K. Kuutti, E. H. Karsten, G. Fitzpatrick, P. Dourish, and K. Schmidt (eds.). *Proceedings of the Eighth European Conference on Computer-Supported Cooperative Work, 14-18 September 2003, Helsinki, Finland*

Applying Cyber-Archaeology

Quentin Jones
New Jersey Institute of Technology, USA
qgjones@acm.org

Abstract. Online spaces that enable public shared inter-personal communications are of significant social and economic importance. This paper outlines a theoretical model and methodology, labeled cyber-archaeology, for researching the relationship between such spaces and the behaviors they contain. The methodology utilizes large-scale field studies into user behavior in online spaces to identify technology-associated user constraints to sustainable patterns of online large-scale shared social interactions.

Empirical research was conducted to assess the validity of both the theoretical model and methodology. It was based on the analysis of 2.65 million messages posted to 600 Usenet newsgroups over a six month period, and 478,240 email messages sent to 487 email lists managed by Listserv software over a 5-month period. Overall, our findings support a key aspect of the model, namely that individual 'information overload' coping strategies have an observable impact on mass-interaction discourse dynamics. Further, that it is possible to demonstrate a link between technology type and information overload impacts through field studies of online behavior.

Cyber-archaeology is discussed in terms of its ability to offer insight into aspects of CMC-tool usability, technology design, and to guide future empirical research.

Introduction

It is widely accepted that the online spaces that enable public shared inter-personal communications are of significant social and economic importance (e.g. Wellman 2001). To date, a very large proportion of research into the behavior of users of online spaces such as interactive email lists, Usenet newsgroups, and bulletin board systems have been in terms of "virtual community" (e.g. Rheingold 1993; Cherny 1999). Researchers have typically utilized one form of social theory or another to guide analysis and paid little attention to the impact on user-behavior of the virtual spaces where shared public online interactions occur. The result is that we have a limited understanding of how the virtual spaces created by different technologies differ in their impact on user interactions. However, it is in the design of these online spaces, rather than user's social networks, where we can often exert the greatest level of control. Therefore, there is a need for CSCW researchers to examine the nature of the relationship between the virtual spaces typically used for shared public online-interactions, their technological platforms, and the behaviors such systems contain. This alternate focus, which is adopted in this paper, shifts the emphasis away from notions of community, and its attention to people and their relations, to the nature of the virtual spaces where shared public interpersonal interactions occur, and the constraints such places impose on online behavior.

To understand the impact of online public interpersonal interaction spaces on behavior an appropriate methodology is needed that is not culture or time specific. Unfortunately, the fact that this question is under-researched means that no clear analytical technique leaps to mind as a preferred or even obvious method of choice. The difficulty in choosing an appropriate methodology is exacerbated by the demand for measures that are relatively culture and time independent, as this means we cannot rely on the in-depth examinations of a small number of unique spaces and associated users' online behavior. This suggests that virtual ethnography in its various forms is not a preferred methodology. At the same time, we cannot use normative statistics to predict in a deterministic fashion the relationship between a technology and user-behavior (Jones 1997). This is because social context determines social outcomes (Spears and Lea 1992), not the enabling technology. For the same reasons it is extremely difficult to design laboratory studies with the necessary ecological validity.

In the sections that follow the cyber-archaeology approach is outlined as a means to understanding and comparing types of computer mediated shared inter-personal interactions spaces. The approach described utilizes large-scale field studies into user behavior in online spaces to identify technology-associated constraints to large-scale interaction dynamics. After outlining the theoretical foundations of the method, a description is given of the empirical research undertaken to assess its validity. This empirical research is based on the analysis of 2.65 million messages posted to 600 Usenet newsgroups over a six month period, and 478,240 email messages sent to 487 email lists managed by Listserv software over a 5-month period.

Theoretical Model

Mobilizing Archaeological Theory for CSCW

Scientists frequently seek to understand new phenomena by using analogies (Steinfeld and Fulk 1987) thereby mobilizing an existing body of knowledge to help explain new phenomena and new situations. Labeling a new phenomenon such as online social structures with a familiar name, such as 'community' or ‘social group’, and others listed in Table I, is useful because it allows authors to effectively communicate and generalize their findings by presenting results in a larger context. Each of the authors listed in Table I is struggling to find terminology to depict the sense of identity and connectedness, which is a feature of cyber-society. While on the surface a number of these metaphors may seem to be interchangeable, each is associated with a different set of assumptions about the significance of various social processes. However, the degree of interchangeability suggests a lack of specificity and a tension between social understanding and research strategy. Further, these analogies are not sufficient for an examination of the impact of the relationship between various computer mediated communication (CMC) technologies and the interactions they support.

Table I. Examples of Metaphors Used for Group-CMC Based Social Structures

Social Structure	**Examples of Authors**
Community	Rheingold 1993
Small Social Group	Sproull and Farj 1997
Social Networks	Wellman 2000
Forum / Discussion Groups	Rojo 1997
Voluntary Network	Butler 2001

The discussion above leads us to distinguish computer mediated social structures from the spaces and places where users gather and perhaps interact online. In Table II metaphors for interaction spaces are listed. Most of these suggest something about underling social processes. However, if the aim is to investigate the relationship between an interaction space and interaction dynamics, as it is here, then we do not want to prejudice research by using terms with connotations about particular social processes. Jones and Rafaeli (2000a) proposed the use of the term virtual public in part, because the label conveys a neutral picture of the social processes that occur within such spaces.

Table II. Examples of Metaphors Used for Places and Spaces

Place	**Examples of Authors**
Chat Rooms	Reid 1991
Team Rooms / Workrooms	Roseman and Greenberg 1996
Conference	Hiltz 1985
Virtual Airport Bar	Doheny-Farina 1996
Cyber-Inns	Coate 1992
Virtual Settlement	Jones 1997
Commons	Kollock and Smith 1996
Virtual Public	Jones 2000a
Information Super highway	Jones 1995

Virtual publics are symbolically delineated computer mediated spaces such as email lists, newsgroup, Internet Relay Chat (IRC) channels etc., 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. The term ‘virtual public’ is adopted here, primarily, for two reasons. Firstly, for our purposes it is important that we distinguish between cybersociety (Jones 1995), virtual communities (Jones 1997), and open public interactions spaces (Jones and Rafaeli 2000a). Secondly, as will become apparent below, we need a simple label to describe online, shared, interpersonal interaction spaces, whose membership and existence is fairly open for both observation and user participation. However, the development of the term of virtual public does not immediately bring to mind any particular research approach.

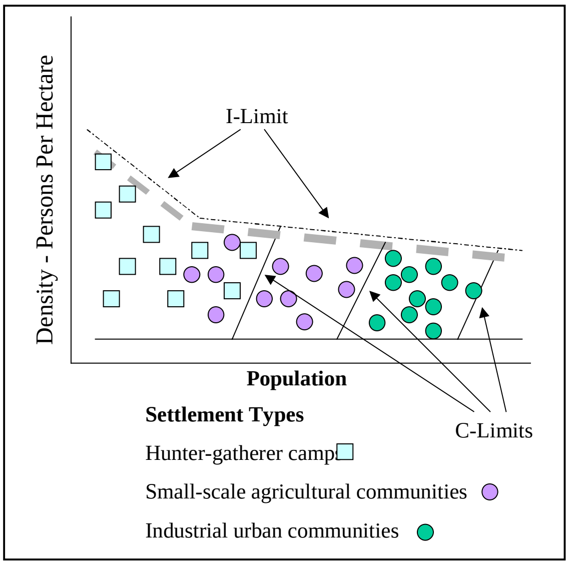
To arrive at a broad understanding of various aspects of human social behavior typically requires examination of findings from a number of research approaches, with the value of each depending upon the issue under investigation. In this case, where an examination is being made of the link between mediated-space and online behavior, a number of other established disciplines that attempt to come to terms with human environments stand out as being of potential value. These include geography, human or bio ecology, architecture, urban design, and archaeology. Each of these disciplines shares and uses techniques and methods originally developed in other fields, including physics and mathematics.

In Jones (1997), and Jones and Rafaeli (2000b), various reasons were listed as to why CSCW researchers might find theories from archaeology of potential value. These include a mutual interest in artifacts, for the archaeologist items like pottery and arrow heads, for HCI researchers items like listserv postings, web site structures, Usenet content, user logs etc. Second, differences between duration of social action and material remains both online and offline. Archeology has had to deal directly with the problem of explanatory scale when examining the relationship between artifacts and society. An understanding of how to deal with this issue is of crucial importance to the construction of valid theories of online behavior. By examining explanatory scale, archaeological theory has been able to produce explanations of the connection between technology and society without recourse to simple technological determinism. Third, although social theory dominates archaeology as it does in CMC research, a significant body of relevant theory exists regarding phenomena that operate over a range of analytical levels. Finally, archaeological theory exists that guides research into the impact of material on human settlements, namely that of Fletcher (1995), that can be adapted to guide CSCW researchers.

Fletchers' (1995) model shows how the material components of settlements play a substantial and essential role in many large-scale transformations of human community life. Material becomes recognizable as an actor without intent, whose operations occur at a scale beyond the limited perceptions of daily community life. Using Fletcher's methodology, cybersociety can also be examined one step removed from social theory, where human intent is not of particular importance and larger-scale relationships between technology and behavior can be observed. Fletcher argues that the starting point for modeling the impact of technology on social structures is to recognize the degree to which material entities can effectively control or aid social life.

Fletcher (1995) mapped various settlement types over the last fifteen thousand years by geographic size and population. What he discovered was a relationship between the upper boundaries of a ratio of community size to residential density, and a society's available technology. Figure 1 below provides a simplified graphical illustration of the results of mapping this relationship. It summarizes the proposed behavioral constraints on the growth of various types of human settlements. The boundaries represent zones rather than rigid, deterministic, instantaneous halt lines. They are indicators of an uncertain range of likelihood within which the behavioral limitations become severe.

Figure 1. Settlement Interaction-Communication Stress Model
Simplified from Fletcher 1995



Density - Persons Per Hectare
I-Limit
Population
C-Limits
Settlement Types
Hunter-gatherer camp □
Small-scale agricultural communities ○
Industrial urban communities ●

The I-Limit in the Figure 1 refers to the interaction limits individuals can cope with and which place a limit on the maximum density of a settlement. For example, hunter-gatherer communities are able to support a higher level of average residential density than industrial urban communities, although their populations are much smaller. The recognition of this relationship has a significant and important impact on our understanding of the growth of human settlement size and population. The C-Limits represent the constraints imposed on population expansion by the maximum extent to which a given assemblage of communication technologies can function adequately. Thus, for example, city populations were able to dramatically increase because of the industrial revolution. Fletcher's approach provides a modeling technique to explore the boundaries of virtual publics.

In a similar fashion to Fletcher's notion of constrained human settlement, and despite technological advances and popular thinking about the limitlessness of cyberspace, it is taken as axiomatic here that each CMC technology acts to enable only a limited range of social interactions. Further, that the range of social interactions enabled and constrained by different CMC technologies will vary. It follows from these assumptions that if we could take the appropriate measures of interactions occurring in different computer mediated shared interpersonal interaction spaces, then we would be able see the ways in which these spaces enable and constrain interactions, and how various types of spaces differ. By applying Fletcher's approach, which is, not time-line or historically focused we hope to mobilize one archaeological theory for CSCW, which we have labeled cyber-archaeology. This will be achieved by first examining the main human constraints to CMC discussed in the literature as limiting online public shared interpersonal interactions, information overload. Second, by outlining a constraints-model of virtual public interaction dynamics and a methodology for testing various associated hypotheses.

Information Overload

The amount of information available to people is growing rapidly so it not surprising that many of us have experienced what is commonly referred to as "information overload", with the associated sensation of being swamped (Shenk 1997). This occurs because the degree to which we can effectively process information is limited by the finite capacity of human cognition. Information overload is defined here as "the state of an individual or system in which excessive communication inputs cannot be processed and utilized, leading to breakdown" (Rogers and Agarwala-Rogers 1975). In the field of psychology where most empirical research into information overload has been conducted, information overload has traditionally been operationalized as information presented at a rate too fast for a person to process (e.g. Gopher and Donchin 1986). In the context of CMC research, information overload has been interpreted in the light of two additional interrelated concepts. First, the delivery of too many communications, that results in individuals receiving more communications than they can respond to. This type of information overload is referred to as 'conversational overload' (Whittaker et. al. 1998). The second, which is termed 'information entropy' (Hiltz and Turoff 1985), is when incoming messages are insufficiently organized by topic or content to be easily recognized as significant or as part of a conversation's history.

Psychologists have recognized for many years that humans have a limited-capacity to store current information in memory (e.g. William James in the 19th century). The analysis of this information overload producing limitation led in the 1950's to foundational work in cognitive psychology. Technologists also recognized early on the need to address the limited ability of people to cope with the vast amounts of information produced in the modern world. Vannevar Bush's landmark paper "As We May Think" (1945) which led to the windows computer interface and the World Wide Web, can in fact be seen as a paper proposing that we need to build better tools for coping with information overload (Simpson et. al. 1996). However, while technologies have helped individuals process more information, and through technologies such as email, to increase the size of their personal social networks (Whittaker, Jones, Terveen 2002), it is shown in this paper that information-processing limits still impact on social interactions observed online.

One of the first scientists to notice the negative social effects of information overload was the sociologist Georg Simmel (1950) who wrote of the overload of sensations in the modern urban world that caused city dwellers to become jaded and develop an incapacity to react to new situations with the appropriate energy. This was followed by the writings of the social psychologist Stanley Milgram who used the concept of information overload to explain bystander behavior (1969). Milgram hypothesized that the bystanders' often disregard events and depersonalize others in their environment as a means of coping with information overload. Since early research into the connection between information overload and city life, researchers have linked information overload to human evolution (Dunbar 1996), settlement size (Fletcher 1995), and as shown below, patterns of inter-personal interaction on the Internet.

As noted in the discussion above, the maximum density and geographic spread of a culture's settlements is also linked to the management of information overload (Fletcher 1995). It is reasoned here that in a similar fashion to interactions in real settlements, sustainable interaction dynamics that occur using virtual publics (Jones and Rafaeli 2000a), such as open interactive email lists, open Internet Relay Chat channels, Usenet newsgroups etc., are constrained by information overload. This occurs because it logically follows that if limitations exist to an individual's ability to effectively process certain virtual public message patterns, then this will impact on the sustainability of such patterns of group-CMC. That is, beyond a particular level of average user communication processing-load, behavioral stress will make existing patterns of public interactive group communication unsustainable. Communication load is the processing effort required to deal with a set of communications.

Individuals can take a range of actions to reduce the impact of information overload resulting from group-CMC. Actions include: making an increased effort; learning new information management techniques to reduce the information overload; failing to respond or attend to certain messages, thereby lowering the growth in communication load; producing simpler responses; storing inputs and responding to them as time permits; making more erroneous responses; and ending all participation in the group communication. It is possible to reduce these seven responses to two primary options for a population of experienced users. The first option is simply to end participation. The second option is to change ongoing communicative behavior. It is hypothesized below that these seven individual responses to overloaded group-CMC can lead to observable impacts of virtual public interaction dynamics.

Although we do not have an exact measure of communication-processing load, we do know that it relates to a number of message system characteristics. Users generally have to make more of an effort to reply coherently to a thread (a chain of messages) than to a single message (Lewis et al. 1997). Higher message interactivity correlates with higher communication-processing load. Similarly, a dense pattern of messages (high frequency of postings) will require quicker and more sustained processing by group members. Therefore, message density will also co-vary with communication-processing load. It is also likely that an increase in interactional-incoherence (e.g. fragmented sentences, disrupted turn adjacency, reversed sequencing, communication lags, agrammatical language, and interactionally disjointed messages) will also increase communication-processing load (Herring 1999).

Abstracting the notion of individual responses to information overload, to impacts on virtual public interaction dynamics, requires a systemic approach to virtual public discourse. This can be achieved by the recognition of virtual public discourse, which is produced by members who are free to vocally participate, lurk, or unsubscribe, as the output of a complex social system (Jones and Rafaeli 2000a, and Jones 2001). Analysis of the systemic nature of social relations can be achieved through a focus on feedback-loops (Forrester 1969).

Figure 3. Virtual Public System Dynamics

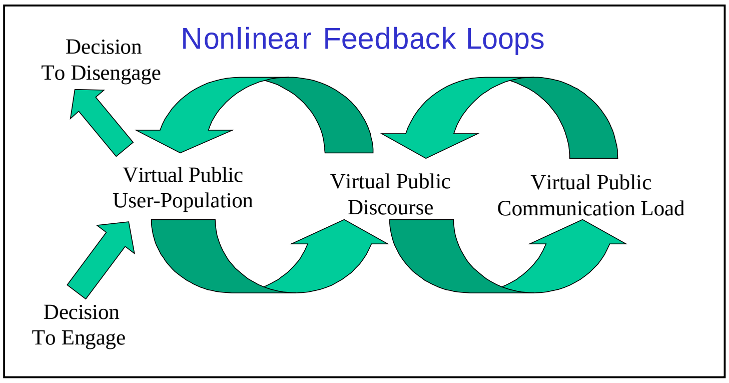


Figure 3 illustrates how the constraints acting on virtual public discourse result in non-linear feedback-loops. It works as follows: An increase in the membership of a virtual public will probably result in an increase in virtual public communication and communication load. However, it will not be possible for individuals to expand their involvement in virtual public communication indefinitely because of limits to the resources available to them to process group communication. Once virtual public communication becomes unmanageable or incoherent to individuals, then, the pattern of their involvement will alter, which in turn will impact on subsequent discourse dynamics.

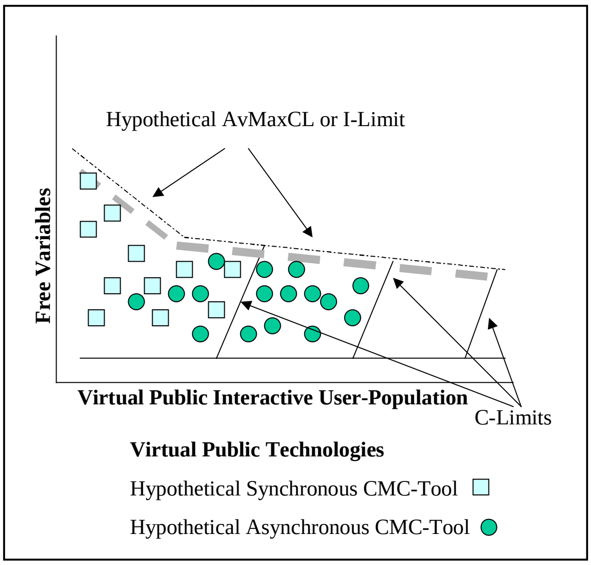
A Method for Comparing Virtual Publics

This section contains an outline of a method for observing empirically some of the impacts of the “Average Maximum Communication Load” [AvMaxCL] individuals are prepared to invest in processing the interaction dynamics of virtual publics. The method is centered on the large-scale observations of “mass interaction”, the shared discourse between hundreds, thousands or more individuals (Whittaker 1996). It is hypothesized that when the structure of group-CMC discourse is close to AvMaxCL, then a further increase in any one part of the communication load function will result in an increased preference by individuals to engage in communicative actions that require less cognitive effort, this will in-turn change the dynamics of the group-CMC in question. For example, at AvMaxCL an increase in the number of interactive users may result in two of the seven individual responses to information overload listed above, shorter reply messages and or a higher turnover of users. In this case, both the proportion of users maintaining an active involvement, and message length, are ‘free variables’ as they are not constrained by the number of active users. Mass interaction, which in many cases is likely to be fairly overloaded, provides a unique opportunity to observe and explore such relationships. This is because the large number of both messages, and users involved in mass interaction should enable observations that may otherwise be hidden by differences between individuals and the social contexts of communication. Not only is it hypothesized that the analysis of mass interaction AvMaxCL effects can lead to an increased understanding of various behavioral implications of a particular type of virtual public technology, but it can also be used for comparative purposes. This is because different CMC-tools are likely to enable different discourse dynamics, so that the way AvMaxCL associated non-linear feedback loops impact on mass interaction should also relate to the CMC-tool type. The theoretical foundations of this aspect of the research are discussed in detail in Jones (2001).

Figure 4 aims to illustrate how the large-scale mapping of virtual public discourse should enable the modeling of the relationship between CMC-technologies and communication load. In so doing it hints at how a research methodology could be developed based on the analysis of virtual public mass interaction. The Figure plots for a number of virtual publics the hypothesized relationships between their user population and various free variables such as average message complexity. This is done in order to graphically display the argument that the stress zones caused by overloaded interactive communication can be identified empirically by mapping active participation in different virtual publics against various components of the communication load function. The plots of the different virtual publics vary widely because of differences in their social context. The hypothetical synchronous virtual publics are plotted as relatively close to the left-axis because they require user co-presence for message exchange, which occurs quickly, and limits group size. On the other hand, the hypothetical plots of asynchronous virtual publics have larger user populations because users can take time to digest and respond to messages. Thus, the Figure displays differences between virtual public technologies.

The approach / model outlined in Figure 4 does not assume that the technology per se will not determine its use. Further, it does not suggest anything about the content of virtual public discourse within the boundaries imposed by technology, nor who will use one virtual public or another. To say something meaningful about the content of discourse, social theory and a focus on context is required. It may also be the case that different types of social aims and social structures for online discourse (e.g. empathetic as opposed to technical; moderated or facilitated, as opposed to unmoderated) may be associated with different stress boundaries. The model does not deal with this issue, although it provides a means for addressing it empirically. The model does not attempt to explain individual variations in the discourse patterns observed; rather it focuses on stress-boundaries, as these provide a key link between technology and discourse structures.

Figure 4. Virtual Public Technology
and Message Processing Capacity


Hypothetical AvMaxCL or I-Limit
Free Variables
Virtual Public Interactive User-Population
C-Limits
Virtual Public Technologies
Hypothetical Synchronous CMC-Tool □
Hypothetical Asynchronous CMC-Tool ●

Hypotheses and Research Methodology

Hypotheses

To test the validity of the theory and method outlined above three hypothesized impacts of individual information overload on virtual public mass interaction dynamics were examined through a study of Usenet Newsgroup postings. A fourth hypothesis was also examined, that the approach can be used to differentiate virtual public technologies, specifically email list and Usenet newsgroup interaction dynamics.

Hypothesis 1: As volume and complexity increase and group-CMC becomes overloaded there will be a decrease in the average complexity of response messages, although this will approach asymptote. The reason for this hypothesized reduction in message complexity is due to the increased effort required by authors to create such messages.

Hypothesis 2: Simple group-CMC messages will be more likely to generate responses than complex messages. When users are confronted with overloaded mass interaction it is hypothesized that they are more likely to fail to respond and / or attend to the messages that are more onerous to process.

Hypothesis 3: As volume and complexity increase and group-CMC becomes overloaded there will be an increased tendency for individuals to end or reduce active participation. Figure 3 above describes this hypothesis, that disengagement is a strategy users will often adopt to cope with overloaded discourse. It follows, then, that on average at average maximum communication load the larger the number of individuals involved in discourse the less stable the participant population.

Hypothesis 4: The collective impact on virtual public mass interaction dynamics of individuals' responses to overloaded group-CMC will relate to virtual public technology type. This hypothesis is explored empirically by examining if the tendency to end or reduce active participation as group-CMC overload increases is linked to virtual public technology type. This measure was chosen because it is a metric that can be easily computed for a wide variety of virtual public technologies.

Methodology

From the above theoretical analysis, it is clear that the method of choice is field research involving the mapping and analysis on a large-scale of naturally occurring patterns of sustained interactive online communication. To implement such field studies and assess the hypotheses outlined we need to: choose appropriate virtual public technologies; collect large data samples; and analyze virtual public interaction dynamics.

Virtual Public Technologies

A number of practical considerations make Usenet Newsgroups a good virtual public technology with which to assess the first three hypotheses. These are: (1) the collection of hundreds to thousands of newsgroup user interactions is relatively straightforward. This is important as prior to this research it was not apparent how large a sample size was required to demonstrate the hypothesized effects. (2) The capture and chaining of inter-user-interactivity data, in this case discussion threads, is straight forward (Liu 1999 outlines why this is not the case with synchronous technologies such as IRC). (3) Anecdotal evidence suggests that a large percentage of Usenet newsgroups are overloaded (Smith 1999, Smith and Fiore 2001). The importance of this third requirement is linked to the need to include in the sample discourse operating at Average Maximum Communication Load (AvMaxCL).

In order to assess the hypothesis that the collective impact of user responses to information overload relates to virtual public technology type, Usenet newsgroups and email list data are compared. This choice was made because of the greater complexities involved in comparing synchronous and asynchronous interactions, and the comparative simplicity of collecting and analyzing email messages.

Data collection

Representative sampling of Usenet discourse is difficult; Whittaker et al's (1998) solution was to produce a randomly stratified sample, of 500 English text based Usenet newsgroups. For this project, 600 newsgroups, using Whittaker et. al.'s approach, 100 of which were moderated, were studied. The full content of 3,293,995 postings were collected over eight months and stored in an Oracle database. The 2,652,552 messages collected over the 6-months from 1st August 1999 to 29th February 2000, were used to conduct this study.

A wide variety of email list management software exists which vary in the ways they deal with subscriptions, postings and user information. Therefore, to reduce potentially unwanted variability, this study focuses on email lists maintained by Listserv. Listserv was the first mailing list management software package. Lsoft, the company that produces Listserv, maintains a database of public Listserv lists called Catalist (www.lsoft.com). On July 28, 1999 there were 24,696 lists contained in Catalist. The lists detailed in Catalist account for approximately 20% of the total number of Listserv lists known to Lsoft. In theory, the provision of Catalist to selected academic researchers makes it possible for researchers to construct a random sample of public Listserv based email discussion lists that are open to the public. For this study 1800 lists were initially extracted from the Catalist database using a stratified random sampling technique, and then a smaller sample was produced through the removal of lists using an iterative process. Lists removed were: non English language based; had a default digest mode of operation; were not active during the entire 5 month period; and did not receive at least 10 messages. This was achieved by subscribing to these lists over a number of weeks. At the end of this process, 478,240 email messages from 487-Listserv email-lists were collected for this study, over 5 months (December 1999 to April 2001). This sample was deemed adequate for the task, although we recognize that ideally the email list sample should exactly parallel the Usenet sample (this difference was the result of the research server being hacked).

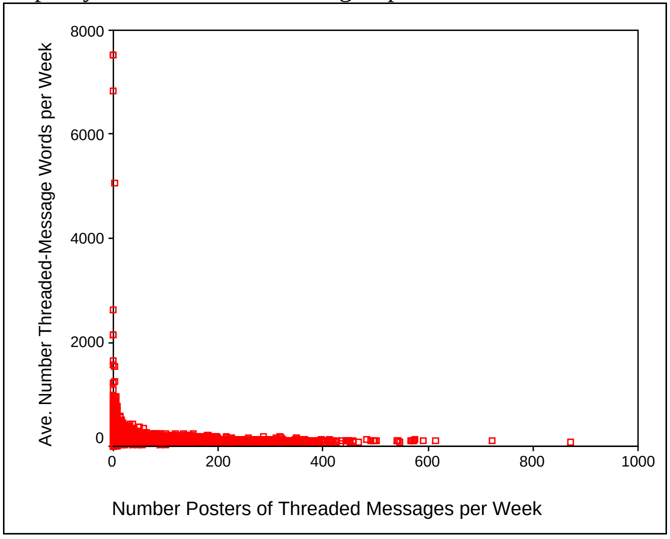
Results

Analysis published elsewhere (Jones, Ravid, and Rafaeli 2002) showed that it was possible to correctly identify Usenet replies in over 99% of cases. Further, that within the data of the Usenet study it was possible to identify approximately 87% of the parent messages of replies. These percentages were deemed adequate for thread reconstruction and the analytical task at hand.

Hypothesis 1: As volume and complexity increase and group-CMC becomes overloaded there will be a decrease in the average complexity of response messages, although this will approach asymptote. This was operationalized as: There will be a decrease in surrogate measures of complexity (e.g. message length) as interactive message communication increases, and or the number of discussion threads increases, although this will approach asymptote.

Figure 5 is a scatter plot of the average number of words in threaded messages (replies) by the number of posters of such messages. The shapes of the curve of this scatter plot, and all the others derived from the various measures of message complexity by various measures of the size of the interactive discussion groups, looked similar. The plot in Figure 5, as do all the other related plots not shown in this paper, displayed the expected relationship between the size of the interactive newsgroup and various surrogate measures of complexity.

Figure 5. Scatter Plot of The Average Number Of Words In Threaded Messages Posted To Newsgroups By The Number Of Newsgroup Posters Of Threaded Messages.



Standard linear regression cannot be used to describe the untransformed relationship between the measures observed because of the Zipf (Zipf 1949, Gunther et. al. 1996) like shape of the curve, with a clearly nonlinear relationship. Further, the curve cannot be tested as a standard Zipf / Power curve because the points on the plot represent means rather than frequencies. Multiple transformations of the variables under study did not succeed in enabling regression modeling using weekly averages for newsgroups. Instead, to perform regression modeling the 1.5 million threaded messages were ranked according to various measures of comparative message complexity. Variables were then computed and matched to individual messages regarding the newsgroup activity during the study week messages were posted. Using these new variables it is possible to see if the number of posters, and or number of interactive threads, is at all predictive of message complexity without the concern of 'regression to the mean'. This approach also allowed for factors such as newsgroup type (e.g. "Comp.", "Misc.", etc.) and message crossposting to be taken into account. Unfortunately, while the ranking and variable matching enabled regression modeling, this approach results in a loss of variance and predictive / explanatory power. As a result, the aim was not to understand the strength of the relationship between newsgroup activity and message size, but rather to simply to see if further support could be found for the notion that group activity related to message complexity. The regression modeling suggested that the newsgroup size (number of threaded messages posted or number of threaded posters) did predict message length (shorter messages being posted to more active groups). Of the five measures of message complexity, each used in different regression models as the dependent variable, the one most strongly predicted by group size was the average number of message lines calculated by the posters client newsreader (F=14836.24, df= 1499124, p < 0.0000). Other influences on message length included the type of newsgroup messages were posted to, the extent of crossposting (messages that were crossposted were longer on average), and the messages' position / depth in a discussion thread (deeper messages were longer overall).

Hypothesis 2: Simple group-CMC messages will be more likely to generate responses than complex messages. There were 593,019 messages that could be considered true unambiguous broadcast or one-way messages. From this sample of one-way messages, 255,697 were found to have initiated (seeded) discussion within their newsgroup during the study period.

As predicted, on average, broadcast messages that seed discourse were significantly smaller/shorter than those that fail to seed discourse. This was consistently the case, no matter how message length was measured. For example, using message header information the mean number of message lines for those that seed discourse was 24.50 lines and for those that did not 62.29 lines.

The outcome of regression modeling was that the following factors are all predictors of a one-way message seeding new discourse: The overall activity of the newsgroup (measured by the number of messages posted per week); All the measures of message complexity such as number of words (examined separately to avoid multi-colinearity); Newsgroup type (e.g. 'talk', 'misc.', etc.); and Moderation status (using a variety of approaches including newsgroup name and newsgroup information center descriptions).

As one would expect, one-way messages posted to larger groups that are more active were more likely to receive a response. Further, all measures of message complexity were found to negatively correlate with seeding discourse (i.e. smaller messages were more likely to seed discourse) and the client header calculation of message length, which is influenced by attachment lengths, was found to be the best predictor. Posting a one-way message to the 'comp' or 'sci' newsgroups resulted in a greater chance of receiving a reply than posting to the other newsgroups, and finally moderation reduced the chances of receiving a reply. The logistic model that appears to best describe the dynamics of discourse seeding was able to predict 63.57% of the cases with a Wald χ^2 of 56559.408, $p > 0.0001$. The findings of the logistic regression modeling argue strongly for the conclusion that smaller messages are more likely to generate ongoing discourse.

Hypothesis 3: As volume and complexity increase and group-CMC becomes overloaded there will be an increased tendency for individuals to end or reduce active participation. For the purposes of this study, poster stability is the percent of posters in a month that also posted in the previous study month. This allowed for the examination of user stability over a 5-month period. Figure 6 displays the number of messages posted to the 578 newsgroups that were active during the first 5 months of the study.

On average only 11.5% of posters sent messages 2 months in a row. Because of the constraints imposed by the proportionality of the stability measure (zero to one hundred) it seems reasonable in this case to also plot on Figure 6 a regression line to highlight the reduction in stability as newsgroup activity increases. The drop in the proportion of individuals involved in sustained discourse is quite strong, with a Spearman's rank correlation coefficient of -.43 ($p < .000$, n=565). When the outliers are removed by only examining the top third of the sample (newsgroup months with more than 2957 messages posted to them, accounting for 1,943,343 of the studies messages) the Spearman's rank correlation coefficient was -0.47 (p < .000, n=192). Linear regression modeling showed the number of posters to be the best predictor of proportional stability, followed by average newsgroup message crossposting (more crossposts results in greater stability), and then newsgroup type (R^2=.43, F=17.7, df=183, P < .001).

Figure 6. Proportional Newsgroup Poster Stability

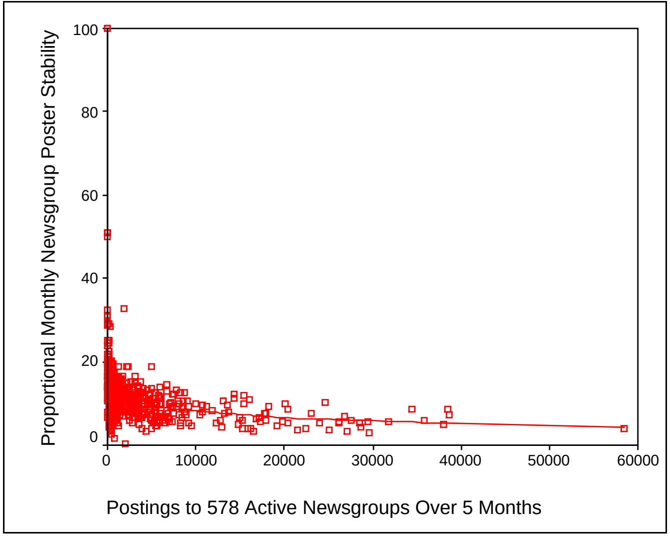
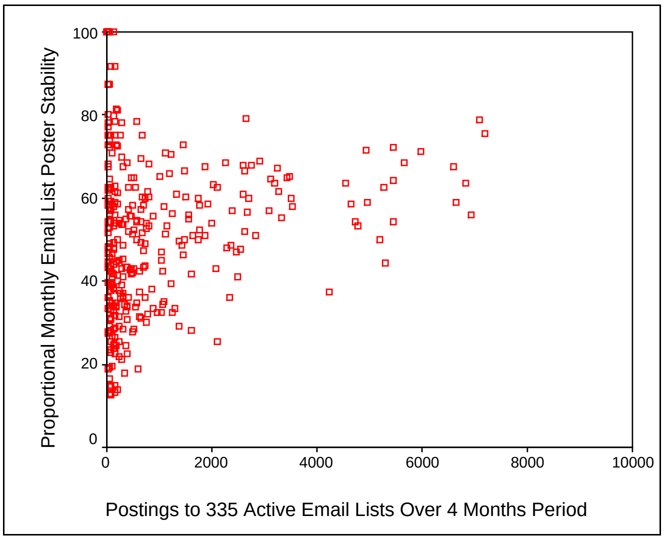


Figure 7. Proportional Listserv List Poster Stability



Hypothesis 4: The collective impact on virtual public mass interaction dynamics of individuals' responses to overloaded group-CMC will relate to virtual public technology type. Of the three hypotheses examined above the easiest both to compute and to compare with other technologies is Hypothesis 3 of ending or reducing active participation as group-CMC overload increases. After an examination of the Listserv list data it was determined that for 335 lists, active data was collected for at least 5 months on a continuous basis, enabling an aggregation of 4-months of poster stability data. This subset of Listserv lists was

then used to create Figure 7, a monthly poster stability plot of Listserv lists. It is equivalent to the Usenet plot except it is over a shorter time-period and has a smaller but sufficient sample size.

The comparison of Figure 7 to Figure 6 highlights just how much more stable the activities of posters are for Listserv email lists than for Usenet newsgroups. This is the case even if we adjust for the shorter time-period used to produce the email list plots of 4 as opposed to 5 months. Note that a number of lists have over 50% of users posting two months in a row, even when over 5000 messages were posted in a 4-month period. Further, unlike the Usenet plot with its -.43 ($p < .000$, $n=565$) Spearman's correlation, no significant Spearman's correlation was found for the data presented in Figure 7.

Discussion and Conclusions

This paper examines the nature of the relationship between the virtual spaces typically used for public online behavior, their technological platforms, and the behaviors such systems contain. A theoretical model and methodology labeled cyber-archaeology was described which utilizes large-scale field studies into user behavior in online spaces to identify technology-associated constraints to sustainable patterns of virtual public interaction dynamics. Applying the label cyber-archaeology was not meant to imply an historical, timeline-orientated perspective, which is just one of many approaches adopted by archaeologists. Rather, the label was given because the approach is based on archaeological theory and focuses on differences between the duration of impacts of social action and material remains (in this case CMC).

Empirical research was undertaken to assess the validity of both the theoretical model and methodology. Overall, the empirical findings support the assertion that individual 'information overload' coping strategies have an observable impact on mass-interaction discourse dynamics. Evidence was also found for the hypothesized link between virtual public technology type and constraints to sustainable patterns of large-scale virtual public interaction dynamics.

While our results are in line with the theoretical model, the findings raise a number of issues. First, our failure to find a decrease in proportional email list poster stability as interactive user populations increased, raises questions as to the validity of the model, which should be addressed. Two possibilities are: 1) Our sample did not contain email lists with a large enough active user population to enable the observation of the hypothesized effect; and 2) That for push technologies such as email lists, to some degree user subscription is akin to posting activity. Whatever the case, it is clear that email lists are typically able to support a much higher level of proportional active poster stability than Usenet newsgroups.

The second and perhaps most important objection that can be raised with regards to this research is the validity of the interpretation of our empirical findings in terms of the collective impact of individual information overload on virtual public interaction dynamics. While collectively the results are supportive of our overload model, obviously it is possible to propose a number of alternative explanations for each of the findings. For example, higher poster turnover in larger groups could relate to the impact of size on social network node structure rather than individual information overload. Although such possibilities are more than reasonable, what is more important is the utility of the method and theory outlined to:

(1) Compare virtual publics and virtual public types;

(2) Make novel predictions that can be examined empirically (good scientific puzzle solving); and

(3) Suggest a further research as outlined by Jones (2002; Chapter 6), Liu (1999), and Ekeblad (1999).

Therefore, while the validity of explanatory model is debatable, it is clear that the significance of the results and method is of greater importance.

The emergence of mass interaction has presented new opportunities to learn about and understand human communication, and information technologies. The availability and persistence of such communications, and the scale at which it operates allows us to explore various system effects on group discourse. To date empirical research into the *systemic* nature of the patterning of social relationships in cyberspace has been relatively rare. Research based on a systems approach to examine internet group communication, such as: the modeling of free riding using the Napster like Gnutella network (Adar and Huberman 2000); modeling the inter-relationship between homepages (Adamic and Adar 2000); exploring the self-organizing nature of email lists (Ekeblad 1999); and showing the World Wide Web to be structured like a small world network (Adamic 1999); have all been undertaken in the last five years. The work described in this paper is the first to explore empirically the impact of systems effects in Usenet discourse.

The recognition of the systemic nature of virtual public discourse allows for the examination of CMC technologies in terms of group-level usability. Currently, it is widely accepted that “reliable measures of overall usability can only be obtained by assessing the effectiveness, efficiency and satisfaction with which representative users carry out representative tasks in representative environments” (Bevan and Macleod 1994). This view supports the use of usability laboratories, and ethnographic methods, which can put user behavior in context. While, not discounting the value of these approaches, using the methodology presented in this paper for comparative purposes represents an alternative approach. This is because it potentially allows us to see and compare the normal range of user interaction dynamics for different types of CMC-technologies.

Not only do the techniques outlined offer insight into aspects of CMC-tool usability, but they also provide insight into technology design. For example, research is currently being undertaken to utilize an understanding of the discourse dynamics of real-time chat channels on IRC in order to build a smart real-time chat channel recommender system (see Terveen and Hill 2001 for a review of recommender systems). The smart recommender system aims to take into account conventional issues such as content as well as notions of group critical mass and overload. Other researchers are also examining how the approach outlined in this paper can be used to inform moderators of online discussion boards.

This research can progress in a number of different directions through:

(1) The large-scale comparative analysis of virtual-public discourse dynamics for a variety of other technologies (IRC, Web based bulletin board systems, Etc.).

(2) A thorough empirical examination of the impact of the various ways in individuals can respond to overloaded virtual public discourse.

(3) The formal proposing of and empirical research into alternative explanations for the underlying causes of consistent patterns of virtual public mass interaction.

(4) Various methodological and theoretical refinements that would result from an examination of related issues such as time scaling, and longitudinal impacts (e.g. Schoberth et. al. 2003); and

(5) The utilization of the knowledge gained into group-level usability to design better CMC-technologies.

The findings of the empirical research suggest that the cyber-archaeology approach is of value. It made verifiable predictions as to the nature of online behavior, specifically that the impact of individual cognitive processing limits are observable in the interaction dynamics of online spaces, and that these impacts differ between technologies. Overall, our theoretical model, methodology, and empirical results suggest a new way of understanding how classes of CMC-technology impact on the discourse they support.

References

Adamic, L. A. 1999. The Small World Web. *Proceedings of the 3rd European Conf on Digital Libraries*. Lecture notes in Computer Science, 443-452. New York: Springer.

Adamic, L., and E. Adar 2000. Friends and neighbors on the Web. PARC Xerox Manuscript, 1501 Page Mill Rd. MS 1U-19, Palo Alto, CA 94304, available online at: http://www.hpl.hp.com/shl/people/eytan/fandn.html

Adar, E., and B. Huberman 2000. Free Riding on Gnutella. *First Monday* 10(5), available online at: http://www/firstmonday.dk

Bevan, N., and M. Macleod, 1999. Usability assessment and measurement. In: *The Management and Measurement of Software Quality,* M. Kelly, ed. Ashgate Technical/Gower Press.

Bush, V. (1945) “As We May Think,” The Atlantic Monthly, July 1945.

Butler, B., 2001. Membership size, communication activity, and sustainability: A resource-based model of online social structures. *Inform Systems Research*. 13(4).

Cherny, L., 1999. *Conversation and Community: Chat in a Virtual World,* Center for the Study of Language and Information, Stanford.

Coate, John., 1992. Innkeeping in Cyberspace, In: *Directions and Implications of Advanced Computing (DIAC-92)*, Computer Professionals for Social Responsibility, Palo Alto, CA., Berkeley, CA, . http://gopher.well.sf.ca.us:70/0/Community/innkeeping.

Doheny-Farina, Stephen, 1996. *The wired neighborhood,* Yale University Press, London.

Dunbar, R., 1996. *Grooming, gossip and the evolution of language,* Harvard University Press, Cambridge, Mas.

Ekeblad, E. 1999. The emergence and decay of multilogue: Self regulation of a scholarly mailnglist. *European Association for Research on Learning and Instruction (EARLI),* Sweden.

Fletcher, R., 1995. *The limits of settlement growth: A theoretical outline,* Cambridge University Press.

Forrester, J. (1969) *Urban Dynamics*. The M.I.T. Press. Cambridge, Mass.

Gopher, D. and E. Donchin, 1986. Handbook of perception and human performance. In: *Cognitive processes and performance,* Vol. 2 (Eds. K. R.B, L. K. and J. P.T.), John Wiley & Sons, New York, pp. 1-49.

Gunther, R., L. Shapiro, P. Wagner. 1996. Zipf's law and the effect of ranking on probability distributions. *International J of Theoretical Physics* 35(2) 395-417.

Herring, S. C. 1999. Interactional coherence in CMC. *Proceedings of the 32nd Hawaii International Conference on System Sciences,* IEEE, Hawaii.

Hiltz, S. R., and M. Turoff 1985. Structuring computer-mediated communication systems to avoid information overload. *Communications of the ACM* 28.

Hiltz, S.R. and M. Turoff, 1978. *The network nation: Human communication via computer,* Addison-Wesley Publishing Company, Inc, London.

Jones Q. 1997. Virtual-communities, virtual-settlements & cyber-archaeology: A theoretical outline. *J of Comp Mediated Communication* 3(3).

Jones Q., and S. Rafaeli 2000a. Time to Split, Virtually: ‘Discourse Architecture’ and ‘Community Building’ as means to Creating Vibrant Virtual Publics. *Electronic Markets: The International Journal of Electronic Commerce and Business Media*. 10(4) 214-223.

Jones Q., and S., Rafaeli. 2000b. What do virtual 'Tells' tell? Placing cybersociety research into a hierarchy of social explanation. *33rd Hawaii International Conference on System Sciences, (Hawaii 2000), Hawaii*, IEEE Press.

Jones Q. 2001. The boundaries of virtual communities: From virtual settlements to the discourse dynamics of virtual publics. PhD Thesis, Graduate School of Business, University of Haifa, Israel.

Jones Q., Ravid G., and Rafaeli S. (2002). “An Empirical Exploration of Mass Interaction System Dynamics: Individual Information Overload and Usenet Discourse.” In: *Proceedings of the 35rd Annual Hawaii International Conference on System Sciences*, IEEE, Big Island, Hawaii.

Jones, S., 1995. Cybersociety: Computer-mediated communication and community. In: *Understanding Community in the Information Age,* Sage, Thousand Oaks, CA, pp. 10-35.

Kollock, P. and M. Smith, 1994. Managing the virtual commons: Cooperation and conflict in computer communities. In: *Computer-Mediated Communication,* (Ed. S. Herring), John Benjamins, Amsterdam.

Lewis, D., and K. Knowles 1997. Threading electronic mail: A preliminary study. *Inform Processing and Management* 33(2) 209-217.

Liu, G.Z., 1999. Virtual community presence in Internet relay chatting, *Journal of Computer-Mediated Communication [online]*, 5 (1). http://www.ascusc.org/jcmc/vol5/issue1/liu.html.

Milgram, S. (1969). Experience in Living in Cities. *Science*, 167: 1461-8.

Reid, E. M., 1991. *Electropolis: Communications and community on Internet Relay Chat,* Honors, History, University of Melbourne. http://www.ee.mu.oz.au/papers/emr/work.html.

Rheingold, H. 1993. *The virtual community: Homesteading on the electronic frontier*. Addison-Wesley, Reading, MA.

Rogers, E. M., & Agarwala-Rogers, R. (1975). Organizational communication. In G. J. Hanneman & W. J. McEwen (Eds.), *Communication and behaviour* (pp. 218-236). Reading, MA: Addison Wesley.

Rojo, Alejandra and Ronald G. Ragsdale, 1997b. Participation in electronic forums: Implications for the design, *Telematics and Informatics,* 14 (1): 83-96.

Roseman, M. and S. Greenberg, 1996. Teamrooms: Network places for collaboration, In: *Computer Supported Collabrative Work,* ACM Inc, Cambridge MA pp. 325-333.

Schmitz, J. and J. Fulk, 1991. Organizational colleagues, media richness, and electronic mail: A test of the social influence model, *Communication Research,* 18: 487-523.

Schoberth, T., Preece J., Armin H., (2003) Online Communities Longitudinal Analysis of Communication Activities Thomas. The Proceedings of the *36th Hawaii International Conference on System Sciences, (HICSS), Hawaii,* IEEE Press.

Shenk, D., 1997. *Data smog - Surviving the information glut,* HarperCollins, New York.

Simpson, R., Renear A., Mylonas E., & van Dam A. (1996) 50 years after "As we may think": the Brown/MIT Vannevar Bush symposium*.* *Interactions,* ACM Press, March 1996 Volume 3 Issue 2.

Smith, M. "Invisible Crowds in Cyberspace: Measuring and Mapping the Social Structure of USENET" in Communities in Cyberspace, edited by Marc Smith and Peter Kollock. London, Routledge Press, 1999.

Smith, M. and A. Fiore. "Visualization Components for Persistent Conversations," *Proceedings of ACM Computer-Human Interaction 2001.*

Spears, R. and M. Lea, 1992. Social influence and the influence of the 'social' in computer-mediated communication. In: *Contexts of computer-mediated communication,* (Ed. M. Lea), Harvester Wheatsheaf, New York, pp. 30-65.

Sproull, L., and S. Faraj 1997. Atheism, sex and databases: The Net as a social technology. Culture of the Internet. S. Kiesler, ed. Lawrence Erlbaum Assoc, Inc., Mahwah, NJ,

Steinfield, C. and Fulk. J., On the role of theory in research on information technologies in organizations. *Communication Research*. 14, 5, 1987, 479-490.

Terveen, L.G and Hill, W. Human-Computer Collaboration in Recommender Systems, in Carroll, J. (ed.), *HCI in the New Millennium* (2001), Addison Wesley.

Wellman, B., 2001. Computer networks as social networks, www.sciencemag.org 293.

Whittaker, S., 1996. Talking to strangers: An evaluation of the factors affecting electronic collaboration, In: *CSCW '96,* ACM, Cambridge, MA pp. 409-418.

Whittaker, S., and C. Sidner, 1996. Email overload: exploring personal information management of email., In: *CHI'96 Conference on Computer Human Interaction*, ACM Press, NY pp. 276-283.

Whittaker, S., Jones, Q., and Terveen, L. (2002). Managing Long Term Conversations: Conversation and Contact Management. In: *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, IEEE, Big Island, Hawaii.

Whittaker, S., L. Terveen, W. Hill, L. Cherny. (1998). The dynamics of mass interaction. *CSCW 98,* ACM Press, Seattle,

Zipf, G. K. 1949. *Human behaviour and the principle of least effort*. Cambridge, MA.