enhance the communication among a team of system administrators. Churchill and Bly (1999) conducted research over longer

periods of time (three years) on the use of MUDs to support ongoing collaborations in the workplace. Considering all of the above, MUDs can be classified into the following four categories:

- Purely gaming MUDs
- Social-oriented MUDs
- Education-oriented MUDs
- Workplace-oriented MUDs

Obviously, there may be cases where some degree of overlapping between these categories can be found. However, this taxonomy provides a framework in which any existing MUD could be integrated.

2.2.2.1 Gaming MUDs. The game-playing-oriented MUDs are probably the most widely used type of MUD application. For the first ten years after the birth of MUDs, their sole purpose was for use in gaming. However, very little research has been done on gaming MUDs. Bartle (1992) provided a very thorough review of Multi-User Adventure (MUA) games. He found evidence of two groups of users that were mostly unaware of each other, one in the UK and one in the US. Also, he provided an interesting genealogy of the evolution and the status quo of MUAs up to 1992.

Muramatsu and Ackerman (1998) looked at gaming MUDs from a social perspective. Although playing the game was the central activity in the researched MUD (called Illusion), social involvement and collaboration were also important. They found the MUD was deeply socially stratified, both formally (mortals, immortals, and various degrees of power to control the game) and informally (levels of the game) and that comparable levels of conflict and cooperation were observed among the users.

2.2.2.2 Social-oriented MUDs. One of the most cited papers in the MUD literature is Pavel Curtis's paper (1996). He was probably the first author who analyzed a social MUD. His work served as an important foundation for subsequent research. Curtis observed certain rhythms in the system's usage, including higher levels of activity and a larger number of users during certain time intervals. Anonymity seemed to be the most significant social factor, having both positive (less social risk, lower inhibitions) and negative (sexual harassment, offensiveness) effects. Small groups of users tended to spend most of the time in conversations and social gravity – defined as the assumption that the more players inside a certain interaction space, the more interesting the interactions – was common. Looking at the community as a whole, the author observed that communities tended to be large in comparison to the number of active users at any given time (there was a large number of inactive users at any time).

One researcher who was deeply involved in the study of MUDs was Amy Bruckman, from MIT Media Labs. In one of her first papers (1992), she looked at the various social and psychological phenomena that appear in "text-based virtual realities," as she called the MUDs. She identified two types of MUDs – adventure-based and non-adventure-based. All the observed MUDs had well-defined hierarchical structures. Although initially social hierarchies were not present, in order to impose equality among players, they were gradually introduced in the MUDs. Her findings showed that identity play, i.e., pretending to be different in the virtual world than in real life; gender swapping, i.e., pretending to be of the opposite sex; and addiction phenomena were quite common. Also, she noted that many MUD users had poor social skills in real life, but they were really different in the virtual worlds.

This gender-swapping phenomenon was later analyzed by Bruckman in more detail (1993). According to her, "gender swapping is one example of how the Internet has the potential to change not just work practice but also culture and values." MUD experiences can help people understand how gender structures human interactions and how easily these phenomena can be understood by experiencing them directly. Her findings showed several interesting facts. Men were often surprised by how they were treated when they posed as women. Female characters were usually overwhelmed with attention. Unwanted attention and sexual advances created an uncomfortable atmosphere for women. Male players often logged on as female characters. Unsolicited offers of assistance were very common, and male characters usually expected some kind of favors from the female characters in return for their assistance.

Cherny (1994) looked at user behavior in a MUD and tried to observe differences in this behavior, based on the gender of the users. The MUD that she examined was JaysHouseMOO (JHM) – a MUD with a population of approximately three fourths men and one fourth women. They found that such differences indeed existed. Men were more likely to use physically violent imagery during conversations, while women were predisposed to more affectionate imagery toward other characters.

Sempsey (1995) reviewed the literature relevant to the psychological and social aspects of MUDs in an attempt to ascertain the current state of knowledge in this area and, if possible, to provide some future directions for the researchers. He suggested that disinhibition (reduction of inhibition) was common in any MUD, as people generally tend to be more relaxed in a virtual environment than in real life. However, disinhibition should not be confounded with uninhibition (complete lack of inhibition) – as MUDs

should be neither chaotic nor anarchic. He noted that most of the literature acknowledged gender-based differences in users' behavior as well as gender-swapping and multiple-identities phenomena. He stated that the most evident conclusion was that more research was needed as too little had been done up to that point, and noted the general bias toward non-experimental research designs. Also, in the author's opinion, it would be difficult to provide a general theoretical research framework for the study of MUDs because of the many possible differences among them.

Schiano and White (1998) conducted online surveys, personal interviews, and studied various logs in order to understand the social aspects of MUD usage. They determined four categories of factors that would be of interest in such an analysis: user and use characteristics; identity; sociality; and spatiality. In general, their findings were in line with those of previous researchers: The proportion of males versus females was consistent with previous estimates (78 percent versus. 22 percent); MUDs could be fairly intimidating to novices; interaction was the primary goal and experience was associated with more time spent socializing and less time exploring. However role-playing and gender swapping occurred only casually. As a whole, the analyzed MUD (LambdaMOO) did not seem to qualify as "a good great place," as defined by Oldenburg (1991) because most of the socializing tended to be done in small groups or in pairs (although this is what typically happens in the good places such as the bars described by Oldenburg). Some features were, however, encountered and the predominant opinion was that "'the great goodness' appeared earlier in the history of the MUD." Finally, LambdaMOO did provide a sense of space and place to its members, even if it was only a text-based environment.

Mynat and O'Day (1998) introduced the concept of network communities as a new genre of collaboration. They were "robust and persistent communities based on a sense of locality that spans both the virtual and physical worlds of their users." They categorized MUDs as a type of network community. They studied Pueblo, a cross-generation, school-centered, text-based MUD; and Jupiter, a hybrid MUD/media space (it provided audio/video links between participants). Their research approach was also non-experimental and consisted of participation and use, ethnographic observations, interviews, and participant observation. They identified several affordances of network communities: persistence (durability across time); periodicity (defined by rhythms and patterns); boundaries; engagement (ways to participate, not only technological but also social); and authoring (the ability to change the medium - creation of objects, rooms, etc). The authors stated that the amount of synchronous communication was an important measure of the activity level of the MUD and that multimodal communication occurred very frequently. Also, each communication modality had its own rhythm, and users moved easily between various types of interaction. Rhythm dynamics were considered fundamental to network communities; one final conclusion was that routines, or intelligible rhythms for individuals and for the community as a whole, were likely to emerge in the future inside network communities.

2.2.2.3 Education-oriented MUDs. The potential of MUDs for supporting education, at various levels, quickly attracted researchers' attention. Bruckman (1994) presented a case study with the experiences of a 43-year-old building contractor in learning to program in a MUD. She used MediaMOO for this case study, and several people who had never programmed before learned to do so in this MUD. Although this was a small

first step in analyzing the impact MUDs can have on a person's education, the author identified possible future directions toward creating MUDs as learning environments for children.

Masinter and Ostrom (1993) imagined a system that would combine two of the most common Internet uses, as they were in 1993: information retrieval and interpersonal communication. Their paper described a system that combined an information retrieval tool (Gopher) with a text-based virtual reality tool (MUD). The Gopher system was implemented in the form of Gopher rooms, Gopher notes, Gopher lists, Gopher slates, and Gopher notebooks. People in the MUD could use those rooms or objects just like any other object in the MUD. The only difference was that they could access Gopher information through them. Overall, this was an interesting idea toward application integration.

Bruckman and Resnick (1993) researched the MediaMOO system - a text-based, networked VR environment designed to enhance professional community among media researchers. They analyzed experiences with the system and highlighted the value of constructionist principles to virtual reality design. The authors argued that text-based virtual environments provide both a shared place and a shared set of activities. They noted that the best interactions typically occurred when people participated in a shared activity and not just a shared context.

Bruckman and Resnick continued to extend their ideas (1995). They argued that "serious exchange of ideas often takes place because of, not in spite of, more informal social interaction." The constructionist philosophy - "learning by doing is better than learning by being told" - was central to the design of MediaMOO. Their goal was to

create a basic skeleton that users could enhance indefinitely by creating new meaningful objects and places - a world built by its inhabitants that was a representation of various places in the real world. The authors also introduced MOOSE Crossing, a MUD for kids that was based on their findings from MediaMOO.

Mehlenbacher et al. (1994) attempted to build a tool that would increase collaboration among professional technical communicators. TechComm-VC was a MOO environment built at North Carolina State University for use in composition and technical writing courses. It used a real world / virtual world metaphor, as the MOO was based on the real campus. The main components of the MOO were: the Library, with the main purpose or providing access to information such as online dictionaries, search engines and various hypertext documents; the Lecture Hall, a room designed for formal talks and discussions with various implemented features, such as the conch – a token that controlled the room, the slide projector, the tape recorder and the tape dispensing machine; and Harrelson Hall, a room designed for informal communication and group work. Building communities and redefining professional productivity were recognized as the top two opportunities offered by the system.

Fanderclai (1995) described the potential of MUDs for use in the educational process. The immediacy of information exchange, the lack of a need for collocation and the blending of work with play were features that, according to the author, could help MUDs "disrupt the hierarchy of the traditional classroom, giving students more power and responsibility." Results of MUD use showed, besides the above-mentioned attributes, the presence of a great deal of incidental learning. The author argued against the use of traditional educational methods, such as on-line lectures, control over the flow

of conversation or over the access to a particular room, etc., because such methods would not be suited to MUD environments. The author also argued that MUDs were not environments that could be controlled, and teachers should not try to control them as most of the potential of this new educational environment would be lost.

Bruckman (1998) described MOOSE Crossing – a MUD designed to be a constructionist learning environment for children ages eight to thirteen. The author stated that the Internet could be used as a context for learning through community-supported collaborative construction. The paper examined several ways in which the MOOSE Crossing community motivated and supported its members' learning experiences. Some of the factors that contributed to the learning experience were role models and the importance of learning from them; situated, ubiquitous project models; emotional and technical support; appreciative audience; and the combination of the local community with the virtual community. Among the factors influencing learning in online communities, the author included online engaging, teaching others, invisibility of some social factors, and spontaneous and scheduled interactions. The paper concluded that, in general, constructionist educational technologies have not yet lived up to their great promise. A certain culture has emerged from within the MOOSE community where the users learned from each other as they were the ones conveying the information.

O'Day et al. (1998) looked at design considerations in moving educational practices from physical to virtual places. They argued that the affordances of virtual and real places could be very different. Among the affordances of the virtual worlds, they identified multi-threaded activities and conversations, ability to keep records of interactions, a sense of place in the virtual world, and the long-term persistence of

artifacts. The authors suggested four design dimensions for virtual classrooms: establishing permeable boundaries of interaction; taking advantage of the persistence; finding and maintaining focus; and learning from experience. They argued that MOOs were a suitable medium to use as virtual classrooms, but design of the educational practice should always be aligned with the affordances of the medium.

2.2.2.4 Workplace-oriented MUDs. Curtis and Nichols (1993) were among the first researchers who looked at MUDs as potential tools to enhance the workplace. Their intention was to extend MUD technology such that it could be used in non-recreational settings. They tried to overcome some of the drawbacks of text-only MUDs in the workplace by providing several new features. These features included support for graphical user interfaces that allowed users to interact with the MUD environment through windows on their screens; shared access for certain window-based applications; audio features that allowed users to hear sounds from the room they were in; video that could enrich the perceived quality of the CMC process, monitor remote places, and allow participation in remote meetings. Of these features, audio was considered to be the most productive addition, as it eased the communication process and was not as ambiguous as text. LambdaMOO was the starting point for the Astro-VR and Jupiter systems. Astro-VR was a social virtual reality system intended for use in the astronomy community and supported real-time multi-user communication, a self-contained email system, links to online astronomical images, editor/viewer for short presentations, collaborative access to various programs and window-based shared editors. The Jupiter system was used at Xerox PARC, and it supported casual interaction, telecommuting, convolving the real and virtual worlds, and administrative support. One of its goals was to tie together not only

the researchers at PARC, but all Xerox employees over the world. The architecture of these two systems was further described in Curtis et al. (1995) – system infrastructure, security/trust model, clients, server, networking, user interface, media coordination system - and Nichols et al.(1995) – the concurrency-control algorithm used to maintain common values for all instances of the shared widgets.

Evard (1993) presented the experiences of a network administration group in using a MUD as a tool for communication (regular means of communication included email, face to face meetings, and newsgroups). The MUD seemed to solve the existing communication problems; was an effective way to hold pre-arranged meetings for remotely located people; was an effective coordination, brain-storming, and problem-solving mechanism; created a social environment that did not exist before; and allowed work to be done more effectively from remote locations. There were also some inherent problems. The MUD had the potential to be a big distraction. It required some time to learn; and the conversation could easily become confused and intertwined.

Guzdial (1997) wanted to offer the power of a command line embedded in a virtual community, and therefore he extended a text-based virtual environment through a small command server. The result was WorkingMan, a MOO through which the users could interact with and control their workstation. It was based on POO (a MOO written in Python programming language), and it was a good example of how new metaphors for user interaction could be developed.

A workshop on the "Design and Use of MUDs for Serious Purposes" was held during the 1996 Computer Supported Cooperative Work (CSCW) conference (Garbis and Waern 1997). The aim of the workshop was to examine a number of questions related to

the design and use of MUDs for serious purposes. There were two main lines of discussion: current use and functions of MUDs; and potential new features that could be added. Three areas of use were identified: education and teaching, crisis action management, and general communication. A consensus was reached, suggesting that no standard body of methods existed for the analysis and evaluation of data gathered in a MUD and that evaluation should be grounded in the practice of the users and not in theoretical speculations. The participants also tried to suggest a name change to CVE - Collaborative Virtual Environments.

Churchill and Bly (1999 1) examined how the use of MUDs could support ongoing, medium-term collaboration in the workplace. A very simple, text-only MUD was used, and the findings showed that all users found it to be of potential benefit for their work collaboration. There were two main purposes for which the MUD was used. Firstly, it was intended to maintain relationships across distance, supporting collaboration between people in different locations; across time, supporting collaboration between people who were not online at the same time; and for people who were inside larger groups as the MUD was predicted to be more efficient than email for group interactions. Secondly, it was intended to provide lightweight conversations for coordination, enabling in this way chance encounters that would not have taken place otherwise. The authors also identified two types of barriers to MUD usage: social barriers and technical barriers. The social barriers included removal of the social cues gained from physical interaction, anonymity issues, fear of managerial perceptions, and feelings of not belonging to the group. The technical barriers were relatively easy to remediate, but there were trade-offs between functionality and lightweightness. Their conclusion was that there were four

interrelated factors central to the success of the MUD in this instance: communication and coordination affordance; technological and usage lightweightness; the existing work practices of the group; and the organizational willingness to accept this type of interactions in the workplace, as part of the daily routine.

Churchill and Bly continued their research on MUDs in the workplace and further studied the use of a text-based virtual environment to support work collaborations (1999 2). They believed interactions could be facilitated by structuring the virtual work space/environment to create a shared "landscape of work artifacts" and that the creation of a feeling of co-presence between non-collocated collaborators was crucial if CMC was to be successful. They identified the representational simplicity (simple activity awareness cues, simple asynchronous messaging) and the ongoing support of work collaborations as two factors that made MUDs useful in the workplace. The authors concluded by saying that MUDs presented a greater potential than realized and that MUD usage had changed in subtle ways but not substantially. Additionally, people felt connected, despite the simplicity of its representation; and MUDs supported sufficient co-presence to enable complex collaborative problem solving in difficult circumstances.

The work of Schafer, Bowman, and Carroll (2002) stands among the latest MUD research. The authors combined a text-based interface with a graphical map representing the environment of a MOO. Their system, called MOOsburg, was based on a real place: Blacksburg, Virginia. The target was the town population, and the overall objective was to support community development within the town. The structure of MOOsburg was based on spaces and landmarks; and the interaction was based on text, clicking, dragging and zooming. Their findings showed that maps allowed users to browse and visit places

while maintaining awareness. They also provided a good interface for place-based navigation.

MUD/MOO systems are not particularly suited for this research, mainly because the navigation issue inside such systems is partly addressed by the geographical metaphors they represent. Users cannot be in more than one place at a single time, and they need to exhaustively explore the entire space in order to get familiar with it. While this may seem reasonable to some people, an approach of this kind can only be successful for small-scale systems; and even in such cases, learning through navigation can be difficult.

2.3 Internet Relay Chat Research

Internet Relay Chat (IRC) is a near-synchronous CMC architecture which allows a certain degree of rehearsability for the sender of the message – the message is sent only when the user hits the carriage return key. IRC systems provide virtual environments where people from all over the world can meet and chat, and create personal and community places. One can find a great diversity of human interests, ideas and issues in the various IRC spaces or chat-channels. There are tens of thousands of IRC channels on thousands of IRC networks at present.

Jarkko Oikarinen is considered to be "the father" of IRC. While at University of Oulu, Finland, in 1988, he started developing a communication program to increase the usability of OuluBox (a public access Bulletin Board System used at University of Oulu). Things evolved into what is known today as IRC, a client-server service for conducting multi-user, real-time chat sessions over the Internet. According to http://www.irc.org, the

birth of IRC occurred in August 1988. Since then, it has grown consistently and has proved to be one of the most popular online chat environments. During the middle of 1989, there were about 40 servers worldwide; in September 1990 this number grew to 117; and in March 1991 there were 135 servers (69 US, 66 non-US) up and running. In 1996, due to various circumstances, such as the high growth rate of the Internet, technical difficulties, and also personal differences, the initial IRC network split into two different networks. Today, there are hundreds, if not thousands, of IRC networks; and millions of people use them on a daily basis (Hinner 2000). The four largest networks, based on the number of connected servers and on the number of users, are EFNet, IRCNet, Undernet, and Dalnet (Gelhausen 2004). Some of the features offered by IRC include: nickname-based services, which provide a high level of anonymity; thousands of interaction spaces (channels) to join; the ability for users to create and register their own channels; private and public channels; moderated and non-moderated channels; group discussions; private discussions (only text-based); and direct file exchanges between users. As mentioned before, IRC is based on a client-server model. Most of the existing server programs run on a Unix-like operating system. However, there exist a variety of different client programs that allow users to connect to IRC servers. Some of them are Unix-based, some are Windows-based or MacOS-based, but the Windows-based clients are predominant. IRC uses a unique protocol that makes it possible to use the same client software to connect easily to different IRC networks.

Although one could find IRC references in areas such as education, distance education (e-learning), or general collaboration among geographically distributed teams (Pilgrim and Leung 1996; Thomas et al. 1996; Neal 1997; Last et al. 2002; Mock 2002),

such work will be excluded from the literature review. The reason is that in those cases IRC was used simply as a communication tool instead of as the main focus of the research.

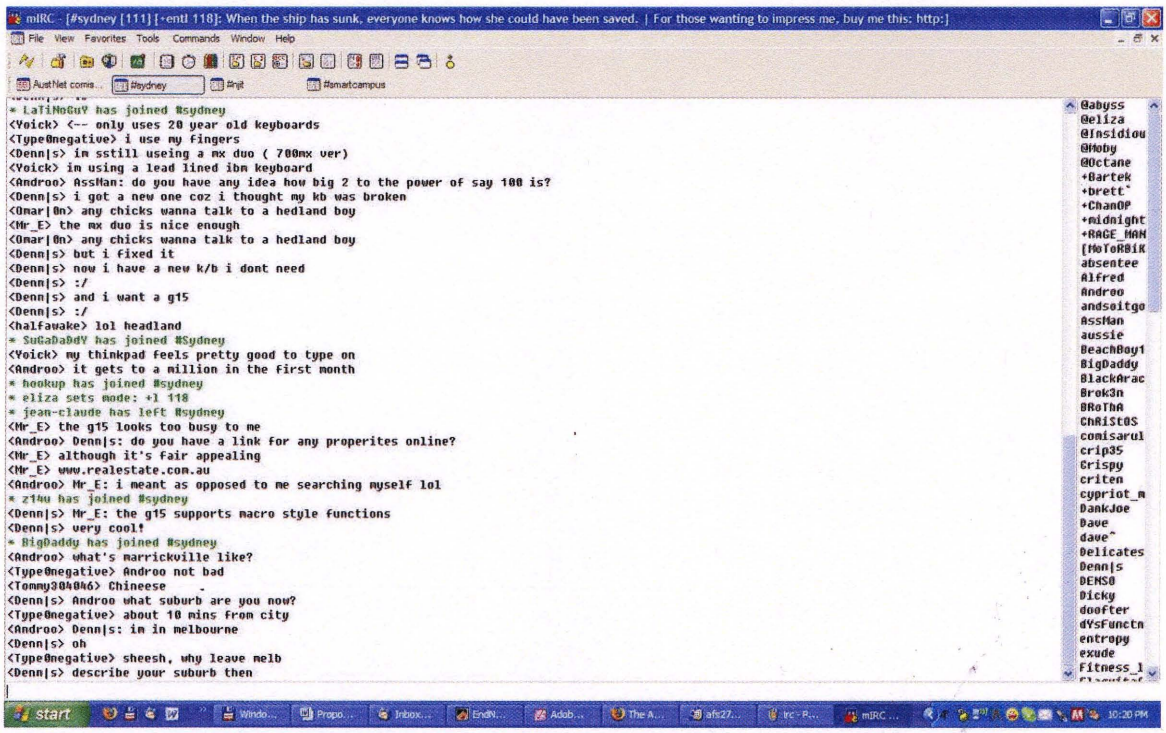

Figure 2.3 A typical IRC client software.

2.3.1 Qualitative Studies of IRC

Most studies of IRC have used an approach that can be labeled as cyber-ethnography (Pacagnella 1997) and looked at IRC from a socio-cultural perspective. They highlighted activities and behaviors that characterize IRC users (Reid 1991) and the specifics of IRC culture (December 1993; Bechar-Israeli 1996; Danet et al. 1996; Danet et al. 1997). Reid (1991) was probably the first to conduct a deep analysis of IRC from a social point of view. While she agreed that the users of IRC in 1991 were not the regular, every-day people one would meet in the streets (mostly because of technology-access issues), her work still provided important insight about the new synchronous CMC system that IRC

represented at that time. Her central thesis was that IRC users "do not shape themselves in conformity with the conventions of social contexts external to the medium, but learn to play their cultural game with them." In her opinion, by simply using the system, IRC users deconstructed the traditional boundaries of social interaction and built communities of their own. These new emergent communities were characterized by heterogeneity and were created "through symbolic strategies and collective beliefs." The author also identified some social issues associated with IRC. On the one hand, IRC offered anonymity, opportunities for gender-switching, reduced self-regulation, disinhibition, and flaming. On the other hand, the IRC community placed social sanctions that counter-balanced the effects of possible user misconducts. The author concluded that to fully understand IRC, researchers must analyze it from an interdisciplinary perspective that takes into account linguistics, sociology, communication theory, anthropology, and history.

Byrne (1994) explored the formation of relationships on IRC. Using surveys, face-to-face (FTF) interviews and log analysis, the author came to the conclusion that relationships formed on IRC had many similarities to FTF relationships, users managed to adapt to both the weaknesses and the strengths of the medium and also found novel ways to convey information and emotional meaning to other users. This socio-communication-oriented study of IRC was limited in size as only one channel was examined. This limitation is common to most of the work that has been done on large-scale CMC systems in general, and on IRC in particular.

Bechar-Israeli (1996) looked at the use of nicknames on IRC. A field study was conducted, and the author was a participant observer. Logs of conversations were

recorded for a period of two weeks and then further analyzed. An important fact that was presented in this study was that, generally, users did not change their nicknames over long periods of time. Nickname play (changing the nicknames) did occur sometimes, but in the long term, people tended to keep their nicknames, because they felt them to be an extension of themselves, a symbol of who they were. The author also mentioned the *nickserv* service offered by some IRC networks. This service allowed a user to register his or her nickname so that only he or she would be able to use it; the service also dealt with some related issues (such as intellectual property and nicknames).

Danet et al. (1996) conducted a study focused on a group of people who performed parodies of well-known theater plays on IRC (by William Shakespeare and Tennessee Williams). They examined the first production of that group – "Hamnet," a parody of *Hamlet*. They analyzed "the substantive and stylistic features of the 'Hamnet' script, the logistics of virtual production, [...the] improvisational play with the Shakespearean canon, the theater game, language itself, the IRC software, and the situation of typed online interaction" following an ethnographic method and drawing on sociolinguistics and discourse analysis.

Danet, Ruedenberg-Wright, and Rosenbaum-Tamari (1997) continued the analysis of "writing, play and performance on IRC." Using content and discourse analysis, they analyzed the text log of a virtual party that occurred inside a particular IRC channel and identified and examined three types of play: identity play, play with frames of interaction, and play with typographic symbols. An important aspect that is worth noticing and considering for future IRC research is the authors' argument that studying public IRC logs should not be ethically problematic for researchers.

December (1993) used IRC as a means to prove "the existence of an emerging discourse culture, based in computer-mediated communication (CMC) systems existing on global computer networks." Later studies focused on the language used in IRC – what did language reveal about the users (Rodino 1997), how did language affect the relationships among users (Paolilo 1999), and how did language help people overcome the limitations of synchronous computer mediated communication systems (Herring 1999).

Rodino (1997) researched interactions on IRC in order to determine the extent to which "research on face-to-face talk and computer-mediated communication can describe gender and its relationship to language". This was yet another study based on qualitative analysis of logged text over a relatively short period of time (40 minutes) in only one channel, although several were monitored. The author also used silent participant observation by lurking inside the channels. No interviews or surveys were used as part of the research.

Rintel, Mulholland, and Pittam (2001) examined how IRC users opened dyadic personal interactions. While previous research determined that IRC was an interpersonal medium, the authors argued that there was a need to understand how interpersonal relationships were formed and developed on IRC. In order to do so, they started with the analysis of openings, i.e., the start of conversations. In accordance with the previous research, the method used was log analysis using qualitative conversation analysis techniques. The authors mentioned the technical (data-collection) and ethical (private/public conversation) problems that occurred when trying to research IRC. They acknowledged that "gathering a 'complete' record of all the interactions undertaken by

IRC users would be virtually impossible." Most analysis done in relation to IRC was qualitative and ethno-methodological. Overall, this is another example of case study of the IRC, which also recognized the difficulties associated with the large-scale analysis of an entire IRC network.

2.3.2 Quantitative Studies of IRC

Paolilo (1999) developed "a social network approach to online language variation and change through qualitative and quantitative analysis of log files of Internet Relay Chat interaction." His analysis revealed "a highly structured relationship between participants' social positions on a channel and the linguistic variants they use." It seemed that a certain subset of all the members in the channel, containing a large proportion of privileged users (channel operators), was sought out for interaction by the rest of the participants. The author captured the entire activity in one channel on an IRC network for a period of 24 hours and then performed quantitative and qualitative analysis on the log. This paper followed the traditional case study IRC research pattern, which measures the activity of a small fraction of an IRC network, in this case, one channel over a relatively short time interval.

Herring (1999) examined a paradox of CMC systems: despite the limitations of such systems, such as lack of coherence due to multiple threads, turn taking, topic changing, lack of simultaneous feedback, and disrupted turn adjacency; their popularity continues to grow. While the author's analysis of IRC showed that indeed "high degree of disrupted adjacency, overlapping exchanges, and topic decay" existed; she suggested that the users' ability to adapt to this medium, together with the advantages presented by loose coherence relative to interactivity and language play were two explanations for the

growth in popularity of CMC systems. The author looked not only at logs from IRC channels, but also at Usenet groups and Listserv email lists. Results from other studies did not find any evidence for incoherence in synchronous chat, but that may be due to the small size of the groups that were studied.

In the past years there has been an increase in statistical analyses of IRC; several Web sites offer statistics of various parameters of many IRC networks (Gelhausen 2004; Hamilton 2004; Hinner 2004). Some of these results have led to published research. For example, Hinner (2000) described a method to collect basic statistics for IRC networks. He also presented those statistics (number of users, channels, and servers) in graphical format (charts). The data was collected at fixed intervals of time (typically 5 minutes) over 21 months from November 1998 until July 2000. The author used a program, commonly known as a "bot," which was permanently connected to the networks and automatically collected the required data.

Haveliwala (2002) stated that most of the Internet's major information sources have been archived and indexed, but IRC was the "glaring exception." He recognized some of the indexing challenges associated with IRC: dynamic channels, flat channel organization, high level of informality, and multiplexed threads of discussion. His goal was to archive some of the more useful technical support-oriented channels on one IRC network and to generate and index useful extracts.

Van Dyke, Lieberman, and Maes (1999) stated a general problem specific to IRC: Since there were thousands of loosely defined groups in which users could participate, finding the groups of most interest was generally problematic. There were no hierarchies for organizing channels, and there were many IRC networks. Their proposed solution

was to augment the user interface with a software agent that would alleviate the information overload problem. They developed "Butterfly" - an agent that sampled the content of IRC channels and made recommendations to the users using a keyword-based model of interest. The channels' content and the user interests were represented through a term vector with positive and negative weights. Channel sampling was done using a scheduled visiting behavior. Due to the constraints of IRC, a user could visit a limited number of channels at the same time – usually between 10 and 20. While present in a channel, the agent built the vector of keywords occurring in the conversation. A main limitation of this approach was that it took a very long time for the agent to build the content vector. Also, the agent was unable to find secret channels or to join private channels. This research showed one more time that IRC users' day-to-day activities are hampered by difficulties in finding appropriate channels. This supports the assertion that users' overall experience with IRC could be greatly improved by the development of better channel selection mechanisms.

Several key points may be extracted from the review of the IRC literature. IRC communities are extremely dynamic, but certain rhythms can be observed. This dynamicity, together with the large numbers of both users and chat-channels, often causes users fundamental problems in navigating and learning the interaction spaces. IRC chat-channels do not have an organized structure; users can be situated in several spaces at the same time and can often feel overloaded or lost. The previous attempts to improve users' navigation failed to provide significant improvements due to various factors such as short sampling time period, small number of analyzed channels, focus on the content of the discourse rather than its dynamics, or lack of attention paid to the rhythms of the IRC

activities. All of the above suggest the importance and the need for a large-scale, comprehensive analysis of the rhythms and group interaction dynamics of an entire IRC network over a long period of time.

2.4 Instant Messaging Research

2.4.1 Instant Messaging Overview and History

Instant Messaging (IM) is a near-synchronous computer-based communication process, which allows people to see if one or several chosen friends, co-workers, or associates are connected to the Internet, and to exchange real-time messages with them. In 1996, Mirablis released the ICQ (I Seek You) software and introduced the "Buddy List" concept – a list of people, friends, family, co-workers, etc., that someone would be interested in and linked to by means of synchronous dyadic chat and/or group chat. Such software later became known as Instant Messaging. In 1997 AOL introduced AIM (AOL Instant Messenger). In 1998 AOL bought Mirablis (the creator of ICQ) and became the dominant player in the instant messaging market. Soon, other players appeared such as MSN Messenger and Yahoo Messenger. The features of instant messaging software also evolved over time. Today, most of them allow video and audio communication, encryption, or some forms of asynchronous communication (mostly in the form of "offline messages" – messages that can be sent anytime and would be seen by the recipient next time he or she logs in to the system).

The features that distinguish IM systems from other synchronous CMC systems are the "buddy list" and the "awareness" information, i.e., the ability to track the current status of people on one's buddy list. Dyadic conversations are predominant in IM

systems, but group interactions are also typically supported. Other features usually offered include finding partners of discussion based on various criteria; file sharing; and stealth/invisible mode, where one is able to see other people who are online, but is not seen by them.

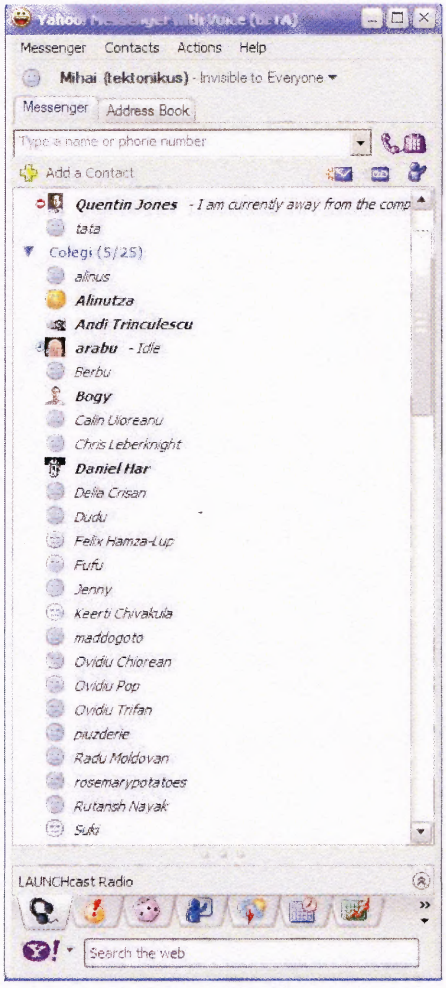

Figure 2.4 A typical IM client software.

IM systems are different from other forms of online chat in the following ways: They do not allow nickname changes unless users create a new account with the provider of the IM system; the list of potential conversation partners is much smaller than in regular synchronous chat systems, limited to either the persons in one's "buddy list" or to the users found by various search methods; and finally, various IM systems follow

different protocols, making cross-application communication difficult. Despite the predominant dyadic nature of most of the conversations that occur in an IM system, group chat is also present. Nearly all of the current IM software typically offers their users two options: (1) to start their own chat-room, which is usually ephemeral, i.e., it will not exist if there are no users inside; or (2) to join an existing chat-room. Generally, the chat-rooms follow the design of IRC chat-channels in that they provide a list of all the connected users, a public discussion space, and the opportunity to start a private conversation with any of the other users. Although the video and audio features are appealing, IM systems' group chat features are rather limited in terms of management of the chat rooms. Typically the number of users allowed inside a particular chat-room is limited. Also, although the names of the rooms are supposed to give a broad idea about the topics of conversations inside them, this is almost never the case. The requirement for users to have an account with the specific IM system in question reduces the number of potential users for the group chat features offered by that particular system. All these factors, together with the outsider's lack of knowledge about what is happening inside the chat-rooms, contribute to the rather low level of group chat usage inside IM systems.

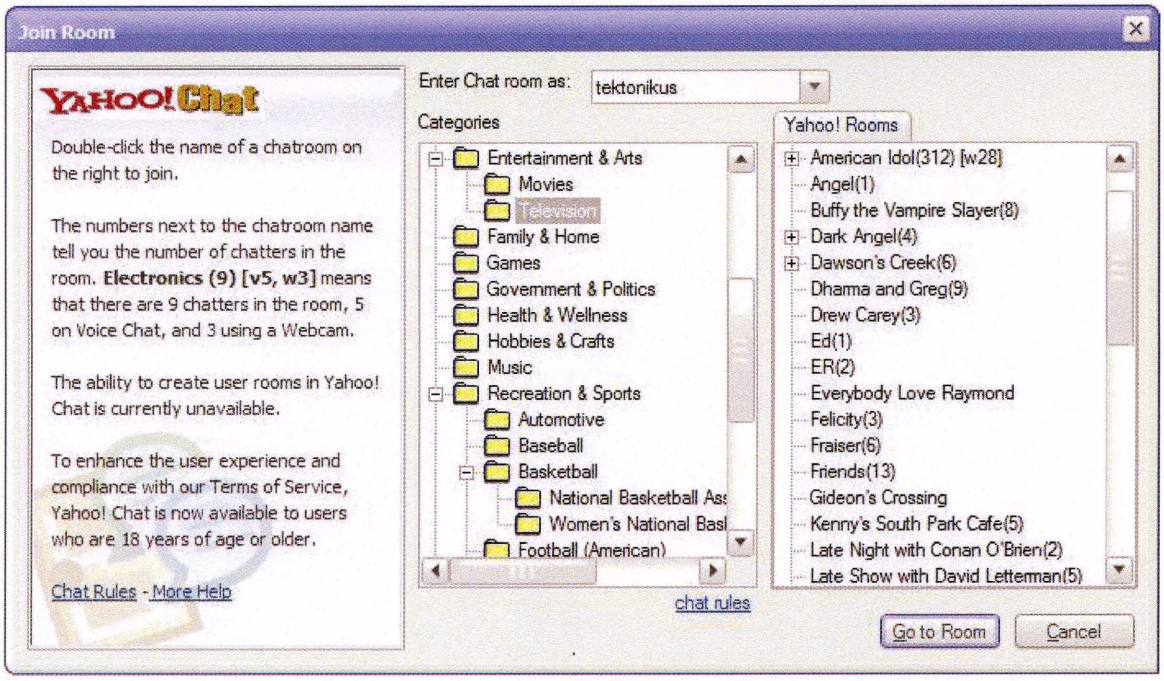

Figure 2.5 Chat-room selection in Yahoo IM client.

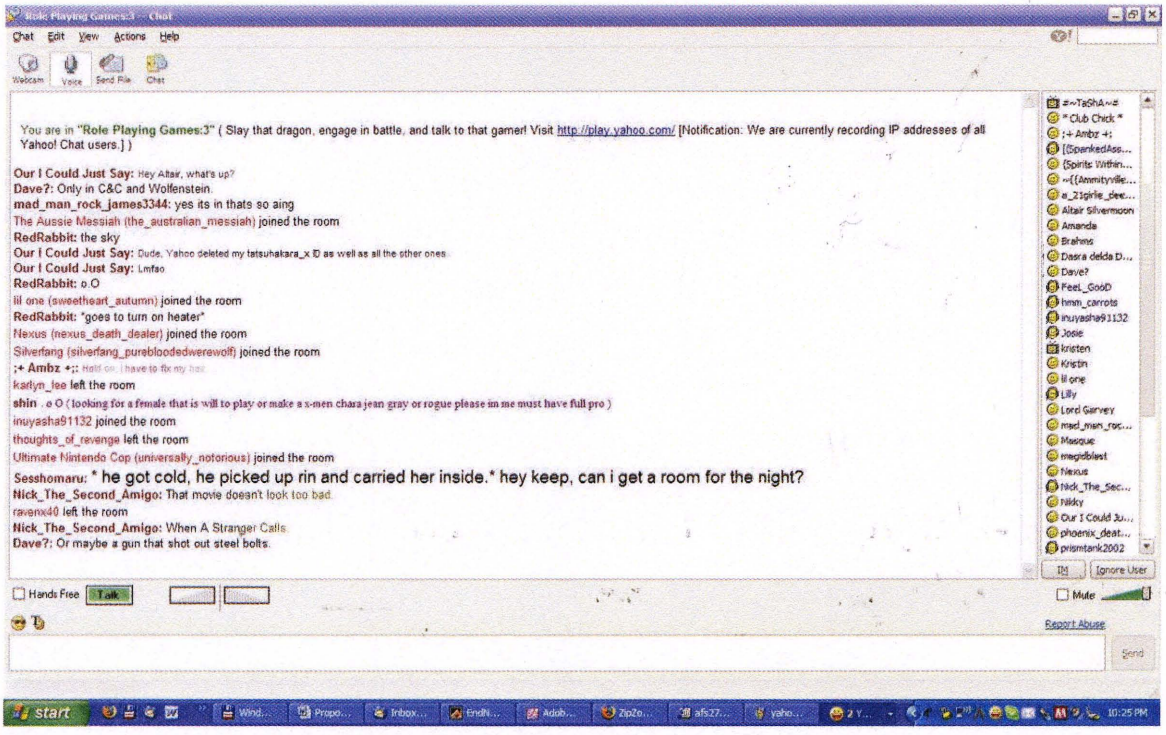

Figure 2.6 A typical Yahoo chat-room.

The current IM research follows two distinct directions. Some researchers are interested in how IM systems are deployed and used in the workplace, while others are focused on the notion of awareness and how IM systems support it. The literature review will be structured based on these two categories.

2.4.2 Instant Messaging in the Workplace

One of the first studies of IM in the workplace identified the communication tasks afforded by such systems (Nardi, Whittaker, and Bradner 2000). The ethnographic study singled out two types of usage for IM systems: interaction and outeraction. Interaction included all the communication processes that involve information exchange, while outeraction was defined as "a set of communicative processes outside of the information exchange, in which people reach out to others in patently social ways to enable information exchange." Relative to the "interaction" mode, the authors found four main functions of instant messaging. These were support quick questions and clarifications related to ongoing work tasks; support coordination and scheduling; coordinate impromptu social meetings; and keep in touch with friends and family. They also observed that instant messaging proved to be very flexible in terms of the work that it supported and also expressive, allowing affective, even intimate, communication. Relative to the "outeraction" mode, the authors determined that instant messaging was used a lot to negotiate conversational ability and establish social connections. They also observed that the concept of "awareness" – the ability of the sender of a message to have an idea about the receiver's status – was strongly related to a successful communication process. Nardi and her colleagues identified the potential of instant messaging for reducing the difficulties usually met in informal communication. Besides rapid

exchanges of information, the authors observed that instant messaging was used to hold intermittent conversations over longer periods of time. This study, widely cited by many IM researchers, provided very interesting and unique (at that time) information. However, like the majority of IM studies, it was limited in the sense that it examined only dyadic interaction, without focusing at all on group interaction.

Herbsleb et al. (2002) examined the introduction of instant messaging in the workplace, but they added to their study the analysis of group conversations (chat in the workplace). They developed their own systems, called Rear View Mirror (RVM), which provided three types of functionality: presence awareness, instant messaging, and group chat. The results showed that group chat was the most used feature, but instant messaging was also used a fair amount. The authors also examined the concept of critical mass and its relation to their systems. They defined critical mass as a group of people who were highly interested in a particular technology and led the way to adopting it. Although their intention was to quickly achieve critical mass for their system, they ended up losing about 90 percent of their potential users. This failure led to a more detailed examination of the factors that influenced the achievement of critical mass. The authors argued, "What constitutes an effective critical mass is subtle and can vary dramatically depending on different definitions of community of interest."

A follow-up study by Handel and Herbsleb (2002) built upon the foundation of their previous research on the RVM systems. They studied the experience of six globally distributed work groups that used the RVM software for a period of 17 months. The focus was on four main research questions: To what extent did the users create groups, join groups, and use group chat? What did the users talk about and to what extent did

they use flaming? Was chat used in different ways, for different purposes, at different times of day? Did different groups use the system in different ways? Hence, they were interested in both group dynamics and discourse content. Their findings showed that users tended to join groups, other than those to which they were initially assigned. Patterns of group chat use were found with succeeding periods of inactivity (usually longer) and high activity (usually shorter). The authors also observed temporal rhythms in the overall use of the systems (the active period was between 9 am and 4 pm, with peaks between 2 pm and 4 pm) for the work-related communication. Finally, the authors observed that all the groups were remarkably similar in their patterns.

Isaacs et al. (2002) examined the "character, function and styles of instant messaging in the workplace." The authors of this study wanted to examine whether the current knowledge about IM usage still held in the context of larger numbers of users (previous studies did not have many subjects). The most important finding was that most of the time people were using instant messaging for work-related tasks, contradicting in this way the perception that IM was commonly used for social interaction in the workplace. The authors also defined two very different styles of use for instant messaging: working together (multi-purpose communications that covered a range of complex activities, were intense, of short duration, and sometimes threaded) and coordinating (short, single purpose communications, minimum threading, and slow-paced).

Another study by the same researchers (Isaacs et al. 2002) determined that "frequent IMers have longer, fast-paced interactions with shorter terms, more threading, and more multitasking relative to infrequent users. Pairs who IM each other often also

have longer interactions than do pairs who interact rarely." The authors argued that "IM integration with voice or video connections may be less important than previously thought."

Hansen and Damm (2002) researched instant messaging in the context of session management in distributed collaboration tools. They examined how people who were using Knight, a tool that supported co-located, collaborative object-oriented modeling, were also using instant messaging software during their interactions. After on-site observations and transcript analysis and interviews, they identified the need for an instant messaging component to be integrated into the initial system, instead of using other software for the purpose of synchronous communication.

Muller et al. (2003) provided the first study of IM based on large samples of user reports as opposed to the previous studies that were based on the analysis of small samples of server logs or on ethnographical methods. The method used was a "survey-based self-report research into people's use of Sametime," an instant messaging software provided by Lotus. The results showed significant decreases in the use of other communication channels. Regarding the use of IM, the results were contradictory to previous IM studies, but the authors' opinion was that in the two-year interval since the first IM studies, the users of IM systems have found additional value in IM.

2.4.3 Awareness and Instant Messaging

Awareness of other people's presence in distributed work groups has been a frequent research topic in the past years. In order for such groups to be efficient and effective, there is a strong need for good coordination. Coordination of a process can be achieved only when every person involved in that process is aware of the status of his or her co-

workers. The early research done in this area focused mostly on audio and video technologies for determining people's status. However, researchers had limited success mostly because of technology issues. More recently, systems that used alternative approaches such as pictorial representations (Isaacs, Tang, and Morris 1996) or line drawings (Greenberg 1996) have been developed.

A more sophisticated system that provided awareness of presence information was ActiveMap (McCarthy and Meidal 1999). The users wore "active badges" – small electronic devices that transmitted the wearer's location to a central server. The position of every active badge could be displayed on the screen of a computer running an application designed especially for this purpose.

Instant messaging systems were the next step in providing awareness information. As the importance of this concept grew, more research has been done in order to find ways to improve others' awareness of an individual without invading the individual's privacy.

Ljungstrand and Hard (2000) described a system called WebWho – a Web-based awareness system that kept track of people's presence. In this case, the location was a computer lab at the university where the system was developed. The users of the system were able to find out the exact location of a particular person, a friend, a colleague, or a co-worker, and also to find unoccupied computers in the lab. It also provided the function of sending anonymous or non-anonymous instant messages to other users. According to the authors, the main difference between WebWho and other awareness systems was that WebWho was "primarily place-centered and only secondary person-centered." The aim of this study was to determine if and how the awareness of presence

affected the content of the instant messages sent among users. They correctly hypothesized that the awareness information would influence the content of the instant messages, but they did not specify in what way.

Tang, Yankelovitch, and Begole (2000) prototyped an instant messaging system called ConNexus (from *Contact Nexus*) that was designed for use in the workplace. Based on research observations, own experience, and conversations with various people; they decided to focus on three design implications for IM in the workplace: (1) provide various awareness tools; (2) integrate IM with other communication media; and (3) design a more natural method for the user interface to support starting, maintaining, and ending conversations. They argued that the awareness cues should be extended in the sense that they should provide information about the activities the users were engaged in, not only about their current location or status. Early results from user experience showed awareness information to be a critical factor in IM.

Tang et al. (2001) continued the work on ConNexus. They developed a series of prototypes intended to facilitate communication by using awareness information. The first prototype of ConNexus was desktop oriented and provided features such as contact lists similar to buddy lists found in most instant messengers, contact toolbars containing detailed information about a particular contact in the contact list, and a tailored set of communication tools appropriate to each member of the contact list. The awareness information provided by this system for a person included online presence of the person, idle time of input devices (mouse and keyboard), and current level of engagement in computer mediated communication activities. Considering the proliferation of mobile devices, the authors extended the system to suit the needs of mobile users and built a

prototype called Awarenex. At the time of publication, several studies were being conducted. However, the authors clearly stated that only longitudinal field studies would give them the answers to the many research questions raised by such a prototype.

Tang and Begole (2003) stated that "IM features and uses are still evolving" and "effective communication requires richer awareness information of current and future reachability, context and availability." They argued that "awareness services" would be a must in the future design of IM systems and presented three research prototypes as solutions for these awareness services: (1) Awarenex, a system that integrated real-time awareness information in order to provide cues about the opportune time to initiate, maintain and end contact; (2) Rhythm Awareness, a prototype that predicted a person's presence and use patterns based on that person's history with the system; and (3) Lilsys, a system that provided awareness information taken from various sensors (sound, phone usage, and computer activity).

Hubbub (Isaacs, Walendowski, and Ranganthan 2002) was a "sound-enhanced mobile instant messenger that supported awareness and opportunistic interactions. The authors examined some of the limitations of the traditional instant messaging systems such as limited awareness information, impossibility of sending messages to persons who were offline, and impossibility of logging on from different locations at the same time. They decided to build a tool that would "provide awareness information among distributed groups, encourage opportunistic conversations, allow people to stay connected as they move among multiple fixed locations [...] and be readily available and easily installed." They developed Hubbub for both desktop computers and mobile devices. The studies they conducted revealed that various sound features of the system improved

people's awareness about other users and increased the number of opportunistic interactions.

Results from another experiment involving the Hubbub system were presented in the work of Isaacs, Walendowski, and Ranganthan (2002). As mentioned above, their goal was "to create a system that would encourage opportunistic interactions and support background information while recognizing the fluidity of people's movement throughout the day." There were 25 participants, and 300 conversations were logged over a 5-month interval. The results showed that Hubbub's features helped people feel more connected with the rest of the group, even to people with whom they would not have normally interacted, and gave them a big sense of being a part of the community.

2.4.4 Other IM Research

While most of the research focused on IM's use in the workplace and on the awareness cues it could provide, there were a few studies that looked at some pure social aspects.

Voida, Newstetter, and Mynatt (2002) discussed some findings they obtained from observations, interviews, and textual analysis of IM log files. They considered IM to be a hybrid genre – a combination between written and verbal communication. They identified five types of tensions that are usually associated with instant messaging: persistence and articulateness tensions, synchronicity tensions, turn-taking and syntax tensions, attention and context tensions, and availability and context tensions.

Grinter and Palen (2002) explored IM as "an emerging feature of the teen life," paying attention to the everyday use of IM and its support for interpersonal communication. The distinctive characteristic of their research was the age of the subjects. While most of the studies were oriented toward adult participants, only

teenagers participated in this experiment. The authors argued that IM communication occurred between "real space friends," that peer pressure was an important factor that affected the embracement of the technology, and that choosing IM over other media was not only an effect predicted by the media richness but was also influenced by other constraints. Domestic rhythms and schedules, as well as privacy issues, were important determinants for teenagers' usage of IM.

In a somewhat related study, Grinter and Eldridge (2003) looked at how teenagers were using Short Message Services (SMS) text messages with their mobile phones. While a comparison between IM and SMS would not be appropriate here, it is worth noting that sending text messages via mobile phones has become a common practice in many countries. This study showed that two of the three primary activities that characterize teenage IM use (Grinter and Palen 2002) also hold in the case of mobile phone text messaging (social chatting and coordinating activities). A third activity discovered here was coordinating communications.

Alvestrand (2002) provided a brief but comprehensive description of today's IM concepts and identified community building to be the most important benefit of such systems.

Chuah (2003) argued that today's IM systems are not anchored enough into the real world. He suggested that a combination between IM software and live streaming of news or sports events would provide a better context that would improve the IM users' experience.

The literature review of IM research highlights several aspects relevant to this research. First, the patterns of dyadic IM conversations differ from those of group

conversations mostly because dyadic interactions occur much more frequently than group interactions inside instant messaging systems. Second, IM interactions are often characterized by certain rhythms, with distinctive periods of activity and inactivity and the rhythms of group interactions seem to be better defined than the rhythms of dyadic interactions. Third, the users of these systems generally appreciate awareness information about the status of their conversation partners.

IM systems are not suited for this research for several reasons. IM conversations, whether dyadic or in groups, are conducted almost exclusively with friends, co-workers, or other acquaintances. Although some systems offer people search capabilities, finding new interaction partners is more difficult when compared, for example, to a system like IRC. Thus, the limited number of potential interaction partners, the inability to easily look for new people with whom to communicate, and the lack of pre-defined interaction spaces for mass interaction make IM systems an unsuitable candidate for this work. The group chat features offered by some IM systems are used to a certain extent and, since their structure is very similar to that of IRC, could represent an alternative. However, the limited chat-room openness and management capabilities, together with the potential data-collection problems associated with proprietary implementation specifications, make IRC a more attractive medium.

2.5 Research on Various Other Chat Systems

Vronay, Smith, and Drucker (1999) identified several main factors that influenced chat efficiency. These factors included the lack of all of the following: recognition, intention indicators, status information; and context. Additional factors were the high signal to noise ratio, typing inefficiency, and the general uselessness of the chat history. While acknowledging the usefulness of thread management research, they focused on the problem of eliminating the conditions that actually led to overlapping threads. They designed and tested two prototypes of new chat user interfaces, but the initial results were disappointing – the users clearly preferred the traditional chat interfaces and resisted the new ones.

Smith, Cadiz, and Burkhalter (2000) described a chat client prototype, Threaded Chat, as a solution to the current chat systems' problems including deficiencies in managing interruptions, managing turn-taking, and conveying comprehension. The system supported a synchronous form of the turn-taking structure characteristic of asynchronous systems such as Usenet newsgroups. The conversations were organized into structures called threads. Each thread started with a "turn" and continued with several "responses." The system was used in an experiment along with a regular, traditional chat system and another experimental prototype. The users rated it significantly lower than the regular chat system. However, the level of task performance was about the same for both systems. Overall, the subjective user ratings for Threaded Chat were very low, but people were able to adapt to the interface and complete the tasks.

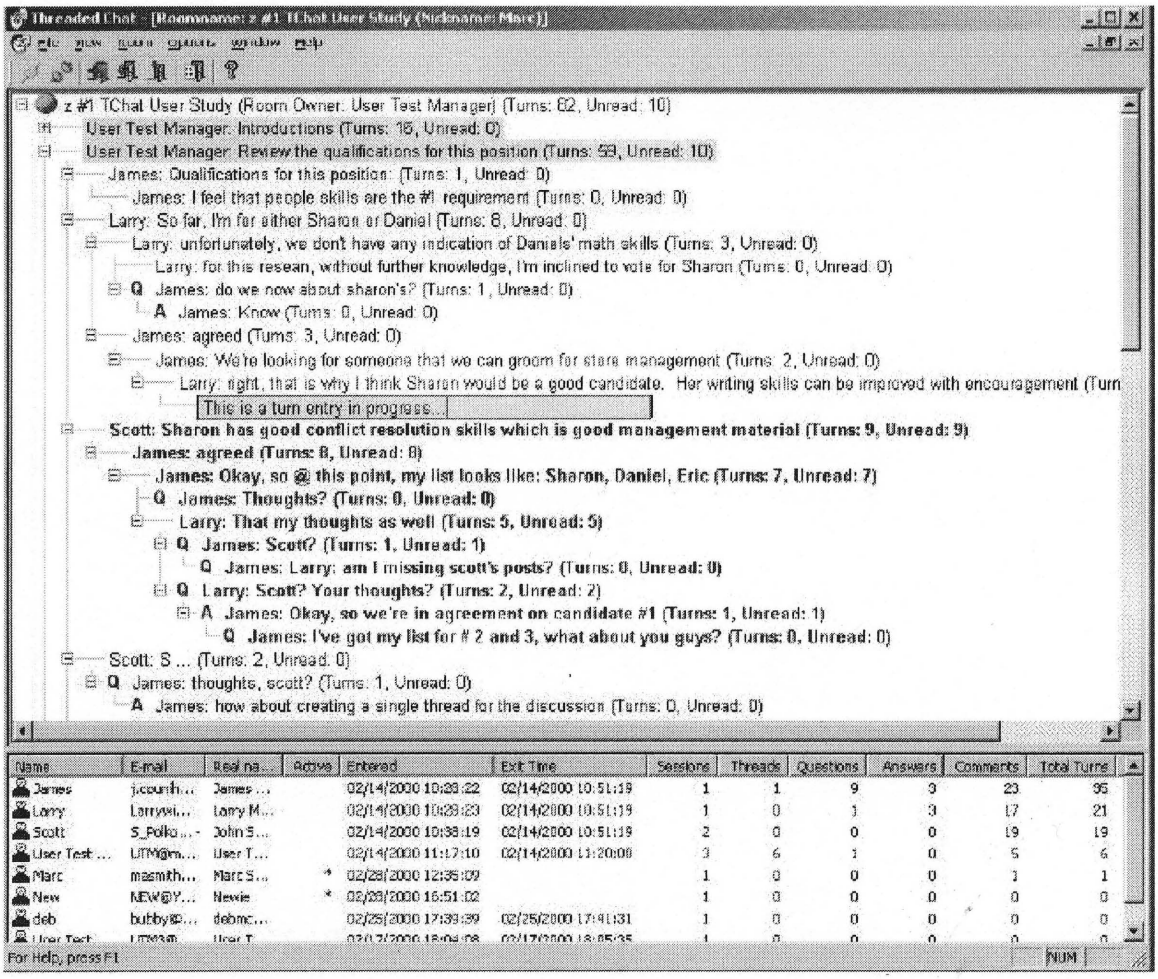

Figure 2.7 The Threaded Chat prototype.
(Source: Smith, Cadiz, and Burkhalter 2000)

Smith, Farnham, and Drucker (2000) performed log file analysis of user behavior in order to illustrate the dynamical structure of social cyberspaces. The paper provided a quantitative analysis of the social dynamics of three chat rooms in the Microsoft V-Chat graphical chat system, a system relying on IRC infrastructure for communication transport, but providing some additional graphical features such as avatars, gestures and positioning relative to the other users. Data on usage patterns and online social interactions were collected using surveys and data logs. The authors clearly stated that very little was known about the social interactions occurring within chat spaces. Log

analysis was seen as a useful complement to ethnographical studies, providing a wide range of measures related to the social structures and the dynamics of the interactions in the medium. The initial findings showed the noisiness of log data, a common problem when dealing with large amounts of data, but they also showed that chat rooms had well-defined rhythms, (which were active mostly in the afternoon).

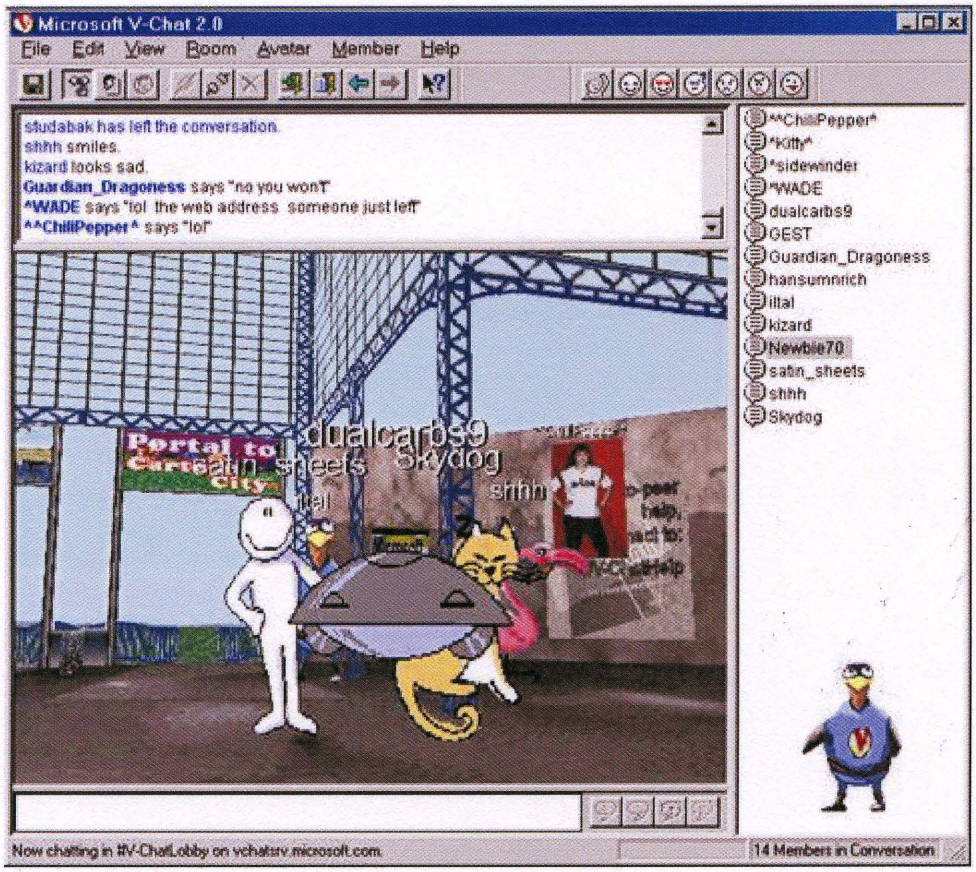

Figure 2.8 Microsoft's V-Chat system.
(Source: Smith, Farnham, and Drucker 2000)

Ribak, Jacovi, and Soroka (2002) developed ReachOut, a "chat-based tool for peer support and community building," trying to improve the distribution of knowledge within business organizations. In doing so, they drew from many other types of CMC systems: newsgroups, Listserv, email lists, and instant messaging. By adding persistency features to a synchronous environment, the authors considered that their system actually

bridged the gap between asynchronous and synchronous systems. Their main design philosophy was that users should turn to peers, instead of looking to acquire some specific information from external sources such as the Internet or a company's intranet. The users would ask questions; each question would have dedicated a chat-room, and the content of all the ongoing discussion in the chat-room would be made persistent and accessible in the future. In this way, the authors argued that not only would people find the necessary information, but their personal social networks would also expand and the entire community would grow. Early results from pilot testing showed encouraging results and ReachOut continued to be extended and further tested.

Jacovi, Soroka, and Ur (2003) presented results drawn from the use of ReachOut in the workplace. They defined six prospective functions that a semi-synchronous chat tool should perform: provide peer support; accelerate decision-making; announce events; facilitate socializing; coordinate teams; and contact colleagues, friends, and family. The authors found supporting evidence for these functions in the evaluation of ReachOut usage over a two-month period at the IBM Haifa Labs. The results proved that all the six functions were fulfilled. Furthermore, two additional functions of the ReachOut system were discovered: It alerted management about an existing problem and raised consciousness for a large group of people about an issue concerning them. Overall, the authors considered ReachOut to be a tool that could be successfully used in the workplace in multiple ways.

Persson (2003) argued that despite the research on avatar-based systems, most synchronous communication was done through text. After enumerating some of the potential causes to support his contention, he proposed a shift from synchronous to

asynchronous use of avatars and presented a messaging system called ExMS that made extensive use of avatars.

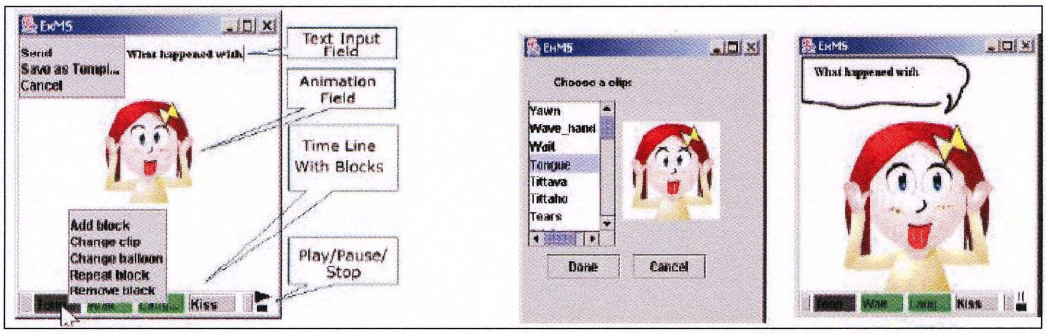

Figure 2.9 The ExMS prototype.
(Source: Persson 2003)

Avatars were used mostly to represent expressiveness such as feelings, reactions, and moods, rather than spatial movement. The system was tested on a very small number of subjects (10 subjects). The conclusion was that a rich, well-designed avatar system could be a useful tool for "quick and dirty" messaging among people.

Kurlander, Skelly, and Salesin (1996) described a system that represented online communications in the form of comics, called Comic Chat. They argued about the importance of improving the text-based chat rooms with graphical features and presented a few of the problems of current avatar-based chat systems. In order to develop their system, they collected and examined numerous chat transcripts, annotated them, presented them to a comic book artist, who then provided them with an illustration. They identified three design elements categories that required automation: characters, balloons, and panels. The paper described in details all the algorithms they used for building the comics. They used the IRC protocol as the foundation of their application as Comic Chat connected to existing IRC servers on the Internet. Overall, the authors were pleased with

the results, and people enjoyed using the system, but they agreed it still needed many improvements.

Dewes, Wichmann, and Feldman (Feldman 2003) developed a methodology to identify the network traffic generated by various synchronous chat systems and to determine the statistical properties of such traffic. They did their analysis by monitoring all the TCP packets sent through their university's network and by extracting the packets that were identified as chat traffic based on various strategies and algorithms. The results showed their methodology was quite successful. Less than 8.3 percent of the existing chat connections were lost during the analyzed period (one week).

None of the systems described above represents a good match for the analysis of mass interaction because they are mostly prototypes and used by a limited number of users.

2.6 Summary

The literature review of the three main types of synchronous CMC systems reveals several important facts that need to be considered. Firstly, it suggests that synchronous CMC systems are sometimes characterized by rhythms and patterns. The work that has been done on rhythms and patterns identification shall be further review in Chapter 4, as these two notions are of highest importance in analyzing the interaction dynamics and determining the group interaction trajectories of large-scale synchronous chat-systems. IRC's uniqueness, when compared to IM systems or to MUD/MOO systems, stems from its dynamicity and its openness relative to the opportunities to meet new people.

Secondly, it can be inferred from previous research that MUD systems and IM systems are not well suited for large-scale research of group interaction dynamics. MUDs seem to have lost researchers' attention, which might suggest they do not represent a major choice for users looking for online synchronous group interactions. Also, since MUDs typically use a spatial metaphor, users cannot be in two or more places at the same time. Additionally, users have to learn the space through navigation, one step at a time, which is a constraint to the research of mass interaction. IM systems typically seem to be used for dyadic conversations with friends or with other known persons. Although group chat features exist and are used to a certain extent, the implementation protocols and the limited chat-room management capabilities would make the identification of group chat dynamics and trajectories problematic from a technical point of view as well as from the perspective of relevancy and generalizability of the results. Furthermore, the number of potential interaction partners in IM systems is limited compared to IRC, and large-scale data-collection is problematic due to the implementation specifications and protocols used by IM systems.

By a process of elimination, the above discussion suggests that IRC is an ideal choice for the proposed research. It is one of the oldest synchronous CMC systems; and despite its simplicity, it is still extremely popular, being used by millions of people on a daily basis. Many of the current chat systems, including the group-chat features of IM systems, were designed using IRC as a model. Finally, previous researchers have emphasized the difficulty of collecting large amounts of data from IRC. Although this is typically true, a method that makes the data-collection process more convenient has been identified and will be applied.

A better design for large-scale synchronous chat systems can only be achieved by acknowledging the current problems the users are facing and then working to solve these problems. Current IRC research presents proof of user problems, such as information overload or inability to find meaningful interaction spaces and/or partners. To start addressing these problems, a detailed analysis of rich, large-scale synchronous chat systems interaction dynamics data sets would be required. The review of the literature clearly identified the lack of such rich data sets. The sampling periods of data sets were typically short; small numbers of channels were observed; the analysis was based mostly on discourse content; and almost no attention was paid to the rhythms of the chat spaces.

CHAPTER 3

SOCIAL RECOMMENDER SYSTEMS

3.1 An Introduction to Recommender Systems

This chapter reviews various tools that help people find their way inside large-scale spaces. Currently, the navigational issues associated with such spaces have been addressed through two main types of tools: social recommender systems and social visualizations. Social recommender systems can be further divided into social matching systems and expertise recommender systems. Social visualizations, although they don't provide recommendations *per se*, help the users to get a better understanding of the activities that characterize the chat spaces and provide support for the users' decision-making processes relative to future interactions inside the chat spaces.

Recommender systems are software tools that attempt to help users satisfy their needs when choosing among various products, services, or other items by providing recommendations based on various algorithms. They represent one approach in dealing with the abundance of or the absence of information. Terveen and Hill (2001) provide a review of the domain, identifying four types of recommender systems: content-based systems, recommendation support systems, social data mining systems, and collaborative filtering systems.

Content-based recommender systems attempt to recommend items that are similar, to some degree, to items previously preferred by the users. They typically use the preferences of the information seeker and attempt to recommend items similar to other items that the user liked in the past. They focus mostly on "algorithms for

learning user preferences and filtering a stream of new items for those that most closely match user preferences."

Recommendation support systems do not provide recommendations by themselves. Instead, they only offer the means for people to share recommendations. They "serve as tools to support people in sharing recommendations, helping both those who produce recommendations and those who look for recommendation."

Social data mining systems try to make recommendations based on the history of users' social activity. By mining various records of social activities such as Usenet postings, emails, hyperlinks, system usage logs, etc., these systems are able to provide useful information to their users and help them find interesting material or interesting people, improving in this way the social navigation.

Collaborative filtering systems attempt to recommend user items that were previously liked by other people with similar tastes or preferences. They "require recommendation seekers to express preferences by rating a dozen or two items, thus merging the roles of recommendation seeker and preference provider." The focus of such systems is "on algorithms for matching people based on their preferences and weighting the interests of people with similar taste to produce a recommendation for the information seeker."

Out of these four types, content-based and collaborative filtering recommender systems are the most widely used. They are sometimes combined into hybrid recommender systems. Such systems typically employ the users/items/ratings model where a rating function is mapped from each user/item pair to some rating value (Terveen and Hill 2001). Lately it has been argued that the users/items/ratings model might not

suffice when dealing with complex and/or dynamic domains. As a result, a multidimensional approach to recommender systems was suggested, where multiple variables are taken into account when building the recommendation algorithm (Adomavicius and Tuzhilin 2001).

3.2 Social Recommendations

In recent years there has been an emergence of recommender systems that focused specifically on social recommendations. As opposed to the traditional recommender systems that were used to recommend and/or sell various products or services such as books, movies, Usenet news, vacation packages, etc. (O'Connor et al. 2001; Schafer, Bowman, and Carroll 2002; Miller et al. 2003), social recommendations focus on different social aspects characteristic to online communities. Social recommendations could be aimed at the individual, trying, for example, to find a match or to increase the size and the strength of a person's social network; or could be aimed at the community, trying, for example, to increase its size and stability over time. There have been two main approaches to provide users of CMC systems with social recommendations: a more direct approach through the use of social matching systems, and an indirect approach through the use of social visualizations.

Social matching systems are a type of recommender systems that "[partially] automate the process of bringing people together" (Terveen and McDonald 2005). In other words, social matching systems attempt to recommend people to people. Terveen and MacDonald argued that such systems "have the potential to increase social interaction and foster collaboration among users within organizational intranets and on

the Internet as a whole." However, despite these potential benefits, they recognized a lack of research in this domain. They tried to precisely define the scope of social matching systems, and, at the same time, describe how they were different from other types of recommender systems. The authors provided a review of the research on social matching and other related systems, identifying a variety of approaches. Social recommenders for information needs attempted to match people "based on their social relationship and an information need." The "Expertise Recommender" (ER) (McDonald and Ackerman 2000; McDonald 2001) was an example of this class of systems. Its authors acknowledged that "locating the expertise necessary to solve difficult problems is a nuanced social and collaborative problem" and finding a person with this expertise could be difficult especially inside an organizational setting. They developed the ER system to facilitate the identification of individuals who possessed the expertise required to solve a particular problem. They conducted fieldwork and discovered sets of heuristics that they further implemented in the algorithms of their system. Expertise recommender systems could be tailored to different recommendation situations (such as different organizations), but they would rely mostly on "profiling techniques that are not common to other recommendation systems." Specifically, great amounts of fieldwork in organizational settings would be required in order to generate the people profiles needed for expertise recommender systems to work.

Terveen and McDonald continued their review with a description of information systems with implicit social mining, which focused on "navigating information spaces with the goal of finding desired facts." They also provided users with pointers to other users who could help when the needed information could not be extracted through the

social mining rules. Opportunistic social matching systems (Svensson et al. 2001; Cohen et al. 2002) are a class of systems in which matching is "based on shared interests, where users' interests are inferred by the system from their current activity or record of past activity." Other approaches related to social matching systems include user modeling systems (Rich 1979), group recommenders (O'Connor et al. 2001), online communities (Preece 1999), awareness systems (Erickson et al. 1999), social visualization (Donath, Karahalios, and Viegas 1999; Smith 1999), and social navigation (Whittaker et al. 1998). After reviewing the social sciences literature relevant to the design of social matching systems, the authors presented a research agenda for this area in the form of a set of claims, and argued that testing those claims should provide vast research opportunities in the future.

> **"Claim 1:** Social matching systems need to use – and users will be willing to supply – relatively sensitive personal information.
>
> **Claim 2:** Social matching algorithms necessarily embody a model of what makes a good match; making that model explicit leads to better matches.
>
> **Claim 3:** Social networks are a useful tool for social matching. While w*hole* (population-based) networks are problematic, *egocentric* (user-centered) networks offer several promising uses and raise interesting research challenges.
>
> **Claim 4:** Creating effective introductions between users is crucial, but requires balancing the effectiveness of the introduction and the disclosure of personal data.
>
> **Claim 5:** Size does matter for a social matching system, but not as much as you might think.
>
> **Claim 6:** Designers must consider possible contexts of interaction between matched users.
>
> > **6a:** Properties of online spaces constrain the possibility for developing interpersonal relationships and group ties.
> >
> > **6b:** Interacting physically offers greater rewards and risks than interacting in a virtual space; when this is an option, systems must support users in exercising this option safely.
>
> **Claim 7:** User feedback for a social match must be relative to a specific role or context; obtaining feedback is much harder than getting user ratings for books, movies, music, etc.

Claim 8: Evaluations of social matching systems should focus on users and their goals."

Although, according to Terveen and McDonald, the main focus of social matching systems is to recommend people to people, one can easily envision the enhancement of such systems to include recommendations about particular interaction spaces to people as a distinct step in the process of bringing people together. Such an approach to social recommender systems was taken by Van Dyke, Lieberman, and Maes (1999). They stated a problem that was common to many large-scale CMC systems, i.e., the existence of thousands of loosely defined groups in which users could participate over the Internet and cited the problems that typically occurred when looking for the most interesting groups. Since no hierarchies for organizing the spaces where these groups typically meet existed, they proposed as a solution the augmentation of the user interface with software agents that would alleviate the information overload problem. Specifically focusing on IRC, they developed "Butterfly" – a software agent that sampled the content of IRC channels and made recommendations to the users using a keyword-based model of interest. The channels' content and the user interests were represented through a term vector with positive and negative weights. Channel sampling was done using a scheduled visiting behavior (due to the limitations of IRC, a user could visit a limited number of channels at the same time – usually between 10 and 20). While present in a channel, the agent built the vector of keywords occurring in the conversation. A main limitation of this approach was that it took a very long time for the agent to build the content vector. Also, the agent was unable to find secret channels or to join private channels. The Butterfly system was oriented toward recommending spaces (chat-channels) to people, as

opposed to the expertise recommender systems or the social-matching systems, which attempted to recommend people to people.

3.3 Social Visualizations

Social visualizations are types of information visualizations that focus on analysis and display of social behavior, typically online social behavior. Such visualizations can offer a glimpse of the underlying social structures associated with online spaces (Donath, Karahalios, and Viegas 1999) and typically display the patterns of user activity (bursts, idles, evolution of the topic, evolution of the users) (Erickson et al. 2002). They can be produced around a "social proxy," i.e., "a minimalist visualization of people and their activities" or at a more aggregate level displaying social network structures (Mutton 2004). By informing the user, such visualizations act as an indirect social recommendation interface.

Social visualizations have been used to map the activities occurring inside various synchronous and asynchronous CMC systems. Viegas and Smith (2004) presented two tools for visualizing activities in Usenet groups: Newsgroup Crowds and Author Lines. Newsgroup Crowds was more group-oriented and represented graphically the population of posters (active users) in a particular newsgroup. Author Lines was oriented toward the individual users and depicted graphically a particular poster's activity across all the newsgroups in which he or she was active. As opposed to the rest of research in this domain, this was a large-scale study in terms of both time and size. It was conducted over a period of one year, and it sampled a large number of Usenet groups. Both these tools helped reveal well-defined temporal patterns of activity. Specifically, relative to the

group activity, the author measured the number of posters inside groups and the overall activity of these posters within each of the analyzed Usenet groups. Relative to the individual activity for each user, the authors computed the number of threads of conversation started by that user, as well as the number of threads to which that user responded across the entire sample of analyzed groups. The authors suggested that such visualizations could be used for future interfaces that would better convey information about the history of the social dynamics inside Usenet groups, leading to better selection and evaluation of content inside newsgroups.

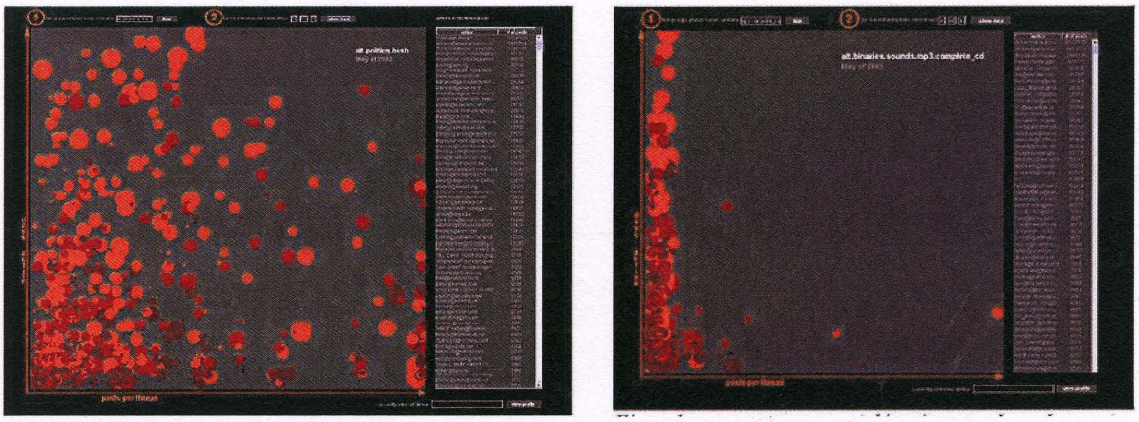

Figure 3.1 Newsgroup Crowds visualization of two newsgroups.
(Source: Viegas and Smith 2004)

Fiore and Smith (2002) used treemaps (Shneiderman 1992) to visualize Usenet newsgroups and concluded that interfaces that implement treemaps would help in the exploration of large-scale social dynamics of CMC systems.

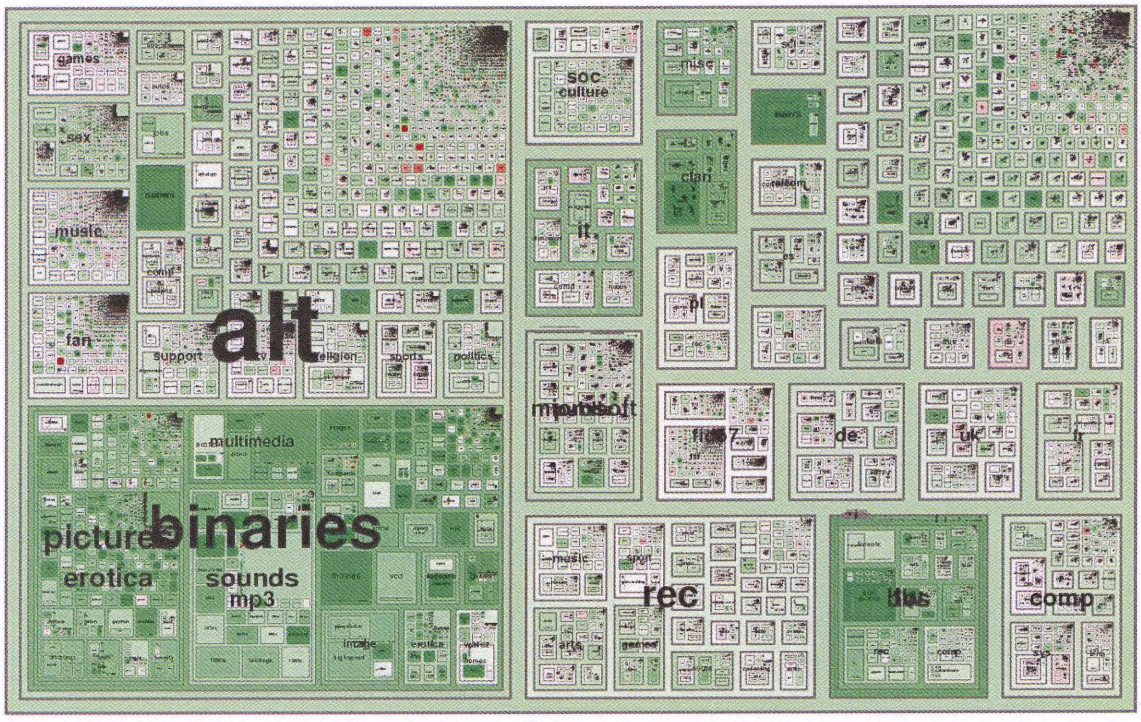

Figure 3.2 TreeMap of all Usenet, March 2000.
(Source: Fiore and Smith 2002)

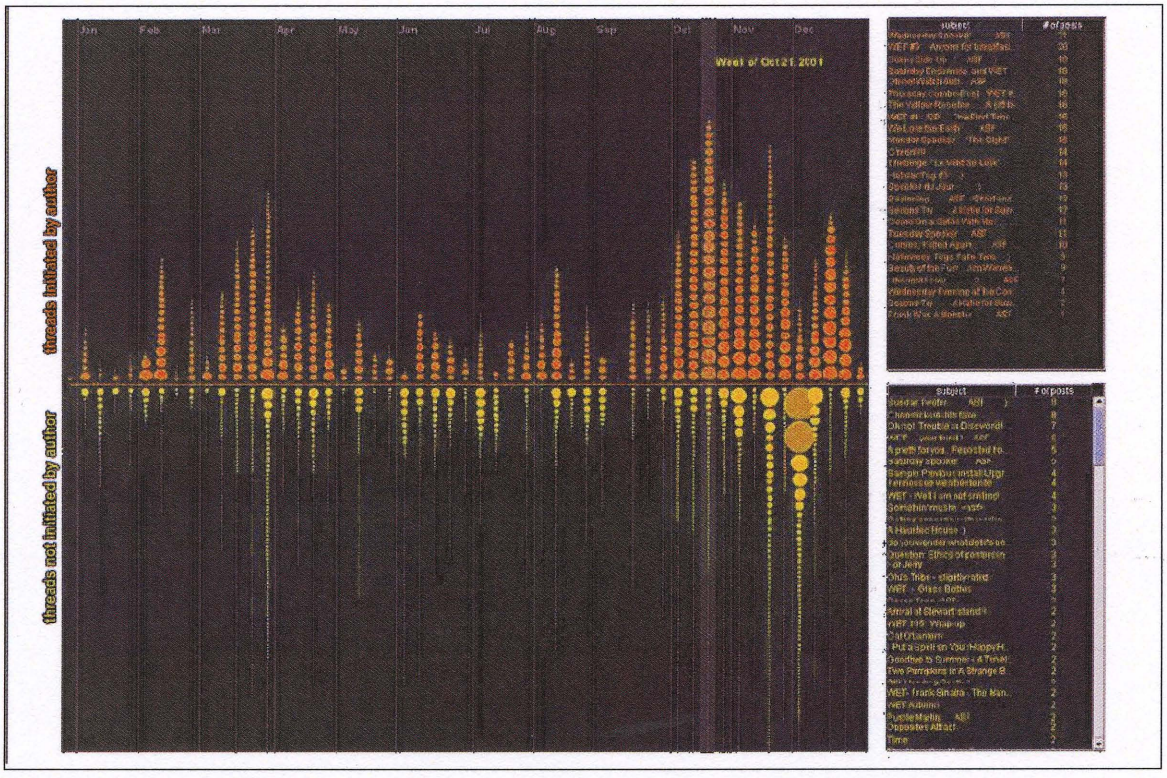

Figure 3.3 AuthorLines visualization of started/responded to threads.
(Source: Viegas and Smith 2004)

Donath, Karahalios, and Viegas (1999) discussed the design of graphical user interfaces that made visible the social structure of the interaction within chat spaces by displaying the patterns of activity such as bursts, idles, evolution of the topic, and evolution of users. They identified the users' need to access such information in a chat room. They took some steps into this direction, by providing visual information about the dynamics of the interactions. They focused on creating representations for highlighting the social information that people needed in order to figure out what was happening in the chat space. They actually built two tools: Chat Circles, which was a graphical user interface for synchronous communication; and Conversation Landscape which was a tool for visualizing the history of the interactions. The authors also referred to Loom, a tool used for visualizing asynchronous systems (Usenet groups). Whether visualizing the activity of synchronous or asynchronous mediums, it is worth noticing that the authors identified the people's need to actually see what was happening in the interaction space in order to improve their experience with the system.

Viegas and Donath (1999) continued their research in designing a "chat system that uses abstract graphics to create a richer, more nuanced communicative environment." They tried to display the dynamics of the conversation and the patterns of interaction and activity that emerged in synchronous conversations. A distinctive feature of their tool was that the users were able to visualize the overall picture of the system, i.e., they were able to see all the other participants. There were no multiple chat-rooms, just one single interaction space. The users were represented as colored circles whose sizes changed according to their activity (amount of participation).

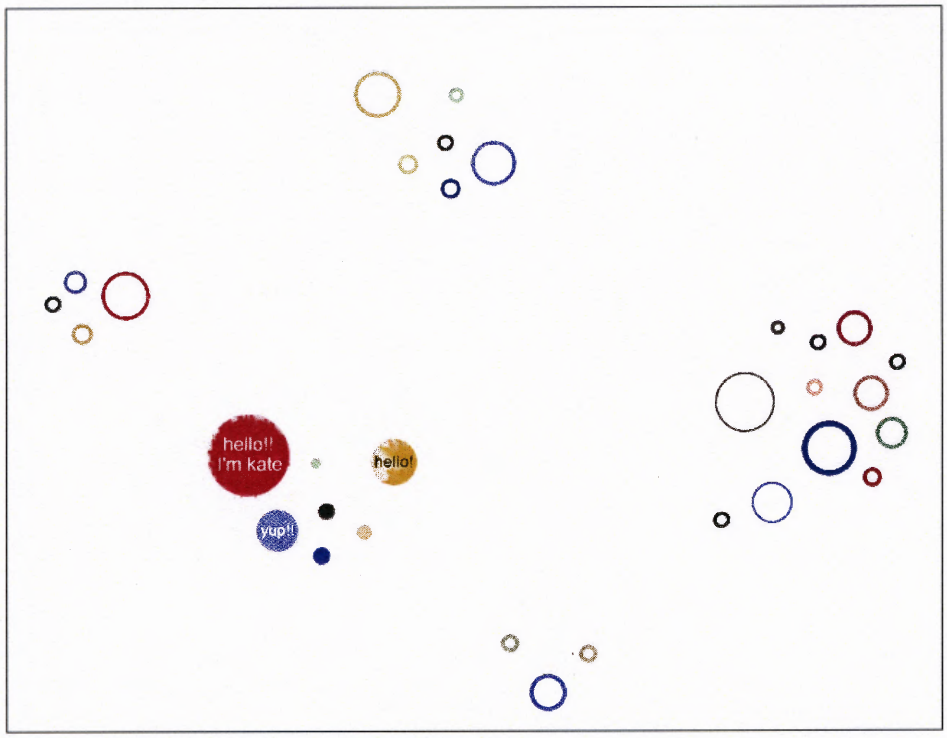

Figure 3.4 Screenshot from a Chat Circles session.
(Source: Donath, Karahalios and Viegas 1999)

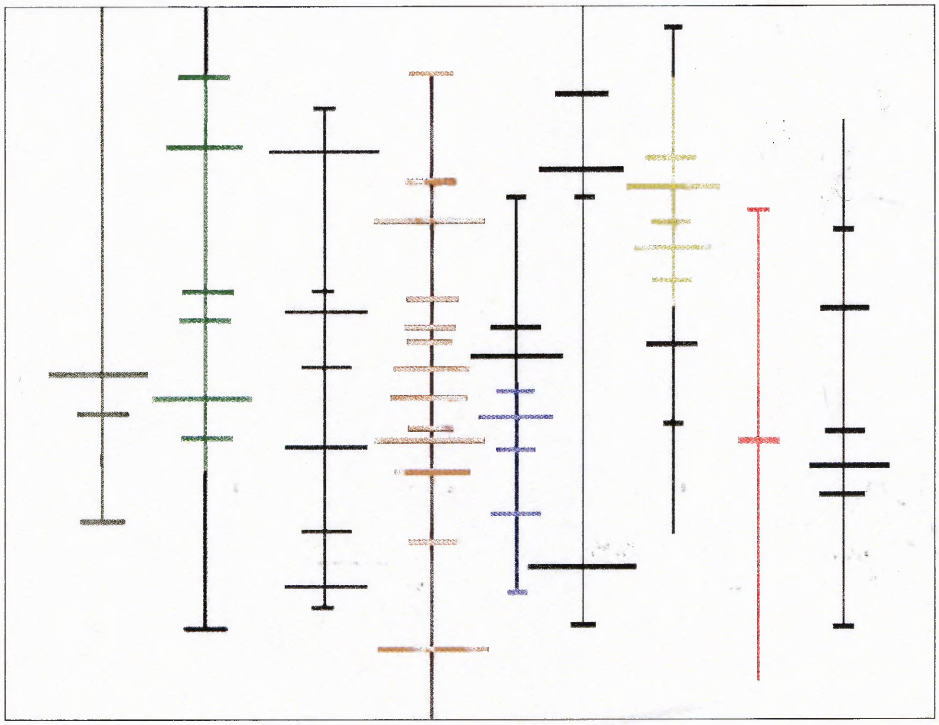

Figure 3.5 ConversationLandscape, the graphical interface to the Chat Circles archives. Vertical lines show the activity of one participant. Horizontal lines represent postings.
(Source: Donath, Karahalios and Viegas 1999)

The users were aware of the total number of users and of the amount of activity in each cluster of conversation, but they were not able to "hear" (see) the discussions going on in a particular place on the screen if their graphical representation (circle) was not close enough to that cluster. The system provided an archive feature with the possibility of recording chats for future reference. The authors also developed an interface for visualizing archives of previous chats. It was called Conversation Landscape and was able to provide graphical information about the patterns of the users. These patterns included periods with high activity versus periods of "lurking." One of the most important aspects of the Chat Circles system was its ability to reveal interaction patterns that occurred over time.

Erickson et al. (1999) started their research from the premise that it would be possible and desirable to design systems that supported social processes. "Socially translucent systems" were defined as systems that provided various social cues that could be perceived by the users, and that afforded awareness and accountability for them. The authors considered translucence to be a fundamental requirement for supporting communication and collaboration interaction among users of CMC systems. Their initial project was called Loops at first and aimed to support "smooth, reflective, and productive conversations through synchronous and asynchronous CMC." The "Babble" system was the first implementation of the basic Loops concepts. The three characteristics that stood at the base of its design were to provide cues from content, to provide social cues, and to support small groups. They also defined the "social proxy" as "a minimalist graphical representation of users which depicts their presence and their activities vis-à-vis the conversation." The social proxy provided basic information about the context of the

conversation such as number of participants, amount of conversational activity and cues about the users' dynamics (for example, how many were leaving or joining the conversation). Results showed that Babble supported opportunistic interactions, group awareness, informality, and sociability and gave its users a feeling of "place." While it merged elements of many types of systems (MUDs, chat, email lists, newsgroups, bulletin boards), Babble wasn't actually any of them; it was unique in its own way.

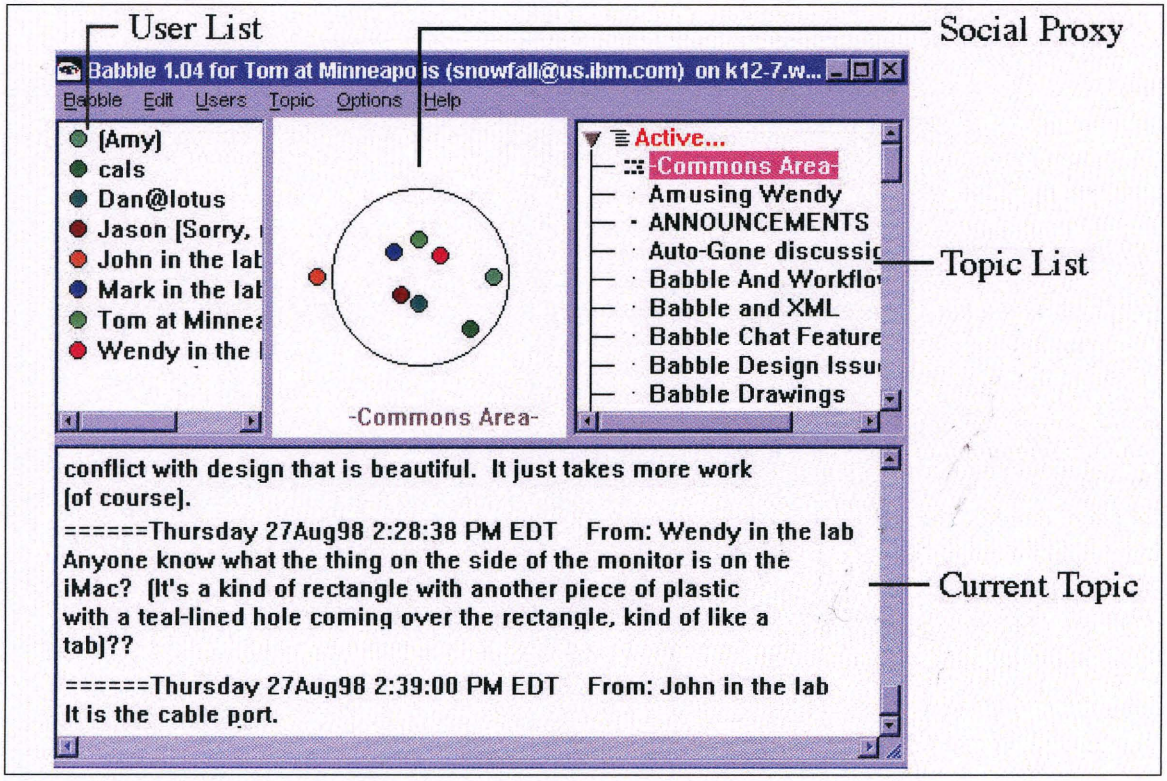

Figure 3.6 The Babble interface.
(Source: Bradner, Kellogg and Erickson 1999)

The Babble system was further researched by Bradner, Kellog, and Erickson (1999). They identified some of the features that distinguished this system from regular chat systems. Babble provided persistent conversations, a social proxy, and it lacked a behavior-enforcing mechanism. The authors found four communicative practices for which Babble was used: waylay, which meant wait for people to be around and then start

a conversation; unobtrusive broadcast, which meant ask a question and wait for somebody to answer, within view of everyone else; staying "in the loop" where one felt connected and aware of what was going on inside the group; and discussion sanctuaries where no outsiders were allowed to use the system, thereby making Babble a safe place to talk. Babble adoption or lack thereof was observed only after four weeks of use. The authors concluded that the relationship between adoption and the communicative practices employed by the group of users was a complex one.

Erickson and Kellogg (2000) argued that people usually make decisions based on other people's activities. They also admitted that generally systems are opaque to any kind of social information and suggested that systems should be designed with an emphasis on making social information available to the users. A system that provided social information to the users could help them be aware of what was going on and help them to be accountable for their actions. Visibility, awareness, and accountability were seen as the key properties of translucent systems. Translucent systems were defined as systems that provided various social cues that could be perceived by the users, and that afforded awareness and accountability for them. Translucence was preferred to transparence because of the tensions between privacy and visibility. They also suggested three approaches in implementing translucent systems: realist systems (teleconferencing, media spaces); mimetic systems (graphical MUDs, Virtual Reality) or VRML (Virtual Reality Modeling Language) worlds; and abstract systems (using some representations of the real world that were not necessarily closely tied to their physical analogs). While they preferred the abstract approach, they also considered the other two to be promising. The Babble prototype represented an instantiation of the abstract approach.

In accordance with the idea that people typically make decisions based on other people's activities, and one's own activities provide information to others; Erickson and his colleagues (Erickson et al. 2002) continued to work on designing systems that allowed their users to perceive social cues about the other participants in the discussion. Although Babble resembled a chat system, it was more than that. Babble also provided persistent conversations and a social proxy. The authors mentioned some other systems that could use social proxies such as online lectures or online auctions. Of course, the issues with socially translucent systems could not be omitted. Trustworthiness and privacy were two major concerns, since making the activities and presence of users available to others could be a sensitive topic. Although Babble was considered a "safe" place because no outsiders had access to it, this situation could change in other circumstances.

Timeline was a system related to Babble (Erickson and Laff 2001). It was a social proxy that allowed visualization of cues about the presence and activity of participants in the Babble system, focusing only on asynchronous participation. The users were able to see and understand the patterns of the community. The authors argued that cues about the users' activities and rhythms of the conversation would be easier to provide in synchronous environments. Timeline was focused only on the asynchronous communications occurring within Babble. Some preliminary results showed that the users were not happy with the implementation but recognized the potential of such a tool in the future.

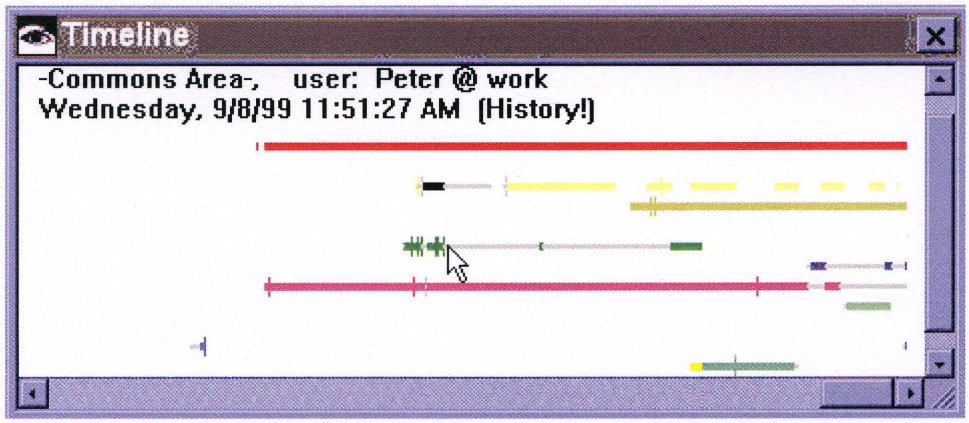

Figure 3.7 The Timeline social proxy.
(Source: Erickson and Laff 2001)

Mutton (2004) described a method for determining the social networks of the users present in an IRC channel and displaying them in a graphical form. He presented the process of inferring the relationships among the users of IRC from a technical point of view, detailing the algorithm for building the users' social networks from the analysis of the text-based interactions occurring inside a chat channel.

3.4 Summary

Social visualizations have been applied to both synchronous and asynchronous CMC systems (Erickson 2003). All of the reviewed studies showed the users' interest in the ongoing activities of their conversation partners and in the current state of the system. They also suggested that the dynamics of a chat system have an important role in the actual system usage and that social visualizations could help users select online spaces for interaction. The methods used today to provide social recommendations are limited. Social matching systems are still in their infancy, and more research is needed to fully understand the design implications of such systems. Furthermore, while social matching systems typically focus on relationships between people, social recommendations could

be extended to include recommending particular interaction spaces to people as a distinct step in the process of bringing people together. Social visualizations techniques have been applied mainly to small groups over small time intervals. The exception is represented by the work of Viegas and Smith on the visualization of large-scale Usenet group activities (2004). However, their research was done on asynchronous CMC systems. The content of group interactions was persistent, and therefore the access to it was not as problematic as it would be in the case of dynamic media such as synchronous chat systems. Social visualizations require a thorough understanding of what is going on inside such systems. This research, through the analysis of the group interaction inside synchronous interaction spaces, will provide the necessary knowledge to build better social visualizations, representing in this way another step forward toward the improvement of social recommendations for synchronous CMC systems.

CHAPTER 4

IDENTIFICATION OF RHYTHMS IN CMC SYSTEMS

4.1 Foundations of Rhythm Identification in CMC Systems

Identification of rhythms is important if people want to know the most appropriate time to find something of interest in a particular CMC system. In the real world, one knows when it is time to go to school or go to work. Currently, the online world does not provide any information to users in the matter of timeliness to join a chat space. This is why there is a need to explore the rhythms of computer mediated communication systems.

The exploration of CMC systems rhythms has expanded in the past few years. It went from simple observations of the presence of rhythms in the users' activities (Curtis 1996; Smith, Farnham, and Drucker 2000; Grinter and Palen 2002) to acknowledging the importance of these rhythms (Mynatt et al. 1998) and addressing them as specific research interests (Smith 1999; Begole et al. 2002; Handel and Herbsleb 2002; Halverson, Erickson, and Sussman 2003; Tyler and Tang 2003; Fisher and Dourish 2004).

One of the first authors who mentioned the rhythms of a CMC system was Curtis (1996). He observed certain rhythms in a MUD system's usage, with higher levels of activity and larger numbers of users during certain time intervals of the day. He looked at the pattern of MUD usage in terms of the number of users versus the time of the day. He observed that the maximum values were recorded between noon and eight in the evening and that the number of users tended to be constant within that time interval. The values dropped for the rest of the day but there were always at least twelve people connected.

Mynat and her colleagues (Mynatt et al. 1998) studied Pueblo, a cross-generation school-centered, text-based MUD; and Jupiter, a hybrid MUD/media space, which provided audio/video links between participants. The authors introduced the concept of network communities, which were defined as "*robust and persistent communities based on a sense of locality that spans both the virtual and physical worlds of their users*" and identified periodicity as one of the affordances of these network communities. Periodicity was defined by combinations of rhythms and patterns in the users' system usage. They noticed that each communication modality (text, video, audio, etc.) had its own rhythm. They also argued that rhythm dynamics were fundamental to network communities and further predicted that routines, intelligible rhythms for individuals and for the community as a whole, were likely to emerge in the future.

Smith, Farnham, and Drucker (2000) performed log file analysis of user behavior inside chat spaces in order to illustrate the dynamical structure of social cyberspaces. Among other findings, they determined that chat rooms had well-defined daily rhythms, in terms of number of people and amount of conversations, with higher activity in the afternoons.

Grinter and Palen (2002) explored IM as "an emerging feature of the teen life," paying attention to the teenagers' everyday use of IM and its support for interpersonal communication. Domestic rhythms and schedules, as well as privacy issues, were found to be important determinants for teenagers' usage of IM. In other words, the participants knew when to expect to find someone online based on the patterns of their usage history of the system.

Tyler and Tang (2003) conducted a study of email responsiveness to "understand how the timing of email responses conveys important information." They found that depending on the time intervals passed between sending and receiving email, people tended to identify the rhythms of their respondents, and therefore began to predict the time frames in which they should expect responses or the time intervals that were best for contacting someone. They suggested that taking into consideration the rhythms of email interaction could lead to better designs for email systems with "better support for mutual negotiation between senders and receivers."

Fisher and Dourish (2004) analyzed the social structures, defined as patterns of collaboration among people (co-workers); and the temporal structures, defined as the ways in which patterns of interaction (among co-workers) changed over time. In doing so, they tried to "uncover usable temporal and social structures from traces of electronic activities." Their initial results showed it was possible to gather information about recurrent patterns of contact from the use of electronic tools. Furthermore, people were likely to recognize those patterns and actually considered them meaningful. One conclusion from this research was that both the social and the temporal rhythms and patterns could be further used to build or enhance CMC systems by providing various awareness tools.

Halverson, Erickson, and Sussman (2003) discovered unique patterns of activity that occurred within a group using a synchronous chat system. Those patterns were "punctuated" by long periods of inactivity followed by bursts of activity. They presented a case study of a geographically distributed group's use of a system (Loops) that was derived from Babble. The observed rhythm was different than the rhythms of all of the

other places where the system (or its predecessor) had been previously deployed. The authors used several methods including a survey of users before deployment of the system; log capture and quantitative activity analysis; content analysis; and semi-structured interviews. The system was considered a success even if it was not used continuously.

Smith (1999) reported the results of Netscan, "a software tool [...] that gathers an ongoing stream of Usenet messages and maintains a database of information drawn from the header of each message." He stated that "an overview of activity in the Usenet has been difficult to assemble. Many basic questions about its size, shape, structure and dynamics have gone unanswered." Netscan helped to find the answer to some of these questions. The results of a 10-week data-collection period showed well defined weekly and daily cycles in Usenet postings. At a weekly level, most of the activity occurred during the working days, with lower levels of activities during the weekend. At a daily level, there were significant differences in the activity of the newsgroups depending on the time of day. The groups were never completely inactive.

Handel and Herbsleb (2002) focused on four main research questions: To what extent did the users of an IM-like application create groups, join groups and use group chat? What did the users talk about, and to what extent did they use flaming? Was chat used in different ways, for different purposes at different times of the day? Did different groups use the system in different ways? The authors were interested in both group dynamics and discourse content. Their findings showed that users tended to join groups other than those to which they were initially assigned. The system's history of use showed patterns of group chat, with succeeding periods of inactivity (usually longer) and

high activity (usually shorter). The authors also observed temporal rhythms in the overall use of the system for the work-related communication. The active period was between 9:00 a.m. and 4:00 p.m., with peaks between 2:00 p.m. and 4:00 p.m., and all the groups were remarkably similar in their patterns.

Begole et al. (2002) were interested in determining the work rhythms of distributed groups. They analyzed visualizations of the awareness histories of several distributed work groups in order to understand meaningful patterns in the users' activities. Using the Awarenex prototype (Tang et al. 2001), they logged all computer interaction activities including keyboard strokes, mouse movements and clicks, location data about where the activities occurred, online calendar appointments, and email activities for 20 users for a period of up to 10 months. The visualizations were produced in the form of actograms displaying the users' daily and weekly activities. The actograms clearly showed the patterns of the users' rhythms, and their analysis revealed several important things. Firstly, the users' work rhythms seemed to be different according to the day of the week. Secondly, the rhythms seemed to be strongly dependent on the location of the users (for example office, home, or mobile location). Thirdly, the patterns tended to vary depending on the users' role within the organization – managers seemed to be more active during the beginning and the end of the workdays and more open to interruptions during the midday. Finally, such visualizations were able to predict, to some extent, the end of users' inactive periods. The authors concluded their paper with some design implications for future systems. Knowing the work rhythms of people could lead to better group coordination applications. Such applications could be improved by providing suggestions about the best times to make contact, predicting

returns from periods of inactivity, augmenting online calendar accuracy, identifying email reading patterns, and restoring cues to negotiate the initiation of contact. The authors mentioned the privacy issues that were inherent to the type of information they collected and suggested careful approaches in the future.

4.2 Summary

To date, there is not a great body of knowledge addressing the rhythms of computer mediated communication systems. Furthermore, the vast majority of the work that has been done so far analyzed the rhythms of small groups over limited time intervals. The exceptions are represented by the works of Smith, who researched the entire Usenet, and of Begole et al., who performed their analysis over a time interval of 10 months. Nothing has been done so far on large-scale synchronous CMC systems. Considering the importance attributed to the rhythms of CMC systems by many authors, by analyzing the rhythms of an entire IRC network this research will provide the foundation for improved navigation inside large synchronous chat spaces, and lead to improved overall designs of such systems.

CHAPTER 5

INTERACTION DYNAMICS OF GROUP TEXT-BASED CMC SYSTEMS

Interaction dynamics can be defined as a description of the activities that are going on inside computer mediated communication spaces. Understanding the interaction dynamics is important from the perspective of social visualizations and social recommendations. There is also a need to determine which of the interaction dynamics of a CMC system are the most relevant, and to understand how they can be used in predicting the activity of synchronous chat spaces.

As it has been shown in the previous chapters, CMC users are generally interested in the activities of their interaction partners. These activities are a part of the interaction dynamics of any CMC system. The term "interaction dynamics" is used to describe general patterns of user interactions in interactive CMC systems. Interaction dynamics have been examined mostly for asynchronous CMC systems (Whittaker et al. 1998; Butler 2001; Jones, Ravid, and Rafaeli 2004), with smaller efforts oriented toward the analysis of interaction dynamics inside synchronous CMC systems (Liu 1999).

5.1 Mass Interaction in Asynchronous CMC Systems

Whittaker et al. (1998) defined the term "mass interaction" as "conversations between hundreds or even thousands of participants." They focused their mass interaction research on Usenet newsgroups, arguing that very little was known about Usenet, which could have been regarded as the world's largest conversation application in 1998. They stated that previous research had been done in the form of small qualitative studies of specific newsgroups, so general questions about mass interaction had not been answered at that time. The authors tried to determine the variables for the demographics, the conversational strategies, and the levels of interactivity of Usenet newsgroups and then to find out the relationships among those variables. Demographics included size of the newsgroup population, familiarity of the participants (how often did they post in a group) and moderation. Conversational strategies included presence of frequently asked questions (FAQ) within the newsgroup, length of messages, and cross-posting, i.e., posting the same message to several groups. Interactivity was defined as the extent of conversational threading. In performing their research, they chose a random sample approach, selecting several hundred Usenet groups. The results showed several things: A significant proportion of users were unfamiliar and posted very rarely; a small proportion of users were responsible for most of the conversation; cross posting was prevalent; initiating messages was common, but it was hard to actually start a conversation; and once a conversation was started, it was likely to attract multiple contributions. The authors concluded that Usenet groups were characterized by moderate and not large amounts of interactivity. They also observed that many attempts to start new interactions did not succeed. Hence, starting mass interaction was problematic. Usenet groups were

characterized by massive participation inequalities; small groups of users were dominant in the discussion. The authors acknowledged that this type of analysis must be complemented by content analysis and user surveys in order to better understand the mass interaction over Usenet groups. An important implication for future research was that conversational overload must be taken into account by the model of mass interaction.

Butler (2001) proposed a theory about the roles of size and communication activities of a community in sustaining online social structures. He stated that previous research showed that simply providing infrastructures for online communication was not enough to guarantee the emergence of social activity. He presented a "resource-based model of the internal dynamics of sustainable social structures." The core premise was that there was a need for a pool of resources to be maintained and social processes to be supported in order to transform the resources into benefits valued by the members of various online or traditional communities. Butler argued that the size of a social structure was an important measure of both source resources and audience resources. Previous research showed that larger groups were likely to provide more valuable benefits to their members and, hence be sustainable over time. However, size could also have a negative effect in terms of an individual's perceptions and attitudes. Earlier research demonstrated that "the undersupply of resources [...] in larger structures is reflected in the general finding that individuals in larger structures tend to be less committed, less satisfied and hence less likely to join or remain members." Two general approaches to manage the effects of size were: to develop internal structures or to use alternative communication technologies. Communication activity was the method through which the benefits were provided; the absence of such activity led to the failure of the social structure.

Communication activity had both positive and negative effects (higher costs - time, attention, energy, knowledge). Similarly, with regard to the case of the group size, there were two general approaches to manage its effects: develop internal structures or use alternative communication technologies. The goal of the model presented by the author was "to describe the dynamics of a wide variety of social structures in a way that allows us to better understand the general impact of information technology." Butler saw the benefit provision as "a necessary condition for sustainability in a wide variety of structures." The method used for data-collection was random sampling of various Listserv email lists. Different measures of size, communication activity and membership change were used for examining the model. The size was defined as the number of users receiving the messages sent to a listserv. The messages sent to the list represented the communication activity, and the number of users who joined or left the list defined the membership change. Results of the data analysis showed that size had both positive and negative effects on the development of social structures, as opposed to previous research which highlighted only single, positive effects. Also, the implications for increasing size seemed to be more complex than initially posited. The positive and negative effects of size on membership might prove that losing members is not necessarily a bad sign overall. The communication activity also had both negative and positive effects on the sustainability of the social structure. Butler concluded that "developing and maintaining sustainable social structures requires that the fundamental problem of balancing the positive and negative impacts of size and communication activity be solved in order to maintain a resource pool for the future while providing benefits for the members in the present."

After stating that claims about the usefulness of the Internet in developing social relationships were still controversial, Cummings, Butler, and Kraut (2002) raised the question of whether online social relationships were better, the same, or worse than relationships sustained be other means. They compared communication over different media, such as face-to-face, email, and phone; and compared the relationships with Internet and non-Internet partners. They also performed an analysis of online social groups, after they collected data from 204 Internet Listserv email lists. The authors argued that previous studies of electronic groups focused solely on the social activities occurring inside the groups, while what really happened in such groups was still under researched and, hence, unknown. They did a sampling of the Listserv email lists and classified them as purely electronic and hybrid. In the case of hybrid email lists the members also met in other settings, not only online. They collected data about membership size in terms of number of members of the list; and data about communication activity in terms of volume (number of messages) and interactivity (length of discussion threads). The results showed that Listserv email lists exhibited little communication; conversations were generally not interactive and a small number of users generated most of the discussion. The results were similar for both types of lists. The authors concluded that Listserv email lists did not appear to be intimate social groups and previous research on online social activity was biased. Highly interactive online social groups were usually the exception and not the rule. However, the authors agreed that things might be different in the case of other technologies, especially synchronous ones (MUDs, IRC). Overall, they concluded that "social places on the internet where close personal relationships are formed and maintained are rare."

Jones and Rafaeli (2000) described a structured approach for understanding the use of collaborative technologies. They defined virtual publics as "symbolically delineated computer mediated spaces, whose existence is relatively transparent and open, that allow groups of individuals to attend and contribute to a similar set of computer-mediated interpersonal interactions."

Jones, Ravid, and Rafaeli (2002) presented results from the analysis of Usenet data, showing the effects of information overload coping strategies over the mass interaction discourse dynamics. Since mass interaction was likely to occur in virtual publics, the authors argued that by modeling the mass interaction, one could achieve a better understanding of the online discourse and the differences between various CMC technologies.

Jones (2003) examined the impact of online public interpersonal interaction spaces on the users' behavior within those spaces. The author argued that the field study was the most appropriate method for such research. He also tried to identify how the user behavior in online spaces affected the interaction dynamics and to identify the technology-associated constraints. He compared two CMC systems - Usenet newsgroups and Listserv email lists - using a method based on the "mass interaction" observations. The empirical findings showed that information overload coping strategies had an impact on the discourse dynamics and that email lists clearly supported a higher level of poster stability than Usenet groups.

Jones, Ravid, and Rafaeli (2004) argued that researchers should examine the type of relationships between the public spaces used for online communication and the interactions that occurred in those spaces, and that the used methodologies must not be

culture-specific or time-specific. They stated that "the particular way in which a technology is used is not determined by the technology itself, but rather is dependent on its social context." The authors recommended performing large-scale field studies when examining the impact of information overload coping strategies over the dynamics of the public discourse in the public spaces, and assumed that any CMC technology could support and enable only a limited range of social interactions.

The review of the literature revealed several categories of variables that were computed to describe the interaction dynamics of asynchronous computer mediated communication systems. These included the proportion of messages that were replied to (Whittaker et al. 1998); the size of user populations over time (Jones, Ravid, and Rafaeli 2004); or the message complexity (Butler 2001). The broad patterns of the interaction dynamics appeared to vary between types of CMC technologies (e.g., email, Usenet, etc.) (Jones 2003).

5.2 Mass Interaction in Synchronous CMC Systems

Very limited work has addressed the interaction dynamics of synchronous chat spaces. Liu (1999) considered IRC to be an environment that would potentially sustain the existence of virtual communities. Based on the Jones's "virtual settlement" theory (1997), Liu hypothesized and then tried to prove that IRC channels had the attributes that would qualify them as virtual communities. His results showed that in theory it would be possible to empirically test for the presence of virtual communities on IRC and not just assume their existence by default. However, some very well-defined parameters would be needed in order to define what a virtual community really represented since the terms

"community" and "virtual community" were among the most controversial terms used in yesterday's and today's research.

5.3 Theoretical Considerations – the Critical Mass Theory

In physics, the critical mass represents the amount of radioactive material needed in order to obtain a nuclear fission explosion. However, the term was successfully adopted in the computer science and information systems literature. In 1968, Licklider and Taylor (1968) saw the critical mass as a small number of people who could effectively contribute toward solving any particular problem online. In 1978, Hiltz and Turoff (1993) observed that computer conferencing systems needed a critical mass of users (usually 8-12) if they were to be successful. However, it was not until 1985 that a fully developed theory of the critical mass was developed (Oliver, Marwell, and Teixeira 1985). Based on Olson's work on various theories of collective action (1965), the authors tried to predict the effects that production functions (relationships between the contributions of the members of a group toward the achievement of the common good) and the group's interest and resource heterogeneity would have on "the probability, the extent and effectiveness of group actions in pursuit of collective good." The critical mass of a group was defined as "getting enough people organized to contribute that some or much of the collective good could be provided." They tested their theory using formal analysis and simulations and concluded that the shape of the production function was the strongest predictor for successful group action. Further, they showed that group success was more likely to occur when heterogeneity of both interests and resources existed among the members of the group.

Oliver and Marwell (2001) reviewed the literature to assess what had happened to the Critical Mass theory since it was first introduced. They identified more than 200 citations, but they characterized most of them (about 66 percent) as "gratuitous at best." Importantly, the majority of the authors who cited this theory saw it as "a species of threshold model," which, by getting enough contributors, would allow a certain tipping point to be passed and lead to unanimous cooperation.

More recently, HCI researchers have shown that critical mass is highly context dependent. Examples include the work of Halverson, Erickson, and Sussman (2003) on persistent chat systems and Grinter and Palen's (2002) exploration of the usage of online calendar systems.

5.4 The Critical Mass Theory and Interactive Media

Markus (1987) used the critical mass theory to explain the diffusion and adoption of interactive media. Arguing that usage of interactive media can have only two states, "all or nothing," she proposed several hypotheses about the relationship between the shape of the production function and the heterogeneity of resources and interests on the one hand, and the achievement of universal media access on the other. Thorn and Connolly (1987), relying on the Critical Mass theory and on some other literature on collective action, studied the contribution of information to "databases," which were essentially archives of computer mediated communication. Seeing the "databases" as interactive media that provided public goods, they tried to determine the factors that influenced users' level of contributions. They produced a conceptual framework, which proposed that reduced contributions occurred because of greater contribution costs; larger groups of

participants; lower value of information to participants; and greater asymmetries in information value and benefits across participants. Rafaeli and LaRose (1993) drew from both Markus' and Thorn and Connolly's work and made several predictions about the success of electronic bulletin boards. Overall, only slight support for the Critical Mass theory was found, but the authors offered a few possible explanations for this situation. Their conclusion was that "the study of interactive technologies needs to proceed beyond the case study level" if one is to better discern the factors that lead to the success or failure of computerized collaborative media.

Although Markus (1987) proposed a critical mass theory that would be applicable to all types of interactive media, most of the recent research has tested this theory focusing exclusively on asynchronous CMC systems (Thorn and Connolly 1987; Rafaeli and LaRose 1993). The results of the existing studies showed that using a "public goods" approach such as the Critical Mass theory in the domain of electronic communication media may be more complex then initially predicted by the theory itself. Thorn and Connolly admitted that both more empirical laboratory work and more theoretical extensions are needed in order to fully demonstrate "the power of 'public goods' thinking for the analysis of organizational communication issues." Rafaeli and LaRose also acknowledged that the picture emerging from the results of their data analysis was that "of a more complex world than predicted by public goods theories." They suggested that further refinements were needed when applying such theories, the Critical Mass theory in particular, to collaborative media.

These authors' findings and suggestions for further research led to the conclusion that the Critical Mass theory lacks a very important component: the time factor (including

here rhythms, patterns, seasonality, etc.). Time can be a crucial variable that should be taken into account when predicting the success or failure of any collective action or achievement of a public good. The Critical Mass theory did not suggest any time frame for measuring the achievement of successful collective action. Instead, it simply stated that given enough resource and interest heterogeneity among the users of a collaborative media, as well as an accelerating shape for the production function, collective action is likely to be obtained. Depending on the size of the community, the type of desired collective action, and many other possible contextual factors; it could take hours, days, weeks, months, or even years for the public goods to be produced and their presence observed.

In the case of computer mediated communication systems, the type of the system, synchronous or asynchronous, is another important variable that may have a significant impact on how time can affect the predictions of the Critical Mass theory. While there are several differences between these two well-known categories of CMC systems, one of the most important is the persistence of the information, in the case of asynchronous CMC, or the volatility of the information, in the case of synchronous CMC. By definition, an asynchronous CMC system does not require the co-presence of users for successful interactions to take place. The pace of interaction is rather slow, and the messages exchanged among the users are stored for long periods of time, being accessible at anyone's convenience. This suggests that the public goods produced within such systems, typically measured in terms of their success, are usually achieved over longer time intervals.

In contrast, information exchanged within synchronous CMC systems is extremely volatile. Users need to be in the right place at the right time in order to successfully interact with each other. Good timing is essential for successful public interactions to occur, since none of the public discourse is stored for future access. It is clear that the success of a chat-room, as a surrogate measure of the public good produced inside it, depends heavily on the time at which interactions typically occur inside it and on the rhythms and patterns of use for that chat-room.

If the contribution rates are considered a measure of the success of CMC systems, as it has been done before, it is clear that different time frames are needed for Usenet groups or email lists on the one hand, and IRC chat-channels on the other. It is reasonable to assume that, with regard to the Critical Mass theory applied to collaborative media, longer time intervals are best suited for research on asynchronous CMC systems, where interactions typically go on for days; and that, as an addition to these longer intervals, shorter, well-defined time periods are necessary when researching synchronous CMC systems.

Previous research has emphasized the importance of the time factor in analyzing the interactions occurring in various types of CMC systems. Rafaeli and LaRose specifically addressed the issue of time as a possible reason for their inability to note the effect of the Critical Mass theory. Nonnecke and Preece (2000) and Halverson, Erickson, and Sussman (2003) showed that the public discourse is often characterized by patterns, where periods of silence and activity alternate. The time factor should be carefully considered in using the Critical Mass theory to predict the success or failure of any synchronous chat system; and therefore it will be addressed in this research.

5.5 Summary

Presently, only limited work has looked at the issue of synchronous chat interaction dynamics. There is a lack of understanding of how the activities conducted inside online chat spaces look. For example, one does not know the meaning of a crowded channel, or an active channel or when channels are overloaded with information or whether they lack critical mass or not. Previous research showed that rhythms and patterns of group interaction were important for the users of CMC systems from at least two perspectives – to find interesting people or to find interesting interaction spaces. Understanding the availability of others, knowing when interaction spaces are active, and knowing the interaction spaces that are preferred by various people are examples of how the identification of the trajectories of group interaction dynamics could help toward these goals. This implies the need for future large-scale research on the dynamics of synchronous CMC interaction spaces.

The review of the critical mass literature showed a consensus regarding the importance of production functions and heterogeneity of resources in the collective action that leads to the public goods. There was also a consensus that online public discourse such as posting to discussion groups can be considered a public good. However, little is known about the shape of the production functions inside synchronous computer mediated communication systems. Presently, nobody has explored and tested the hypothesis presented by the Critical Mass theory related to the effect of the shape of the production functions, and the heterogeneity of resources on the amount of collective action achieved within CMC interaction spaces. To further complicate matters, the theory itself does not incorporate any temporal aspects such as weekly, daily, and/or

hourly weekly patterns or rhythms of engagement, which need to be systematically accounted for in any empirical investigation of production functions and synchronous group interaction trajectories.

This lack of consideration for the time factor makes the application of the Critical Mass theory problematic when it comes to the identification of collective action trajectories. In predicting the likelihood of successful group interaction, it highlights the need to distinguish short-term trajectories from long-term trajectories, as well as the difficulty in determining the most relevant time intervals for both these types of trajectories. The identification of short-term trajectories of group interaction inside chat spaces should provide information about the likelihood of sustained discussions for the very near future. The identification of long-term trajectories should determine whether the interaction spaces are likely to survive or die over longer time intervals. To be able to make the distinction between the short-term and the long-term trajectories and to understand the time intervals that can be considered relevant for both these types of trajectories (i.e., how long does one need to look at a particular group to determine its sustainability over time, or what is the nearest time interval for which one can predict the likelihood of sustained group interaction), empirical examinations of these issues in the context of large-scale synchronous chat systems need to be conducted. This will create better understanding for adapting the Critical Mass theory to such computer mediated communication environments.

There is a need for more efficient navigation through online social spaces, especially through synchronous ones. Providing users with recommendations about where to go could be a possible solution to this navigation problem. However, to make

such recommendations, one should first be able to predict where successful, sustained group interactions are likely to occur. Therefore, one needs to understand whether the Critical Mass theory can offer information about identifying the group interaction trajectories of online synchronous spaces.

CHAPTER 6

RESEARCH QUESTIONS, ASSOCIATED HYPOTHESES AND PROPOSED METHODS

The aim of this dissertation is to explore, empirically and theoretically, whether it is possible to predict the likelihood of sustained group interaction inside a large-scale synchronous CMC system, particularly inside an IRC network. This is important because such prediction algorithms may be used to design systems to benefit both individual users, by providing them real-time recommendations about where to find successful group discourse, and managers of group spaces, by providing them vital information about the health of their communities.

The literature review revealed that the very first step in attempting to make such predictions is to explore and understand the general dynamics of the CMC system. These dynamics encompass all the main characteristics of the system and their variation over time. For IRC, they include factors like the total number of chat-channels and the different types of chat-channels, the total number of users, the total number of publicly active users (posters) or the total number of messages, as well as various other measures that describe the general activity of the IRC network.

The analysis of the dynamics of the IRC system provides insight for predicting future group interaction occurring inside the IRC network. There are two types of predictions that are of interest: short-term predictions and long-term predictions. Short-term predictions target the immediate level of activity of IRC channels, while long-term predictions focus on the survival of the IRC channels over longer intervals of time.

Considering all the above, four main research questions were addressed in this dissertation. Each research question and its associated hypotheses are presented in Table 6.1 and will be detailed in the following sections.

Table 6.1 Research Questions and Hypotheses

Research Question	Hypotheses
What does mass interaction on an IRC network look like?	
What are the boundaries to chat-channel interaction dynamics?	Message density, defined as the number of messages per poster in an IRC channel, will vary with the user population up to a limited user pool. Beyond that point, the message density will remain constant. The cap on message density will constrain the number of posters co-present in an IRC channel.
Considering the dynamic nature of chat networks, when and to what extent is it possible to predict short-term channel activity?	For any publicly active channel and for any short-term time interval for which it is intended to predict the level of channel activity, there will be three main categories of factors that will have an impact on the accuracy of the predictions: (1) the trajectories of channel activity during various previous time periods; (2) the trajectories of network activity during various time periods; and (3) the seasonality of the channels, i.e., rhythms information about each individual channel. The level of predictability of a publicly active channel for any particular week can be estimated as high, low, or perfect by using various descriptive statistics of that channel, computed for the one-month period preceding the week for which predictions are attempted.
What are the early predictors of channel survival?	The long-term survivability of any newly born publicly active channel can be predicted using four categories of factors: (1) the level of channel activity during various time intervals; (2) the trajectories of channel activity during various time intervals; and (3) the heterogeneity of the channel's population during various time intervals; and d) the type of production functions for various time intervals.

6.1 Research Question 1: What Does Mass Interaction on an IRC Network Look Like?

To answer this question, a series of descriptive statistics about the IRC network will be provided. Specifically, detailed descriptions of various aspects of the mass interaction that typically occurs on a medium-sized IRC network will be presented. One year's worth of detailed descriptive statistics data about the IRC network as a whole, the users of the IRC network, and the individual channels of the IRC network will be reported. The following subsections describe the reason that IRC was chosen as the synchronous CMC system to be researched, the selection process of the particular IRC network that was analyzed in this dissertation, the data-collection mechanisms that were employed, and the variables included in each category of descriptive statistics.

6.1.1 Method

6.1.1.1 Selection of the IRC Network. The literature review of the current research on synchronous chat systems revealed several important aspects of IRC. IRC is one of the oldest synchronous CMC systems, and despite its simplicity, is still extremely popular. It is being used by millions of people on a daily basis. Many of the current chat systems, including the group chat features of IM systems, were designed using IRC as a model. Even though such systems do not follow the original IRC protocol, and some provide better graphics and audio and video capabilities, the basic functionality and architecture of the interaction spaces are still based on those of IRC in that they provide a list of current participants, a public area for group discussion, and the opportunity for starting private discussions among their users. Finally, IRC is the most easily accessible medium in terms of opportunities for large-scale data-collection. Although previous researchers

have emphasized the difficulty of collecting large amounts of data from IRC, which is typically true, there exists a method that makes the data-collection process more convenient. This method consists of linking a server to an existing IRC network – not a trivial task – and was the method used in this research.

While thousands of IRC networks exist all over the world, many factors needed to be carefully considered before starting the application process for getting linked to one of them: the size of the network (number of users and channels); the amount of activity on the network; the predominant language spoken on the network; the geographical distribution of the network (where are the other servers located); the reliability of the network; the bandwidth requirements; the possibility of having to deal with distributed denial of services (DDoS) attacks; and the hardware and software requirements for data-collection, storage and analysis. The search for an IRC network that would best suit this research began in the spring of 2003. It took 18 months, a lot of effort, and many failures to finally find a match. The Austnet IRC network was chosen for the following reasons: (1) it had a medium size (an average of 4,000 users and 2,500 channels at any time), which made it easier to analyze (from a technical point of view) compared to large networks (with tens of thousands of users and channels); (2) it was a distributed network, consisting of servers located in the US, Europe, Asia, and Australia; (3) English was the predominant language; and (4) its management committee agreed to link a new server to the network.

6.1.1.2 Data Collection. As shown by the literature review, research on synchronous communication systems has been done using various approaches: virtual ethnography (Reid 1991; Bruckman 1992; Danet et al. 1996; Nardi et al. 2000), participant

observations (Bruckman 1992; Bechar-Israeli 1996), log analysis (Danet et al. 1997; Herring 1999; Isaacs et al. 2002), questionnaires (Muller et al. 2003), interviews (Hansen and Damm 2002; Voida et al. 2002), and case studies (Paolillo 1999). Most of the previous IRC studies have looked only at a very small fraction (one or, at most, several channels) of the total size of a typical IRC network, (which can range from hundreds to tens of thousands of channels and/or users). Also, the time frame for these studies was extremely limited, ranging from one hour to one day. There were studies that spanned longer periods of time, but these studies were still small scale and dealt only with IM systems, MUD systems, or Usenet newsgroups.

Millions of people interact on hundreds of thousands of IRC channels on a daily basis (Hinner 2000), hence the activities conducted here can be categorized as mass interaction (Whittaker et al. 1998). However, small-scale studies are not able to clearly identify the broad patterns of user interactions necessary for a complete understanding of the overall dynamics of the system. To address such issues in mass interaction, large-scale field studies are needed (Whittaker et al. 1998; Butler 2001; Jones, Ravid, and Rafaeli 2004). Accordingly, this dissertation used the approach in favor of collecting data about the entire IRC network as opposed to selecting a random sample of channels and users. There are several reasons for this rationale. First, IRC channels usually have punctuated patterns of evolution in that the number of users and the amount of channel activity change over time. These patterns may be very different between one channel and another, and are impossible to determine without long-term observations. Also, the rates of growth and decay for IRC channels, in other words the lifetime of a channel, usually vary from case to case. The constraints on the boundaries and the stability of channels

can also be very different from case to case. In some situations three users may be sufficient for long-term channel stability, while in other cases, channels may eventually die even with fifteen users at any given time. On any IRC network, the social contexts may differ among channels because of the particularities of the groups of people using them. For example, the social context of a Linux help channel is not likely to be very similar to the social context of a flirt-oriented channel. The same is true for the activities conducted inside channels; the dynamics of file-sharing channels are far different from those of game-playing channels or conversation-oriented channels. The point that needs to be made is that every channel may have some features that make it unique. Even if channels share some similarities, their differences may be much more important. The dynamic nature of synchronous chat-channels makes it impossible to use even a stratified random sample. To build such a sample, one would have to categorize the channels according to some criteria, but those criteria are only obtained from a large-scale analysis of an IRC System. Therefore, the random sample approach is simply not feasible for understanding the dynamics of an IRC network.

Collecting large amounts of IRC data has always been difficult. Especially when dealing with large IRC networks, the data-collection process can be truly cumbersome, and this may be one of the reasons that mass interaction on IRC has never been researched. The biggest technical difficulty encountered by previous researchers was the inability of regular users to log activities on a large number of channels (due to the architecture and the implementation of IRC). Other impediments included the inability to get various statistical information about the network without special administrative rights, connectivity problems (delays between users, network splits, DDoS attacks, etc.), and

channel access issues (users can get "kicked" out or "banned" from a channel, sometimes without any noticeable reason). Linking a server to the Austnet IRC network presents two main advantages: it allows overcoming many of the problems encountered by previous IRC researchers; and it enables the collection of large amounts of data that would have been virtually impossible to obtain otherwise.

The data will be collected using two approaches. The first is through the use of custom-written IRC bots that will continuously monitor the channel spaces and will collect data at specific time intervals. A bot, which is short for "robot," is defined as any program that, once started by a human person, can connect to the IRC network and perform various tasks such as (but not limited to) joining channels, posting messages independently and automatically, or collecting data without the need of further human action. IRC bots will acquire information about the total number of channels and the total number of users of the network. The second approach will be a combination of open-source Transmission Control Protocol (TCP) traffic-monitoring software and custom-written programs that will parse the data collected by the traffic monitoring software and will extract the information relevant to the number of publicly active users and the number of messages exchanged in the public interaction spaces of the chat-channels.

6.1.1.3 Data Analysis.

6.1.1.3.1 Global System Dynamics. The monthly and the yearly values for each of the following variables will be reported for the time interval, February 1, 2005 – January 31, 2006.

- Total number of channels
- Total number of active channels
- Total number of publicly active channels
- Total number of new channels per month
- Total number of users
- Total number of posters
- Total number of messages
- Proportional monthly stability of active channels. For each month, this value is given by the percentage of previous month's active channels that continue to be active during the current month
- Proportional monthly stability of publicly active channels. For each month, this value is given by the percentage of the previous month's publicly active channels that continue to be active during the current month
- Proportional monthly stability of active channels. For each month, this value is given by the percentage of the previous month's active channels that continue to be active during the current month
- Proportional user stability. For each month, this value is given by the percentage of the previous month's users that return during the current month
- Proportional poster stability. For each month, this value is given by the percentage of the previous month's posters that return during the current month

6.1.1.3.2 User-related Descriptive Statistics. The dynamics of the users of the IRC network will be examined in two different ways. First, user activity will be explored at the level of the entire network over a period of one year, with monthly breakdowns. The following key measures will be computed:

- Total number of months users visited the IRC network
- Total number of months users were publicly active inside the IRC network
- Total number of channels visited by users

- Total number of channels in which users were publicly active
- Total number of public messages posted by users to the IRC channels

Second, a smaller sample of users that were engaged in public discussions during an interval of one month will be examined, and the following variables will be computed:

- Average number of channels visited by a user during a session
- Average number of channels a user was publicly active in during a session
- Average time spent by a user during a session
- Average time spent by a user until the first posted public message
- Average number of days a user connected to the network
- Average number of days a user was publicly active

6.1.1.3.3 Channel-related Descriptive Statistics. The dynamics of the channels of the IRC network will be examined in two different ways. First, channel activity will be explored at the level of the entire network over a period of one year, with monthly breakdowns. The following key measures will be computed:

- Total number of months channels were visited
- Total number of months in which channels supported public discussions
- Total number of users that visited the channels
- Total number of posters that visited the channels
- Total number of public messages posted to the channels

Second, a smaller sample of channels that were engaged in public discussions during an interval of one month will be examined, and the following variables will be computed:

- Average number of days a channel existed
- Average number of users per channel
- Average number of daily users per channel
- Average number of posters per channel
- Average number of daily posters per channel
- Average number of messages per channel
- Average number of daily messages per channel
- Average user return time (after how long do users typically return to a channel)
- Average channel user diversity
- Average channel poster diversity
- Average daily user diversity
- Average daily poster diversity

The diversity variables measure the heterogeneity of channels' user and poster populations. Specifically, the user or poster diversity for a particular time interval represents the percentage of users or posters that are present or active in a channel during that interval, with respect to a larger time period. For example, if during a particular hour a channel has 5 unique posters present, and during the day the channel has 15 unique posters; the "diversity" for that hour is equal to 5 / 15 = 0.33. Two types of diversity will be computed: the average diversity for any hour with respect to that day; and the average diversity for any day (daily diversity) with respect to that larger week.

6.2 Research Question 2: What Are the Boundaries to Chat-channel Interaction Dynamics?

IRC channels are highly dynamic with levels of activity, number of users, and topics that are constantly changing. As a result, navigation of IRC networks can be a challenge. Ideally recommendation tools may be developed to help users find channels of relevance; however, this would require that the level of short-term activity of thousands of highly dynamic social environments could be predicted with some degree of reliability.

Currently one does not know the extent to which reliable short-term predictions about channel activity can be made. The research described in this section is aimed at addressing this situation. While it is known that IRC does not limit the number of users or postings in a channel, little is understood about the boundaries imposed by the users' capabilities. This section outlines how empirical research will address the boundaries imposed by user information processing constraints.

The information-processing constraints theory (Jones 1997, Jones and Rafaeli 1999) argues that one influence on a user's participation in computer mediated communication is the level of information overload to which the user is exposed when using the system. Prior research on asynchronous CMC systems has shown that the level of activity within such a system can only rise to a certain level. After this level is reached, due to the effects of information overload, the activity either remains constant or decreases. This research question will aim to identify the maximum level of activity that can be reached inside a synchronous CMC system such as IRC.

6.2.1 Hypotheses

The following hypotheses were formulated using the information-processing constraints theory:

- Message density, defined as the number of messages per poster in an IRC channel, will vary with the user population up to a limited user pool. Beyond that point, the message density will remain constant.
- The cap on message density will constrain the number of posters co-present in an IRC channel.

6.2.2 Method

6.2.2.1 Data Collection. The descriptive statistics of the IRC network presented in the previous chapter revealed the existence of a large number of publicly active channels during each month over the one-year data-collection period. The analysis of all these channels would have been virtually impossible, because of both time constraints and processing power constraints. Therefore, a more manageable dataset will be selected for the identification of channel activity boundaries. This dataset will be selected through a stratified random sampling of all the channels that were publicly active during August 2005. The month of August will be analyzed because it was in the middle of the data-collection period. A detailed description of the sampling procedure will be detailed in Chapter 8, subsection 8.2.1.

6.2.2.2 Data Analysis. The variables and measures to be used in the analysis of the sample of channels are briefly presented in Table 6.2 and will be detailed in Chapter 8, subsection 8.2.2. The IRC system will be sampled three times per hour and for each of these intervals the number of users, messages, and posters will be recorded.

Table 6.2 Variables Used to Measure the Boundaries of Channel Activity

Variable	Description
Observed Users	All the people logged into a channel and using it for either private or public conversations
Observed Users Max	The maximum number of users as the representative value of the three hourly measurements
Observed Messages	All the public postings sent to the IRC channels in the sample
Observed Messages Max	The maximum number of messages as the representative value of the three hourly measurements
Observed Posters	Those users who posted messages in public to the entire group of channel users
Observed Posters Max	The maximum number of posters as the representative value of the three hourly measurements
OMperOU_max	The mean of the maximum number of messages per user within an hour of activity
OMperOP_max	The mean of the maximum number of messages per poster within an hour of activity
OPperOU_max	The ratio of participants who posted messages in public within an hour of activity

6.3 Research Question 3: Considering The Dynamic Nature Of Chat Networks, When And To What Extent Is It Possible To Predict Short-term Channel Activity?

To date, no empirical work has investigated the extent to which short-term measures of activity can be reliably predicted for synchronous spaces such as IRC channels. Regression modeling, both linear and nonlinear, can be used to address this shortfall. Specifically, this section will examine which factors, extracted from the analysis of IRC channel interaction dynamics, can be used to predict short-term chat-channel activity reliably, accurately, and effectively.

6.3.1 Short-term Channel Activity Predictability

6.3.1.1 Hypothesis. It is hypothesized that for any publicly active channel and for any short-term time interval for which it is intended to predict the level of channel activity, there will be three main categories of factors that will have an impact on the accuracy of the predictions: (1) the trajectories of channel activity during various previous time

periods; (2) the trajectories of network activity during various time periods; and (3) the seasonality of the channels, i.e., rhythms information about each individual channel.

6.3.1.2 Method

6.3.1.2.1 Data Collection. The analysis will be performed on the same stratified random sample of channels initially described in subsection 6.2.1, whose selection will be detailed in Chapter 8, subsection 8.2.1.

6.3.1.2.2 Data Analysis. Regression analysis will be used to understand the general short-term predictability of the channels' levels of activity. Regression analysis is a technique used for the modeling and analysis of numerical data consisting of values of a dependent variable and of independent variables: it examines the relation of the dependent variable to one or more specified independent variables.

For this research, the number of posters will considered a measure of the overall activity of chat-channels and therefore will be used as the dependent variable in the regression models. The independent variables in the regression analyses are briefly presented in Table 6.3 and will be detailed in Chapter 9, subsection 9.2.2.

In this research, the notion of "short-term interval" will be equivalent to a "20-minute interval"; and, consequently, for any given time in any 20-minute interval, short-term predictions will be defined as predictions for the immediate 20-minute interval. The reasons for selecting this particular interval will be described in more details in Chapter 9, subsection 9.2.1.

Guided by the results of the regression models, a new variable that will maximize the overall predictability, referred to as the BestPredictor variable will be computed. The accuracy of the models will be determined by exploring the correlations between the

BestPredictor and the actual values of the dependent variable – in this case the actual number of posters.

Table 6.3 Independent Variables for the Linear and Nonlinear Regression Models

Variable	Description
AvgOP_Prev3_20	The average of the observed number of posters during the previous three 20-minute time intervals for each channel in the sample
AvgOP_PrevHr_Nwrk	The average number of observed posters per channel for the entire network for the previous hour
AvgOP_Prev3_20_Nwrk	The average number of observed posters per channel for the entire network during the previous three 20-minute time intervals
AvgOP_Prev3wks	The average number of observed posters for the closest three 20-minutes intervals (just before, current and just after) at the same time during the previous three weeks
AvgOP_Prev12wks_Nwrk	The average number of observed posters per channel for the entire network for the closest three 20-minutes intervals (just before, current and just after) at the same time during the previous 12 weeks
Slope	The slope of the line determined by the observed values for the previous three 20-minute time intervals for each channel
SP (Seasonality Predictor)	The value predicted by a time series analysis of the observed values per channel
TC (Trajectory Coefficient)	A correlation coefficient between "time" and the observed number of posters during the last hour

6.3.2 Identification of Factors that Influence Channel Predictability

It is reasonable to assume that some channels will be easier to predict than others. This research question attempts to identify some of the characteristics that separate highly predictable channels from unpredictable channels.

6.3.2.1 Hypothesis

It is hypothesized that the level of predictability of a publicly active channel for any particular week can be estimated as high, low, or perfect by using various descriptive statistics of that channel, computed for the one-month period preceding the week for which predictions are attempted.

6.3.2.2 Method

6.3.2.2.1 Data Collection. The analysis will be performed on the same stratified random sample of channels initially described in subsection 6.2.1; selection will be detailed in Chapter 8, subsection 8.2.1.

6.3.2.2.2 Data Analysis. Logistic regression will be used to explore the characteristics that separate channels into three main categories: channels with high predictability, channels with low predictability, and channels with perfect predictability. A channel's degree of predictability will be given by the correlation coefficients computed between the best predictor produced by the linear or nonlinear regression model and the observed values of the dependent variable – the higher the value of the correlation coefficient, the higher the predictability of the channel.

Logistic regression is a model used to predict the probability of an event's occurrence and is useful for situations where there is a need to predict the presence or absence of a characteristic or outcome, based on values of a set of predictor variables. The predictors that will be entered in the logistic regression are presented in Table 6.4.

Table 6.4 Predictor Variables for the Logistic Regression Model

Variable	Description
SurvivalTime	Total number of days the channel existed in August 2005
AvgUserRetTime	Average number of minutes between two user sessions
AvgUsers	Average number of users per any 20-minute interval
AvgDailyUsers	Average number of users per day
AvgPosters	Average number of posters per any 20-minute interval
AvgDailyPosters	Average number of posters per day
AvgMessages	Average number of messages per any 20-minute interval
AvgDailyMessages	Average number of messages per day
AvgUserDiv	Average user diversity computed for any 20-minute interval with respect to the day
AvgDailyUserDiv	Average user diversity computed for any day, with respect to the month of August 2005
AvgPosterDiv	Average poster diversity computed for any 20-minute interval with respect to the day
AvgDailyPosterDiv	Average poster diversity computed for any day, with respect to the month of August 2005
Users	Total number of users in August 2005
Posters	Total number of posters in August 2005
Messages	Total number of messages in August 2005
AvgDailyMessagesPerPoster	Daily average number of messages per poster
MessagesPerPoster	Average number of messages per poster for August 2005
DaysVisited	Number of days the channel was visited in August 2005
DaysActive	Number of days public discussions occurred in the channel in August 2005
AvgDailyUserStability	Average daily user stability for August 2005
AvgDailyPosterStability	Average daily poster stability for August 2005

6.4 Research Question 4: What are the Early Predictors of Channel Survival?

Presently, no empirical work has investigated the extent to which the long-term survivability of synchronous spaces such as IRC channels can be reliably predicted. In theory, survival analysis methods could be used to address this shortfall. Specifically, this section will examine which factors, extracted from the analysis of IRC channel interaction dynamics, can be used to predict the long-term survivability of chat-channels reliably, accurately and effectively.

6.4.1 Hypothesis

It is hypothesized that the long-term survivability of any newly born publicly active channel can be predicted using four categories of factors: (1) the level of channel activity

during various time intervals; (2) the trajectories of channel activity during various time intervals; and (3) the heterogeneity of the channel's population during various time intervals; and d) the type of production functions for various time intervals.

6.4.2 Method

6.4.2.1 Data Collection. The analysis will be performed on the set of IRC channels that were "born" during the month of July 2005.

6.4.2.2 Data Analysis. Two important notions need to be considered before analyzing the data: the birth of a channel and the death of a channel.

A channel will be considered "born" the first day when that channel hosts at least three posters exchanging at least four public messages during one 20-minute interval. A channel will be considered "dead" if four weeks of non activity have passed since the last day that channel hosted at least three posters who exchanged at least four public messages during any 20-minute interval. A channel will be considered to be non-active during a particular day if less than three posters are publicly active in that channel during all the 20-minute intervals of that day.

Cox regression analysis will be used to identify the factors that may predict the long-term survivability of chat-channels. Cox regression is a survival analysis method for modeling time-to-event data in the presence of censored cases, which also allows including predictor variables (covariates) in the models. The predictors that will be entered in the Cox regression are briefly presented in Table 6.5 and will be detailed in Chapter 11. For each channel, these predictors will be computed for four intervals: the first two hours of life, the first day of life, the first week of life, and the first month of life.

Table 6.5 Predictor Variables for the Cox Regression Models

Variable	Description
Users	Number of users
Posters	Number of posters
Lurkers	Number of lurkers (users who are not posters)
Messages	Number of messages
PosterDiv	Poster diversity
PosterTrajectory	Posters trajectory
MessageTrajectory	Messages trajectory
PFM	Type of production function for messages

CHAPTER 7

DESCRIPTIVE STATISTICS

This chapter presents a variety of descriptive statistics that provide a broad summary of the dataset and the relationships between several variables of interest. The next section describes the Austnet IRC network and the data-collection process. The results section is divided into three subsections: (1) Austnet System Dynamics; (2) User-related Descriptive Statistics; (3) Channel-related Descriptive Statistics. System dynamics focus on the main characteristics of the entire IRC network and their variation over time. Such characteristics include the total number of channels, the total number of users, the total number of messages, and the total number of posters, as well as various other measures related to them such as their stability over time. User dynamics focus on the main characteristics of the individual users of the IRC network such as the number of channels they visited or the amount of time they spent on IRC. The channel dynamics focus on the main characteristics of the individual channels existing in the IRC network such as the number of people that visited them, the number of days they were visited or the number of messages that were posted inside them.

7.1 Method

7.1.1 The Austnet IRC Network and Data Collection

At any time during the data-collection period, provided there were no network connectivity issues, the IRC network that was researched had, on average, approximately 2,500 channels and 4,000 users. Compared to other existing IRC networks, Austnet was a medium-size IRC network. Data was collected for one year, from February 1, 2005, to January 31, 2006. During this period, the network consisted of ten servers distributed as follows: three servers in Australia; two servers in Asia; four servers in the United States; and one server in Europe (this particular server was disconnected at a point close to the middle of the study period and was not linked back to the network). The servers were linked together in various configurations, depending on factors such as the status of the internet connections in the area where the servers were located or the number of users coming from different regions of the globe. While most of the time all of the servers were online, there were occasions when connectivity problems caused one or several servers to be split from the rest of the IRC network. The server used in this research did not elude such problems, which are quite common in any IRC network. The status of the data-collection process for the 365 days is summarized in what follows:

- For 12 days no data was collected at all. This was caused by severe Distributed Denial of Services (DDoS) attacks on the server, which brought down the entire university network for brief periods, and forced the telecommunications department to take the server completely off the network until the attacks subsided;

- For 41 days only partial data was collected. There were three main reasons for this: (1) the server was disconnected from the rest of the IRC network due to communication issues with the two servers that were the regular links to the network; (2) some of the other exiting servers were disconnected from the IRC network due to various technical problems (mainly DDoS attacks), reducing in this way the size of the IRC network. Typically the periods of time the network

was not whole did not last for too long, but over the course of one year they added up to a significant amount; and (3) the traffic monitoring programs and/or the data-collecting bots were partially disabled and/or disconnected, due to various other reasons (memory leaks, software bugs, code revisions, and code maintenance);

- Complete data was collected for 312 days.

The data was collected using two approaches. The first was through the use of custom-written IRC bots that continuously monitored the channel spaces and collected data at specific time intervals. A bot, which is short for "robot", is defined as any program that, once started by a human person, can connect to the IRC network and perform various tasks such as (but not limited to) joining channels, posting messages independently and automatically or collecting data, without the need of further human action. This resulted in information about the total number of channels and the total number of users of the network. The second approach was a combination of open-source TCP traffic monitoring software and custom-written programs that parsed the data collected by the traffic monitoring software and extracted the information relevant to the number of active users (posters) and the number of messages exchanged in the public interaction spaces of the chat-channels. In addition, a keyword-based algorithm was developed for the identification of postings and other actions that were taken by various IRC bots or other automated scripts.

The data was collected by the custom-written programs in text format and occupied an amount of approximately 4.3 GB of hard disk space. The data collected by the traffic monitoring software occupied approximately 34.27 GB of hard disk space. However, this data was collected and archived in zip files. The total amount of raw text data that was collected in this manner occupied approximately 171.35 GB of hard disk

space. As mentioned above the entire collected data, i.e., both the data collected by the bots and the data collected by the traffic monitoring software, was further processed to extract the information relevant to this research. This information was stored into a MS SQL Server 2000 database. At the end, the database contained 159 tables, which were all used in one way or another in the analysis. The MS SQL database occupied 56.40 GB of hard disk space.

7.1.2 Data Analysis

The data was analyzed using the SPSS statistical software. The most common methods used were descriptive statistics, frequency analysis, histograms and other various types of graphs.

7.2 Results

7.2.1 Austnet System Dynamics

This section uses tables and plots to summarize the breakdown of the most important variables related to the system dynamics of the IRC network. First, a detailed description of the monthly and yearly values for seven variables is provided: the number of channels, the number of active channels, the number of publicly active channels, the number of new channels per month, the number of users, the number of posters, and the number of public user messages.

Table 7.1 shows the monthly values for the number of channels, active channels, publicly active channels, new channels, users, posters, and messages. Active channels are defined as channels that were visited by users, but public interactions did not occur. Publicly active channels are defined as channels that were visited by users, and public

interactions did occur inside them. Public interaction only includes messages that are sent to the public space of a chat channel. New channels are defined as channels that did not exist during the previous month. Users are defined as any nicknames that joined at least one channel of the IRC network. Posters are defined as any nicknames that were not identified as bots and that posted at least one message to the public interaction space of at least one channel. The first monthly average was computed taking the outliers for May and December into consideration, while for the second monthly average these months were disregarded. Figure 7.1 a) shows the variation of the total number of channels, the total number of active channels, the total number of publicly active channels, and the total number of new channels across the entire study period. One can observe a constant decrease of all of these four variables. Figure 7.1 b) shows the variation of the daily average values for the number of channels, active channels, and publicly active channels. The daily averages were computed using only the days for which complete data was collected. A constant decline in all three variables can be observed over the year.

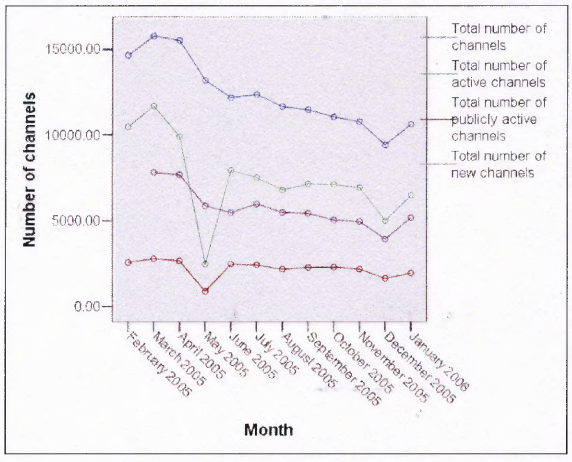

Figure 7.1 a) Variation of the number of IRC channels between February 2005 – January 2006.

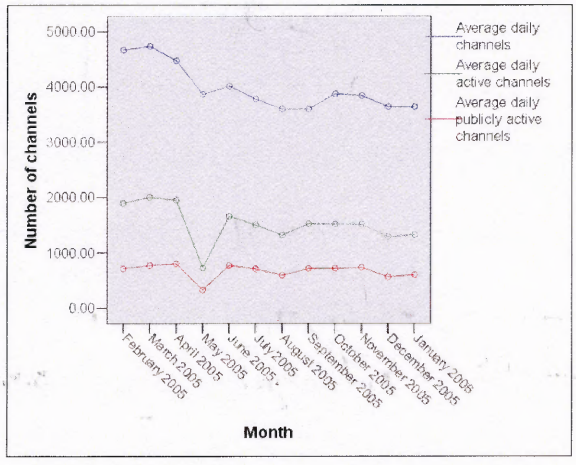

Figure 7.1 b) Variation of the average daily number of IRC channels between February 2005 – January 2006.

Table 7.1 Global System Descriptive Statistics Expressed as Numbers

Month	Total number of channels	Total number of active channels	Total number of publicly active channels	Total number of new channels per month	Total number of users	Total number of posters	Total number of user messages
February	14,646	10,469	2,583	N/A	168,171	81,412	9,303,036
March	15,774	11,676	2,807	7,811	184,709	91,182	11,088,763
April	15,506	9,899	2,678	7,672	171,034	84,424	10,315,247
May	13,180	2,453	905	5,864	64,402	24,258	1,501,508[1]
June	12,177	7,911	2,483	5,478	167,079	67,509	8,778,209
July	12,358	7,491	2,438	5,976	163,682	73,056	9,688,051
August	11,652	6,781	2,186	5,482	149,771	64,063	7,854,781
September	11,463	7,114	2,290	5,428	161,283	70,942	9,987,007
October	11,051	7,090	2,319	5,039	189,232	68,788	10,196,745
November	10,764	6,905	2,192	4,962	142,572	67,169	10,869,283
December	9,403	4,987	1,647	3,933	92,749	41,169	4,572,892[1]
January	10,609	6451	1,960	5,174	128,055	62,449	9,296,404
MonthlyAvg1	11,816	6,708	2,110	5,501	142,986	62,383	8,905,422
MonthlyAvg2	12,600	8,179	2,394	5,891	162,559	73,099	9,737,753
Total	**61,337**	**35,167**	**7,180**	**N/A**	**1,115,141**	**489,561**	**103,451,926**

One explanation for this descending trend, especially in the case of the total number of channels and the total number of active channels, could be the effect of security measures that were employed by the administrators of the IRC network in April 2005. At that time the network suffered numerous attacks from "botnets." A botnet is basically a network of malicious bots, installed through means such as viruses or Trojan horses on the computers of unprotected Internet users. Botnets could include hundreds or even thousands of infected computers and are usually, but not always, controlled by spammers or people interested in conducting DDoS attacks. The person in control of such a botnet could interact with the bots through IRC. For example, all the bots would join a channel on an IRC network, and the person controlling them would then give them instructions by communicating with them via private messages or public messages sent to that channel. The Austnet IRC network was confronted with such problems especially in

[1] The rather low values (compared to the rest of the months) for the all the variables during the months of May and December were caused by some technical difficulties that made it impossible to collect data for 21 days in May and 10 days in December

February and March of 2005. Therefore, in April 2005 a range of security measures were implemented that eventually prevented malicious bots from connecting to the servers of the IRC network and maintaining the channels they were programmed to join and use.

Another explanation for the decrease in the total number of channels over time may relate to the mechanisms through which IRC channels come into existence and/or disappear. From Table 7.1 it may be observed that during each month of the analyzed period, a significant proportion of the total number of channels that existed in the IRC network were not visited by users. An IRC channel is created when a user decides to join it. Provided that there is no other channel with the same name, the new channel is created, and it will exist while there is at least one user present inside it. Once the last user leaves, the channel disappears. However, there are two scenarios in which a channel will not disappear even when no users are present. One is when a user-controlled bot (a bot that is run by a user of the IRC network from his computer or from a remote computer) remains inside the channel at all times, preventing in this way the disappearance of the channel. However, this is not a dependable method because various factors could cause the bot to disconnect from the network, leading to the channel's disappearance. A more reliable approach is to register a channel with the IRC network. In this way, that channel will continue to exist even if no users and no user-controlled bots are present. Usually, the IRC servers provide a service that keeps track of all the registered channels by permanently keeping a server-controlled bot inside them. This is how it is possible to have channels on the network that are never visited by users and this is the explanation for the difference between the total number of channels and the total number of active channels. However, there are limitations to how long a channel can be

registered. Depending on the settings of the registration service, a channel can be automatically un-registered if a certain amount of time has passed, and no users have visited the channel during that period. Therefore, one could presume that the registration of many channels - registered before the start of the data-collection period - expired at some point after February 2005, leading to the decrease in the number of total channels. Figure 7.1 also shows that the number of publicly active channels seemed to decrease at a lower rate, staying almost constant throughout the entire year (if the months of May and December are disregarded due to the technical difficulties encountered in the data-collection process).

Figure 7.2 a) plots the monthly variation of the total number of users and posters during the entire study period. Figure 7.2 b) plots the monthly variation of the daily average number of users and posters computed using only the days for which complete data was collected. A decline can be observed for both variables, especially in the case of the daily averages.

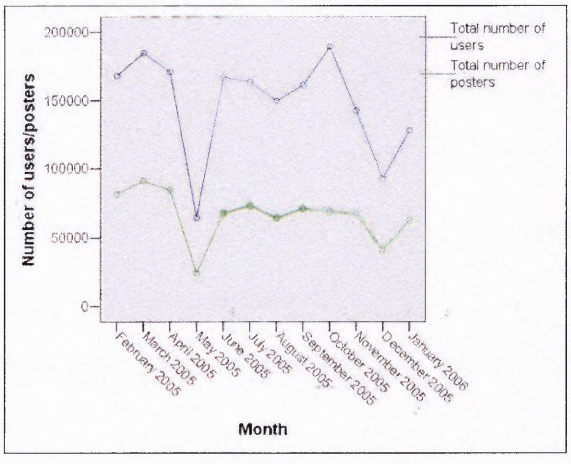

Figure 7.2 a) Users and posters by month.

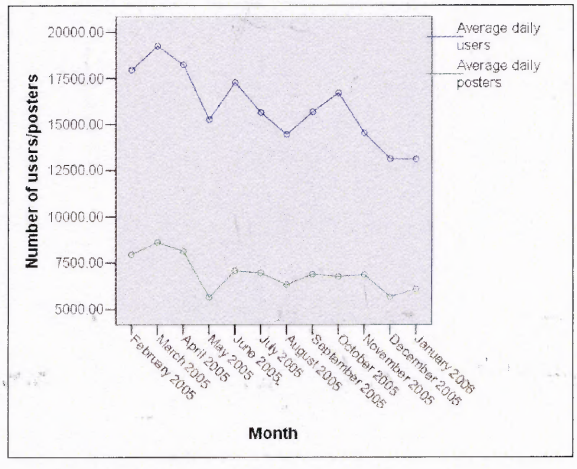

Figure 7.2 b) Average daily users and posters by month.

It would be reasonable to assume that many of the variables presented in Table 7.1 are highly correlated with each other; and the plots presented above tend to confirm this assumption. Correlations are typically explained by either the Pearson correlation coefficient or by the Spearman correlation coefficient. Pearson's correlation coefficient is a measure of linear association. Two variables can be perfectly related; however, if the relationship is not linear, Pearson's correlation coefficient is not an appropriate statistic for measuring their association. One way to reduce the likelihood of this problem is to examine the relationship via the Spearman correlation coefficient, which examines the ranks of ordinal data rather than the values themselves. Spearman's correlation will be used in this research, as it is generally the preferred approach with this sort of data: assumptions about the normality of distributions are not required.

Table 7.2 presents the correlations between the system variables described in Table 7.1. High positive correlation values such as those between the number of new channels and the total number of channels; between the number of posters and the number of publicly active channels; or between the number of users and the number of active channels are to be expected as an increase or a decrease in either of the variables from the aforementioned pairs should be related to an increase or a decrease in the other variable. Medium-high correlation values such as those between the number of messages and the number of publicly active channels; between the total number of channels and the number of active channels; between the total number of channels and the number of publicly active channels; or between the number of messages and the number of posters could indicate some interesting characteristics of the IRC network and lead to further research topics.

Table 7.2 Spearman Correlations Coefficients between System Variables

			Total no. of channels	Total no. of active channels	Total no. of publicly active channels	Total no. of new channels	Total no. of users	Total no. of posters	Total no of messages
Spearman's rho	Total no of channels	Correlation Coefficient	1.000	.692(*)	.692(*)	.955(**)	.517	.671(*)	-.133
		Sig. (2-tailed)	.	.013	.013	.000	.085	.017	.681
		N	12	12	12	11	12	12	12
	Total no. of active channels	Correlation Coefficient	.692(*)	1.000	.986(**)	.582	.839(**)	.951(**)	.259
		Sig. (2-tailed)	.013	.	.000	.060	.001	.000	.417
		N	12	12	12	11	12	12	12
	Total no. of publicly active channels	Correlation Coefficient	.692(*)	.986(**)	1.000	.564	.888(**)	.951(**)	.294
		Sig. (2-tailed)	.013	.000	.	.071	.000	.000	.354
		N	12	12	12	11	12	12	12
	Total no. of new channels	Correlation Coefficient	.955(**)	.582	.564	1.000	.391	.573	-.227
		Sig. (2-tailed)	.000	.060	.071	.	.235	.066	.502
		N	11	11	11	11	11	11	11
	Total no of users	Correlation Coefficient	.517	.839(**)	.888(**)	.391	1.000	.853(**)	.280
		Sig. (2-tailed)	.085	.001	.000	.235	.	.000	.379
		N	12	12	12	11	12	12	12
	Total no. of posters	Correlation Coefficient	.671(*)	.951(**)	.951(**)	.573	.853(**)	1.000	.371
		Sig. (2-tailed)	.017	.000	.000	.066	.000	.	.236
		N	12	12	12	11	12	12	12
	Total no of messages	Correlation Coefficient	-.133	.259	.294	-.227	.280	.371	1.000
		Sig. (2-tailed)	.681	.417	.354	.502	.379	.236	.
		N	12	12	12	11	12	12	12

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

For example since an increase in the number of messages was not very highly correlated with the number of users or the number of posters, one could hypothesize that information overload may be the reason for this behavior. The issue of information overload in chat-channels will be addressed in more detail in Chapter 9. Medium correlation values such as those between the number of new channels and the number of active channels or between the number of new channels and the number of publicly active channels indicate that only in about half the time of the analyzed period, an increase in the number of new channels was related to an increase in overall channel activity (active channels or publicly active channels). This could mean that many of the new channels that appeared in the IRC network were not visited by users and died over time. The survival capabilities of new channels will be addressed in more detail in Chapter 10.

Table 7.3 Global System Descriptive Statistics Expressed as Percentages

Month	% of users who are posters	% of active channels from total channels	% of publicly active channels from active channels	Active channels monthly proportional stability	Publicly active channels monthly proportional stability	Users monthly proportional stability	Posters monthly proportional stability
February	48.41	71.48	24.67	N/A	N/A	N/A	N/A
March	49.36	74.02	24.04	53.87%	70.69%	35.04%	32.70%
April	49.36	63.83	27.05	45.94%	66.94%	31.74%	29.71%
May	37.67	18.61	36.89	19.82%	31.00%	19.00%	15.00%
June	40.40	64.96	31.39	72.28%	83.54%	40.89%	41.28%
July	44.63	60.61	32.55	53.39%	70.96%	29.15%	32.64%
August	42.77	58.19	32.24	52.96%	65.87%	28.13%	28.69%
September	43.99	62.06	32.19	56.94%	72.87%	30.63%	31.80%
October	36.35	64.15	32.71	55.86%	72.58%	28.64%	30.34%
November	47.11	64.14	31.75	53.48%	70.12%	24.36%	30.99%
December	44.39	53.03	33.03	46.00%	61.54%	24.83%	23.00%
January	48.77	60.80	30.38	62.74%	77.90%	39.29%	38.51%
MonthlyAvg[1]	**44.43**	**59.65**	**30.74**	**52.12%**	**67.64%**	**30.16%**	**30.42%[1]**
MonthlyAvg[2]	**48.77**	**64.81**	**29.89**	**53.20%**	**74.71%**	**29.67%**	**30.98%[2]**

[1] Computed for all the months
[2] Computed only for the months during which complete data was collected

Table 7.3 reports the number of posters, the number of active channels, and the number of publicly active channels as percentages of the total number of users, the total number of channels, and the total number of active channels, as well as the monthly proportional stability values for active channels, publicly active channels, users, and posters. To get a better image of how these percentages varied over the course of one year, they are also presented in graphical form in Figures 7.3 (a - d).

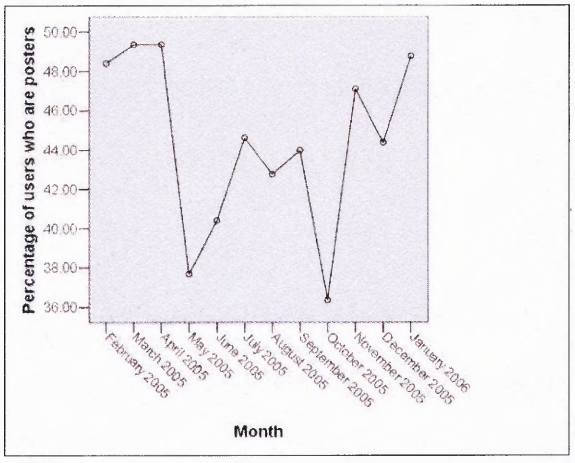

Figure 7.3 a) Number of posters expressed as a percentage of the total number of users.

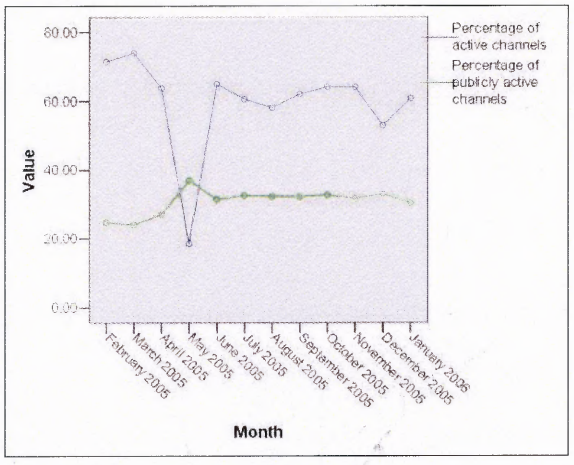

Figure 7.3 b) Number of active channels and publicly active channels expressed as percentages.

From Figure 7.3 a) it may be observed that during every month of the analyzed period, the number of posters varied between 36 and 49 percent of the total number of users of the network, with a monthly average value of 44 or 48 percent, depending on whether the two problematic months were taken into account. Figure 7.3 b) plots the number of active channels expressed as a percentage of the total number of channels and the number of publicly active channels expressed as a percentage of the number of active channels. If the outliers corresponding to the months of May 2005 and December 2005 are disregarded, it may be observed that the number of active channels, i.e., the number of channels that were visited by users, represented a relatively high percentage of the

total number of channels. The lowest value was recorded during August, while the highest values occurred in February and March. The monthly average, disregarding May and December was approximately 65 percent, while the monthly average for the entire year was approximately 60 percent. In other words, during any month about 60 to 65 percent of the total number of channels existing on the IRC network were visited by users, while the rest were completely inactive.

The number of publicly active channels also remained constant throughout the year. The lowest values were recorded in February and March, when approximately 24 percent of the active channels hosted public discussions. During the following months this percentage grew slightly just above 30 percent and remained close to this value for the rest of the year. The monthly averages that were computed (one disregarding May and December and one taking all months into account) were very close to each other: 30.74 percent and 29.89 percent. In conclusion, during any month of the analyzed period, approximately 30 percent of the channels that were visited by users actually hosted public conversations.

Figure 7.3 c) describes the variation of the proportional monthly stability of the users and posters, while Figure 7.3 d) describes the variation of the proportional monthly stability of the active and publicly active channels of the IRC network.

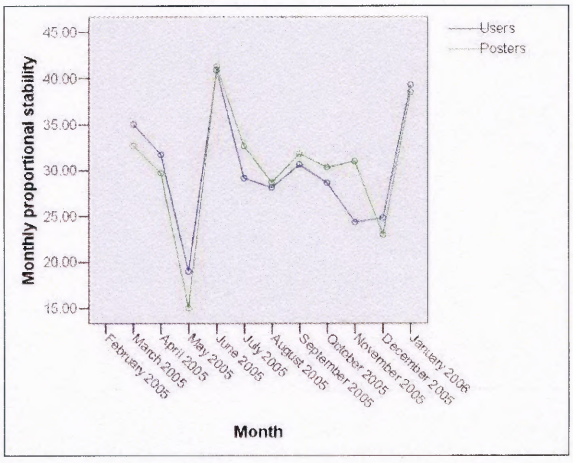

Figure 7.3 c) Users and posters monthly proportional stability.

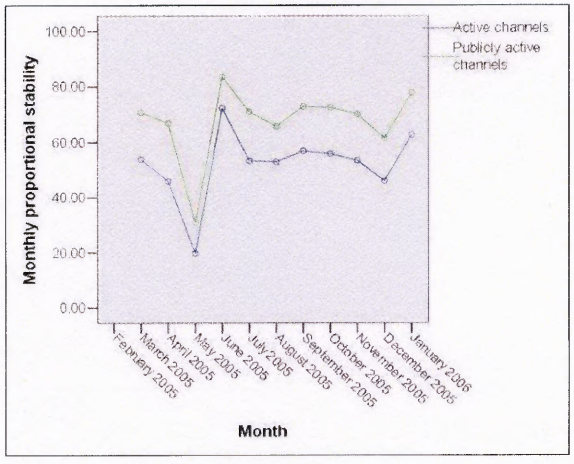

Figure 7.3 d) Active channels and publicly active channels monthly proportional stability.

The proportional monthly stability is a measure that gives an understanding of the rate at which the network maintains some of its characteristics from month to month. In the case of users and posters, the proportional monthly stability shows how many of the users and posters that were active on the network during a particular month would also return during the next month. Both monthly averages that were computed for the users and the posters had similar values, approximately 30 percent. This means that only 30 percent of the users from one month would visit the network during the next month and that only 30 percent of the posters from one month would be involved in public interactions during the next month. The average monthly proportional stability of active channels was just above 50 percent, meaning that about half the channels that were visited during a month would be likely to be visited the following month. The publicly active channels had the highest stability values, averaging 67 percent or 74 percent, depending on whether or not the outlier months were taken into consideration. This high stability value suggests that many of the publicly active channels are well-established,

well-known to the users of the IRC network, and tend to sustain public interactions over longer periods of time.

7.2.2 User-Related Descriptive Statistics

This section examines via tables, graphs, plots and correlations various characteristics of IRC users. In Subsection 7.2.2.1 the focus is at the level of the entire network and the time span is the entire year, with monthly breakdowns. This subsection looks at some key measures of the activity throughout the year such as the total number of months users were active, the total number of channels that were visited by users, or the total number of public messages posted by users to the IRC channels. Subsection 7.2.2.2 focuses on the month of August and on the subset of users that were engaged in public discussions during that interval.

7.2.2.1 General Characteristics of the Users of the IRC Network. This subsection describes the general characteristics of the human users of the IRC network. The users of any IRC network can be broadly categorized into human users – regular people using the network – and non-human users – various bots or other programs that connect to the network. The behavior of the bots can influence the network both positively and negatively. Friendly bots are programs that help users in the management of the channels, preventing misbehavior and enforcing the rules for decent public interaction. Other friendly bots are used in game playing. Trivia bots are the most popular example, but there are other game-playing bots as well, such as IRC horse racing bots. Other tasks usually performed by friendly bots may include (but are not limited to) the following: network management, security management, and statistical data-collection. Unfriendly bots are usually responsible for spam, virus distribution, DDoS attacks, etc. Such bots can

cause significant problems, which, if not taken seriously and addressed in time, can have an extremely negative impact on both the users of the network and the organizations hosting the IRC servers.

Since this research focuses specifically on the behavior of human users, the need for a mechanism that has the ability to separate the messages sent to the public interaction spaces by bots from the messages sent by humans becomes very important. To this end, an algorithm was developed that performed this separation. The algorithm parsed all the collected message data and attempted to distinguish real users from bots based on various patterns of keywords, phrases, and special characters that were identified as typically used by bots. To refine the algorithm, several iterations were performed. To improve the selection mechanism, random samples of messages attributed to both real users and bots were examined between iterations. Since the review of messages might raise some ethical questions, it should be mentioned that absolutely no connection was made between any identifier (nickname or Internet Protocol [IP] address) of the authors of the messages and the content of the messages. The analysis was performed simply to determine various patterns that would enable a better distinction between the messages generated by bots and by humans.

In the end, an expert's review revealed that the algorithm correctly identified over 99 percent of all the messages. Less than 1 percent of the messages attributed by the algorithm to bots were actually originated by human users, while less than 1 percent of the messages attributed by the algorithm to human users were actually originated by bots.

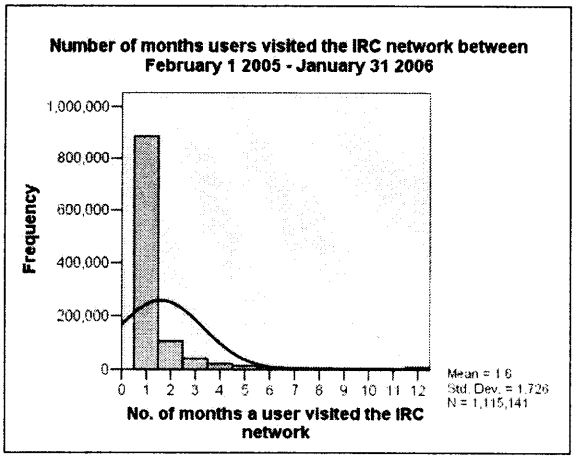

Figure 7.4 a) Histogram of number of visited months per user.

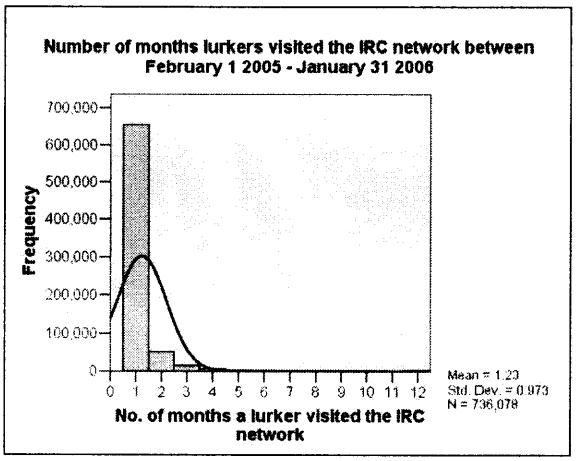

Figure 7.4 b) Histogram of number of visited months per lurker.

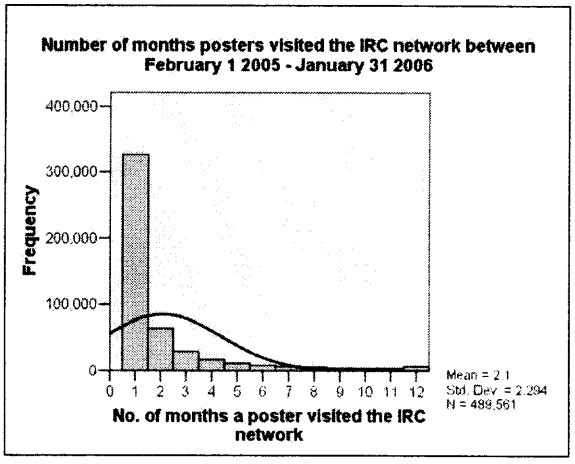

Figure 7.4 c) Histogram of number of visited months per poster.

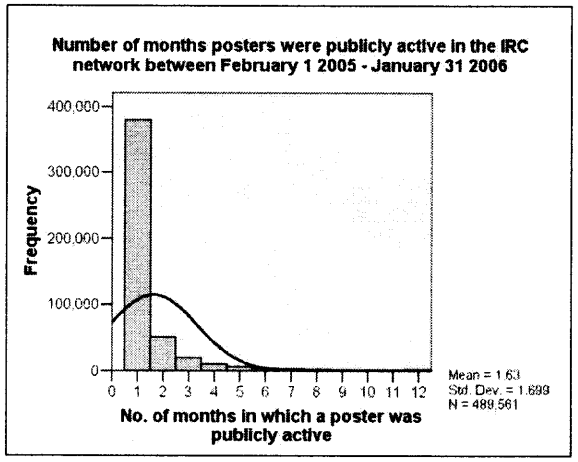

Figure 7.4 d) Histogram of number of active months per user.

The histograms in Figures 7 (a - d) reveal that over the course of one year the vast majority of users who visited the IRC during any month did not return during the next month. The users of the IRC network were divided into two main categories: lurkers and posters. Lurkers were defined as those users who did not participate in public conversations in the IRC channels; posters were those users who were publicly active and sent at least one message to the public interaction space of at least one channel of the IRC

network. While attempts were made to discard all the data that was related to IRC bots from this analysis, it was not possible to completely eliminate it. As more information is available, it is easier to differentiate between bots and humans. The public messages sent by users were the most important pieces of information in determining whether the source of those messages was a bot or a human. Therefore, separating bot posters from human posters was easier than separating bot lurkers from human lurkers.

The histograms show that most of the lurkers and posters tended to visit the IRC network for short periods of time. Table 7.4 presents some interesting descriptive statistics for users, lurkers, and posters. The mean, median, and mode values for the number of visited months per user, the number of visited months per lurker, and the number of visited/active months per posters confirm what the histograms suggested, i.e., the typical IRC users did not visit the network for more than one month. The percentile distributions show very clearly that 75 percent of all the users visited the network for only one month, while only 1 percent of the population remained stable during the course of the entire year. The lurker population and the poster population followed similar patterns with 75 percent of the lurkers not visiting for more than one month and 75 percent of the posters not visiting for more than two months. However, a significant difference between lurkers and posters can be observed in the case of the smaller, more stable segments of the populations: while approximately 5 percent of the poster population visited the network for at least eight months, only 1 percent of the lurker population visited the network for more than five months. Another interesting finding is that the 10 percent of the poster population that visited the network for five or more months was not active during the entire visited periods. Only 5 percent of the poster

population was actively engaged in public interactions for five or more months. This suggests that in order to obtain a stable poster population for IRC channels, a leading period of several months is needed. During such periods, small percentages of the total number of lurkers may become interested in public interaction and become posters.

Table 7.4 Descriptive Statistics – Number of Months Users Visited the IRC Network

	Mean	Median	Mode	StDev	Percentiles						N
					25%	50%	75%	90%	95%	99%	
Number of months users visited the network	1.60	1	1	1.726	1	1	1	3	5	11	1,115,141
Number of months lurkers visited the network	1.23	1	1	0.973	1	1	1	2	2	5	736,078
Number of months posters visited the network	2.10	1	1	2.294	1	1	2	5	8	12	489,561
Number of months posters were active in the network	1.63	1	1	1.699	1	1	1	3	5	11	489,561

In what follows, the relationships between users and the channels they visited are examined in more detail. Figures 7.5 (a - l) display the monthly distributions of the number of channels visited by users during the study period. In all the cases the distribution is highly skewed, showing that most of the users visited very few channels during any given month.

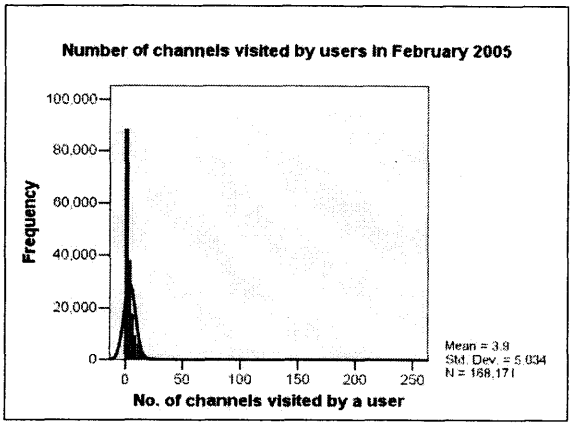

Figure 7.5 a) Histogram of channels visited by users in 02/2005.

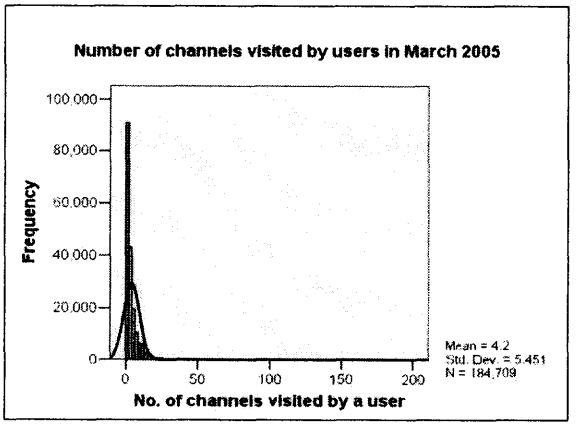

Figure 7.5 b) Histogram of channels visited by users in 03/2005.

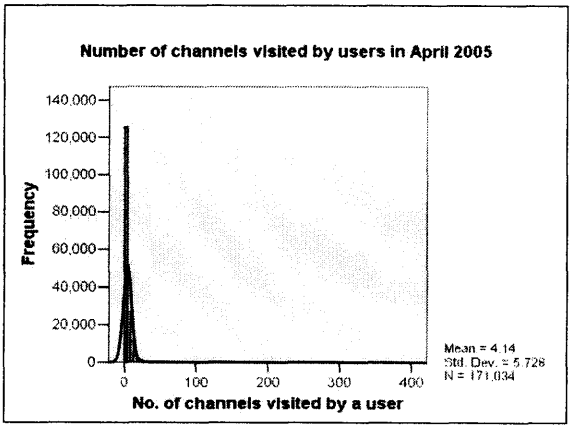

Figure 7.5 c) Histogram of channels visited by users in 04/2005.

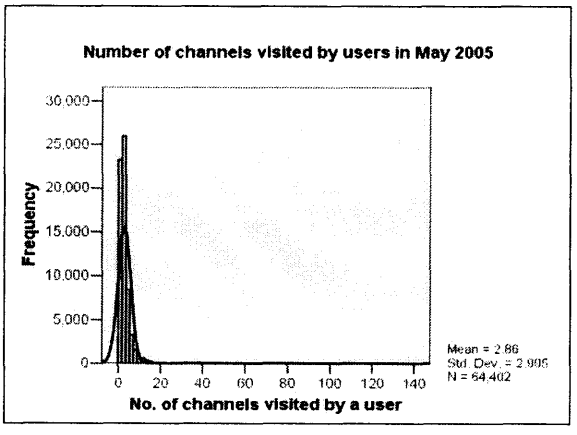

Figure 7.5 d) Histogram of channels visited by users in 05/2005.

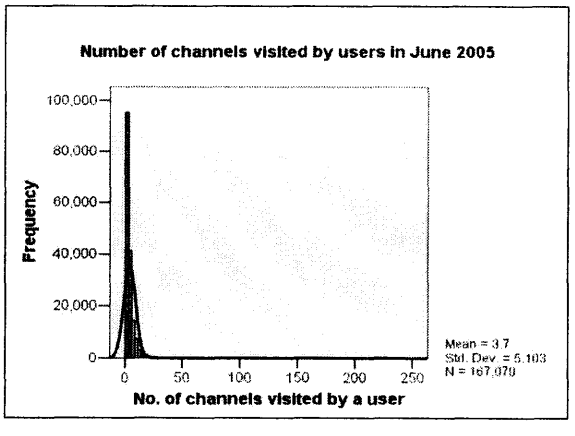

Figure 7.5 e) Histogram of channels visited by users in 06/2005.

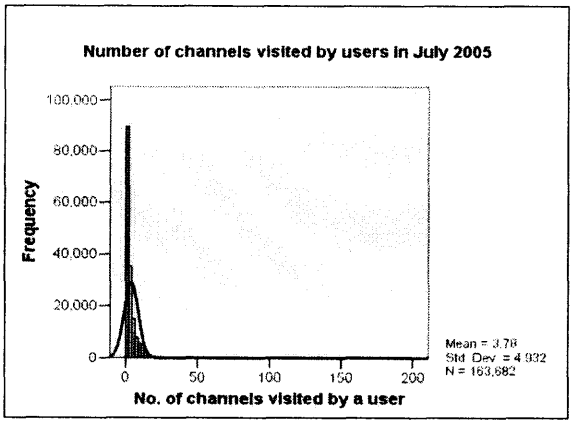

Figure 7.5 f) Histogram of channels visited by users in 07/2005.

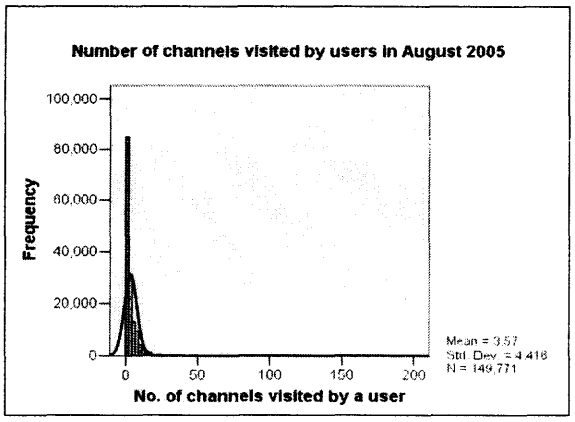

Figure 7.5 g) Histogram of channels visited by users in 08/2005.

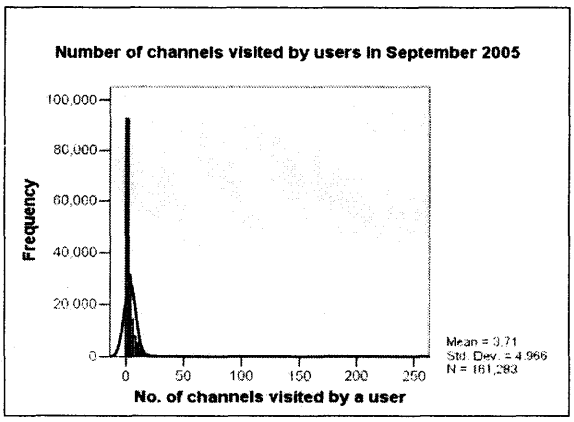

Figure 7.5 h) Histogram of channels visited by users in 09/2005.

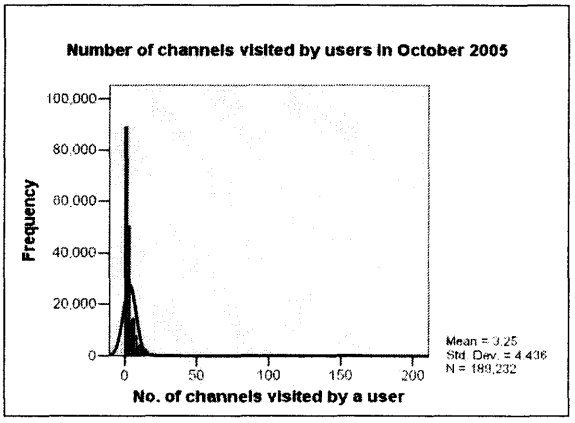

Figure 7.5 i) Histogram of channels visited by users in 10/2005.

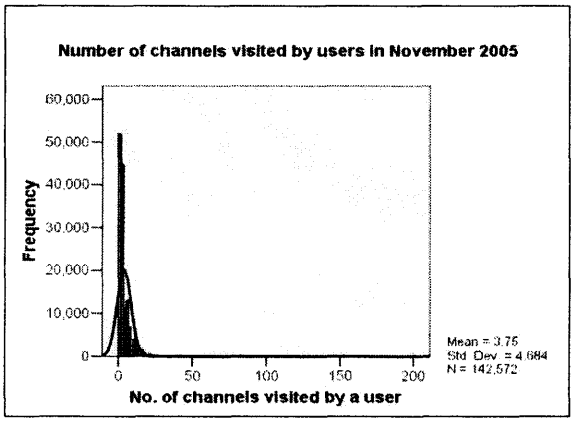

Figure 7.5 j) Histogram of channels visited by users in 11/2005.

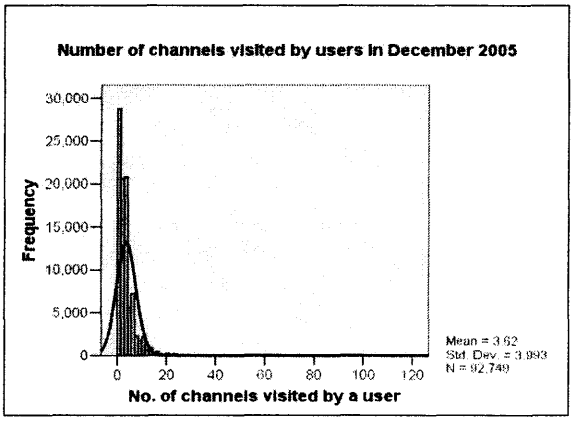

Figure 7.5 k) Histogram of channels visited by users in 12/2005.

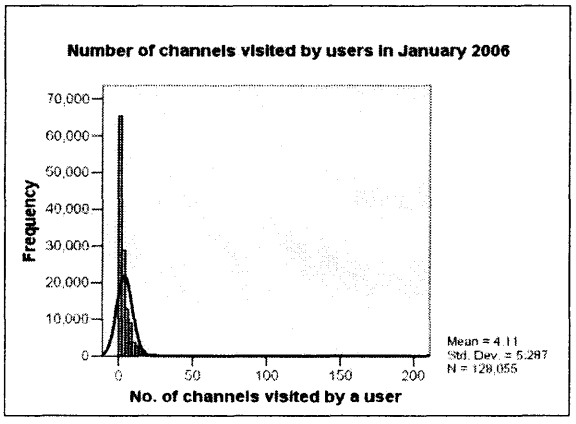

Figure 7.5 l) Histogram of channels visited by users in 01/2006.

Table 7.5 displays the most important descriptive statistics of this variable.

Table 7.5 Descriptive Statistics – Channels Visited by Users per Month

	Mean	Median	Mode	StDev	Min	Max	Range	Percentiles						Total users
								25%	50%	75%	90%	95%	99%	
Feb	3.90	2	1	5.034	1	210	209	1	2	5	8	12	22	168,171
Mar	4.20	3	1	5.451	1	195	194	1	3	5	9	13	24	184,709
Apr	4.14	2	1	5.728	1	372	371	1	2	5	9	13	25	171,034
May	2.86	2	1	2.995	1	132	131	1	2	3	6	8	15	64,402
Jun	3.70	2	1	5.103	1	243	242	1	2	4	8	12	22	167,079
Jul	3.78	2	1	4.932	1	198	197	1	2	4	8	12	22	163,682
Aug	3.57	2	1	4.416	1	187	186	1	2	4	8	11	20	149,771
Sep	3.71	2	1	4.966	1	202	201	1	2	4	8	12	22	161,283
Oct	3.25	2	1	4.436	1	167	166	1	2	4	7	11	20	189,232
Nov	3.75	2	1	4.684	1	167	166	1	2	4	9	12	22	142,572
Dec	3.62	2	1	3.993	1	119	118	1	2	4	8	10	20	92,749
Jan	4.11	2	1	5.287	1	182	181	1	2	5	9	13	25	128,055
Total	3.93	2	1	7.467	1	729	728	1	2	4	8	13	31	1,115,141

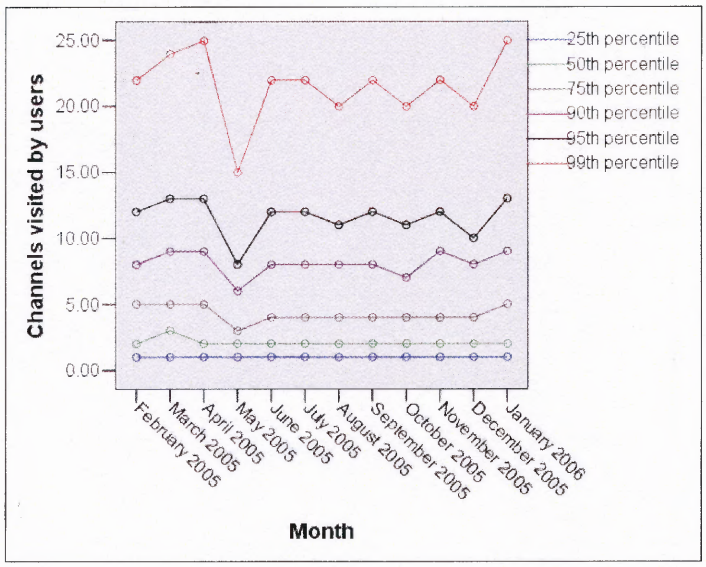

Figure 7.6 Percentile categories for the number of channels visited by users.

Figure 7.6 plots the values of the six percentile categories found in Table 7.5. It may be noted that these values remained relatively constant throughout the year. Overall, 50 percent of the users visited two channels at the most; and 90 percent of the users visited eight channels at the most.

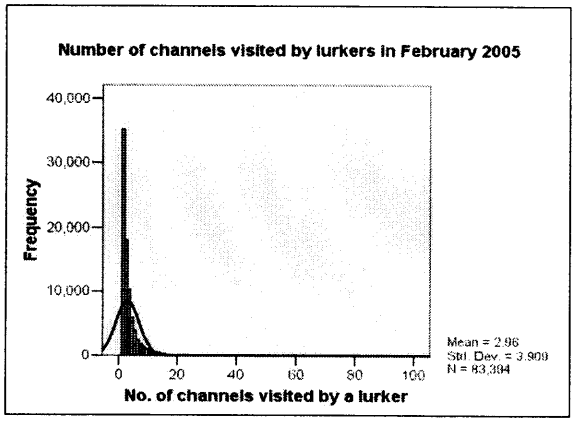

Figure 7.7 a) Histogram of channels visited by lurkers in 02/2005.

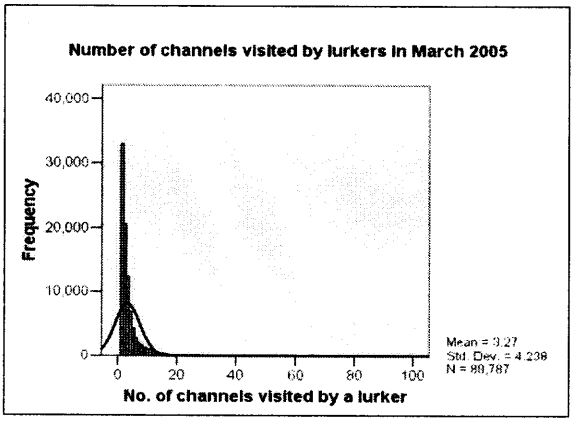

Figure 7.7 b) Histogram of channels visited by lurkers in 03/2005.

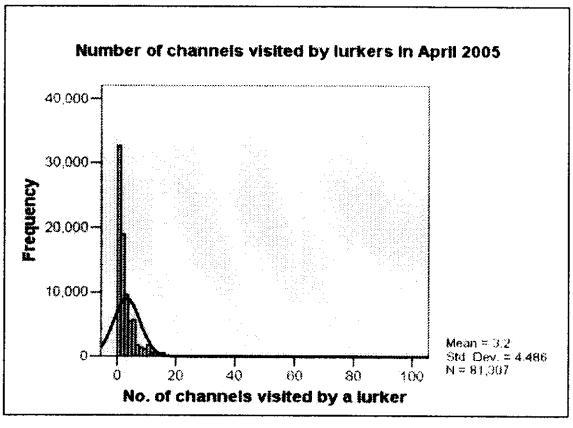

Figure 7.7 c) Histogram of channels visited by lurkers in 04/2005.

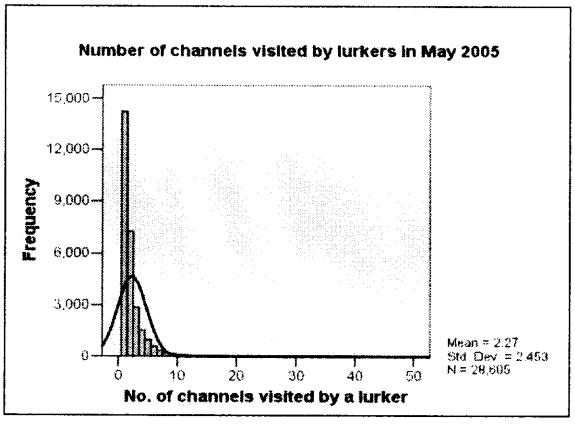

Figure 7.7 d) Histogram of channels visited by lurkers in 05/2005.

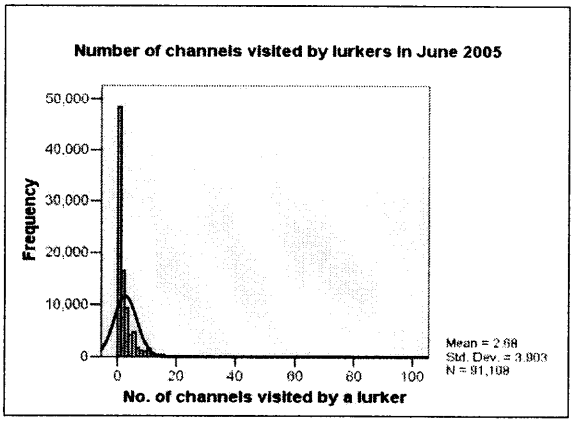

Figure 7.7 e) Histogram of channels visited by lurkers in 06/2005.

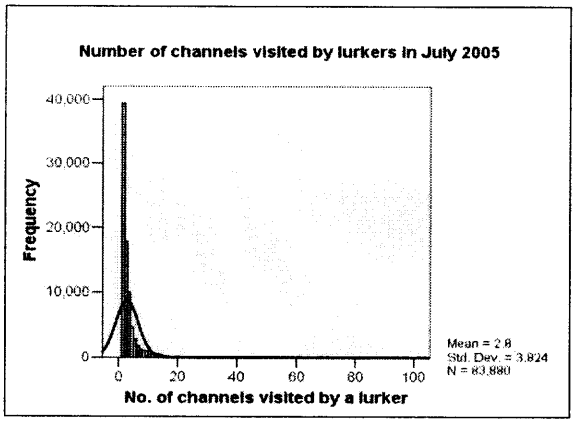

Figure 7.7 f) Histogram of channels visited by lurkers in 07/2005.

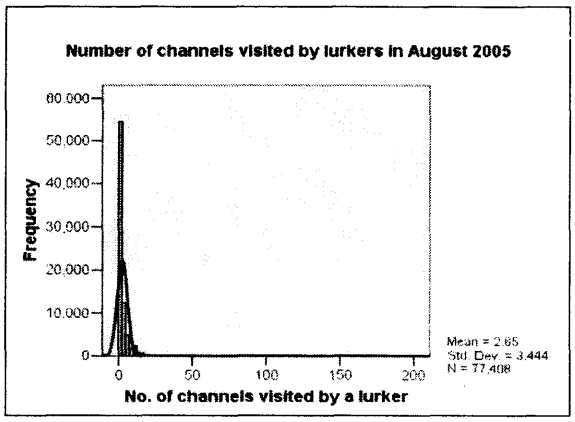

Figure 7.7 g) Histogram of channels visited by lurkers in 08/2005.

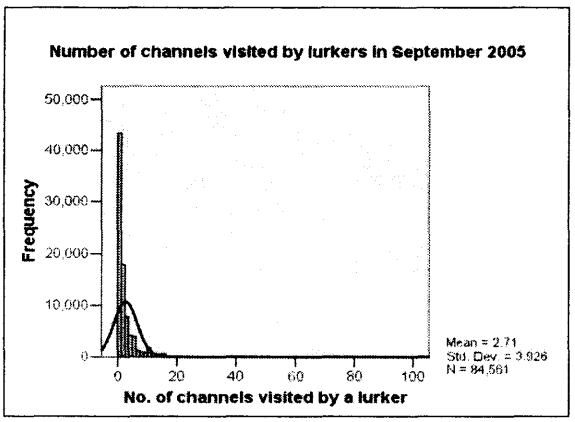

Figure 7.7 h) Histogram of channels visited by lurkers in 09/2005.

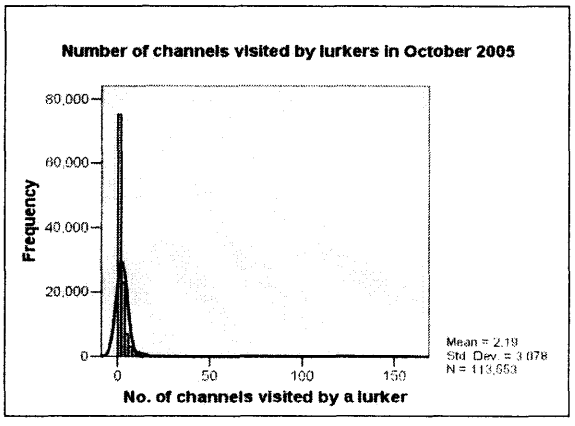

Figure 7.7 i) Histogram of channels visited by lurkers in 10/2005.

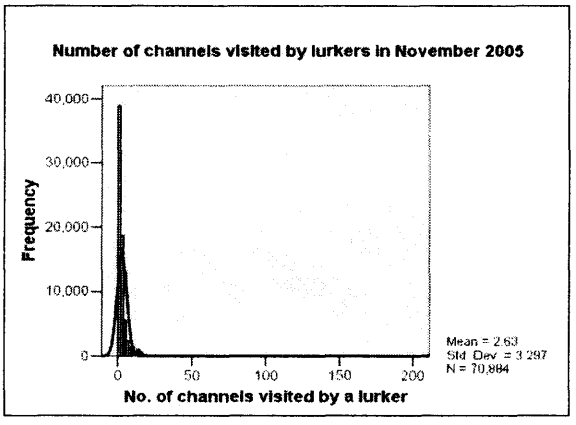

Figure 7.7 j) Histogram of channels visited by lurkers in 11/2005.

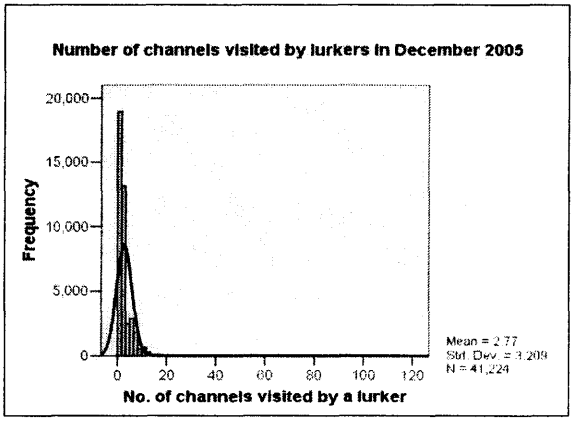

Figure 7.7 k) Histogram of channels visited by lurkers in 12/2005.

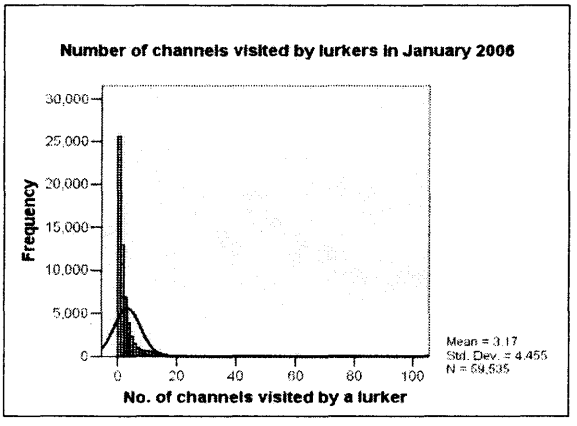

Figure 7.7 l) Histogram of channels visited by lurkers in 01/2006.

Figures 7.7 (a - l) display the monthly distributions of the number of channels visited by lurkers during the study period. As in the case of channels visited by users, the distribution is highly skewed, showing that most of the lurkers visited very few channels during any given month. Table 7.6 displays the most important descriptive statistics of this variable. Figure 7.8 plots the values of the six percentile categories found in Table 7.6 and shows, among other things, that for the entire year 50 percent of the lurkers visited a single channel, and 90 percent of the lurkers visited at most five channels.

Table 7.6 Descriptive Statistics – Channels Visited by Lurkers per Month

	Mean	Median	Mode	StDev	Min	Max	Range	Percentiles						Total lurkers
								25%	50%	75%	90%	95%	99%	
Feb	2.96	2	1	3.909	1	89	88	1	2	3	6	9	15	83,394
Mar	3.27	2	1	4.238	1	90	89	1	2	4	7	10	17	89,787
Apr	3.2	2	1	4.486	1	91	90	1	2	3	7	10	17	81,307
May	2.27	2	1	2.453	1	41	40	1	2	2	4	6	12	28,605
Jun	2.68	1	1	3.903	1	97	96	1	1	3	6	9	15	91,108
Jul	2.80	2	1	3.824	1	85	84	1	2	3	6	10	15	83,880
Aug	2.65	1	1	3.444	1	187	186	1	1	3	6	9	15	77,408
Sep	2.71	1	1	3.926	1	97	96	1	1	3	6	10	16	84,561
Oct	2.19	1	1	3.078	1	157	156	1	1	2	4	7	14	113,553
Nov	2.63	1	1	3.297	1	167	166	1	1	3	6	10	15	70,884
Dec	2.77	2	1	3.209	1	114	113	1	2	3	6	8	15	41,224
Jan	3.17	2	1	4.455	1	82	81	1	2	3	7	11	18	59,535
Total	2.58	1	1	4.081	1	207	206	1	1	3	5	8	18	736,078

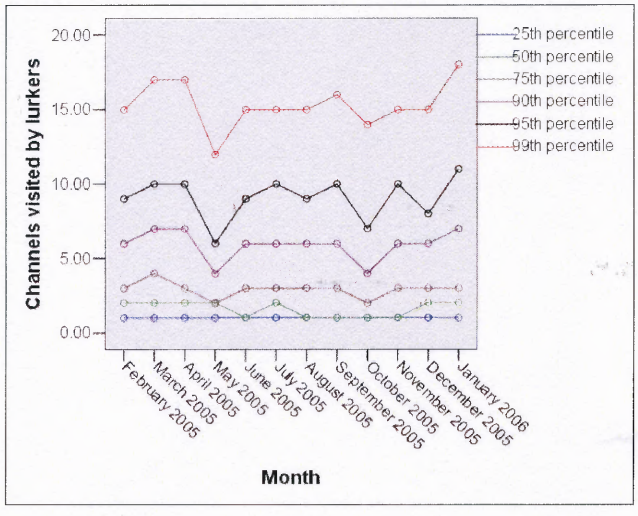

Figure 7.8 Percentile categories for the number of channels visited by lurkers.

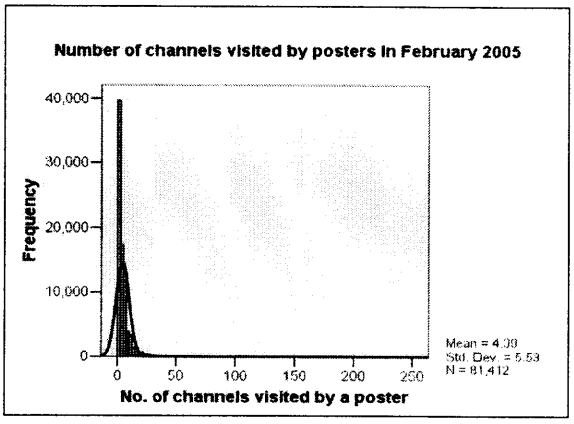

Figure 7.9 a) Histogram of channels visited by posters in 02/2005.

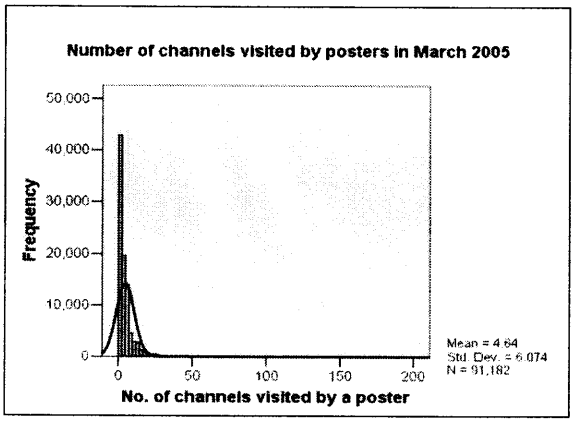

Figure 7.9 b) Histogram of channels visited by posters in 03/2005.

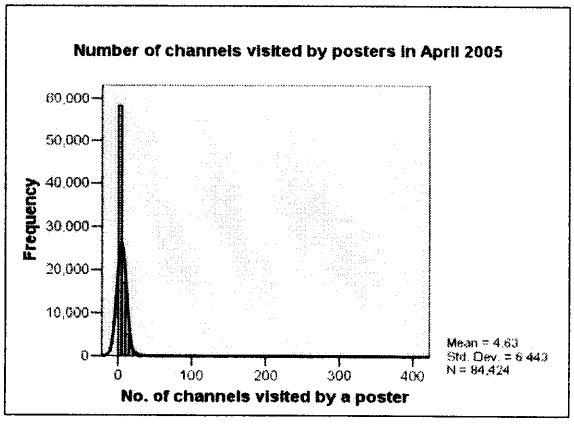

Figure 7.9 c) Histogram of channels visited by posters in 04/2005.

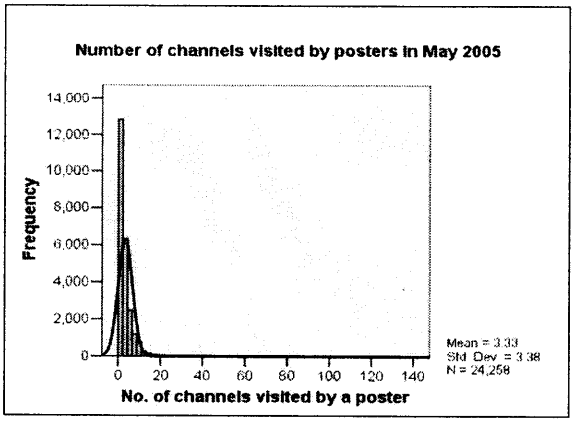

Figure 7.9 d) Histogram of channels visited by posters in 05/2005.

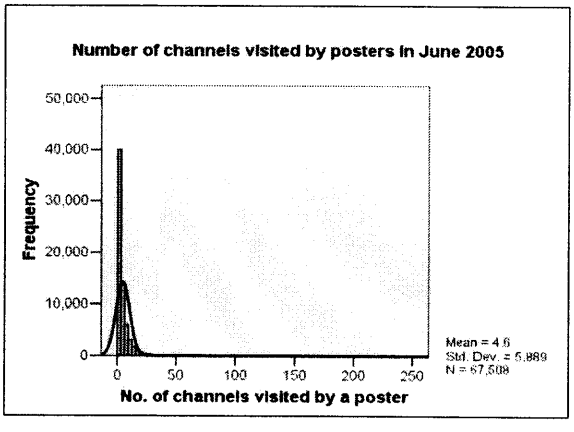

Figure 7.9 e) Histogram of channels visited by posters in 06/2005.

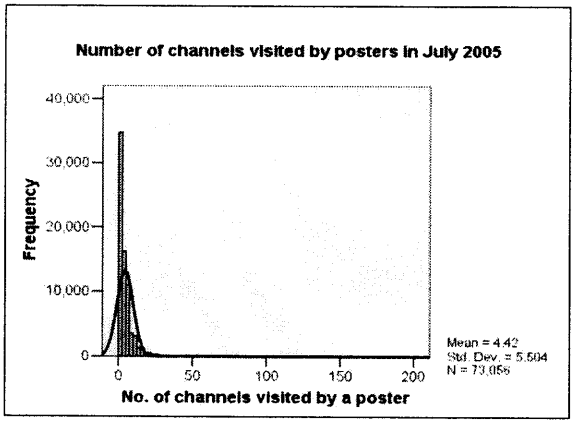

Figure 7.9 f) Histogram of channels visited by posters in 07/2005.

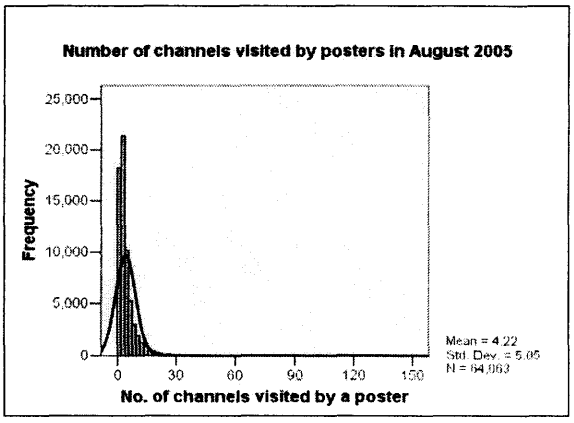

Figure 7.9 g) Histogram of channels visited by posters in 08/2005.

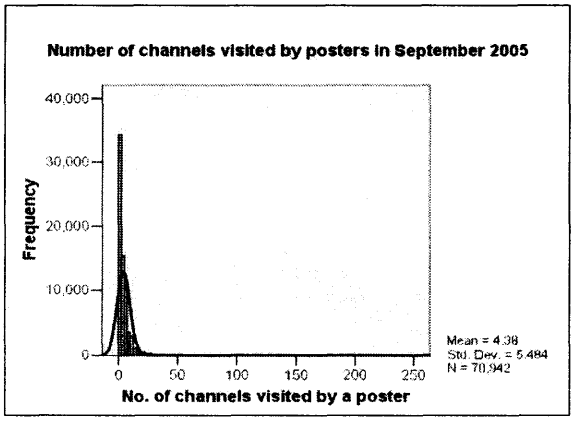

Figure 7.9 h) Histogram of channels visited by posters in 09/2005.

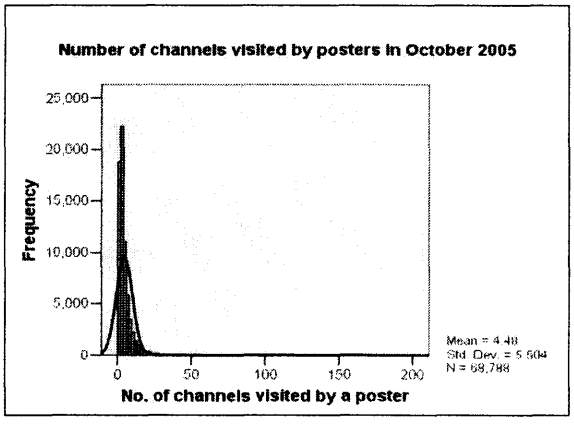

Figure 7.9 i) Histogram of channels visited by posters in 10/2005.

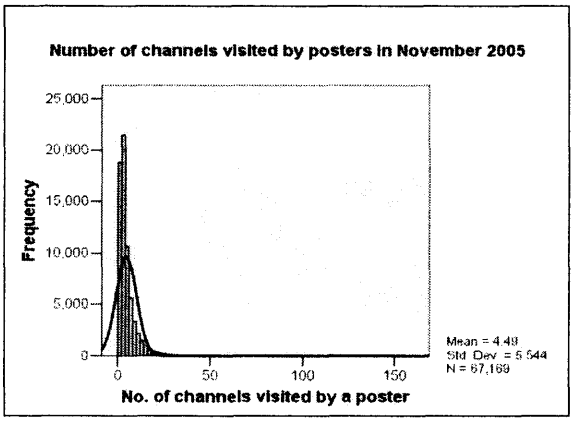

Figure 7.9 j) Histogram of channels visited by posters in 11/2005.

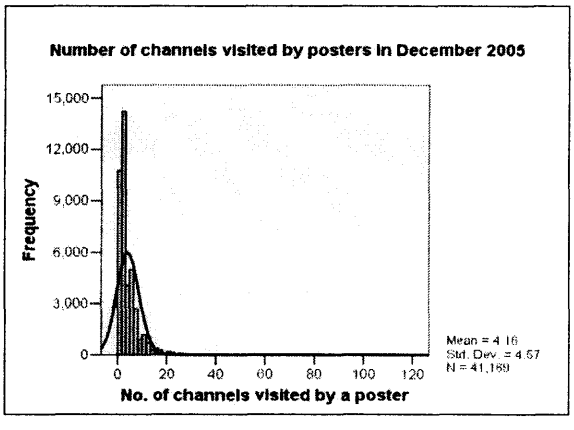

Figure 7.9 k) Histogram of channels visited by posters in 12/2005.

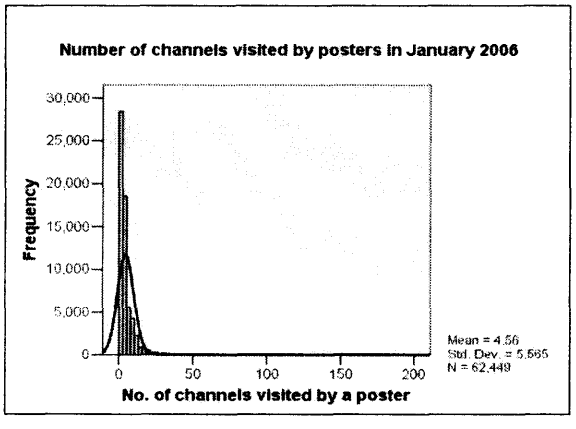

Figure 7.9 l) Histogram of channels visited by posters in 01/2006.

Figures 7.9 (a - l) display the monthly distributions of the number of channels visited by posters during the study period. As in the previous cases of channels visited by users and lurkers, the distribution is highly skewed, showing that most of the posters visited very few channels during any given month. Table 7.7 displays the most important descriptive statistics of this variable. Figure 7.10 plots the values of the six percentile categories in Table 7.7 and shows that over the year, 50 percent of the posters visited at most three channels, and 90 percent of the posters visited at most twelve channels.

Table 7.7 Descriptive Statistics – Channels Visited by Posters per Month

								Percentiles						
	Mean	Median	Mode	StDev	Min	Max	Range	25%	50%	75%	90%	95%	99%	Total posters
Feb	4.39	3	1	5.530	1	210	209	1	3	5	10	14	26	81,412
Mar	4.64	3	1	6.074	1	195	194	1	3	6	10	15	28	91,182
Apr	4.63	3	1	6.443	1	372	371	1	3	5	10	15	29	84,424
May	3.33	2	1	3.380	1	132	131	1	2	4	7	9	16	24,258
Jun	4.60	3	1	5.889	1	243	242	1	3	5	10	14	27	67,508
Jul	4.42	3	1	5.504	1	198	197	1	3	5	10	14	26	73,056
Aug	4.22	3	1	5.050	1	141	140	1	3	5	9	13	24	64,063
Sep	4.38	3	1	5.484	1	202	201	1	3	5	10	14	26	70,942
Oct	4.48	3	1	5.504	1	167	166	1	3	5	10	14	26	68,788
Nov	4.49	3	1	5.544	1	157	156	1	3	5	10	14	26	67,169
Dec	4.16	3	1	4.570	1	119	118	1	3	5	9	12	23	41,169
Jan	4.56	3	1	5.565	1	182	181	1	3	5	10	14	27	62,449
Total	5.41	3	1	9.942	1	729	728	1	3	6	12	19	44	489,561

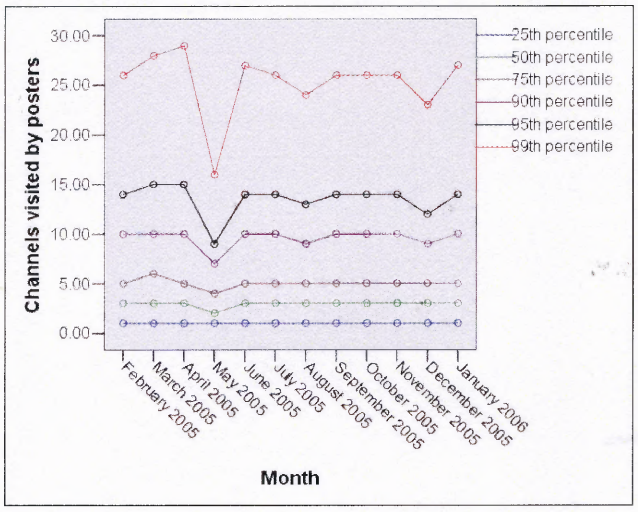

Figure 7.10 Percentile categories for the number of channels visited by posters.

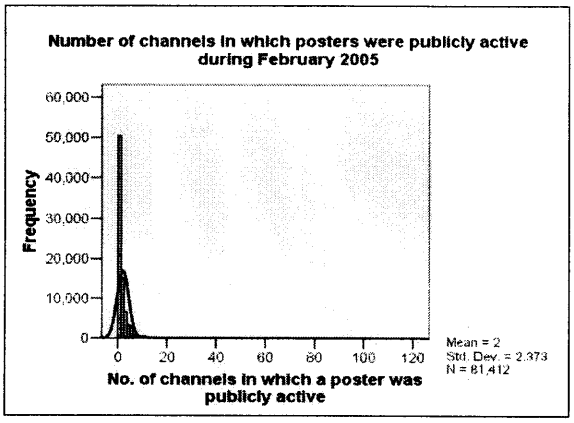

Figure 7.11 a) Histogram of active channels per poster in 02/2005.

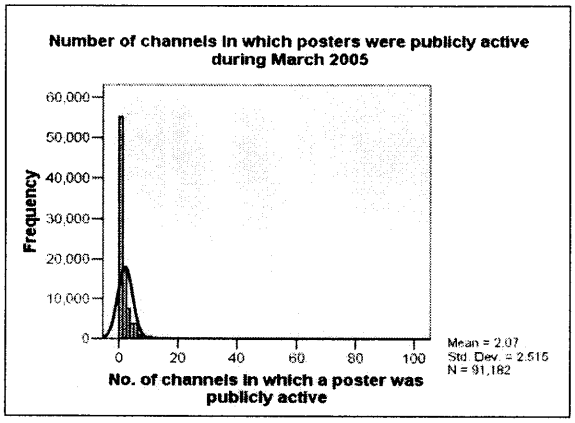

Figure 7.11 b) Histogram of active channels per poster in 03/2005.

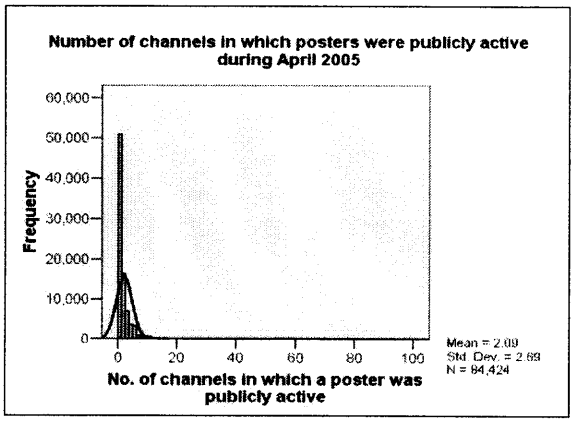

Figure 7.11 c) Histogram of active channels per poster in 04/2005.

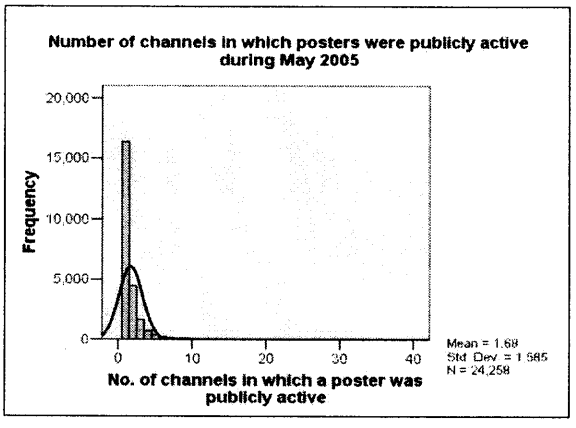

Figure 7.11 d) Histogram of active channels per poster in 05/2005.

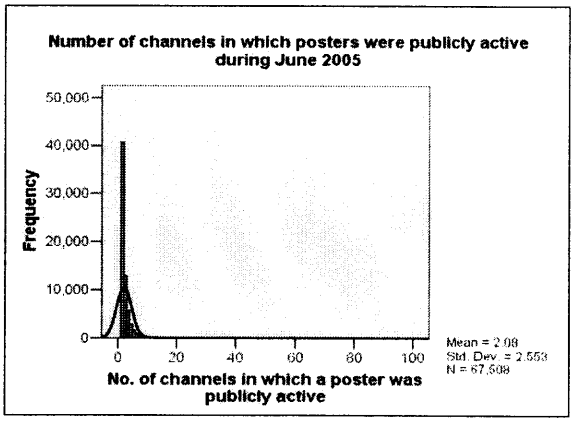

Figure 7.11 e) Histogram of active channels per poster in 06/2005.

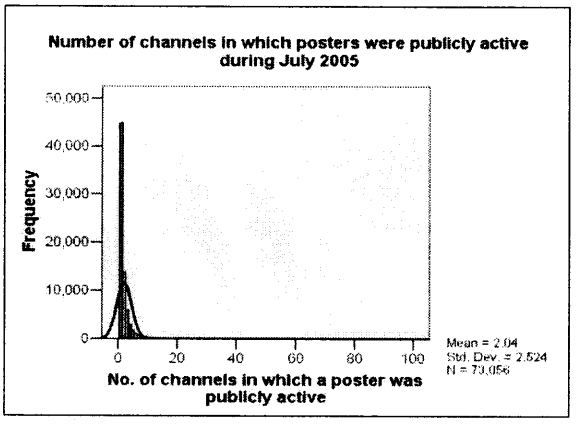

Figure 7.11 f) Histogram of active channels per poster in 07/2005.

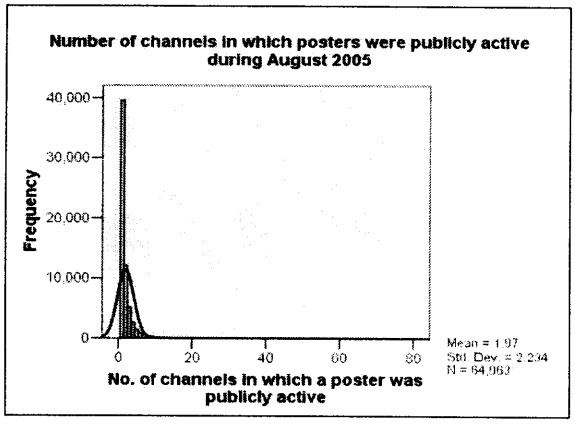

Figure 7.11 g) Histogram of active channels per poster in 08/2005.

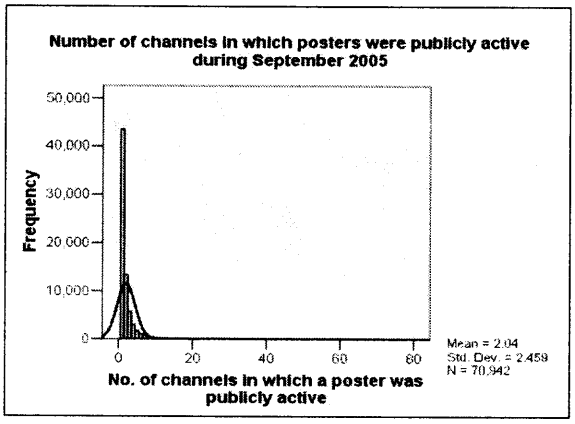

Figure 7.11 h) Histogram of active channels per poster in 09/2005.

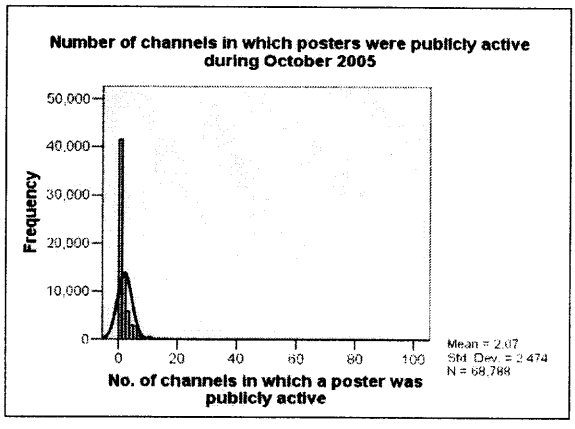

Figure 7.11 i) Histogram of active channels per poster in 10/2005.

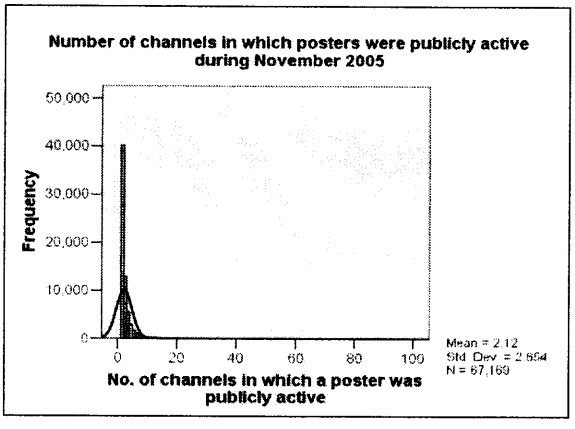

Figure 7.11 j) Histogram of active channels per poster in 11/2005.

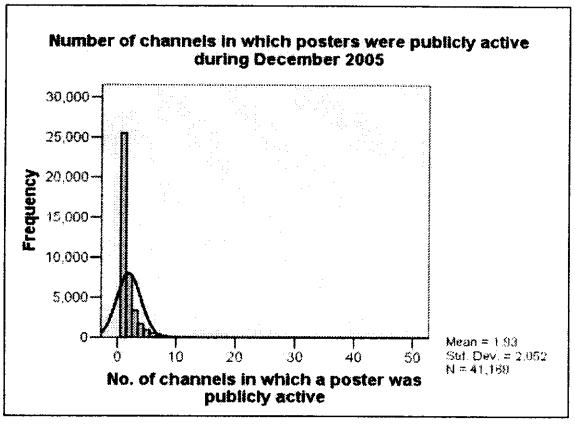

Figure 7.11 k) Histogram of active channels per poster in 12/2005.

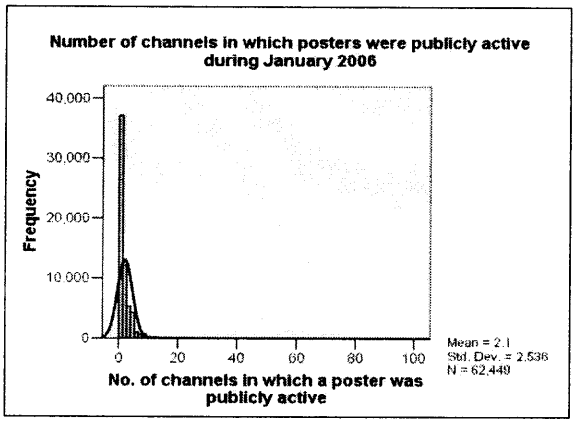

Figure 7.11 l) Histogram of active channels per poster in 01/2006.

Figures 7.11 (a - l) display the monthly distributions of the number of channels in which posters were active during the study period. As in the previous cases, the distribution is highly skewed, showing that most of the posters were active in very few channels during any given month. Table 7.8 displays the most important descriptive statistics of this variable. Figure 7.12 plots the values of the six percentile categories found in Table 7.8, and shows that over the year, 50 percent of the posters were active in a single channel, and 90 percent of the posters were active in four channels at the most.

Table 7.8 Descriptive Statistics – Channels in which Posters Were Active per Month

	Mean	Median	Mode	StDev	Min	Max	Percentiles						Total posters
							25%	50%	75%	90%	95%	99%	
Feb	2.00	1	1	2.373	1	104	1	1	2	4	6	12	81,412
Mar	2.07	1	1	2.515	1	99	1	1	2	4	6	12	91,182
Apr	2.09	1	1	2.690	1	97	1	1	2	4	6	13	84,424
May	1.68	1	1	1.585	1	39	1	1	2	3	4	8	24,258
Jun	2.09	1	1	2.553	1	86	1	1	2	4	6	13	67,508
Jul	2.04	1	1	2.524	1	81	1	1	2	4	6	12	73,056
Aug	1.97	1	1	2.234	1	76	1	1	2	4	5	11	64,063
Sep	2.04	1	1	2.459	1	78	1	1	2	4	6	12	70,942
Oct	2.07	1	1	2.474	1	94	1	1	2	4	6	12	68,788
Nov	2.12	1	1	2.654	1	87	1	1	2	4	6	13	67,169
Dec	1.93	1	1	2.052	1	43	1	1	2	4	5	11	41,169
Jan	2.10	1	1	2.536	1	92	1	1	2	4	6	13	62,449
Total	2.37	1	1	4.125	1	379	1	1	2	4	7	18	489,561

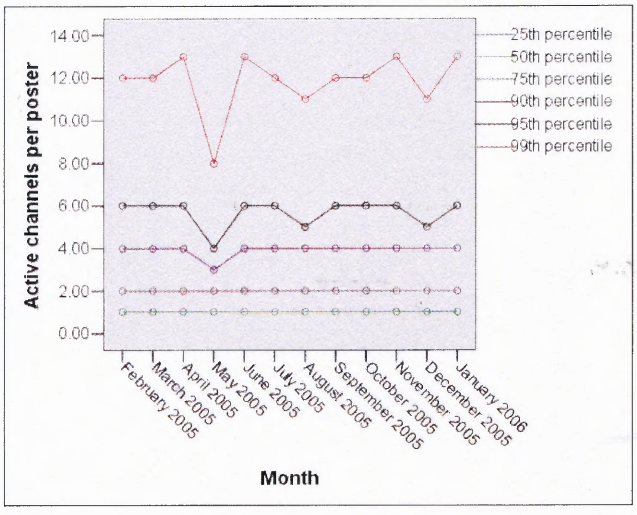

Figure 7.12 Percentile categories for the no. of channels in which posters were active.

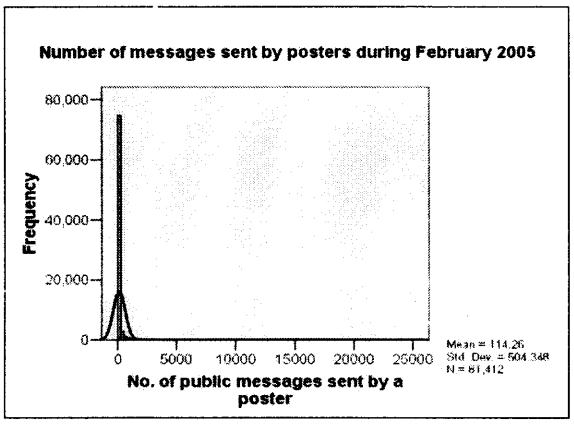

Figure 7.13 a) Histogram of messages per poster in 02/2005.

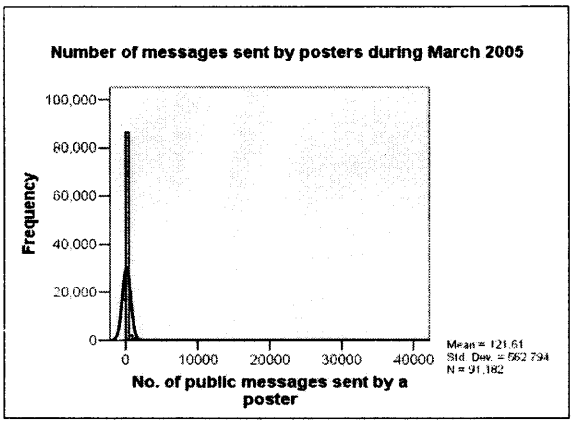

Figure 7.13 b) Histogram of messages per poster in 03/2005.

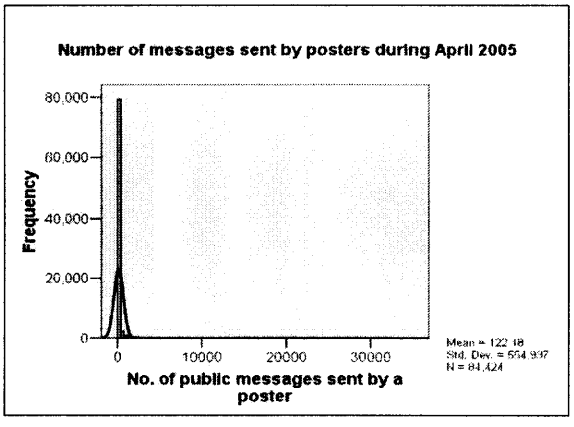

Figure 7.13 c) Histogram of messages per poster in 04/2005.

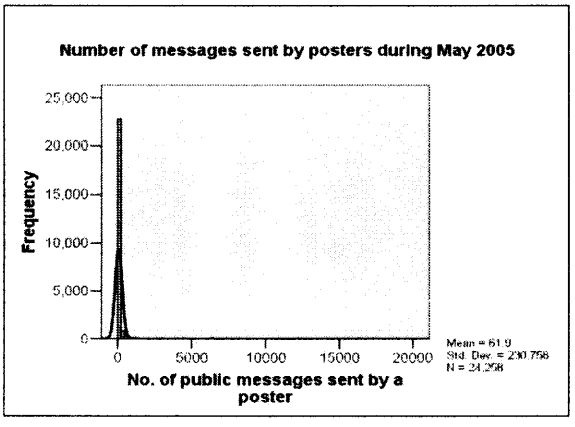

Figure 7.13 d) Histogram of messages per poster in 05/2005.

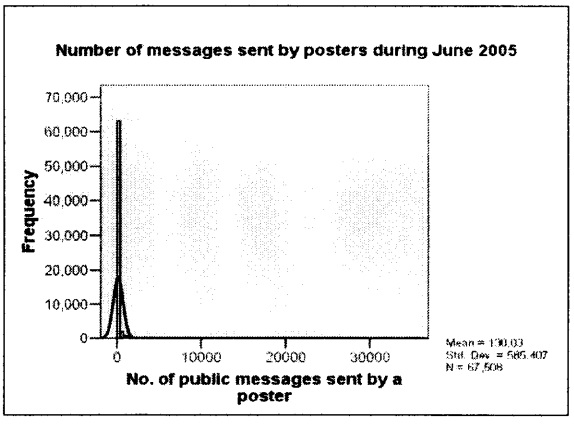

Figure 7.13 e) Histogram of messages per poster in 06/2005.

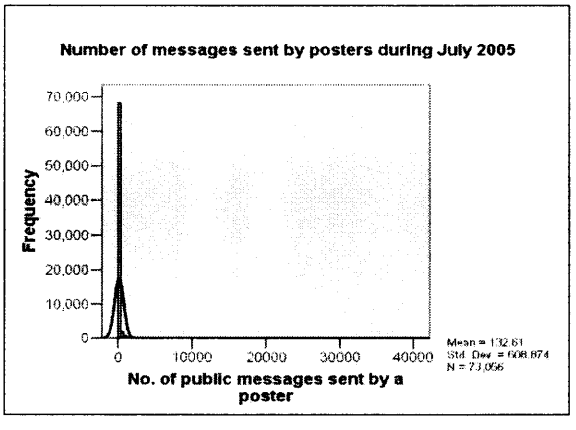

Figure 7.13 f) Histogram of messages per poster in 07/2005.

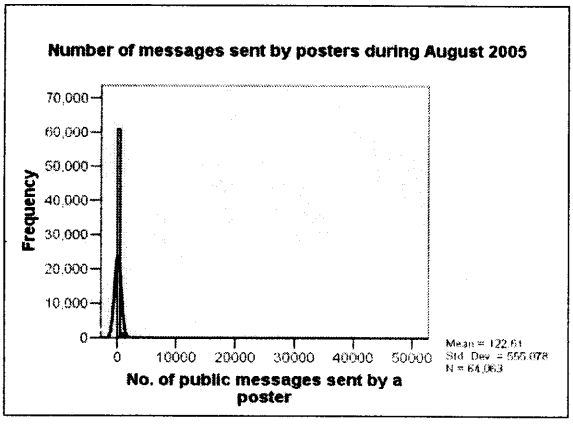

Figure 7.13 g) Histogram of messages per poster in 08/2005.

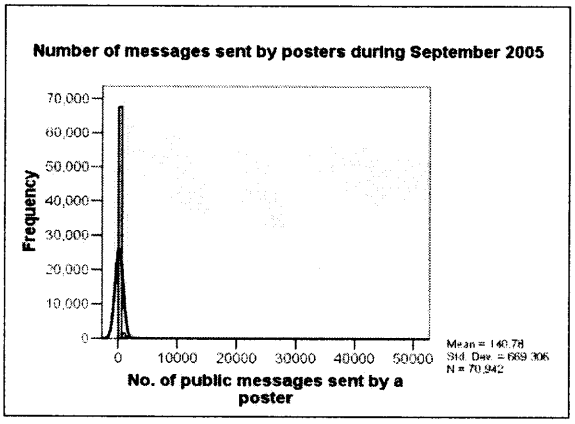

Figure 7.13 h) Histogram of messages per poster in 09/2005.

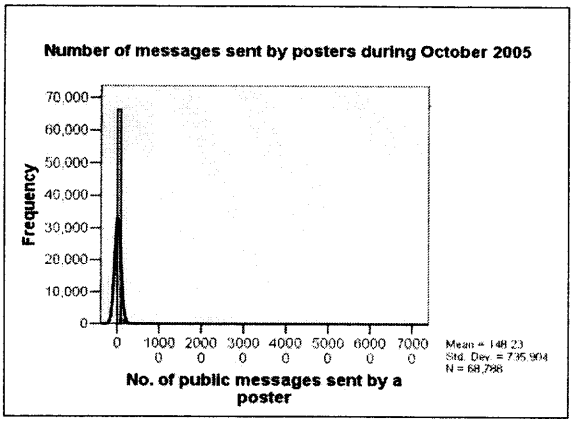

Figure 7.13 i) Histogram of messages per poster in 10/2005.

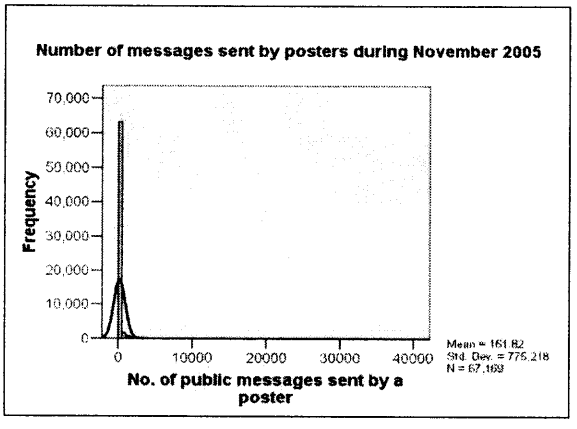

Figure 7.13 j) Histogram of messages per poster in 11/2005.

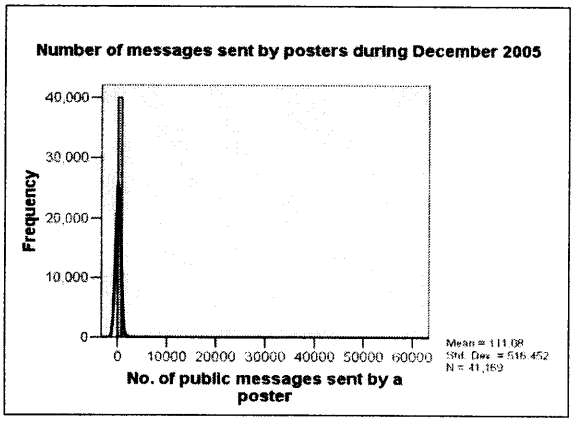

Figure 7.13 k) Histogram of messages per poster in 12/2005.

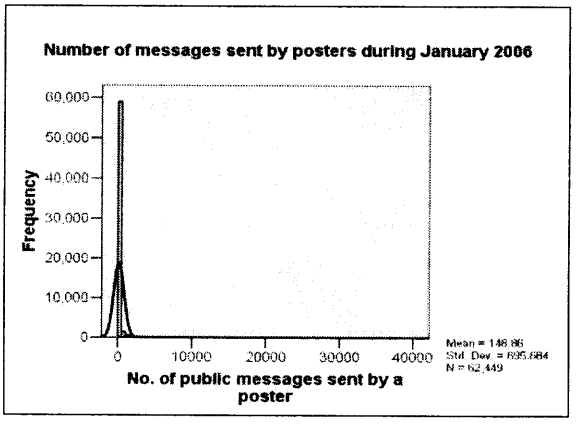

Figure 7.13 l) Histogram of messages per poster in 01/2006.

Figures 7.13 (a - l) display the monthly distributions of the number of public messages sent by posters during the study period. As in the previous cases of visited channels and publicly active channels, the distribution is highly skewed, showing that most of the posters were responsible for small numbers of public messages during any given month. Table 7.9 displays the most important descriptive statistics of this variable.

Table 7.9 Descriptive Statistics – Public Messages Sent by Posters per Month

	Mean	Median	Mode	StDev	Min	Max	Percentiles						Total posters
							25%	50%	75%	90%	95%	99%	
Feb	114.3	10	1	504.348	1	22233	2	10	45	191	470	2088	81,412
Mar	121.6	10	1	562.794	1	37326	3	10	46	194	480	2255	91,182
Apr	122.2	10	1	554.937	1	29667	3	10	46	194	489	2272	84,424
May	61.9	8	1	230.758	1	16089	2	8	35	131	280	950	24,258
Jun	130.0	9	1	585.407	1	30259	2	9	45	208	552	2494	67,508
Jul	132.6	9	1	608.874	1	33560	2	9	43	197	537	2586	73,056
Aug	122.6	9	1	555.078	1	41292	2	9	43	198	510	2305	64,063
Sep	140.8	9	1	669.306	1	46416	2	9	44	205	573	2767	70,942
Oct	148.2	9	1	735.904	1	67511	2	9	46	219	601	2920	68,788
Nov	161.8	10	1	775.218	1	37049	3	10	49	230	645	3195	67,169
Dec	111.0	9	1	516.452	1	55647	2	9	47	205	498	1936	41,169
Jan	148.9	10	1	695.684	1	35802	2	10	48	216	604	2930	62,449
Total	211.3	8	1	1910.322	1	201532	2	8	38	156	415	4081	489,561

Figure 7.14 plots the values of the six percentile categories found in Table 7.9. The values remained relatively constant throughout the year. For the entire year, 50 percent of the posters sent at most eight messages, and 90 percent of the posters sent at most 156 messages.

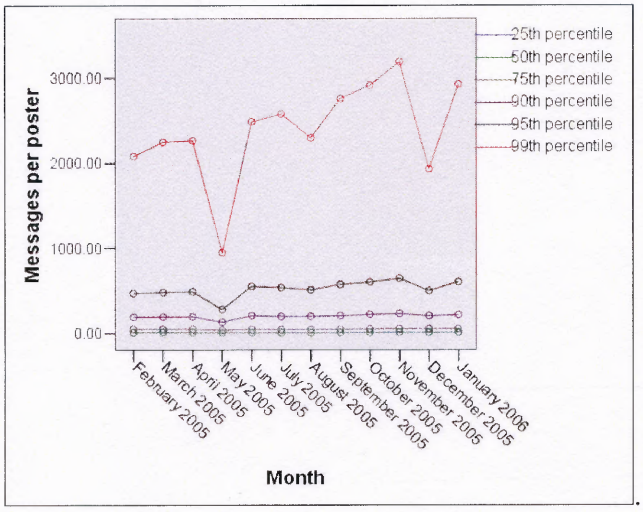

Figure 7.14 Percentile categories for the number of messages sent by posters.

Table 7.9 as well as Figures 7.13 and 7.14 show that the distribution of messages per poster in an IRC network is similar to the distributions found in asynchronous CMC systems such as Usenet groups, in that a small number of users are responsible for most of the mass interaction (Whitaker et al. 1998). Table 7.10 summarizes the number of messages sent by the most active posters, expressed as percentages of the total number of monthly or yearly messages.

Table 7.10 Top Posters and Messages Percentages

	Top 10% posters	Top 5% posters	Top 3% posters	Top 1% posters
February	81.6%	68.7%	58.1%	36.1%
March	82.3%	70.1%	59.5%	37.4%
April	82.2%	69.9%	59.2%	36.9%
May	74.4%	59.1%	48.0%	28.1%
June	83.5%	70.6%	59.6%	36.6%
July	84.5%	72.4%	61.5%	37.6%
August	83.2%	70.4%	59.5%	36.4%
September	85.2%	73.1%	62.2%	38.5%
October	85.0%	73.1%	62.2%	38.7%
November	85.5%	73.8%	63.1%	39.3%
December	80.2%	66.1%	54.7%	32.5%
January	84.8%	72.8%	61.8%	38.2%
Monthly average	82.7%	70.0%	59.1%	36.3%
Yearly	91.6%	85.8%	80.3%	63.3%

Figure 7.15 plots the values of the four most active poster categories found in Table 7.10. The values remained relatively constant throughout the year if the outliers for the months of May 2005 and December 2005 are not considered.

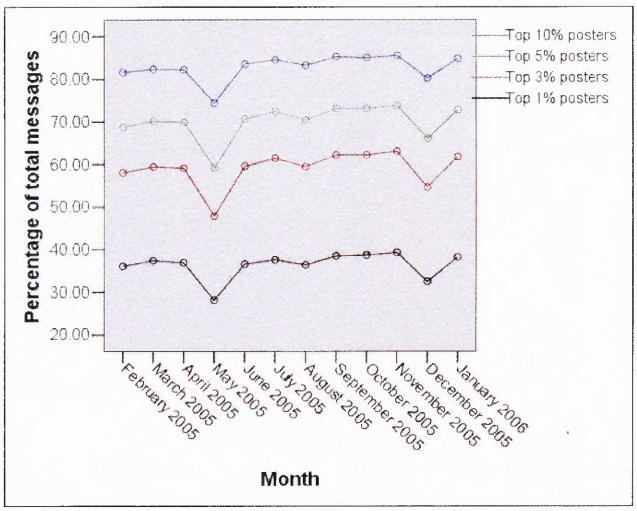

Figure 7.15 Percentage of total messages originated by most active posters.

It may be noted that during any given month, the most active 10 percent of the posters were responsible for approximately 80 percent of the total number of monthly public messages. The most active 5 percent of the posters were responsible for approximately 70 percent of the total number of monthly public messages. The most active 3 percent of the posters were responsible for approximately 59 percent of the total number of monthly public messages; and the most active 1 percent of the posters were responsible for approximately 36 percent of the total number of monthly public messages. The percentage of messages for the most active poster categories increases significantly if the entire year is considered. The most active 10 percent of the posters were responsible for approximately 91 percent of the total number of public messages. The most active 5 percent of the posters were responsible for approximately 86 percent of the total number of public messages. The most active 3 percent of the posters were

responsible for approximately 80 percent of the total number of public messages; and the most active 1 percent of the posters were responsible for approximately 63 percent of the total number of public messages.

Figures 7.16 (a - c) describe the relationships among several user-related variables over the entire study period. Specifically, the number of messages, the number of channels a user visited, the number of channels a user was active in, the number of months a user visited the network, and the number of months a user was publicly active are examined. For each pair of variables the Spearman's correlation coefficient was also computed. As mentioned in Section 7.2, both the Pearson and Spearman correlation coefficients can be used to measure the correlation between two variables. However, since the Spearman correlation coefficient examines the ranks of ordinal data rather than the values themselves, it is the one preferred in this analysis. The highest observed correlation was between the number of channels a user visited (active channels) and the number of channels a user was active in (publicly active channels). In this case Spearman's rho was 0.675, suggesting that as users visited more channels, they also tended to be publicly active in those channels. However, since the correlation is not very high, there were still a significant number of users who preferred to be publicly active in only a few of the channels they visited. A medium correlation was observed between the number of messages and the number of publicly active channels: Spearman's rho = 0.591. Only in about half the cases, an increase in the number of messages was correlated with an increase in the number of channels in which users were publicly active. The lowest observed correlation was between the number of messages and the number of visited channels: Spearman's rho = 0.392. For all three cases, n=489,561, $p < .001$.

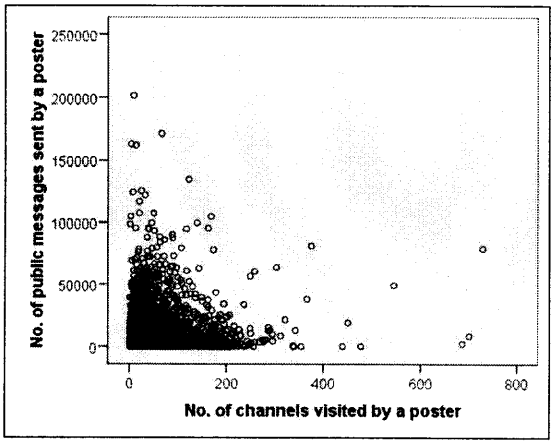

Figure 7.16 a) Messages by active channels per poster.

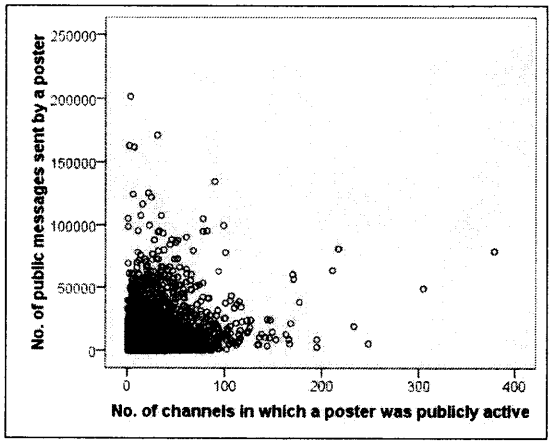

Figure 7.16 b) Messages by publicly active channels per poster.

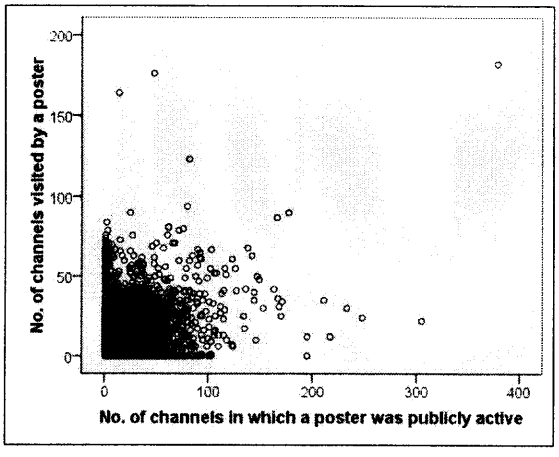

Figure 7.16 c) Active channels by publicly active channels per poster.

7.2.2.2 Dynamics of a Sample of Posters. This section describes the behavior of IRC posters. Several variables of interest were computed: the average time spent by posters in an IRC session; the average time spent by posters before starting to publicly interact after connecting to the network; the average number of days posters visited the network during a month; the average number of days posters were publicly active during a month; the average number of channels posters visited during an IRC session; and the average number of channels in which posters were publicly active during an IRC session. Table 7.11 presents the descriptive statistics of all these variables. The sample of posters was

selected from the month of August 2005 because it was in the middle of the data-collection period. Some of the variables described in Table 7.11 were computed for the entire month of August 2005, while other variables were computed only for the first week of August 2005. A smaller time interval was chosen because of the large amounts of both time and computing power required to compute some of these variables, given the size of the dataset.

Table 7.11 Poster Dynamics Data

Measure	Mean	Median	Mode	StDev	Percentiles						Sample size
					25%	50%	75%	90%	95%	99%	
Average number of channels visited by a poster during a session[1]	2.33	2	1	2.08199	1	2	3	5	6	11	13,043
Average number of channels in which a poster was publicly active during a session[1]	1.2	1	1	0.87839	1	1	1	2	3	5	10,648
Average time per session spent by a poster on IRC (in minutes)[2]	263.8	39	10	556.782	12	39	141	916	1,619	2,704	57,098
Average time until the first posted public message (in minutes)[2]	21.01	20	20	57.907	20	20	20	20	20	20	33,019
Average number of days a poster connected to the network[2]	4.36	2	1	5.955	1	2	5	13	19	27	64,063
Average number of days a poster was publicly active[2]	2.94	1	1	4.577	1	1	2	7	13	25	64,063

The trend that was visible in the general user data (described in subsection 7.2.2.1) can also be observed for this sample of posters. Most of them visited and were publicly active in very few channels during an IRC session. Approximately 75 percent of the August posters visited three or fewer channels and were publicly active in only one

[1] Computed for the first week of August 2005
[2] Computed for the entire month of August 2005

channel per session. Only 1 percent of the August posters visited eleven or more channels and were publicly active in five or more channels during a session. Posters seemed to be publicly active in approximately half of the channels they visited during a session during this month. Also, 75 percent of the August posters visited the network for five or fewer days and were publicly active for at most two days. Only 1 percent of the August posters were active for most of the month: they visited the network for 27 days or more and they were publicly active for at least 25 days. An expected high correlation was observed between the number of days posters connected to the network and the number of days posters were actually publicly active: Spearman's rho = 0.814, n=64,063, $p < .01$. This suggests that posters tend to be publicly active in most of the days they connect to the network. The correlation between the average number of channels posters visited during a session and the average number of channels in which posters were publicly active during a session was rather low: Spearman's rho = 0.378, n=10,648, $p < .01$. The histograms in Figures 7.17 (a – f) display the distributions of the variables presented in Table 7.11.

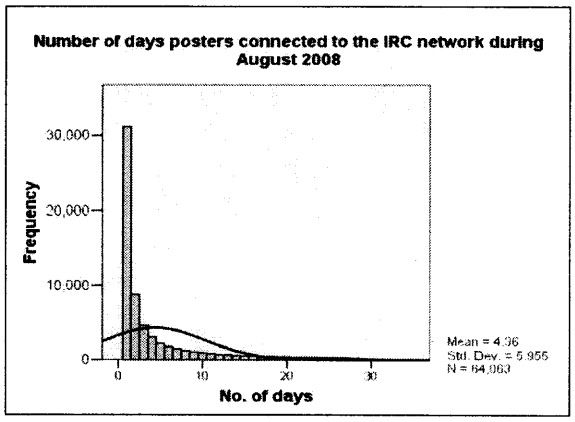

Figure 7.17 a) Histogram of number of visited days per poster in 08/2005.

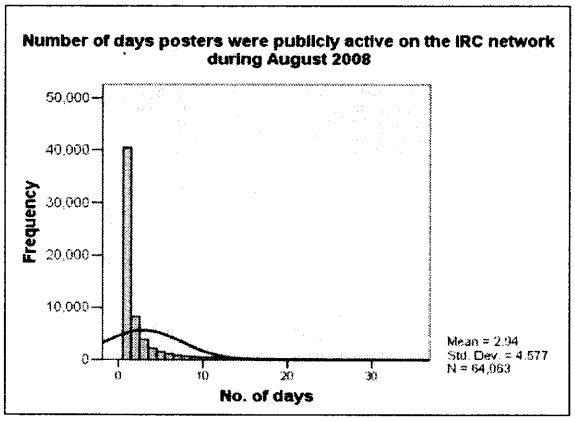

Figure 7.17 b) Histogram of number of publicly active days per poster in 08/2005.

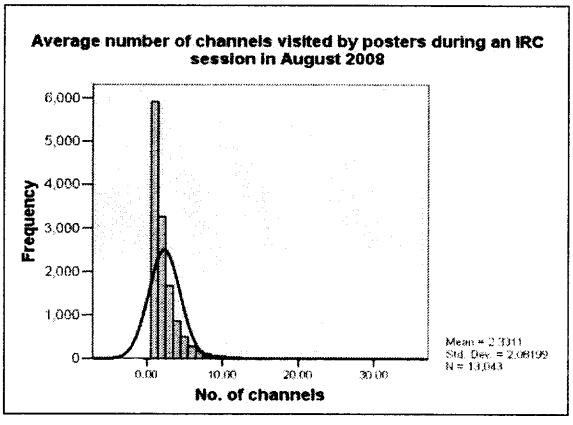

Figure 7.17 c) Histogram of visited channels per session in 08/2005.

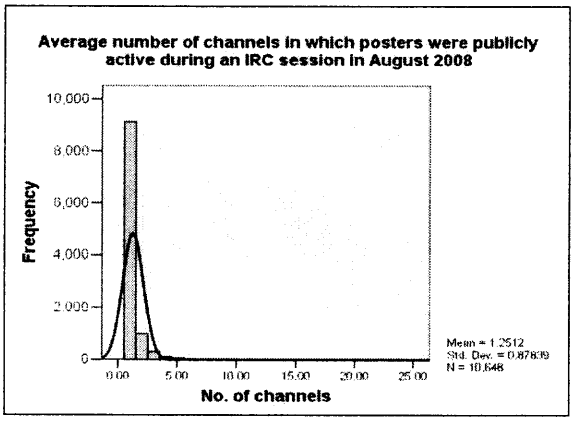

Figure 7.17 d) Histogram of publicly active channels per session in 08/2005.

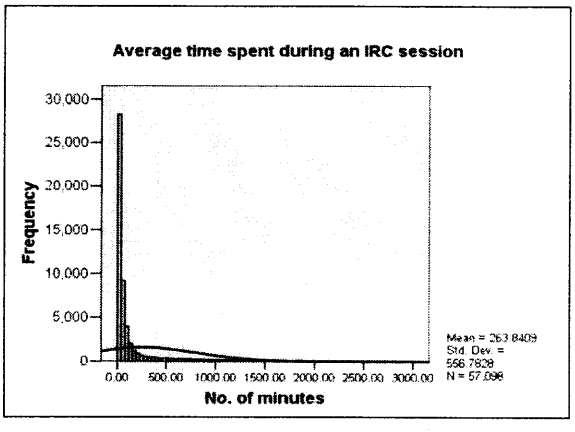

Figure 7.17 e) Histogram of average time per session in 08/2005.

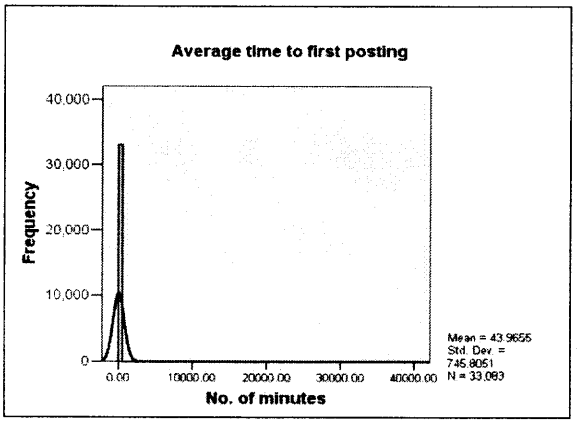

Figure 7.17 f) Histogram of average time to first posting in 08/2005.

7.2.3 Channel-related Descriptive Statistics

This section looks at various characteristics of the IRC chat-channels by examining tables, graphs, plots, and correlations. In Subsection 7.2.3.1 the focus is at the level of the entire network, and the time span is the entire year with monthly breakdowns. Some key measures of the channel activity throughout the year are presented such as the total number of months the channels were visited, the total number of users and posters that visited the channels, or the total number of public messages posted to the IRC channels. In Subsection 7.2.3.2 the focus is on the month of August and on the subset of channels that were publicly active during that interval.

7.2.3.1 General Characteristics of the IRC Chat-channels. This subsection describes the general characteristics of the channels of the IRC network. Two main categories of channels were considered: active channels (channels that were visited by users, but no public interaction occurred), and publicly active channels (channels where public messages were exchanged among users). The channels that simply existed but were not visited by users were excluded from the analysis since they presented no interest. The monthly breakdowns demonstrate if the dynamics of the channels changed over the course of one year.

Figures 7.18 (a – c) show the histograms for the number of months active channels and publicly active channels were visited, and for the number of months publicly active channels sustained public interactions.

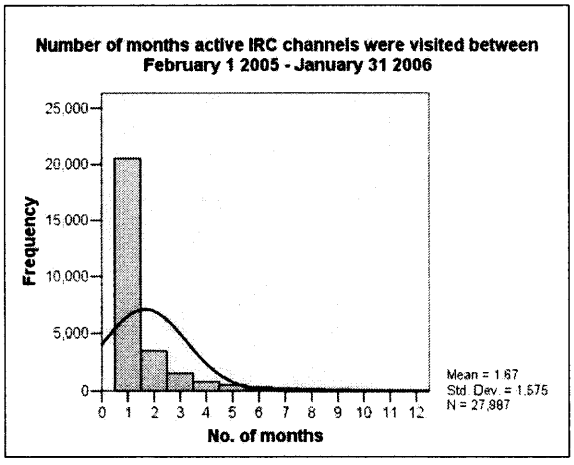

Figure 7.18 a) Histogram of number of visited months per active channel.

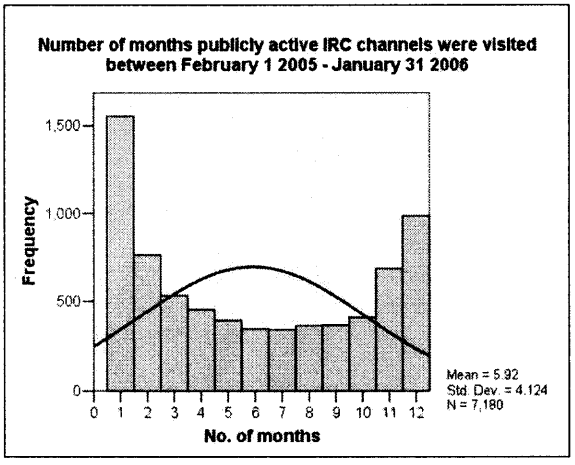

Figure 7.18 b) Histogram of number of visited months per publicly active channel.

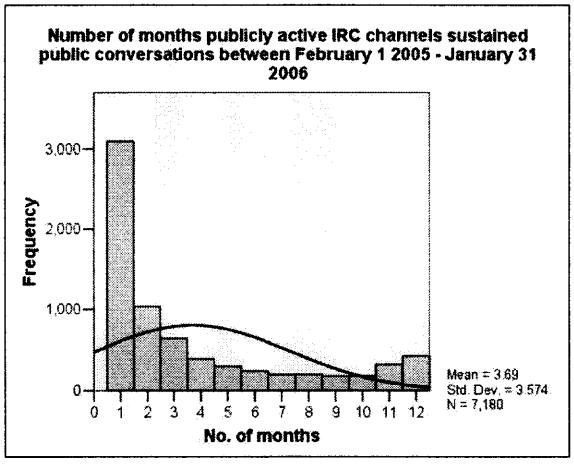

Figure 7.18 c) Histogram of number of publicly active months per publicly active channel

It may be noted that the vast majority of the active channels kept their active status for short periods of time, and very few were able to maintain their user population for more than three months. The situation was very different for publicly active channels. While there was still a large number of publicly active channels that were visited for only a month or two, there was a comparable number of channels that were visited for the entire year, as well as many other channels that were visited anywhere between three and eleven months. With respect to the number of months publicly active channels actually

hosted public interactions, one can again observe that most of them did so for short periods of time. However, there were some channels that remained publicly active for the entire year, as well as others that were publicly active for several months. This distribution suggests that channels where public discussions occur are more likely to attract users than channels that lack such interactions.

Table 7.12 Descriptive Statistics – Number of Months IRC Channels Were Active

	Mean	Median	Mode	StDev	Percentiles						Total channels
					25%	50%	75%	90%	95%	99%	
Number of months active channels were visited	1.67	1	1	1.575	1	1	2	3	5	9	27,987
Number of months publicly active channels were visited	5.92	5	1	4.124	2	5	10	12	12	12	7,180
Number of months publicly active channels hosted public	3.69	2	1	3.574	1	2	5	11	12	12	7,180

Table 7.12 presents some descriptive statistics for active and publicly active channels: the mean, median, and mode values for the number of months channels were visited or sustained public interactions. These values confirm what the histograms suggested. The typical active channels did not attract users for more than a few months as opposed to the typical publicly active channels, which were more successful at maintaining a user base for longer periods of time. The percentile distributions show very clearly that 75 percent of all the active channels were visited for two months at the most, while only 1 percent of the channel base remained active for nine or more months. It may also be observed that 10 percent of the publicly active channels were visited and sustained public interactions during (almost) the entire study period. Also, 25 percent of the publicly active channels were visited for at least 10 months and hosted public discussions for at least half the time of this interval.

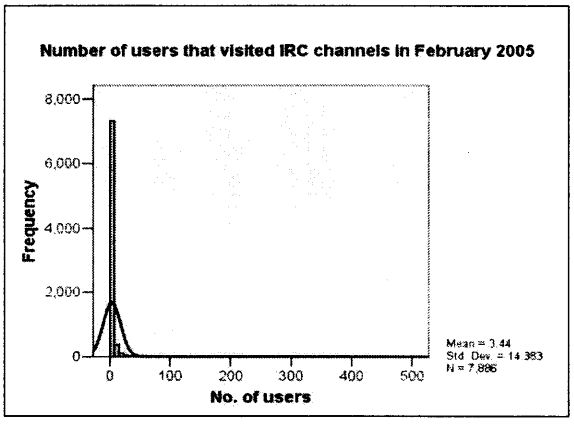

Figure 7.19 a) Histogram of users per active channel in 02/2005.

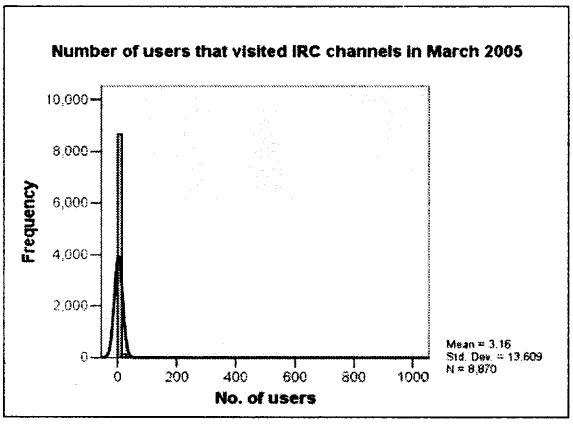

Figure 7.19 b) Histogram of users per active channel in 03/2005.

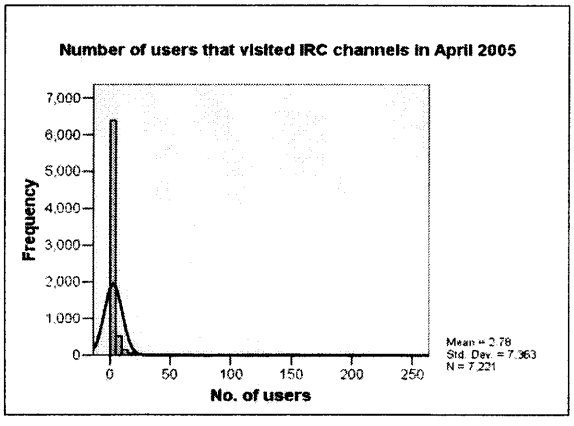

Figure 7.19 c) Histogram of users per active channel in 04/2005.

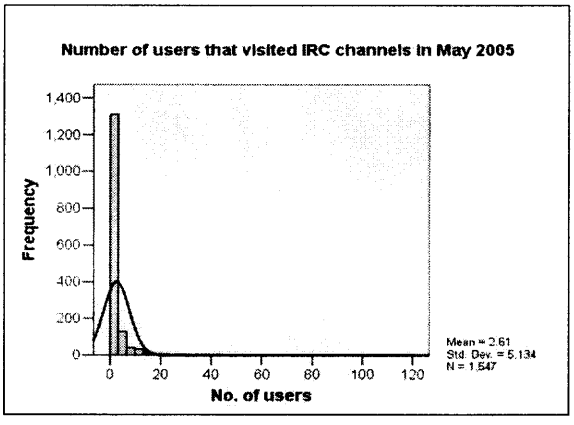

Figure 7.19 d) Histogram of users per active channel in 05/2005.

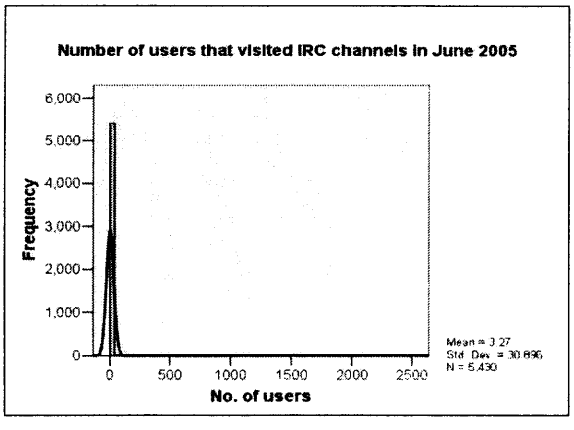

Figure 7.19 e) Histogram of users per active channel 06/2005.

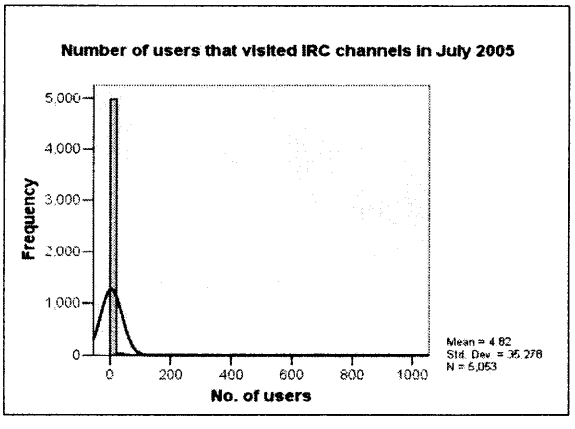

Figure 7.19 f) Histogram of users per active channel in 07/2005.

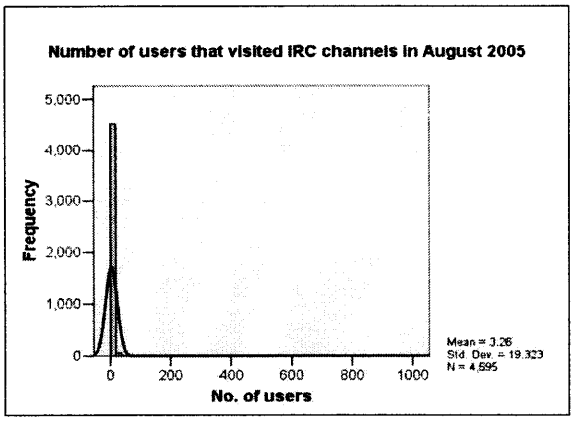

Figure 7.19 g) Histogram of users per active channel in 08/2005.

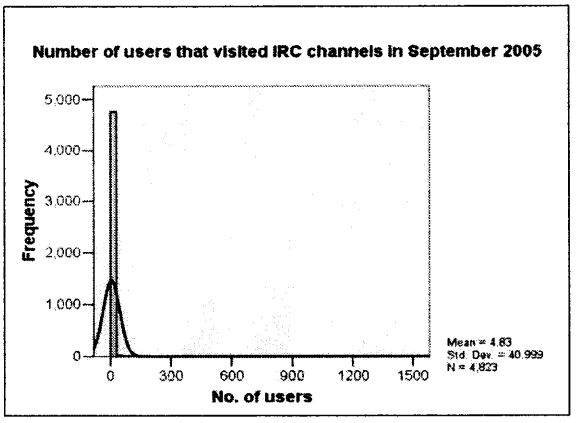

Figure 7.19 h) Histogram of users per active channel in 09/2005.

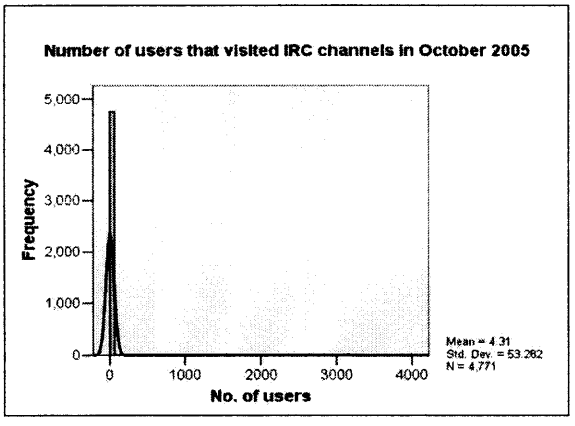

Figure 7.19 i) Histogram of users per active channel in 10/2005.

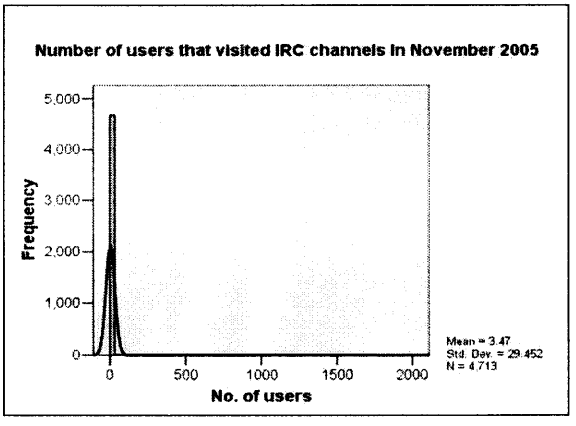

Figure 7.19 j) Histogram of users per active channel in 11/2005.

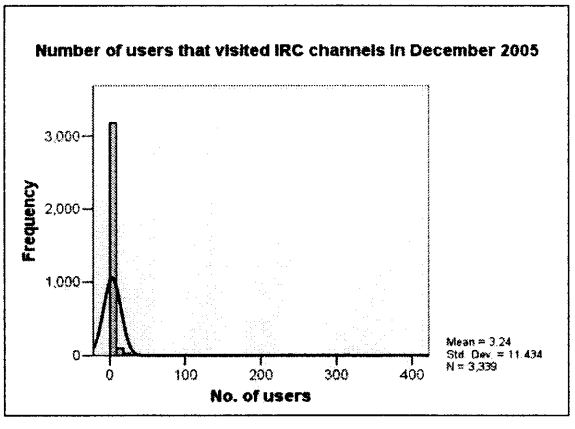

Figure 7.19 k) Histogram of users per active channel in 12/2005.

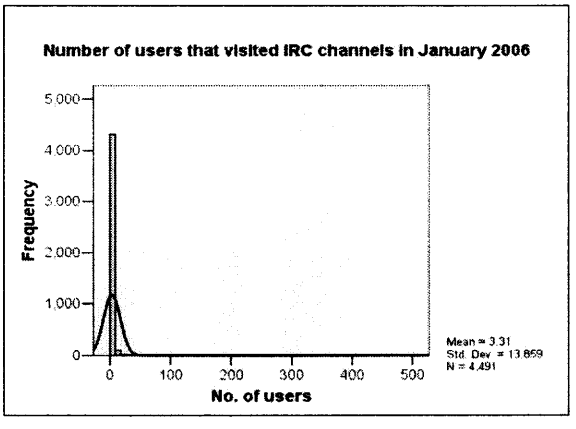

Figure 7.19 l) Histogram of users per active channel in 01/2006.

Figures 7.19 (a - l) display the monthly distributions of the number of users who visited active channels during the study period. In all the cases the distribution is highly skewed, showing that most of the channels were visited by low numbers of users during any given month. Table 7.13 displays the most important descriptive statistics of this variable. Figure 7.20 plots the values of the six percentile categories found in Table 7.13. It is apparent that over the year 50 percent of the channels were visited by only one user, and 90 percent of the channels were visited by five or fewer users.

Table 7.13 Descriptive Statistics – Number of Users Who Visited Active Channels

	Mean	Median	Mode	StDev	Min	Max	Range	Percentiles						Total channels
								25%	50%	75%	90%	95%	99%	
Feb	3.44	1	1	14.383	1	412	411	1	1	3	6	10	29	7,886
Mar	3.16	1	1	13.609	1	828	827	1	1	2	5	9	25	8,870
Apr	2.78	1	1	7.363	1	243	242	1	1	2	5	9	26	7,221
May	2.61	1	1	5.134	1	118	117	1	1	2	5	9	20	1,547
Jun	3.27	1	1	30.896	1	2113	2112	1	1	2	5	8	24	5,430
Jul	4.82	1	1	35.278	1	968	967	1	1	3	6	9	29	5,053
Aug	3.26	1	1	19.323	1	812	811	1	1	2	5	9	25	4,595
Sep	4.83	1	1	40.999	1	1471	1470	1	1	3	5	9	39	4,823
Oct	4.31	1	1	53.282	1	3269	3268	1	1	2	5	8	24	4,771
Nov	3.47	1	1	29.452	1	1714	1713	1	1	2	5	8	25	4,713
Dec	3.24	1	1	11.434	1	366	365	1	1	2	6	9	27	3,339
Jan	3.31	1	1	13.859	1	416	415	1	1	2	5	8	39	4,491
Total	3.57	1	1	24.995	1	1522	1521	1	1	2	5	9	30	27,987

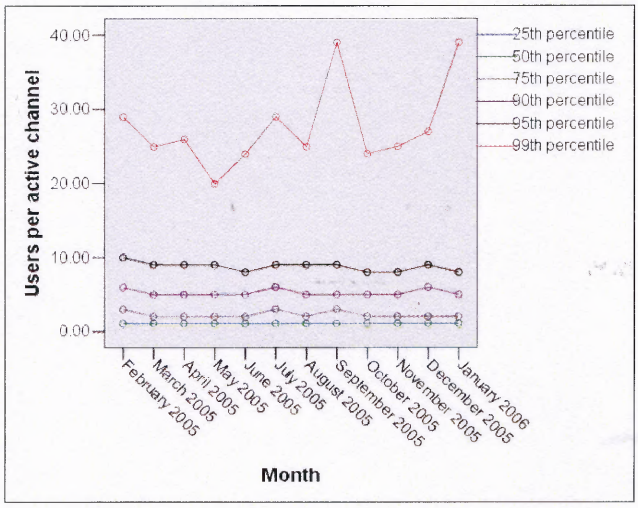

Figure 7.20 Percentile categories for the number of users per active channel.

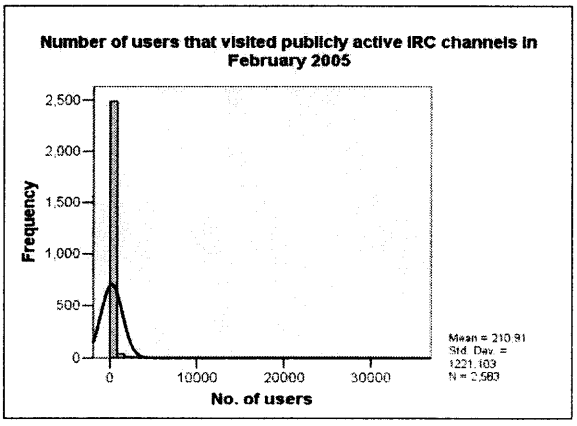

Figure 7.21 a) Histogram of users per publicly active channel in 02/2005.

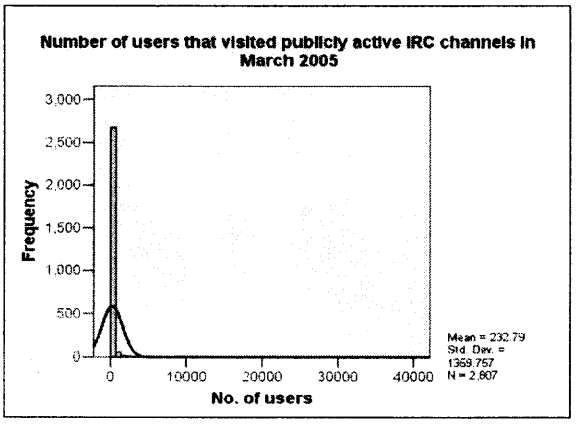

Figure 7.21 b) Histogram of users per publicly active channel in 03/2005.

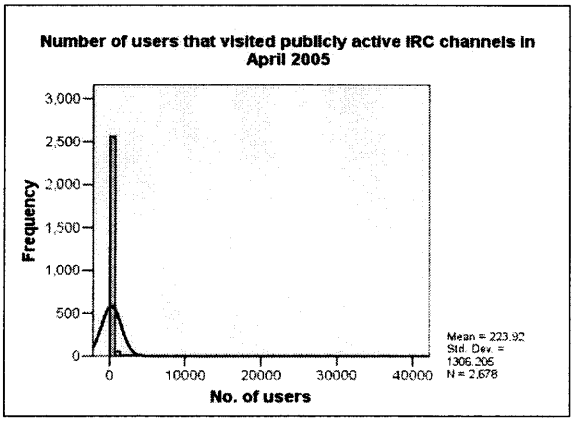

Figure 7.21 c) Histogram of users per publicly active channel in 04/2005.

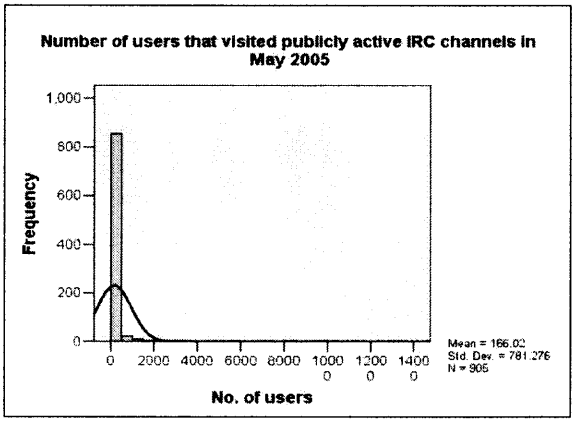

Figure 7.21 d) Histogram of users per publicly active channel in 05/2005.

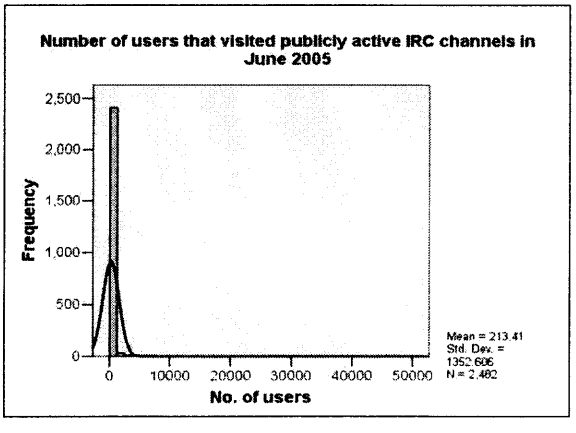

Figure 7.21 e) Histogram of users per publicly active channel 06/2005.

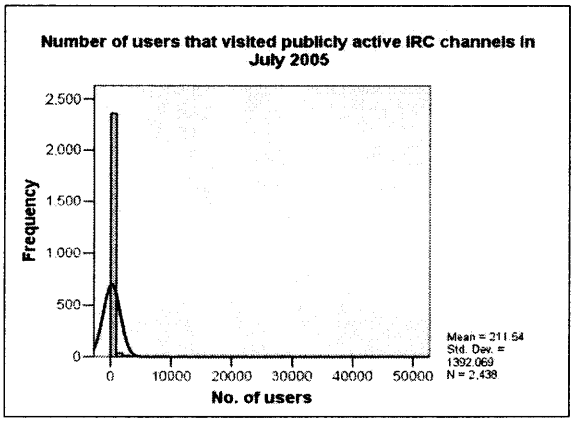

Figure 7.21 f) Histogram of users per publicly active channel in 07/2005.

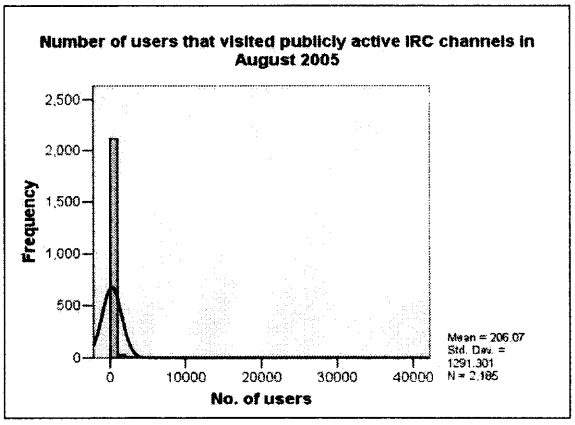

Figure 7.21 g) Histogram of users per publicly active channel in 08/2005.

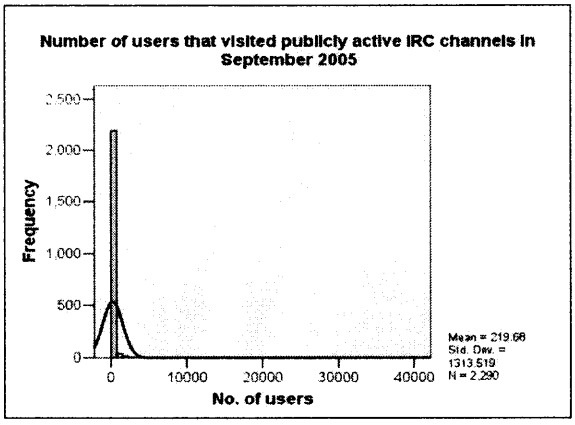

Figure 7.21 h) Histogram of users per publicly active channel in 09/2005.

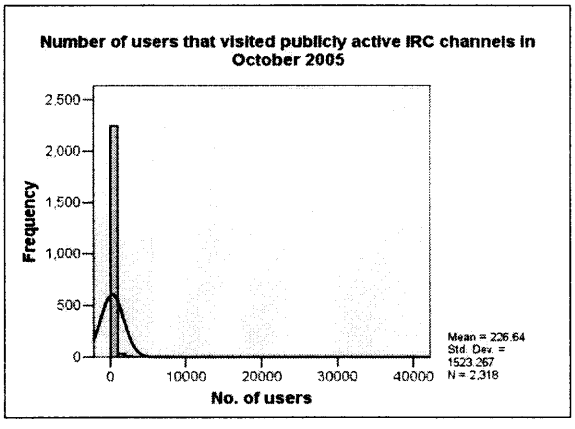

Figure 7.21 i) Histogram of users per publicly active channel in 10/2005.

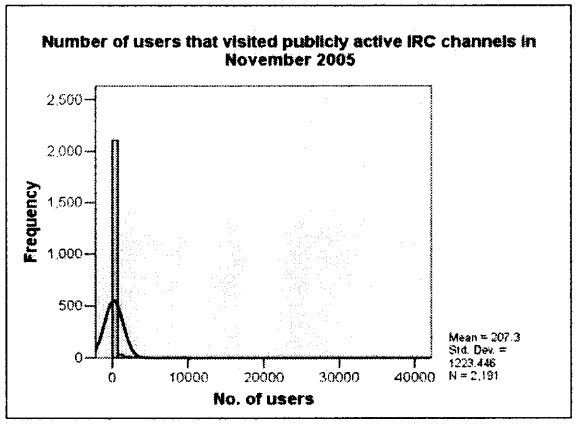

Figure 7.21 j) Histogram of users per publicly active channel in 11/2005.

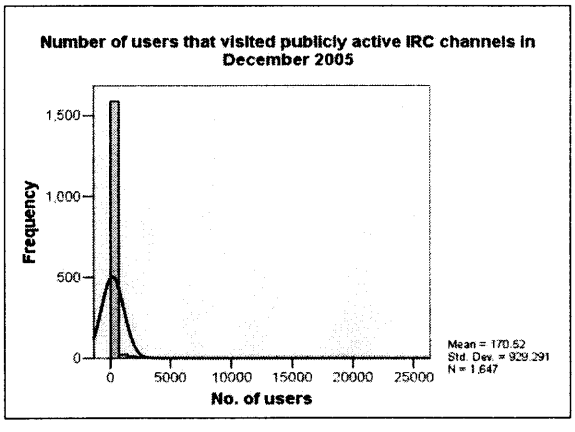

Figure 7.21 k) Histogram of users per publicly active channel in 12/2005.

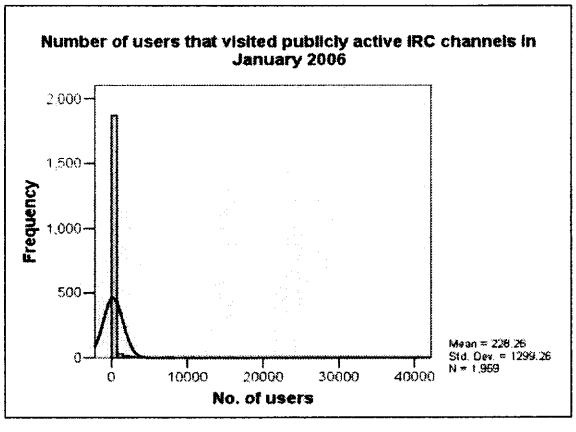

Figure 7.21 l) Histogram of users per publicly active channel in 01/2006.

Figures 7.21 (a - l) display the monthly distributions of the number of users per publicly active channels during the study period. In all the cases the distribution is highly skewed, showing that most of the channels were visited by low numbers of users during any given month. Table 7.14 displays the most important descriptive statistics of this variable. Figure 7.22 plots the values of the six percentile categories found in Table 7.14. Among other things, it shows that over the year 50 percent of the channels were visited by at most 29 users, and 90 percent of the channels were visited by 450 or fewer users.

Table 7.14 Descriptive Statistics – No. of Users Who Visited Publicly Active Channels

							Percentiles						
	Mean	Median	Mode	StDev	Min	Max	25%	50%	75%	90%	95%	99%	Total chans
Feb	210.91	22	2	1221.103	1	31,207	9	22	73	234	565	4,490	2,583
Mar	232.79	23	4	1359.757	1	34,775	9	23	73	264	628	4,802	2,807
Apr	223.92	22	2	1306.225	1	34,648	8	22	75	271	629	4,405	2,678
May	166.02	21	4	781.276	1	13,882	8	21	63	213	614	3,232	905
Jun	213.41	23	3	1352.606	1	47,201	9	23	72	250	611	4,459	2,482
Jul	211.54	23	3	1392.069	1	44,169	9	23	69	233	566	4,000	2,438
Aug	206.07	22	4	1291.301	1	35,866	8	22	62	224	519	3,964	2,185
Sep	219.68	23	4	1313.519	1	34,872	9	23	70	238	602	4,520	2,290
Oct	226.64	23	3	1523.267	1	39,040	9	23	69	235	508	4,329	2,318
Nov	207.30	24	2	1223.446	1	34,855	9	24	71	245	562	4,450	2,191
Dec	170.52	21	6	929.291	1	22,840	9	21	58	207	491	4,011	1,647
Jan	228.26	23	6	1299.260	1	33,591	8	23	68	241	657	5,228	1,959
Total	529.06	29	2	5268.428	1	269,696	9	29	110	450	1,098	8,275	7,180

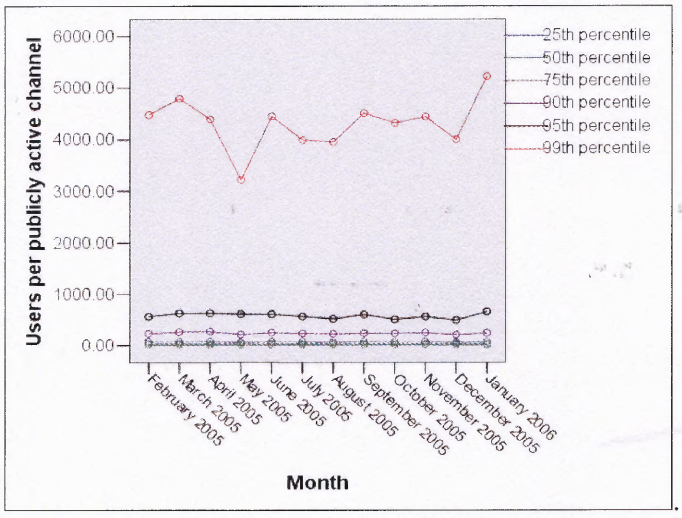

Figure 7.22 Percentile categories for the number of users per publicly active channel.

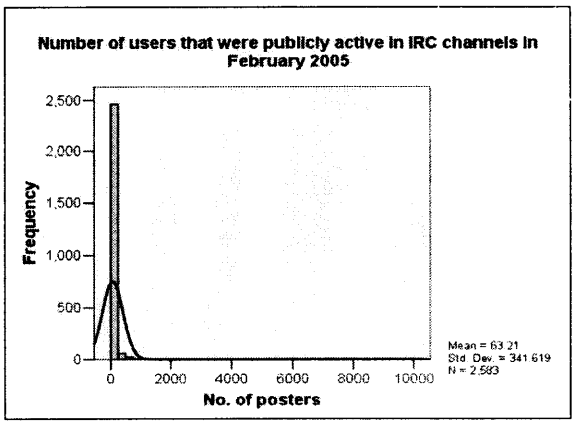

Figure 7.23 a) Histogram of posters per publicly active channel in 02/2005.

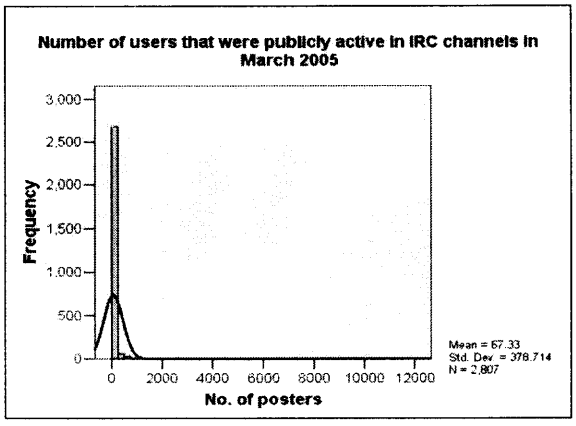

Figure 7.23 b) Histogram of posters per publicly active channel in 03/2005.

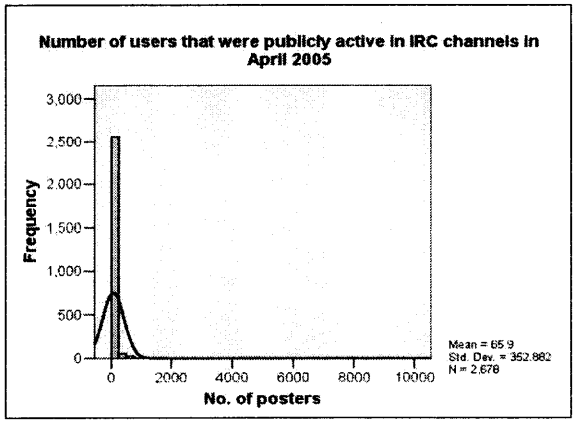

Figure 7.23 c) Histogram of posters per publicly active channel in 04/2005.

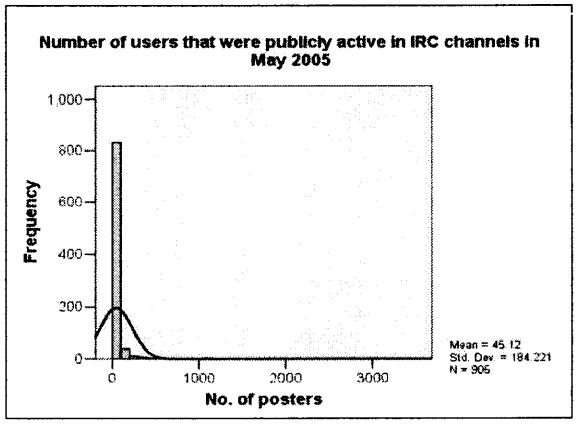

Figure 7.23 d) Histogram of posters per publicly active channel in 05/2005.

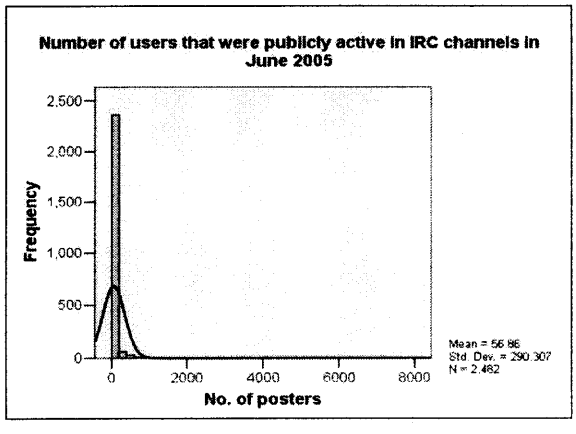

Figure 7.23 e) Histogram of posters per per publicly active channel 06/2005.

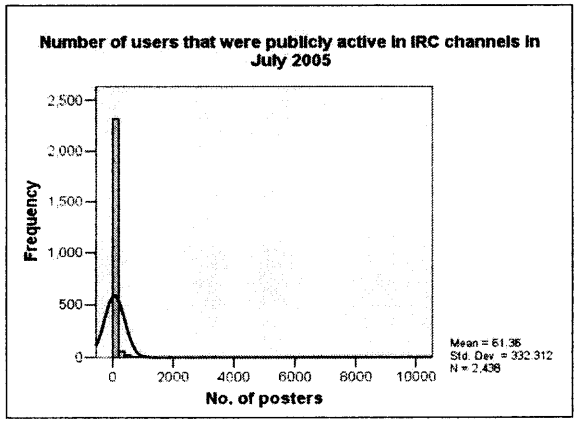

Figure 7.23 f) Histogram of posters per publicly active channel in 07/2005.

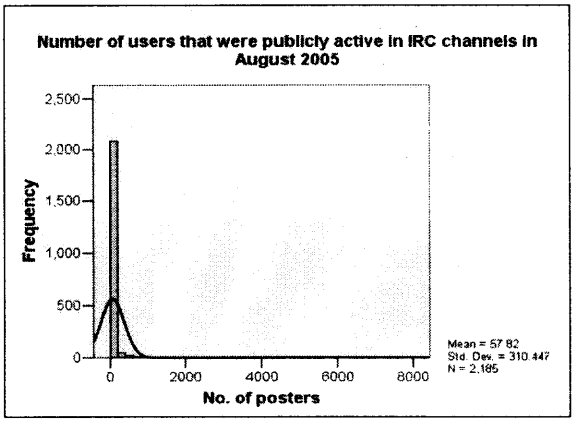

Figure 7.23 g) Histogram of posters per publicly active channel in 08/2005.

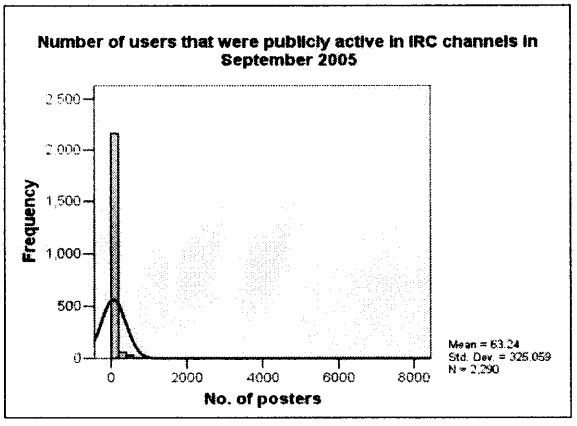

Figure 7.23 h) Histogram of posters per publicly active channel in 09/2005.

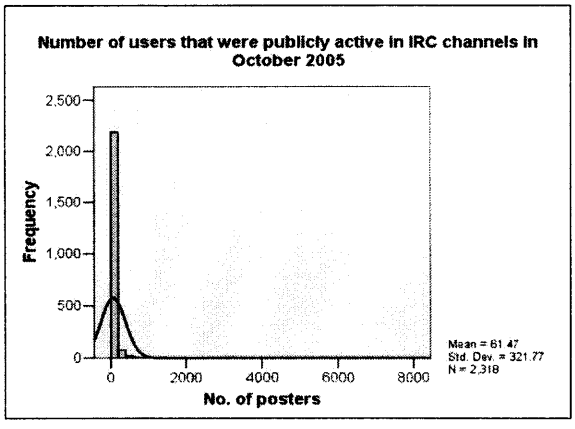

Figure 7.23 i) Histogram of posters per publicly active channel in 10/2005.

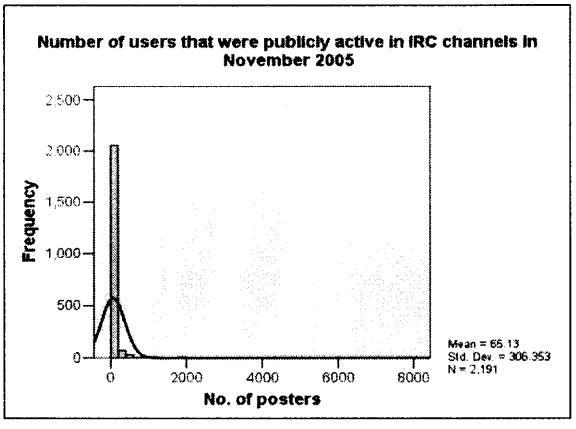

Figure 7.23 j) Histogram of posters per publicly active channel in 11/2005.

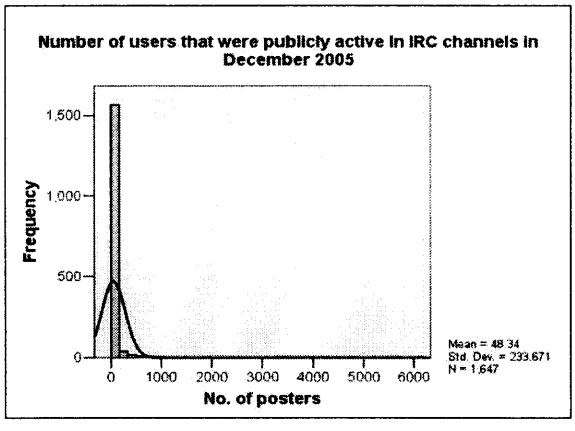

Figure 7.23 k) Histogram of posters per publicly active channel in 12/2005.

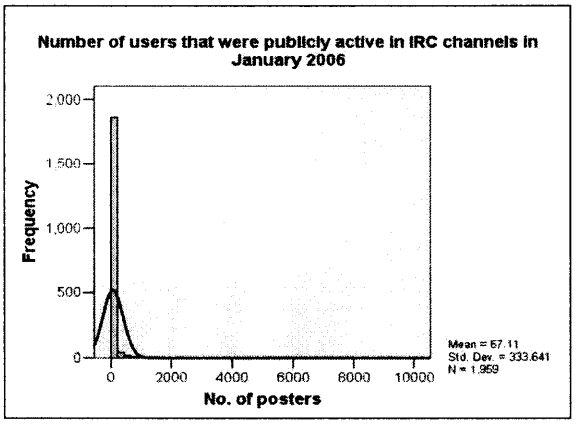

Figure 7.23 l) Histogram of posters per publicly active channel in 01/2006.

Figures 7.23 (a - l) display the monthly distributions of the number of posters per publicly active channels during the study period. The distributions are highly skewed, showing that most of the channels were visited by low numbers of posters during any given month. Table 7.15 displays the most important descriptive statistics of this variable. Figure 7.24 plots the values of the six percentile categories found in Table 7.15. Among other things, it may be noted that over the year 50 percent of the channels had at most eight posters and, 90 percent of the channels had at most 171 posters.

Table 7.15 Descriptive Statistics – Number of Posters per Publicly Active Channel

	Mean	Median	Mode	StDev	Min	Max	Range	Percentiles						Total chans
								25%	50%	75%	90%	95%	99%	
Feb	63.21	7	1	341.619	1	9,151	9,150	2	7	28	101	222	998	2,583
Mar	67.33	7	1	378.714	1	10,381	10,380	2	7	30	105	227	1,018	2,807
Apr	65.90	8	1	352.882	1	9,288	9,287	2	8	31	114	218	946	2,678
May	45.12	8	1	184.221	1	3,001	3,000	3	8	26	79	167	762	905
Jun	56.86	8	1	290.307	1	7,072	7,071	2	8	29	95	195	827	2,482
Jul	61.36	8	1	332.312	1	8,349	8,348	2	8	29	98	190	1,013	2,438
Aug	57.82	8	1	310.447	1	7,472	7,471	2	8	27	95	186	857	2,185
Sep	63.24	8	1	325.059	1	7,630	7,629	2	8	30	101	225	882	2,290
Oct	61.47	7	1	321.770	1	7,640	7,639	2	7	29	96	228	988	2,318
Nov	65.13	8	1	306.353	1	7,736	7,735	2	8	31	110	254	1,071	2,191
Dec	48.34	7	1	233.671	1	5,579	5,578	2	7	23	78	166	802	1,647
Jan	67.11	7	1	333.641	1	8,206	8,205	2	7	30	102	222	1,073	1,959
Total	161.98	8	1	1476.191	1	64,670	64,669	2	8	38	171	434	2,623	7,180

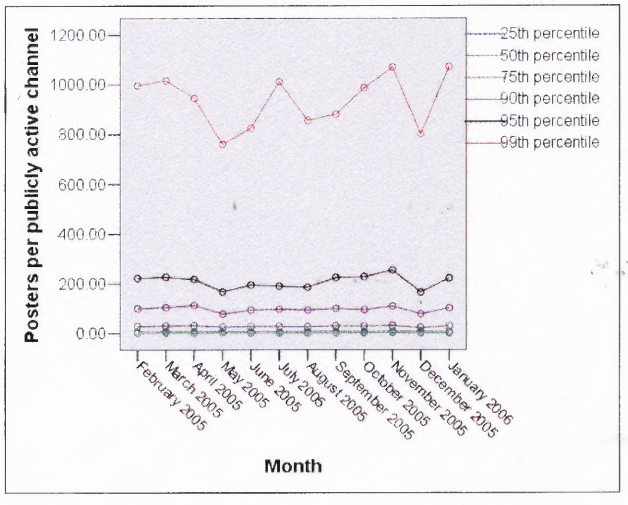

Figure 7.24 Percentile categories for the number of posters per publicly active channel.

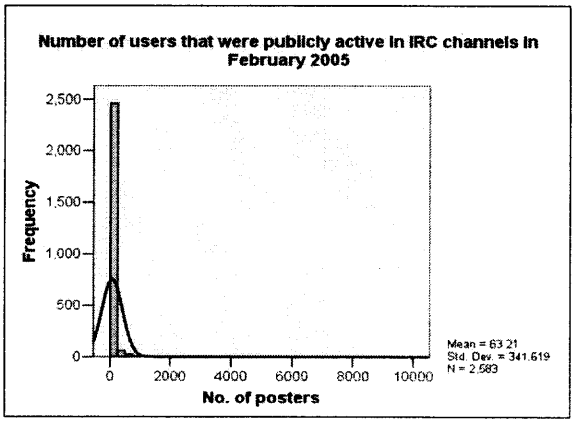

Figure 7.25 a) Histogram of messages per publicly active channel in 02/2005.

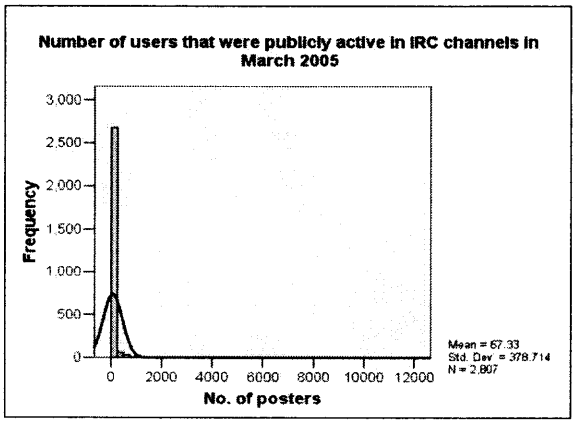

Figure 7.25 b) Histogram of messages per publicly active channel in 03/2005.

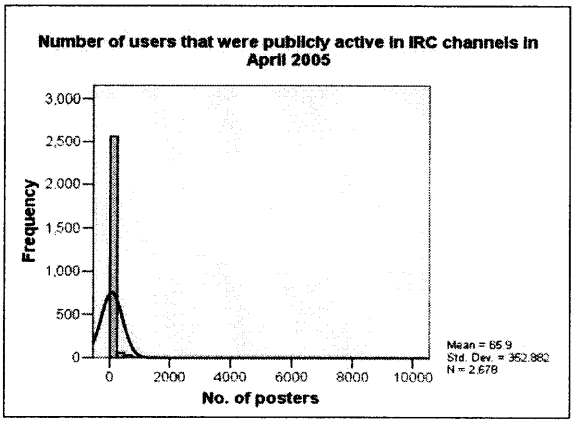

Figure 7.25 c) Histogram of messages per publicly active channel in 04/2005.

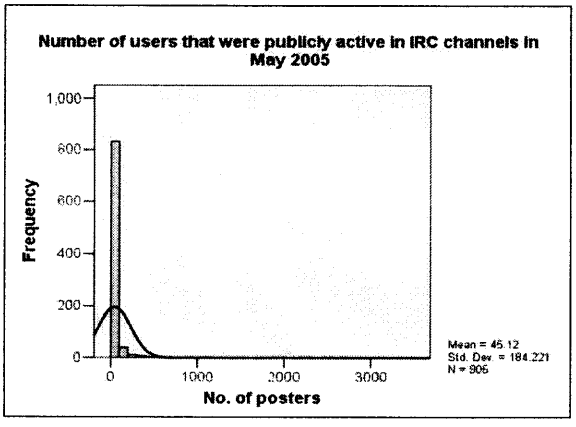

Figure 7.25 d) Histogram of messages per publicly active channel in 05/2005.

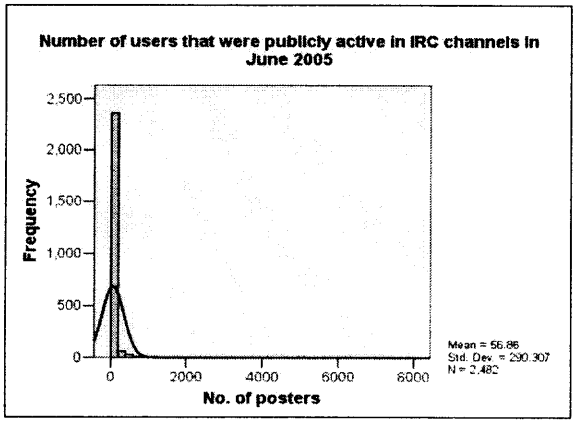

Figure 7.25 e) Histogram of messages per publicly active channel 06/2005.

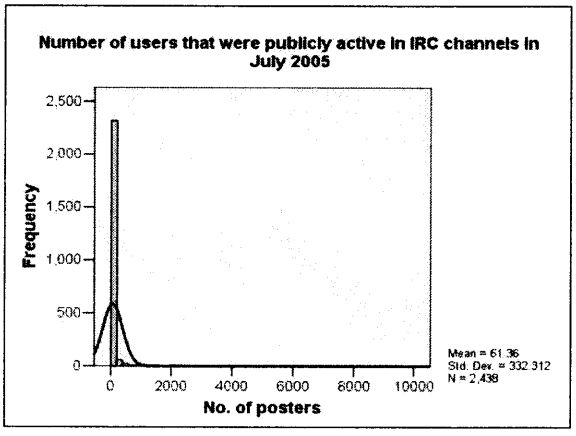

Figure 7.25 f) Histogram of messages per publicly active channel in 07/2005.

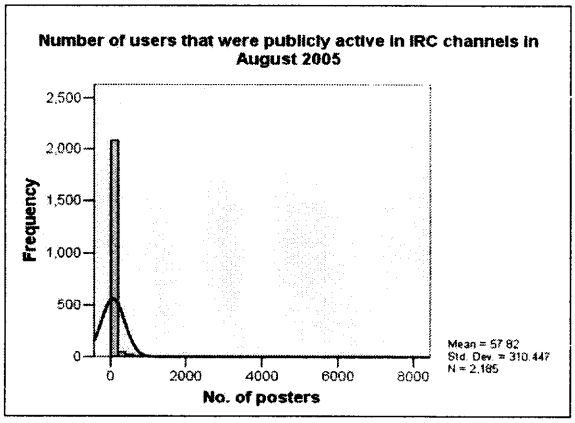

Figure 7.25 g) Histogram of messages per publicly active channel in 08/2005.

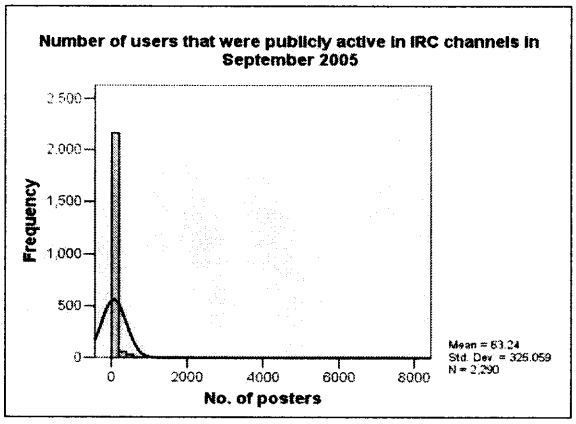

Figure 7.25 h) Histogram of messages per publicly active channel in 09/2005.

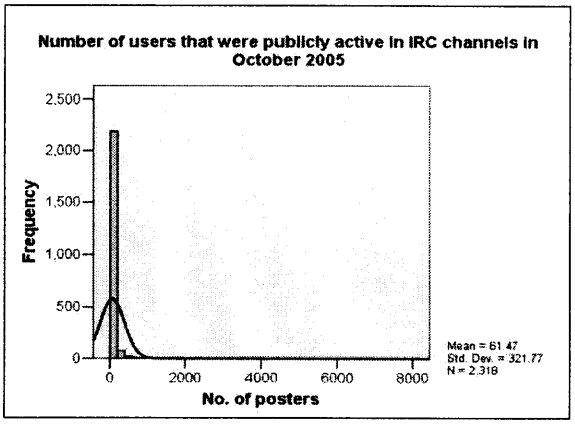

Figure 7.25 i) Histogram of messages per publicly active channel in 10/2005.

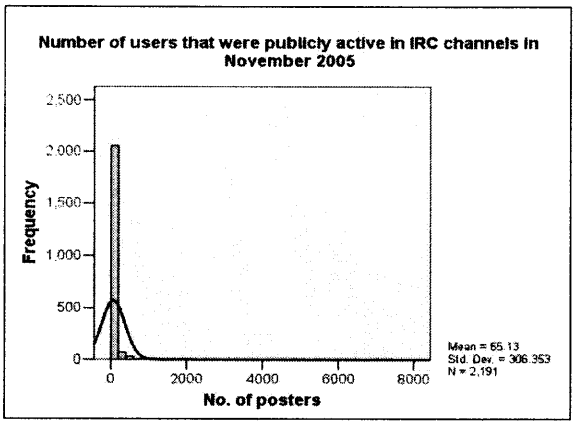

Figure 7.25 j) Histogram of messages per publicly active channel in 11/2005.

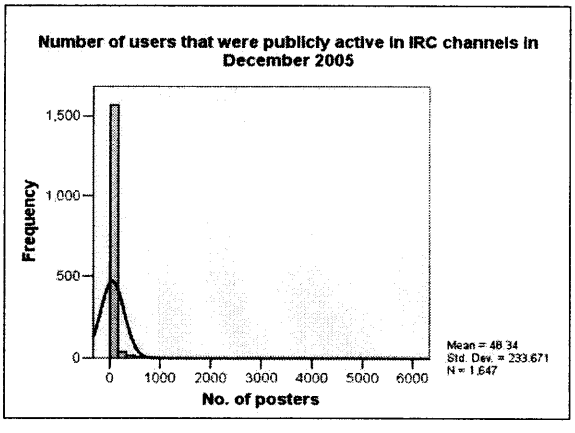

Figure 7.25 k) Histogram of messages per publicly active channel in 12/2005.

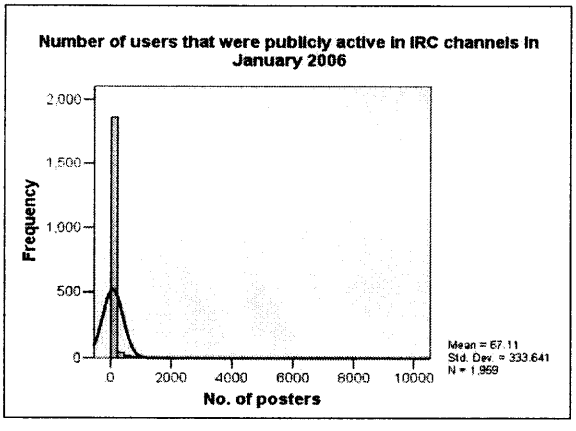

Figure 7.25 l) Histogram of messages per publicly active channel in 01/2006.

Figures 7.25 (a - l) display the monthly distributions of the number of messages per publicly active channels during the study period. The distributions are highly skewed, showing that most of the channels had low amounts of public interaction during any given month. Table 7.16 displays the most important descriptive statistics of this variable. Figure 7.26 plots the values of the six percentile categories found in Table 7.16. Among other things, it may be noted that over the year 50 percent of the channels had 160 messages at the most, and 90 percent of the channels had 11,491 messages at the most.

Table 7.16 Descriptive Statistics – Number of Messages per Publicly Active Channel

						Percentiles						
	Mean	Median	Mode	Min	Max	25%	50%	75%	90%	95%	99%	Chans
Feb	3601.64	134	1	1	615,738	15	134	962	5,066	13,187	66,052	2,583
Mar	3950.78	133	1	1	696,968	14	133	976	5,569	13,701	68,076	2,807
Apr	3852.29	137	1	1	515,197	14	137	1,006	6,241	15,249	69,172	2,678
May	1659.27	111	1	1	107,165	17	111	696	2,933	6,453	34,106	905
Jun	3537.06	152	1	1	383,705	15	152	1,194	6,198	14,707	69,717	2,482
Jul	3974.15	145	1	1	452,846	13	145	1,217	6,486	15,859	74,776	2,438
Aug	3595.13	140	1	1	427,404	16	140	1,033	5,912	14,140	74,427	2,185
Sep	4361.50	151	1	1	531,192	14	151	1,199	6,924	19,065	86,492	2,290
Oct	4399.23	147	1	1	545,708	14	147	1,270	6,837	17,209	91,945	2,318
Nov	4961.15	176	1	1	504,191	15	176	1,437	8,965	18,978	124,247	2,191
Dec	2776.66	109	1	1	286,088	11	109	793	4,157	10,232	64,010	1,647
Jan	4745.73	163	1	1	479,567	13	163	1,234	7,585	17,600	96,409	1,959
Total	14409.39	160	1	1	5,545,769	17	160	1,495	11,491	36,619	273,948	7,180

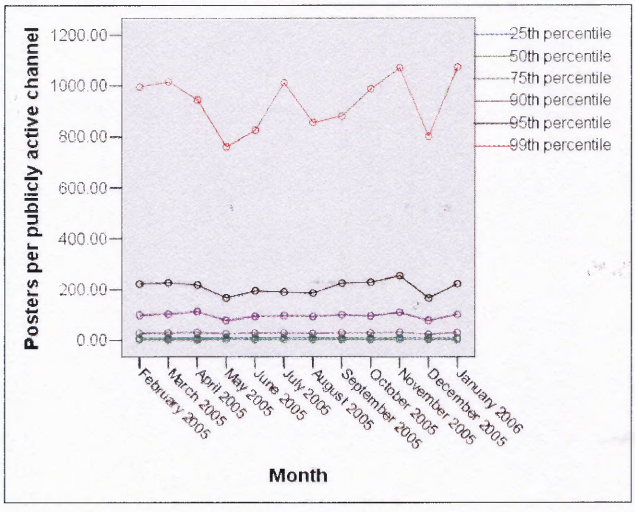

Figure 7.26 Percentiles for the number of messages per publicly active channel.

Table 7.16 as well as Figures 7.25 and 7.26 shows that over the study period a small number of publicly active IRC channels hosted most of the public group interaction, in the same way as a small number of posters were responsible for most of the mass interaction. Table 7.17 summarizes the number of messages sent to the most publicly active channels; there are expressed as percentages of the total number of monthly or yearly messages.

Table 7.17 Top Channels and Messages Percentages

	Top 10% channels	Top 5% channels	Top 3% channels	Top 1% channels
February	87.37	76.07	66.38	45.95
March	87.53	76.56	67.42	46.37
April	86.10	73.50	63.36	41.52
May	80.39	67.56	57.82	36.68
June	83.92	70.54	59.83	38.69
July	85.46	73.56	63.14	40.70
August	85.81	73.30	63.21	40.67
September	85.92	73.24	62.40	40.81
October	86.43	74.53	64.74	42.92
November	85.33	72.58	62.73	41.60
December	86.12	75.11	64.98	39.01
January	86.39	74.48	63.56	38.85
Monthly average	85.56	73.41	63.29	41.14
Yearly	94.17	86.68	79.22	60.77

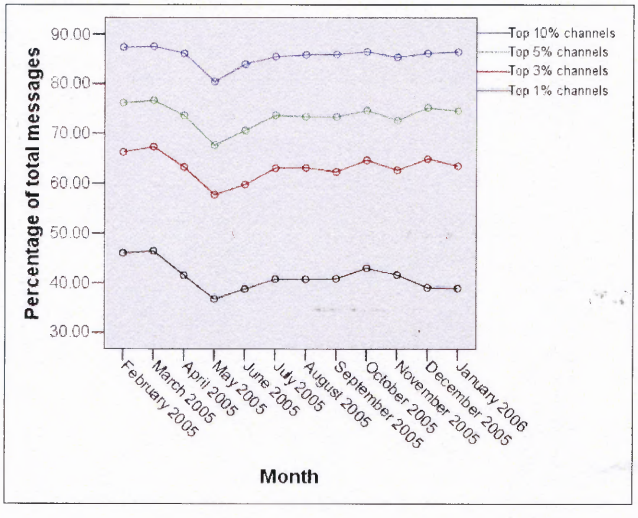

Figure 7.27 Percentage of total messages sent to the most publicly active channels.

Figure 7.27 plots the values of the four most active publicly active channels categories found in Table 7.17. The values remained relatively constant throughout the year, if the outliers for the months of May 2005 and December 2005 are not considered. It may be noticed that during any given month, the top 10 percent of the publicly active channels hosted approximately 85 percent of the total number of monthly public messages; the top 5 percent of the publicly active channels hosted approximately 73 percent of the total number of monthly public messages; the top 3 percent of the publicly active channels hosted approximately 63 percent of the total number of monthly public messages; and the top 1 percent of the publicly active channels hosted approximately 41 percent of the total number of monthly public messages. Throughout the entire year, the percentage of messages per top most publicly active channel categories increases significantly: the top 10 percent of the publicly active channels hosted approximately 94 percent of the total number of public messages; the top 5 percent of the publicly active channels hosted approximately 86 percent of the total number of public messages; the top 3 percent of the publicly active channels hosted approximately 79 percent of the total number of public messages; and the top 1percent of the publicly active channels hosted approximately 60 percent of the total number of public messages.

Figures 7.28 (a - c) describe the relationships between various channel-related variables such as the number of messages, the number of users, and the number of posters over the entire study period. Common sense suggests that an increase in the number of users of an IRC channel would typically be correlated with increases in both the number of posters and the number of messages. The Spearman correlation coefficients computed for each pair of variables confirm these assumptions. The highest observed correlation

was between the number of messages and the number of posters in an IRC channel. In this case, Spearman's rho was 0.907. Another obvious high correlation was observed between the number of users and the number of posters of IRC channels: Spearman's rho = 0.861. The lowest correlation was observed between the number of messages and the number of users: Spearman's rho = 0.727. For all three cases, n=7,180, $p < .001$.

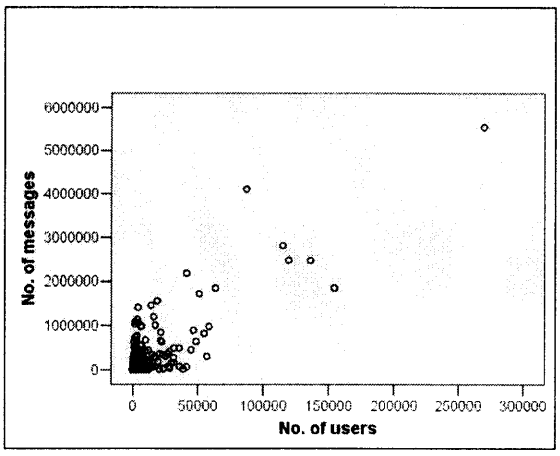

Figure 7.28 a) Messages by users per publicly active channel.

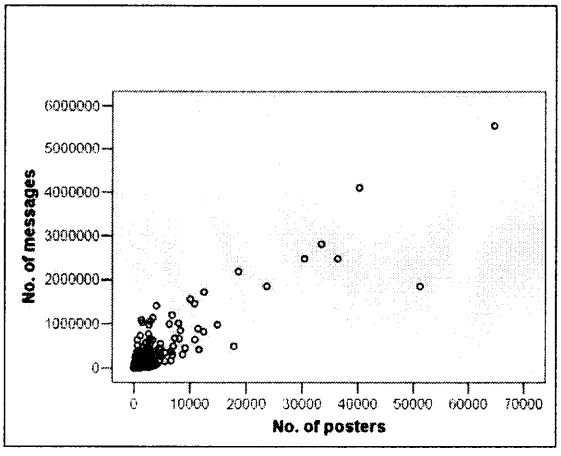

Figure 7.28 b) Messages by posters per publicly active channel.

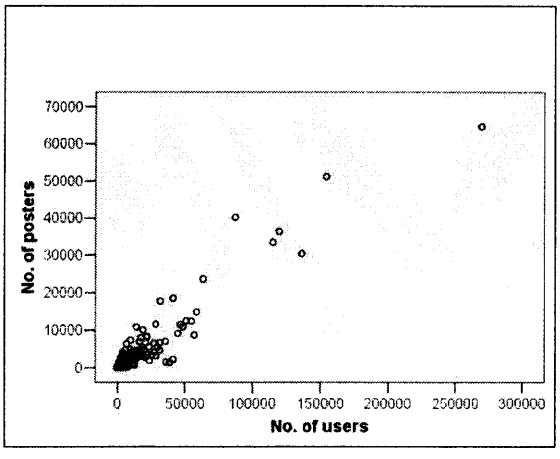

Figure 7.28 c) Users by posters per publicly active channel.

7.2.3.2 Dynamics of a Sample of Channels. The subsection describes the dynamics of publicly active IRC chat-channels. Several measures were computed; these are presented in Table 7.18 along with some of their descriptive statistics. The sample of channels was drawn from August 2005, for the same reason as in the case of the sample of posters – it was in the middle of the data-collection period. All of the variables described in Table 7.18 were computed for the entire month of August 2005. The smallest time interval for which it was possible to collect and/or compute the values of measures like the number of users, the number of posters, the number of messages, the user diversity, or the poster diversity in a reliable fashion was one third of an hour or 20 minutes. Two types of averages were computed: the average value for any 20-minute time interval during August, and the average value for any day of the month (daily average value).

User diversity and poster diversity measured the heterogeneity of the channels' user and poster populations. Specifically, the user or poster diversity for a particular time interval represented the percentage of users or posters that were present or active in a channel during that interval with respect to the larger time period. For example, if during a particular hour a channel had five unique posters present, and during that day the channel had 15 unique posters, the "diversity" for that hour was 5 / 15 = 0.33. Two types of diversity were computed: the average diversity for any hour, with respect to that particular day; and the average diversity for any day (daily diversity) with respect to that particular week.

Table 7.18 Channel Dynamics Data

Measure	Mean	Median	Mode	Percentiles						Sample size
				25%	50%	75%	90%	95%	99%	
No. of days in existence	8.58	2.00	1.00	1	2	13	31	31	31	2,186
Average no. of users per channel	3.02	1.00	1.00	1	1	2	3	7	27	2,1862
Average no. of daily users per channel	4.39	1.00	1.00	1	1	2	5	10	42	2,186
Average no. of posters per channel	6.05	3.00	1.00	2	3	5	11	18	57	2,186
Average no. of daily posters per channel	7.36	3.00	1.00	2	3	6	14	21	71	2,186
Average no. of messages per channel	72.47	18.00	1.00	5	18	48	136	270	1,054	2,186
Average no. of daily messages per channel	170.52	36.00	1.00	8	36	123	352	595	2,477	2,186
Average channel user diversity	0.8113	0.9003	1.00	0.4375	0.7033	1	1	1	1	2,186
Average daily channel user diversity	0.5302	0.4286	1.00	0.2143	0.4286	1	1	1	1	2,186
Average channel poster diversity	0.8547	0.9375	1.00	0.7500	0.9375	1	1	1	1	2,186
Average daily channel poster diversity	0.4962	0.3750	1.00	0.2000	0.3750	1	1	1	1	2,186
Average user return time (in minutes)	3063.96	2816.00	4135.0	1,989	2,816	3,817	5,020	6,963	8,473	113

7.3 Summary of Descriptive Statistics

The descriptive statistics presented in this chapter showed the following:

- Over the course of one year, only 57 percent of the total number of channels were visited by users;
- Over the course of one year, 20 percent of the channels that were visited by users hosted public interactions;
- Approximately half of the channels that were active during any given month were likely to be active during the following month;
- Approximately 75 percent of the channels that were publicly active during any given month were likely to be publicly active during the following month;
- Approximately 30 percent of the users who visited the network during any given month were likely to visit the network during the following month;
- Approximately 30 percent of the posters who were publicly active during any given month were likely to be publicly active during the following month;
- Most of the users (both lurkers and posters) visited the IRC network for short periods of time;
- Most of the users (both lurkers and posters) visited a small number of channels (both over the entire year and during any given session);
- Overall, posters visited more channels than lurkers;
- Most of the posters were publicly active in a small number of channels (both over the entire year and during any given session);
- A small proportion of posters generated the vast majority of messages: over the year, 10 percent of the posters were responsible for 91 percent of the messages;
- Publicly active channels were visited significantly more often than active channels in terms of both size of the population (number of users) and length of time (number of months);
- Approximately half of the publicly active channels were visited for at least five months while approximately 25 percent of the publicly active channels hosted public interactions for at least 5 months;

- A small number of channels were host to the vast majority of public interactions: over the year, 94 percent of the public messages were sent to 10 percent of the publicly active channels.

In sum, the descriptive statistics presented in this chapter demonstrated that the analyzed IRC network was a dynamic, constantly changing environment with a great deal of constant turnover in users and channels. Thus, it was interesting to see if the apparently "chaotic" behavior could in fact be better understood and even predicted. The next chapters address this issue by describing various models that focus on predicting the activity of IRC chat-channels.

CHAPTER 8

EFFECTS OF INFORMATION PROCESSING CONSTRAINTS ON THE BOUNDARIES OF CHANNEL ACTIVITY

While IRC does not limit the number of users, posters, or messages in a channel; little is known about the boundaries imposed on these variables by the users' information processing capabilities. This chapter attempts to identify empirically the upper information-processing limits that constrain the community interaction dynamics seen in chat-channels. The information-processing constraints theory (Jones 1997; Jones and Rafaeli 1999) argues that one of the main influences on a user's participation in computer mediated communication is the level of information overload to which the user is exposed when using the system. Prior research on asynchronous CMC systems has shown that the level of activity within such a system can only rise up to a certain level. After this level is reached, due to the effects of information overload, the activity either remains constant or decreases.

This chapter identifies the maximum level of activity measured in users, posters, and messages that can be reached inside a synchronous CMC system such as IRC.

8.1 Hypotheses

The following hypotheses were formulated using the information-processing constraints theory (Jones 1997; Jones and Rafaeli 1999):

- Message density, defined as the number of messages per poster in an IRC channel, will vary with the user population up to a limited user pool. Beyond that point, the message density will remain constant.

- The cap on message density will constrain the number of posters co-present in an IRC channel.

8.2 Method

8.2.1 Data Collection

The descriptive statistics of the IRC network presented in the previous chapter revealed the existence of a large number of publicly active channels during each month over the one-year data-collection period. The analysis of all these channels would have been virtually impossible due to both time constraints and processing power constraints. Therefore, it was necessary to select a more manageable dataset to use for the identification of channel activity boundaries. This dataset was selected through a stratified random sampling of all the channels that were publicly active during August 2005. The month of August was chosen because it was in the middle of the data-collection period. From an experienced IRC observer perspective, it made sense to stratify the random sample of channels based on their size (the number of users that visited them) and their intensity (how often the channels were visited during a particular time interval). While there are other possible approaches for the stratification of channels, the lack of previous research in this area made it necessary to select these two particular variables as a starting point.

During the selected month, there were a total of 2,186 channels with public user postings. The channels that were not visited by at least a minimal group of posters (this minimal group was defined as at least three posters) were eliminated, leaving 1,124 channels that had three or more posters. The stratified random sample was selected from this set of 1,124 channels. Two variables were computed for each channel in this set: the total number of users that visited the channel during the interval February 1 to August 31, 2005, and the total number of days the channel was visited by users during the same interval. Afterwards, a frequencies analysis was performed, which led to the formation of subgroups of small, medium, and large channels characterized by low, medium, or high intensity. Nine categories of channels were obtained: small channels with low intensity; small channels with medium intensity; small channels with high intensity; medium channels with low intensity; medium channels with medium intensity; medium channels with high intensity; large channels with low intensity; large channels with medium intensity; and large channels with high intensity. The frequencies analysis resulted in the following delimiters for the nine subgroups of channels: small channels had between 4 and 98 users; medium channels had between 99 and 444 users; large channels had between 445 and 175,141 users; low intensity channels were visited for less than 45 days between February 1 and August 31 (less than a month and a half); medium intensity channels were visited between 45 and 90 days (between a month and a half and three months); and high intensity channels were visited more than 90 visited days between February 1 and August 31 (more than three months). Ten channels were randomly selected from each of the nine categories, resulting in a stratified random sample of 90

channels. Table 8.1 shows the mean values of various descriptive statistics for all the channels in these subgroups.

Table 8.1 Descriptive Statistics of Sample Channels for the Month of August 2005

Size	Small			Medium			Large		
Intensity	**Low**	**Med.**	**High**	**Low**	**Med.**	**High**	**Low**	**Med.**	**High**
Total users	30.3	29.2	22.8	153.4	110.2	63.6	541.2	330.8	3922.0
Total posters	13.0	18.0	15.1	54.0	61.0	39.1	149.5	120.1	1021.0
Total messages	623.9	2298.0	2267.4	3711.0	2472.1	2787.3	5978.2	4873.4	46871.3
Number of days visited	8	16	22	16	17	22	18	24	29
Number of active days	5	11	16	10	12	17	13	19	26
Year-to-date number of visited days	20	70	110	29	70	121	31	72	180
Avg. user return time (minutes)	2499.8	2886.5	3394.5	3219.1	3016.19	2717.3	3539.3	2378.0	3173.2
Avg. no. of users	4.3	4.6	4.1	11.0	8.5	7.1	20.3	18.2	160.2
Avg. no. of posters	4.0	5.1	4.2	7.1	7.4	6.7	14.19	11.2	56.1
Avg. no. of daily posters	6.1	5.4	3.9	11.6	9.7	7.8	22.5	12.2	71.2
Avg. no. of messages	45.2	57.4	67.7	85.8	67.6	64.1	131.9	111.2	671.6
Avg. no. of daily messages	157.3	175.0	130.9	253.1	158.8	160.6	323.8	197.7	1535.8
Avg. daily user diversity	0.2839	0.2234	0.2136	0.1581	0.1578	0.1709	0.1500	0.10322	0.0963
Avg. daily poster diversity	0.4214	0.3013	0.2514	0.2465	0.2056	0.2057	0.2520	0.1208	0.1089
Avg. daily poster stability	0.2715	0.4414	0.4985	0.3118	0.2711	0.4015	0.2108	0.2911	0.3053

8.2.2 Data Analysis

The sample of channels was analyzed using the following variables and measures:

- Observed Users (OU) - all the people logged into a channel and using it for either private or public conversations;

- Observed Users Max (OU_max) - the IRC system was sampled three times per hour (three 20-minute intervals), and the number of users in each of these intervals (as well as messages and posters defined below) was recorded. Then the data was aggregated on an hourly basis, and this new variable was created by taking the maximum number of users of the three measurements as the representative value for that hour;

- Observed Messages (OM) - all the public postings sent to the IRC channels in the sample. It should be stressed that the message data analyzed only pertained to messages sent in public to the chat-channel group interface. Messages sent privately among users were not analyzed;

- Observed Messages Max (OM_max) - a new variable obtained by taking the maximum number of messages as the representative value of the three hourly measurements;

- Observed Posters (OP) - those users who posted messages in public to the entire group of channel users;

- Observed Posters Max (OP_max) - a new variable obtained by taking the maximum number of posters as the representative value of the three hourly measurements;

- OMperOU_max - a ratio of OM_max to OU_max. This is a measure of message density since it describes the mean of the maximum number of messages per user within an hour of activity;

- OMperOP_max - a ratio of OM_max to OP_max. This is a measure of message density since it describes the mean of the maximum number of messages per poster within an hour of activity;

- OPperOU_max - a ratio of OP_max to OU_max. This measure indicates the ratio of participants who posted messages in public within an hour of activity.

The entire dataset collected over one month for the 90 channels consisted of 200,880 observations (3 observations per hour x 24 hours x 31 days x 90 channels). This data was aggregated to 66,960 observations of hourly maximum points for each variable.

The maximum was preferred over other statistics because it represented the potential of the system and because it was the recommended value for analysis of information overload according to Jones et al. (2004). The size and complexity of the data, and the interdependence within the data (repeated measures over time and within channels) prevented the use of standard inferential statistical tests such as correlation and regression. The wealth of data allowed the observation of phenomena based on exploration of the distributions of the variables and their descriptive statistics.

8.3 Results

This study used information on individual behavior to learn about constraints on community activity that result from information overload. Therefore, a general understanding of the system was required. With respect to this, a comprehensive set of system-related, user-related, and channel-related descriptive statistics of the IRC network was provided in the previous chapter. This section continues with the presentation of various indicators of information overload.

In exploring the raw user data, it was evident that the values of message density could be divided into several intervals and averaged. This would provide a clear picture of overload. Table 8.2 reports the mean of maximum message density for ranges of users and suggests that something happens to the message density when the number of users reaches about 30. Figure 8.1 provides a plot of users' hourly communication density (OMperOU_max) versus population size.

Table 8.2 Mean Message Density for Ranges of Users

Range of Users	Mean Message Density
2-9	52.99
10-19	21.67
20-29	12.63
30-39	4.51
40-49	4.77
50-59	3.72
60-100	2.69
110-160	2.91
170-220	1.30
220<	0.64

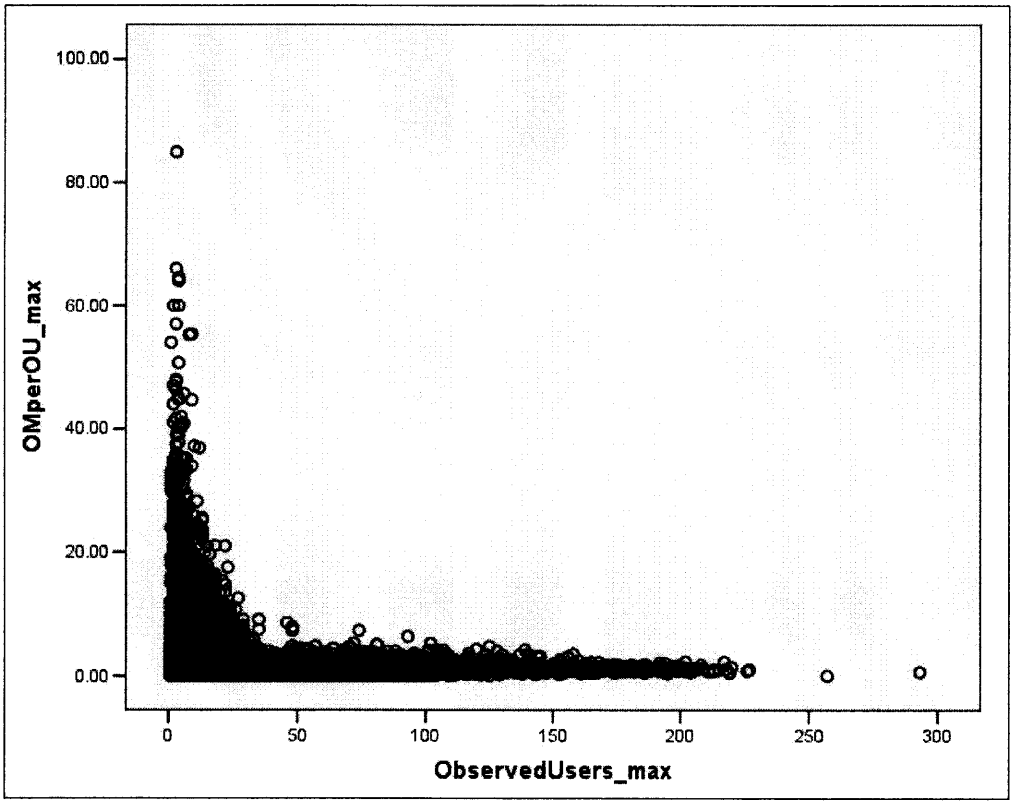

Figure 8.1 Users' public message density versus number of users.

After exploring the raw poster data, it was also evident that the values of message density could be divided into several intervals and averaged. Table 8.3 reports the mean message density for ranges of posters and shows that the limit of active public posters in

the system is 39. Figure 8.2 provides a higher resolution of a similar situation depicting the hourly message density based on the posters.

Table 8.3 Mean Message Density for Ranges of Posters

Range of Posters	Mean Message Density
2-9	65.97
10-19	29.59
20-29	19.33
30-39	11.73
39<	0

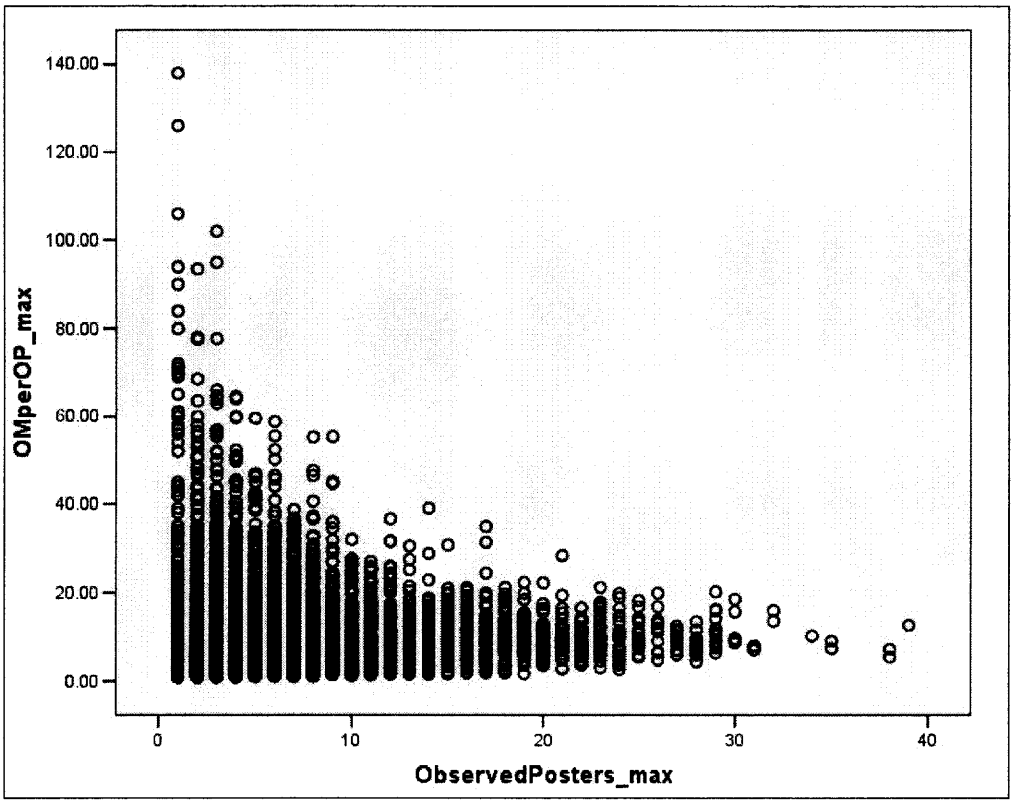

Figure 8.2 Posters' public message density versus number of posters.

Plotting the maximum values of messages versus the posters (Figure 8.3) showed that communication activity increases up to a certain point. Figure 8.3 complements Figure 8.2 in revealing that the maximum message communication load is reached when the posters range between 15 and 30. The data points in the above plots represent every hour of activity during August 2005, for all 90 channels in the sample.

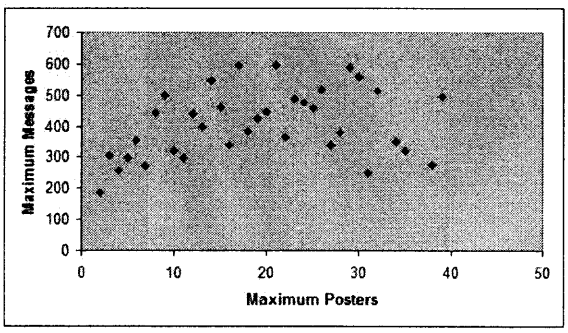

Figure 8.3 Maximum public messages versus maximum posters.

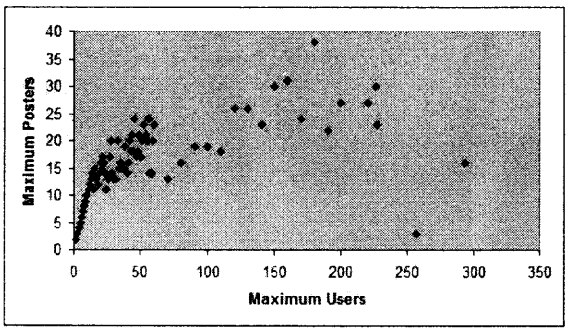

Figure 8.4 Maximum posters versus maximum users.

Finally, the chart of the maximum number of posters versus the maximum number of users is very telling. Figure 8.4 shows a linear increase in the number of posters up to a certain, fairly small, number of users. Then the rate of increase diminishes and levels off at about 30 posters for the range of 150-220 users. Beyond that number, users refrain from posting public messages.

In Figures 8.3 and 8.4 each data point represents a number of channels (different number for each point) that had the same set of data. These figures show the maximum system potential to answer the question "what is the upper limit to activity in IRC channels?" For example, when the maximum number of posters in a channel is 2, then the maximum number of messages that was observed is 187. This means there could have been channels with 2 posters and less than 187 messages, but not more than that. When the maximum number of posters in a channel was 3, then the maximum number of messages that was observed was 306.

8.4 Summary

This chapter identified the social constraints that emerge naturally in synchronous CMC. While the IRC system did not pose restrictions on the number of users, posters, or messages; constraints on all dimensions emerged as a result of information overload.

High message density was possible when the number of users and posters was small (as seen in Tables 8.3 and 8.4). As the number of participants, either users or posters, increased; the message density declined until community boundaries were reached.

Figures 8.1 and 8.4 suggest that the limit of the user community was less than 300 concurrent users in one chat-channel, while the limit of the channel poster community was less than 40 posters (Figures 8.2 and 8.3).

Figure 8.3 suggests that the upper limit to the message volume was reached even before the upper limit of posters. A maximum of about 600 public messages per chat-channel per 20-minute interval was observed. Roughly, this means that when the poster population was about 30, posters could not absorb more than 20 messages per poster within 20 minutes, or 30 messages per minute for the entire channel group.

As predicted by the information-processing model, the boundaries to the rate of posting identified for IRC public channel communication were much lower than those found in Usenet (see appendix of Jones, Ravid and Rafaeli 2004 for details).

Community size was much smaller in synchronous CMC than in asynchronous CMC. High message density was possible only at low values of observed users. Beyond a certain number of users, about 30 according to Figure 8.1 and Table 8.2, message density remained low and declined until the community stopped growing altogether.

CHAPTER 9

SHORT TERM ACTIVITY PREDICTABILITY

To date, no empirical work has investigated the extent to which short-term measures of activity can be reliably predicted for synchronous spaces such as IRC channels. This chapter addresses this shortfall through regression modeling, both linear and nonlinear, and describes the best prediction models that can be obtained for various types of IRC channels.

Specifically, the research question that is discussed in this section examines which factors, extracted from the analysis of IRC channel interaction dynamics, can be used to predict short-term chat-channel activity as reliably, accurately and effectively as possible. Channel activity is a surrogate measure for the group interactions occurring inside IRC chat-channels. Channel activity can easily be operationalized in terms of many distinct measures such as the overall number of potential contributors (e.g., number of users per channel); the overall number of actual contributors (e.g., number of posters per channel); the overall number of public messages; the rates of contribution per user (e.g., number of messages per user); the rates of contribution per poster (e.g., number of messages per poster); the complexity of the contributions (e.g., the number of words per message); the proportion of messages that receive replies; or the number of distinct threads of conversation. While all these variables represent the level of group interaction as they clearly indicate the amount and the intensity of the public activity of any IRC chat-channel, none were addressed in any previous work. Therefore, a starting point was needed. This research focuses on one variable, specifically the number of

actual contributors, and tries to make short-term predictions about the number of posters that would be active in an IRC chat-channel during a particular time interval.

9.1 Hypothesis

Considering the above, it is hypothesized that for any publicly active channel and for any short-term time interval for which the level of channel activity is predicted, there will be three main categories of factors that will have an impact on the accuracy of the predictions: (1) the trajectories of channel activity during various previous time periods; (2) the trajectories of network activity during various time periods; and (3) the seasonality of the channels, i.e., rhythms information about each individual channel.

9.2 Method

9.2.1 Data Considerations

The analysis was performed on the same stratified random sample of channels that was used in the previous chapter. The selection process was described in details in Chapter 8, subsection 8.2.1.

A very important issue that needed to be addressed was the notion of "short term." This notion is dependent on the medium as, for example, asynchronous communication such as emails or Usenet newsgroups are different from synchronous communication such as IRC. For IRC, "short-term" is relative to the length of the conversations occurring inside chat-channels as well as to the time spent by users in a chat session. Public conversations are exchanges of messages between people who are co-present in an IRC channel. Ideally, the duration of the co-presence of any two users

engaged in a public conversation would be a great indicator for the definition of "short-term" interaction. Unfortunately, it was impossible to compute the values of such a co-presence variable. Therefore, the next best indicator for the value of "short-term" interactions was the time spent during an IRC session by the users of the IRC network. The user-related descriptive statistics presented in the previous chapter showed that the most common (modal) value of the average time spent by a user in a session was 10 minutes, and that 50 percent of the users spent 40 minutes or less during an IRC session.

Thus, considering the dynamic nature of chat as well as the constraints of the data-collection method, which allowed the collection of only three measurements per hour, for the rest of this research the notion of "short-term interval" will be equivalent to "20-minute interval." Every hour is divided into exactly three 20-minute time intervals (the first 20 minutes of the hour, the middle 20 minutes of the hour, and the last 20 minutes of the hour). For any given time, in any 20-minute interval, short-term predictions are defined as predictions for the immediate 20-minute interval.

9.2.2 Variables and Measures

Currently there is a lack of research in the area of predicting activity in synchronous systems. Therefore, no well-known predictor variables exist that could be used to predict the activity of IRC channels. Consequently, there is a need to choose a starting point from which to make such predictions. Considering the above discussion and the knowledge about IRC, gained from examining the descriptive statistics presented in the previous chapter; several predicting variables were defined:

- AvgOP_Prev3_20 - the average of the observed number of posters during the previous three 20-minute time intervals for each channel in the sample;

- AvgOP_PrevHr_Nwrk - the average number of observed posters per channel for the entire network for the previous hour, i.e., the total number of posters divided by the total number of publicly active channels;

- AvgOP_Prev3_20_Nwrk - the average number of observed posters per channel for the entire network during the previous three 20-minute time intervals. This differs from the second predictor in that it was computed after taking the three 20-minute intervals that formed the previous hour into account, i.e. the predicted value was the average of the previous three intervals rather than the average of the previous hour;

- AvgOP_Prev3wks - the average number of observed posters for the closest three 20-minutes intervals (just before, current, and just after) at the same time during the previous three weeks for each channel in the sample. For example, for the interval 5:01 p.m. – 5:20 p.m. of Monday, September 5, 2005; this predictor was computed as the average of the intervals 4:41 p.m. – 5.00 p.m., 5:01 p.m. – 5:20 p.m. and 5:21p.m. – 5:40 p.m. for August 29, 2005, August 22, 2005, and August 15, 2005, which were the previous three Mondays;

- AvgOP_Prev12wks_Nwrk - the average number of observed posters per channel for the entire network for the closest three 20-minutes intervals (just before, current, and just after) at the same time during the previous 12 weeks;

- Slope - the slope of the line determined by the observed values for the previous three 20-minute time intervals for each channel; it is a basic indicator of the amount by which the number of posters varied during the previous hour;

- SP (Seasonality Predictor) - the value predicted by a time series analysis of the observed values per channel during the interval August 1, 2005 to August 31, 2005;

- TC (Trajectory Coefficient) - a correlation coefficient between "time" and the observed number of posters during the last hour, which gives a general idea about the direction of the conversation (up, down, or constant).

With respect to the hypothesis presented in Section 9.1, AvgOP_Prev3_20, Slope, and TC provide information about the trajectories of channels activity; AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, and AvgOP_Prev12wks_Nwrk provide

information about the trajectories of network activity; and AvgOP_Prev3wks and SP provide information about the rhythms of the channels.

9.2.3 Data Analysis

To explore the short-term predictability of channel engagement, the number of posters was selected as the surrogate measure for the overall activity of chat-channels. Then, the aim was to understand the general predictability of the number of posters for each channel in the selected sample during the interval September 1, 2005 to September 7, 2005. Since the sample of channels was selected from the month of August 2005, and many descriptive statistics for these channels were computed for this month; it made sense to choose an interval that immediately followed to make predictions. This is why the first week of September was selected. Regression Analysis was the method chosen to make short-term predictions about the number of posters present in a channel.

Regression analysis is a technique used for the modeling and analysis of numerical data consisting of values of a dependent variable and of independent variables: it examines the relation of the dependent variable to one or more specified independent variables. In this research, the dependent variable was called ObservedPosters and represented the observed number of posters for each 20-minute interval, i.e., the actual number of posters that were recorded during each 20-minute interval during the first week of September 2005. The independent variables were the eight predictors described is section 9.3.1. Regression analysis was used to examine the utility of these independent variables with respect to predicting the value of the dependent variable.

Linear regression is a form of regression analysis in which observational data are modeled by a function that is a linear combination of the model parameters and depends

on one or more independent variables. In simple linear regression the model function represents a straight line.

Nonlinear regression is a form of regression analysis in which observational data are modeled by a function that is a nonlinear combination of the model parameters and depends on one or more independent variables.

Both linear and nonlinear regression analysis were used to predict the values of the dependent variable based on the set of eight independent variables.

While for six of the eight predicting variables there was only one straightforward method of computation (AvgOP_Prev3_20, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wks, AvgOP_Prev12wks_Nwrk, and Slope); for the other two (SP and TC), there were multiple possibilities of computation.

The seasonality predictor (SP) was a variable derived from the seasonality analysis of the ObservedPosters variable for each channel during the month of August 2005. For each channel in the sample, a seasonal decomposition was performed in he SPSS software for August 2005. The seasonal decomposition produced four variables:

- SAF: seasonal adjustment factors, representing seasonal variation. Seasonal factors could be used as input to an exponential smoothing model;
- SAS: seasonally adjusted series, representing the original series with seasonal variations removed;
- STC: smoothed trend-cycle component, a smoothed version of the seasonally adjusted series that showed both trend and cyclic components;
- ERR: the residual component of the series for a particular observation.

To predict the ObservedPosters values for the first week of September, an exponential smoothing of the ObservedPosters variable was performed, taking into account the SAF variable as the seasonal factor. This led to the SP variable, which

basically represents the predicted number of posters as produced by the exponential smoothing of the seasonal adjusted series. There are two types of seasonal decomposition: multiplicative and additive. In a multiplicative seasonal decomposition the seasonal component is a factor by which the seasonally adjusted series is multiplied to yield the original series. In an additive seasonal decomposition the seasonal adjustments are added to the seasonally adjusted series to obtain the observed values.

Both these types of seasonal decomposition were used to compute the seasonality predictor, resulting in two distinct possible values for the SP variable. Furthermore, there were three days in August that were problematic in terms of data-collection. During these three days, the values of the ObservedPosters variable were lower than normal because of various connectivity issues of the IRC server. Therefore, two different approaches were used to compute the SP variables. First, the missing values were replaced using the "mean of nearby points" method, and then the seasonal decomposition and exponential smoothing were performed on the data. Second, the problematic days were completely excluded from the analysis; and the seasonal decomposition and the exponential smoothing were performed on the reduced dataset.

In conclusion, there were four possible instances of the SP variable: one resulted from the multiplicative seasonal decomposition on the full data set (SP1); one resulted from the additive seasonal decomposition on the full data set (SP2); one resulted from the multiplicative seasonal decomposition on the reduced data set (SP3); and one resulted from the additive seasonal decomposition on the reduced data set (SP4).

The trajectory coefficient (TC) variable was computed for each 20-minute interval in the first week of September 2005 as the correlation coefficient between time and the

ObservedPosters variable during the previous three 20-minute intervals. The starting time of each interval was expressed as the number of seconds that had elapsed since midnight Coordinated Universal Time of January 1, 1970, not counting leap seconds – a representation of time widely used in many operating systems. TC's values ranged from -1 to 1; it was used as an indicator of the general trajectory of the observed number of posters during the previous hour. There were two possible instances for this variable, one given by the Pearson correlation coefficient (TC1) and one given by the Spearman correlation coefficient (TC2).

Of the eight independent variables, four of them had at least one negative value and/or one zero value (AvgOP_Prev3_20, AvgOP_Prev3wks, Slope, and both instances of TC) while the dependent variable (ObservedPosters) also had a significant number of zero values. To avoid potential problems caused by the non-positive values, transformations were performed on these variables, making all of them positive. The modified variables were named AvgOP_Prev3_20Mod, AvgOP_Prev3wksMod, SlopeMod, TCxMod (where x was 1 or 2 depending on the type of computation), and ObservedPostersMod.

Finally, considering all of the above, two regression models were produced for both the linear regression analysis and for the non-linear regression analysis. The first model included all the initial, non-transformed independent and dependent variables; the second model included the initial independent variables that had only positive values, together with the dependent and independent variables that needed to be transformed.

Guided by the results of the regression models described above, a new variable that maximized overall predictability, referred to as the BestPredictor (BP) variable, was computed.

To avoid multicollinearity problems, only one of the four possible SP instances and only one of the two possible TC instances were entered into the regression models at a time, together with the other six independent variables (AvgOP_Prev3_20, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wks, AvgOP_Prev12wks_Nwrk, and Slope). These combinations resulted in a total of eight values for BestPredictor variable for each model, for both the linear regression analysis and the non-linear regression analysis.

Table 9.1 summarizes the variables used in each regression model.

Table 9.1 Summary of the Regression Models

Model	Variation	Independent variables	Dependent variable	Best predictor
1	1	AvgOP_Prev3_20, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wks, AvgOP_Prev12wks_Nwrk, Slope, SP1, TC1	ObservedPosters	BP1
	2	AvgOP_Prev3_20, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wks, AvgOP_Prev12wks_Nwrk, Slope, SP1, TC2	ObservedPosters	BP2
	3	AvgOP_Prev3_20, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wks, AvgOP_Prev12wks_Nwrk, Slope, SP2, TC1	ObservedPosters	BP3
	4	AvgOP_Prev3_20, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wks, AvgOP_Prev12wks_Nwrk, Slope, SP2, TC2	ObservedPosters	BP4
	5	AvgOP_Prev3_20, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wks, AvgOP_Prev12wks_Nwrk, Slope, SP3, TC1	ObservedPosters	BP5
	6	AvgOP_Prev3_20, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wks, AvgOP_Prev12wks_Nwrk, Slope, SP3, TC2	ObservedPosters	BP6
	7	AvgOP_Prev3_20, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wks, AvgOP_Prev12wks_Nwrk, Slope, SP4, TC1	ObservedPosters	BP7
	8	AvgOP_Prev3_20, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wks, AvgOP_Prev12wks_Nwrk, Slope, SP4, TC4	ObservedPosters	BP8
2	1	AvgOP_Prev3_20Mod, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wksMod, AvgOP_Prev12wks_Nwrk, SlopeMod, SP1, TC1Mod	ObservedPostersMod	BP1
	2	AvgOP_Prev3_20Mod, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wksMod, AvgOP_Prev12wks_Nwrk, SlopeMod, SP1, TC2Mod	ObservedPostersMod	BP2
	3	AvgOP_Prev3_20Mod, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wksMod, AvgOP_Prev12wks_Nwrk, SlopeMod, SP2, TC1Mod	ObservedPostersMod	BP3
	4	AvgOP_Prev3_20Mod, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wksMod, AvgOP_Prev12wks_Nwrk, SlopeMod, SP2, TC2Mod	ObservedPostersMod	BP4
	5	AvgOP_Prev3_20Mod, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wksMod, AvgOP_Prev12wks_Nwrk, SlopeMod, SP3, TC1Mod	ObservedPostersMod	BP5
	6	AvgOP_Prev3_20Mod, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wksMod, AvgOP_Prev12wks_Nwrk, SlopeMod, SP3, TC2Mod	ObservedPostersMod	BP6
	7	AvgOP_Prev3_20Mod, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wksMod, AvgOP_Prev12wks_Nwrk, SlopeMod, SP4, TC1Mod	ObservedPostersMod	BP7
	8	AvgOP_Prev3_20Mod, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wksMod, AvgOP_Prev12wks_Nwrk, SlopeMod, SP4, TC2Mod	ObservedPostersMod	BP8

9.3 Results

9.3.1 Linear Regression

9.3.1.1 First Regression Model. To explore channel predictability, the correlations between the various BestPredictor variables and the ObservedPosters variable were examined for both regression models described in subsection 9.2.3, using the Spearman correlation coefficient. To compare the results of general predictions, obtained from regression equations created for the entire sample of channels, to results of more specific predictions, obtained from regression equations that targeted particular subgroups of channels; the linear regression analysis was carried out at three main levels. There were: for all the channels in the sample; for all three subgroups of channels based on size (small channels, medium channels, and large channels); and for all three subgroups of channels based on intensity (low-intensity channels, medium-intensity channels, and high-intensity channels).

Table 9.2 describes the regression equations for each BestPredictor variable and the corresponding R^2 values, computed for all the channels in the sample. It may be observed that four regression equations produced an R^2 value of 0.785, while the other four regression equations produced an R^2 value of 0.790. Theoretically, the BestPredictor variable produced by any of these latter four regression equations could be used as the single best predicting variable for the number of posters present inside a channel. However, to make the prediction model as simple as possible, one single variation of the linear regression model was selected as the final prediction model. This was variation number 6, which included seasonality predictor SP3 and trajectory coefficient TC2 and produced BP6 as the best predictor. The SP3 predictor was chosen because it produced

the highest R^2 value when entered as the only independent variable in a linear regression equation with the ObservedPosters as the dependent variable. Given that using either TC1 or TC2 with the best prediction model yielded the same R^2 value, the decision to use TC2 was randomly taken.

Table 9.2 Regression Equations for Best Predictors

Regression equation	R^2
BP1 = -.508 + .598*AvgOP_Prev3_20 - .034*AvgOP_PrevHr_Nwrk + .064*AvgOP_Prev3_20_Nwrk + .011*AvgOP_Prev3wks + .024*AvgOP_Prev12wks_Nwrk + .341*SP1 - .537*Slope + .218*TC1	0.790
BP2 = -.508 + .597*AvgOP_Prev3_20 - .034*AvgOP_PrevHr_Nwrk + .064*AvgOP_Prev3_20_Nwrk + .011*AvgOP_Prev3wks + .024*AvgOP_Prev12wks_Nwrk + .342*SP1 - .544*Slope + .203*TC2	0.790
BP3 = -.415 + .665*AvgOP_Prev3_20 - .034*AvgOP_PrevHr_Nwrk + .066*AvgOP_Prev3_20_Nwrk + .014*AvgOP_Prev3wks + .316*SP2 - .433*Slope + .274*TC1	0.785
BP4 = -.415 + .665*AvgOP_Prev3_20 - .034*AvgOP_PrevHr_Nwrk + .066*AvgOP_Prev3_20_Nwrk + .014*AvgOP_Prev3wks + .316*SP2 - .439*Slope + .263*TC2	0.785
BP5 = -.458 + .593*AvgOP_Prev3_20 + .039*AvgOP_Prev3_20_Nwrk + .010*AvgOP_Prev3wks + .331*SP3 - .556*Slope + .201*TC1	0.790
BP6 = -.458 + .592*AvgOP_Prev3_20 + .039*AvgOP_Prev3_20_Nwrk + .010*AvgOP_Prev3wks + .331*SP3 - .563*Slope + .187*TC2	0.790
BP7 = -.395 + .678*AvgOP_Prev3_20 - .035*AvgOP_PrevHr_Nwrk + .066*AvgOP_Prev3_20_Nwrk + .015*AvgOP_Prev3wks + .300*SP4 - .430*Slope + .278*TC1	0.785
BP8 = -.395 + .677*AvgOP_Prev3_20 - .035*AvgOP_PrevHr_Nwrk + .066*AvgOP_Prev3_20_Nwrk + .015*AvgOP_Prev3wks + .300*SP4 - .437*Slope + .266*TC2	0.785

The results obtained from the best linear regression model are described in Tables 9.3 through 9.6. Both the stepwise forward and the backward forward methods were, used, and they produced exactly the same results.

Table 9.3 Summary of Best Linear Regression Prediction Model for All Channels

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
	.889	.790	.790	.955

Predictors: (Constant), TC2, AvgOP_Prev3wks, AvgOP_Prev3_20, Slope, AvgOP_Prev3_20_Nwrk, SP3

Table 9.4 ANOVA for Best Linear Regression Prediction Model for All Channels

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	155728.525	6	25954.754	28472.426	.000
	Residual	41342.666	45353	.912		
	Total	197071.190	45359			

Predictors: (Constant), TC2, AvgOP_Prev3wks, AvgOP_Prev3_20, Slope, AvgOP_Prev3_20_Nwrk, SP3
Dependent Variable: ObservedPosters

Table 9.5 Coefficients for Best Linear Regression Prediction Model for All Channels

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-.458	.026		-17.494	.000
AvgOP_Prev3_20	.592	.005	.512	117.267	.000
AvgOP_Prev3_20_Nwrk	.039	.008	.011	4.943	.000
AvgOP_Prev3wks	.010	.001	.018	7.621	.000
SP3	.331	.004	.345	81.676	.000
Slope	-.563	.013	-.120	-43.141	.000
TC2	.187	.019	.025	9.835	.000

Dependent Variable: ObservedPosters

Table 9.6 Variables Excluded from Best Linear Regression Prediction Model for All Channels

Model	Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
					Tolerance
AvgOP_Prev12wks_Nwrk	.000(b)	-.122	.903	-.001	.399
AvgOP_PrevHr_Nwrk	-.008(b)	-1.635	.102	-.008	.218

Predictors in the Model: (Constant), TC2, AvgOP_Prev3wks, AvgOP_Prev3_20, Slope, AvgOP_Prev3_20_Nwrk, SP3
Dependent Variable: ObservedPosters

Overall channel predictability was measured by the correlation coefficients between the best predictor BP6 produced by the best model and the ObservedPosters dependent variable. Both Pearson and Spearman correlation coefficients were computed.

The Pearson correlation coefficient computed between BP6 and ObservedPosters had the highest value among all the eight best predictors (r = 0.884). The Spearman correlation coefficient computed between BP6 and ObservedPosters was slightly lower than the coefficients computed between some of the other best predictor variables (rho = 0.662).

However, the differences between the Spearman correlation coefficient obtained from the best prediction model and the Spearman correlation coefficients obtained from the other seven models were quite low, never larger than 0.015.

Table 9.7 breaks down the results of the best linear regression prediction model by subgroups, and presents the Spearman correlation coefficients for each of the nine types of channels.

Table 9.7 Best Model Correlation Coefficients for All Channels Grouped by Type

Type				BP6
1	Spearman's rho	ObservedPosters	Correlation Coefficient	.492(**)
			Sig. (2-tailed)	.000
			N	5040
2	Spearman's rho	ObservedPosters	Correlation Coefficient	.570(**)
			Sig. (2-tailed)	.000
			N	5040
3	Spearman's rho	ObservedPosters	Correlation Coefficient	.427(**)
			Sig. (2-tailed)	.000
			N	5040
4	Spearman's rho	ObservedPosters	Correlation Coefficient	.607(**)
			Sig. (2-tailed)	.000
			N	5040
5	Spearman's rho	ObservedPosters	Correlation Coefficient	.589(**)
			Sig. (2-tailed)	.000
			N	5040
6	Spearman's rho	ObservedPosters	Correlation Coefficient	.370(**)
			Sig. (2-tailed)	.000
			N	5040
7	Spearman's rho	ObservedPosters	Correlation Coefficient	.572(**)
			Sig. (2-tailed)	.000
			N	5040
8	Spearman's rho	ObservedPosters	Correlation Coefficient	.520(**)
			Sig. (2-tailed)	.000
			N	5040
9	Spearman's rho	ObservedPosters	Correlation Coefficient	.838(**)
			Sig. (2-tailed)	.000
			N	5040

** Correlation is significant at the 0.01 level (2-tailed).

Table 9.8 breaks down the results based on the size of the channels in the sample, while Table 9.9 breaks down the results based on the intensity of the channels in the sample.

Table 9.8 Best Model Correlation Coefficients for All Channels Grouped by Size

Size				BP6
large	Spearman's rho	ObservedPosters	Correlation Coefficient	.757(**)
			Sig. (2-tailed)	.000
			N	15120
medium	Spearman's rho	ObservedPosters	Correlation Coefficient	.513(**)
			Sig. (2-tailed)	.000
			N	15120
small	Spearman's rho	ObservedPosters	Correlation Coefficient	.516(**)
			Sig. (2-tailed)	.000
			N	15120

Table 9.9 Best Model Correlation Coefficients for All Channels Grouped by Intensity

Intensity				BP6
high	Spearman's rho	ObservedPosters	Correlation Coefficient	.729(**)
			Sig. (2-tailed)	.000
			N	15120
low	Spearman's rho	ObservedPosters	Correlation Coefficient	.345(**)
			Sig. (2-tailed)	.000
			N	15120
medium	Spearman's rho	ObservedPosters	Correlation Coefficient	.338(**)
			Sig. (2-tailed)	.000
			N	15120

For all these cases BP6 was computed using the same regression equation, i.e., there were no specific regression equations for different types or subgroups of channels. The values for different types of channels ranged from a minimum of 0.370 (type 6) to a maximum of 0.838 (type 9). For subgroups of channels, correlation coefficients were higher for large channels compared to medium and small channels; and higher for high-intensity channels compared to medium- and low-intensity channels. The tables that report the correlation coefficients between the observed posters and the rest of the computed best predictors can be found in the Appendix.

** Correlation is significant at the 0.01 level (2-tailed).

Table 9.10 presents the regression equations for each BestPredictor variable and the corresponding R^2 values, computed for all the small channels.

Table 9.10 Regression Equations and R^2 Values for Small Channels

Regression equation	R^2
BP1 = -.102 + .628*AvgOP_Prev3_20 + .013*AvgOP_Prev12wks_Nwrk + .071*SP1 - .221*Slope + .302*TC1	0.367
BP2 = -.101 + .627*AvgOP_Prev3_20 + .013*AvgOP_Prev12wks_Nwrk + .071*SP1 - .225*Slope + .297*TC2	0.367
BP3 = -.131 + .624*AvgOP_Prev3_20 + .010*AvgOP_Prev12wks_Nwrk + .107*SP2 - .225*Slope + .298*TC1	0.368
BP4 = -.131 + .622*AvgOP_Prev3_20 + .010*AvgOP_Prev12wks_Nwrk + .107*SP2 - .228*Slope + .293*TC2	0.368
BP5 = -.096 + .629*AvgOP_Prev3_20 + .013*AvgOP_Prev12wks_Nwrk + .066*SP3 - .222*Slope + .301*TC1	0.367
BP6 = -.096 + .627*AvgOP_Prev3_20 + .013*AvgOP_Prev12wks_Nwrk + .065*SP3 - .226*Slope + .296*TC2	0.367
BP7 = -.131 + .624*AvgOP_Prev3_20 + .010*AvgOP_Prev12wks_Nwrk + .106*SP4 - .225*Slope + .298*TC1	0.369
BP8 = -.131 + .622*AvgOP_Prev3_20 + .010*AvgOP_Prev12wks_Nwrk + .106*SP4 - .228*Slope + .293*TC2	0.369

The best prediction models, which produced the highest R^2 value (0.369), were obtained from variation numbers 7 and 8, which used the seasonality predictor SP4. Given that using either TC1 or TC2 yielded the same R^2 value, the decision to use TC2 was randomly taken. Hence, BP8 was selected in this case as the best predictor.

The results obtained from the best linear regression model are described in Tables 9.11 through 9.14. Both the stepwise forward and the backward forward methods were used, and they produced exactly the same results.

Table 9.11 Summary of Best Linear Regression Prediction Model for Small Channels

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
	.606	.368	.367	.306

Predictors: (Constant), TC2, SP4, AvgOP_Prev3_20, Slope, AvgOP_Prev12wks_Nwrk

Table 9.12 ANOVA for Best Linear Regression Prediction Model for Small Channels

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	822.988	5	164.598	1756.482	.000
Residual	1416.313	15114	.094		
Total	2239.301	15119			

Predictors: (Constant), TC2, SP4, AvgOP_Prev3_20, Slope, AvgOP_Prev12wks_Nwrk
Dependent Variable: ObservedPosters

Table 9.13 Coefficients for Best Linear Regression Prediction Model for Small Channels

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-.131	.019		-6.934	.000
AvgOP_Prev3_20	.622	.012	.469	53.299	.000
AvgOP_Prev12wks_Nwrk	.010	.004	.018	2.774	.006
SP4	.106	.014	.051	7.586	.000
Slope	-.228	.016	-.137	-14.450	.000
TC2	.293	.016	.137	18.544	.000

Dependent Variable: ObservedPosters

Table 9.14 Variables Excluded from Best Linear Regression Prediction Model for Small Channels

Model	Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
					Tolerance
AvgOP_Prev3_20_Nwrk	-.012	-1.169	.242	-.010	.404
AvgOP_PrevHr_Nwrk	-.020	-1.470	.142	-.012	.231
AvgOP_Prev3wks	-.010	-1.598	.110	-.013	.999

Predictors in the Model: (Constant), TC2, SP4, AvgOP_Prev3_20, Slope, AvgOP_Prev12wks_Nwrk
Dependent Variable: ObservedPosters

Overall channel predictability was measured by the correlation coefficients between the best predictor BP8 produced by the best model and the ObservedPosters dependent variable. Both Pearson and Spearman correlation coefficients were computed.

The Pearson correlation coefficient computed between BP8 and ObservedPosters had the highest value among all the eight best predictors (r = 0.561). The Spearman correlation coefficient computed between BP8 and ObservedPosters was slightly lower than some of the coefficients computed between the other seven BP variables and the ObservedPosters (rho = 0.517). However, the differences between the Spearman

correlation coefficient obtained from the best prediction model and the Spearman correlation coefficients obtained from the other seven models were almost negligible, never larger than 0.003.

Table 9.15 breaks down the results of the best linear regression prediction model and presents the Spearman correlation coefficients for each of the three types of channels corresponding to the subgroup of small channels. The tables that report the correlation coefficients between the observed posters and the rest of the computed best predictors can be found in the Appendix.

Table 9.15 Best Model Correlation Coefficients for All Small Channels, Grouped by Type

Type				BP8
1	Spearman's rho	ObservedPosters	Correlation Coefficient	.536(**)
			Sig. (2-tailed)	.000
			N	5040
2	Spearman's rho	ObservedPosters	Correlation Coefficient	.580(**)
			Sig. (2-tailed)	.000
			N	5040
3	Spearman's rho	ObservedPosters	Correlation Coefficient	.407(**)
			Sig. (2-tailed)	.000
			N	5040

Table 9.16 presents the regression equations for each BestPredictor variable and the corresponding R^2 values, computed for all the medium channels.

Table 9.16 Regression Equations and R^2 Values for Medium Channels

Regression equation	R^2
BP1 = -.284 + .647*AvgOP_Prev3_20 + .150*SP1 - .445*Slope + .310*TC1	0.404
BP2 = -.284 + .646*AvgOP_Prev3_20 + .150*SP1 - .458*Slope + .306*TC2	0.404
BP3 = -.334 + .572*AvgOP_Prev3_20 + .335*SP2 - .429*Slope + .318*TC1	0.421
BP4 = -.334 + .571*AvgOP_Prev3_20 + .335*SP2 - .431*Slope + .316*TC2	0.421
BP5 = -.270 + .649*AvgOP_Prev3_20 + .140*SP3 - .457*Slope + .308*TC1	0.404
BP6 = -.271 + .648*AvgOP_Prev3_20 + .140*SP3 - .460*Slope + .305*TC2	0.404
BP7 = -.299 + .569*AvgOP_Prev3_20 + .304*SP4 - .435*Slope + .315*TC1	0.423
BP8 = -.300 + .568*AvgOP_Prev3_20 + .305*SP4 - .437*Slope + .312*TC2	0.423

** Correlation is significant at the 0.01 level (2-tailed).

The best prediction models, which produced the highest R^2 value (0.369), were obtained from variation numbers 7 and 8, which used the seasonality predictor SP4. Given that using either TC1 or TC2 yielded the same R^2 value, the decision to use TC2 was randomly taken. Hence, BP8 was selected in this case as the best predictor.

The results obtained from the best linear regression model are described in Tables 9.17 through 9.20. Both the stepwise forward and the backward forward methods were used, and they produced exactly the same results.

Table 9.17 Summary of Best Linear Regression Prediction Model for Medium Channels

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
	.650	.422	.422	.729

Predictors: (Constant), TC2, AvgOP_Prev3_20, SP4, Slope

Table 9.18 ANOVA for Best Linear Regression Prediction Model for Medium Channels

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	5877.269	4	1469.317	2763.903	.000(e)
	Residual	8035.279	15115	.532		
	Total	13912.548	15119			

Predictors: (Constant), TC2, AvgOP_Prev3_20, SP4, Slope
Dependent Variable: ObservedPosters

Table 9.19 Coefficients for Best Linear Regression Prediction Model for Medium Channels

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-.300	.015		-20.187	.000
AvgOP_Prev3_20	.568	.010	.436	55.424	.000
SP4	.305	.012	.173	24.531	.000
Slope	-.437	.021	-.169	-21.212	.000
TC2	.312	.027	.081	11.492	.000

Dependent Variable: ObservedPosters

Table 9.20 Variables Excluded from Best Model for Medium Channels

Model	Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
					Tolerance
AvgOP_Prev3wks	.001(d)	.134	.894	.001	.998
AvgOP_PrevHr_Nwrk	.008(d)	1.232	.218	.010	.955
AvgOP_Prev3_20_Nwrk	.009(d)	1.415	.157	.012	.964
AvgOP_Prev12wks_Nwrk	.010(d)	1.558	.119	.013	.942

Predictors in the Model: (Constant), TC2, AvgOP_Prev3_20, SP4, Slope
Dependent Variable: ObservedPosters

** Correlation is significant at the 0.01 level (2-tailed).

Overall channel predictability was measured by the correlation coefficients between the best predictor BP8, produced by the best model, and the ObservedPosters dependent variable. Both Pearson and Spearman correlation coefficients were computed.

The Pearson correlation coefficient computed between BP8 and ObservedPosters had the highest value among all the eight best predictors (r = 0.642). The Spearman correlation coefficient computed between BP8 and ObservedPosters was slightly lower than some of the coefficients computed between the other seven BP variables and the ObservedPosters (rho = 0.575). However, the differences between the Spearman correlation coefficient obtained from the best prediction model and the Spearman correlation coefficients obtained from the other seven models were quite low, never larger than 0.033.

Table 9.21 breaks down the results of the best linear regression prediction model and presents the Spearman correlation coefficients for each of the three types of channels corresponding to the subgroup of medium channels.

Table 9.21 Best Model Correlation. Coefficients for All Medium Channels, Grouped by Type

Type				BP8
4	Spearman's rho	ObservedPosters	Correlation Coefficient	.575(**)
			Sig. (2-tailed)	.000
			N	5040
5	Spearman's rho	ObservedPosters	Correlation Coefficient	.600(**)
			Sig. (2-tailed)	.000
			N	5040
6	Spearman's rho	ObservedPosters	Correlation Coefficient	.545(**)
			Sig. (2-tailed)	.000
			N	5040

** Correlation is significant at the 0.01 level (2-tailed).

Table 9.22 presents the regression equations for each BestPredictor variable and the corresponding R^2 values, computed for all the large channels.

Table 9.22 Regression Equations and R^2 Values for Large Channels

Regression equation	R^2
BP1 = -.686 + .573*AvgOP_Prev3_20 - .061*AvgOP_PrevHr_Nwrk + .180*AvgOP_Prev3_20_Nwrk + .028*AvgOP_Prev3wks+ .342*SP1 - .638*Slope + .118*TC1	0.818
BP2 = -.686 + .572*AvgOP_Prev3_20 - .061*AvgOP_PrevHr_Nwrk + .180*AvgOP_Prev3_20_Nwrk + .028*AvgOP_Prev3wks+ .342*SP1 - .648*Slope + .096*TC2	0.818
BP3 = -.575 + .668*AvgOP_Prev3_20 - .096*AvgOP_PrevHr_Nwrk + .189*AvgOP_Prev3_20_Nwrk + .038*AvgOP_Prev3wks+ .289*SP2 - .534*Slope + .177*TC1	0.811
BP4 = -.575 + .667*AvgOP_Prev3_20 - .096*AvgOP_PrevHr_Nwrk + .189*AvgOP_Prev3_20_Nwrk + .038*AvgOP_Prev3wks+ .289*SP2 - .542*Slope + .160*TC2	0.811
BP5 = -.617 + .566*AvgOP_Prev3_20 - .072*AvgOP_PrevHr_Nwrk + .177*AvgOP_Prev3_20_Nwrk + .027*AvgOP_Prev3wks+ .333*SP3 - .662*Slope + .094*TC1	0.819
BP6 = -.617 + .565*AvgOP_Prev3_20 - .072*AvgOP_PrevHr_Nwrk + .177*AvgOP_Prev3_20_Nwrk + .027*AvgOP_Prev3wks+ .333*SP3 - .672*Slope + .072*TC2	0.819
BP7 = -.549 + .680*AvgOP_Prev3_20 - .100*AvgOP_PrevHr_Nwrk + .191*AvgOP_Prev3_20_Nwrk + .040*AvgOP_Prev3wks+ .273*SP4 - .537*Slope + .176*TC1	0.810
BP8 = -.549 + .680*AvgOP_Prev3_20 - .100*AvgOP_PrevHr_Nwrk + .191*AvgOP_Prev3_20_Nwrk + .040*AvgOP_Prev3wks+ .273*SP4 - .545*Slope + .159*TC2	0.810

The best prediction models, which produced the highest R^2 value (0.819), were obtained from variation numbers 5 and 6, which used the seasonality predictor SP3. Given that using either TC1 or TC2 yielded the same R^2 value, the decision to use TC2 was randomly taken. Hence, BP6 was selected in this case as the best predictor.

The results obtained from the best linear regression model are described in Tables 9.23 through 9.26. Both the stepwise forward and the backward forward methods were used, and they produced exactly the same results.

Table 9.23 Summary of Best Linear Regression Prediction Model for Large Channels

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
	.905	.818	.818	1.437

Predictors: (Constant), TC2, AvgOP_Prev3wks, AvgOP_Prev3_20, Slope, AvgOP_Prev3_20_Nwrk, SP3, AvgOP_PrevHr_Nwrk

Table 9.24 ANOVA for Best Linear Regression Prediction Model for Large Channels

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	140487.765	7	20069.681	9724.865	.000
	Residual	31187.376	15112	2.064		
	Total	171675.140	15119			

Predictors: (Constant), TC2, AvgOP_Prev3wks, AvgOP_Prev3_20, Slope, AvgOP_Prev3_20_Nwrk, SP3, AvgOP_PrevHr_Nwrk
Dependent Variable: ObservedPosters

Table 9.25 Coefficients for Best Linear Regression Prediction Model for Large Channels

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-.617	.068		-9.082	.000
AvgOP_Prev3_20	.565	.009	.493	66.044	.000
AvgOP_PrevHr_Nwrk	-.072	.032	-.017	-2.281	.023
AvgOP_Prev3_20_Nwrk	.177	.044	.030	4.060	.000
AvgOP_Prev3wks	.027	.003	.035	7.852	.000
SP3	.333	.007	.361	48.010	.000
Slope	-.672	.025	-.124	-26.404	.000
TC2	.072	.039	.008	1.840	.066

Dependent Variable: ObservedPosters

Table 9.26 Variables Excluded from Best Linear Regression Model for Large Channels

Model	Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
					Tolerance
AvgOP_Prev12wks_Nwrk	.002	.235	.814	.002	.228

Predictors in the Model: (Constant), TC2, AvgOP_Prev3wks, AvgOP_Prev3_20, Slope, AvgOP_Prev3_20_Nwrk, SP3, AvgOP_PrevHr_Nwrk
Dependent Variable: ObservedPosters

Overall channel predictability was measured by the correlation coefficients between the best predictor BP6, produced by the best model, and the ObservedPosters dependent variable. Both Pearson and Spearman correlation coefficients were computed. The Pearson correlation coefficient computed between BP6 and ObservedPosters had the highest value among all the eight best predictors ($r = 0.901$).

The Spearman correlation coefficient computed between BP6 and ObservedPosters was also higher than all of the coefficients computed between the other seven BP variables and the ObservedPosters (rho = 0.734). However, the differences between the Spearman correlation coefficient obtained from the best prediction model and the Spearman correlation coefficients obtained from the other seven models were negligible, never larger than 0.009. Table 9.27 breaks down the results of the best linear regression prediction model and presents the Spearman correlation coefficients for each of the three types of channels corresponding to the subgroup of large channels. The tables that report the correlation coefficients between the observed posters and the rest of the computed best predictors can be found in the Appendix.

Table 9.27 Best Model Correlation Coefficients for All Large Channels, Grouped by Type

Type				BP6
7	Spearman's rho	ObservedPosters	Correlation Coefficient	.531(**)
			Sig. (2-tailed)	.000
			N	5040
8	Spearman's rho	ObservedPosters	Correlation Coefficient	.465(**)
			Sig. (2-tailed)	.000
			N	5040
9	Spearman's rho	ObservedPosters	Correlation Coefficient	.833(**)
			Sig. (2-tailed)	.000
			N	5040

** Correlation is significant at the 0.01 level (2-tailed).

Table 9.28 presents the regression equations for each BestPredictor variable and the corresponding R^2 values, computed for all the low-intensity channels.

Table 9.28 Regression Equations and R^2 Values for Low-Intensity Channels

Regression equation	R^2
BP1 = -.204 + .619*AvgOP_Prev3_20 + .045*AvgOP_Prev12wks_Nwrk + .064*SP1 - .510*Slope + .246*TC1	0.375
BP2 = -.204 + .619*AvgOP_Prev3_20 + .044*AvgOP_Prev12wks_Nwrk + .064*SP1 - .513*Slope + .240*TC2	0.375
BP3 = -.245 + .589*AvgOP_Prev3_20 + .029*AvgOP_Prev12wks_Nwrk + .147*SP2 - .473*Slope + .267*TC1	0.383
BP4 = -.245 + .589*AvgOP_Prev3_20 + .029*AvgOP_Prev12wks_Nwrk + .147*SP2 - .476*Slope + .261*TC2	0.383
BP5 = -.199 + .620*AvgOP_Prev3_20 + .044*AvgOP_Prev12wks_Nwrk + .060*SP3 - .511*Slope + .245*TC1	0.375
BP6 = -.199 + .619*AvgOP_Prev3_20 + .044*AvgOP_Prev12wks_Nwrk + .060*SP3 - .514*Slope + .239*TC2	0.375
BP7 = -.242 + .588*AvgOP_Prev3_20 + .029*AvgOP_Prev12wks_Nwrk + .147*SP4 - .470*Slope + .268*TC1	0.384
BP8 = -.242 + .588*AvgOP_Prev3_20 + .029*AvgOP_Prev12wks_Nwrk + .147*SP4 - .473*Slope + .262*TC2	0.384

The best prediction models, which produced the highest R^2 value (0.384), were obtained from variation numbers 7 and 8, which used the seasonality predictor SP4. Given that using either TC1 or TC2 yielded the same R^2 value, the decision to use TC1 was randomly taken. Hence, BP7 was selected in this case as the best predictor.

The results obtained from the best linear regression model are described in Tables 9.29 through 9.32. Both the stepwise forward and the backward forward methods were used, and they produced exactly the same results.

Table 9.29 Summary of Best Linear Regression Prediction Model for Low-Intensity Channels

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
	.619(d)	.384	.383	.715

Predictors: (Constant), TC1, SP4, AvgOP_Prev3_20, Slope, AvgOP_Prev12wks_Nwrk

Table 9.30 ANOVA for Best Linear Regression Prediction Model for Low-Intensity Channels

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	4805.516	5	961.103	1881.405	.000
Residual	7720.887	15114	.511		
Total	12526.403	15119			

Predictors: (Constant), TC1, SP4, AvgOP_Prev3_20, Slope, AvgOP_Prev12wks_Nwrk
Dependent Variable: ObservedPosters

Table 9.31 Coefficients for Best Linear Regression Prediction Model for Low-Intensity Channels

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-.242	.033		-7.414	.000
AvgOP_Prev3_20	.588	.010	.450	59.668	.000
AvgOP_Prev12wks_Nwrk	.029	.009	.021	3.234	.001
SP4	.147	.010	.108	15.388	.000
Slope	-.473	.021	-.184	-22.763	.000
TC1	.262	.029	.066	9.192	.000

Dependent Variable: ObservedPosters

Table 9.32 Variables Excluded from Best Linear Regression Model for Low-Intensity Channels

Model	Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
					Tolerance
AvgOP_PrevHr_Nwrk	.011	.858	.391	.007	.232
AvgOP_Prev3_20_Nwrk	.010	.990	.322	.008	.404
AvgOP_Prev3wks	.008	1.217	.224	.010	.995

Predictors in the Model: (Constant), TC1, SP4, AvgOP_Prev3_20, Slope, AvgOP_Prev12wks_Nwrk
Dependent Variable: ObservedPosters

Overall channel predictability was measured by the correlation coefficients between the best predictor BP7, produced by the best model, and the ObservedPosters dependent variable. Both Pearson and Spearman correlation coefficients were computed. The Pearson correlation coefficient computed between BP7 and ObservedPosters had the highest value among all the eight best predictors ($r = 0.613$). The Spearman correlation coefficient computed between BP7 and ObservedPosters was also higher than all of the coefficients computed between the other seven BP variables and the ObservedPosters ($rho = 0.600$). However, the differences between the Spearman correlation coefficient obtained from the best prediction model and the Spearman correlation coefficients

obtained from the other seven models were quite low, never larger than 0.020. Table 9.33 breaks down the results of the best linear regression prediction model and presents the Spearman correlation coefficients for each of the three types of channels corresponding to the subgroup of low-intensity channels. The tables that report the correlation coefficients between the observed posters and the rest of the computed best predictors can be found in the Appendix.

Table 9.33 Best Model Correlation. Coefficients for All Low-Intensity Channels, Grouped by Type

Type				BP7
1	Spearman's rho	ObservedPosters	Correlation Coefficient	.513(**)
			Sig. (2-tailed)	.000
			N	5040
4	Spearman's rho	ObservedPosters	Correlation Coefficient	.615(**)
			Sig. (2-tailed)	.000
			N	5040
7	Spearman's rho	ObservedPosters	Correlation Coefficient	.575(**)
			Sig. (2-tailed)	.000
			N	5040

Table 9.34 presents the regression equations for each BestPredictor variable and the corresponding R^2 values, computed for all the medium-intensity channels.

Table 9.34 Regression Equations and R^2 Values for Medium-Intensity Channels

Regression equation	R^2
BP1 = -.268 + .745*AvgOP_Prev3_20 + .215*SP1 - .253*Slope + .389*TC1	0.507
BP2 = -.268 + .743*AvgOP_Prev3_20 + .214*SP1 - .261*Slope + .376*TC2	0.507
BP3 = -.250 + .713*AvgOP_Prev3_20 + .253*SP2 - .267*Slope + .377*TC1	0.515
BP4 = -.250 + .711*AvgOP_Prev3_20 + .253*SP2 - .274*Slope + .364*TC2	0.515
BP5 = -.255 + .748*AvgOP_Prev3_20 + .206*SP3 - .254*Slope + .389*TC1	0.507
BP6 = -.254 + .746*AvgOP_Prev3_20 + .206*SP3 - .261*Slope + .375*TC2	0.507
BP7 = -.219 + .706*AvgOP_Prev3_20 + .226*SP4 - .276*Slope + .373*TC1	0.517
BP8 = -.220 + .704*AvgOP_Prev3_20 + .226*SP4 - .283*Slope + .359*TC2	0.517

** Correlation is significant at the 0.01 level (2-tailed).

The best prediction models, which produced the highest R^2 value (0.517), were obtained from variation numbers 7 and 8, which used the seasonality predictor SP4. Given that using either TC1 or TC2 yielded the same R^2 value, the decision to use TC1 was randomly taken. Hence, BP7 was selected in this case as the best predictor.

The results obtained from the best linear regression model are described in Tables 9.35 through 9.38. Both the stepwise forward and the backward forward methods were used, and they produced exactly the same results.

Table 9.35 Summary of Best Linear Regression Prediction Model for Medium-Intensity Channels

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
	.719	.517	.517	.517

Predictors: (Constant), TC1, AvgOP_Prev3_20, SP4, Slope

Table 9.36 ANOVA for Best Linear Regression Prediction Model for Medium-Intensity Channels

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	4316.329	4	1079.082	4043.833	.000
	Residual	4033.383	15115	.267		
	Total	8349.712	15119			

Predictors: (Constant), TC2, SP4, AvgOP_Prev3_20, Slope, AvgOP_Prev12wks_Nwrk
Dependent Variable: ObservedPosters

Table 9.37 Coefficients for Best Linear Regression Prediction Model for Medium-Intensity Channels

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-.219	.012		-18.940	.000
AvgOP_Prev3_20	.706	.009	.561	74.631	.000
SP4	.226	.010	.140	22.668	.000
Slope	-.276	.018	-.119	-15.279	.000
TC1	.373	.020	.122	18.597	.000

Dependent Variable: ObservedPosters

Table 9.38 Variables Excluded from Best Linear Regression Model for Medium-Intensity Channels

Model	Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
					Tolerance
AvgOP_Prev3wks	.000	.002	.998	.000	1.000
AvgOP_PrevHr_Nwrk	.007	1.242	.214	.010	.957
AvgOP_Prev3_20_Nwrk	.003	.448	.654	.004	.964
AvgOP_Prev12wks_Nwrk	.009	1.584	.113	.013	.940

Predictors in the Model: (Constant), TC1, AvgOP_Prev3_20, SP4, Slope
Dependent Variable: ObservedPosters

Overall channel predictability was measured by the correlation coefficients between the best predictor BP7, produced by the best model, and the ObservedPosters dependent variable. Both Pearson and Spearman correlation coefficients were computed. The Pearson correlation coefficient computed between BP7 and ObservedPosters had the highest value among all the eight best predictors (r = 0.706). The Spearman correlation coefficient computed between BP7 and ObservedPosters was slightly lower than some of the coefficients computed between the other seven BP variables and the ObservedPosters (rho = 0.569). However, the differences between the Spearman correlation coefficient obtained from the best prediction model and the Spearman correlation coefficients obtained from the other seven models were negligible, never larger than 0.008. Table 9.39 breaks down the results of the best linear regression prediction model and presents the Spearman correlation coefficients for each of the three types of channels corresponding to the subgroup of low-intensity channels. The tables that report the correlation coefficients between the observed posters and the rest of the computed best predictors can be found in the Appendix.

Table 9.39 Best Model Correlation. Coefficients for All Medium-Intensity Channels, Grouped by Type

Type				BP7
2	Spearman's rho	ObservedPosters	Correlation Coefficient	.580(**)
			Sig. (2-tailed)	.000
			N	5040
5	Spearman's rho	ObservedPosters	Correlation Coefficient	.613(**)
			Sig. (2-tailed)	.000
			N	5040
8	Spearman's rho	ObservedPosters	Correlation Coefficient	.501(**)
			Sig. (2-tailed)	.000
			N	5040

Table 9.40 presents the regression equations for each BestPredictor variable and the corresponding R^2 values, computed for all the high-intensity channels.

Table 9.40 Regression Equations and R^2 Values for High-Intensity Channels

Regression equation	R^2
BP1 = -.745 + .554*AvgOP_Prev3_20 - .074*AvgOP_PrevHr_Nwrk + .201*AvgOP_Prev3_20_Nwrk + .024*AvgOP_Prev3wks+ .362*SP1 - .644*Slope + .099*TC1	0.828
BP2 = -.746 + .553*AvgOP_Prev3_20 - .074*AvgOP_PrevHr_Nwrk + .201*AvgOP_Prev3_20_Nwrk + .024*AvgOP_Prev3wks+ .362*SP1 - .652*Slope + .080*TC2	0.828
BP3 = -.648 + .647*AvgOP_Prev3_20 - .095*AvgOP_PrevHr_Nwrk + .210*AvgOP_Prev3_20_Nwrk + .034*AvgOP_Prev3wks+ .313*SP2 - .538*Slope + .155*TC1	0.822
BP4 = -.648 + .647*AvgOP_Prev3_20 - .095*AvgOP_PrevHr_Nwrk + .210*AvgOP_Prev3_20_Nwrk + .034*AvgOP_Prev3wks+ .313*SP2 - .544*Slope + .143*TC2	0.822
BP5 = -.674 + .547*AvgOP_Prev3_20 - .085*AvgOP_PrevHr_Nwrk + .199*AvgOP_Prev3_20_Nwrk + .024*AvgOP_Prev3wks+ .351*SP3 - .672*Slope + .070*TC1	0.829
BP6 = -.677 + .545*AvgOP_Prev3_20 - .086*AvgOP_PrevHr_Nwrk + .200*AvgOP_Prev3_20_Nwrk + .024*AvgOP_Prev3wks+ .352*SP3 - .700*Slope	0.829
BP7 = -.625 + .662*AvgOP_Prev3_20 - .097*AvgOP_PrevHr_Nwrk + .212*AvgOP_Prev3_20_Nwrk + .035*AvgOP_Prev3wks+ .295*SP4 - .545*Slope + .152*TC1	0.821
BP8 = -.625 + .661*AvgOP_Prev3_20 - .097*AvgOP_PrevHr_Nwrk + .212*AvgOP_Prev3_20_Nwrk + .035*AvgOP_Prev3wks+ .295*SP4 - .552*Slope + .138*TC2	0.821

** Correlation is significant at the 0.01 level (2-tailed).

The best prediction models, which produced the highest R^2 value (0.829), were obtained from variation numbers 5 and 6, which used the seasonality predictor SP3. BP6 was selected in this case as the best predictor because its regression equation had a smaller number of independent variables than the regression equation for BP5.

The results obtained from the best linear regression model are described in Tables 9.41 through 9.44. Both the stepwise forward and the backward forward methods were used, and they produced exactly the same results.

Table 9.41 Summary of Best Linear Regression Prediction Model for High-Intensity Channels

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
	.910	.829	.829	1.382

Predictors: (Constant), AvgOP_Prev3wks, AvgOP_Prev3_20, Slope, AvgOP_Prev3_20_Nwrk, SP3, AvgOP_PrevHr_Nwrk

Table 9.42 ANOVA for Best Linear Regression Prediction Model for High-Intensity Channels

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	139869.928	6	23311.655	12206.356	.000
	Residual	28862.754	15113	1.910		
	Total	168732.681	15119			

Predictors: (Constant), AvgOP_Prev3wks, AvgOP_Prev3_20, Slope, AvgOP_Prev3_20_Nwrk, SP3, AvgOP_PrevHr_Nwrk
Dependent Variable: ObservedPosters

Table 9.43 Coefficients for Best Linear Regression Prediction Model for High-Intensity Channels

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-.677	.065		-10.365	.000
AvgOP_Prev3_20	.545	.008	.476	64.132	.000
AvgOP_PrevHr_Nwrk	-.086	.030	-.020	-2.818	.005
AvgOP_Prev3_20_Nwrk	.200	.042	.034	4.759	.000
AvgOP_Prev3wks	.024	.003	.031	7.036	.000
SP3	.352	.007	.383	50.338	.000
Slope	-.700	.020	-.125	-34.272	.000

Dependent Variable: ObservedPosters

Table 9.44 Variables Excluded from Best Linear Regression Model for High-Intensity Channels

Model	Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
AvgOP_Prev12wks_Nwrk	.007	.994	.320	.008	.229
TC2	.005	1.280	.200	.010	.623

Predictors in the Model: (Constant), AvgOP_Prev3wks, AvgOP_Prev3_20, Slope, AvgOP_Prev3_20_Nwrk, SP3, AvgOP_PrevHr_Nwrk
Dependent Variable: ObservedPosters

Overall channel predictability was measured by the correlation coefficients between the best predictor BP6 produced by the best model and the ObservedPosters dependent variable. Both Pearson and Spearman correlation coefficients were computed. The Pearson correlation coefficient computed between BP6 and ObservedPosters had the highest value among all the eight best predictors (r = 0.908). The Spearman correlation coefficient computed between BP6 and ObservedPosters was slightly lower than some of the coefficients computed between the other seven BP variables and the ObservedPosters (rho = 0.718). However, the differences between the Spearman correlation coefficient obtained from the best prediction model and the Spearman correlation coefficients obtained from the other seven models were quite low, never larger than 0.022. Table 9.45 breaks down the results of the best linear regression prediction model and presents the Spearman correlation coefficients for each of the three types of channels corresponding to the subgroup of low-intensity channels. The tables that report the correlation coefficients between the observed posters and the rest of the computed best predictors can be found in the Appendix.

Table 9.45 Best Model Correlation. Coefficients for All High-Intensity Channels, Grouped by Type

Type				BP6
3	Spearman's rho	ObservedPosters	Correlation Coefficient	.382(**)
			Sig. (2-tailed)	.000
			N	5040
6	Spearman's rho	ObservedPosters	Correlation Coefficient	.370(**)
			Sig. (2-tailed)	.000
			N	5040
9	Spearman's rho	ObservedPosters	Correlation Coefficient	.833(**)
			Sig. (2-tailed)	.000
			N	5040

Table 9.46 summarizes the correlations obtained from the best linear regression model for each of the nine categories of channels in the sample. For some types of channels the predictions produced by the best model for all the channels were better than the predictions produced by the best models for subgroups of channels (types 3, 8, and 9); while for other types of channels the reverse was true (types 1, 2, 4, 5, 6, and 7). Overall, the easiest channels to predict were the large, highly intensive channels.

Table 9.46 Summary of Best Model Correlations Obtained for Each Type of Channel

Channel type	Correlations computed for all channels	Correlations computed for sub groups of channels based on size	Correlations computed for sub-groups of channels based on intensity
1	0.492	0.536	0.513
2	0.570	0.580	0.580
3	0.427	0.407	0.382
4	0.607	0.575	0.615
5	0.589	0.600	0.613
6	0.370	0.545	0.370
7	0.572	0.531	0.575
8	0.520	0.465	0.501
9	0.838	0.833	0.833

** Correlation is significant at the 0.01 level (2-tailed).

9.3.1.2 Second Regression Model. As mentioned in subsection 9.2.3, four of the independent variables used in the regression analysis had at least one negative value and/or one zero value (AvgOP_Prev3_20, AvgOP_Prev3wks, Slope, and both instances of TC), while the dependent variable (ObservedPosters) also had a significant number of zero values. To understand whether these non-positive values would have an impact on the results of the regression analysis; transformations were performed on these variables, making all of them positive before entering them into the regression model. The end result showed that the transformations did not have any impact on the resulting correlations between the computed best predictors and the actual observed posters. The exact same operations from the first regression model were performed on the new set of independent variables, and the resulting correlation coefficients were exactly the same for each type of channel and for each sub-group of channels.

9.3.2 Nonlinear Regression

Nonlinear regression attempts to find a nonlinear model of the relationship between the dependent variable and a set of independent variables. Unlike traditional linear regression, which is restricted to estimating linear models, nonlinear regression can estimate models with arbitrary relationships between independent and dependent variables. The SPSS software provides 11 curve estimation regression models: Linear, Logarithmic, Inverse, Quadratic, Cubic, Power, Compound, S-curve, Logistic, Growth, and Exponential. Of these, the Inverse and S models cannot be calculated if the independent variables contain values of zero, while the Logarithmic and Power models cannot be calculated if the independent variables contain negative values. Also, the Compound, Power, S, Growth, Exponential, Logarithmic, and Logistic models cannot be calculated when the dependent variable contains non-positive values. Considering these restrictions, the nonlinear regression analysis was conducted using the transformed set of variables described in subsection 9.2.3. To determine the best fit between the set of independent variables and the dependent variable, curve estimation procedures were run for each of the eight independent variables; then regression equations were created combining the best fit models produced for each independent variable. Tables 9.47 through 9.58 present the curve fit models computed for each independent variable for all the channels in the sample; while Table 9.59 summarizes the best fit model for each independent variable.

Table 9.47 Curve Fit Models for AvgOP_Prev3_20Mod

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.743	130914.322	1	45358	.000	.049	.995		
Logarithmic	.634	78604.209	1	45358	.000	.909	3.723		
Inverse	.451	37243.451	1	45358	.000	7.555	-6.634		
Quadratic	.741	67321.730	2	45357	.000	-.173	1.203	-.013	
Cubic	.736	45537.176	3	45356	.000	.068	.943	.024	-.001
Compound	.619	73574.514	1	45358	.000	.858	1.231		
Power	.700	105758.826	1	45358	.000	1.009	.894		
S	.604	69298.748	1	45358	.000	1.754	-1.755		
Growth	.619	73574.514	1	45358	.000	-.153	.208		
Exponential	.619	73574.514	1	45358	.000	.858	.208		
Logistic	.619	73574.514	1	45358	.000	1.166	.813		

Dependent Variable: ObservedPostersMod

Table 9.48 Curve Fit Models for AvgOP_PrevHr_Nwrk

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.013	609.950	1	45358	.000	.238	.304		
Logarithmic	.013	590.318	1	45358	.000	-.091	1.129		
Inverse	.012	554.057	1	45358	.000	2.483	-3.954		
Quadratic	.013	307.065	2	45357	.000	.675	.071	.030	
Cubic	.013	307.152	2	45357	.000	.756	.000	.050	-.002
Compound	.016	715.075	1	45358	.000	.852	1.078		
Power	.015	705.516	1	45358	.000	.783	.282		
S	.015	675.373	1	45358	.000	.399	-.996		
Growth	.016	715.075	1	45358	.000	-.160	.075		
Exponential	.016	715.075	1	45358	.000	.852	.075		
Logistic	.016	715.075	1	45358	.000	1.174	.928		

Dependent Variable: ObservedPostersMod

Table 9.49 Curve Fit Models for AvgOP_Prev3_20_Nwrk

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.011	506.706	1	45358	.000	.202	.384		
Logarithmic	.011	494.805	1	45358	.000	.088	1.170		
Inverse	.010	467.763	1	45358	.000	2.519	-3.366		
Quadratic	.011	253.665	2	45357	.000	.400	.255	.020	
Cubic	.011	253.665	2	45357	.000	.400	.255	.020	.000
Compound	.012	555.455	1	45358	.000	.853	1.096		
Power	.012	557.161	1	45358	.000	.826	.283		
S	.012	541.079	1	45358	.000	.402	-.826		
Growth	.012	555.455	1	45358	.000	-.159	.092		
Exponential	.012	555.455	1	45358	.000	.853	.092		
Logistic	.012	555.455	1	45358	.000	1.173	.912		

Dependent Variable: ObservedPostersMod

Table 9.50 Curve Fit Models for AvgOP_Prev3wksMod

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.153	8217.441	1	45358	.000	1.026	.219		
Logarithmic	.246	14791.385	1	45358	.000	1.083	1.654		
Inverse	.232	13712.110	1	45358	.000	5.120	-4.106		
Quadratic	.324	10798.556	2	45357	.000	.160	.963	-.031	
Cubic	.324	7231.629	3	45356	.000	.013	1.112	-.046	.000
Compound	.114	5822.579	1	45358	.000	1.057	1.044		
Power	.225	13152.437	1	45358	.000	1.061	.361		
S	.251	15230.598	1	45358	.000	1.012	-.976		
Growth	.114	5822.579	1	45358	.000	.055	.043		
Exponential	.114	5822.579	1	45358	.000	1.057	.043		
Logistic	.114	5822.579	1	45358	.000	.946	.958		

Dependent Variable: ObservedPostersMod

Table 9.51 Curve Fit Models for AvgOP_Prev12wks_Nwrk

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	B1	b2	b3
Linear	.017	794.078	1	45358	.000	-.028	.404		
Logarithmic	.017	767.924	1	45358	.000	-.333	1.395		
Inverse	.016	729.852	1	45358	.000	2.760	-4.618		
Quadratic	.017	403.714	2	45357	.000	.957	-.162	.079	
Cubic	.017	403.714	2	45357	.000	.957	-.162	.079	.000
Compound	.022	999.131	1	45358	.000	.788	1.109		
Power	.021	984.706	1	45358	.000	.726	.360		
S	.021	954.027	1	45358	.000	.481	-1.203		
Growth	.022	999.131	1	45358	.000	-.239	.103		
Exponential	.022	999.131	1	45358	.000	.788	.103		
Logistic	.022	999.131	1	45358	.000	1.269	.902		

Dependent Variable: ObservedPostersMod

Table 9.52 Curve Fit Models for SlopeMod

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.191	10738.049	1	45358	.000	32.467	-2.050		
Logarithmic	.180	9982.651	1	45358	.000	73.597	-26.563		
Inverse	.046	2185.698	1	45358	.000	-4.598	91.281		
Quadratic	.212	6099.003	2	45357	.000	80.971	-8.876	.239	
Cubic	.228	4475.229	3	45356	.000	33.232	2.077	-.579	.020
Compound	.266	16452.211	1	45358	.000	4891.690	.576		
Power	.246	14763.515	1	45358	.000	258454684.970	-7.080		
S	.063	3042.303	1	45358	.000	-1.476	24.384		
Growth	.266	16452.211	1	45358	.000	8.495	-.552		
Exponential	.266	16452.211	1	45358	.000	4891.690	-.552		
Logistic	.266	16452.211	1	45358	.000	.000	1.737		

Dependent Variable: ObservedPostersMod

Table 9.53 Curve Fit Models for TC1Mod

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.009	413.448	1	45358	.000	.032	.705		
Logarithmic	.000	3.663	1	45358	.056	1.359	.121		
Inverse	.004	190.729	1	45358	.000	.715	1.417		
Quadratic	.170	4624.070	2	45357	.000	13.966	-13.543	3.570	
Cubic	.170	4633.955	2	45357	.000	10.130	-6.866	.000	.595
Compound	.014	621.267	1	45358	.000	.774	1.218		
Power	.000	1.062	1	45358	.303	1.135	.015		
S	.008	373.823	1	45358	.000	-.095	.452		
Growth	.014	621.267	1	45358	.000	-.257	.197		
Exponential	.014	621.267	1	45358	.000	.774	.197		
Logistic	.014	621.267	1	45358	.000	1.293	.821		

Dependent Variable: ObservedPostersMod

Table 9.54 Curve Fit Models for TC2Mod

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.009	399.532	1	45358	.000	.052	.695		
Logarithmic	.000	3.161	1	45358	.075	1.364	.113		
Inverse	.004	188.633	1	45358	.000	.720	1.407		
Quadratic	.164	4448.516	2	45357	.000	13.744	-13.297	3.504	
Cubic	.164	4463.847	2	45357	.000	9.989	-6.750	.000	.585
Compound	.013	612.636	1	45358	.000	.775	1.217		
Power	.000	1.179	1	45358	.278	1.135	.016		
S	.008	362.908	1	45358	.000	-.091	.445		
Growth	.013	612.636	1	45358	.000	-.255	.196		
Exponential	.013	612.636	1	45358	.000	.775	.196		
Logistic	.013	612.636	1	45358	.000	1.291	.822		

Dependent Variable: ObservedPostersMod

Table 9.55 Curve Fit Models for SP1

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.683	97773.338	1	45358	.000	.286	.827		
Logarithmic	.561	57853.374	1	45358	.000	.992	3.529		
Inverse	.303	19699.604	1	45358	.000	6.206	-5.096		
Quadratic	.682	49511.916	2	45357	.000	.102	.999	-.009	
Cubic	.681	33007.272	3	45356	.000	.098	1.003	-.010	1.58E-005
Compound	.491	43683.850	1	45358	.000	.917	1.174		
Power	.504	46040.134	1	45358	.000	1.041	.764		
S	.327	22022.650	1	45358	.000	1.268	-1.210		
Growth	.491	43683.850	1	45358	.000	-.087	.160		
Exponential	.491	43683.850	1	45358	.000	.917	.160		
Logistic	.491	43683.850	1	45358	.000	1.090	.852		

Dependent Variable: ObservedPostersMod

Table 9.56 Curve Fit Models for SP2

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.654	85866.386	1	45358	.000	.188	.879		
Logarithmic	.533	51773.686	1	45358	.000	.908	3.390		
Inverse	.322	21547.498	1	45358	.000	6.294	-5.331		
Quadratic	.650	43667.018	2	45357	.000	-.021	1.071	-.011	
Cubic	.652	29568.456	3	45356	.000	.280	.749	.035	-.002
Compound	.507	46680.853	1	45358	.000	.891	1.193		
Power	.521	49404.021	1	45358	.000	1.017	.766		
S	.380	27756.031	1	45358	.000	1.341	-1.322		
Growth	.507	46680.853	1	45358	.000	-.115	.177		
Exponential	.507	46680.853	1	45358	.000	.891	.177		
Logistic	.507	46680.853	1	45358	.000	1.122	.838		

Dependent Variable: ObservedPostersMod

Table 9.57 Curve Fit Models for SP3

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.684	97985.000	1	45358	.000	.334	.793		
Logarithmic	.561	58013.260	1	45358	.000	1.027	3.495		
Inverse	.280	17638.016	1	45358	.000	5.865	-4.684		
Quadratic	.684	50020.933	2	45357	.000	.107	1.006	-.011	
Cubic	.684	33365.477	3	45356	.000	.047	1.070	-.019	.000
Compound	.483	42389.253	1	45358	.000	.927	1.165		
Power	.497	44850.307	1	45358	.000	1.049	.752		
S	.298	19280.329	1	45358	.000	1.180	-1.104		
Growth	.483	42389.253	1	45358	.000	-.076	.152		
Exponential	.483	42389.253	1	45358	.000	.927	.152		
Logistic	.483	42389.253	1	45358	.000	1.079	.859		

Dependent Variable: ObservedPostersMod

Table 9.58 Curve Fit Models for SP4

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.644	82048.490	1	45358	.000	.214	.862		
Logarithmic	.530	51072.900	1	45358	.000	.926	3.371		
Inverse	.322	21541.632	1	45358	.000	6.325	-5.342		
Quadratic	.640	42120.904	2	45357	.000	-.039	1.095	-.013	
Cubic	.643	28429.452	3	45356	.000	.215	.826	.024	-.001
Compound	.501	45625.328	1	45358	.000	.896	1.190		
Power	.517	48558.924	1	45358	.000	1.021	.761		
S	.377	27401.854	1	45358	.000	1.344	-1.320		
Growth	.501	45625.328	1	45358	.000	-.110	.174		
Exponential	.501	45625.328	1	45358	.000	.896	.174		
Logistic	.501	45625.328	1	45358	.000	1.117	.840		

Dependent Variable: ObservedPostersMod

Table 9.59 Best Models and Corresponding R^2 Values for All Channels

Independent variable	Best model	R^2
AvgOP_Prev3_20Mod	Linear	.743
AvgOP_PrevHr_Nwrk	Compound	.016
AvgOP_Prev3_20_Nwrk	Compound	.012
AvgOP_Prev3wksMod	Quadratic	.323
AvgOP_Prev12wks_Nwrk	Compound	.022
SlopeMod	Compound	.266
TC1Mod	Quadratic	.169
TC2Mod	Quadratic	.164
SP1	Linear	.683
SP2	Linear	.654
SP3	Linear	.684
SP4	Linear	.644

Based on the best models reported in Table 9.59, a nonlinear regression equation was created for each independent variable. As in the linear regression analysis, to avoid multicollinearity problems only one of the four possible instances of the seasonality predictor variable and only one of the two possible instances of the trajectory coefficient variable were entered into the regression models at a time, together with the other six independent variables (AvgOP_Prev3_20Mod, AvgOP_PrevHr_Nwrk, AvgOP_Prev3_20_Nwrk, AvgOP_Prev3wksMod, AvgOP_Prev12wks_Nwrk, and SlopeMod). The best nonlinear regression prediction model included seasonality predictor SP3 and trajectory coefficient TC1, as their best fit models had the highest R^2 values. The results of this best nonlinear regression prediction model are presented in Tables 9.60 and 9.61.

Table 9.60 Parameter Estimates for Best Nonlinear Prediction Model

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
a	.606	.005	.596	.616
b	-.023	.013	-.049	.003
c	.049	.014	.022	.075
d	.047	.005	.037	.057
e	-.001	.000	-.002	-.001
f	.017	.011	-.005	.038
g	-3.407	.023	-3.451	-3.362
h	1.611	.102	1.411	1.812
i	-.368	.026	-.419	-.317
j	.318	.004	.310	.327

The Parameter Estimates table summarizes the model-estimated value of each parameter. The small standard errors with respect to the value of the estimates suggest that one can be confident in the computed estimates.

Table 9.61 ANOVA for Best Nonlinear Prediction Model

Source	Sum of Squares	df	Mean Squares
Regression	250065.764	10	25006.576
Residual	41221.236	45350	.909
Uncorrected Total	291287.000	45360	
Corrected Total	197071.190	45359	

Dependent variable: ObservedPostersMod
R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .791.

The ANOVA Table provides a breakdown of the sum of squares, a measure of variability in the dependent variable, for this model. The Regression row displays information about the variation accounted for by the model, while the Residual row displays information about the variation that is not accounted for by the model. The Uncorrected Total represents the entire variability in the dependent variable, while the Corrected Total is adjusted to reflect only the variability about average values of the dependent variable. The Residual Sum of Squares and Corrected Total are used to compute R^2. The obtained R^2 value (0.791) shows that the model accounts for approximately 79.1 percent of the variability in the dependent variable.

Overall channel predictability was measured by the correlation coefficients between the best predictor (BP) produced by the model and the ObservedPosters dependent variable. Both Pearson and Spearman correlation coefficients were computed. The Pearson correlation coefficient computed between BP and ObservedPosters was r = 0.886, while the Spearman correlation coefficient computed between BP and ObservedPosters was rho = 0.667.

Table 9.62 breaks down the results of the best nonlinear regression prediction model and presents the Spearman correlation coefficients for each of the nine types of channels.

Table 9.62 Best Model Correlation Coefficients for All Channels Grouped by Type

Type				BP
1	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.513(**)
			Sig. (2-tailed)	.000
			N	5040
2	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.575(**)
			Sig. (2-tailed)	.000
			N	5040
3	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.427(**)
			Sig. (2-tailed)	.000
			N	5040
4	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.619(**)
			Sig. (2-tailed)	.000
			N	5040
5	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.596(**)
			Sig. (2-tailed)	.000
			N	5040
6	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.368(**)
			Sig. (2-tailed)	.000
			N	5040
7	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.553(**)
			Sig. (2-tailed)	.000
			N	5040
8	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.514(**)
			Sig. (2-tailed)	.000
			N	5040
9	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.839(**)
			Sig. (2-tailed)	.000
			N	5040

** Correlation is significant at the 0.01 level (2-tailed).

Table 9.63 breaks down the results based on the size of the channels in the sample, while Table 9.64 breaks down the results based on the intensity of the channels in the sample.

Table 9.63 Best Model Correlation Coefficients for All Channels Grouped by Size

Size				BP
large	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.757(**)
			Sig. (2-tailed)	.000
			N	15120
medium	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.522(**)
			Sig. (2-tailed)	.000
			N	15120
small	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.521(**)
			Sig. (2-tailed)	.000
			N	15120

Table 9.64 Best Model Correlation Coefficients for All Channels Grouped by Intensity

Intensity				BP
high	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.737(**)
			Sig. (2-tailed)	.000
			N	15120
low	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.589(**)
			Sig. (2-tailed)	.000
			N	15120
medium	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.565(**)
			Sig. (2-tailed)	.000
			N	15120

The values for different types of channels ranged from a minimum of 0.368 (type 6) to a maximum of 0.839 (type 9). For subgroups of channels, correlation coefficients were higher for large channels compared to medium and small channels; and higher for high-intensity channels compared to medium- and low-intensity channels.

For comparison purposes, seven other predictor variables were computed using all the possible combinations of independent variables, as was done in the linear regression analysis described in the previous section. The tables that report the correlation coefficients between the observed posters and the rest of the computed predictors can be found in the Appendix.

** Correlation is significant at the 0.01 level (2-tailed).

Table 9.59 revealed that the best curve fit models for three of the independent variables had very low values, smaller than 0.1. Consequently, those independent variables were eliminated from the regression model and a new BestPredictor variable was computed. The model used the following independent variables: AvgOP_Prev3_20Mod, AvgOP_Prev3wksMod, SlopeMod, SP3, and TC1Mod. The results of this best nonlinear regression prediction model are presented in Tables 9.65 and 9.66.

Table 9.65 Parameter Estimates for the Reduced Best Nonlinear Prediction Model

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
a	.607	.005	.597	.617
b	.049	.005	.039	.059
c	-.001	.000	-.002	-.001
d	-3.368	.022	-3.410	-3.326
e	1.560	.102	1.360	1.760
f	-.354	.026	-.404	-.303
g	.318	.004	.310	.326

The Parameter Estimates Table summarizes the model-estimated value of each parameter. The small standard errors with respect to the value of the estimates suggest that one can be confident in the computed estimates.

Table 9.66 ANOVA for Reduced Best Nonlinear Prediction Model

Source	Sum of Squares	df	Mean Squares
Regression	250032.819	10	25003.282
Residual	41254.181	45350	.910
Uncorrected Total	291287.000	45360	
Corrected Total	197071.190	45359	

Dependent variable: ObservedPostersMod

a R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .791.

The obtained R^2 value (0.791) shows that the model accounts for approximately 79.1 percent of the variability in the dependent variable. It may be observed that the R^2 value produced by the reduced model is the same as the one produced by the full model.

Overall channel predictability was measured by the correlation coefficients between the best predictor produced by the model and the ObservedPosters dependent variable. Both Pearson and Spearman correlation coefficients were computed. The Pearson correlation coefficient computed between BP and ObservedPosters was r = 0.886, while the Spearman correlation coefficient computed between BP and ObservedPosters was rho = 0.660.

Table 9.67 breaks down the results of the best nonlinear regression prediction model and presents the Spearman correlation coefficients for each of the nine types of channels.

Table 9.67 Reduced Best Model Correlation Coefficients for All Channels Grouped by Type

Type				BP
1	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.524(**)
			Sig. (2-tailed)	.000
			N	5040
2	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.575(**)
			Sig. (2-tailed)	.000
			N	5040
3	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.430(**)
			Sig. (2-tailed)	.000
			N	5040
4	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.615(**)
			Sig. (2-tailed)	.000
			N	5040
5	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.597(**)
			Sig. (2-tailed)	.000
			N	5040
6	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.334(**)
			Sig. (2-tailed)	.000
			N	5040
7	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.551(**)
			Sig. (2-tailed)	.000
			N	5040
8	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.517(**)
			Sig. (2-tailed)	.000
			N	5040
9	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.839(**)
			Sig. (2-tailed)	.000
			N	5040

** Correlation is significant at the 0.01 level (2-tailed).

Table 9.68 breaks down the results based on the size of the channels in the sample, while Table 9.69 breaks down the results based on the intensity of the channels in the sample.

Table 9.68 Reduced Best Model Correlation Coefficients for All Channels Grouped by Size

Size				BP
large	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.758(**)
			Sig. (2-tailed)	.000
			N	15120
medium	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.501(**)
			Sig. (2-tailed)	.000
			N	15120
small	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.523(**)
			Sig. (2-tailed)	.000
			N	15120

Table 9.69 Reduced Best Model Correlation Coefficients for All Channels Grouped by Intensity

Intensity				BP
high	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.725(**)
			Sig. (2-tailed)	.000
			N	15120
low	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.587(**)
			Sig. (2-tailed)	.000
			N	15120
medium	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.567(**)
			Sig. (2-tailed)	.000
			N	15120

The values for different types of channels ranged from a minimum of 0.334 (type 6) to a maximum of 0.839 (type 9). For subgroups of channels, correlation coefficients were higher for large channels compared to medium and small channels; and higher for high-intensity channels compared to medium- and low-intensity channels.

For comparison purposes, seven other predictor variables were computed using all the possible combinations of independent variables, as was done in the linear regression analysis described in the previous section. The tables that report the correlation coefficients between the observed posters and the rest of the computed predictors can be found in the Appendix.

** Correlation is significant at the 0.01 level (2-tailed).

The last nonlinear regression model that was built used an even smaller set of independent variables including the following:

- AvgOP_Prev3_20Mod - the average of the observed number of posters during the previous three 20-minute time intervals for each channel in the sample;
- SlopeMod - the slope of the line determined by the observed values for the previous three 20-minute time intervals for each channel; it is a basic indicator of the amount by which the number of posters varied during the previous hour;
- TC1Mod - a correlation coefficient between "time" and the observed number of posters during the last hour, which gave a general idea about the direction of the conversation (up, down, or constant).

These particular variables were considered because all of them were representative for the channel activity that occurred during the hour preceding the interval for which the predictions were made. Making predictions based on the most recent activity makes sense in the context of a system such as an IRC network, whose scale and dynamicity make long-term data-collection, as well as the associated processing of the data, problematic. The results of this minimal nonlinear regression prediction model are presented in Tables 9.70 and 9.71.

Table 9.70 Parameter Estimates for the Minimal Best Nonlinear Prediction Model

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
a	.968	.003	.963	.974
b	-3.504	.023	-3.548	-3.459
c	2.717	.108	2.505	2.928
d	-.639	.027	-.693	-.586

The Parameter Estimates Table summarizes the model-estimated value of each parameter. The small standard errors with respect to the value of the estimates suggest that one can be confident in the computed estimates.

Table 9.71 ANOVA for Minimal Best Nonlinear Prediction Model

Source	Sum of Squares	df	Mean Squares
Regression	243649.737	4	60912.434
Residual	47637.263	45356	1.050
Uncorrected Total	291287.000	45360	
Corrected Total	197071.190	45359	

Dependent variable: ObservedPostersMod

a R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .758.

The obtained R^2 value (0.758) shows that the model accounts for approximately 75.8 percent of the variability in the dependent variable. It may be observed that the R^2 value produced by the reduced model is very close to the ones produced by the full model and the reduced model. This suggests that the independent variables that describe the recent activity of chat-channels are the most important predictors for future short-term activity.

Overall channel predictability was measured by the correlation coefficients between the best predictor produced by the model and the ObservedPosters dependent variable. Both Pearson and Spearman correlation coefficients were computed. The Pearson correlation coefficient computed between BP and ObservedPosters was r = 0.869, while the Spearman correlation coefficient computed between BP and ObservedPosters was rho = 0.694. Table 9.72 breaks down the results of the best nonlinear regression prediction model and presents the Spearman correlation coefficients for each of the nine types of channels.

Table 9.72 Minimal Best Model Correlation Coefficients for All Channels Grouped by Type

Type				BP
1	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.484(**)
			Sig. (2-tailed)	.000
			N	5040
2	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.582(**)
			Sig. (2-tailed)	.000
			N	5040
3	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.457(**)
			Sig. (2-tailed)	.000
			N	5040
4	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.597(**)
			Sig. (2-tailed)	.000
			N	5040
5	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.609(**)
			Sig. (2-tailed)	.000
			N	5040
6	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.554(**)
			Sig. (2-tailed)	.000
			N	5040
7	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.572(**)
			Sig. (2-tailed)	.000
			N	5040
8	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.505(**)
			Sig. (2-tailed)	.000
			N	5040
9	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.842(**)
			Sig. (2-tailed)	.000
			N	5040

Table 9.73 breaks down the results based on the size of the channels in the sample, while Table 9.74 breaks down the results based on the intensity of the channels in the sample.

Table 9.73 Minimal Best Model Correlation Coefficients for All Channels Grouped by Size

Size				BP
large	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.762(**)
			Sig. (2-tailed)	.000
			N	15120
medium	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.590(**)
			Sig. (2-tailed)	.000
			N	15120
small	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.534(**)
			Sig. (2-tailed)	.000
			N	15120

** Correlation is significant at the 0.01 level (2-tailed).

Table 9.74 Minimal Best Model Correlation Coefficients for All Channels Grouped by Intensity

Intensity				BP
high	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.783(**)
			Sig. (2-tailed)	.000
			N	15120
low	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.589(**)
			Sig. (2-tailed)	.000
			N	15120
medium	Spearman's rho	ObservedPostersMod	Correlation Coefficient	.569(**)
			Sig. (2-tailed)	.000
			N	15120

The values for different types of channels ranged from a minimum of 0.457 (type 3) to a maximum of 0.842 (type 9). For subgroups of channels, correlation coefficients were higher for large channels compared to medium and small channels; and higher for high-intensity channels compared to medium- and low-intensity channels.

Tables 9.75, 9.76 and 9.77 summarize the best predictions produced by the nonlinear regression models.

** Correlation is significant at the 0.01 level (2-tailed).

Table 9.75 Best Overall Predictions for All Channels

Overall correlation coefficients obtained from the best nonlinear regression models created for all channels		
All predictors	Reduced set of predictors	Minimal set of predictors
0.667	0.660	.694

Table 9.76 Nonlinear Regression Best Predictions Summary by Channel Type

	Best correlation coefficients obtained from nonlinear regression equations created for all channels		
Channel Type	All predictors	Reduced set of predictors	Minimal set of predictors
1	0.513	0.524	0.484
2	0.575	0.575	0.582
3	0.427	0.430	0.457
4	0.619	0.615	0.597
5	0.596	0.597	0.609
6	0.368	0.334	0.554
7	0.553	0.551	0.572
8	0.514	0.517	0.505
9	0.839	0.839	0.842

Table 9.77 Nonlinear Regression Best Predictions Summary by Channel Subgroup

	Best correlation coefficients obtained from nonlinear regression equations created for all channels		
Channel Size	**All predictors**	**Reduced set of predictors**	**Minimal set of predictors**
Large	0.757	0.758	0.762
Medium	0.522	0.501	0.590
Small	0.521	0.523	0.534
Channel Intensity			
High	0.737	0.725	0.783
Low	0.589	0.587	0.589
Medium	0.565	0.567	0.569

9.4 Summary

This chapter presented the results of several linear and nonlinear regression models, which were conducted to assess whether it is possible to make short-term predictions about the activity of IRC chat-channels.

Both the linear and the nonlinear regression models used various combinations of the independent variables to produce a best predictor variable. The accuracy of the predictions was measured using Spearman correlation coefficients between the best predictor computed by the regression models and the actual values of observed posters for each channel in the sample.

For the best linear regression model the overall correlation coefficient between the best predictor and the observed posters was 0.662. The results of the best linear regression model are shown again in Table 9.78.

Table 9.78 Coefficients for Best Linear Regression Prediction Model for All Channels

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-.458	.026		-17.494	.000
AvgOP_Prev3_20	.592	.005	.512	117.267	.000
AvgOP_Prev3_20_Nwrk	.039	.008	.011	4.943	.000
AvgOP_Prev3wks	.010	.001	.018	7.621	.000
SP3	.331	.004	.345	81.676	.000
Slope	-.563	.013	-.120	-43.141	.000
TC2	.187	.019	.025	9.835	.000

For the best nonlinear regression model the overall correlation coefficient between the best predictor and the observed posters was 0.694. Also, it must be noted that the best nonlinear regression model used a minimal set of predictors. The best predictor in this case was computed using the following equation:

$$BP = 14.015 + (a * AvgOP\_Prev3\_20Mod) + (b * SlopeMod^{0.576}) +$$

$$+ (c * TC1 + d * TC1^2).$$

Table 9.79 reports the parameter estimates for the best nonlinear regression model described in the equation above.

Table 9.79 Parameter Estimates for the Minimal Best Nonlinear Prediction Model

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
a	.968	.003	.963	.974
b	-3.504	.023	-3.548	-3.459
c	2.717	.108	2.505	2.928
d	-.639	.027	-.693	-.586

The best nonlinear regression model only included predictors related to the previous hour of channel activity. This suggests that in a very dynamic synchronous medium such as IRC, accurate predictions about future activity may be made by only taking into account information pertinent to the channels' activity during the most recent hour rather than looking at historic data over longer periods of time.

CHAPTER 10

IDENTIFICATION OF FACTORS THAT INFLUENCE CHANNEL PREDICTABILITY

This chapter uses logistic regression to identify the factors that can distinguish highly predictable channels from unpredictable channels.

The linear and the nonlinear regression analyses described in the previous chapter revealed that the activity of large, highly intensive channels was the easiest to predict; while the rest of the channel categories had very similar overall degrees of predictability. This suggests that although the categorization of channels based on the total number of users (computed from the beginning of the data-collection period); and on the intensity of channel activity measured by the total number of days a channel was visited (also computed from the beginning of the data-collection period) was suited for selecting a random sample of channels; a channel's predictability level cannot be determined simply by observing the category to which that channel belongs.

Also, categorizing channels in the above manner may present scalability problems when addressing the entire IRC network versus a small subset of channels; and therefore may not be suited for larger-scale predictions. Thus, there is a need to identify other characteristics of chat-channels that would improve the short-term predictions, while also reducing the costs involved in making them.

10.1 Hypothesis

In light of this discussion, it is hypothesized that the level of predictability of a publicly active channel for any particular week can be estimated as high, low, or perfect by using various descriptive statistics of that channel computed for the one-month period preceding the week for which predictions are attempted.

10.2 Method

10.2.1 Data Considerations

The analysis was performed on the same stratified random sample of channels that was used in the previous chapter. The selection process was described in detail in Chapter 8, subsection 8.2.1. The descriptive statistics used as predictors for a channel's predictability level were described in Chapter 7, subsection 7.2.3.2.

10.2.2 Data Analysis

There are two methods that are suitable for this type of research: discriminant analysis and logistic regression.

Discriminant analysis is useful for situations where a predictive model of group membership needs to be built, based on the observed characteristics of each case. The procedure generates discriminant functions based on linear combinations of the predictor variables that provide the best discriminations between the groups.

Logistic regression is a model used for predicting the probability of an event's occurrence. It is useful for situations where there is a need to predict the presence or absence of a characteristic or outcome, based on values of a set of predictor variables. It is similar to a linear regression model, but it is suited to models where the dependent

variable is dichotomous. According to the documentation of the SPSS statistical software package, logistic regression is applicable to a broader range of research situations than discriminant analysis. Some authors prefer logistic regression to discriminant analysis because they consider it more flexible in its assumptions and in the types of data that can be analyzed. Also, as opposed to discriminant analysis, logistic regression can handle both categorical and continuous variables; and the predictors do not have to be normally distributed, linearly related, or of equal variance within each group (Tabachnick and Fidell 1996).

Considering the above, logistic regression was used to further explore the characteristics that separate channels into three main categories: channels with high predictability, channels with low predictability, and channels with perfect predictability.

Table 10.1 Predictor Variables for the Logistic Regression Model

Variable	Description	Acronym
SurvivalTime	Total number of days the channel existed in August 2005	ST
AvgUserRetTime	Average number of minutes between two user sessions	AURT
AvgUsers	Average number of users per any 20-minute interval	AU
AvgDailyUsers	Average number of users per day	ADU
AvgPosters	Average number of posters per any 20-minute interval	AP
AvgDailyPosters	Average number of posters per day	ADP
AvgMessages	Average number of messages per any 20-minute interval	AM
AvgDailyMessages	Average number of messages per day	ADM
AvgUserDiv	Average user diversity computed for any 20-minute interval with respect to the day	AUD
AvgDailyUserDiv	Average user diversity computed for any day, with respect to the month of August 2005	ADUD
AvgPosterDiv	Average poster diversity computed for any 20-minute interval with respect to the day	APD
AvgDailyPosterDiv	Average poster diversity computed for any day, with respect to the month of August 2005	ADPD
Users	Total number of users in August 2005	USR
Posters	Total number of posters in August 2005	POS
Messages	Total number of messages in August 2005	MSG
AvgDailyMessagesPerPoster	Daily average number of messages per poster	ADMP
MessagesPerPoster	Average number of messages per poster for August 2005	MPP
DaysVisited	Number of days the channel was visited in August 2005	DV
DaysActive	Number of days public conversations occurred in the channel in August 2005	DA
AvgDailyUserStability	Average daily user stability for August 2005	ADUS
AvgDailyPosterStability	Average daily poster stability for August 2005	ADPS

There were 90 channels selected in the sample. For these 90 channels various descriptive statistics were computed for the month of August 2005. These descriptive statistics, presented in Table 10.1, were used as predictor variables in the logistic regression model.

Of the 90 channels in the sample, 20 channels did not sustain public interactions at all during the first week of September 2005 (the time interval for which predictions were made). Both the linear and the nonlinear regression models successfully predicted the value zero for the BestPredictor variable for all the cases when the number of observed posters was actually zero. Therefore, these 20 channels, which were perfectly predictable in the sense that no activity was predicted by the models in 100 percent of the cases, were included in the "perfect predictability" category. Individual correlation coefficients between the BestPredictor variables produced by the best nonlinear regression model and the ObservedPosters were computed for all the other 70 channels. After sorting them based on these coefficients, the top 20 channels (rho >.615) were included in the "high predictability" category, while the bottom 20 channels (rho < .325) were included in the "low predictability" category (Table 10.2).

Table 10.2 Predictability Categories

High predictability			Low predictability channels			Perfect predictability		
Channel	Type	rho	Channel	Type	rho	Channel	Type	rho
teens	9	0.838	meadowheights	8	0.324	bbw	1	1.0
sydney	9	0.834	wasteland	5	0.308	bondage	1	1.0
gaysex	9	0.789	pussylounge	5	0.253	gloryboys	1	1.0
trogdor	2	0.755	elfornicato	3	0.243	house_atreides	1	1.0
s3xcrimes	5	0.725	mandycam	3	0.221	sub-zero	1	1.0
hott	8	0.716	wap	4	0.220	wiznet	1	1.0
unix	9	0.703	ford	6	0.218	blacksun	2	1.0
tasmania	6	0.697	exodus	9	0.195	chill_lounge	2	1.0
hattrick	2	0.696	swingers	5	0.160	cage	3	1.0
hushhush	5	0.687	Java	6	0.149	heavenly	3	1.0
jedi	6	0.685	europeanz	5	0.145	delici0us	4	1.0
vietsingle	7	0.679	chatcafe	1	-0.002	lol	4	1.0
philosophyphd	4	0.678	privacy	4	-0.002	melb_teen	4	1.0
toot_toot	3	0.665	linux	5	-0.002	noobs	4	1.0
oblivion	9	0.643	pirates	2	-0.003	curtin	6	1.0
slut	7	0.643	barbie	8	-0.005	hotplay	7	1.0
lov3	7	0.642	nofuknidea	7	-0.006	school	7	1.0
politicslounge	4	0.641	ripcurl	8	-0.006	anxienty	8	1.0
gangstaz	5	0.629	mushroom	3	-0.008	sxc	8	1.0
lesbian	9	0.616	buds	7	-0.009	united	8	1.0

10.3 Results

The logistic regression attempted to find the variables that could distinguish channels with a high degree of predictability from channels with a low degree of predictability (subsection 10.3.2); channels with a low degree of predictability from channels that were perfectly predictable (subsection 10.3.3); and channels with a high degree of predictability from channels that were perfectly predictable (subsection 10.3.4).

The SPSS software reports the results of a logistic regression in table format. Specifically, eight tables are produced when running a logistic regression.

The first table reports the omnibus tests of model coefficients and provides a test of the joint predictive ability of all the covariates in the model.

The second table reports the model summary and includes various pseudo R^2 statistics such as the Cox and Snell R^2 statistic and the Nagelkerke R^2 statistic. The R^2

statistic, which measures the variability in the dependent variable explained by a linear regression model, cannot be computed for logistic regression models. The pseudo R^2 statistics are designed to have similar properties to the true R^2 statistic and are based on a comparison of the likelihood of the current model to the "null" model (one without any predictors). Larger pseudo R^2 statistics indicate that more of the variation is explained by the model from a minimum of 0 to a maximum of 1.

The third table reports the Hosmer-Lemeshow goodness-of-fit statistic. Goodness-of-fit statistics help determine whether the model adequately describes the data. The Hosmer-Lemeshow statistic indicates a poor fit if the significance value is smaller than 0.05.

The fourth table reports the contingency for the Hosmer-Lemeshow statistic. This statistic is the most reliable test of model fit for SPSS binary logistic regression because it aggregates the observations into groups of "similar" cases. The statistic is then computed based upon these groups.

The fifth table, called the classification table, reports the practical results of using the logistic regression model. From step to step, the improvement in classification indicates how well the model performed. A better model should correctly identify a higher percentage of the cases.

The sixth table is the parameter estimates table, and it summarizes the effect of each predictor. It includes the Wald statistic, which is computed as the ratio of the coefficient to its standard error, squared. If the significance level of the Wald statistic is small (less than 0.05), then the parameter is useful to the model. The predictors and

coefficient values shown in the last row of the table are used by the procedure to make predictions.

The seventh table reports the potential effects of removing the variables chosen by the regression method from the model. The most important thing to note from this table is the significance value of the -2 log-likelihood ratio. All the variables that were chosen should have significant changes in -2 log-likelihood. This indicates that the removal of the variables from the model influences its ability to make accurate predictions. The change in -2 log-likelihood is generally more reliable than the Wald statistic.

The eighth table reports the variables that were removed from the model.

Table 10.3 Spearman Correlation Coefficients for All Predictor Variables

	ST	AURT	AU	ADU	AP	ADP	AM	ADM	AUD	ADUD	APD	ADPD	USR	POS	MSG	ADMP	MPP	DV	DA	ADUS	ADPS
ST	1.00	-.550	.375	.314	.291	.238	.172	.042	.095	-.417	.186	-.332	.344	.289	.141	-.168	-.169	.435	.360	.160	-.223
AURT	-.550	1.00	-.239	-.168	-.232	-.197	-.242	-.183	-.236	.189	-.209	.248	-.159	-.223	-.214	-.119	-.095	-.260	-.248	-.204	.041
AU	.375	-.239	1.00	.942	.824	.838	.544	.443	.053	-.808	-.014	-.828	.956	.918	.557	-.089	-.103	.846	.768	-.244	-.122
ADU	.314	-.168	.942	1.00	.786	.840	.459	.403	-.185	-.736	-.126	-.763	.965	.899	.486	-.171	-.196	.746	.666	-.269	-.135
AP	.291	-.232	.824	.786	1.00	.957	.827	.768	.043	-.727	.104	-.714	.807	.932	.814	.282	.270	.817	.789	-.231	.114
ADP	.238	-.197	.838	.840	.957	1.00	.762	.738	-.063	-.676	-.112	-.671	.835	.952	.770	.210	.208	.739	.722	-.263	.115
AM	.172	-.242	.544	.459	.827	.762	1.00	.960	.236	-.451	.079	-.542	.489	.698	.974	.712	.691	.648	.793	-.047	.403
ADM	.042	-.183	.443	.403	.768	.738	.960	1.00	.153	-.295	-.062	-.430	.398	.643	.970	.767	.747	.500	.696	-.021	.493
AUD	.095	-.236	.053	-.185	.043	-.063	.236	.153	1.00	-.136	.198	-.131	-.082	-.015	.216	.318	.350	.259	.277	.074	.121
ADUD	-.417	.189	-.808	-.736	-.727	-.676	-.451	-.295	-.136	1.00	-.185	.867	-.875	-.809	-.407	.153	.201	-.817	-.639	.550	.284
APD	.186	-.209	-.014	-.126	.104	-.112	.079	-.062	.198	-.185	1.00	-.085	-.059	-.040	.020	.119	.097	.227	.127	.013	-.032
ADPD	-.332	.248	-.828	-.763	-.714	-.671	-.542	-.430	-.131	.867	-.085	1.00	-.864	-.848	-.561	.012	.068	-.804	-.816	.306	.159
USR	.344	-.159	.956	.965	.807	.835	.489	.398	-.082	-.875	-.059	-.864	1.00	.929	.495	-.166	-.202	.807	.701	-.398	-.198
POS	.289	-.223	.918	.899	.932	.952	.698	.643	-.015	-.809	-.040	-.848	.929	1.00	.716	.096	.073	.811	.794	-.321	.008
MSG	.141	-.214	.557	.486	.814	.770	.974	.970	.216	-.407	.020	-.561	.495	.716	1.00	.710	.699	.625	.835	-.001	.473
ADMP	-.168	-.119	-.089	-.171	.282	.210	.712	.767	.318	.153	.119	.012	-.166	.096	.710	1.00	.964	.085	.341	.157	.668
MPP	-.169	-.095	-.103	-.196	.270	.208	.691	.747	.350	.201	.097	.068	-.202	.073	.699	.964	1.00	.100	.356	.215	.754
DV	.435	-.260	.846	.746	.817	.739	.648	.500	.259	-.817	.227	-.804	.807	.811	.625	.085	.100	1.00	.556	-.164	-.026
DA	.360	-.248	.768	.666	.789	.722	.793	.696	.277	-.639	.127	-.816	.701	.794	.835	.341	.356	.556	1.00	-.011	.221
ADUS	.160	-.204	-.244	-.269	-.231	-.263	-.047	-.021	.074	.550	.013	.306	-.398	-.321	-.001	.157	.215	-.164	-.011	1.00	.280
ADPS	-.223	.041	-.122	-.135	.114	.115	.403	.493	.121	.284	-.032	.159	-.198	.008	.473	.668	.754	-.026	.221	.280	1.00

Table 10.3 presents the Spearman correlation coefficients among all the predictor variables. Examining these correlations is important to determine the existence of multicollinearity and to avoid the effects caused by it. In logistic regression models, multicollinearity is a result of strong correlations between independent variables. The existence of multicollinearity inflates the variances of the parameter estimates. That may result, particularly for small and moderate sample sizes, in the lack of statistical significance of individual independent variables, while the overall model may be strongly significant. Multicollinearity may also result in wrong signs and magnitudes of regression coefficient estimates, and consequently in incorrect conclusions about relationships between independent and dependent variables.

Table 10.3 reveals many strong correlations between several of the independent variables. This was to be expected considering that many of the independent variables were various measures related to the number of users, posters, and messages that characterized IRC channels and were likely to be highly correlated.

Two main methods of logistic regression were used in this analysis: the backward stepwise method and the forward stepwise method. A backward stepwise method starts with a model that includes all the predictors. At each step, the predictor that contributes the least is removed from the model until all of the predictors in the model are significant. A forward stepwise method starts with a model that does not include any of the predictors. At each step, the predictor with the largest score statistic whose significance value is less than a specified value (by default 0.05) is added to the model. The variables left out of the analysis at the last step all have significance values larger than 0.05, so no

more are added. If both methods choose the same variables, one can be fairly confident that the chosen model is a good one.

10.3.1 Effects of Multicollinearity

Simply to illustrate the potential effects of multicollinearity, a backward stepwise logistic regression was conducted using all the independent variables with the goal of identifying the best predictors that can separate channels with "high" predictability from channels with "low" predictability. The dependent variable was the predictability of channels, particularly the "high predictability/low predictability" dichotomy. The results of this backward stepwise logistic regression are presented in Tables 10.4 through 10.6.

The Hosmer-Lemeshow statistic indicates that it took six steps to produce the final model, and the model adequately fitted the data at each step.

Table 10.4 Hosmer-Lemeshow Goodness-of-Fit Statistic

Step	Chi-square	df	Sig.
1	.000	7	1.000
2	.000	7	1.000
3	.000	7	1.000
4	.000	7	1.000
5	.000	7	1.000
6	.000	7	1.000

The classification table (Table 10.5) shows the practical results of using the logistic regression model. For each case, the predicted response is *high* or *low*, based on each case's model-predicted probability being greater than the cutoff value specified (default value is 0.5). In this case, the model's predictions were 100 percent correct in all six steps.

Table 10.5 Classification Table

Observed			Predicted		
			Predictability		
			high	low	Percentage Correct
Step 1	Predictability	high	20	0	100.0
		low	0	20	100.0
	Overall Percentage				100.0
Step 2	Predictability	high	20	0	100.0
		low	0	20	100.0
	Overall Percentage				100.0
Step 3	Predictability	high	20	0	100.0
		low	0	20	100.0
	Overall Percentage				100.0
Step 4	Predictability	high	20	0	100.0
		low	0	20	100.0
	Overall Percentage				100.0
Step 5	Predictability	high	20	0	100.0
		low	0	20	100.0
	Overall Percentage				100.0
Step 6	Predictability	high	20	0	100.0
		low	0	20	100.0
	Overall Percentage				100.0

The parameter estimates table (Table 10.6) summarizes the effect of each predictor. If the significance level of the Wald statistic is small (less than 0.05), then the parameter is useful to the model.

Table 10.6 Parameter Estimates Table

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 6	AvgDailyPosters	-10.969	3015.526	.000	1	.997	.000
	AvgMessages	.322	486.194	.000	1	.999	1.380
	AvgDailyMessages	-.225	591.416	.000	1	1.000	.798
	AvgUserDiv	-38.382	53526.056	.000	1	.999	.000
	AvgDailyUserDiv	-263.449	972145.285	.000	1	1.000	.000
	AvgPosterDiv	-355.086	114089.696	.000	1	.998	.000
	AvgDailyPosterDiv	110.294	508355.358	.000	1	1.000	7.946E+047
	Users	.080	81.124	.000	1	.999	1.083
	Posters	.313	370.278	.000	1	.999	1.368
	Messages	.003	22.973	.000	1	1.000	1.003
	AvgDailyMessagesPerPoster	.984	3626.286	.000	1	1.000	2.674
	MessagesPerPoster	-.031	496.944	.000	1	1.000	.969
	DaysVisited	7.191	2379.325	.000	1	.998	1327.525
	DaysActive	-5.888	3878.580	.000	1	.999	.003
	AvgDailyUserStability	53.380	149441.389	.000	1	1.000	152316.000
	AvgDailyPosterStability	-8.110	73762.110	.000	1	1.000	.000
	Constant	344.284	143004.480	.000	1	.998	3.315E+149

Table 10.6 reports only the predictors and coefficient values from the last step of the logistic regression. It may be noted that although the predictions were 100 percent accurate, all the independent variables lacked statistical significance – a result of their multicollinearity.

Such problems can and must be avoided by selecting a set of independent variables that are not highly correlated with each other. Table 10.3 helps identify suitable combinations of independent variables to be used in the logistic regression models.

10.3.2 High Predictability/Low Predictability Channel Differentiation

Both a forward stepwise logistic regression and a backward stepwise logistic regression were conducted to identify the best predictors for distinguishing channels with a high degree of predictability from channels with a low degree of predictability. Multiple combinations of independent variables were attempted for both regressions. The tables below report the results of the best models, i.e., the models that produced the highest percentage of correct predictions. Tables 10.7 through 10.14 present the results of the forward stepwise logistic regression.

Table 10.7 Omnibus Tests of Model Coefficients for the Forward Stepwise Logistic Regression

		Chi-square	df	Sig.
Step 1	Step	11.290	1	.001
	Block	11.290	1	.001
	Model	11.290	1	.001
Step 2	Step	11.518	1	.001
	Block	22.808	2	.000
	Model	22.808	2	.000

Table 10.8 Model Summary for the Forward Stepwise Logistic Regression

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	44.162	.246	.328
2	32.644	.435	.579

Table 10.9 Hosmer-Lemeshow Statistic for the Forward Stepwise Logistic Regression

Step	Chi-square	df	Sig.
1	7.506	7	.378
2	4.234	7	.752

Table 10.10 Contingency Table for the Hosmer-Lemeshow Statistic for the Forward Stepwise Logistic Regression

		Predictability of the channel = high		Predictability of the channel = low		
		Observed	Expected	Observed	Expected	Total
Step 1	1	5	4.410	0	.590	5
	2	3	4.003	2	.997	5
	3	4	3.402	1	1.598	5
	4	2	2.185	2	1.815	4
	5	2	2.120	3	2.880	5
	6	1	1.713	4	3.287	5
	7	1	1.041	3	2.959	4
	8	2	.595	1	2.405	3
	9	0	.530	4	3.470	4
Step 2	1	5	4.855	0	.145	5
	2	4	3.612	0	.388	4
	3	3	3.179	1	.821	4
	4	2	2.749	2	1.251	4
	5	3	2.777	2	2.223	5
	6	1	1.546	3	2.454	4
	7	2	.809	3	4.191	5
	8	0	.331	4	3.669	4
	9	0	.143	5	4.857	5

Table 10.11 Classification Table for the Forward Stepwise Logistic Regression

			Predicted		
			Predictability		
Observed			high	low	Percentage Correct
Step 1	Predictability	high	14	6	70.0
		low	4	16	80.0
	Overall Percentage				75.0
Step 2	Predictability	high	16	4	80.0
		low	4	16	80.0
	Overall Percentage				80.0

Table 10.12 Parameter Estimates for the Forward Stepwise Logistic Regression

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	DaysActive	-.145	.050	8.354	1	.004	.865
	Constant	2.424	.903	7.213	1	.007	11.290
Step 2	DaysActive	-.627	.204	9.433	1	.002	.534
	DaysVisited	.503	.189	7.110	1	.008	1.653
	Constant	.043	1.147	.001	1	.970	1.044

Table 10.13 Model if Term Removed for the Forward Stepwise Logistic Regression

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	DaysActive	-27.726	11.290	1	.001
Step 2	DaysActive	-26.520	20.396	1	.000
	DaysVisited	-22.081	11.518	1	.001

Table 10.14 Variables Excluded from the Model in the Forward Stepwise Logistic Regression

			Score	df	Sig.
Step 1	Variables	SurvivalTime	4.638	1	.031
		AvgUserRetTime	.093	1	.761
		AvgDailyMessagesPerPoster	3.575	1	.059
		DaysVisited	7.185	1	.007
		AvgUserDiv	.073	1	.787
		AvgPosterDiv	.028	1	.867
		AvgDailyUserStability	.208	1	.648
		AvgDailyPosterStability	3.420	1	.064
	Overall Statistics		11.668	8	.167
Step 2	Variables	SurvivalTime	.786	1	.375
		AvgUserRetTime	.012	1	.912
		AvgDailyMessagesPerPoster	1.546	1	.214
		AvgUserDiv	1.360	1	.243
		AvgPosterDiv	2.150	1	.143
		AvgDailyUserStability	.025	1	.875
		AvgDailyPosterStability	.858	1	.354
	Overall Statistics		4.782	7	.687

The value of the Hosmer-Lemeshow goodness-of-fit statistic in the final step (>.05) indicates that the model adequately fitted the data. The classification table indicates that predictions were successful in 80 percent of the cases. The model correctly predicted that 16 out of 20 channels belonged to the correct category (either "high predictability" or "low predictability"). The model identified the independent variables *DaysActive* and *DaysVisited* as the best predictors; and the parameter estimates table shows that they were statistically significant (.002 and .008). Both variables chosen by the model had significant changes in -2 log-likelihood; and the significance level of the Wald statistic was small enough to indicate that the parameters were useful to the model.

Tables 10.15 – 10.22 present the results of the backward stepwise logistic regression, which was completed in eight steps. Due to space considerations, some of the tables below report only the results produced by the final step of the model.

Table 10.15 Omnibus Tests of Model Coefficients for the Backward Stepwise Logistic Regression

		Chi-square	Df	Sig.
Step 1	Step	29.050	9	.001
	Block	29.050	9	.001
	Model	29.050	9	.001
Step 2	Step	-.030	1	.863
	Block	29.020	8	.000
	Model	29.020	8	.000
Step 3	Step	-.068	1	.795
	Block	28.953	7	.000
	Model	28.953	7	.000
Step 4	Step	-.901	1	.343
	Block	28.052	6	.000
	Model	28.052	6	.000
Step 5	Step	-.426	1	.514
	Block	27.626	5	.000
	Model	27.626	5	.000
Step 6	Step	-1.184	1	.277
	Block	26.442	4	.000
	Model	26.442	4	.000
Step 7	Step	-1.279	1	.258
	Block	25.164	3	.000
	Model	25.164	3	.000
Step 8	Step	-2.356	1	.125
	Block	22.808	2	.000
	Model	22.808	2	.000

Table 10.16 Model Summary for the Backward Stepwise Logistic Regression

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	26.402	.516	.688
2	26.431	.516	.688
3	26.499	.515	.687
4	27.400	.504	.672
5	27.826	.499	.665
6	29.010	.484	.645
7	30.288	.467	.623
8	32.644	.435	.579

Table 10.17 Hosmer-Lemeshow Statistic for the Backward Stepwise Logistic Regression

Step	Chi-square	df	Sig.
1	6.371	8	.606
2	4.630	8	.796
3	5.326	8	.722
4	5.334	8	.721
5	8.598	8	.377
6	3.917	8	.864
7	5.846	8	.665
8	4.234	7	.752

Table 10.18 Contingency Table for the Hosmer-Lemeshow Statistic for the Backward Stepwise Logistic Regression

		Predictability of the channel = high		Predictability of the channel = low		
		Observed	Expected	Observed	Expected	Total
Step 8	1	4	3.513	0	.487	4
	2	3	3.343	1	.657	4
	3	2	2.992	2	1.008	4
	4	3	2.548	1	1.452	4
	5	3	2.067	1	1.933	4
	6	1	1.701	3	2.299	4
	7	1	1.394	3	2.606	4
	8	1	1.102	3	2.898	4
	9	2	.822	2	3.178	4
	10	0	.518	4	3.482	4

Table 10.19 Classification Table for the Backward Stepwise Logistic Regression

Observed			Predicted		
			Predictability		
			high	low	Percentage Correct
Step 1	Predictability	high	17	3	85.0
		low	6	14	70.0
	Overall Percentage				77.5
Step 2	Predictability	high	17	3	85.0
		low	6	14	70.0
	Overall Percentage				77.5
Step 3	Predictability	high	17	3	85.0
		low	6	14	70.0
	Overall Percentage				77.5
Step 4	Predictability	high	17	3	85.0
		low	4	16	80.0
	Overall Percentage				82.5
Step 5	Predictability	high	16	4	80.0
		low	5	15	75.0
	Overall Percentage				77.5
Step 6	Predictability	high	17	3	85.0
		low	5	15	75.0
	Overall Percentage				80.0
Step 7	Predictability	high	16	4	80.0
		low	5	15	75.0
	Overall Percentage				77.5
Step 8	Predictability	high	16	4	80.0
		low	4	16	80.0
	Overall Percentage				80.0

Table 10.20 Parameter Estimates for the Backward Stepwise Logistic Regression

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	SurvivalTime	.131	.114	1.318	1	.251	1.139
	AvgUserRetTime	.000	.001	.062	1	.803	1.000
	AvgDailyMessagesPerPoster	-.007	.042	.030	1	.863	.993
	DaysActive	-.824	.326	6.403	1	.011	.439
	DaysVisited	.711	.316	5.064	1	.024	2.036
	AvgUserDiv	-5.606	5.628	.992	1	.319	.004
	AvgPosterDiv	-5.936	4.617	1.653	1	.199	.003
	AvgDailyUserStability	6.495	6.470	1.008	1	.315	661.535
	AvgDailyPosterStability	-3.375	5.036	.449	1	.503	.034
	Constant	3.017	6.292	.230	1	.632	20.437
Step 2	SurvivalTime	.133	.114	1.349	1	.245	1.142
	AvgUserRetTime	.000	.001	.067	1	.796	1.000
	DaysActive	-.835	.323	6.694	1	.010	.434
	DaysVisited	.721	.313	5.314	1	.021	2.056
	AvgUserDiv	-5.934	5.339	1.235	1	.266	.003
	AvgPosterDiv	-6.118	4.508	1.841	1	.175	.002
	AvgDailyUserStability	6.613	6.463	1.047	1	.306	744.965
	AvgDailyPosterStability	-3.866	4.184	.854	1	.355	.021
	Constant	3.251	6.124	.282	1	.596	25.805
Step 3	SurvivalTime	.121	.104	1.356	1	.244	1.128
	DaysActive	-.847	.321	6.957	1	.008	.429
	DaysVisited	.733	.310	5.585	1	.018	2.082
	AvgUserDiv	-6.246	5.189	1.449	1	.229	.002
	AvgPosterDiv	-6.329	4.448	2.024	1	.155	.002
	AvgDailyUserStability	6.017	5.980	1.012	1	.314	410.336
	AvgDailyPosterStability	-3.545	3.888	.831	1	.362	.029
	Constant	4.562	3.521	1.679	1	.195	95.809
Step 4	SurvivalTime	.131	.094	1.930	1	.165	1.140
	DaysActive	-.828	.292	8.022	1	.005	.437
	DaysVisited	.713	.288	6.121	1	.013	2.041
	AvgUserDiv	-5.096	4.530	1.266	1	.261	.006
	AvgPosterDiv	-6.029	4.315	1.953	1	.162	.002
	AvgDailyUserStability	2.556	3.951	.418	1	.518	12.884
	Constant	3.141	2.995	1.099	1	.294	23.120
Step 5	SurvivalTime	.119	.089	1.809	1	.179	1.127
	DaysActive	-.781	.270	8.348	1	.004	.458
	DaysVisited	.673	.274	6.021	1	.014	1.959
	AvgUserDiv	-4.704	4.419	1.134	1	.287	.009
	AvgPosterDiv	-5.356	4.107	1.701	1	.192	.005
	Constant	3.393	2.979	1.297	1	.255	29.758
Step 6	SurvivalTime	.085	.078	1.179	1	.278	1.089
	DaysActive	-.704	.241	8.551	1	.003	.495
	DaysVisited	.593	.247	5.783	1	.016	1.809
	AvgPosterDiv	-6.152	3.918	2.466	1	.116	.002
	Constant	2.487	2.784	.798	1	.372	12.023
Step 7	DaysActive	-.734	.245	9.016	1	.003	.480
	DaysVisited	.655	.248	6.996	1	.008	1.926
	AvgPosterDiv	-5.401	3.795	2.026	1	.155	.005
	Constant	3.342	2.576	1.684	1	.194	28.279
Step 8	DaysActive	-.627	.204	9.433	1	.002	.534
	DaysVisited	.503	.189	7.110	1	.008	1.653
	Constant	.043	1.147	.001	1	.970	1.044

Table 10.21 Model if Term Removed for the Backward Stepwise Logistic Regression

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 8	DaysActive	-26.520	20.396	1	.000
	DaysVisited	-22.081	11.518	1	.001

Table 10.22 Variables Excluded from the Model in the Backward Stepwise Logistic Regression

			Score	df	Sig.
Step 2	Variables	AvgDailyMessagesPerPoster	.030	1	.863
	Overall Statistics		.030	1	.863
Step 3	Variables	AvgUserRetTime	.067	1	.796
		AvgDailyMessagesPerPoster	.035	1	.853
	Overall Statistics		.096	2	.953
Step 4	Variables	AvgUserRetTime	.001	1	.977
		AvgDailyMessagesPerPoster	.501	1	.479
		AvgDailyPosterStability	.846	1	.358
	Overall Statistics		.921	3	.820
Step 5	Variables	AvgUserRetTime	.028	1	.866
		AvgDailyMessagesPerPoster	.171	1	.679
		AvgDailyUserStability	.429	1	.513
		AvgDailyPosterStability	.128	1	.720
	Overall Statistics		1.253	4	.869
Step 6	Variables	AvgUserRetTime	.046	1	.830
		AvgDailyMessagesPerPoster	.455	1	.500
		AvgUserDiv	1.194	1	.274
		AvgDailyUserStability	.265	1	.607
		AvgDailyPosterStability	.152	1	.697
	Overall Statistics		2.294	5	.807
Step 7	Variables	SurvivalTime	1.237	1	.266
		AvgUserRetTime	.299	1	.584
		AvgDailyMessagesPerPoster	.708	1	.400
		AvgUserDiv	.429	1	.513
		AvgDailyUserStability	.223	1	.637
		AvgDailyPosterStability	.693	1	.405
	Overall Statistics		3.022	6	.806
Step 8	Variables	SurvivalTime	.786	1	.375
		AvgUserRetTime	.012	1	.912
		AvgDailyMessagesPerPoster	1.546	1	.214
		AvgUserDiv	1.360	1	.243
		AvgPosterDiv	2.150	1	.143
		AvgDailyUserStability	.025	1	.875
		AvgDailyPosterStability	.858	1	.354
	Overall Statistics		4.782	7	.687

The value of the Hosmer-Lemeshow goodness-of-fit statistic in the final step (>.05) indicates that the model adequately fitted the data. The classification table indicates that predictions were successful in 80 percent of the cases. The model correctly predicted that 16 out of 20 channels belonged to the correct category (either "high

predictability", or "low predictability"). The model identified the independent variables *DaysActive* and *DaysVisited* as the best predictors, and the parameter estimates table shows that they were statistically significant (.002 and .008). Both variables chosen by the model had significant changes in -2 log-likelihood and the significance level of the Wald statistic was small enough to indicate that the parameters were useful to the model.

Since both the forward and the backward methods produced the same results, one can be fairly confident that the model was a good one. The independent variables *DaysActive* and *DaysVisited* successfully categorized channels into "high predictability" and "low predictability" categories in 80 percent of the cases.

10.3.3 Low Predictability/Perfect Predictability Channel Differentiation

Both a forward and a backward stepwise logistic regression were conducted to identify the best predictors for distinguishing channels with a low degree of predictability from channels with a perfect degree of predictability. Multiple combinations of independent variables were attempted for both regressions. Tables 10.23 – 10.30 below present the results of the best forward stepwise logistic regression model.

Table 10.23 Omnibus Tests of Model Coefficients for the Forward Stepwise Logistic Regression

		Chi-square	df	Sig.
Step 1	Step	7.932	1	.005
	Block	7.932	1	.005
	Model	7.932	1	.005

Table 10.24 Model Summary for the Forward Stepwise Logistic Regression

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	47.520	.180	.240

Table 10.25 Hosmer-Lemeshow Statistic for the Forward Stepwise Logistic Regression

Step	Chi-square	df	Sig.
1	12.654	7	.081

Table 10.26 Contingency Table for the Hosmer-Lemeshow Statistic for the Forward Stepwise Logistic Regression

		Predictability of the channel = low		Predictability of the channel = full		
		Observed	Expected	Observed	Expected	Total
Step 1	1	3	3.599	1	.401	4
	2	4	4.386	2	1.614	6
	3	5	3.073	0	1.927	5
	4	2	1.569	1	1.431	3
	5	1	1.934	3	2.066	4
	6	0	1.917	5	3.083	5
	7	4	1.859	2	4.141	6
	8	1	1.243	4	3.757	5
	9	0	.419	2	1.581	2

Table 10.27 Classification Table for the Forward Stepwise Logistic Regression

			Predicted		
			Predictability		
Observed			low	perfect	Percentage Correct
Step 1	Predictability	low	14	6	70.0
		perfect	4	16	80.0
	Overall Percentage				75.0

Table 10.28 Parameter Estimates for the Forward Stepwise Logistic Regression

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	DaysActive	-.158	.065	5.946	1	.015	.854
	Constant	1.486	.677	4.821	1	.028	4.419

Table 10.29 Model if Term Removed for the Forward Stepwise Logistic Regression

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	DaysActive	-27.726	7.932	1	.005

Table 10.30 Variables Excluded from the Model in the Forward Stepwise Logistic Regression

			Score	df	Sig.
Step 1	Variables	SurvivalTime	1.395	1	.238
		AvgUserRetTime	1.819	1	.177
		AvgDailyMessagesPerPoster	.030	1	.862
		AvgUserDiv	.558	1	.455
		AvgPosterDiv	1.588	1	.208
		AvgDailyUserStability	.054	1	.817
		AvgDailyPosterStability	.786	1	.375
		DaysVisited	.011	1	.916

The value of the Hosmer-Lemeshow goodness-of-fit statistic in the final step (>.05) indicates that the model adequately fitted the data. The classification table shows that predictions were successful in 75 percent of the cases. The model correctly predicted that 14 out of 20 channels belonged to the "low predictability" category and 16 out of 20 channels belonged to the "perfect predictability" category. Only one independent variable was identified as the best predictor: *DaysActive*. The parameter estimates table shows that it was statistically significant (.015). The variable chosen by the model had significant changes in -2 log-likelihood; and the significance level of the Wald statistic was small enough to indicate that the parameter was useful to the model.

Tables 10.31 – 10.38 present the results of the backward stepwise logistic regression. Due to space considerations, some of the tables below report only the results produced by the final step of the model.

Table 10.31 Omnibus Tests of Model Coefficients for the Backward Stepwise Logistic Regression

		Chi-square	df	Sig.
Step 1	Step	17.513	9	.041
	Block	17.513	9	.041
	Model	17.513	9	.041
Step 2	Step	-.157	1	.692
	Block	17.356	8	.027
	Model	17.356	8	.027
Step 3	Step	-.686	1	.408
	Block	16.671	7	.020
	Model	16.671	7	.020
Step 4	Step	-.767	1	.381
	Block	15.903	6	.014
	Model	15.903	6	.014
Step 5	Step	-1.012	1	.314
	Block	14.892	5	.011
	Model	14.892	5	.011
Step 6	Step	-1.972	1	.160
	Block	12.919	4	.012
	Model	12.919	4	.012
Step 7	Step	-1.074	1	.300
	Block	11.845	3	.008
	Model	11.845	3	.008
Step 8	Step	-2.396	1	.122
	Block	9.449	2	.009
	Model	9.449	2	.009
Step 9	Step	-1.517	1	.218
	Block	7.932	1	.005
	Model	7.932	1	.005

Table 10.32 Model Summary for the Backward Stepwise Logistic Regression

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	37.939	.355	.473
2	38.095	.352	.469
3	38.781	.341	.454
4	39.548	.328	.437
5	40.560	.311	.414
6	42.533	.276	.368
7	43.606	.256	.342
8	46.002	.210	.281
9	47.520	.180	.240

Table 10.33 Hosmer-Lemeshow Statistic for the Backward Stepwise Logistic Regression

Step	Chi-square	df	Sig.
1	15.381	8	.052
2	5.828	8	.666
3	13.193	8	.105
4	12.682	8	.123
5	20.165	8	.010
6	10.600	8	.225
7	15.147	8	.056
8	10.694	8	.220
9	5.002	4	.287

Table 10.34 Contingency Table for the Hosmer-Lemeshow Statistic for the Backward Stepwise Logistic Regression

		Predictability of the channel = low		Predictability of the channel = full		Total
		Observed	Expected	Observed	Expected	
Step 9	1	13	12.159	7	7.841	20
	2	1	2.277	3	1.723	4
	3	2	2.004	2	1.996	4
	4	3	2.032	2	2.968	5
	5	0	1.101	4	2.899	4
	6	1	.427	2	2.573	3

Table 10.35 Classification Table for the Backward Stepwise Logistic Regression

Observed			Predicted		
			Predictability		
			low	perfect	Percentage Correct
Step 1	Predictability	low	15	5	75.0
		perfect	3	17	85.0
	Overall Percentage				80.0
Step 2	Predictability	low	15	5	75.0
		perfect	3	17	85.0
	Overall Percentage				80.0
Step 3	Predictability	low	16	4	80.0
		perfect	4	16	80.0
	Overall Percentage				80.0
Step 4	Predictability	low	16	4	80.0
		perfect	5	15	75.0
	Overall Percentage				77.5
Step 5	Predictability	Low	16	4	80.0
		perfect	4	16	80.0
	Overall Percentage				80.0
Step 6	Predictability	low	14	6	70.0
		perfect	5	15	75.0
	Overall Percentage				72.5
Step 7	Predictability	low	14	6	70.0
		perfect	8	12	60.0
	Overall Percentage				65.0
Step 8	Predictability	low	12	8	60.0
		perfect	6	14	70.0
	Overall Percentage				65.0
Step 9	Predictability	low	14	6	70.0
		perfect	4	16	80.0
	Overall Percentage				75.0

Table 10.36 Parameter Estimates for the Backward Stepwise Logistic Regression

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	SurvivalTime	-.124	.085	2.102	1	.147	.884
	AvgUserRetTime	.000	.000	1.025	1	.311	1.000
	AvgDailyMessagesPerPoster	-.047	.046	1.053	1	.305	.954
	AvgUserDiv	5.133	3.186	2.595	1	.107	169.519
	AvgPosterDiv	-3.773	2.978	1.606	1	.205	.023
	AvgDailyUserStability	3.322	3.674	.818	1	.366	27.709
	AvgDailyPosterStability	3.256	3.166	1.058	1	.304	25.947
	DaysVisited	.042	.106	.158	1	.691	1.043
	DaysActive	-.256	.171	2.229	1	.135	.774
	Constant	2.513	2.269	1.226	1	.268	12.343
Step 2	SurvivalTime	-.118	.085	1.946	1	.163	.889
	AvgUserRetTime	.000	.000	.905	1	.342	1.000
	AvgDailyMessagesPerPoster	-.050	.046	1.192	1	.275	.951
	AvgUserDiv	5.495	3.092	3.158	1	.076	243.350
	AvgPosterDiv	-3.501	2.915	1.442	1	.230	.030
	AvgDailyUserStability	2.823	3.479	.659	1	.417	16.825
	AvgDailyPosterStability	3.093	3.070	1.015	1	.314	22.048
	DaysActive	-.206	.110	3.524	1	.060	.814
	Constant	2.514	2.288	1.207	1	.272	12.357
Step 3	SurvivalTime	-.109	.086	1.603	1	.206	.897
	AvgUserRetTime	.000	.000	.750	1	.387	1.000
	AvgDailyMessagesPerPoster	-.051	.047	1.158	1	.282	.950
	AvgUserDiv	5.362	2.979	3.238	1	.072	213.052
	AvgPosterDiv	-2.968	2.708	1.201	1	.273	.051
	AvgDailyPosterStability	3.407	2.995	1.294	1	.255	30.183
	DaysActive	-.194	.105	3.440	1	.064	.823
	Constant	2.674	2.332	1.315	1	.252	14.502
Step 4	SurvivalTime	-.131	.083	2.460	1	.117	.877
	AvgDailyMessagesPerPoster	-.054	.046	1.360	1	.244	.948
	AvgUserDiv	5.827	3.014	3.738	1	.053	339.278
	AvgPosterDiv	-2.596	2.629	.975	1	.323	.075
	AvgDailyPosterStability	3.167	2.876	1.212	1	.271	23.729
	DaysActive	-.180	.098	3.373	1	.066	.835
	Constant	3.600	2.106	2.923	1	.087	36.608
Step 5	SurvivalTime	-.158	.081	3.802	1	.051	.854
	AvgDailyMessagesPerPoster	-.065	.044	2.154	1	.142	.937
	AvgUserDiv	4.876	2.746	3.153	1	.076	131.079
	AvgDailyPosterStability	3.690	2.768	1.777	1	.183	40.052
	DaysActive	-.216	.092	5.452	1	.020	.806
	Constant	3.283	2.077	2.498	1	.114	26.650
Step 6	SurvivalTime	-.141	.075	3.523	1	.061	.868
	AvgDailyMessagesPerPoster	-.035	.036	.944	1	.331	.966
	AvgUserDiv	4.493	2.695	2.779	1	.096	89.415
	DaysActive	-.159	.071	4.957	1	.026	.853
	Constant	3.051	1.962	2.418	1	.120	21.129
Step 7	SurvivalTime	-.112	.066	2.870	1	.090	.894
	AvgUserDiv	3.594	2.444	2.162	1	.142	36.375
	DaysActive	-.160	.071	5.103	1	.024	.852
	Constant	2.319	1.707	1.846	1	.174	10.165
Step 8	SurvivalTime	-.067	.059	1.318	1	.251	.935
	DaysActive	-.141	.067	4.472	1	.034	.868
	Constant	3.116	1.676	3.457	1	.063	22.546
Step 9	DaysActive	-.158	.065	5.946	1	.015	.854
	Constant	1.486	.677	4.821	1	.028	4.419

Table 10.37 Model if Term Removed for the Backward Stepwise Logistic Regression

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 9	DaysActive	-27.726	7.932	1	.005

Table 10.38 Variables Excluded from the Model in the Backward Stepwise Logistic Regression

			Score	df	Sig.
Step 2	Variables	DaysVisited	.159	1	.690
	Overall Statistics		.159	1	.690
Step 3	Variables	AvgDailyUserStability	.673	1	.412
		DaysVisited	.004	1	.952
	Overall Statistics		.848	2	.654
Step 4	Variables	AvgUserRetTime	.789	1	.375
		AvgDailyUserStability	.494	1	.482
		DaysVisited	.017	1	.898
Step 5	Variables	AvgUserRetTime	.526	1	.468
		AvgPosterDiv	1.006	1	.316
		AvgDailyUserStability	.264	1	.608
		DaysVisited	.089	1	.765
Step 6	Variables	AvgUserRetTime	.382	1	.536
		AvgPosterDiv	1.693	1	.193
		AvgDailyUserStability	.512	1	.474
		AvgDailyPosterStability	1.880	1	.170
		DaysVisited	.168	1	.682
Step 7	Variables	AvgUserRetTime	.634	1	.426
		AvgDailyMessagesPerPoster	.978	1	.323
		AvgPosterDiv	2.214	1	.137
		AvgDailyUserStability	.244	1	.622
		AvgDailyPosterStability	.423	1	.516
		DaysVisited	.005	1	.944
Step 8	Variables	AvgUserRetTime	1.299	1	.254
		AvgDailyMessagesPerPoster	.231	1	.630
		AvgUserDiv	2.332	1	.127
		AvgPosterDiv	.665	1	.415
		AvgDailyUserStability	.382	1	.536
		AvgDailyPosterStability	.613	1	.434
		DaysVisited	.163	1	.687
	Overall Statistics		6.989	7	.430
Step 9	Variables	SurvivalTime	1.395	1	.238
		AvgUserRetTime	1.819	1	.177
		AvgDailyMessagesPerPoster	.030	1	.862
		AvgUserDiv	.558	1	.455
		AvgPosterDiv	1.588	1	.208
		AvgDailyUserStability	.054	1	.817
		AvgDailyPosterStability	.786	1	.375
		DaysVisited	.011	1	.916

The value of the Hosmer-Lemeshow goodness-of-fit statistic in the final step (>.05) indicates that the model adequately fitted the data. The classification table shows that predictions were successful in 75 percent of the cases. The model correctly predicted

that 14 out of 20 channels belonged to the "low predictability" category and 16 out of 20 channels belonged to the "perfect predictability" category. Only one independent variable was identified as the best predictor: *DaysActive*. The parameter estimates table shows that it was statistically significant (.015). The variable chosen by the model had significant changes in -2 log-likelihood; and the significance level of the Wald statistic was small enough to indicate that the parameter was useful to the model.

Since both the forward and the backward methods produced the same results, one can be fairly confident that the chosen model was a good one. The independent variable *DaysActive* can be used to successfully categorize channels into "low predictability" and "perfect predictability" categories in 75 percent of the cases.

10.3.4 High Predictability/Perfect Predictability Channel Differentiation

Both a forward and a backward stepwise logistic regression were conducted to identify the best predictors for differentiating channels with a high degree of predictability from channels with a perfect degree of predictability. Multiple combinations of independent variables were attempted for both regressions. Tables 10.39 – 10.46 below present the results of the best forward stepwise logistic regression model.

Table 10.39 Omnibus Tests of Model Coefficients for the Forward Stepwise Logistic Regression

		Chi-square	df	Sig.
Step 1	Step	28.026	1	.000
	Block	28.026	1	.000
	Model	28.026	1	.000

Table 10.40 Model Summary for the Forward Stepwise Logistic Regression

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	27.426	.504	.672

Table 10.41 Hosmer-Lemeshow Statistic for the Forward Stepwise Logistic Regression

Step	Chi-square	df	Sig.
1	5.616	8	.690

Table 10.42 Contingency Table for the Hosmer-Lemeshow Statistic for the Forward Stepwise Logistic Regression

		Predictability of the channel = high		Predictability of the channel = full		Total
		Observed	Expected	Observed	Expected	
Step 1	1	5	4.967	0	.033	5
	2	4	3.895	0	.105	4
	3	3	3.598	1	.402	4
	4	3	3.607	2	1.393	5
	5	2	1.400	1	1.600	3
	6	1	1.005	3	2.995	4
	7	2	.723	2	3.277	4
	8	0	.473	4	3.527	4
	9	0	.265	5	4.735	5
	10	0	.067	2	1.933	2

Table 10.43 Classification Table for the Forward Stepwise Logistic Regression

			Predicted		
			Predictability		
Observed			high	full	Percentage Correct
Step 1	Predictability	high	17	3	85.0
		full	3	17	85.0
	Overall Percentage				85.0

Table 10.44 Parameter Estimates for the Forward Stepwise Logistic Regression

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	DaysActive	-.284	.085	11.154	1	.001	.753
	Constant	3.645	1.104	10.891	1	.001	38.270

Table 10.45 Model if Term Removed for the Forward Stepwise Logistic Regression

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	DaysActive	-27.726	28.026	1	.000

Table 10.46 Variables Excluded from the Model in the Forward Stepwise Logistic Regression

			Score	df	Sig.
Step 1	Variables	SurvivalTime	3.103	1	.078
		AvgUserRetTime	.604	1	.437
		AvgDAilyMessagesPerPoster	2.789	1	.095
		AvgUserDiv	.210	1	.647
		AvgPosterDiv	.645	1	.422
		AvgDailyUserStability	.000	1	.985
		AvgDailyPosterStability	2.456	1	.117
		DaysVisited	3.109	1	.078

The value of the Hosmer-Lemeshow goodness-of-fit statistic in the final step (>.05) indicates that the model adequately fitted the data. The classification table shows that predictions were successful in 85 percent of the cases. The model correctly predicted that 17 out of 20 channels belonged to the "high predictability" category and 17 out of 20 channels belonged to the "perfect predictability" category. Only one independent variable was identified as the best predictor: *DaysActive*. The parameter estimates table shows that it was statistically significant (.001). The variable chosen by the model had significant changes in -2 log-likelihood; and the significance level of the Wald statistic was small enough to indicate that the parameter was useful to the model.

Tables 10.47 – 10.54 present the results of the backward stepwise logistic regression. Due to space considerations, some of the tables below report only the results produced by the final step of the model.

Table 10.47 Omnibus Tests of Model Coefficients for the Backward Stepwise Logistic Regression

		Chi-square	df	Sig.
Step 1	Step	41.403	8	.000
	Block	41.403	8	.000
	Model	41.403	8	.000
Step 2	Step	-.587	1	.444
	Block	40.816	7	.000
	Model	40.816	7	.000
Step 3	Step	-.690	1	.406
	Block	40.127	6	.000
	Model	40.127	6	.000
Step 4	Step	-.485	1	.486
	Block	39.642	5	.000
	Model	39.642	5	.000
Step 5	Step	-2.425	1	.119
	Block	37.217	4	.000
	Model	37.217	4	.000
Step 6	Step	-2.894	1	.089
	Block	34.323	3	.000
	Model	34.323	3	.000
Step 7	Step	-3.276	1	.070
	Block	31.048	2	.000
	Model	31.048	2	.000
Step 8	Step	-3.022	1	.082
	Block	28.026	1	.000
	Model	28.026	1	.000

Table 10.48 Model Summary for the Backward Stepwise Logistic Regression

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	14.049(a)	.645	.860
2	14.636(a)	.640	.853
3	15.325(a)	.633	.844
4	15.810(a)	.629	.838
5	18.234(b)	.606	.807
6	21.128(b)	.576	.768
7	24.404(c)	.540	.720
8	27.426(c)	.504	.672

Table 10.49 Hosmer-Lemeshow Statistic for the Backward Stepwise Logistic Regression

Step	Chi-square	df	Sig.
1	1.067	8	.998
2	12.736	8	.121
3	8.124	8	.421
4	11.802	8	.160
5	10.174	8	.253
6	9.017	8	.341
7	11.682	8	.166
8	8.794	3	.132

Table 10.50 Contingency Table for the Hosmer-Lemeshow Statistic for the Backward Stepwise Logistic Regression

		Predictability of the channel = high		Predictability of the channel = full		Total
		Observed	Expected	Observed	Expected	
Step 8	1	14	12.068	7	8.932	21
	2	0	2.675	5	2.325	5
	3	3	1.874	1	2.126	4
	4	1	1.938	4	3.062	5
	5	2	1.445	3	3.555	5

Table 10.51 Classification Table for the Backward Stepwise Logistic Regression

Observed			Predicted		
			Predictability		
			high	full	Percentage Correct
Step 1	Predictability of the channel	high	17	3	85.0
		full	1	19	95.0
	Overall Percentage				90.0
Step 2	Predictability of the channel	high	18	2	90.0
		full	1	19	95.0
	Overall Percentage				92.5
Step 3	Predictability of the channel	high	17	3	85.0
		full	1	19	95.0
	Overall Percentage				90.0
Step 4	Predictability of the channel	high	17	3	85.0
		full	1	19	95.0
	Overall Percentage				90.0
Step 5	Predictability of the channel	high	19	1	95.0
		full	2	18	90.0
	Overall Percentage				92.5
Step 6	Predictability of the channel	high	17	3	85.0
		full	2	18	90.0
	Overall Percentage				87.5
Step 7	Predictability of the channel	high	17	3	85.0
		full	3	17	85.0
	Overall Percentage				85.0
Step 8	Predictability of the channel	high	17	3	85.0
		full	3	17	85.0
	Overall Percentage				85.0

Table 10.52 Parameter Estimates for the Backward Stepwise Logistic Regression

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	SurvivalTime	.342	.203	2.842	1	.092	1.408
	AvgUserRetTime	.001	.001	1.557	1	.212	1.001
	AvgDailyMessagesPerPoster	-.070	.095	.550	1	.458	.932
	AvgUserDiv	-4.524	5.731	.623	1	.430	.011
	AvgPosterDiv	-7.570	6.520	1.348	1	.246	.001
	AvgDailyUserStability	11.701	13.089	.799	1	.371	120670.932
	AvgDailyPosterStability	-8.386	6.855	1.496	1	.221	.000
	DaysActive	-.454	.197	5.289	1	.021	.635
	Constant	4.955	4.271	1.346	1	.246	141.903
Step 2	SurvivalTime	.345	.188	3.351	1	.067	1.412
	AvgUserRetTime	.001	.001	2.734	1	.098	1.001
	AvgUserDiv	-4.469	5.646	.627	1	.429	.011
	AvgPosterDiv	-8.664	6.152	1.983	1	.159	.000
	AvgDailyUserStability	12.842	14.233	.814	1	.367	377899.192
	AvgDailyPosterStability	-10.881	7.268	2.241	1	.134	.000
	DaysActive	-.473	.209	5.123	1	.024	.623
	Constant	4.779	4.328	1.220	1	.269	118.999
Step 3	SurvivalTime	.287	.167	2.945	1	.086	1.333
	AvgUserRetTime	.001	.001	2.576	1	.109	1.001
	AvgPosterDiv	-8.833	5.738	2.370	1	.124	.000
	AvgDailyUserStability	6.686	9.921	.454	1	.500	801.266
	AvgDailyPosterStability	-8.827	5.878	2.255	1	.133	.000
	DaysActive	-.407	.163	6.243	1	.012	.665
	Constant	3.996	4.089	.955	1	.329	54.362
Step 4	SurvivalTime	.326	.171	3.652	1	.056	1.385
	AvgUserRetTime	.001	.001	3.326	1	.068	1.001
	AvgPosterDiv	-8.593	5.823	2.178	1	.140	.000
	AvgDailyPosterStability	-6.885	5.102	1.821	1	.177	.001
	DaysActive	-.406	.160	6.464	1	.011	.666
	Constant	3.861	4.275	.816	1	.366	47.533
Step 5	SurvivalTime	.301	.146	4.245	1	.039	1.352
	AvgUserRetTime	.001	.000	3.517	1	.061	1.001
	AvgPosterDiv	-6.302	4.395	2.057	1	.152	.002
	DaysActive	-.423	.169	6.270	1	.012	.655
	Constant	1.094	2.881	.144	1	.704	2.986
Step 6	SurvivalTime	.208	.101	4.224	1	.040	1.232
	AvgUserRetTime	.001	.000	3.006	1	.083	1.001
	DaysActive	-.437	.147	8.899	1	.003	.646
	Constant	-1.405	2.178	.416	1	.519	.245
Step 7	SurvivalTime	.120	.073	2.708	1	.100	1.127
	DaysActive	-.363	.111	10.636	1	.001	.695
	Constant	1.686	1.424	1.403	1	.236	5.400
Step 8	DaysActive	-.284	.085	11.154	1	.001	.753
	Constant	3.645	1.104	10.891	1	.001	38.270

Table 10.53 Model if Term Removed for the Backward Stepwise Logistic Regression

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 8	DaysActive	-27.726	28.026	1	.000

Table 10.54 Variables Excluded from the Model in the Backward Stepwise Logistic Regression

			Score	df	Sig.
Step 2	Variables	AvgDailyMessagesPerPoster	.568	1	.451
	Overall Statistics		.568	1	.451
Step 3	Variables	AvgDailyMessagesPerPoster	.597	1	.440
		AvgUserDiv	.672	1	.412
	Overall Statistics		1.126	2	.569
Step 4	Variables	AvgDailyMessagesPerPoster	.540	1	.462
		AvgUserDiv	.125	1	.723
		AvgDailyUserStability	.470	1	.493
	Overall Statistics		1.661	3	.646
Step 5	Variables	AvgDailyMessagesPerPoster	1.453	1	.228
		AvgUserDiv	.115	1	.735
		AvgDailyUserStability	.096	1	.756
		AvgDailyPosterStability	2.232	1	.135
	Overall Statistics		4.123	4	.390
Step 6	Variables	AvgDailyMessagesPerPoster	2.103	1	.147
		AvgUserDiv	.965	1	.326
		AvgPosterDiv	2.526	1	.112
		AvgDailyUserStability	.161	1	.688
		AvgDailyPosterStability	1.480	1	.224
	Overall Statistics		5.885	5	.318
Step 7	Variables	AvgUserRetTime	3.958	1	.047
		AvgDailyMessagesPerPoster	3.516	1	.061
		AvgUserDiv	.507	1	.476
		AvgPosterDiv	2.177	1	.140
		AvgDailyUserStability	.099	1	.753
		AvgDailyPosterStability	1.464	1	.226
Step 8	Variables	SurvivalTime	3.103	1	.078
		AvgUserRetTime	.604	1	.437
		AvgDailyMessagesPerPoster	3.405	1	.065
		AvgUserDiv	.210	1	.647
		AvgPosterDiv	.645	1	.422
		AvgDailyUserStability	.000	1	.985
		AvgDailyPosterStability	2.456	1	.117

The value of the Hosmer-Lemeshow goodness-of-fit statistic in the final step (>.05) indicates that the model adequately fitted the data. The classification table shows that predictions were successful in 85 percent of the cases. The model correctly predicted that 17 out of 20 channels belonged to the "high predictability" and 17 out of 20 channels belonged to the "perfect predictability" category. Only one independent variable was identified as the best predictor: *DaysActive*. The parameter estimates table shows that it was statistically significant (.001). The variable chosen by the model had significant changes in -2 log-likelihood; and the significance level of the Wald statistic was small enough to indicate that the parameter was useful to the model.

Since both the forward and backward methods produced the same results, one can be fairly confident that the chosen model was a good one. The independent variable *DaysActive* can be used to successfully categorize channels into "high predictability" and "perfect predictability" categories in 85 percent of the cases.

10.4 Summary

The aim of this chapter was to find the factors that can be used to distinguish highly predictable channels from unpredictable channels. Three categories of predictability were considered: high predictability, low predictability, and perfect predictability.

Of the 90 channels in the sample, 20 channels did not sustain public interactions at all during the first week of September 2005 (the time interval for which predictions were made). Both the linear and the nonlinear regression models described in Chapter 9 successfully predicted the value zero for the BestPredictor variable for all the cases when the number of observed posters was actually zero. Therefore, these 20 channels, which were perfectly predictable in the sense that no activity was predicted by the models in 100 percent of the cases, were included into the "perfect predictability" category. Individual correlation coefficients between the BestPredictor variables produced by the best nonlinear regression model and the ObservedPosters were computed for all the other 70 channels. After sorting them based on this coefficient, the top 20 channels were included in the "high predictability" category, while the bottom 20 channels were included in the "low predictability" category

Several descriptive statistics were computed for all the channels and were then entered as independent variables into three logistic regression models. These models

attempted to find which of the descriptive statistics would best differentiate high predictability channels from low-predictability channels, low-predictability channels from perfect predictability channels, and perfect-predictability channels from high-predictability channels. The logistic regression models revealed the following:

- When trying to determine whether the activity of a chat-channel during a particular week would have a high degree or a low degree of predictability, the best indicators were the number of days the channel was visited and the number of days the channel sustained public interactions during the previous month. The predictions were successful in 80 percent of the cases;

- When trying to determine whether the activity of a chat-channel during a particular week would have a low degree of predictability or would be perfectly predictable, the best indicator was the number of days the channel sustained public interactions during the previous month. The predictions were successful in 75 percent of the cases;

- When trying to determine whether the activity of a chat-channel during a particular week would have a high degree of predictability or would be perfectly predictable, the best indicator was the number of days the channel sustained public interactions during the previous month. The predictions were successful in 85 percent of the cases.

CHAPTER 11

IDENTIFICATION OF FACTORS THAT INFLUENCE CHANNEL SURVIVABILITY

To date, little is known about the initial conditions that lead to the formation of groups in synchronous spaces such as IRC channels, as well as about the subsequent conditions necessary for those groups to evolve and be sustained over longer periods of time. The notion of critical mass if often used when discussing the long-term sustainability of groups and the general consensus is that a group needs in its early stages a certain critical mass of members in order to become and remain successful over longer periods of time. While researchers have argued recently that critical mass is highly context dependent (Halverson, Erickson, and Sussman 2003; Grinter and Palen's 2002), very little empirical work has been done to explore this issue and its implications for long-term group survival.

Many authors consider critical mass a species of threshold model, in which a minimum number of contributors is necessary for a certain tipping point to be passed, leading thus to unanimous cooperation (Oliver and Marwell 2001). Others describe it simply as a group of people highly interested in a particular technology, leading the way to adopting it (Herbsleb et al. 2002). The Critical Mass theory (Oliver, Marwell, and Teixeira 1985) provides a more complex theoretical model. The definition of the term "critical mass" – a small segment of a group's population that behaves differently from the typical group members by making big contributions to the group's collective action while the majority of the group remains inactive – is not very different from previous ones. However, one of the theory's most important contributions is the argument that a

group's level of heterogeneity, together with the shapes of various production functions, defined as relationships between resources contributed by the group and the collective output of that group, can be used to distinguish between the likelihood of longer-term success of the group.

The driving research question for this chapter explores the Critical Mass theory's ability to help in predicting the long-term sustainability of groups in synchronous spaces. Specifically, it asks whether it is possible to predict IRC channels' chances of survival by looking at some of the initial starting conditions that characterize the overall activity of the channels, at the trajectories of the channel activity occurring inside over various time intervals in the initial stages of the channels' lives, at the population's level of heterogeneity during various time intervals and at the channels' production functions computed for the same time intervals.

In theory, survival analysis methods could be used to address this question. This chapter will examine whether it is possible to distinguish between the likelihood of IRC channels' survival over time based on variables extracted from the analysis of IRC channel interaction dynamics, on their heterogeneity of population, on their trajectories of activity, and on their production functions, all computed for four different time intervals.

11.1 Hypothesis

Considering the above, it is hypothesized that the long term survivability of any newly born publicly active channel can be predicted using four categories of factors: (1) the level of channel activity during various time intervals; (2) the trajectories of channel activity during various time intervals; (3) the heterogeneity of the channel's population during various time intervals; and (4) the type of production functions for various time intervals.

11.2 Method

11.2.1 Data Considerations

The analysis was performed on the set of IRC channels that were "born" during the month of July 2005. Two important notions needed to be considered beforehand: the birth of a channel and the death of a channel.

Because of the current lack of research in this area, no well-known definitions pertaining to these terms currently exist. Consequently, there was a need to clearly define them, from the perspective of this work. Therefore, a channel was considered "born" the first day when that channel hosted at least three posters who exchanged at least four public messages during the same 20-minute interval; and a channel was considered "dead" if four weeks of non activity have passed since the last day that channel hosted at least three posters who exchanged at least four public messages during the same 20-minute interval. A channel was considered to be non-active during a particular day if less than three posters were publicly active in that channel during all 20-minute intervals of that day. The main reason for defining the birth and death of chat-channels in terms of the

supported level of activity was because of the interest presented by this level of activity. Chapter 7 revealed that channels can easily be created and can exist for long periods of time after their creation without being visited at all before they would eventually disappear. Therefore, a channel's creation day and disappearance day may not be relevant indicators for the actual life of the channel. The definitions provided above are better suited for this research because they examine the life and death of a channel based on the presence or absence of public activity inside that channel, and not based simply on the presence or the absence of the channel itself on the IRC network.

11.2.2 Data Analysis

In order to explore the long term survivability of IRC channels, all the channels that were born during July 2005 were identified. Then, the lifetime of each channel was computed as the number of days between the birth and the death of that channel. A total of 282 channels were born during that month. Out of those channels, only 8 were still alive, according to the above definition, at the end of the data-collection period (January 31, 2006); the other 274 died at some point during the second half of the year for which data was collected.

Then, the aim was to understand how to distinguish the channels that survived from the channels that did not survive. To do so, Cox regression analysis was used.

Cox regression (sometimes called proportional hazards regression) is a method for investigating the effect of several variables upon the time a specified event takes to happen. In the context of an outcome such as death this is known as Cox regression for survival analysis. The method does not assume any particular "survival model" but it is

not truly non-parametric because it does assume that the effects of the predictor variables upon survival are constant over time and are additive in one scale.

Cox regression is used for modeling time-to-event data in the presence of censored cases (censored cases are cases for which the event of interest has not been recorded). However, as opposed to other time-to-event modeling methods such as the Kaplan-Meier survival analysis, it allows the inclusion of predictor variables (covariates) in the models. Cox regression will handle the censored cases correctly, and it will provide estimated coefficients for each of the covariates, allowing the assessment of the impact of multiple covariates in the same model.

In this research, the event of interest was the death of the channels, which was observed for 274 cases. Eight cases were censored – the ones corresponding to the channels that continued to be active after the end of the data-collection period.

Four Cox regression models were created, corresponding to four different time intervals for which the predictors of survivability were computed. These time intervals were (1) the first two hours of life; (2) the first day of life; (3) the first week of life; and (4) the first two weeks of life.

The objective was to determine whether the survival of channels can be predicted by looking at the initial starting conditions that characterized the overall activity of the channels; at the trajectories of the channel activity occurring inside them; at the level of heterogeneity of the channels' populations; and at the channels' production functions, computed for each of the four time intervals mentioned above.

Table 11.1 describes the variables entered into each Cox regression model. The number of users, posters, lurkers and messages measured the overall channel activity; the

posters trajectory (PT) and messages trajectory (MT) variables measured the trajectories of channel activity; and the poster diversity (PosterDiv) variable measured the heterogeneity of the channel poster population. The dependent variable was the lifespan of the channels, computed as the number of days between the birth and the death.

The possible values of the PT and message trajectory MT variables ranged from -1 to 1 and they indicated how the number of posters and the number of messages varied over time, during the various time intervals for which they were computed. For example, a value of 1 in a channel's poster trajectory measure for its first two hours of life would indicate that the number of posters for that channel continuously increased with every 20-minute interval since the channel's birth. As another example, a value of -1 in a channel's MT variable for its first day of life would indicate that the number of messages for that channel continuously decreased with every hour that passed since the channel's birth. The PT and MT variables were computed for each channel as the Spearman correlation coefficients between time and the number of posters or messages observed in that channel for each of the four time intervals. Each interval had a different number of data points that were used in computing the correlation coefficients. The first two hours of a channel's life had six data points for which the number of posters and messages were computed, each corresponding to a 20-minute interval. The first day of a channel's life had 24 data points, each corresponding to an hour; the first week of a channel's life had 7 data points; and the first two weeks of a channel's life had 14 data points, each corresponding to a day. The time was expressed as the number of seconds that have elapsed since midnight Coordinated Universal Time of January 1, 1970 until the starting time of the data point interval.

The possible values for the PosterDiv variables ranged from 1 to 100 and they indicated how heterogeneous or homogeneous the poster population of a channel was during a particular time interval, with respect to a larger time interval. A channel was considered more homogeneous if its poster population stayed relatively constant as time passed, and more heterogeneous if its poster population changed significantly over time. The maximum value of 100 indicates a fully homogeneous population, while the minimum value of 1 indicates a population with the highest level of heterogeneity.

The poster diversity for the first two hours of life was computed as the percentage value represented by the number of posters present in the channel during this interval reported to the total number of posters that visited the channel during its first day of life. The poster diversity for the first day of life was computed as the percentage value represented by the number of posters present in the channel during this interval, reported to the total number of posters that visited the channel during its first week of life. The poster diversity for the first week and the first two weeks of life was computed as the percentage value represented by the number of posters present in the channel during those intervals, reported to the total number of posters that visited the channel during its first month of life. For example, consider a channel that had 3 posters during its first two hours, 10 posters during its first day, 20 posters in its first week, 25 posters during its first two weeks and 30 posters during its first month. In this case, *PosterDiv2Hrs* = 3/10 = 30%, *PosterDivFirstDay* = 10/20 = 50%, *PosterDivFirstWeek* = 20/30 = 66% and *PosterDivFirstTwoWeeks* = 25/30 = 83%. Here, the values of the computed diversity variables show that initially the population was more heterogeneous, but with the passage of time it became more homogeneous.

Table 11.1 Predictor Variables for the Cox Regression Models

Model	Variables	Description
1 First two hours of life	UsersFirst2Hrs	Total number of users during the first two hours of life
	PostersFirst2Hrs	Total number of posters during the first two hours of life
	LurkersFirst2Hrs	Total number of lurkers (non-posters) during the first two hours of life
	MessagesFirst2Hrs	Total number of messages during the first two hours of life
	PosterDivFirst2Hrs	Poster diversity during the first two hours of life, computed with respect to the first day of life
	PTFirst2Hrs	Posters trajectory during the first two hours of life
	MTFirst2Hrs	Messages trajectory during the first two hours of life
	PFFirst2Hrs	Type of production function during the first two hours of life
2 First day of life	UsersFirstDay	Total number of users during the first day of life
	PostersFirstDay	Total number of posters during the first day of life
	LurkersFirstDay	Total number of lurkers (non-posters) during the first day of life
	MessagesFirstDay	Total number of messages during the first day of life
	PosterDivFirstDay	Poster diversity during the first day of life, computed with respect to the first week of life
	PTFirstDay	Posters trajectory during the first day of life
	MTFirstDay	Messages trajectory during the first day of life
	PFFirstDay	Type of production function during the first day of life
3 First week of life	UsersFirstWeek	Total number of users during the first week of life
	PostersFirstWeek	Total number of posters during the first week of life
	LurkersFirstWeek	Total number of lurkers (non-posters) during the first week of life
	MessagesFirstWeek	Total number of messages during the first week of life
	PosterDivFirstWeek	Poster diversity during the first week of life, computed with respect to the first month of life
	PTFirstWeek	Posters trajectory during the first week of life
	MTFirstWeek	Messages trajectory during the first week of life
	PFFirstWeek	Type of production function during the first week of life
4 First two weeks of life	UsersFirst2Weeks	Total number of users during the first two weeks of life
	PostersFirst2Weeks	Total number of posters during the first two weeks of life
	LurkersFirst2Weeks	Total number of lurkers (non-posters) during the first two weeks of life
	MessagesFirst2Weeks	Total number of messages during the first two weeks of life
	PosterDivFirst2Weeks	Poster diversity during the first two weeks of life, compute with respect to the first month of life
	PTFirst2Weeks	Posters trajectory during the first two weeks of life
	MTFirst2Weeks	Messages trajectory during the first two weeks of life
	PFFirst2Weeks	Type of production function during the first two weeks of life

The production functions were computed based on the definition provided by the Critical Mass theory. The theory defined production functions as the relationships between resources contributed by a group and the collective output of that group, and argued that they can be used to distinguish among the likelihood of longer-term group

success. Based on the shape of the graphs obtained by plotting the number of resources by the amount of group success, the Critical Mass theory described two types of production functions: accelerating and decelerating.

In the case of IRC chat-channels, the number of users present in the channel was considered a surrogate measure for the group resources, while the number of messages was considered a surrogate measure for the amount of group success achieved. Twelve categories of production functions were identified after plotting the number of users by the number of messages, for each of the 282 channels. These twelve categories are presented in Table 11.2 and their corresponding shapes are described in Figures 11.1 (a – l).

Table 11.2 Categories of Production Functions

Category	Shape of the production function
0	Constant
1	Linear ascending
2	Linear descending
3	Accelerating ascending
4	Decelerating ascending
5	S-shaped ascending
6	Accelerating descending
7	Decelerating descending
8	S-shaped descending
9	Parabola
10	Inverse parabola
11	Variable/Unidentified

For each channel, the shape of the production functions for all the four intervals was determined by plotting the number of users by the number of messages for that channel, using the same data points that were used to compute the trajectory measures described above.

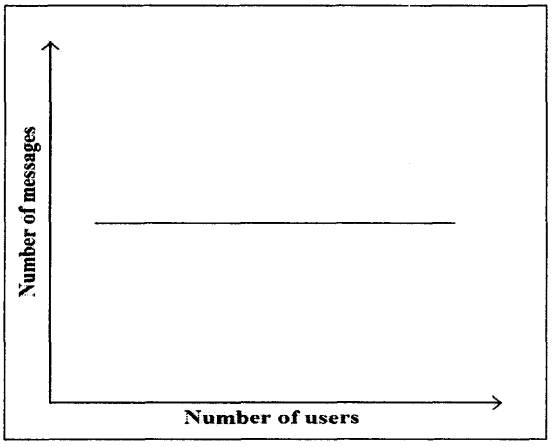

Figure 11.1 a) Constant production function – Category 0.

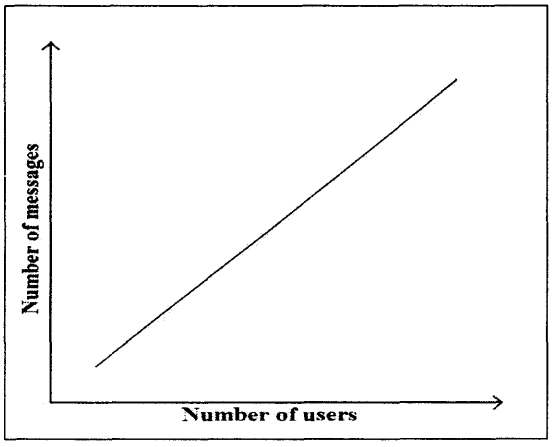

Figure 11.1 b) Linear ascending production function – Category 1.

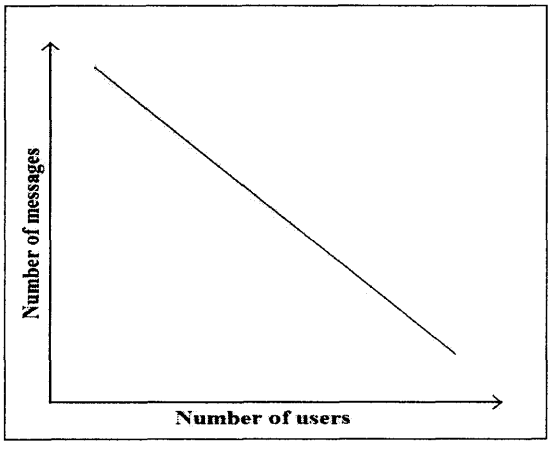

Figure 11.1 c) Linear descending production function – Category 2.

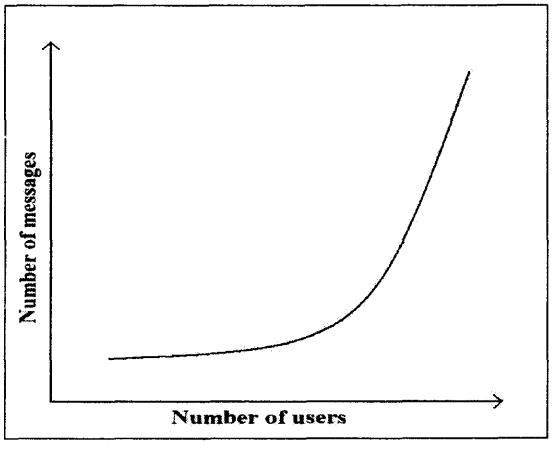

Figure 11.1 d) Accelerating ascending production function – Category 3.

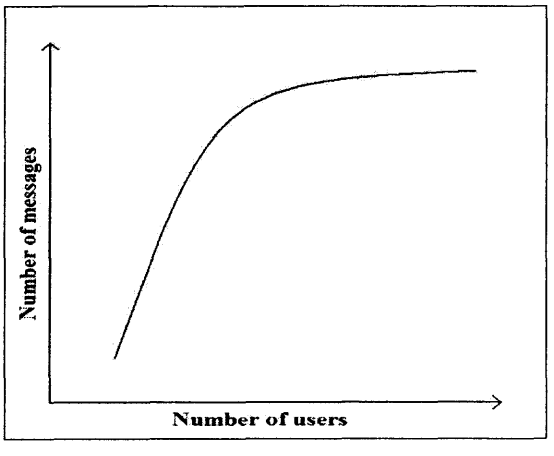

Figure 11.1 e) Decelerating ascending production function – Category 4.

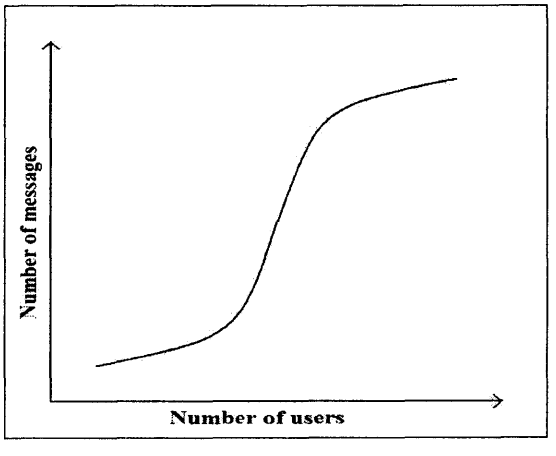

Figure 11.1 f) S-shaped ascending production function – Category 5.

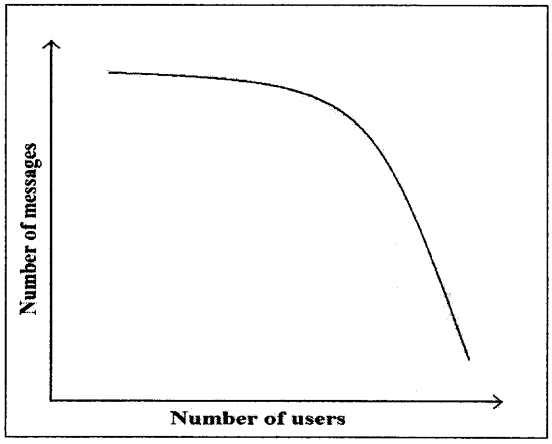

Figure 11.1 g) Accelerating descending production function – Category 6.

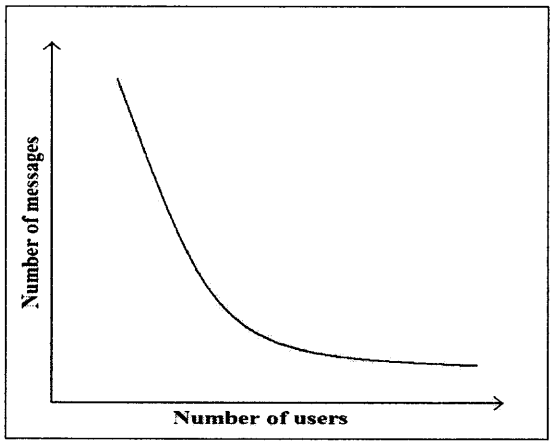

Figure 11.1 h) Decelerating descending production function – Category 7.

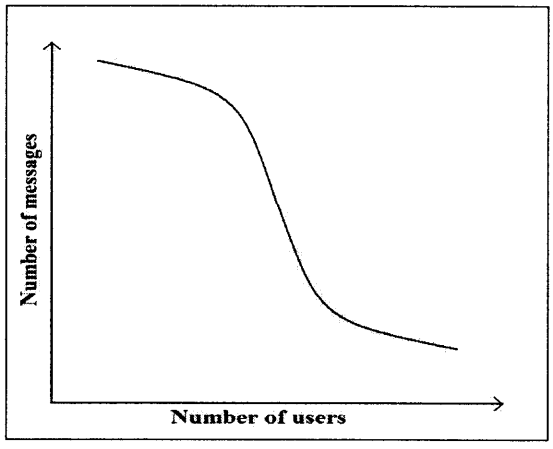

Figure 11.1 i) S-shaped descending production function – Category 8.

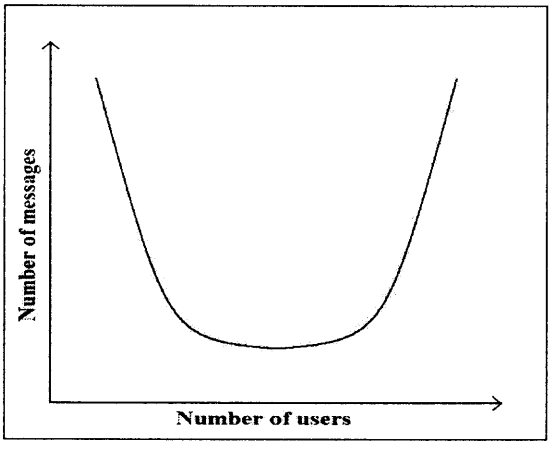

Figure 11.1 j) Parabola production function – Category 9.

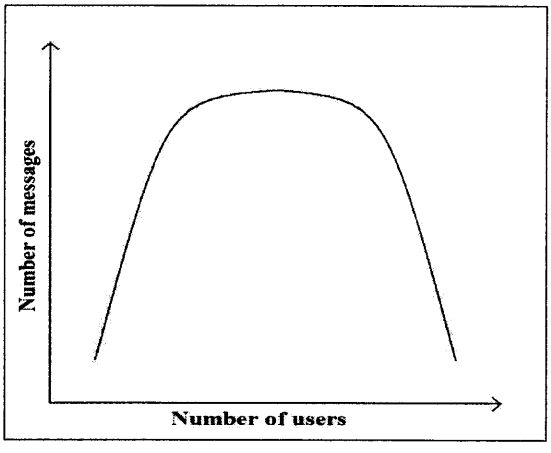

Figure 11.1 k) Inverse parabola production function – Category 10.

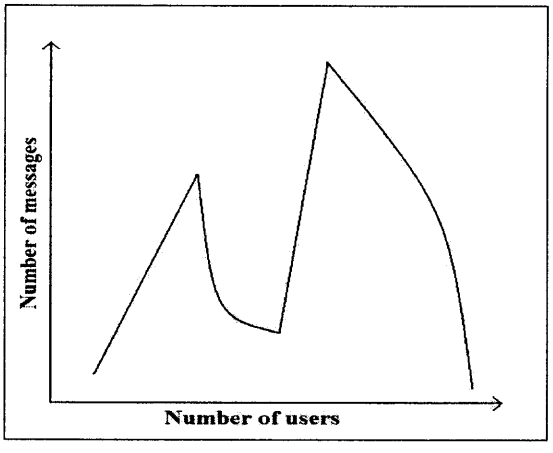

Figure 11.1 l) Variable/unidentified production function – Category 11.

Table 11.3 reports the number of channels whose production functions had one of the corresponding twelve shapes described in Figure 11.1, for each time interval.

Table 11.3 Number of Channels per Categories of Production Functions

Interval	Cat0	Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7	Cat8	Cat9	Cat10	Cat11
First two hours	39	64	42	46	16	10	10	4	2	8	17	24
First day	26	55	7	84	11	14	4	3	0	4	15	59
First week	48	63	15	99	1	8	1	0	1	2	8	36
First two weeks	38	44	11	108	3	11	2	0	0	1	4	60

It can be observed that some of the production function categories identified in the manner described above were not very common. In order to make the analysis easier and more relevant, the twelve categories of production functions were grouped into four broader types: constant, ascending, descending, and variable. Table 11.4 describes these four types in terms of the categories they included, while Table 11.5 reports the number of channels characterized by each type of production function during each of the analyzed time intervals.

Table 11.4 Broader Types of Production Functions

Type	Included categories
Constant	Constant
Ascending	Linear ascending, accelerating ascending, decelerating ascending, S-shaped ascending
Descending	Linear descending, accelerating descending, decelerating descending, S-shaped descending
Variable	Parabola, inverse parabola, variable/unidentified

Table 11.5 Number of Channels per Types of Production Functions

Interval	Constant	Ascending	Descending	Variable
First two hours	39	136	58	49
First day	26	164	14	78
First week	48	171	17	46
First two weeks	38	166	13	65

11.3 Results

The model-building process took place in two blocks. In the first block, a forward stepwise algorithm was employed and the following variables were entered: the number of users, the number of posters, the number of lurkers, the number of messages, the poster diversity, the posters trajectory, and the messages trajectory. In the second block, the categorical variable used to represent the type of production function was added to the model (see Table 11.1 for the exact names of the variables used in the four regression models corresponding to each time interval).

The basic model offered by the Cox regression procedure is the proportional hazards model, which assumes that the time to event and the covariates are related through a particular equation. The hazard function is a measure of the potential for the event to occur at a particular time *t*, given that the event did not yet occur. Larger values of the hazard function indicate greater potential for the event to occur. The baseline hazard function measures this potential independently of the covariates. The shape of the hazard function over time is defined by the baseline hazard, for all cases. The covariates simply help to determine the overall magnitude of the function

First, some descriptive statistics of the variables used by the survival analysis are reported. Then, the following results produced by the Cox regression procedure available in the SPSS software are presented, for each of the analyzed intervals:

- The omnibus tests of model coefficients table – measures of how well the model performed (for both blocks)
- The final variables that were used in the regression equation
- The final variables that were not used in the regression equation
- The covariate means and pattern values

- The plot of the basic survival curve – a visual display of the cumulated model-predicted time to death for the "average" channel

- The plot of the basic hazard curve – a visual display of the cumulative model-predicted potential to die for the "average" channel

- The plot of the survival curves for each covariate pattern – a visual representation of the effect of the production function type categorical variable on the channels' survival

- The plot of the hazard curves for each covariate pattern – a visual representation of the effect of the production function type categorical variable on the channels' potential to die

Table 11.6 reports the case processing summary for all four regression models and it shows that 8 cases of the total of 282 were censored. These cases represented the channels that did not die. They were not used in the computation of the regression coefficients, but were used in the computation of the baseline hazard.

Table 11.6 Case Processing Summary

		N	Percent
Cases available in analysis	Event(a)	274	97.2%
	Censored	8	2.8%
	Total	282	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		282	100.0%

a Dependent Variable: Lifespan

11.3.1 Descriptive Statistics

Table 11.7 reports descriptive statistics for most of the variable entered into the four Cox regression models.

Table 11.7 Descriptive Statistics for the Variables in the Cox Regression Models

	Mean	Median	Mode	Min	Max	Percentiles		
						25%	50%	75%
Lifespan	17	1	1	1	203	1	1	15
UsersFirst2Hrs	9	5	3	3	113	3	5	8
PostersFirst2Hrs	6	4	3	3	36	3	4	7
MessagesFirst2Hrs	123	66	4	4	1001	22	66	140
LurkersFirst2Hrs	4	1	0	0	81	0	1	3
PosterDivFirst2Hrs	86	100	100	8	100	75	100	100
UsersFirstDay	15	6	3	3	293	4	6	12
PostersFirstDay	8	4	3	3	87	3	4	8
MessagesFirstDay	217	96	4	4	2202	28	96	234
LurkersFirstday	8	2	0	0	206	0	2	5
PosterDivFirstDay	82	100	100	8	100	67	100	100
UsersFirstWeek	27	10	4	3	683	5	10	23
PostersFirstWeek	13	6	4	3	184	4	6	13
MessagesFirstWeek	410	132	4	4	6249	43	132	413
LurkersFirstWeek	15	3	0	0	499	1	3	9
PosterDivFirstWeek	93	100	100	18	100	94	100	100
UsersFirst2Weeks	33	11	3	3	799	6	11	29
PostersFirst2Weeks	15	6	3	3	217	3	6	18
MessagesFirst2Weeks	539	135	4	4	8145	46	135	474
LurkersFirst2Weeks	19	5	0	0	582	1	5	12
PosterDivFirst2Weeks	92	100	100	25	100	88	100	100

It can easily be observed that half of the new channels that appeared in July 2005 did not last more than a day and had very few users, posters, lurkers, and messages. It is likely that such channels were created by very small groups of users who decided to get together for short periods of time to discuss something in a more private environment, rather than in the open spaces of other already existing channels. The channels disappeared after those discussions were resolved and the users left. It may also be noted that a vast number of channels had more homogeneous populations, rather than

heterogeneous, for all four intervals. In part this was to be expected, considering that half the new channels lasted for only one day, but it might also suggest that a newly born channel has difficulties in diversifying its population during the first month of life.

Tables 11.8 – 11.11 report the Spearman correlation coefficients among the variables used in each Cox regression model. Interestingly, the *Lifespan* variable, which was the dependent variable, was negatively correlated with the *PosterDiversity* during all the four intervals, and the correlation coefficients were quite high. This shows that channels that survived longer were likely to be more heterogeneous than channels that survived for shorter periods. Although correlation does not imply causation, this relationship is worth exploring further.

The correlation coefficients between the lifespan of channels and the number of users, posters, and messages were less than 0.5 during the first two hours and the first day, and grew above this value for the first week and the first two weeks.

High correlations were observed in each time interval between the number of users and the number of posters, and between the number of posters and the number of messages. This was to be expected as it makes sense to assume that the more users visit a channel, the more likely to have more posters in that channel, or that the more posters become active in a channel, the more messages they will send to the public interaction space.

Table 11.8 Correlation Coefficients for the First Two Hours

			Number of users during the first 2 hours of life	Number of posters during the first 2 hours of life	Number of lurkers during the first 2 hours of life	Number of messages during the first 2 hours of life	Poster diversity	Lifespan
Spearman's rho	Number of users during the first 2 hours of life	Correlation Coefficient	1.000	.691(**)	.555(**)	.390(**)	-.257(**)	.212(**)
		Sig. (2-tailed)	.	.000	.000	.000	.000	.000
		N	282	282	282	282	282	282
	Number of posters during the first 2 hours of life	Correlation Coefficient	.691(**)	1.000	.119(*)	.596(**)	-.266(**)	.277(**)
		Sig. (2-tailed)	.000	.	.046	.000	.000	.000
		N	282	282	282	282	282	282
	Number of lurkers during the first 2 hours of life	Correlation Coefficient	.555(**)	.119(*)	1.000	-.113	-.140(*)	.104
		Sig. (2-tailed)	.000	.046	.	.058	.019	.083
		N	282	282	282	282	282	282
	Number of messages during the first 2 hours of life	Correlation Coefficient	.390(**)	.596(**)	-.113	1.000	-.156(**)	.175(**)
		Sig. (2-tailed)	.000	.000	.058	.	.009	.003
		N	282	282	282	282	282	282
	Poster diversity	Correlation Coefficient	-.257(**)	-.266(**)	-.140(*)	-.156(**)	1.000	-.476(**)
		Sig. (2-tailed)	.000	.000	.019	.009	.	.000
		N	282	282	282	282	282	282
	Lifespan	Correlation Coefficient	.212(**)	.277(**)	.104	.175(**)	-.476(**)	1.000
		Sig. (2-tailed)	.000	.000	.083	.003	.000	.
		N	282	282	282	282	282	282

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Table 11.9 Correlation Coefficients for the First Day

			Number of users during the first day of life	Number of posters during the first day of life	Number of lurkers during the first day of life	Number of messages during the first day of life	Poster diversity	Lifespan
Spearman's rho	Number of users during the first day of life	Correlation Coefficient	1.000	.769(**)	.712(**)	.479(**)	-.341(**)	.355(**)
		Sig. (2-tailed)	.	.000	.000	.000	.000	.000
		N	282	282	282	282	282	282
	Number of posters during the first day of life	Correlation Coefficient	.769(**)	1.000	.289(**)	.663(**)	-.341(**)	.425(**)
		Sig. (2-tailed)	.000	.	.000	.000	.000	.000
		N	282	282	282	282	282	282
	Number of lurkers during the first day of life	Correlation Coefficient	.712(**)	.289(**)	1.000	.083	-.241(**)	.194(**)
		Sig. (2-tailed)	.000	.000	.	.166	.000	.001
		N	282	282	282	282	282	282
	Number of messages during the first day of life	Correlation Coefficient	.479(**)	.663(**)	.083	1.000	-.245(**)	.354(**)
		Sig. (2-tailed)	.000	.000	.166	.	.000	.000
		N	282	282	282	282	282	282
	Poster diversity	Correlation Coefficient	-.341(**)	-.341(**)	-.241(**)	-.245(**)	1.000	-.658(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.	.000
		N	282	282	282	282	282	282
	Lifespan	Correlation Coefficient	.355(**)	.425(**)	.194(**)	.354(**)	-.658(**)	1.000
		Sig. (2-tailed)	.000	.000	.001	.000	.000	.
		N	282	282	282	282	282	282

** Correlation is significant at the 0.01 level (2-tailed).

Table 11.10 Correlation Coefficients for the First Week

			Number of users during the first week of life	Number of posters during the first week of life	Number of lurkers during the first week of life	Number of messages during the first week of life	Poster diversity	Lifespan
Spearman's rho	Number of users during the first week of life	Correlation Coefficient	1.000	.828(**)	.837(**)	.606(**)	-.362(**)	.504(**)
		Sig. (2-tailed)	.	.000	.000	.000	.000	.000
		N	282	282	282	282	282	282
	Number of posters during the first week of life	Correlation Coefficient	.828(**)	1.000	.485(**)	.766(**)	-.384(**)	.575(**)
		Sig. (2-tailed)	.000	.	.000	.000	.000	.000
		N	282	282	282	282	282	282
	Number of lurkers during the first week of life	Correlation Coefficient	.837(**)	.485(**)	1.000	.310(**)	-.284(**)	.320(**)
		Sig. (2-tailed)	.000	.000	.	.000	.000	.000
		N	282	282	282	282	282	282
	Number of messages during the first week of life	Correlation Coefficient	.606(**)	.766(**)	.310(**)	1.000	-.326(**)	.555(**)
		Sig. (2-tailed)	.000	.000	.000	.	.000	.000
		N	282	282	282	282	282	282
	Poster diversity	Correlation Coefficient	-.362(**)	-.384(**)	-.284(**)	-.326(**)	1.000	-.602(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.	.000
		N	282	282	282	282	282	282
	Lifespan	Correlation Coefficient	.504(**)	.575(**)	.320(**)	.555(**)	-.602(**)	1.000
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.
		N	282	282	282	282	282	282

** Correlation is significant at the 0.01 level (2-tailed).

Table 11.11 Correlation Coefficients for the First Two Weeks

			Number of users during the first two weeks of life	Number of posters during the first two weeks of life	Number of lurkers during the first two weeks of life	Number of messages during the first two weeks of life	Poster diversity	Lifespan
Spearman's rho	Number of users during the first two weeks of life	Correlation Coefficient	1.000	.845(**)	.820(**)	.618(**)	-.282(**)	.562(**)
		Sig. (2-tailed)	.	.000	.000	.000	.000	.000
		N	282	282	282	282	282	282
	Number of posters during the first two weeks of life	Correlation Coefficient	.845(**)	1.000	.483(**)	.776(**)	-.274(**)	.647(**)
		Sig. (2-tailed)	.000	.	.000	.000	.000	.000
		N	282	282	282	282	282	282
	Number of lurkers during the first two weeks of life	Correlation Coefficient	.820(**)	.483(**)	1.000	.294(**)	-.177(**)	.315(**)
		Sig. (2-tailed)	.000	.000	.	.000	.003	.000
		N	282	282	282	282	282	282
	Number of messages during the first two weeks of life	Correlation Coefficient	.618(**)	.776(**)	.294(**)	1.000	-.217(**)	.623(**)
		Sig. (2-tailed)	.000	.000	.000	.	.000	.000
		N	282	282	282	282	282	282
	Poster diversity	Correlation Coefficient	-.282(**)	-.274(**)	-.177(**)	-.217(**)	1.000	-.545(**)
		Sig. (2-tailed)	.000	.000	.003	.000	.	.000
		N	282	282	282	282	282	282
	Lifespan	Correlation Coefficient	.562(**)	.647(**)	.315(**)	.623(**)	-.545(**)	1.000
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.
		N	282	282	282	282	282	282

** Correlation is significant at the 0.01 level (2-tailed).

11.3.2 Cox Regression Results for the First Two Hours of Life

Table 11.12 Omnibus Tests of Model Coefficients for the First Block

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1	2680.694	21.886	1	.000	24.172	1	.000	24.172	1	.000

Variable(s) Entered at Step Number 1: PosterDivFirst2Hrs1

Table 11.13 Omnibus Tests of Model Coefficients for the Second Block

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
2678.354	24.279	4	.000	2.340	3	.505	2.340	3	.505

Table 11.14 Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
PosterDivFirst2Hrs	.013	.003	18.403	1	.000	1.013
PFFirst2Hrs			2.273	3	.518	
PFFirst2Hrs1)	-.093	.184	.256	1	.613	.911
PFFirst2Hrs(2)	-.130	.211	.382	1	.537	.878
PFFirst2Hrs(3)	-.316	.224	1.983	1	.159	.729

Table 11.15 Variables not in the Equation

	Score	df	Sig.
PostersFirst2Hrs	.352	1	.553
UsersFirst2Hrs	.603	1	.438
LurkersFirst2Hrs	.614	1	.433
MessagesFirst2Hrs	.980	1	.322
PTFirst2Hrs	1.211	1	.271
MTFirst2Hrs	1.723	1	.189

Table 11.16 Covariate Means and Pattern Values

	Mean	Pattern			
		1	2	3	4
PostersFirst2Hrs	5.773	5.773	5.773	5.773	5.773
UsersFirst2Hrs	11.227	11.227	11.227	11.227	11.227
LurkersFirst2Hrs	6.394	6.394	6.394	6.394	6.394
MessagesFirst2Hrs	122.752	122.752	122.752	122.752	122.752
PosterDivFirst2Hrs	86.060	86.060	86.060	86.060	86.060
PTFirst2Hrs	-.500	-.500	-.500	-.500	-.500
MTFirst2Hrs	-.440	-.440	-.440	-.440	-.440
PFFirst2Hrs(1)	.482	.000	1.000	.000	.000
PFFirst2Hrs(2)	.206	.000	.000	1.000	.000
PFFirst2Hrs(3)	.174	.000	.000	.000	1.000

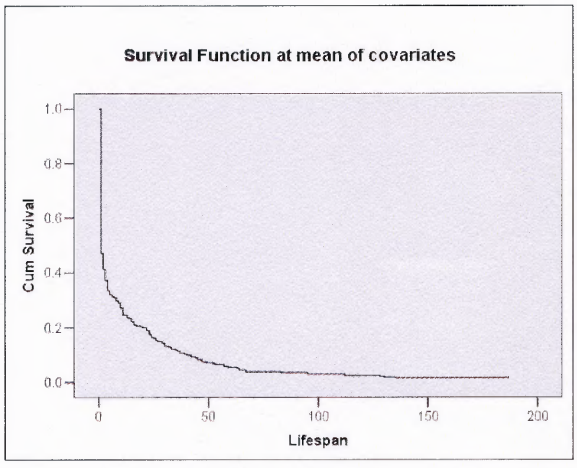

Figure 11.2 a) Survival curve for the first two hours of life.

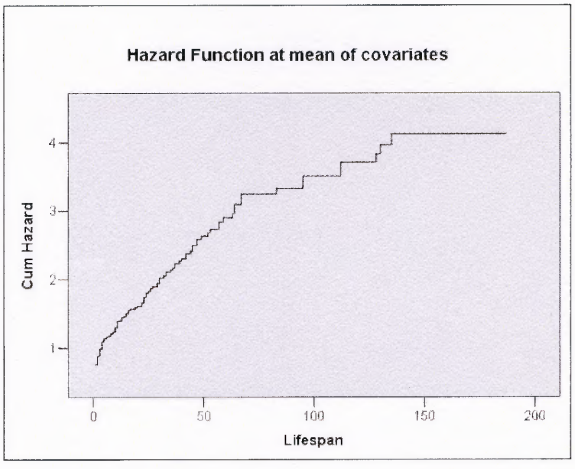

Figure 11.2 b) Hazard curve for the first two hours of life.

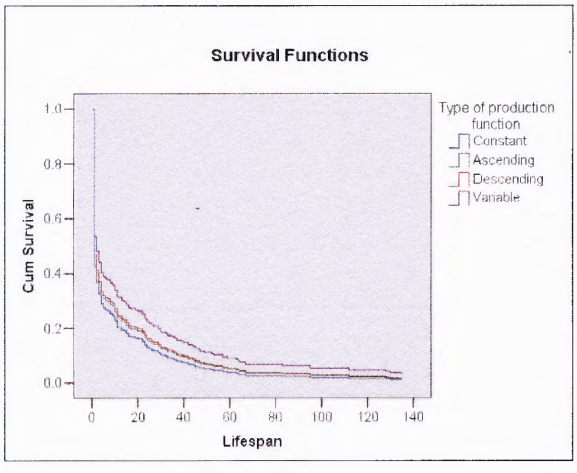

Figure 11.2 c) Survival curves for types of production functions during the first two hours of life.

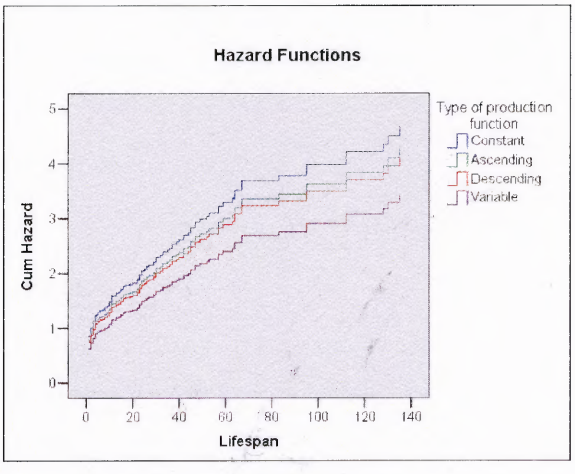

Figure 11.2 d) Hazard curves for types of production functions during the first two hours of life.

Table 11.12 shows that from all the variables entered into the first block of the regression model, only the poster diversity contributed significantly to it (significance of change less than 0.05). Table 11.13 shows that the addition of the type of production function to the model as a categorical variable did not contribute to the model (significance of change larger than 0.05). Table 11.14 shows the regression coefficients for the variables used in the final step of the model. Exp(B) represents the predicted change in the hazard for a unit increase in the predictor. In this case, the value of Exp(B)

for *PosterDivFirst2Hrs* means that the channels' death hazard increases by (100% * 1.013) – 100% = 1.3% for each increase of 1% in the diversity of the channel poster population. The death hazard for a channel whose diversity measure witnesses a raise of 10% is increased by (100% * (1.013^10)) – 100% = 13.78%. In other words, the larger the value of the poster diversity measure, the higher the risk of death (remember that a diversity value of 100 signifies a fully homogenous channel while lower values imply greater heterogeneity).

The regression coefficients for the first three levels of *PFFirst2Hrs* were relative to the reference category, which corresponded to the *Constant* production function type. The regression coefficient for the first category, corresponding to channels with *Ascending* production functions, suggests that the hazard for this type of channels is 0.911 times that of channels characterized by *Constant* production functions. Similarly, the hazard for channels with *Descending* production functions is 0.878 times the hazard of channels with *Constant* production functions, and the hazard for channels with *Variable* production functions is 0.729 times the hazard of channels with *Constant* production functions. However, the significance values for all these coefficients are greater than 0.10, so any observed differences between these channels categories could be due to chance. Table 11.15 shows that the variables left out of the model all had score statistics with significance values greater than 0.05.

Table 11.16 describes the four covariate patterns that correspond to the types of production functions, each with otherwise "average" covariates. This table is a useful reference when looking at the survival plots, which are constructed for the mean values and each covariate pattern. Note, however, that the "average" channel doesn't actually

exist when looking at the means of indicator variables for categorical predictors. Even with all scale predictors, it is unlikely to find a channel whose covariate values are all close to the mean.

The basic survival curve shown in Figure 11.2 a) is a visual display of the model-predicted time to death for the "average" channel. The horizontal axis shows the time to event (the lifespan of the channel), while the vertical axis shows the probability of survival. Thus, any point on the survival curve shows the probability that the "average" channel will stay alive past that time. Past 65 days, the survival curve becomes less smooth. There are fewer channels who have survived for that long, so there is less information available, and thus the curve is blocky. The plot of the survival curves for each covariate pattern shown in Figure 11.2 c) shows the effect of the "production function type" category. Although not statistically significant, as explained above, the channels with *Variable* and *Descending* production functions seemed to have been likely to survive longer than channels with *Ascending* production functions, while channels with *Constant* production functions seemed to have had the lowest chance of survival.

The basic hazard curve shown in Figure 11.2 b) is a visual display of the cumulative model-predicted potential to die for the "average" channel. The vertical axis shows the cumulative hazard, equal to the negative log of the survival probability. Past 65 days, the hazard curve, like the survival curve, becomes less smooth, for the same reason. The plot of the hazard curves for each covariate pattern described in Figure 11.2 d) shows the effect of the "production function type" category. Channels with *Constant* and *Ascending* production functions had higher hazard curve because they had a greater potential to not survive.

11.3.3 Cox Regression Results for the First Day of Life

Table 11.17 Omnibus Tests of Model Coefficients for the First Block

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1	2663.338	37.266	1	.000	41.528	1	.000	41.528	1	.000
2	2653.221	46.898	2	.000	10.117	1	.001	51.645	2	.000

Variable(s) Entered at Step Number 1: PosterDivFirstDay
Variable(s) Entered at Step Number 2: MessagesFirstDay

Table 11.18 Omnibus Tests of Model Coefficients for the Second Block

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
2648.997	52.013	5	.000	4.224	3	.238	4.224	3	.238

Table 11.19 Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
MessagesFirstDay	-.001	.000	7.421	1	.006	.999
PosterDivFirstDay	.015	.003	27.601	1	.000	1.015
PFFirstDay			4.310	3	.230	
PFFirstDay(1)	-.275	.218	1.583	1	.208	.760
PFFirstDay(2)	-.516	.341	2.297	1	.130	.597
PFFirstDay(3)	-.450	.240	3.508	1	.061	.638

Table 11.20 Variables not in the Equation

	Score	df	Sig.
PostersFirstDay	.101	1	.751
UsersFirstDay	.322	1	.571
LurkersFirstDay	.430	1	.512
PTFirstDay	1.482	1	.223
MTFirstDay	2.294	1	.130

Table 11.21 Covariate Means and Pattern Values

	Mean	Pattern			
		1	2	3	4
PostersFirstDay	8.209	8.209	8.209	8.209	8.209
UsersFirstDay	17.876	17.876	17.876	17.876	17.876
LurkersFirstDay	10.280	10.280	10.280	10.280	10.280
FirstDay	216.975	216.975	216.975	216.975	216.975
PosterDivFirstDay	82.954	82.954	82.954	82.954	82.954
PTFirstDay	-.369	-.369	-.369	-.369	-.369
MTFirstDay	-.353	-.353	-.353	-.353	-.353
PFFirstDay(1)	.582	.000	1.000	.000	.000
PFFirstDay(2)	.050	.000	.000	1.000	.000
PFFirstDay(3)	.277	.000	.000	.000	1.000

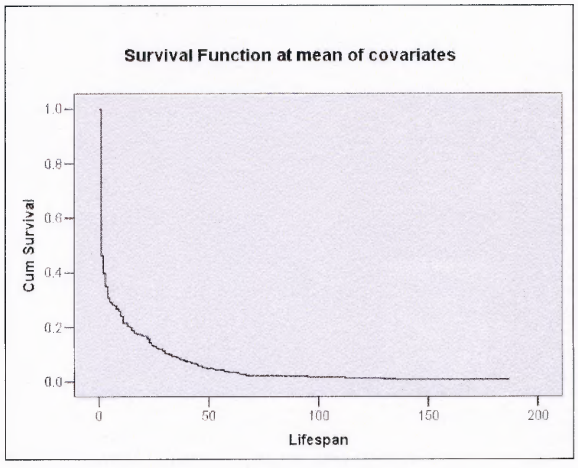

Figure 11.3 a) Survival curve for the first day of life.

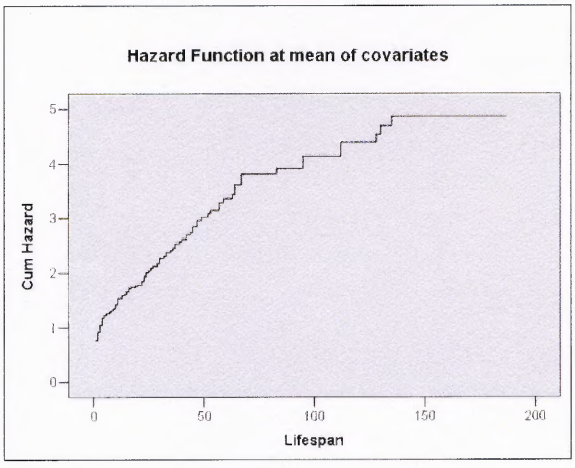

Figure 11.3 b) Hazard curve for the first day of life.

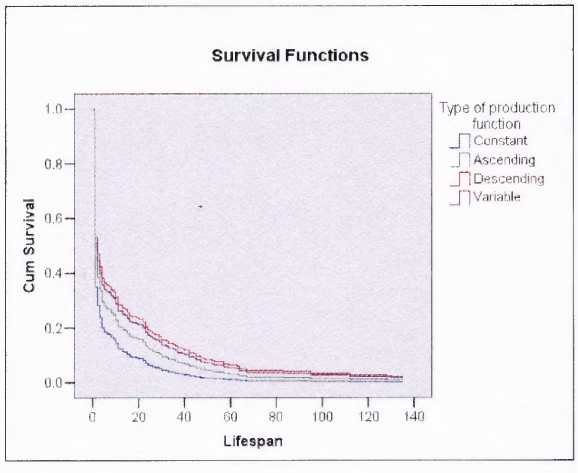

Figure 11.3 c) Survival curves for types of production functions during the first day of life.

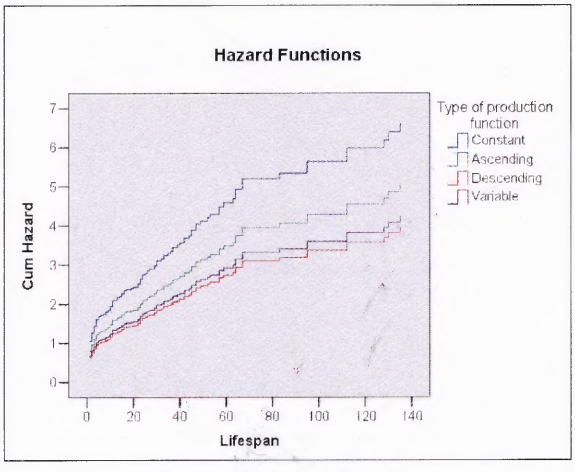

Figure 11.3 d) Hazard curves for types of production functions during the first day of life.

Table 11.17 shows that from all the variables entered into the first block of the regression model, only the poster diversity and the number of messages contributed significantly to it (significance of change less than 0.05). Table 11.18 shows that the addition of the type of production function to the model as a categorical variable did not contribute to the model (significance of change larger than 0.05). Table 11.19 shows the regression coefficients for the variables used in the final step of the model. Exp(B) represents the predicted change in the hazard for a unit increase in the predictor. The

value of Exp(B) for *PosterDivFirstDay* means that the channels' death hazard increases by (100% * 1.015) – 100% = 1.5% for each increase of 1% in the diversity of the channel poster population. The value of Exp(B) for *MessagesFirstDay* means that the channels' death hazard decreases by 100% - (100% * .999) = 0.1% for every new message. The death hazard for a channel which would have 100 more messages in its first day of life would decrease by 100% - (100% * (0.999^100)) = 9.5%. In other words, the larger the value of the poster diversity measure, the higher the risk of death, and the more messages, the lower the risk of death.

The regression coefficients for the first three levels of *PFFirstDay* were relative to the reference category, which corresponded to the *Constant* production function type. The regression coefficients suggest that the hazard for channels with *Ascending* production functions is 0.760 times the hazard of channels characterized by *Constant* production functions, the hazard for channels with *Descending* production functions is 0.597 times the hazard of channels with *Constant* production functions, and the hazard for channels with *Variable* production functions is 0.638 times the hazard of channels with *Constant* production functions. However, the significance values for all these coefficients are greater than 0.10, so any observed differences between these channels categories could be due to chance. Table 11.20 shows that the variables left out of the model all had score statistics with significance values greater than 0.05.

Table 11.21 describes the four covariate patterns that correspond to the types of production functions, each with otherwise "average" covariates. This table is a useful reference when looking at the survival plots, which are constructed for the mean values and each covariate pattern. As before, note that the "average" channel doesn't actually

exist when looking at the means of indicator variables for categorical predictors. Even with all scale predictors, it is unlikely to find a channel whose covariate values are all close to the mean.

The basic survival curve shown in Figure 11.3 a) is a visual display of the model-predicted time to death for the "average" channel. The vertical axis shows the probability of survival. Thus, any point on the survival curve shows the probability that the "average" channel will stay alive past that time. Past 65 days, the survival curve becomes less smooth because of the fewer channels that have survived for that long. The plot of the survival curves for each covariate pattern shown in Figure 11.3 c) gives a visual representation of the effect of the "production function type" category. Although not statistically significant, just as before, it is worth noticing that channels with *Descending* and *Variable* production functions seemed to have been more likely to survive than channels with *Ascending* production functions, while channels with *Constant* production functions seemed to have had the lowest chance of survival.

The basic hazard curve shown in Figure 11.3 b) is a visual display of the cumulative model-predicted potential to die for the "average" channel. The vertical axis shows the cumulative hazard, equal to the negative log of the survival probability. Past 65 days, the hazard curve, like the survival curve, becomes less smooth, for the same reason. The plot of the hazard curves for each covariate pattern shown in Figure 11.3 d) gives a visual representation of the effect of the "production function type" category. Channels with *Constant* and *Ascending* production functions had higher hazard curves because they had a greater potential to not survive.

11.3.4 Cox Regression Results for the First Week of Life

Table 11.22 Omnibus Tests of Model Coefficients for the First Block

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1	2660.736	34.146	1	.000	44.130	1	.000	44.130	1	.000
2	2624.757	63.915	2	.000	35.979	1	.000	80.109	2	.000

Variable(s) Entered at Step Number 1: PosterDivFirstWeek
Variable(s) Entered at Step Number 2: MessagesFirstWeek

Table 11.23 Omnibus Tests of Model Coefficients for the Second Block

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
2622.231	67.983	5	.000	2.526	3	.471	2.526	3	.471

Table 11.24 Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
MessagesFirstWeek	-.001	.000	22.376	1	.000	.999
PosterDivFirstWeek	.026	.005	26.225	1	.000	1.026
PFFirstWeek			2.526	3	.471	
PFFirstWeek(1)	-.169	.171	.978	1	.323	.845
PFFirstWeek(2)	-.335	.285	1.382	1	.240	.716
PFFirstWeek(3)	-.316	.224	1.977	1	.160	.729

Table 11.25 Variables not in the Equation

	Score	df	Sig.
PostersFirstWeek	.044	1	.834
UsersFirstWeek	.021	1	.884
LurkersFirstWeek	.025	1	.873
PTFirstWeek	.867	1	.352
MTFirstWeek	.366	1	.545

Table 11.26 Covariate Means and Pattern Values

	Mean	Pattern			
		1	2	3	4
PostersFirstWeek	12.684	12.684	12.684	12.684	12.684
UsersFirstWeek	29.543	29.543	29.543	29.543	29.543
LurkersFirstWeek	17.379	17.379	17.379	17.379	17.379
MessagesFirstWeek	409.060	409.060	409.060	409.060	409.060
PosterDivFirstWeek	92.816	92.816	92.816	92.816	92.816
PTFirstWeek	-.568	-.568	-.568	-.568	-.568
MTFirstWeek	-.557	-.557	-.557	-.557	-.557
PFFirstWeek(1)	.606	.000	1.000	.000	.000
PFFirstWeek(2)	.060	.000	.000	1.000	.000
PFFirstWeek(3)	.163	.000	.000	.000	1.000

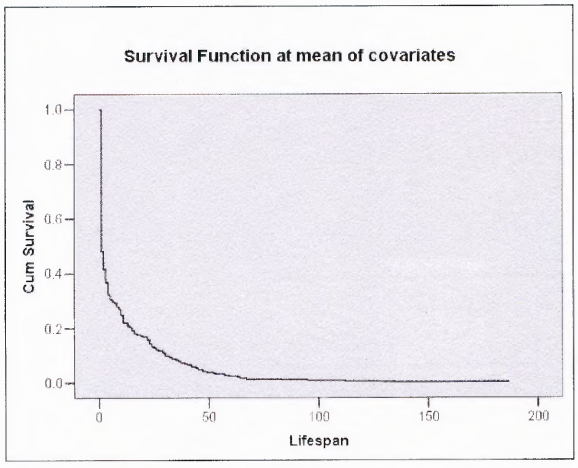

Figure 11.4 a) Survival curve for the first week of life.

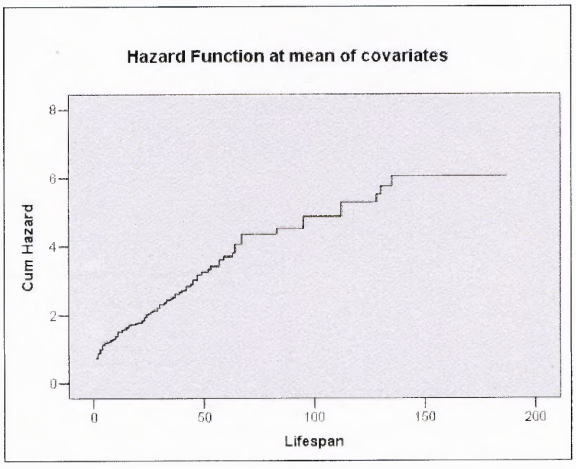

Figure 11.4 b) Hazard curve for the first week of life.

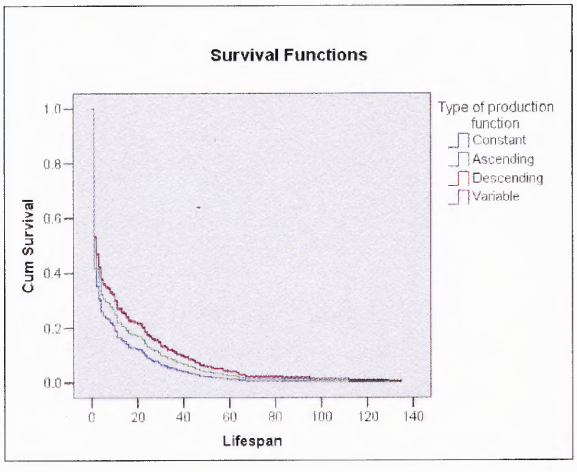

Figure 11.4 c) Survival curves for types of production functions during the first week of life.

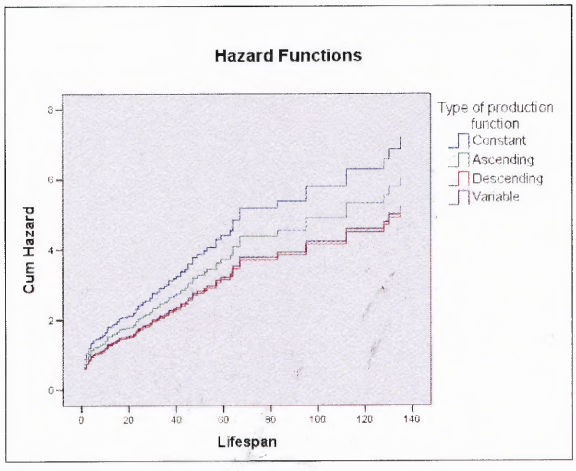

Figure 11.4 d) Hazard curves for types of production functions during the first week of life.

Table 11.22 shows that from all the variables entered into the first block of the regression model, only the poster diversity and the number of messages contributed significantly to it (significance of change less than 0.05). Table 11.23 shows that the addition of the type of production function to the model as a categorical variable did not contribute to the model (significance of change larger than 0.05). Table 11.24 shows the regression coefficients for the variables used in the final step of the model. Exp(B) represents the predicted change in the hazard for a unit increase in the predictor. The

value of Exp(B) for *PosterDivFirstWeek* means that the channels' death hazard increases by (100% * 1.026) – 100% = 2.6% for each increase of 1% in the diversity of the channel poster population. The value of Exp(B) for *MessagesFirstWeek* means that the channels' death hazard decreases by 100% - (100% * .999) = 0.1% for every new message. Same as before, the larger the value of the poster diversity measure, the higher the risk of death, and the more messages, the lower the risk of death.

The regression coefficients for the first three levels of *PFFirstWeek* were relative to the reference category, which corresponded to the *Constant* production function type. The regression coefficients suggest that the hazard for channels with *Ascending* production functions is 0.845 times the hazard of channels characterized by *Constant* production functions, the hazard for channels with *Descending* production functions is 0.716 times the hazard of channels with *Constant* production functions, and the hazard for channels with *Variable* production functions is 0.729 times the hazard of channels with *Constant* production functions. However, the significance values for all these coefficients are greater than 0.10, so any observed differences between these channels categories could be due to chance. Table 11.25 shows that the variables left out of the model all had score statistics with significance values greater than 0.05.

Table 11.26 describes the four covariate patterns that correspond to the types of production functions, each with otherwise "average" covariates. This table is a useful reference when looking at the survival plots, which are constructed for the mean values and each covariate pattern. As before, note that the "average" channel doesn't actually exist when looking at the means of indicator variables for categorical predictors. Even

with all scale predictors, it is unlikely to find a channel whose covariate values are all close to the mean.

The basic survival curve shown in Figure 11.4 a) is a visual display of the model-predicted time to death for the "average" channel. The vertical axis shows the probability of survival. Thus, any point on the survival curve shows the probability that the "average" channel will stay alive past that time. Past 65 days, the survival curve becomes less smooth because of the fewer channels that have survived for that long. The plot of the survival curves for each covariate pattern shown in Figure 11.4 c) gives a visual representation of the effect of the "production function type" category. Although not statistically significant, just as before, it is worth noticing that channels with *Descending* and *Variable* production functions seemed to have been more likely to survive than channels with *Ascending* production functions, while channels with *Constant* production functions seemed to have had the lowest chance of survival.

The basic hazard curve shown in Figure 11.4 b) is a visual display of the cumulative model-predicted potential to die for the "average" channel. The vertical axis shows the cumulative hazard, equal to the negative log of the survival probability. Past 65 days, the hazard curve, like the survival curve, becomes less smooth, for the same reason. The plot of the hazard curves for each covariate pattern shown in Figure 11.4 d) gives a visual representation of the effect of the "production function type" category. Channels with *Constant* and *Ascending* production functions had higher hazard curves because they had a greater potential to not survive.

11.3.5 Cox Regression Results for the First Two Weeks of Life

Table 11.27 Omnibus Tests of Model Coefficients for the First Block

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1	2642.968	48.466	1	.000	61.899	1	.000	61.899	1	.000
2	2588.838	82.193	2	.000	54.130	1	.000	116.029	2	.000

Variable(s) Entered at Step Number 1: PosterDivFirst2Weeks
Variable(s) Entered at Step Number 2: MessagesFirst2Weeks

Table 11.28 Omnibus Tests of Model Coefficients for the Second Block

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
2583.742	88.726	5	.000	5.096	3	.050	5.096	3	.050

Table 11.29 Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
MessagesFirst2Weeks	-.001	.000	30.331	1	.000	.999
PosterDivFirst2Weeks	.033	.005	42.651	1	.000	1.034
PFFirst2Weeks			5.397	3	.045	
PFFirst2Weeks(1)	-.367	.189	3.777	1	.050	.692
PFFirst2Weeks(2)	-.530	.328	2.609	1	.106	.589
PFFirst2Weeks(3)	-.473	.220	4.638	1	.031	.623

Table 11.30 Variables not in the Equation

	Score	df	Sig.
PostersFirst2Weeks	.998	1	.318
UsersFirst2Weeks	1.175	1	.278
LurkersFirst2Weeks	1.042	1	.307
PTFirst2Weeks	1.679	1	.195
MTFirst2Weeks	1.742	1	.187

Table 11.31 Covariate Means and Pattern Values

	Mean	Pattern			
		1	2	3	4
PostersFirst2Weeks	14.727	14.727	14.727	14.727	14.727
UsersFirst2Weeks	35.511	35.511	35.511	35.511	35.511
LurkersFirst2Weeks	21.230	21.230	21.230	21.230	21.230
MessagesFirst2Weeks	537.624	537.624	537.624	537.624	537.624
PosterDivFirst2Weeks	91.876	91.876	91.876	91.876	91.876
PTFirst2Weeks	-.434	-.434	-.434	-.434	-.434
MTFirst2Weeks	-.414	-.414	-.414	-.414	-.414
PFFirst2Weeks(1)	.589	.000	1.000	.000	.000
PFFirst2Weeks(2)	.046	.000	.000	1.000	.000
PFFirst2Weeks(3)	.230	.000	.000	.000	1.000

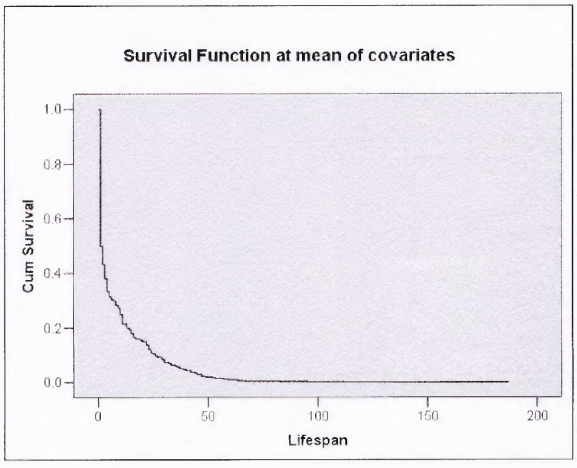

Figure 11.5 a) Survival curve for the first two weeks of life.

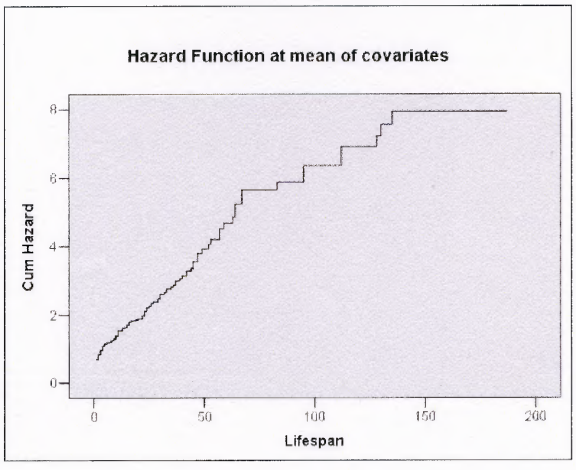

Figure 11.5 b) Hazard curve for the first two weeks of life.

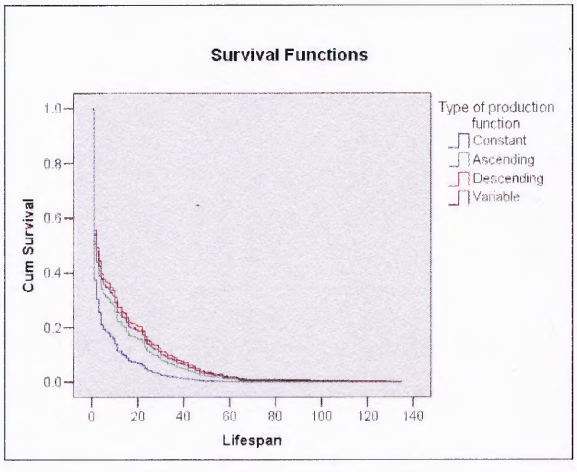

Figure 11.5 c) Survival curves for types of production functions during the first two weeks of life.

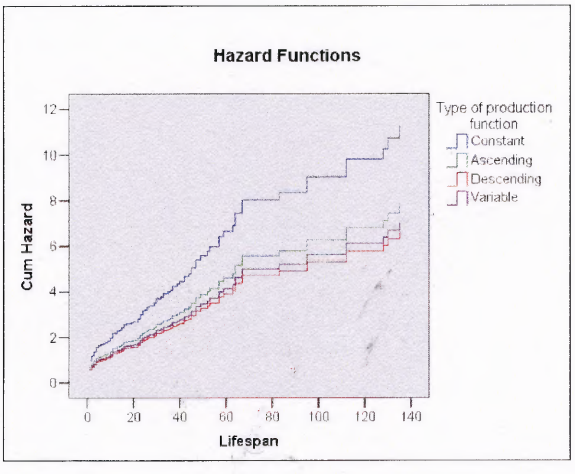

Figure 11.5 d) Hazard curves for types of production functions during the first two weeks of life.

Table 11.27 shows that from all the variables entered into the first block of the regression model, only the poster diversity and the number of messages contributed significantly to it (significance of change less than 0.05). Table 11.28 shows that in this case the addition of the type of production function to the model as a categorical variable contributed to the model (significance of change less than 0.05). Table 11.29 shows the regression coefficients for the variables used in the final step of the model. Exp(B) represents the predicted change in the hazard for a unit increase in the predictor. The

value of Exp(B) for *PosterDivFirst2Weeks* means that the channels' death hazard increases by (100% * 1.034) – 100% = 3.4% for each increase of 1% in the diversity of the channel poster population. The value of Exp(B) for *MessagesFirst2Weeks* means that the channels' death hazard decreases by 100% - (100% * .999) = 0.1% for every new message. Same as before, the larger the value of the poster diversity measure, the higher the risk of death, and the more messages, the lower the risk of death.

The regression coefficients for the first three levels of *PFFirst2Weeks* were relative to the reference category, which corresponded to the *Constant* production function type. The regression coefficients suggest that the hazard for channels with *Ascending* production functions is 0.692 times the hazard of channels characterized by *Constant* production functions, the hazard for channels with *Descending* production functions is 0.589 times the hazard of channels with *Constant* production functions, and the hazard for channels with *Variable* production functions is 0.623 times the hazard of channels with *Constant* production functions.

The significance value of the regression coefficient for channels with *Descending* production functions was slightly larger than 0.10, so any observed differences between this category and the reference category could be due to chance. However, the other two regression coefficients had significance values smaller than 0.05, indicating that channels with *Ascending* and *Variable* productions functions were statistically different from channels with *Constant* production functions.

Table 11.30 shows that the variables left out of the model all had score statistics with significance values greater than 0.05.

Table 11.31 describes the four covariate patterns that correspond to the types of production functions, each with otherwise "average" covariates. This table is a useful reference when looking at the survival plots, which are constructed for the mean values and each covariate pattern. As before, note that the "average" channel doesn't actually exist when looking at the means of indicator variables for categorical predictors. Even with all scale predictors, it is unlikely to find a channel whose covariate values are all close to the mean.

The basic survival curve shown in Figure 11.5 a) is a visual display of the model-predicted time to death for the "average" channel. The vertical axis shows the probability of survival. Thus, any point on the survival curve shows the probability that the "average" channel will stay alive past that time. Past 65 days, the survival curve becomes less smooth because of the fewer channels that have survived for that long. The plot of the survival curves for each covariate pattern shown in Figure 11.5 c) gives a visual representation of the effect of the "production function type" category.

In this case the channels with *Ascending* and *Variable* production functions were statistically different from channels with Constant production functions. Channels with *Variable* production functions were the most likely to survive, followed by channels with *Ascending* production functions, while channels with *Constant* production functions were the less likely to survive.

The basic hazard curve shown in Figure 11.5 b) is a visual display of the cumulative model-predicted potential to die for the "average" channel. The vertical axis shows the cumulative hazard, equal to the negative log of the survival probability. Past 65 days, the hazard curve, like the survival curve, becomes less smooth, for the same reason.

The plot of the hazard curves for each covariate pattern shown in Figure 11.5 d) gives a visual representation of the effect of the "production function type" category. Channels with *Constant* and *Ascending* production functions had higher hazard curves because they had a greater potential to not survive.

Figures 11.6 a) and b) show scatter plots of poster diversity and number of messages (the only survival predictors identified by the model) by lifespan (the dependent variable), with the channels grouped by the type of production function.

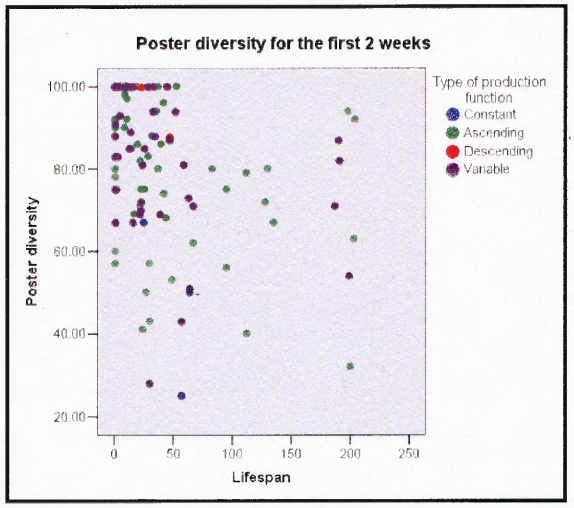

Figure 11.6 a) Poster diversity by lifespan for the first two weeks of life.

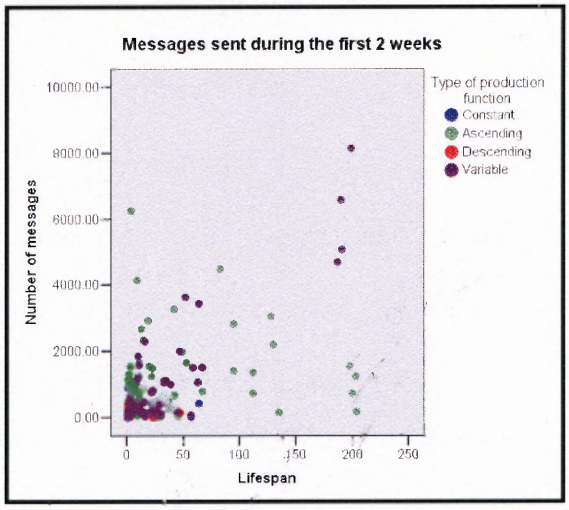

Figure 11.6 b) Number of messages by lifespan for the first two weeks of life.

11.4 Summary

The aim of this chapter was to assess whether it is possible to predict an IRC channel's chances of survival by looking at some of the initial starting conditions that characterize the overall activity of that channel, at the trajectories of the channel activity occurring inside the channel over various time intervals in the initial stages of its life, at the level of heterogeneity of the channel's user population, and at the channel's production functions computed for the same time intervals.

To predict the survivability of IRC channels, Cox regression models were created for four time intervals in the initial stages of the channels' lives: the first two hours, the first day, the first week, and the first two weeks.

The trajectories of channel activity did not predict the likelihood of channels' survival at all for any of the four intervals analyzed.

The heterogeneity of the channels' user populations (measured by the PosterDiv variable) was found to be a strong indicator for the channels' likelihood of survival by all the four models. Channels with heterogeneous populations were more likely to survive than channels with homogeneous populations. An increase of 1 percent in the PosterDiv variable in the first two hours of life implied a decrease of 1.3 percent in a channel's chance of survival. An increase of 1 percent in the PosterDiv variable in the first day of life implied a decrease of 1.5 percent in a channel's chance of survival. An increase of 1 percent in the PosterDiv variable in the first week of life implied a decrease of 2.6 percent in a channel's chance of survival. An increase of 1 percent in the PosterDiv variable in the first two weeks of life implied a decrease of 3.4 percent in a channel's chance of survival. This clearly indicates that the longer a channel's population stayed

homogeneous, the less likely was for that channel to survive. Lower heterogeneity may be an indication of higher poster turnover. Higher poster turnover may indicate persistent interest in the specific channel by more and more new posters. For example, a PosterDiv value of 100 percent for a channel during any of the analyzed time intervals suggests that the channel had constant/stable group of posters throughout that interval. An explanation could be a small group of people who resolved some chat among them and then didn't meet again. The opposite of this case was when a channel had a lower value for the PosterDiv variable. Such cases could be explained by more turnover of posters, therefore a persistent interest in the channel, hence the longer lifespan.

The number of messages was found to be a good indicator of the channels' likelihood of survival in three of the four cases. While the number of messages in the first two hours of life was not significant enough, the number of messages sent to a channel during its first day of life, first week of life and first two weeks of life were identified as good predictors by the Cox regression models. In all three cases, an increase of 1 message implied an increase of 0.1 percent in a channel's chance of survival, while 100 more messages sent to the channel implied an increase of 9.5 percent in the channel's chance of survival.

The shape of the channels' production functions was identified as a good predict only for the longest analyzed interval. Four types of production functions were used: constant, ascending, descending, and variable. The results of Cox regression model for the first two weeks of life indicated that channels with constant production functions were the most likely to die, while channels with variable production functions were the most likely to survive. The likelihood of survival of channels with ascending production

functions was slightly lower than that of channels with variable production functions. The model failed to produce statistically significant results for the channels with descending production functions, but this might have been due to the fact that there were only a few channels characterized by this type of production function, compared to the rest of the channels.

Overall, the Cox regression procedure produced a suitable model for predicting IRC channels' survivability. The use of separate blocks for fitting the model allowed guaranteeing that the production function categorical variable would be added to the final model, while still taking advantage of the stepwise techniques for choosing the other predictors. The best results were obtained when the regression model used the predictors that were computed for the immediate two-week period that followed a channel's creation.

CHAPTER 12

SUMMARY, CONTRIBUTIONS AND FUTURE RESEARCH

This chapter highlights the key findings of this dissertation, presents its contributions to the information systems research community, and outlines how each of the contributions could be extended and used by researchers in the future. These contributions encompass four broad areas: 1) Fundamental knowledge about Internet Relay Chat dynamics; 2) Methodologies for the capture and analysis of synchronous computer mediated communication dynamics; 3) Information systems theory; and 4) Fundamentals for synchronous social interaction space recommendation systems.

The presentation of these key findings, in the context of the four areas of contributions mentioned above, is structured as follows: Section 12.1 describes innovative methodologies developed to enable large-scale data-collection of synchronous CMC; Section 12.2 presents fundamental knowledge about Internet Relay Chat dynamics; Section 12.3 discusses how the empirical findings extend the Information-processing constraints model and the Critical Mass theory; Section 12.4 outlines how all these findings can be used as baseline data for the construction of synchronous social interaction space recommendation systems. Finally, Section 12.5 addresses the limitations of this work and Section 12.6 discusses potential topics for future research.

12.1 Data Capture and Analysis

Collecting large amounts of synchronous chat data has always been difficult. Especially when dealing with large IRC networks, the data-collection process can be truly cumbersome and this may be one of the reasons why mass interaction on IRC has never been researched. The most important technical difficulty encountered by previous researchers was the impossibility for regular users to log the activities of a large number of channels for longer time intervals (due to the architecture and the implementation of IRC). Other impediments included the inability to get various statistical information about the network without special administrative rights; connectivity problems (delays between users, network splits, DDoS attacks etc.); and channel access issues (users being "kicked" out or "banned" from channels, sometimes without any noticeable reason).

The solution to such problems was to setup an IRC server and link it to the Austnet IRC network. This allowed the collection of large amounts of data that would have been virtually impossible to obtain otherwise. A key issue was the linking of the server to the IRC network. While it would be relatively easy to simply set up a stand-alone IRC server, it is practically impossible to attract the tens of thousands of users required for the analysis of IRC mass interaction. The linking of a server to a well-established IRC network provided access to an already existing user base, which made the analysis of mass interaction easier. Having unrestricted access to an IRC server linked to an IRC network allowed capturing virtually all the data that passed through the network during the entire one-year data-collection period, with the exception of the intervals when the server was separated from the rest of the network due to various connectivity problems.

The data was collected using two approaches. The first was through the use of custom-written IRC bots that continuously monitored the channel spaces and collected data at specific time intervals. A bot, which is short for "robot", is defined as any program that, once started by a human person, can connect to the IRC network and perform various tasks such as (but not limited to) joining channels, posting messages independently and automatically, or collecting data, without the need of further human action. This resulted in information about the total number of channels and the total number of users of the network. The bots ran as background processes on the same machine on which the IRC server was installed.

The second approach was a combination of open-source TCP traffic monitoring software (Ethereal) and custom-written programs. The traffic monitoring software ran in the background on the same machine on which the IRC server software was installed and continuously collected all the TCP packets that flowed through the server and contained IRC-related data. The custom-written programs also ran as background processes on the IRC server machine; they parsed the data collected by the traffic monitoring software and extracted the information relevant to the total number of active channels, publicly active channels, publicly active users (posters), and messages exchanged in the public interaction spaces of the chat-channels.

Finally, a keyword-based algorithm was developed for the identification of postings and other actions that were taken by various IRC bots or other automated scripts. Since this research focused specifically on the behavior of human users, the need for a mechanism to separate the messages sent by bots from the messages sent by humans was very important. Hence, an algorithm was developed to perform this separation. The

algorithm parsed all the collected message data and attempted to distinguish real users from bots based on various patterns of keywords, phrases, and special characters that were identified as typically being used by bots. To refine the algorithm, several iterations were performed and random samples of messages attributed to both real users and bots were examined between iterations. To avoid any ethical issues, absolutely no connection was made between any identifier (nickname or Internet Protocol address) of the authors of the messages and the content of the messages. The analysis was performed simply to determine various patterns that would enable a better distinction between the messages generated by bots and the messages generated by humans. In the end, an expert's review revealed that the algorithm correctly identified over 99 percent of all the messages. Less than 1 percent of the messages attributed by the algorithm to bots were actually originated by human users, while less than 1 percent of the messages attributed by the algorithm to human users were actually originated by bots. This algorithm was run on a separate machine, using a copy of the dataset that was captured.

In the end, the data collected by the custom-written programs (in text format) occupied an amount of approximately 4.3 GB of hard disk space. The data collected by the traffic monitoring software occupied approximately 34.27 GB of hard disk space. However, this data was collected and archived in zip files. The total amount of raw text data that was collected in this manner occupied approximately 171.35 GB of hard disk space. As mentioned above, the entire collected data, i.e., both the data collected by the bots and the data collected by the traffic monitoring software was further processed to extract the information relevant to this research. This information was stored into a MS SQL Server 2000 database. At the completion of the research, the database contained 159

tables, which were all used in one way or another in the analysis, and occupied approximately 56 GB of hard disk space.

The first major contribution of this dissertation is the rich, large corpus of IRC data collected over a period of one year, which will be anonymized and made available online for future IRC research. Prior to this work, no such repository existed, as most of the previous research on synchronous CMC systems was conducted over small time intervals and looked at very limited numbers of synchronous interaction spaces. In the same way the Netscan system (Smith 1999) provided detailed reports about the activity of Usenet newsgroups, this corpus will be able to provide detailed reports about all the IRC channels that were present on the Austnet IRC network between February 1, 2005 and January 31, 2006; and about the users who joined the network during this interval. Furthermore, it will also be able to offer rich data for visualizing group interaction dynamics in synchronous chat, and compare them to the group interaction dynamics observed in Usenet newsgroups (Viegas and Smith 2004); across a number of different communication modalities, such as email or instant messaging (Cowell et al. 2006); or in face-to-face meetings (DiMicco, Hollenbach, and Bender 2006).

12.2 IRC Interaction Dynamics

Detailed descriptive statistics about IRC interaction dynamics were presented in Chapter 7. In what follows, the major findings are highlighted.

Over the course of one year, 43 percent of the total number of channels were visited by users, and of these channels only 20 percent hosted public interactions.

Approximately half of the channels that were active during any given month were likely to be active during the following month, while approximately 75 percent of the channels that were publicly active during any given month were likely to be publicly active during the following month.

Approximately 30 percent of the users who visited the network during any given month were likely to return during the following month, and likewise approximately 30 percent of the posters who were publicly active during any given month were likely to be publicly active during the following month.

Most of the users visited the IRC network for short periods of time and joined a small number of channels (both over the entire year, and during any given session). Similarly, most of the posters were publicly active in a small number of channels (both over the entire year, and during any given session).

A small proportion of posters generated the vast majority of public messages; over the year 10 percent of the posters were responsible for 91 percent of the public messages, while during any particular month 10 percent of that month's posters generated 80 – 85 percent of the public messages.

Publicly active channels were significantly more visited than active channels, in terms of both size of the population (number of users) and length of time (number of months). Approximately half of the publicly active channels were visited for five months at the least, while approximately 25 percent of the publicly active channels hosted public interactions for five months at the least.

A small number of channels were host to the vast majority of public interactions; over the year 94 percent of the public messages were sent to 10 percent of the channels

that have been publicly active throughout the year, while during any particular month, 80 – 87 percent of the public messages were sent to 10 percent of the channels that were publicly active over that month.

Until this work, researchers did not have a good understanding about mass interaction occurring inside large-scale synchronous CMC systems over long periods of time. The descriptive statistics data reported in this dissertation closes this research gap by presenting new, unique knowledge about the dynamics of such systems. The second major contribution of this thesis consists in the description of the overall ecology of an entire IRC network over a period of one year. This includes individual user behavior data, individual channel activity data, and general system evolution data. In sum, the descriptive statistics demonstrated that the analyzed IRC network was a dynamic, constantly changing environment with great turnover in users and channels. They also demonstrated that a small number of users were responsible for the vast majority of public interaction inside the IRC network, and that only a small subset of the channels existing in the IRC network were actually constantly visited and publicly active. These results support the findings obtained for Usenet newsgroups (Whittaker et al. 1998) or Internet forums (Soroka and Rafaeli 2006). Consequently, it is clear now that the asymmetry of user activity is common in both synchronous and asynchronous CMC systems.

12.3 Extending Information Systems Theory

This dissertation addressed and extended two theories: the Information-processing constraints theory (Jones and Rafaeli 1999) and the Critical Mass theory (Oliver, Marwell, and Teixeira 1985).

The Information-processing constraints theory conjectures that the level of information overload to which people are exposed when using a system influences their participation in computer mediated communication, and that the level of activity within an asynchronous CMC system can only rise up to a certain level. After that level is reached, due to the effects of information overload, the activity either remains constant or decreases. The Information-processing constraints theory has been used and cited mostly in relation to asynchronous CMC systems (Jones, Ravid, and Rafaeli 2004; Raban and Rafaeli 2007; Lampe, Johnston, and Resnick 2007). This research extended this theory to synchronous CMC systems, as it empirically identified the upper information-processing limits that constrain the community interaction dynamics seen in IRC chat-channels.

The Critical Mass theory seeks to predict the probability, extent and effectiveness of group action in pursuit of a public good. This research extended this social theory by providing a method to adapt two of its independent variables (production functions and group heterogeneity), and its dependent variable (achievement of a public good) to synchronous CMC systems; and a method to test whether its hypotheses hold in the case of groups using synchronous interaction spaces.

12.3.1 Information-processing Constraints Theory and IRC

While the IRC system did not pose itself restrictions on the number of users, posters, or messages, constraints on all three dimensions emerged as a result of information overload.

High message density was possible when the number of users and posters was small. As the number of participants, either users or posters, increased, the message density declined until community boundaries were reached.

The limit of the user community was identified to be less than 300 concurrent users in one chat-channel, while the limit of the channel poster community was identified to be less than 40 posters.

The upper limit to the message volume was reached even before the upper limit of posters. A maximum of approximately 600 public messages per chat-channel per 20-minute interval was observed. Roughly, this means that when the poster population was about 30, posters could not absorb more than 20 messages per poster within 20 minutes, or 30 messages per minute for the entire channel group.

Figures 12.1 and 12.2 illustrate these findings.

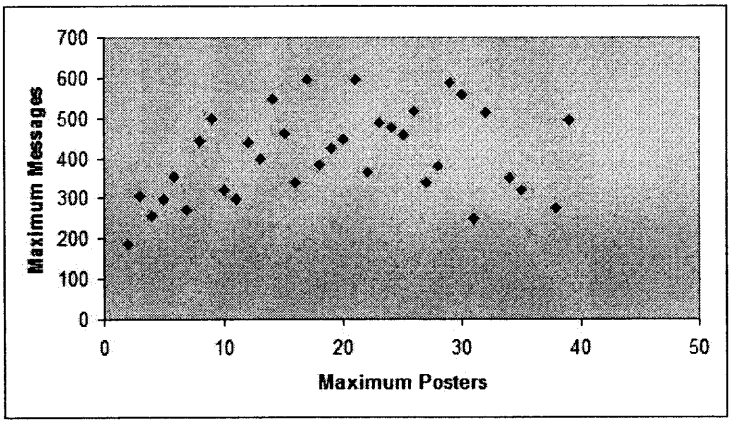

Figure 12.1 Maximum public messages versus maximum posters

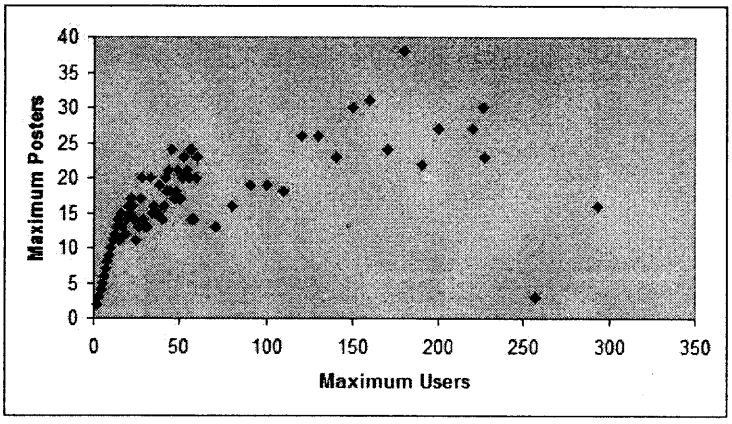

Figure 12.2 Maximum posters versus maximum users.

In the above figures each data point represents a number of channels (different number for each point) that had the same set of data. These figures show the maximum system potential to answer the question "what is the upper limit to activity in IRC channels?" For example, when the maximum number of posters in a channel is 2, then the maximum number of messages that was observed is 187. This means there could have been channels with 2 posters and less than 187 messages, but not more than that.

High message density was possible only at low values of observed users. Beyond a certain number of users, about 30 according to the findings, message density remained low and declined until the community stopped growing altogether.

As predicted by the Information-processing constraints theory, the boundaries to the rate of posting identified for IRC public channel communication were much lower than those found in Usenet (see Appendix of Jones, Ravid, and Rafaeli 2004 for details). Butler's work on email lists (2001) shows a mean community size of 168, which indicates that the maximum size is much higher. This provides support for the prediction that the number of interactive users in community spaces supported by synchronous CMC will be much smaller than the number of interactive users in community spaces

supported by asynchronous CMC. While in some context co-presence may be an explanation for this, the number of individuals participating in the Austnet IRC network implies that availability of potential participants is not likely to be the major problem. This suggests that information processing constraints, together with the information overload issues they give rise to, are significant factors in shaping the interaction structures present in this otherwise relatively unconstrained communication environment.

There are several possible explanations for the clear community boundaries that were identified. First, there is probably a human cognitive and practical limit to reading, absorbing, and reacting to messages in a given time frame. In addition, it is possible to have witnessed a form of diffusion of responsibility (or social loafing); when many people are in a chat-channel, specific individuals may feel less inclined to provide responses in the public interaction space because they know someone else is likely to respond. In a small group people may feel more responsible and may also get more credit and appreciation for their participation. All of these factors contribute to the interplay between individual limitations, technology features, and emergent interaction structures.

Prior to this research, the Information-processing constraints theory was only used in relation to asynchronous CMC systems (Jones, Ravid, and Rafaeli 2004; Raban and Rafaeli 2007; Lampe, Johnston, and Resnick 2007); and it demonstrated that typically the level of information overload to which people are exposed when using such a system influences their participation in computer mediated communication. Previous research has also shown that a group's sustainability is influenced by the publicly active members, as well as by the publicly inactive members of the group (Koh et al. 2007), but continuous increases in membership size and communication activity may affect both

positively and negatively the sustainability of groups in asynchronous CMC systems (Butler 2001). However, very little was known about these issues with respect to synchronous CMC systems. The third main contribution of this research is the testing of the Information-processing constraints theory in the context of an IRC network. This work was the first endeavor of this kind, and it showed clearly that the model is a solid one, and that its predictions are easily observable in the case of dynamic synchronous interaction spaces such as IRC channels.

12.3.2 Critical Mass Theory and IRC

To date, little is known about the initial conditions that lead to the formation of groups in synchronous spaces such as IRC channels, as well as about the subsequent conditions necessary for those groups to evolve and be sustained over longer periods of time. The notion of critical mass if often used when discussing the long-term sustainability of groups and the general consensus is that a group needs in its early stages a certain critical mass of members in order to become and remain successful over longer periods of time.

The Critical Mass theory of sociologists Oliver, Marwell, and Teixeira (1985) provides a theoretical model for predicting a group's success over time. Based on Olson's work on various theories of collective action (1965), this theory tries to predict the effects that production functions and the group's interest and resource heterogeneity would have on "the probability, the extent and effectiveness of group actions in pursuit of collective good." One of the theory's most important conjectures is that a group's level of heterogeneity, together with the shapes of various production functions, defined as relationships between resources contributed by the group and the collective output of that

group, can be used to distinguish between the likelihood of longer-term success of the group. This social theory was used in relation to CMC systems by several researchers.

Markus (1987) used it to explain the diffusion and adoption of interactive media. Arguing that usage of interactive media can have only two states, "all or nothing," she proposed several hypotheses about the relationship between the shape of the production functions and the heterogeneity of resources and interests on the one hand, and the achievement of universal media access on the other. Thorn and Connolly (1987), relying on the Critical Mass theory and on some other literature on collective action, studied the contribution of information to "databases," which were essentially archives of computer mediated communication. Seeing the "databases" as interactive media that provided public goods, they tried to determine the factors that influenced users' level of contributions. They produced a conceptual framework, which proposed that reduced contributions occurred because of greater contribution costs; larger groups of participants; lower value of information to participants; and greater asymmetries in information value and benefits across participants. Rafaeli and LaRose (1993) drew from both Markus' and Thorn and Connolly's work and made several predictions about the success of electronic bulletin boards. Overall, only slight support for the Critical Mass theory was found, but the authors offered a few possible explanations for this situation. Their conclusion was that "the study of interactive technologies needs to proceed beyond the case study level" if one is to better discern the factors that lead to the success or failure of computerized collaborative media.

Although Markus (1987) proposed a critical mass theory that would be applicable to all types of interactive media, most, if not all, of the recent research has used this

theory focusing exclusively on asynchronous CMC systems (Thorn and Connolly 1987; Rafaeli and LaRose 1993, Fulk et al. 2004). The results of the existing studies showed that using a "public goods" approach such as the Critical Mass theory in the domain of electronic communication media may be more complex then initially predicted by the theory itself. Thorn and Connolly admitted that both more empirical laboratory work and more theoretical extensions are needed in order to fully demonstrate "the power of 'public goods' thinking for the analysis of organizational communication issues." Rafaeli and LaRose acknowledged that the picture emerging from the results of their data analysis was that "of a more complex world than predicted by public goods theories." They suggested that further refinements were needed when applying such theories, the Critical Mass theory in particular, to collaborative media.

Fulk et al. (1996) looked at connectivity (defined as the capability to link people together both socially and physically), and communality (defined as a group's joint holding of a single body of information) as specific forms of public goods in interactive systems. Monge et al. (1998) developed a public-goods based theory that described the production of collective action in alliance-based inter-organizational communication and information systems. Fulk et al. (2004) developed and tested a model of how "individual-level factors interact with perceptions of collective action in influencing individuals' motivations to contribute privately controlled information to a collective repository."

The general consensus among researchers is that in the case of CMC systems the public goods produced by a group as a result of collective action are the communication and the information functions the systems provide, rather than the systems themselves.

This dissertation adapted the Critical Mass theory to synchronous CMC systems and tested its ability to help predict the long-term sustainability of group interaction in synchronous interaction spaces. Specifically, it proved that it is possible to predict IRC channels' chances of survival by looking at the channels' level of heterogeneity and at the channels' production functions during their initial stages of life.

The degree of homogeneity or heterogeneity for an IRC channel was measured by "poster diversity" variables, whose values ranged from 1 to 100. A channel was considered more homogeneous if its poster population stayed relatively constant as time passed (i.e., every time the channel was active the same users were present); and more heterogeneous if its poster population changed significantly over time (i.e., every time the channel was active, different users were present). A maximum value of 100 for the poster diversity variable would indicate a fully homogeneous population, while a minimum value of 1 would indicate a population with the highest level of heterogeneity.

The production functions were computed based on the definition provided by the Critical Mass theory. The theory defined production functions as the relationships between resources contributed by a group and the collective output of that group. In the case of IRC channels, the number of users present in the channel was considered a measure for the available group resources, while the number of public messages exchanged was considered a measure for the amount of group success achieved as a result of the collective action of the group.

To predict the long-term survivability of IRC channels, Cox regression models were created for four time intervals in the initial stages of the channels' lives: the first two hours, the first day, the first week and the first two weeks. These models used as

predictors various variables that measured the channels' overall activity, the trajectories of the channels' activity, the level of heterogeneity of the channels' user populations, and the channels' types of production functions.

Both of the independent variables drawn from the Critical Mass theory were helpful in the predictions. The heterogeneity of the channels' user populations was found to be a strong indicator for the channels' likelihood of survival by all the four models. Channels with heterogeneous populations were more likely to survive than channels with homogeneous populations.

The shape of the channels' production functions was identified as a good predictor only for the longest analyzed interval. Four types of production functions were used: constant, ascending, descending, and variable. The results of the Cox regression model for the first two weeks of life indicated that channels with constant production functions were the most likely to die, while channels with variable production functions were the most likely to survive. The likelihood of survival of channels with ascending production functions was slightly lower than that of channels with variable production functions. The model failed to produce statistically significant results for the channels with descending production functions, but this might have been due to the fact that there were only a few channels characterized by this type of production function, compared to the rest of the channels.

In addition, the number of messages was found to be a good indicator of the channels' likelihood of survival in three of the four cases. While the number of messages in the first two hours of life was not significant, the number of messages sent to a channel during its first day of life, first week of life and first two weeks of life were identified as

good predictors by the Cox regression models. The more messages that were exchanged in a channel, the higher the chances of survival were for that channel. The trajectories of channel activity did not predict the likelihood of channels' survival at all for any of the four intervals analyzed.

Figures 12.3 and 12.4 show scatter plots of poster diversity and number of messages, which were the survival predictors identified by the best model, by lifespan, the dependent variable, with the channels grouped by the type of production function.

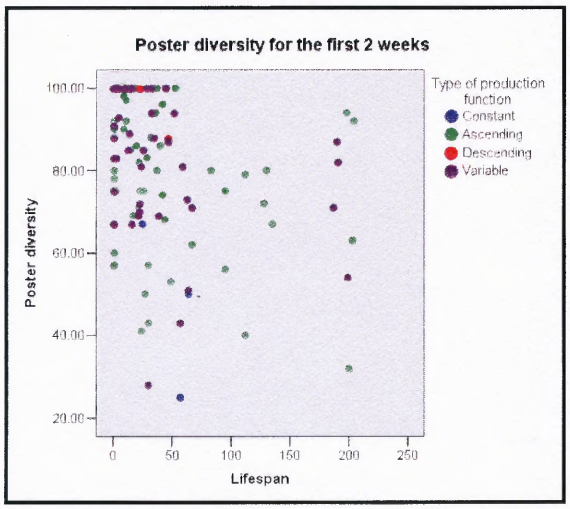

Figure 12.3 Poster diversity by lifespan for the first 2 weeks of life.

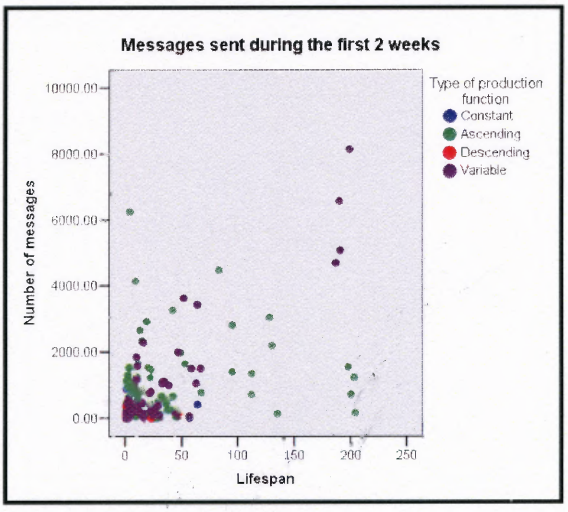

Figure 12.4 Number of messages by lifespan for the first 2 weeks of life.

Overall, the Cox regression procedure that used the predictors computed for the first two weeks of a channel's life produced the best model for predicting IRC channels' survivability. This suggests that various channel characteristics and activity measures for the immediate two-week period following a channel's creation can be used as strong predictors of its long-term sustainability. Also, the results showed that the Critical Mass theory can be successfully applied to predict the long-term sustainability of synchronous interaction spaces. The shape of the productions functions and the heterogeneity of the

channels' user populations during the channels' first two weeks of life were strong indicators of the channels' likelihood of survival.

To date, the use of the Critical Mass theory in the context of interactive media has been limited, both in terms of the types of systems used to test the theory, and in terms of relevance of the findings. The general consensus of previous studies was that using a "public goods" approach such as the Critical Mass theory in the domain of electronic communication media may be more complex then initially predicted by the theory itself; and that further refinements would be needed when applying such theories to CMC systems. The formation of groups in online spaces, and the necessary initial conditions that may lead to their long-term sustainability have always been topics of interest for researchers. However, very limited work actually addressed the formation of groups and their sustainability over time in large-scale CMC systems. Particularly in the case of synchronous CMC systems, this kind of knowledge was virtually nonexistent. This dissertation is the first endeavor to address this research gap, by successfully adapting the Critical Mass theory to synchronous CMC media, and proving that its predictions can hold in the case of IRC channels. Using two of its sets of independent variables – the group heterogeneity and the shape of the production functions – computed for the initial stages of IRC channels' lives, the hypotheses of the Critical Mass theory were proven to be correct.

12.4 Baseline Data for Synchronous CMC Recommendation Systems

Traditionally, recommendation systems have been categorized into four broad types: content-based systems, recommendation support systems, social data mining systems, and collaborative filtering systems (Terveen and Hill 2001). Of these, the content-based systems and the collaborative filtering systems are the most common, and they are sometimes combined into hybrid recommendation systems (Adomavicius and Tuzhilin 2005; Degemmis, Lops, and Semeraro 2007). Such systems typically employ the users/items/ratings model where a rating function is mapped from each user/item pair to some rating value (Terveen and Hill 2001). It has been argued recently that the users/items/ratings model might not suffice when dealing with complex and/or dynamic domains. As a result, a multidimensional approach to recommendation systems was suggested, where multiple variables should be taken into account when building the recommendation algorithm (Adomavicius and Tuzhilin 2001; Adomavicius et al. 2005). Lately, there has been an emergence of recommendation systems that focus specifically on social recommendations. As opposed to the traditional recommender systems that were used to recommend and/or sell various products or services such as books, movies, Usenet news, vacation packages, etc. (O'Connor et al. 2001; Schafer, Bowman, and Carroll 2002; Miller et al. 2003), social recommendations focus on different social aspects characteristic to online communities. These systems can be divided into two main categories: social matching systems, which attempt to "(partially) automate the process of bringing people together" (Terveen and McDonald 2005); and social space recommendation systems, which attempt to match people with interaction spaces (Van Dyke, Lieberman, and Maes 1999).

A recommendation system for synchronous CMC would fall into this latter category, as it would attempt to recommend chat spaces to users. To build such a system, there is a need for extensive baseline profile data on both user behavior and interaction space activity. This baseline data would serve as input for accurate prediction algorithms for the short-term activity and the long-term sustainability of the spaces. Understanding the short-term activity predictability of interaction spaces would help in providing instant recommendations. Understanding the long-term sustainability of interaction spaces would help in recommending only spaces that have a good chance of survival, and in determining when recommendations would lead to a more functional group in the long run.

The results of the analysis of the IRC interaction dynamics provided a large amount of baseline data about the users and the channels of the researched IRC network. The extension of the Information-processing constraints theory showed that constraints on the number of users, posters, and messages emerged as a result of information overload in IRC chat-channels, and determined the actual upper boundaries of these variables. The extension of the Critical Mass theory proved that it is possible to predict IRC channels' long-term sustainability by looking at the channels' level of heterogeneity and at the channels' production functions during their initial stages of life.

To understand whether it is possible to make short-term predictions (i.e., predictions for the next 20-minute interval) about the activity of IRC chat-channels, several linear and nonlinear regression models were created for a sample of channels. Both the linear and the nonlinear regression models used various combinations of independent variables to produce a best predictor variable. The accuracy of the

predictions was measured using Spearman correlation coefficients between the best predictor computed by the regression models and the actual values of observed posters, for each channel in the sample.

For the best linear regression model the overall correlation coefficient between the best predictor and the observed posters was 0.662. The results of the best linear regression model are reviewed in Table 12.1.

Table 12.1 Coefficients for Best Linear Regression Prediction Model for All Channels

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-.458	.026		-17.494	.000
AvgOP_Prev3_20	.592	.005	.512	117.267	.000
AvgOP_Prev3_20_Nwrk	.039	.008	.011	4.943	.000
AvgOP_Prev3wks	.010	.001	.018	7.621	.000
SP3	.331	.004	.345	81.676	.000
Slope	-.563	.013	-.120	-43.141	.000
TC2	.187	.019	.025	9.835	.000

AvgOP_Prev3_20 represents the average of the observed number of posters during the previous three 20-minute time intervals for each channel in the sample; *AvgOP_Prev3_20_Nwrk* represents the average number of observed posters per channel for the entire network during the previous three 20-minute time intervals; *AvgOP_Prev3wks* represents the average number of observed posters for the closest three 20-minutes intervals (just before, current and just after) at the same time during the previous three weeks, for each channel in the sample; *Slope* represents the slope of the line determined by the observed values for the previous three 20-minute time intervals for each channel and it is a basic indicator of the amount by which the number of posters varied during the previous hour; *SP3* represents a seasonality predictor whose value was predicted by a multiplicative decomposition of a time series analysis of the observed values per channel during the interval August 1 2005 – August 31 2005; *TC2* represents a

Spearman correlation coefficient between "time" and the observed number of posters during the last hour, offering a general idea about the direction of the conversation (up, down or constant).

For the best nonlinear regression model the overall correlation coefficient between the best predictor and the observed posters was 0.694. Also, very important to notice, the best nonlinear regression model used a minimal set of predictors. The best predictor in this case was computed using the following equation:

$$BP = 14.015 + (a * AvgOP\_Prev3\_20Mod) + (b * SlopeMod^{0.576}) +$$

$$+ (c * TC1 + d * TC1^2).$$

AvgOP_Prev3_20Mod represents the modified value of *AvgOP_Prev3_20* and *SlopeMod* represents the modified value of *Slope*. Transformations were necessary in order to eliminate the non-positive values that would have cause potential problems for the nonlinear regression model.

Table 12.2 reviews the parameter estimates for the best nonlinear regression model described in the equation above.

Table 12.2 Parameter Estimates for the Minimal Best Nonlinear Prediction Model

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
a	.968	.003	.963	.974
b	-3.504	.023	-3.548	-3.459
c	2.717	.108	2.505	2.928
d	-.639	.027	-.693	-.586

Overall, both regression models were found to be useful in the prediction of short-term channel activity. However, the nonlinear regression model produced slightly better predictions using a smaller set of predictors. The best nonlinear regression model included only predictors related to the previous hour of channel activity. This suggests that in a very dynamic synchronous medium such as IRC, accurate predictions about

future activity can be made taking into account only information pertinent to the channels' activity during the most recent hour, rather than looking at historic data over longer periods of time.

To further improve the short-term predictions of channel activity, while also reducing the costs involved in making such predictions, there was a need to identify the factors that can be used to distinguish highly predictable channels from unpredictable channels. Based on the results of the best nonlinear regression model three categories of predictability were considered: high predictability, low predictability, and perfect predictability. Several descriptive statistics were computed for all the channels and were then entered as independent variables into three logistic regression models. These models attempted to find which of the descriptive statistics would best differentiate high-predictability channels from low-predictability channels, low-predictability channels from perfect-predictability channels, and perfect-predictability channels from high-predictability channels.

The first logistic regression model revealed that when trying to determine whether the activity of a chat-channel during a particular week would have a high degree or a low degree of predictability, the best indicators were the number of days the channel was visited and the number of days the channel sustained public interactions during the previous month. The predictions were successful in 80 percent of the cases.

The second logistic regression model revealed that when trying to determine whether the activity of a chat-channel during a particular week would have a low degree of predictability or would be perfectly predictable, the best indicator was the number of

days the channel sustained public interactions during the previous month. The predictions were successful in 75 percent of the cases.

The third logistic regression model revealed that when trying to determine whether the activity of a chat-channel during a particular week would have a high degree or predictability or would be perfectly predictable, the best indicator was the number of days the channel sustained public interactions during the previous month. The predictions were successful in 85 percent of the cases.

The combination of all these results – the IRC network dynamics, the information overload data related to IRC channels, together with the algorithms for predicting the short-term activity and the long-term sustainability of IRC channels, can be considered baseline data for the construction of synchronous social interaction space recommendation systems.

This dissertation contributed to future efforts toward building real-time recommendation systems for synchronous CMC systems in several ways. First, it identified the upper boundaries to group interaction inside the synchronous interaction spaces of an IRC network. The understanding of such boundaries is important in order to avoid recommending chat spaces that are approaching, or have already reached the maximum level of activity with which their users can cope. Second, it showed how to differentiate predictable IRC channels from unpredictable IRC channels, and it determined the best metrics for predicting short-term activity inside IRC channels. Third, it offered insight into how to automatically profile both the users and the interaction spaces of a large-scale synchronous CMC system.

12.5 Limitations

The first limitation that needs to be addressed is the relevance of the findings with respect to other IRC networks or other large-scale synchronous chat systems. Thousands of IRC networks exist worldwide and their sizes vary greatly. Some host tens of thousands of channels and are visited by hundreds of thousands of users on a daily basis, while others have much lower numbers, ranging from tens to hundreds, of both channels and users. The network analyzed in this research had, at any time during the data-collection period provided there were no network connectivity issues, on average, approximately 2,500 channels and 4,000 users. Compared to other existing IRC networks these numbers qualified it as medium-sized. While common characteristics can certainly be found in many of the existing networks, it is likely that differences among them also exist. Ideally, similar research should be conducted both on large and small networks, as well as on other medium-sized networks and other types of large-scale synchronous chat systems in order to validate the findings reported in this dissertation.

The second limitation relates to the noise of the data, with respect to data generated by bot users versus data generated by human users. While it was possible to create an algorithm that correctly distinguished bot posters from human posters in over 99 percent of the cases, identifying bot lurkers from human lurkers was impossible. This was because in the absence of any publicly posted messages, the collected data did not provide any indicators about the nature of the users of the network. While the impact of this was not likely to be a major one, it should however be mentioned that the variables that described the total number of users may have included a small proportion of bots, and not only human users.

The third possible limitation of this research relates to the variable used as a measure of channel activity, to the variables chosen as predictors for all the regression models, and to the definitions used to describe the birth and death of channels.

Channel activity is a surrogate measure for the group interactions occurring inside IRC chat-channels. Channel activity can easily be operationalized in terms of many distinct measures, such as the overall number of potential contributors (e.g., number of users per channel); the overall number of actual contributors (e.g., number of posters per channel); the overall number of public messages; the rates of contribution per user (e.g., number of messages per user); the rates of contribution per poster (e.g., number of messages per poster); the complexity of the contributions (e.g., the number of words per message); the proportion of messages that receive replies; or the number of distinct threads of conversation. While all these variables represent the level of group interaction, as they clearly indicate the amount and the intensity of the public activity of any IRC chat-channel, none of them were addressed by any previous work. Therefore, a starting point needed to be chosen. This research focused on one variable, specifically the number of actual contributors, but further studies should be conducted using other measures of channel activity in order to validate the findings reported in this dissertation.

The current lack of research in the area of predicting activity in synchronous systems, both short-term and long-term, implied that there were no well-known variables that could be used to predict the activity of IRC channels. Consequently, there was a need to choose a point from which to start working toward making such predictions. Overall, the chosen predictors were successful in accomplishing the tasks they were intended to perform. However, the set of predictors can certainly be further improved.

Similarly, because of the same lack of research, the "birth" and "death" of channels needed to be defined based only on empirical observations and common sense. The main reason for defining the birth and death of chat-channels in terms of the level of activity supported was because of the interest presented by this level of activity. A channel's creation day and disappearance day were considered less relevant mostly because channels could easily be created and could exist for long periods of time after their creation without being visited at all, before they would eventually disappear.

The last limitation relates to the procedure used to create the dataset that was used for most of this research. This dataset was selected through a stratified random sampling of all the channels that were publicly active during August 2005. The stratification was based on the channels' number of visitors and number of visited days. While there are other possible approaches for the stratification of channels, the lack of previous research in this area made it necessary to select these two particular variables as a starting point. In the absence of time constraints and processing power constraints, a larger set of channels should be analyzed to maximize the relevance of the findings. This larger set should include either several stratified random samples, based on different stratification methods, or, in the ideal case, all the channels that were publicly active during a particular month.

12.6 Future Research

The rich, large corpus of information containing the one-year data will be made available to other researchers as it offers a plenitude of future research opportunities related to synchronous chat. Among these opportunities, one could mention: social networks research; introduction management research; analyzing the formation of relationships on IRC; analyzing the use of language on IRC; analyzing the use of nicknames on IRC; analyzing the differences between bot users activity and human users activity; analyzing the private interactions among IRC users.

The profiling of both users and chat-channels could further be extended. Detailed profile data could be created for various subsets of users and channels (for example the characteristics of the most active users and channels could be compared to the characteristics of less active users and channels) in order to better understand the dynamics of IRC networks.

The Information-processing constraints theory could further be extended by a more direct comparison between various types of synchronous and asynchronous CMC systems (such as IRC channels, Usenet newsgroups, or Listserv email lists). Furthermore, the collected data could permit the testing of Butler's resource-based theory of sustainable online social structures (2001) – which proved that membership size and communication activity influence positively and negatively the sustainability of groups in asynchronous CMC systems – in the context of large-scale synchronous chat systems.

Social visualizations of both group activity and individual activity in synchronous CMC systems represent another potential area of future research. Using the collected

data, tools such as those described by Viegas and Smith (2004) for visualizing activity in Usenet newsgroups could be created for visualizing activity inside IRC channels.

The use of the Critical Mass theory in the context of synchronous interaction spaces can also be further explored. Originally, the theory offered hypotheses about only two types of production functions: accelerating and decelerating. This dissertation identified twelve categories of production functions for IRC channels, which were grouped into four broader types. Examining each of these twelve categories and their effect on the group sustainability more deeply, for both new and already established channels, would be an interesting endeavor. Also, using other variables as measures of the public goods produced as a result of collective action in IRC channels could offer interesting insight into how the Critical Mass theory can be applied to large-scale synchronous CMC systems.

Finally, building a recommender system for IRC chat-channels would be the most natural continuation of this research. Since IRC networks often contain thousands of chat-channels, populated by even larger numbers of users, navigating through them is difficult. In part, this is because only relatively impoverished mechanisms exist for user navigation. This research could easily represent the foundation for various types of real-time recommender algorithms. The data-collection method described in this dissertation could provide information relevant for both automated user profiling (what channels do users visit, how many channels do users visit, how much time do users spend during a session, how often do users visit the network etc.) and automated space profiling (when are channels most active, how many users visit them, how often are channels visited etc.). Real-time recommendations could be provided to a user either implicitly, based on

mining the profile data and identifying channels with similar characteristics to those that were previously visited by the user; or explicitly, based on various parameter-based requests. Knowing the boundaries of channel activity would help avoid recommending channels that are approaching or have already reached the maximum level of activity that can be supported by users. An accurate mechanism for making short-term predictions would provide significant insight about when to recommend and when not to recommend a particular channel, while an accurate mechanism for making long-term predictions would help in recommending only spaces that have a good chance of survival, and in determining when recommendations would actually lead to more functional groups in the long run.

APPENDIX

CORRELATION COEFFICIENTS TABLES

This appendix contains all the tables that report the correlation coefficients obtained between the observed posters and the eight computed best predictors, for the linear and non linear regression models described in Chapter 9.

Table A.1 Linear Regression Correlation Coefficients for All Channels

			BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
Spearman's rho	ObservedPosters	Correlation Coefficient	.663(**)	.662(**)	.675(**)	.674(**)	.664(**)	.662(**)	.677(**)	.676(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
		N	45360	45360	45360	45360	45360	45360	45360	45360

Table A.2 Linear Regression Correlation Coefficients for All Channels Grouped by Type

Type				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
1	Spearman's rho	ObservedPosters	Correlation Coefficient	.467(**)	.492(**)	.503(**)	.503(**)	.492(**)	.492(**)	.503(**)	.493(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
2	Spearman's rho	ObservedPosters	Correlation Coefficient	.573(**)	.573(**)	.562(**)	.559(**)	.573(**)	.570(**)	.561(**)	.559(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
3	Spearman's rho	ObservedPosters	Correlation Coefficient	.428(**)	.427(**)	.434(**)	.433(**)	.428(**)	.427(**)	.435(**)	.433(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
4	Spearman's rho	ObservedPosters	Correlation Coefficient	.611(**)	.609(**)	.578(**)	.576(**)	.610(**)	.607(**)	.582(**)	.581(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
5	Spearman's rho	ObservedPosters	Correlation Coefficient	.591(**)	.590(**)	.603(**)	.601(**)	.590(**)	.589(**)	.601(**)	.596(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
6	Spearman's rho	ObservedPosters	Correlation Coefficient	.358(**)	.364(**)	.535(**)	.534(**)	.372(**)	.370(**)	.541(**)	.539(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
7	Spearman's rho	ObservedPosters	Correlation Coefficient	.573(**)	.574(**)	.568(**)	.567(**)	.575(**)	.572(**)	.568(**)	.567(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
8	Spearman's rho	ObservedPosters	Correlation Coefficient	.521(**)	.521(**)	.476(**)	.478(**)	.522(**)	.520(**)	.478(**)	.477(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
9	Spearman's rho	ObservedPosters	Correlation Coefficient	.838(**)	.837(**)	.830(**)	.830(**)	.838(**)	.838(**)	.832(**)	.832(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040

** Correlation is significant at the 0.01 level (2-tailed).

Table A.3 Linear Regression Correlation Coefficients for All Channels Grouped by Size

Size				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
large	Spearman's rho	ObservedPosters	Correlation Coefficient	.758(**)	.757(**)	.742(**)	.742(**)	.758(**)	.757(**)	.744(**)	.743(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120
medium	Spearman's rho	ObservedPosters	Correlation Coefficient	.510(**)	.511(**)	.575(**)	.573(**)	.515(**)	.513(**)	.578(**)	.575(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120
small	Spearman's rho	ObservedPosters	Correlation Coefficient	.518(**)	.518(**)	.516(**)	.514(**)	.518(**)	.516(**)	.515(**)	.513(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120

Table A.4 Linear Regression Correlation Coefficients for All Channels Grouped by Intensity

Intensity				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
high	Spearman's rho	ObservedPosters	Correlation Coefficient	.728(**)	.728(**)	.766(**)	.765(**)	.730(**)	.729(**)	.768(**)	.767(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120
low	Spearman's rho	ObservedPosters	Correlation Coefficient	.346(**)	.346(**)	.351(**)	.351(**)	.345(**)	.345(**)	.349(**)	.349(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120
medium	Spearman's rho	ObservedPosters	Correlation Coefficient	.339(**)	.339(**)	.336(**)	.336(**)	.338(**)	.338(**)	.334(**)	.334(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120

** Correlation is significant at the 0.01 level (2-tailed).

Table A.5 Linear Regression Correlation Coefficients for All Small Channels

			BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
Spearman's rho	ObservedPosters	Correlation Coefficient	.520(**)	.519(**)	.516(**)	.518(**)	.520(**)	.518(**)	.516(**)	.517(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
		N	15120	15120	15120	15120	15120	15120	15120	15120

Table A.6 Linear Regression Correlation Coefficients for All Small Channels, Grouped by Type

Type				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
1	Spearman's rho	ObservedPosters	Correlation Coefficient	.536(**)	.536(**)	.536(**)	.536(**)	.536(**)	.536(**)	.536(**)	.536(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
2	Spearman's rho	ObservedPosters	Correlation Coefficient	.581(**)	.581(**)	.579(**)	.581(**)	.581(**)	.579(**)	.579(**)	.580(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
3	Spearman's rho	ObservedPosters	Correlation Coefficient	.409(**)	.408(**)	.407(**)	.407(**)	.409(**)	.409(**)	.407(**)	.407(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040

Table A.7 Correlation Coefficients for All Medium Channels

			BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
Spearman's rho	ObservedPosters	Correlation Coefficient	.598(**)	.599(**)	.574(**)	.574(**)	.599(**)	.598(**)	.576(**)	.576(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
		N	15120	15120	15120	15120	15120	15120	15120	15120

** Correlation is significant at the 0.01 level (2-tailed).

Table A.8 Linear Regression Correlation Coefficients for All Medium Channels Grouped by Type

Type				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
4	Spearman's rho	ObservedPosters	Correlation Coefficient	.606(**)	.605(**)	.572(**)	.572(**)	.608(**)	.606(**)	.575(**)	.575(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
5	Spearman's rho	ObservedPosters	Correlation Coefficient	.618(**)	.617(**)	.603(**)	.603(**)	.618(**)	.617(**)	.600(**)	.600(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
6	Spearman's rho	ObservedPosters	Correlation Coefficient	.561(**)	.564(**)	.540(**)	.540(**)	.562(**)	.562(**)	.545(**)	.545(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040

Table A.9 Linear Regression Correlation Coefficients for All Large Channels

			BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
Spearman's rho	ObservedPosters	Correlation Coefficient	.732(**)	.732(**)	.726(**)	.725(**)	.733(**)	.734(**)	.727(**)	.727(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
		N	15120	15120	15120	15120	15120	15120	15120	15120

Table A.10 Linear Regression Correlation Coefficients for All Large Channels Grouped by Type

Type				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
7	Spearman's rho	ObservedPosters	Correlation Coefficient	.530(**)	.532(**)	.541(**)	.539(**)	.532(**)	.531(**)	.541(**)	.541(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
8	Spearman's rho	ObservedPosters	Correlation Coefficient	.460 (**)	.461(**)	.443(**)	.446(**)	.466(**)	.465(**)	.443(**)	.446(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
9	Spearman's rho	ObservedPosters	Correlation Coefficient	.832 (**)	.831(**)	.827(**)	.827(**)	.832(**)	.833(**)	.829(**)	.829(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040

** Correlation is significant at the 0.01 level (2-tailed).

Table A.11 Linear Regression Correlation Coefficients for All Low-Intensity Channels

			BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
Spearman's rho	ObservedPosters	Correlation Coefficient	.581(**)	.582(**)	.595(**)	.595(**)	.581(**)	.580(**)	.600(**)	.599(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
		N	15120	15120	15120	15120	15120	15120	15120	15120

Table A.12 Linear Regression Correlation Coefficients for All Low-Intensity Channels Grouped by Type

Type				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
1	Spearman's rho	ObservedPosters	Correlation Coefficient	.475(**)	.475(**)	.503(**)	.503(**)	.475(**)	.475(**)	.513(**)	.503(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
4	Spearman's rho	ObservedPosters	Correlation Coefficient	.603(**)	.604(**)	.614(**)	.612(**)	.605(**)	.604(**)	.615(**)	.614(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
7	Spearman's rho	ObservedPosters	Correlation Coefficient	.553(**)	.554(**)	.568(**)	.569(**)	.552(**)	.551(**)	.575(**)	.575(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040

Table A.13 Linear Regression Correlation Coefficients for All Medium-Intensity Channels

			BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
Spearman's rho	ObservedPosters	Correlation Coefficient	.577(**)	.576(**)	.564(**)	.564(**)	.577(**)	.575(**)	.595(**)	.567(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
		N	15120	15120	15120	15120	15120	15120	15120	15120

** Correlation is significant at the 0.01 level (2-tailed).

Table A.14 Linear Regression Correlation Coefficients for All Medium-Intensity Channels Grouped by Type

Type				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
2	Spearman's rho	ObservedPosters	Correlation Coefficient	.583(**)	.582(**)	.580(**)	.579(**)	.583(**)	.578(**)	.580(**)	.581(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
5	Spearman's rho	ObservedPosters	Correlation Coefficient	.616(**)	.616(**)	.608(**)	.608(**)	.617(**)	.616(**)	.613(**)	.611(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
8	Spearman's rho	ObservedPosters	Correlation Coefficient	.518(**)	.518(**)	.493(**)	.493(**)	.518(**)	.517(**)	.501(**)	.498(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040

Table A.15 Linear Regression Correlation Coefficients for All High-Intensity Channels

			BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
Spearman's rho	ObservedPosters	Correlation Coefficient	.717(**)	.717(**)	.738(**)	.738(**)	.717(**)	.718(**)	.740(**)	.739(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
		N	15120	15120	15120	15120	15120	15120	15120	15120

Table A.16 Linear Regression Correlation Coefficients for All High-Intensity Channels Grouped by Type

Type				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
3	Spearman's rho	ObservedPosters	Correlation Coefficient	.386(**)	.386(**)	.359(**)	.361(**)	.382(**)	.382(**)	.358(**)	.359(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
6	Spearman's rho	ObservedPosters	Correlation Coefficient	.369(**)	.369(**)	.480(**)	.479(**)	.370(**)	.370(**)	.483(**)	.482(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
9	Spearman's rho	ObservedPosters	Correlation Coefficient	.831(**)	.831(**)	.825(**)	.825(**)	.832(**)	.833(**)	.827(**)	.827(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040

** Correlation is significant at the 0.01 level (2-tailed).

Table A.17 Nonlinear Regression Correlation Coefficients for All Channels

			BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.664(**)	.664(**)	.676(**)	.676(**)	.667(**)	.668(**)	.677(**)	.677(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
		N	45360	45360	45360	45360	45360	45360	45360	45360

Table A.18 Nonlinear Regression Correlation Coefficients for All Channels Grouped by Type

Type				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
1	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.513(**)	.513(**)	.524(**)	.524(**)	.513(**)	.513(**)	.536(**)	.536(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
2	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.577(**)	.576(**)	.572(**)	.570(**)	.575(**)	.580(**)	.573(**)	.571(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
3	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.428(**)	.425(**)	.421(**)	.421(**)	.427(**)	.428(**)	.422(**)	.421(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
4	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.615(**)	.614(**)	.573(**)	.574(**)	.619(**)	.616(**)	.578(**)	.579(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
5	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.590(**)	.591(**)	.604(**)	.602(**)	.596(**)	.598(**)	.599(**)	.599(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
6	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.356(**)	.357(**)	.537(**)	.538(**)	.368(**)	.367(**)	.543(**)	.543(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
7	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.553(**)	.553(**)	.561(**)	.561(**)	.553(**)	.557(**)	.564(**)	.563(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
8	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.512(**)	.515(**)	.462(**)	.461(**)	.514(**)	.519(**)	.463(**)	.462(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
9	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.839(**)	.839(**)	.832(**)	.832(**)	.839(**)	.840(**)	.833(**)	.833(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040

** Correlation is significant at the 0.01 level (2-tailed).

Table A.19 Nonlinear Regression Correlation Coefficients for All Channels Grouped by Size

Size				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
large	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.757(**)	.757(**)	.742(**)	.742(**)	.757(**)	.758(**)	.743(**)	.743(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120
medium	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.513(**)	.513(**)	.574(**)	.574(**)	.522(**)	.522(**)	.576(**)	.576(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120
small	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.523(**)	.521(**)	.518(**)	.517(**)	.521(**)	.524(**)	.519(**)	.518(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120

Table A.20 Nonlinear Regression Correlation Coefficients for All Channels Grouped by Intensity

Intensity				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
high	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.733(**)	.733(**)	.768(**)	.769(**)	.737(**)	.737(**)	.770(**)	.770(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120
low	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.588(**)	.587(**)	.575(**)	.575(**)	.589(**)	.590(**)	.578(**)	.578(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120
medium	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.563(**)	.564(**)	.549(**)	.547(**)	.565(**)	.569(**)	.548(**)	.547(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120

** Correlation is significant at the 0.01 level (2-tailed).

Table A.21 Nonlinear Regression Correlation Coefficients for All Channels Obtained Using a Reduced Set of Independent Variables

			BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.659(**)	.660(**)	.677(**)	.677(**)	.660(**)	.661(**)	.678(**)	.678(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
		N	45360	45360	45360	45360	45360	45360	45360	45360

Table A.22 Nonlinear Regression Correlation Coefficients for All Channels Grouped by Size Obtained Using a Reduced Set of Independent Variables

Size				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
large	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.757(**)	.757(**)	.743(**)	.743(**)	.758(**)	.758(**)	.743(**)	.743(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120
medium	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.500(**)	.502(**)	.576(**)	.577(**)	.501(**)	.502(**)	.577(**)	.577(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120
small	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.520(**)	.520(**)	.522(**)	.520(**)	.523(**)	.524(**)	.522(**)	.520(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120

Table A.23 Nonlinear Regression Correlation Coefficients for All Channels Grouped by Intensity Obtained Using a Reduced Set of Independent Variables

Intensity				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
high	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.725(**)	.725(**)	.769(**)	.769(**)	.725(**)	.725(**)	.770(**)	.770(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120
low	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.588(**)	.590(**)	.576(**)	.577(**)	.587(**)	.589(**)	.580(**)	.580(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120
medium	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.560(**)	.563(**)	.550(**)	.549(**)	.567(**)	.568(**)	.548(**)	.547(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	15120	15120	15120	15120	15120	15120	15120	15120

** Correlation is significant at the 0.01 level (2-tailed).

Table A.24 Nonlinear Regression Correlation Coefficients for All Channels Grouped by Type, Obtained Using a Reduced Set of Independent Variables

Type				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
1	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.536(**)	.524(**)	.536(**)	.536(**)	.524(**)	.524(**)	.536(**)	.536(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
2	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.571(**)	.571(**)	.576(**)	.573(**)	.575(**)	.577(**)	.576(**)	.573(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
3	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.427(**)	.430(**)	.425(**)	.424(**)	.430(**)	.428(**)	.425(**)	.425(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
4	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.619(**)	.620(**)	.575(**)	.578(**)	.615(**)	.616(**)	.578(**)	.580(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
5	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.590(**)	.593(**)	.602(**)	.603(**)	.597(**)	.600(**)	.599(**)	.598(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
6	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.333(**)	.334(**)	.542(**)	.543(**)	.334(**)	.334(**)	.547(**)	.547(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
7	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.550(**)	.552(**)	.562(**)	.561(**)	.551(**)	.554(**)	.566(**)	.565(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
8	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.506(**)	.515(**)	.463(**)	.463(**)	.517(**)	.516(**)	.460(**)	.461(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
9	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.839(**)	.839(**)	.831(**)	.832(**)	.839(**)	.840(**)	.833(**)	.833(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040

** Correlation is significant at the 0.01 level (2-tailed).

Table A.25 Nonlinear Regression Correlation Coefficients for All Channels Obtained Using Only Independent Variables Relative to the Previous Hour of Channel Activity

			BP1	BP2
Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.694	.694
		Sig. (2-tailed)	.000	.000
		N	45360	45360

Table A.26 Nonlinear Regression Correlation Coefficients for All Channels Grouped by Type Obtained Using Only Independent Variables Relative to the Previous Hour of Channel Activity

Type				BP1	BP2
1	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.484(**)	.475(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
2	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.582(**)	.580(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
3	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.457(**)	.456(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
4	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.597(**)	.599(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
5	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.609(**)	.608(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
6	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.554(**)	.554(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
7	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.572(**)	.572(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
8	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.505(**)	.506(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
9	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.842(**)	.841(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040

Table A.27 Nonlinear Regression Correlation Coefficients for All Channels Grouped by Size Obtained Using Only Independent Variables Relative to the Previous Hour of Channel Activity

Size				BP1	BP2
large	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.762(**)	.762(**)
			Sig. (2-tailed)	.000	.000
			N	15120	15120
medium	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.590(**)	.590(**)
			Sig. (2-tailed)	.000	.000
			N	15120	15120
small	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.534(**)	.532(**)
			Sig. (2-tailed)	.000	.000
			N	15120	15120

Table A.28 Nonlinear Regression Correlation Coefficients for All Channels Grouped by Intensity Obtained Using Only Independent Variables Relative to the Previous Hour of Channel Activity

Intensity				BP1	BP2
high	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.783(**)	.783(**)
			Sig. (2-tailed)	.000	.000
			N	15120	15120
low	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.589(**)	.589(**)
			Sig. (2-tailed)	.000	.000
			N	15120	15120
medium	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.569(**)	.569(**)
			Sig. (2-tailed)	.000	.000
			N	15120	15120

Table A.29 Nonlinear Regression Correlation Coefficients for All Small Channels

			BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.533(**)	.534(**)	.532(**)	.531(**)	.533(**)	.534(**)	.529(**)	.529(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
		N	15120	15120	15120	15120	15120	15120	15120	15120

Table A.30 Nonlinear Regression Correlation Coefficients for All Small Channels Grouped by Type

Type				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
1	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.524(**)	.524(**)	.524(**)	.524(**)	.524(**)	.524(**)	.524(**)	.536(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
2	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.587(**)	.587(**)	.585(**)	.584(**)	.587(**)	.587(**)	.582(**)	.582(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
3	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.441(**)	.441(**)	.441(**)	.437(**)	.441(**)	.441(**)	.437(**)	.434(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040

Table A.31 Nonlinear Regression Correlation Coefficients for All Small Channels Obtained Using Only Independent Variables Relative to the Previous Hour of Channel Activity

			BP1	BP2
Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.532(**)	.532(**)
		Sig. (2-tailed)	.000	.000
		N	15120	15120

Table A.32 Nonlinear Regression Correlation Coefficients for All Small Channels Obtained Using Only Independent Variables Relative to the Previous Hour of Channel Activity and Grouped by Type

Type				BP1	BP2
1	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.524(**)	.524(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
2	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.582(**)	.582(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
3	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.444(**)	.444(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040

Table A.33 Nonlinear Regression Correlation Coefficients for All Medium Channels

			BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.525(**)	.525(**)	.576(**)	.576(**)	.514(**)	.513(**)	.575(**)	.576(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
		N	15120	15120	15120	15120	15120	15120	15120	15120

Table A.34 Nonlinear Regression Correlation Coefficients for All Medium Channels Grouped by Type

Type				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
4	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.540(**)	.541(**)	.609(**)	.608(**)	.530(**)	.529(**)	.605(**)	.604(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
5	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.567(**)	.568(**)	.578(**)	.579(**)	.559(**)	.558(**)	.580(**)	.582(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
6	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.459(**)	.458(**)	.526(**)	.527(**)	.442(**)	.442(**)	.529(**)	.530(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040

Table A.35 Nonlinear Regression Correlation Coefficients for All Medium Channels Obtained Using Only Independent Variables Relative to the Previous Hour of Channel Activity

			BP1	BP2
Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.572(**)	.572(**)
		Sig. (2-tailed)	.000	.000
		N	15120	15120

Table A.36 Nonlinear Regression Correlation Coefficients for All Medium Channels Obtained Using Only Independent Variables Relative to the Previous Hour of Channel Activity and Grouped by Type

Type				BP1	BP2
4	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.600(**)	.601(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
5	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.577(**)	.577(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
6	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.527(**)	.527(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040

Table A.37 Nonlinear Regression Correlation Coefficients for All Large Channels

			BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.704(**)	.719(**)	.684(**)	.687(**)	.710(**)	.723(**)	.682(**)	.687(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
		N	15120	15120	15120	15120	15120	15120	15120	15120

Table A.38 Nonlinear Regression Correlation Coefficients for All Large Channels Grouped by Type

Type				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
7	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.438(**)	.482(**)	.439(**)	.444(**)	.450(**)	.485(**)	.433(**)	.442(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
8	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.394(**)	.426(**)	.331(**)	.343(**)	.410(**)	.439(**)	.332(**)	.338(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
9	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.814(**)	.818(**)	.805(**)	.806(**)	.817(**)	.819(**)	.805(**)	.808(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040

Table A.39 Nonlinear Regression Correlation Coefficients for All Large Channels Obtained Using Only Independent Variables Relative to the Previous Hour of Channel Activity

			BP1	BP2
Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.707(**)	.693(**)
		Sig. (2-tailed)	.000	.000
		N	15120	15120

Table A.40 Nonlinear Regression Correlation Coefficients for All Large Channels Obtained Using Only Independent Variables Relative to the Previous Hour of Channel Activity and Grouped by Type

Type				BP1	BP2
7	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.431(**)	.440(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
8	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.372(**)	.368(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
9	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.814(**)	.804(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040

Table A.41 Nonlinear Regression Correlation Coefficients for All Low-Intensity Channels

			BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.559(**)	.558(**)	.576(**)	.577(**)	.553(**)	.553(**)	.576(**)	.577(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
		N	15120	15120	15120	15120	15120	15120	15120	15120

Table A.42 Nonlinear Regression Correlation Coefficients for All Low-Intensity Channels Grouped by Type

Type				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
1	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.306(**)	.306(**)	.351(**)	.351(**)	.287(**)	.287(**)	.351(**)	.351(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
4	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.582(**)	.582(**)	.594(**)	.594(**)	.577(**)	.577(**)	.594(**)	.594(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
7	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.536(**)	.535(**)	.555(**)	.556(**)	.532(**)	.533(**)	.555(**)	.556(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040

Table A.43 Nonlinear Regression Correlation Coefficients for All Low-Intensity Channels Obtained Using Only Independent Variables Relative to the Previous Hour of Channel Activity

			BP1	BP2
Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.577(**)	.577(**)
		Sig. (2-tailed)	.000	.000
		N	15120	15120

Table A.44 Nonlinear Regression Correlation Coefficients for All Low-Intensity Channels Obtained Using Only Independent Variables Relative to the Previous Hour of Channel Activity and Grouped by Type

Type				BP1	BP2
1	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.470(**)	.470(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
4	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.597(**)	.597(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
7	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.545(**)	.546(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040

Table A.45 Nonlinear Regression Correlation Coefficients for All Medium-Intensity Channels

			BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.571(**)	.572(**)	.560(**)	.559(**)	.570(**)	.571(**)	.565(**)	.563(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
		N	15120	15120	15120	15120	15120	15120	15120	15120

Table A.46 Nonlinear Regression Correlation Coefficients for All Medium-Intensity Channels Grouped by Type

Type				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
2	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.594(**)	.594(**)	.586(**)	.590(**)	.593(**)	.592(**)	.587(**)	.588(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
5	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.597(**)	.601(**)	.612(**)	.606(**)	.599(**)	.599(**)	.610(**)	.604(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
8	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.511(**)	.511(**)	.471(**)	.473(**)	.506(**)	.511(**)	.488(**)	.488(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040

Table A.47 Nonlinear Regression Correlation Coefficients for All Medium-Intensity Channels Obtained Using Only Independent Variables Relative to the Previous Hour of Channel Activity

			BP1	BP2
Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.572(**)	.572(**)
		Sig. (2-tailed)	.000	.000
		N	15120	15120

Table A.48 Nonlinear Regression Correlation Coefficients for All Medium-Intensity Channels Obtained Using Only Independent Variables Relative to the Previous Hour of Channel Activity and Grouped by Type

Type				BP1	BP2
2	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.586(**)	.586(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
5	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.611(**)	.611(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
8	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.508(**)	.508(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040

Table A.49 Nonlinear Regression Correlation Coefficients for All High-Intensity Channels

			BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.699(**)	.710(**)	.678(**)	.687(**)	.698(**)	.711(**)	.678(**)	.687(**)
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
		N	15120	15120	15120	15120	15120	15120	15120	15120

Table A.50 Nonlinear Regression Correlation Coefficients for All High-Intensity Channels Grouped by Type

Type				BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8
3	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.208(**)	.238(**)	.145(**)	.156(**)	.207(**)	.250(**)	.148(**)	.156(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
6	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.354(**)	.369(**)	.312(**)	.335(**)	.353(**)	.371(**)	.310(**)	.336(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040
9	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.799(**)	.805(**)	.791(**)	.796(**)	.799(**)	.805(**)	.792(**)	.796(**)
			Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
			N	5040	5040	5040	5040	5040	5040	5040	5040

Table A.51 Nonlinear Regression Correlation Coefficients for All High-Intensity Channels Obtained Using Only Independent Variables Relative to the Previous Hour of Channel Activity

			BP1	BP2
Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.707(**)	.718(**)
		Sig. (2-tailed)	.000	.000
		N	15120	15120

Table A.52 Nonlinear Regression Correlation Coefficients for All High-Intensity Channels Obtained Using Only Independent Variables Relative to the Previous hour of Channel Activity and Grouped by Type

Type				BP1	BP2
3	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.227(**)	.282(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
6	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.389(**)	.434(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040
9	Spearman's rho	ObservedPostersPlus1	Correlation Coefficient	.801(**)	.806(**)
			Sig. (2-tailed)	.000	.000
			N	5040	5040

REFERENCES

1. Adomavicius, G. and Tuzhilin, A. (2001). "Extending Recommender Systems: A Multidimensional Approach." Proceedings of the International Joint Conference on Artificial Intelligence, Workshop on Intelligent Techniques for Web Personalization, August 04-06, Seattle, Washington, USA.

2. Adomavicius, G. and Tuzhilin, A. (2005). "Personalization technologies: a process-oriented perspective." Communications of the ACM 48(10), pp. 83-90.

3. Adomavicius, G., Sankaranarayanan, R., Sen, S., and Tuzhilin, A. (2005). "Incorporating contextual information in recommender systems using a multidimensional approach." ACM Transactions on Information Systems 23(1), pp. 103-145.

4. Alvestrand, H. (2002). "Instant Messaging and Presence on the Internet." Brief published on the *Internet Society* website. Retrieved January 1, 2006 from: http://www.isoc.org/briefings/009/briefing09.pdf.

5. Bartle, R. (1990). "Interactive Multi-User Computer Games." Report commissioned by British Telecom Research Laboratories. Retrieved November 10 2005 from: http://www.mud.co.uk/richard/imucg.htm.

6. Bechar-Israeli, H. (1996). "From <Bonehead> to <cLoNeHeAd>: Nicknames, Play and Identity on Internet Relay Chat." Journal of Computer Mediated Communication 1(2).

7. Begole, J., Tang, J., Smith R. and Yankelovich, N. (2002). "Work Rhythms: Analyzing Visualizations of Awareness Histories of Distributed Groups." Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work, pp. 334-343, November 16-20, 2002, New Orleans, Louisiana, USA.

8. Bradner, E., Kellogg, W. and Erickson, T. (1999). "The Adoption and Use of 'Babble': A Field Study of Chat in the Workplace." Proceedings of the Sixth European Conference on Computer Supported Cooperative Work, pp. 139-158, August 1999, Copenghagen, Denmark.

9. Bruckman, A. (1992). "Identity Workshop: Emergent Social and Psychological Phenomena in Text-Based Virtual Reality." MIT Media Labs unpublished article. Retrieved November 10 2005 from: http://www.cc.gatech.edu/~asb/papers/old-papers.html.

10. Bruckman, A. (1993). "Gender Swapping on the Internet." Proceedings of the 1993 Conference on Internet Society, August 17-20, San Francisco, California, USA.

11. Bruckman, A. (1994). "Programming for Fun: MUDs as a Context for Collaborative Learning." Presented at the National Educational Computing Conference, June 1994, Boston, Massachusetts, USA.

12. Bruckman, A. (1998). "Community Support for Constructional Learning." Computer Supported Cooperative Work 7(1-2), pp. 47-86.

13. Bruckman, A. and Resnick, M. (1993). "Virtual Professional Community: Results from the MediaMOO Project." Presented at the Third International Conference on Cyberspace, May 1993, Austin, Texas, USA.

14. Bruckman, A. and Resnick, M. (1995). "The MediaMOO Project: Constructionism and Professional Community." Convergence: The International Journal of Research into New Media Technologies 1(1), pp. 94-109.

15. Butler, B. (2001). "Membership Size, Communication Activity, and Sustainability: A Resource-Based Model of Online Social Structures." Information Systems Research 12(4), pp. 346-362.

16. Byrne, E. (1994). "Formation of Relationships on Internet Relay Chat." A thesis submitted to the University of Western Sydney, Nepean as partial fulfillment for the degree of Bachelor of Arts in Applied Communication Studies. Retrieved November 10 2005 from: http://www.irchelp.org/irchelp/communication-research/academic/byrne-e-cyberfusion-1993/thesis1.html.

17. Cherny, L. (1994). "Gender Differences in Text-Based Virtual Reality." Proceedings of the Third Berkeley Women and Language Conference, Berkeley, California, USA.

18. Chuah, M. (2003). "Reality Instant Messaging: Injecting a Dose of Reality into Online Chat." CHI 2003 Extended Abstracts on Human Factors in Computing Systems, pp. 926 - 927, April 05-10, Ft. Lauderdale, Florida, USA.

19. Churchill, E. and Bly, S. (1999). "It's all in the words: supporting work activities with lightweight tools." Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work, pp. 40-49, November 14-17, Phoenix, Arizona, USA.

20. Churchill, E. and Bly, S. (1999). "Virtual Environments at Work: ongoing use of MUDs in the Workplace." Proceedings of the International Joint Conference on Work Activities Coordination and Collaboration, pp. 99-108, February 22-25, San Francisco, California, USA.

21. Cohen, D., Jacovi, M., Maarek, Y. S. Soroka, V. (2002). "Livemaps for collection awareness." International Journal of Human-Computer Studies 56(1): pp. 7-23.

22. Cowell, A. J., Gregory, M. L., Bruce, J., Haack, J., Love, D., Rose, S., and Andrew, A. H. (2006). "Understanding the dynamics of collaborative multi-party discourse." Information Visualization 5(4), pp. 250-259.

23. Cummings, J., Butler, B. and Kraut, R. (2002). "The Quality of Online Social Relationships." Communications of the ACM 45(7): pp. 103-108.

24. Curtis, P. (1996). "Mudding: Social Phenomena in text-based virtual realities." Internet Dreams: Archetypes, Myths and Metaphors, MIT Press: pp. 265-292.

25. Curtis, P., M. Dixon, R. Frederick and D. Nichols (1995). "The Jupiter Audio/Video Architecture: Secure Multimedia in Network Places." Proceedings of the Third ACM International Conference on Multimedia, pp. 79-90, November 05-09, San Francisco, California, USA.

26. Curtis, P. and D. Nichols (1993). "MUDs Grow Up: Social Virtual Reality in the Real World." Xerox PARC unpublished report Retrieved November 10 2005 from: ftp://parcftp.xerox.com/pub/MOO/papers/MUDsGrowUp.ps.

27. Danet, B., Ruedenberg-Wright, L. Rosenbaum-Tamari, Y. (1997). "Hmmmm... Where's That Smoke Coming From? Writing, Play and Performance on Internet Relay Chat." Journal of Computer Mediated Communication 2(4).

28. Danet, B., Wachenhauser, T., Bechar-Israeli, H., Cividalli, A. and Rosenbaum-Tamari, Y. (1996). "Curtain Time 20:00 GMT: Experiments with Virtual Theater on Internet Relay Chat." Journal of Computer Mediated Communication 1(2).

29. December, J. (1993). "Characteristics of Oral Culture in Discourse on the Net." Presented at the Twelfth Annual Penn State Conference on Rhetoric and Composition, July 8, University Park, Pennsylvania, USA.

30. Degemmis, M., Lops, P., and Semeraro, G. (2007). "A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation." User Modeling and User-Adapted Interaction 17(3), pp. 217-255.

31. Dewes, C., Wichmann, A. and Feldman, A. (2003). "An Analysis of Internet Chat Systems." Proceedings of the Third ACM SIGCOMM Conference on Internet Measurement, pp. 51-64, October 27-29, Miami Beach, Florida, USA.

32. DiMicco, J. M., Hollenbach, K. J., and Bender, W. (2006). "Using visualizations to review a group's interaction dynamics." CHI 2006 Extended Abstracts on Human Factors in Computing Systems, pp. 706-711, April 22 – 27, Montréal, Québec, Canada.

33. Donath, J., Karahalios, K. and Viegas, F. (1999). "Visualizing Conversation." Journal of Computer Mediated Communication 4(4).

34. Dourish, P. (1998). "Introduction: The State of Play." Computer Supported Cooperative Work 7(1-2), pp. 1-7.

35. Erickson, T. (2003). "Designing Visualizations of Social Activity: Six Claims." CHI 2003 extended abstracts on Human Factors in Computing Systems, pp. 846-847, April 05-10, Ft. Lauderdale, Florida, USA.

36. Erickson, T., Halverson, C., Kellogg, W., Laff, M. and Wolf, T. (2002). "Social Translucence: Designing Social Infrastructures that Make Collective Activity Visible." Communications of the ACM 45(4), pp. 40-44.

37. Erickson, T. and Kellogg, W. (2000). "Social Translucence: An Approach to Designing Systems that Support Social Processes." ACM Transactions on Computer-Human Interaction 7(1), pp. 59-83.

38. Erickson, T. and Laff, M. (2001). "The Design of the 'Babble' Timeline: A Social Proxy for Visualizing Group Activity Over Time." CHI 2001 extended abstracts on Human Factors in Computing Systems, pp. 329-330, March 31-April 05, Seattle, Washington, USA.

39. Erickson, T., Smith, D., Kellogg, W., Laff, M., Richards, J. and Bradner, E. (1999). "Socially Translucent Systems: Social Proxies, Persistent Conversation and the Design of 'Babble'." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: the CHI is the Limit, pp. 72-29, May 15-20, Pittsburgh, Pennsylvania, USA.

40. Evard, R. (1993). "Collaborative Networked Communication: MUDs as System Tools." Proceedings of the 7th USENIX conference on System Administration, pp. 1-8, November 01-05, Monterey, California, USA.

41. Fanderclai, T. L. (1995). "MUDs in Education: New Environments, New Pedagogies." Computer-Mediated Communication Magazine, 2(1).

42. Fiore, A. T. and Smith, M. A. (2002). "Treemap Visualizations of Newsgroups." Interactive poster presented at the IEEE Symposium on Information Visualization, Boston, Massachusetts.

43. Fisher, D. and Dourish, P. (2004). "Social and Temporal Structures in Everyday Collaboration." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 551-558, April 24-29, Vienna, Austria.

44. Fulk, J., Flanagin, A., Kalman, M., Monge, P.R , and Ryan, T., (1996). "Connective and communal public goods in interactive communication systems." Communication Theory 6, pp. 60-87.

45. Fulk, J., Heino, R., Flanagin, A. J., Monge, P. R., and Bar, F. (2004). "A Test of the Individual Action Model for Organizational Information Commons." Organization Science 15(5), pp. 569-585

46. Garbis, C. and Waern, Y. (1997). "Design and Use of MUDs for Serious Purposes: Report from a workshop at the CSCW conference in Boston 16th November '96." ACM SIGCHI Bulletin, 29(3), pp. 31-33.

47. Garcia, A. C. and Jacobs, J. B. (1999). "The eyes of the beholder: Understanding the turn-taking system in quasi-synchronous computer-mediated communication." Research on Language and Social Interaction 32(4), pp. 337-367.

48. Gelhausen, A. (2004). "Internet Relay Chat Statistics". Retrieved on November 10 2005 from http://irc.netsplit.de.

49. Greenberg, S. (1996). "Peepholes: Low Cost Awareness of One's Community." Conference companion on Human Factors in Computing Systems: common ground, pp. 206-207, April 13-18, Vancouver, British Columbia, Canada.

50. Grinter, R. and Eldridge, M. (2003). "Wan2tlk?: Everyday Text Messaging." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 441-448, April 05-10, Ft. Lauderdale, Florida, USA.

51. Grinter, R. and Palen, L. (2002). "Instant Messaging in Teen Life." Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work, pp. 21-30, November 16-20, New Orleans, Louisiana, USA.

52. Guzdial, M. (1997). "A Shared Command Line in a Virtual Space: The Working Man's MOO." Proceedings of the 10th Annual ACM Symposium on User Interface Software and Technology, pp. 73-74, October 14-17, Banff, Alberta, Canada.

53. Halverson, C., Erickson, T. and Sussman, J. (2003). "What Counts as Success? Punctuated Patterns of Use in a Persistent Chat Environment." Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, pp. 180-189, November 09-12, 2003, Sanibel Island, Florida, USA.

54. Hamilton, J. (2004). "IRC search engine." Statistics of IRC networks available on the World Wide Web: http://searchirc.com/.

55. Handel, M. and Herbsleb, J. (2002). "What is Chat Doing in the Workplace?" Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work, pp. 1-10, November 16-20, 2002, New Orleans, Louisiana, USA.

56. Hansen, K. M. and Damm, C. H. (2002). "Instant Collaboration: Using Context-Aware Instant Messaging for Session Management in Distributed Collaboration Tools." Proceedings of the Second Nordic Conference on Human-Computer Interaction, pp. 279-282, October 19-23, 2002, Aarhus, Denmark.

57. Haveliwala, T. (2002). "Search Facilities for Internet Relay Chat." Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, p. 395, July 14-18 Portland, Oregon, USA.

58. Herbsleb, J., Atkins, A., Boyer, D., Handel, M. and Finholt, T. (2002). "Introducing Instant Messaging and Chat in the Workplace." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing our World, Changing Ourselves, pp. 171-178, April 20-25, Minneapolis, Minnesota, USA.

59. Herring, S. (1999). "Interactional Coherence in CMC." Journal of Computer Mediated Communication 4(4).

60. Hiltz, S. R. and Turoff, M. (1993). *The network nation: human communication via computer.* MIT Press, Cambridge, Massachusetts, USA.

61. Hinner, K. (2000). "Statistics of Major IRC Networks: Methods and Summary of User Count." Journal of Media Culture 3(4).

62. Hinner, K. (2004). "IRC Statistics." Statistics of IRC networks available on the World Wide Web: http://www.hinner.com/ircstat.

63. Isaacs, E., Kamm, C., Schiano, D., Walendowski, A. and Whittaker, S. (2002). "Characterizing Instant Messaging from Recorded Logs." CHI 2002 extended abstracts on Human Factors in Computing Systems, pp. 720-721, April 20-25, Minneapolis, Minnesota, USA.

64. Isaacs, E., Tang, J. and Morris, T. (1996). "Piazza: A desktop environment supporting impromptu and planned interactions." Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work, pp. 315-324, November 16-20, Boston, Massachusetts, USA.

65. Isaacs, E., Walendowski, A. and Ranganthan, D. (2002). "Hubbub: A sound-enhanced mobile instant messenger that supports awareness and opportunistic interactions." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing our World, Changing Ourselves, pp. 179-186, April 20-25, Minneapolis, Minnesota, USA.

66. Isaacs, E., Walendowski, A. and Ranganthan, D. (2002). "Mobile Instant Messaging through Hubbub." Communications of the ACM 45(9), pp. 68-72.

67. Isaacs, E., Walendowski, A., Whittaker, S., Schiano, D., and Kamm, C. (2002). "The Character, Functions and Styles of Instant Messaging in the Workplace." Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work, pp. 11-20, November 16-20, New Orleans, Louisiana, USA.

68. Jacovi, M., Soroka, V. and Ur, S. (2003). "Why do we ReachOut: functions of a semi-persistent peer-support tool." Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, pp. 161-169, November 09-12, Sanibel Island, Florida, USA.

69. Jones, Q. (1997). "Virtual Communities, Virtual Settlements and Cyber-Archaeology: A Theoretical Outline." Journal of Computer Mediated Communication 3(3).

70. Jones, Q. (2003). "Applying Cyber Archaeology." Proceedings of the Eighth European Conference on Computer Supported Cooperative Work, pp. 41-60, September 14-18, Helsinki, Finland.

71. Jones, Q. and Rafaeli, S. (1999). "User population and User Contributions to Virtual Publics: A Systems Model." Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work, pp. 239-248, November 14-17, Phoenix, Arizona, United States.

72. Jones, Q. and Rafaeli, S. (2000). "What do Virtual 'Tells' Tell? Placing Cybersociety Research Into a Hierarchy of Social Explanation." Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 1, p. 1011, January 04-07, Maui, Hawaii, USA.

73. Jones, Q., Ravid, G. and Rafaeli, S. (2002). "An Empirical Exploration of Mass Interaction System Dynamics: Individual Information Overload and Usenet Discourse." Proceedings of the 35th Annual Hawaii International Conference on System Sciences -Volume 4, p. 114.2, January 07-10, Big Island, Hawaii, USA.

74. Jones, Q., Ravid, G. and Rafaeli, S. (2004). "Information Overload and the Message Dynamics of Online Interaction Spaces: a Theoretical Model and Empirical Exploration." Information Systems Research 15(2), pp. 194-210.

75. Kerr, E. and Hiltz. S. R. (1982). *Computer-Mediated Communication Systems: Status and Evaluation.* Academic Press, New York.

76. Koh, J., Kim, Y., Butler, B., and Bock, G. (2007). "Encouraging participation in virtual communities." Communications of the ACM 50(2), pp. 68-73.

77. Kurlander, D., Skelly, T. and Salesin, D. (1996). "Comic Chat." Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, pp. 225-236, August 04-09, New Orleans, Louisiana, USA.

78. Lampe, C. A., Johnston, E., and Resnick, P. (2007). "Follow the reader: filtering comments on slashdot." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1253-1262, April 28 – May 03, San Jose, California, USA.

79. Last, M., Daniels, M., Hause, M. and Woodroffe, M. (2002). "Learning from students: continuous improvement in international collaboration." Proceedings of the 7th Annual Conference on Innovation and Technology in Computer Science Education, pp. 136-140, June 24-28, Aarhus, Denmark.

80. Licklider, J. C. R. and Taylor, R. W. (1968). "The Computer as a Communication Device." Science and Technology, April 1968.

81. Liu, G. (1999). "Virtual Community Presence in Internet Relay Chatting." Journal of Computer Mediated Communication 5(5).

82. Ljungstrad, P. and Hard, Y. (2000). "Awareness of Presence, Instant Messaging and WebWho." ACM SIGGROUP Bulletin 21(3), pp. 21-27.

83. Markus, L. (1987). "Toward a 'Critical Mass' Theory of Interactive Media: Universal Access, Interdependence and Diffusion." Communication Research 14(5), pp. 491-511.

84. Masinter, L. and Ostrom, E. (1993). "Collaborative Information Retrieval: Gopher from MOO." Proceedings of the 1993 Conference on Internet Society, August 17-20, San Francisco, California, USA.

85. McCarthy, J. S. and Meidal, E. S. (1999). "ACTIVE MAP: A visualization tool for location awareness to support informal interactions." Proceedings of the First International Symposium on Handheld and Ubiquitous Computing, pp. 158-170, September 27-29, Karlsruhe, Germany.

86. McDonald, D. W. (2001). "Evaluating Expertise Recommendations." Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work, pp. 214-223, September 30-October 03, Boulder, Colorado, USA.

87. McDonald, D. W. and Ackerman, M. S. (2000). "Expertise Recommender: A Flexible Recommendation System and Architecture." Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work, pp. 315-324, November 14-18, Seattle, Washington, USA.

88. Mehlenbacher, B., Hardin, B., Barrett, C. and Clagett, J. (1994). "Multi-User Domains and Virtual Campuses: Implications for Computer-Mediated Collaboration and Technical Communication." Proceedings of the 11th Annual International Conference on Systems Documentation, pp. 209-222, October 05-08, Waterloo, Ontario, Canada.

89. Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A., and Riedl, J. (2003). "MovieLens Unplugged: Experiences with an Occasionally Connected Recommender System." Proceedings of the 8th International Conference on Intelligent User Interfaces, January 12-15, pp .263-266, Miami, Florida, USA.

90. Mock, K. (2002). "The Use of Internet Tools to Supplement Communication in the Classroom. " Journal of Computer Sciences in Colleges 17(2), pp. 14-21.

91. Monge, P.R., Fulk, J., Kalman, M., Flanagin, A., Parnassa, C., and Rumsey, S. (1998). "Production of collective action in alliance-based interorganizational communication and information systems." Organization Science 9, pp. 411-433.

92. Muller, M., Raven, M. E., Kogan, S., Millen, D., and Carey, K. (2003). "Introducing Chat into Business Organizations: Toward an Instant Messaging Maturity Model." Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work, pp. 50-57, November 09-12, 2003, Sanibel Island, Florida, USA.

93. Muramatsu, J. and Ackerman, M. (1998). "Computing, Social Activity and Entertainment: A Field Study of a Game MUD." Computer Supported Cooperative Work 7(1-2), pp. 87-122.

94. Mutton, P. (2004). "Inferring and Visualizing Social Networks on Internet Relay Chat." Proceedings of the Eighth International Conference on Information Visualization, pp. 35-43, July 14-16, London, United Kingdom.

95. Mynatt, E., O'Day, V., Adler, A. and Ito, M. (1998). "Network Communities: Something Old, Something New, Something Borrowed...." Computer Supported Cooperative Work 7(1-2), pp. 87-122.

96. Nardi, B., Whittaker, S. and Bradner, E. (2000). "Interaction and Outeraction: Instant Messaging in Action." Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, pp. 79-88, December 02-06, Philadelphia, Pennsylvania, USA.

97. Neal, L. (1997). "Virtual Classrooms and Communities." Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work: the Integration Challenge, pp. 81-90, November 16-19, Phoenix, Arizona, USA.

98. Newhagen, J. and Rafaeli, S. (1996). "Why Communication Researchers Should Study the Internet: A Dialogue." Journal of Computer Mediated Communication 1(4).

99. Nichols, D., Curtis, P., Dixon, M. and Lamping, J. (1995). "High-Latency, Low-Bandwidth Windowing in the Jupiter Collaboration System." Proceedings of the 8th annual ACM Symposium on User Interface and Software Technology, pp. 111-120, November 15-17, Pittsburgh, Pennsylvania, USA.

100. Nonnecke, B. and Preece, J. (2000). "Persistence and Lurkers in Discussion Lists: A Pilot Study." Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 3, p. 3031, January 04-07, Maui, Hawaii, USA.

101. O'Connor, M., Cosley, D., Konstan, J. A. and Riedl, J. (2001). "PolyLens: A recommender system for groups of users." Proceedings of the Seventh European Conference on Computer Supported Cooperative Work, pp. 199-218, September 16-20, Bonn, Germany.

102. O'Day, V., Bobrow, D., Bobrow, K., Shirley, M., Hughes, B. and Walters, J. (1998). "Moving Practice: From Classrooms to MOO Rooms." Computer Supported Cooperative Work 7(1-2), pp. 9-45.

103. Oldenburg, R. (1991). *The Great Good Place.* Marlowe & Company, New York.

104. Oliver, P. and G. Marwell (2001). "Whatever Happened to Critical Mass Theory: A Retrospective and Assessment." Sociological Theory 19(3), pp. 292-311.

105. Oliver, P., Marwell, G. and Teixeira, R. (1985). "A Theory of the Critical Mass. I. Interdependence, Group Heterogeneity and the Production of Collective Action." The American Journal of Sociology 91(3), pp. 522-556.

106. Olson, M. J. (1965). "The Logic of Collective Action: Public Goods and the Theory of Groups." Harvard University Press, Cambridge, Massachusetts.

107. Pacagnella, L. (1997). "Getting the Seats of Your Pants Dirty: Strategies for Ethnographic Research on Virtual Communities." Journal of Computer Mediated Communication 3(1).

108. Paolillo, J. (1999). "The Virtual Speech Community: Social Network and Language Variation on IRC." Journal of Computer Mediated Communication 4(4).

109. Persson, P. (2003). "ExMS: an Animated and Avatar-based Messaging System for Expressive Peer Communication." Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, pp. 22-24, November 09-12, Sanibel Island, Florida, USA.

110. Pilgrim, C. J. and Leung, Y. K. (1996). "Appropriate use of the Internet in computer science courses." Proceedings of the First Conference on Integrating Technology into Computer Science Education, pp. 81-86, June 02-06, Barcelona, Spain.

111. Preece, J. (1999). "Empathic communities: Balancing emotional and factual communication." Interdisciplinary Journal of Human-Computer Interaction 12(1), pp. 63-77.

112. Raban, D. R. and Rafaeli, S. (2007). "Investigating ownership and the willingness to share information online." Computers in Human Behavior 23(5), pp. 2367-2382.

113. Rafaeli, S. and LaRose, R. J. (1993). "Electronic Bulletin Boards and 'Public Good' Explanations of Collaborative Mass Media." Communication Research 20(2), pp. 277-297.

114. Reid, E. (1991). "Electropolis: Communication and Community on Internet Relay Chat." An adaptation from an Honors thesis written at the University of Melbourne (Australia) in 1991. Retrieved November 10 2005 from: http://www.irchelp.org/irchelp/misc/electropolis.html.

115. Ribak, A., M. Jacovi and V. Soroka (2002). "Ask Before You Search: Peer Support and Community Building with ReachOut." Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work, pp. 126-135, November 16-20, New Orleans, Louisiana, USA.

116. Rich, E. (1998). "User modeling via stereotypes." Readings in intelligent User interfaces. Morgan Kaufmann Publishers Inc., San Francisco, California, USA.

117. Rintel, E. S., Mulholland, J., and Pittam, J. (2001). "First Things First: Internet Relay Chat Openings." Journal of Computer Mediated Communication 6(3).

118. Rodino, M. (1997). "Breaking out of Binaries: Reconceptualizing Gender and its Relationship to Language in Computer-Mediated Communication." Journal of Computer Mediated Communication 3(3).

119. Schafer, W., Bowman, D. and Carroll, M. (2002). "Map-Based Navigation in a Graphical MOO." Crossroads 9(1), pp. 8-15.

120. Schiano, D. and White, S. (1998). "The First Noble Truth of Cyberspace: People are People (Even When They MOO)." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 352-359, April 18-23, 1998, Los Angeles, California, USA.

121. Sempsey, J. (1995). "The Psycho-Social Aspects of Multi-User Dimensions in Cyberspace." A literature review. Retrieved November 10 2005 from: http://www.brandeis.edu/pubs/jove/HTML/v2/sempsey.html.

122. Shneiderman, B. (1992). "Tree visualization with treemaps: a 2-d space-filling approach." ACM Transactions on Graphics 11(1), pp. 92-99.

123. Smith, M. (1999). "Invisible Crowds in Cyberspace: Mapping the Social Structure of the USENET." Communities in Cyberspace: Perspectives on New Forms of Social Organization, Routledge Press, London, UK.

124. Smith, M., Cadiz, J. J. and Burkhalter, B. (2000). "Conversation Trees and Threaded Chats." Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, pp. 97-105, December 02-06, Philadelphia, Pennsylvania, USA.

125. Smith, M., Farnham, S. and Drucker, M. (2000). "The Social Life of Small Graphical Chat Spaces." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 462-469, April 01-06, The Hague, The Netherlands.

126. Soroka, V. and Rafaeli, S. (2006). "Invisible participants: how cultural capital relates to lurking behavior." Proceedings of the 15th international Conference on World Wide Web, pp. 163-172, May 23 – 26, Edinburgh, Scotland.

127. Svensson, M., Hook, K., Laaksolahti, J. and Waern, A. (2001). "Social navigation of food recipes." Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pp. 341-348, March 2001, Seattle, Washington, USA.

128. Tabachnick, B.G. and Fidell, L.S. (1996). *Using Multivariate Statistics.* HarperCollins, New York.

129. Tang, J. and Begole, J. (2003). "Beyond Instant Messaging." Queue 1(8), pp. 28-37.

130. Tang, J., Yankelovich, N., Begole, J., Van Kleek, M., Li, F. and Bhalodia, J. (2001). "ConNexus to Awarenex: Extending awareness to mobile users." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 221-228, March 31, - April 5, Seattle, Washington, USA.

131. Tang, J., Yankelovitch, N. and Begole, B. (2000). "ConNexus: Instant Messaging for the Workplace." Sun Microsystems laboratory report. Retrieved November 10 2005 from: http://research.sun.com/netcomm/papers.html.

132. Terveen, L. and Hill, W. (2001). "Beyond Recommender Systems: Helping People Help Each Other." HCI In The New Millennium, J. Caroll, Addison-Wesley.

133. Terveen, L. and McDonald, D. W. (2005). "Social Matching: A Framework and Research Agenda." ACM Transactions on Computer-Human Interaction, 12(3), pp. 401-434.

134. Thomas, P., Carswell, L., Petre, M., Poniatowska, B., Price, B. and Emms, J. (1996). "Distance education over the Internet." ACM SIGCSE Bulletin, 28(SI), pp. 147-149.

135. Thorn, B. and Connolly, T. (1987). "Discretionary Data Bases: A Theory and Some Experimental Findings." Communication Research 14(5), pp. 512-528.

136. Tyler, J. and Tang, J. (2003). "When Can I expect an Email Response: A Study of Rhythms in Email Usage." Proceedings of the Eighth European Conference on Computer Supported Cooperative Work, pp. 239-258, September 14-18, Helsinki, Finland.

137. Van Dyke, N., Lieberman, H. and Maes, P. (1999). "Butterfly: A Conversation-Finding Agent for Internet Relay Chat." Proceedings of the Fourth International Conference on Intelligent User Interfaces, pp. 39-41, January 05-08, Los Angeles, California, USA.

138. Viegas, F. and Donath, J. (1999). "Chat Circles". Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: the CHI is the Limit, pp. 9-16, May 15-20, Pittsburgh, Pennsylvania, USA.

139. Viegas, F. and Smith, M. (2004). "Newsgroup Crowds and AuthorLines: Visualizing the Activity of Individuals in Conversational Cyberspaces." Proceedings of the 37th Annual Hawaii International Conference on System Sciences, p. 40109.2, January 05-08, Waikoloa, Hawaii, USA.

140. Voida, A., Newstetter, W. and Mynatt, E. (2002). "When Conventions Collide: The tensions of Instant Messaging Attributed." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing Our World, Changing Ourselves, pp. 187-194, April 20-25, 2002, Minneapolis, Minnesota, USA.

141. Vronay, D., Smith, M. and Drucker, S. (1999). "Alternative Interfaces for Chat." Proceedings of the Twelfth Annual ACM Symposium on User Interface Software and Technology, pp. 19-26, November 07-10, Asheville, North Carolina, USA.

142. Whittaker, S., Terveen, L., Hill, W. and Cherny, L. (1998). "The dynamics of mass interaction." Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work, pp. 257-264, November 14-18, Seattle, Washington, USA.