THE DYNAMICS OF CYBERSPACE:
EXAMINING AND MODELLING ONLINE SOCIAL STRUCTURE

Brian S. Butler

Carnegie Mellon University

April 21, 1999

This dissertation is submitted to the
Graduate School of Industrial Administration at Carnegie Mellon University
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Committee Members:

Dr. Kathleen Carley (Co-chair)
Dr. Robert Kraut (Co-chair)
Dr. Sandra Slaughter
Dr. Richard Moreland (Independent Reader)

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

CARNEGIE MELLON UNIVERSITY

GRADUATE SCHOOL OF INDUSTRIAL ADMINISTRATION

DISSERTATION

Submitted in partial fulfillment of the requirements

for the degree of DOCTOR OF PHILOSOPHY

INDUSTRIAL ADMINISTRATION (INFORMATION SYSTEMS)

Title THE DYNAMICS OF CYBERSPACE: EXAMINING AND MODELLING ONLINE SOCIAL STRUCTURE

Presented by BRIAN S. BUTLER

Accepted by Kathleen Carley 4-23-99

4-23-99

Dr. Kathleen Carley (Co-Chair)
Dr. Robert Kraut (Co-Chair)
Date

Approved by the Dean

28-April-1999

Douglas M. Dunn Dean Date

THE DYNAMICS OF CYBERSPACE: EXAMINING AND MODELLING ONLINE SOCIAL STRUCTURE

Contents

Paper One Appendices

Paper Three Appendices

ACKNOWLEDGEMENTS

Praise be to the name of God for ever and ever, wisdom and power are his.
He changes times and seasons;
he sets up kings and deposes them.
He gives wisdom to the wise and knowledge to the discerning
He reveals deep and hidden things;
he knows what lies in darkness, and light dwells within him.
I thank and praise you, O God of my fathers:
You have given me wisdom and power,
you have made known to me what we asked of you...

Daniel 2:20-23a

I also thank....

The members of my committee (Robert Kraut, Kathleen Carley, Sandra Slaughter and Richard Moreland) and my parents (Jim and Barbara Butler) for pushing me to do the work that I am capable of,

Shyam Sunder, for encouraging me to ask 'different' questions,

Jackie Cavendish and the other staff members at GSIA for being encouraging and helpful as I navigated the paperwork of graduate school, teaching, and research,

Michelle for listening to me talk, teaching me how to write, reading and proofreading papers, and generally being supportive even when I probably didn't deserve it,

Paul 'Big Paul' Mathew for being a good friend for me and my whole family,

and

'Little' Paul for repeatedly telling his daddy to work on his thesis and generally being a really good kid.

This work was funded by NSF grant #CSS9711548.

THE DYNAMICS OF CYBERSPACE: EXAMINING AND MODELLING ONLINE SOCIAL STRUCTURE

ABSTRACT

It has been proposed that online social structures represent new forms of organizing which are fundamentally different from traditional social structures. However, while there is a growing body of empirical research that considers behavioral aspects of online activity, research on online social structure structural remains largely anecdotal. This work consists of three papers that combine previous studies of traditional social structures, empirical analysis of longitudinal data from a sample of Internet listservs, and computational modeling to examine the dynamics of social structure development in networked environments.

The first paper (Title: When is a Group not a Group: An Empirical Examination of Metaphors for Online Social Structure) empirically examines the appropriateness of metaphors which have been used in popular and academic discussions of online social structure. The structural features implied by the metaphors are compared with data from a random sample of e-mail based Internet listservs. The results indicate that the most commonly applied metaphor ('small group') does not accurately represent the membership and communication features observed in these online social structures. Furthermore, there is evidence that the characterization of online structures in these terms has significantly biased the selection of cases and stories in the current literature. The empirical results also suggest that the metaphor of 'voluntary associations' is more accurate and hence is better foundation for theorizing about online social structure.

In the second paper (Title: Membership Size, Communication Activity, and Sustainability: The Internal Dynamics of Networked Social Structures) presents a resource-based theory of social structures. This model implies that structural features, such as size and communication activity, play both positive and negative roles in the sustainability of a social structure. Prior work has argued that networked communication technologies will significantly reduce the negative impact of size and communication activity, resulting in fundamentally different social structures. However, analysis of the longitudinal data from the e-mail based Internet listservs indicates that size and communication activity continue to have both positive and negative effects. This suggests that while the use of networked communication technologies may alter the form of communication, balancing the positive and negative impacts of membership size and communication activity remains a fundamental problem underlying the development of sustainable social structures.

The third paper (Title: Communication Cost, Attitude Change and Membership Maintenance: A Model of Technology and Social Structure Development) integrates processes of individual belief change and member movement in a dynamic model of online social structure development. Contributed messages create a composite signal, providing members with information about the benefits of membership. This information changes members' beliefs about the structure and affects their willingness to remain members. The processes of communication, individual belief change, and membership maintenance form a cycle that underlies the development of the collective. Communication costs, a feature of the communication infrastructure, affect a social structure's development by moderating the process of member belief change. A dynamic, multi-agent computational model of social structure development was implemented, calibrated, and validated using the listserv data. Analysis of the model implies that reduced communication costs, as are expected in networked environments, slow down the development process, resulting in online social structures which have more (and more diverse) members while being less stable than traditional face-to-face associations.

When is a Group not a Group:
An Empirical Examination of Metaphors for Online Social Structure

Brian S. Butler

April 21, 1999

When is a Group not a Group:
An Empirical Examination of Metaphors for Online Social Structure

Abstract

As extensive computer-mediated communication infrastructures have emerged, both within organizations and in the public sphere, researcher and practitioner interest in networked social structures has increased. One of most common online social structures is the asynchronous electronic collective, in which text-based computer mediated communication systems enable members to broadcast messages to a targeted audience. There are many metaphors that have been applied to these structures, including community, group, forum, and conference. While on the surface these metaphors may seem to be interchangeable, each metaphor is associated with a different set of assumptions about the features and processes of these collectives. Although they have implicitly been the basis for much discussion of these structures, there has been little empirical research that has explicitly compared the various metaphors for online social structure.

A review of fieid studies of asynchronous voluntary electronic collectives is presented to characterize the metaphors that have been used to describe these social structures. The representations implied by these metaphors are then compared with data from a random sample of e-mail based Internet listservs. In addition, because of the role these metaphors play in discussions comparing traditional and online social structures, pure online collectives and hybrids that combine networked and traditional communication infrastructures are compared. The results indicate that although it is common in studies of computer-mediated communication, the metaphor of 'small groups' does not accurately represent the membership and communication features observed in online social collectives. Furthermore, there is evidence that the characterization of these structures as small groups has biased the existing set of empirical studies. The empirical results suggest that voluntary associations are a more appropriate metaphor, providing a more accurate description and hence better foundation for theorizing about social structures in networked environments.

Networked environments are an increasingly common part of everyday life. Many business, education, and government organizations have invested heavily in the creation of internal communication infrastructures. Similarly, one of the fastest growing segments of the telecommunication industry revolves around the developing public data network known as the Internet. Various systems have been developed within these infrastructures to support social activity. Technologies such as electronic mail and the World Wide Web (WWW) support social activity by allowing members to send messages. Video conferencing and text-based conferencing systems enable individuals who are geographically distant to interact. Whether at work or at home it is more and more likely that people are part of a networked communication system.

Since the early 1980's, when the earliest computer mediated communication systems were created, researchers have been intrigued by the potential of networked technologies to support, and perhaps change, the way people interact. From this interest has developed an extensive body of research focused on how individuals behave in on-line social environments. In an effort to guide the design of new technologies, much of this work has addressed questions about how social behavior in networked environments might differ from that observed in more traditional face-to-face contexts (e.g. McGuire, Kiesler, and Siegel, 1987; DeSanctis and Gallupe, 1987). In most cases studies have considered how individual behavior in traditional and on-line social contexts compare, and on that basis attempted to infer how the traditional and on-line social structures will differ (Sproull and Kiesler, 1990). However, while the studies of individual behavior in on-line social settings have developed a solid empirical foundation, discussions about the impact of these new technologies on social structure (e.g. Daft and Lewin, 1993) remain based primarily on anecdotes, conjecture, and limited case studies.

Although there are an increasing number of studies that focus on describing examples of on-line social structure, overall this literature provides a weak foundation for theorizing about on-line social structure because it focuses on demonstrating that certain behaviors are possible in networked environments. As computer-mediated technologies developed, the theoretical position that text-based communication media were inherently unsuited for supporting complex social interaction was advanced (Daft and Lengel, 1986). This theory, known as media richness theory, led to the early conclusion that text-based networked environments would be unable to support may types of communication activity. A major thrust of computer-mediated communication research has been to examine the claims of this theory. Whether implictly or explicitly, past studies of on-line social structures have generally focused on refuting media richness models by documenting the capability of networked environments to support a wide variety of social behaviors (Table 1).[1]

	Technology	**Member Population**	**Duration**	**Number of Groups**	**Primary Method**
Baym, 1993	USENET	Soap opera fans	1 month	1	Participant Observation
Bikson and Eveland (1990)	E-mail	Corporate employees and retirees	1 year		Survey, Archival
Collins and Berge (1997)	E-mail Lists (Internet)	Varied	-	8	Survey
Faraj and Sproull (1994)	USENET	Varied	-		Archival
Finholt and Sproull (1990)	E-mail Lists (Organizational)	Varied	6 weeks	5	Archival
Freeman (1984)	Specialized	Social Network Researchers	-	1	Sociometric Survey
Garramone, Harris, and Anderson (1986)	BBS	Political Constituents	-	1	Survey
Garramone, Harris, and Pizante (1986)	BBS	Political Constituents	-	1	Survey
Ha (1995)	E-mail Lists (Internet)	Marketing Professionals and Academics	-	4	Survey

[1]

Hagel and Armstrong(1997)	Internet	Consumers	-	Multiple	Anecdotal
Hiltz (1985)	Specialized	Academic Researchers	1 year +	6	Archival and Surveys
Hof, Browder, Elstrom (1997)	Internet	Varied	-	Multiple	Anecdotal
Korenman and Wyatt (1996)	E-mail Lists (Internet)	Women's Studies Academics	1 year +	1	Archival and Survey
Lally (1995)	USENET	MBA Students	-	122	Survey
Meyers (1987)	BBS	Unspecified	2 months	1	Survey, Archival, and Interviews
Ogan (1993)	E-mail Lists (Internet)	Turkish Nationals	1 month	1	Archival
Rafaeli and LaRose (1993)	BBS	Varied	-	126	Survey
Rafaeli (1986)	BBS	Students	6 weeks	1	Survey, Archival
Rheingold (1993)	BBS and Internet	Varied	-	Multiple	Anecdotal
Rice and Love (1987)	CompuServ	Medical Professionals and Students	6 weeks	1	Archival
Rice (1982)	Specialized	Academic researchers	24 months	10	Archival
Roberts (1998)	USENET	Varied	-	30	Survey
Rojo (1995)	E-mail Lists (Internet)	Varied	1 year +	11	Archive and Survey
Smith (1997)	USENET	Varied	3 weeks	4000+	Archival
Sproull and Faraj (1997)	USENET	Varied	?	< 10	Archival
Sproull and Kiesler (1990)	E-Mail	Business Organization Members	-	Multiple	Archival
Sudweeks (1995)	E-mail	Communication Researchers	2 years+	1	Archival, Survey and Interview
Whittaker (1996)	Lotus Notes (Organizational)	Varied	90 days+	20	Archival and Interviews
Zenhousem and Wong (1997)	E-mail Lists (Internet)	Varied	Varied	10	Archival

Table 1: Example Studies of On-line Social Structures

However, while prior field studies have addressed questions about the types of behavior that *can* occur in networked social environments, they have had less to say about what *does* happen. Researchers have typically chosen online sites for study based on personal interest in the content (e.g. Baym, 1995; Ha, 1995) or because the structures were expected to exhibit the social phenomena of interest (Finholt and Sproull, 1990). While these studies are useful existence proofs for online behavior, it is unlikely that they provide a realistic description of 'normal' operation of online social structures. Similarly, anecdotal accounts are likely to be biased, with casual observers noticing and reporting interesting events, and providing little or no information about the features of mundane (or failed) structures. Thus while the studies in this area provide glimpses networked social structures, they are, at best, a questionable foundation for theorizing about the development and operation of on-line social strutures.

This study adds to this body of research, developing its basis for generalization by providing a systematic characterization of a random sample of one type of on-line social structure, e-mail based Internet listservs. In addition, we also contribute to the study of networked social environments by empirically comparing features of a set of pure online social structures with those of hybrid structures that combine networked and traditional infrastructures. Hypotheses about differences in size, membership change, communication volume, interactivity, and participation distribution are proposed and tested in order to assess the consequences of different communication infrastructures for the nature of social structures.

Another influence on on-line social research has been the application of the small group as a dominant metaphor for characterizing on-line social structures. The metaphors used to characterize social structures are important because each one embodies a set of assumptions about the features, processes, and impacts of the structures. Each metaphor partially describes a social structure, and different models focus attention on different aspects of that structure. For instance, 'community' implies a sense of identity that 'conference' does not. Thinking about 'discussion forums' suggests that there is extensive interaction among the participants, while 'mass media' is likely to have distinct producers and audience members. These are just few examples of how the selection of a metaphor leads to assumptions about the nature and operation of networked social structures.

Labeling a new phenomenon such as online social structures with a familiar name is useful because it allows researchers to effectively communicate and generalize their findings, by presenting a focused result in the context of a larger framework. Many metaphors have been used to characterize the social structures that have arisen in networked environments (Table 2).

Community, Virtual Community		Baym (1993) Rheingold (1993) Roberts (1998) Hiltz (1985) Hagel and Armstrong (1997) Hof, Browder, Elstrom (1997)
Social Group		Faraj and Sproull (1994) Finholt and Sproull (1990) Hiltz (1985) Korenman and Wyatt (1996) Sudweeks (1995) Zenhousern and Wong (1997) Sproull and Kiesler (1990)
Social Network		Rice (1982) Wellman (1997)
Discussion Forum, Discussion Group		Berge (1994, 1995) Collins and Berge (1997) Rojo (1995) Ha (1995)
Conference		Freeman (1984) Hiltz (1985)
Shared Information Space, Information Source		Whittaker (1996) Lally (1995)
Public Good, Virtual Commons		Rafaeli and LaRose (1993) Kollock and Smith (1996) Kollock (1997)
Mass Media, Communication Media		Rafaeli (1986) Garramone, Harris, and Anderson (1986) Garramone, Harris, and Pizante (1986) Ogan (1993) Rafaeli and LaRose (1993)

Table 2: Metaphors for On-line Social Structures

However, it is also important to critically examine whether the characterization implied by a metaphor is appropriate. Roberts (1998) and Baym (1993) find evidence of community-like elements in their studies of USENET groups, supporting the anecdotal reports of Rheingold (1993) and other popular authors. Finholt and Sproull (1990) and Sproull and Keisler (1990) report behaviors that are similar to those found in small groups. However, the accuracy of the metaphors that underlie discussions of online social structures remains largely unconsidered. Rarely are the prototype structures implied by metaphors compared with one another or with empirical descriptions of online social structures. Consequently, it is often unclear whether the assumptions embedded in discussions of online social structure are consistent with the features and operation of naturally occurring networked social structures.

Many studies of online social behavior have adopted the model of the small groups, and as a result implicitly assumed that small groups provide an appropriate metaphor for online social structure. Conceptualizing social structure in terms of small groups provides a theoretical foundation that makes it logistically and methodologically easier to study the behavior of individuals in on-line social contexts. Small, task oriented groups communicating synchronously are easier to recreate in the controlled setting of a laboratory than other social structures which operate over longer time spans (weeks vs. hours), have less precisely defined goals, and sporadic participation.

However, while research based on the model of small groups has provided valuable insights into individual behavior in computer-mediated environments (e.g. McGuire, Kiesler, and Siegel, 1987; DeSanctis and Gallupe, 1987) it remains unclear whether is it the best foundation for describing the nature of online structures. As applied in most studies of online communication, the model of small groups assumes that while the perceptions, attitudes, and behaviors of individuals may change, the nature of a social structure remains essentially fixed. Small groups are assumed to be set up, operate, and then they end, typically within a short time span. Small groups are seen as the context for individual behavior, and not as entities which themselves exist in a larger context. If membership composition is considered at all, it is treated as a causal factor, not an emergent outcome to be explained. Questions are asked about the performance of a group – but not its existance. Likewise, the consquences and causes of membership movement, in the form of new member entry and member loss, received little attention (c.f. McGrath and Hollingshead, 1994: Chapter 5). When considering behaviors, attitudes, and perceptions of individuals in on-line environments it is likely that equating social structure and small groups is appropriate. However, as we move from questions about how individuals behave to questions about how the social structures operate it is necessary to reconsider whether the metaphor of small groups is appropriate, or whether some other model might serve as a better foundation for characterizing on-line social structure.

The analysis presented here examines two alternative metaphors for online social structure and asks which one provides a more appropriate foundation for studies of online social structure. The models, small groups and voluntary associations, were chosen as representative of two broad classes of metaphors used in the exploratory studies of online social structure. 'Small groups' are most commonly thought of as having fixed, limited membership (< 10 people), high levels of interaction, limited duration, and well defined goals or activities. In contrast, 'voluntary associations', which include social clubs, discussion forums, volunteer organizations, professional societies, conferences, and communities, are expected to have larger, more variable membership; highly uneven, and often non-interactive, participation; extended, if not unlimited duration; and informal, often ambiguously defined, objectives. The appropriateness of these metaphors is tested by comparing structural features seen in a sample of e-mail based Internet listservs, such as membership size and variability, communication volume and structure, and the distribution of participation, with those expected in a prototypical small group and voluntary association.

Sample Selection and Data Collection Methods

The on-line social structures considered in this work are unmanaged, e-mail based Internet listservs. These on-line social collectives[2] utilize Internet-based e-mail and a centrally maintained mailing list to enable individuals to broadcast text-messages to other members. E-mail based collectives were chosen as representative of online social structure because they are known to be prevalent in both private (Finholt and Sproull, 1990) and public network infrastructures such as the Internet.

Although there may be an individual who is responsible to maintaining the mailing list (i.e. the listowner), the selected social structures are unmanaged. Listowners take no formal steps to restrict membership or message content. These collectives are expected to be representative of social structures that operate in environments when there is little active intervention. Hence, the results can be seen as providing a baseline against which the impact of management strategies, such as moderation and member screening, might be evaluated.

Sample Selection

The public networked environment of the Internet includes e-mail collectives with various topical emphases, attracting members from a wide range of communities and organizations. From the population of approximately 70,000 collectives, an initial sample of 1066 was created. The initial sample was stratified by topic type to ensure that it spanned a reasonable range of topics and member communities. One third of the sampled collectives focused on work-related topics. One third focused on personal topics (hobbies, lifestyles, etc.). The remaining third involved topics that mixed work-related and personal interests (e.g. geographic locations).

The initial sample was subjected to a multiple stage confirmation process (see Appendix A for more details). Actively managed collectives, including moderated listservs and those with formal new member screening, were eliminated. This selection process also verified that each listserv was mechanically functional, able to provide the needed data, and available for inclusion in the study (See Table 3 for a summary of the reasons for elimination from the sample). The result was a set of 284 listservs, which fell to 204 as collectives were eliminated during data collection[3].

	Number Eliminated
Listowner chose not to participate	227
Inoperable server or group	120
Inaccessible membership data	143
Exclusive membership	86
Course-related groups	73
Moderated groups	53
Broadcast groups	51
Non-English groups	22
Sensitive topic/groups	21
Non-standard message/membership formats	13
Gateways and non-e-mail lists	9
Unable to contact the listowner	8
Incomplete addresses	6
Duplicates	4
No description available	2

Table 3: Reasons for Elimination of a Listserv from Initial Sample

To verify that the final sample spanned the intended range of topics and populations, the degree to which each listserv's focused on work-related, personal, or academic concerns was

[2] The term collective is used to refer to the online social structures. The terms 'group' and 'association' are used to refer to the prototypes implies by the small group and voluntary association metaphors.

assessed. These measures were constructed by asking coders to read a short description of each listserv and indicate on a scale from 1 (low) to 5 (high) the likelihood that a substantial portion of each collective's membership participates for work-related, personal, and academic reasons (three measures for each listserv). Inter-rater reliability was found to be acceptable, with Cronbach alphas of 0.79, 0.88, and 0.78 respectively, and although the sample was not evenly distributed among the three categories, the final sample includes a wide range of topics and membership communities.

Within the final sample listservs were classified as either pure or hybrid collectives. Pure online collectives operate completely in the networked environment. In contrast, hybrid collectives use computer-mediated communication technology to supplement traditional communication activities, such as meetings or print communications (Finholt and Sproull, 1990). Multiple judges were used to assess this feature of each collective. Based on short descriptions, coders assessed, on a 1 (low) to 5 (high) scale, the likelihood that each online collective also used traditional, non-networked, communication activities. Inter-rater reliability was found to be acceptable, with a Cronbach alpha of 0.82. The evaluations were averaged to create a single assessment of each collective's infrastructure. The listservs were then classified either as hybrid or pure based on whether their assessment was above or below the median value for the sample.

Data Collection

For a 130 day period, between July 23, 1997 and November 30, 1997, data on communication activity and membership was collected for each listserv. The communication data consisted of all e-mail messages distributed to members. To collect these messages a

[3] The collected data for listservs eliminated during data collection was archived; however it is not included in the analyses presented here.

project account was created and added to each listserv's membership list. This account then received a copy of all messages. The sender identification field was encoded and the messages archived[4]. The data collection process resulted in an archive of all communication activity that occurred within the selected collectives during the observation period.

Once a day during the data collection period, a message was automatically sent requesting the listervs' membership lists. As the lists were received, individual contact information was encoded and the data stored. This process generated a record of the membership changes that occurred in the sampled collectives during the observation period. The message and membership archives are the basis of the various measures of collective structure and activity used in the following analyses. Each section will describe the relevant measures and how they were constructed from this raw data.

Membership Size

Membership size, as indicated by the number of people who are exposed to a collective's communication activity, is one of the most prominent ways that the metaphors of small groups and voluntary associations differ. Groups are thought of as relatively small social structures, with membership of between 2 and 7 individuals (Forsyth, 1990). Studies of both casual and formal groups have found that group size is distributed according to a j-shaped distribution (e.g. truncated exponential or Poisson distribution), with median values of 2 or 3 (Bakeman and Beck, 1974; Burgess, 1984; Coleman and James, 1961; Desportes and Lemaine, 1988; Dunbar, 1993; James, 1953; Tucker and Friedman, 1972). In contrast, studies of community associations (McPherson, 1983a [Mean: 188, Median: 40]) and youth gangs (Thrasher, 1927 [Mean 31,

[4] Identifying information in both the message and membership data was encrypted in order to address concerns

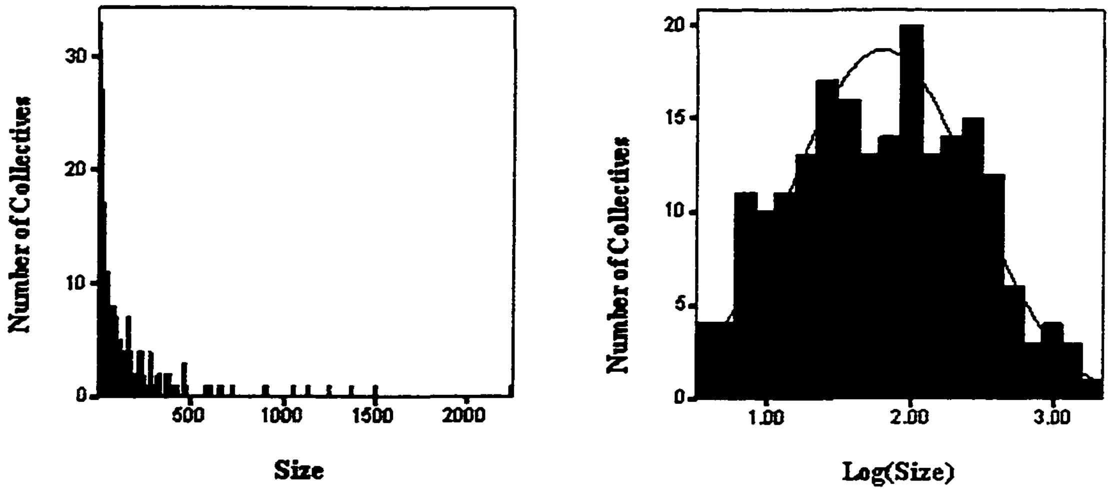
Median 16]) found that voluntary associations are larger and that their sizes are log-normally distributed[5].

Collective size is measured by counting the number of members on each listserv's e-mail distribution list on the first day of the observation period. This characterizes size in terms of the number of people who are exposed to the collective's communication activity at that time. While this measure may increase the observed size by counting individuals who receive message but do not read them, it is conceptually equivalent to counting the number of people who attend traditional meetings, a common measure of size in studies of social structure in non-networked environments.

The distribution of listserv sizes is well characterized by a log-normal distribution (Figure 1)

about illicit data use.

[5] The distribution of voluntary association sizes is also similar to the distribution of business firm sizes (Quandt, 1966; Simon, 1957; Simon and Bonini, 1958; Collins, 1973).



(a) (b)

Mean:	163	Maximum:	2245
Median:	64		
Std Dev.:	279	Minimum:	3

Figure 1: Membership Size Distribution

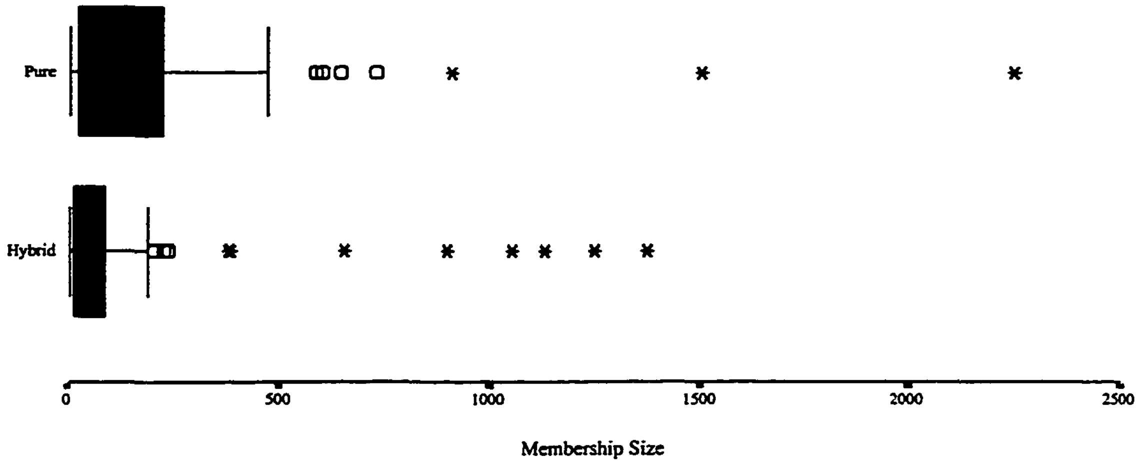
Comparison of the observed distribution of membership sizes with the metaphor prototypes indicates that the online collectives are more similar to voluntary associations than small groups. The listservs are significantly larger than the 2 to 7 range that is expected of groups. However, the mean size of 163 and median size of 64 is comparable to the sizes seen for voluntary associations. Also, as seen for voluntary associations (McPherson, 1983a; Thrasher, 1927) the distribution of membership size among online collectives is log-normal (Figure 1b).

In some discussions of online social activity it has been conjectured that asynchronous computer-mediated communication infrastructures are capable of supporting larger structures than traditional social infrastructures (e.g. Rheingold 1993, Finholt and Sproull, 1990). While this is clearly true when comparing the online collectives (with a mean size of 163 members) with traditional small groups (with sizes in the range of 2 to 7), it is less apparent when comparing them with traditional voluntary associations (with sizes in the 100's and 1000's).

Nonetheless, the hypothesis that online collective are larger than traditional voluntary associations is supported by a comparison of the median online collective size (64) and the median size observed in McPherson's (1983a) analysis of voluntary associations within a several Midwest communities[6] (40).

Differences between traditional and networked social environments are also expected to lead to size differences between pure and hybrid social structures. Pure online social collectives exist entirely within the networked environment. As a result, it is argued, they are less affected by the logistical problems that inhibit growth in traditional social environments (Rheingold, 1993; Finholt and Spoull, 1990). In contrast, hybrid collectives combine networked and traditional communication structures, and are more likely to be subject to the costs and constraints faced by traditional groups and associations (c.f. Hare, Blumberg, Davies, and Kent, 1994; p.147; McPherson,1983a). Thus, if networked social structures are expected to be larger than those operating in traditional environments, then pure online collectives should be larger than hybrid collectives.

[6] The medians were compared instead of means because both sets of data are highly skewed and non-normal.



N:	131	Maximum:	2245
Mean:	174	Median:	91
Std Dev.:	271	Minimum:	3

Pure Online Collectives

N:	73	Maximum:	1373
Mean:	140	Median:	34
Std Dev.:	295	Minimum:	4

Hybrid Collectives

Figure 2: Membership Size in Pure and Hybrid Social Collectives

Pure online collectives tend to be larger than hybrid structures (Figure 2). The difference in membership size is significant in the predicted direction (Wilcoxon test: $p < 0.001$), implying that pure network structures will be larger than hybrid collectives. These results also support claims that networked environments will support larger structures than tradition social environments.

Membership Change

The metaphors of small groups and voluntary associations also differ with respect to membership change. Small groups are seen as having fixed, or at least highly stable, membership. A group is characterized in terms of its members. If the membership significantly changes, it is perceived to be a different group. In contrast, voluntary associations routinely experience high levels of member movement. During the lifetime of an association, many people come and go (c.f. McPherson, 1983a). As a result, both the size and composition of a voluntary association can change significantly over time.

Three measures are used to characterize membership change: percentage growth in membership, percentage loss of members, and percentage change in membership. Percentage growth is the number of new members[7] who arrived during the observation period, relative to a collective's initial size. Percentage loss is the number of members who left the group over the same time period, normalized by the collective's initial size. Percentage change combines the measures of member growth and loss to describe the net change in size during the observation period.

Membership change is the norm in the sampled online collectives (Figure 3). More than 75% of listservs had new members during the observation period. Over 50% had members leave.

	% Growth		% Loss		% Change	
	Observed	*Annual*[8]	*Observed*	*Annual*	*Observed*	*Annual*
Mean	22.5%	63.2%	13.6%	38.2%	8.9%	25.0%
Std Dev	23.8%	66.8%	17.4%	48.9%	12.2%	34.3%
Maximum	117.0%	328.5%	98.0%	275.2%	94.0%	263.9%
Q3	34.8%	97.7%	20.5%	57.6%	14.4%	40.4%
Median	14.8%	41.6%	8.1%	22.7%	3.4%	9.5%
Q1	3.7%	10.4%	0.0%	0.0%	0.0%	0.0%

[7] This measure also includes individuals who rejoin after an absence. However, since returning individuals are relatively rare (less than 10% of all recorded 'new' members) no special treatment is given to these individuals.
[8] Annual rates were determined by extrapolating the observed change rates to a 365 day year (365 / 130 * observed value).

Minimum	0.0%	0.0%	0.0%	0.0%	-41.1%	-115.4%
Collectives with 0%	33		53		41	

N = 204

Figure 3: Membership Change Distributions

The online social structures were characterized by significant membership flows, which when operating together resulted in a generally positive change in membership size.

Just as it affects size, the composition of a collective's infrastructure is also expected to affect the rate of membership change. Hybrid collectives, because of they are linked with traditional social infrastructures, should be able to recruit members more effectively. References to the listservs in face-to-face meetings, conferences, or print publications, all raise awareness of the online social activity among a targeted population of individuals who are likely to be interested. In contrast, pure online collectives typically must rely on interpersonal word of mouth or untargeted advertising through the WWW. Thus, hybrid collectives are expected to have higher rate of membership growth than pure online social collectives.

The connection with traditional communication activity also may affect the rate at which members leave hybrid collectives. Traditional social structures require that members make greater investments of time, energy, and attention, than in pure online collectives. The higher costs make it more likely that the individuals will leave traditional or hybrid collectives than pure networked social structures. Thus because of the costs incurred, pure online collectives are expected to have lower membership loss rates than hybrid collectives.

The expected differences between the membership change processes in pure and hybrid collectives were not observed in the listserv data (Table 2).

		Pure	Hybrid	Difference
		(N = 131)	(N = 73)	
% Growth				
	Mean	21.9%	23.2%	-1.3%
	Median	14.8%	16.7%	-1.9%
% Loss				
	Mean	13.5%	13.7%	-0.2%
	Median	8.1%	8.1%	0.0%
% Change				
	Mean	8.4%	10%	-1.6%
	Median	4%	2%	2.0%

Table 2: Membership Change in Pure and Hybrid Social Collectives

Although the growth rates differ in the predicted direction (i.e. the hybrid collective's growth rate is higher) the difference is not statistically significant (Wilcoxon, p > 0.1). There is also no significant difference between the membership loss rates for the two collective types.

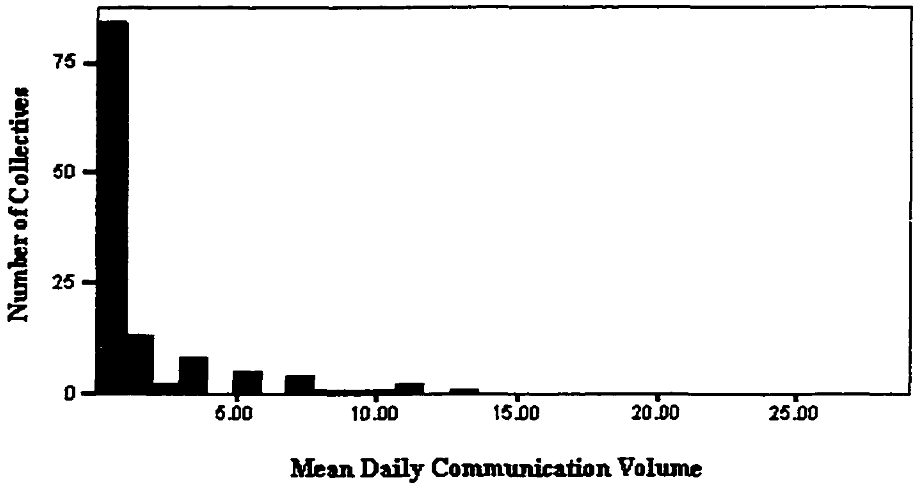
Communication Activity

Communication among members underlies coordination, social support, information sharing, and other social process, such as identity or norm formation, which are essential to the operation of any social structure. Yet the amount and structure of communication implied by the metaphors of small groups and voluntary associations differ. Small groups are seen engaging in limited sessions involving high levels of interactive communication. Some theorists have defined small groups as collections of individuals[9] who influence one another through interaction (for review see Forsyth, 1990: pp. 6-8), highlighting the importance of communication in these social structures. The expectation with small groups is that they involve members in a limited session with high levels of communication activity. In contrast, voluntary associations, with their long lifespans, are expected to involve a lower volume of communication activity, often

[9] Forsyth (1990) is an example of researcher who focus on groups as social structures. There is also a body of research that conceptualizes groups as psychological constructs. These "minimal group" studies are based on the idea that a group is defined by members (and selected non-members) perceptions - irrespective of social activity.

making use of structured meetings, informal gatherings, and print media to maintain communication among the members.

Communication activity volume in online collectives is measured in terms of the average number of messages per day. In a listserv, each message represents a member taking a 'turn' in a conversation. Among the sampled collectives there is significant variation in the communication volume. However, it does not appear that the norm is high levels of activity. One third of the listservs had no communication activity during the observation period (Figure 4).



Mean:	1.635	Maximum:	29.121
Std Dev.:	3.543	Median:	0.277
		Minimum:	0.007

Figure 4: Distribution of Mean Daily Communication Activity

Among the online social collectives that were active during the observation period, the mean daily communication volume is concentrated at the low end, with a median that is the equivalent of one message every 3.6 days[10].

[10] The distribution of mean daily communication activity among the active online social collectives has the following features:

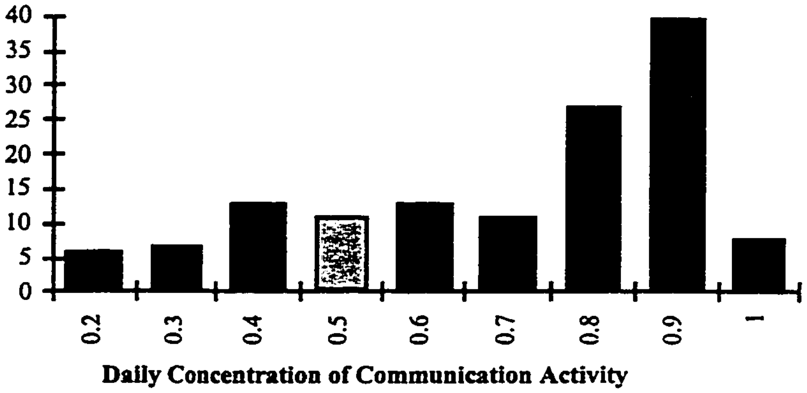
N:	136	Maximum:	29.121

Another difference between the small group and voluntary association metaphors is the expectations regarding the distribution of communication activity. Small group sessions are assumed to be communication oriented, to the point that a collection of people who came together but did not talk to one another would probably not be considered a group (Forsyth, 1990). This assumption leads to the characterization of small groups as having high levels of ongoing communications activity. In contrast, voluntary associations are characterized as having relatively uneven communication flows. For example, the amount of communication activity in a professional organization might be low with 'bursts' of activity occurring around intermittent meetings, conferences, and print publications.

To characterize the distribution of communication activity in the online social collectives, a Gini coefficient was calculated with each day as a category. The Gini coefficient is a value between 0 and 1 (inclusive) which describes the concentration of items in a set of categories. A low value indicates that items (e.g. messages) are spread evenly across the categories (e.g. days). A high value indicates that they are highly concentrated, with a few of the categories (e.g. days) accounting for a large number of the items (e.g. messages). This provides an overall measure of the degree to which communication activity seen during the observation period is evenly (or unevenly) distributed.

In online social collectives, communication activity is not evenly distributed over the observation period (Figure 5).

Mean:	1.635	Median:	0.277
Std Dev.:	3.543	Minimum:	0.007



N:	136	Maximum:	1
Mean:	0.751	Median:	0.833
Std Dev.:	0.227	Minimum:	0.205

Figure 5: Communication Activity Concentration

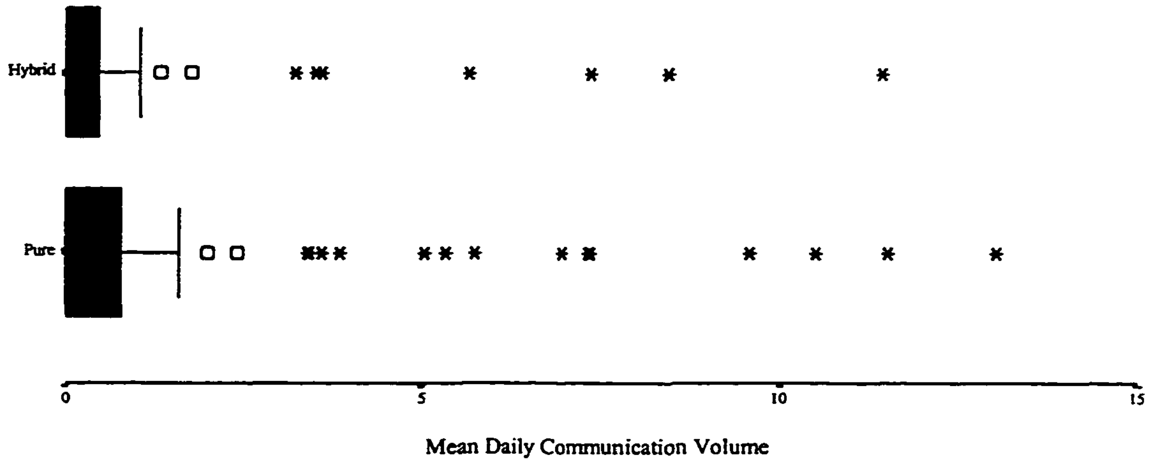
The mean value of 0.751 and the left-skewed distribution among the sampled collectives indicates that communication activity in these social structures tends to occur in a few bursts rather than evenly over time. Overall, the online social collectives are best characterized as having low volumes of highly concentrated communication activity.

The expected effect of combining networked and traditional modes of communication on a collective's communication activity is unclear. Some work suggests that communication activity will be greater in hybrid contexts and other results imply that online social activity will greater in pure networked environments. E-mail is often used in organizational settings to coordinate activities and share information in support of other off-line communication activities (meetings, project collaborations, etc.). The presence of a relationship supported by face-to-face communication is expected to increase the ability of individuals to use text-based communication media. A significant relationship has been found between who people interact with in traditional settings and who they communicate with via E-mail (Rice, Grant, Schmitz, and Torobin, 1990).

For this reason hybrid infrastructures might be expected to see higher levels of communication activity than pure collectives.

On the other hand, hybrid structures operate in a context that provides members with alternative means for interacting as a collective (Finholt and Sproull, 1990). Members of hybrid collectives have multiple communication media to choose from, while the participants in pure online collectives have little choice but to use the networked communication tools. To the degree that communication media are substitutes, the availability of traditional communication opportunities may reduce use of the online communication. In contrast to the above argument, this characterization of online communication implies that hybrid social collectives will have lower volumes of online social activity than pure networked collectives.

Although graphically there is some evidence that activity in pure online collectives may be greater than in hybrid collectives (Figure 6), the difference is not statistically significant (Wilcoxon Test: $p > 0.1$).



N:	131	Maximum[11]:	29.131	N:	73	Maximum:	11.439
Mean:	1.251	Median:	0.078	Mean:	0.799	Median:	0.077
Std Dev.:	3.414	Minimum:	0.000	Std Dev.:	2.009	Minimum:	0.000

Pure Online Collectives *Hybrid Collectives*

Figure 6: Communication Volume in Pure and Hybrid Social Collectives

Pure and hybrid collectives also differ in terms of the proportion of structures that saw no activity during the observation period (Table 3).

	No Activity	Activity	
Hybrid	29 (40%)	44 (60%)	73
Pure	39 (30%)	92 (70%)	131
	68	136	204

Table 3: Proportion of Online Collectives with No Communication Activity

However, a Fisher exact probability test (p = 0.165 > 0.1) indicates that there is no significant relationship between the collective type (hybrid vs. pure) and the proportion of collectives that see no activity. These results suggest that the presence or absence of traditional infrastructure elements does not significantly affect the volume of communication activity in networked social structures.

Group Communication Structure

Small groups and voluntary associations also differ in terms of the structure of communication activity. Small groups are generally seen as being interactive, with members taking turns in an ongoing stream of interrelated conversation (Bonito, 1997; Hollingshead and Bonito, 1998). In these contexts, individual members hear and respond directly to the comments of others. In contrast, communication activity within voluntary associations is expected to be more episodic. Although there are still likely to be themes and topics that are common throughout the stream of communication, because of logistical and temporal constraints, there are significantly fewer explicit responses.

Message activity in online collectives has the potential to have an interactive structure. Discussion threads, formed by a set of messages that share a common subject line, are a common communication structure that is considered to be indicative of interaction in asynchronous online environments (Sproull and Faraj, 1997). The proportion of a collective's messages which receive no reply (i.e. solo messages) and the average number of messages within a discussion thread, (including solo messages as threads of length 1) are thus two values which provide an indication of the level of interaction (Sproull and Faraj, 1997) . These measures characterize the 'public' interactivity or the explicit structural interaction present within the group communication. They do not capture interaction that takes place through traditional communication media, personal e-mail outside a collective's communication infrastructure, or members' perceptions of interactivity (Koreman and Wyatt, 1996). However, as Finholt and Sproull (1990) note, communication features such as these are important to consider because

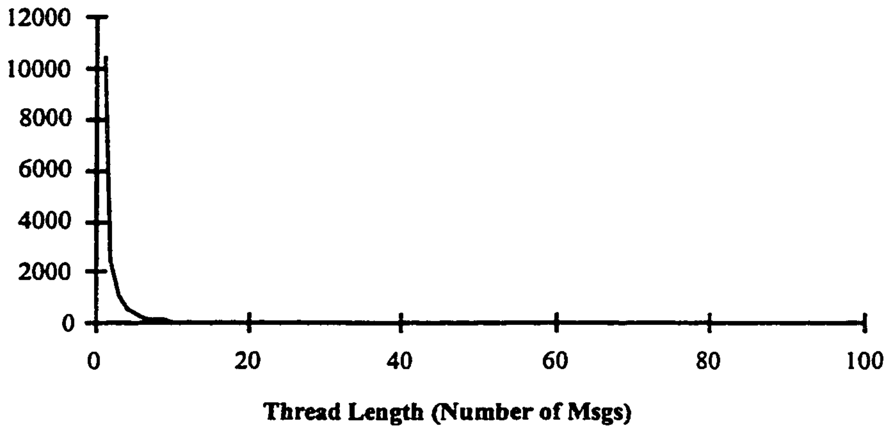
[11] The data point with this maximum value (29.131) was excluded from the figure to allow for more effective comparison of the category distributions.

they are highly visible, and hence are likely to play a significant role in individuals' perception and behaviors.

Solo messages are identified by first categorizing each message as either a new message or a reply to an earlier message. This categorization is performed based on the contents of the message subject line. Subject lines that begin with 're:' are classified as replies[12]. All other messages are categorized as new messages. Discussion threads are identified by removing the 're:' and matching the first 40 characters of each reply subject line with the subjects of previously distributed new messages. Thread length was determined by counting the number of messages within each identified thread. Solo messages are discussion threads that have a length of one. The proportion of solo messages was computed by dividing the number of single message threads by the total number of messages distributed within the collective during the observation period. This value provides a measure of the interactivity of a collective's communication, with lower proportions of solo messages indicating higher interactivity. Average thread length also serves as a measure of interactivity. An online collective with shorter threads sees relatively less public interaction while longer threads indicate that the communication activity regularly includes explicit interaction.

Overall, 32% (10439 ÷ 32373) of the recorded messages are solo messages. This suggests that, within the sampled collectives, extended public interaction is somewhat unusual. The distribution of thread length also supports this characterization with at least 75% of the observed threads involving only 1 or 2 messages (Figure 7).

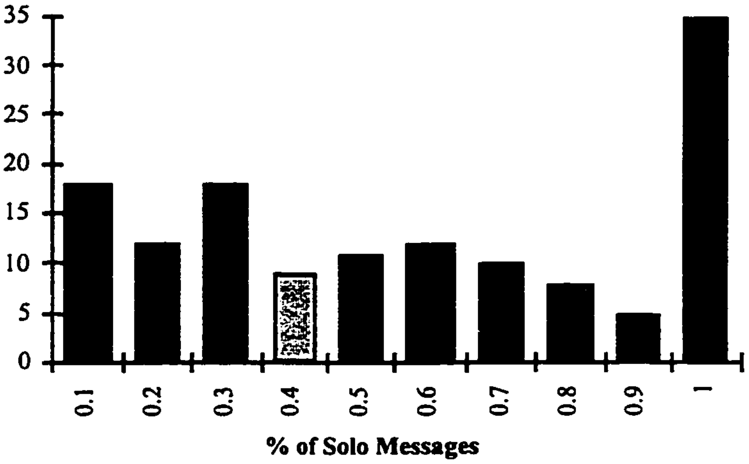
[12] To test the reliability of this classification rule a sample of 500 messages were classified manually. The error rate of the automatic classification rule ('re:' in the subject line) was less than 1%.



N:	15,905	Maximum:	95
Mean:	2.036	Median:	1
Std Dev.:	2.838	Minimum:	1

Figure 7: Distribution of Discussion Thread Lengths

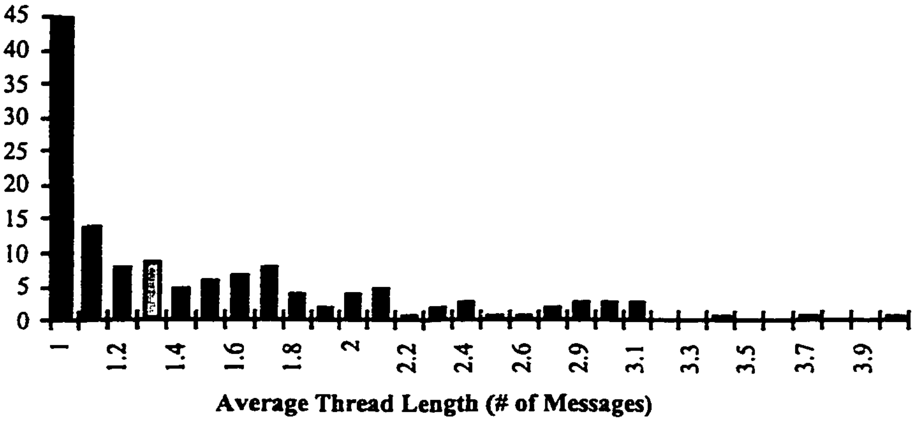
The general lack of explicit interaction is also reflected in the measures of collective interactivity. In over half of the sampled collectives, a majority (> 50%) of the communication activity was solo messages (Figure 8).



N:	136	Maximum:	100%
Mean:	60.8%	Median:	60%
Std Dev.:	31.2%	Minimum:	10.2%

Figure 8: Distribution of the Proportion of Solo Messages

Also, thread lengths within the online collective was short, with most (more than 75%) of the listservs having mean thread lengths of less than two messages (Figure 9).



N:	136	Maximum:	4.047
Mean:	1.580	Median:	1.333
Std Dev.:	0.686	Minimum:	1

Figure 9: Distribution of Mean Discussion Thread Length

The sampled online collectives are not highly interactive, at least in terms of the structure of the group communication. A significant portion of all messages are solo messages (32%) and a majority of the structures have more than half of their communication in the form of solo messages. Overall thread lengths are short (1 or two messages) and most online collectives can be characterized as tending to have short public discussions (<= 2 messages).

In hybrid groups the availability of alternative communication opportunities is likely to affect the structure of communication activity. In many cases, complex interaction can be accomplished more efficiently and effectively in a face-to-face setting. Thus, the availability of face-to-face interaction should reduce the average complexity of online communication, resulting in shorter messages and less explicit interaction in the hybrid collectives. In addition, common activities, experiences, culture, and shared physical spaces all provide mechanisms for more efficient communication. Shared knowledge of a physical space, for example, allows things to be referenced, and hence discussed, more succinctly. The combined effect of shared context and the availability of alternative communication opportunities suggests that social activity in hybrid collectives should be more compact and less interactive than in pure networked collectives.

A significant difference in the mean message length in the two sub-populations supports the prediction that activity in hybrid collectives will be more compact than in pure online collectives (Wilcoxon test: $p = < 0.0001$) (Table 4).

	Pure	Hybrid
Message Length	403	312
(Number of Words)	(N=23,906)	(N=8,467)
Interactivity		
Proportion of Solo Messages	60%	61%
	(N = 94)	(N = 44)
Average Thread Length	1.61	1.50
	(N = 94)	(N = 44)

Table 4: Communication Structure in Hybrid and Pure Online Collectives

However, while there is a small difference in the predicted direction for the average thread lengths in hybrid and pure collectives, it is not statistically significant (1.50 vs. 1.61: Wilcoxon test: $p > 0.1$). There is also no difference between the hybrid and pure collectives in terms of the proportion of communication activity accounted for by solo messages (61% vs. 60%). Thus while the composition of a collective's infrastructure may affect features of individual messages, there is no evidence that it significantly alters the overall structure of the communication stream with respect to public interactivity.

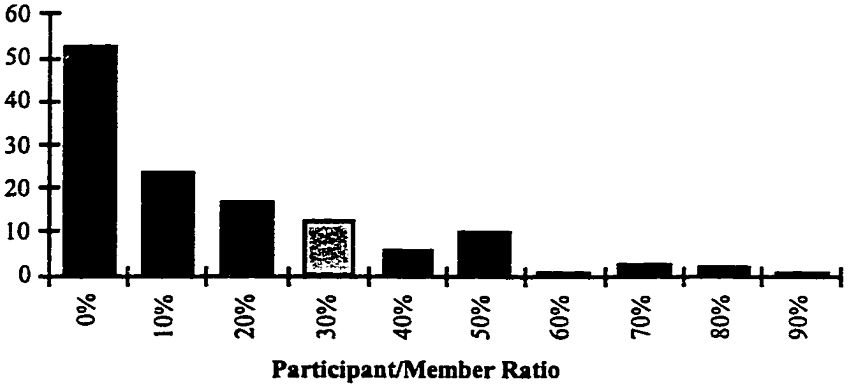
Participation Patterns

Both small groups and voluntary associations are known to exhibit uneven participation distributions. In both small groups and voluntary associations, it is common for a small percentage of membership to account for a majority of the communication activity (Bales, Strodtbeck, Mills, and Rosenborough, 1951; Warner and Hilander, 1964; Skvoretz, 1988). However, small groups and voluntary associations differ in the degree to which active participation is unequal. Within small group sessions it is not uncommon for the top one or two active participants to account for 50-75% of the communication activity (Bales, Strodtbeck, Mills, and Rosenborough, 1951; Skvoretz, 1988), while the least active members contribute relatively little. However, while there is clearly an unequal distribution of activity, it is generally not the case that a substantial portion of the membership is silent. In contrast, within voluntary associations large segments of the membership may be passive participants, not contributing at all to the communication activity (Warner and Hilander, 1964; Warner, 1965). In these structures, there is typically a pronounced dichotomy between the active core members and the silent periphery (Lyon, 1974).

The participation structure of the online collectives has been characterized in terms of three measures: participation ratio, a Gini coefficient for the distribution of participation among the active participants, and the proportion of activity accounted for by the two most active members. The participation ratio, or proportion of members who contribute at least one message (i.e. active members), captures the degree to which the membership as a whole is actively involved in a collective's on-line social activity. The concentration of communication among the active participants was described with a Gini coefficient calculated for each online collective. The Gini coefficient is a value between 0 and 1 (inclusive) that indicates the degree to which messages are concentrated among the active participants. A low value indicates that communication is equally distributed among the active participants and a high value indicates that it is concentrated, with a few individuals contributing most of the messages[13]. The distribution of communication among the participants was also measured by determining the proportion of communication activity that is accounted for by the two most active participants. This measure was constructed because it was expected that the features of the high end of the participation distribution have been highlighted in prior research on groups and associations, but the Gini coefficient is known to be insensitive to differences at the extremes (Smolensky, 1994). For comparison these measures were also calculated for a set of 30 traditional small groups working on a discussion task (data was originally reported in Skovretz, 1988).

The participation structure of online social collectives tends to be highly concentrated among a relatively small number of members. The participation ratios are skewed to the right with most of the collectives having fewer than 20% of the membership actively participating (Figure 10).

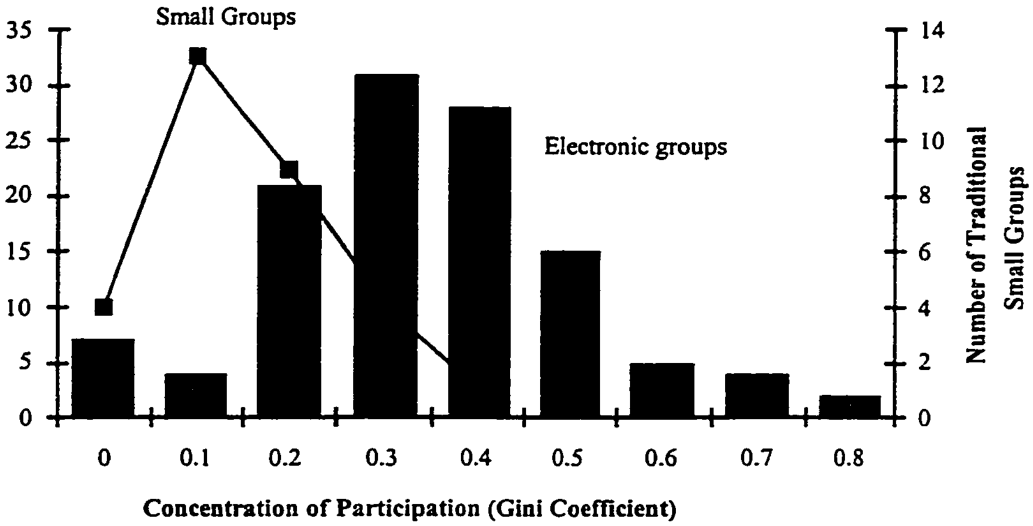
[13] The range of possible Gini coefficient values is not [0,1]. Use of only active members (i.e. whose who sent at least one message) ensures that the extreme situation (i.e. all messages being sent by one person with the other included individuals sending no messages), and hence the extreme value of 1, cannot occur.



N:	130	Maximum:	95.7%
Mean:	21.9%	Median:	15.3%
Std Dev.:	21.2	Minimum:	0.1%

Figure 10: Participation Ratio Distribution[14]

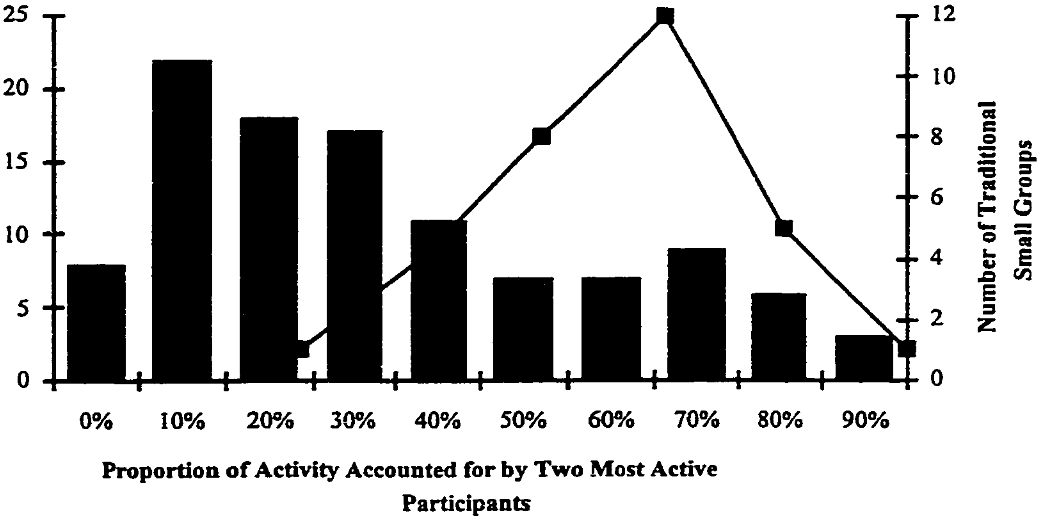
Likewise, among the active participants the activity was not evenly distributed (Figure 11)



N:	117	Maximum:	0.882
Mean:	0.385	Median:	0.378
Std Dev.:	0.172	Minimum:	0.0

(a) Gini coefficient

[14] Six cases with participation ratios greater than 100% are not included in this figure. This occurred in several small collectives that experienced significant membership growth during the observation period.



N:	117	Maximum:	100%
Mean:	63.8%	Median:	62.0%
Std Dev.:	22.6%	Minimum:	< 0.1%

(b) Proportion of the activity accounted for by the top two participants

Figure 11: Distribution of Participation Concentration

In comparison to the Skvoretz (1988) small group data, the online social collectives have more concentrated participation patterns. In contrast to the 30 traditional small groups studied by Skvoretz, in which the participation ratios were all higher than 80%, the online collectives had a majority of members who did not actively participate. Furthermore, among the active participants, the small groups had a participation pattern that was more equally distributed than the pattern seen online (Figure 11a). However, while in the online collectives participation was overall more concentrated, at the extreme the difference was reversed with the top participants accounting for a lower proportion of the activity in the online collectives than in the small groups (Figure 10b).

In prior work it has been argued that one of the valuable features of networked environments is their potential to equalize participation (Dubrovsky, Kiesler, and Sethna, 1991).

By eliminating the logistical problem of blocking (Gallupe, Dennis, Cooper, Valacich, Bastianutti, and Nunamaker, 1992) and reducing the social cues (Sproull and Kiesler, 1986), these technologies are expected to reduce the factors which can lead to participation inequality (Skvoretz, 1988). However, subsequent research has shown that participation differentials can persist due to status differences (Saunders, Robey, and Vaverek, 1994; Weisband, Schnieder, Connolly, 1995) and differences in individuals' expectations regarding participation (Rojo, 1995). As a result, the distribution of participation in online social collectives can remain concentrated among a small subset of the members.

Hybrid collectives are more likely to be subject to known status differences among the members, an important aspect of the process by which status is linked to participation (Weisband, Schnieder, and Connolly, 1995). Therefore, it is expected that participation in hybrid collectives will tend to be more concentrated than in pure online collectives.

The differences observed in these results suggest that while hybrid groups may be more concentrated among the active members, a greater proportion of their members actively participate in online social activity (Table 5).

		Pure Online Collectives	*Hybrid Collectives*
Participation Ratio		(N= 94)[15]	(N = 44)
	Mean	28%	52%
	Median	15%	19%
Concentration among active members (Gini coefficient)		(N= 72)	(N= 36)
	Mean	0.3815	0.3920
	Median	0.3765	0.3957
Proportion of Messages Sent by Two Most Active Participants		(N = 72)	(N = 36)
	Mean	36%	42%
	Median	31%	37%

Table 5: Participation Concentration in Hybrid and Pure Online Collectives

However, while these results suggest that there are differences between the participation structures in hybrid and pure online collectives, they are not statistically significant (Wilcoxon test: $p > 0.1$ in all three cases).

Discussion

These results imply that listservs are best described as large, dynamic social structures in which a core of active participants generates relatively low levels of sporadic communication (Table 6).

[15] The number of listservs in each condition varies due to differences in the number of participants. For example, listservs with no message activity are excluded. Likewise, participation concentration measures (e.g. the Gini coefficients) cannot be calculated for listservs with only one participant.

	Small Groups	Voluntary Associations	Online Collectives
Membership			
Size	3-10	30+	44
Growth and Loss	Little or none	Constant	Significant
Communication Activity			
Volume	High	Low	Low
Distribution	Constant	Sporadic/Bursty	Sporadic/Bursty
Structure	Interactive/Conversational	Episodic	Episodic
Participation Structure			
Overall	Full membership	Dichotomous – Active Core and Passive Periphery	Dichotomous – Active Core and Passive Periphery
Distribution among Active Participants	Concentrated	Concentrated	Concentrated

Table 6: Comparison of Small Groups, Voluntary Associations, and Online Collectives

In terms of membership size and change, communication volume and structure, and participation Internet listservs are more like voluntary associations than small groups. These findings highlight a bias in prior studies of online social activity. While the goals of verifying the existence of recognizable social behaviors in networked environments have been well served by focusing on highly active, interactive examples of online social structures, at least for e-mail based Internet structures, these cases do not seem to be representative. For example, interactivity is a common theme in many descriptions of on-line social activity (Rheingold, 1993; Baym, 1995; Hof, Browder, and Elstrom, 1997). Cases often highlight different types of interaction that can occur in e-mail based social contexts (Sproull and Kiesler, 1990; Finholt and Sproull, 1990). However, the results presented here imply that while interactivity can occur in these contexts, its is more the exception than the norm. Another feature that is common to most of the structures described in prior work is a reasonably high volume of communication activity. In some cases this is acknowledged as an explicit selection criteria (Finholt and Sproull, 1990) while in others it is a result of researchers desire to work with clearly visible social phenomena (e.g. Rheingold, 1993; Baym, 1993; Ogan 1993). However, while this strategy is effective for documenting the types of social processes that *can* occur in networked environments, the results presented above imply that prior work many have unintentionally presented a biased description of the social activity that is *likely* to occur in a networked social environment.

This bias is significant because of the effect that it has on discussions, both academic and popular, about online social structures. For example, contrary to discussion of the problems of developing electronic social collectives which focus on minimizing the consequences of unwanted communication (e.g. Kollock and Smith, 1996; Kollock, 1997), these results imply that a major problem facing developers is prompting some appropriate level of communication. That is, while " free riding" behavior, in which individual contribute unwanted messages, may be a problem in some cases, it seems that a more common problem is collective silence. From a practical standpoint, this implies that rather than focusing on controlling contributions, developers should devote their attention to encouraging participation. More fundamentally, this suggests that to better ground the discussion of networked social environments additional work should be done to document the characteristics of online social activity and structures in a variety of contexts.

Clearly one limitation of this work is that it only considers one type of online social structure: listservs. While listservs can be considered representative of a large class of on-line social structures, including WWW conferencing systems, USENET, and other structures based on asynchronous communication infrastructure, it is possible that other on-line social structures might be more "group-like". Specifically, on-line social structure which make use of synchronous communication technology, such as Multi-User Dungeon (MUDs) or Chat rooms, might be expected to have features (size, communication activity, participation, etc.) more like small groups and less like voluntary associations. Also, while the comparison of pure and hybrid listservs revealed few differences, it is also possible that online social structures existing within the context of a single organization might also operate differently. Additional population studies would further the study of computer-mediated communication by providing a better understanding the types of social structures which arise within different communication infrastructures.

These results also call attention to the assumptions underlying discussions of "new" forms of organizing. Technology impact claims are made with respect to some baseline. Discussions of technology enabling "new forms" of social structure, implicitly make assumptions about what "old forms" of social structure looked like. If the metaphor of small groups is used then the baseline is likely to be a prototypical traditional small group. This leads to the conclusion that in most cases on-line social structures are indeed a new form of organizing. However, if the baseline of voluntary association then the validity of the novelty claim is less apparent. As the analysis of pure and hybrid online collectives indicated, there is evidence for some differences between social structures that make use of different communication infrastructures. However, overall, the listservs had membership, communication, and participation features that were generally similar than those expected from the prototypical voluntary association. Thus claims about the impact of communication technology on social structure may over state the novelty because they implicitly compare apples (online social collectives) and oranges (traditional small groups), rather than two types of apples (online social collectives and traditional voluntary organizations).

A radical interpretation of these findings would suggest that small groups should not be used as a foundational metaphor for the study of on-line social structure. The conceptual framework underlying small group research has embedded in it assumptions about size, membership stability, levels of communication activity, interactivity, and participation structures. However, while these assumptions may be adequate for characterizing the context for examining individual social behavior it cannot be assumed that they are appropriate when the structures is itself the object of study. Thus, while studies of online social activity based on the small group paradigm can provide valuable insight into individual behavior in online social context, applying that model in discussions about the operation of naturally-occurring networked social structures must be done critically, if at all.

In contrast, conceptualizing many on-line social structures as associations or organization may be more appropriate than seeing them as meetings or social gatherings (i.e. small group sessions). While there are technologies, such as Chat Rooms and MUDs, which enable synchronous communication session, the most common communication infrastructures (e-mail, USENET, and the WWW), are asynchronous like the infrastructure used by the listservs considered in this study. This suggests that the literature on voluntary associations and organizations in sociology (e.g. Warner and Hilander, 1964; Warner, 1965; McPherson,1983, McPherson and Smith-Lovin, 1988; McPherson, 1990; McPherson and Rotolo, 1996) and organization theory (Wilderom and Miner, 1991), is fruitful reference discipline for researchers interested in studying online social structure. Characterization of on-line social structures as associations or organizations also raises questions for future research. The tendency towards low levels of explicit public interaction leads to questions about whether online voluntary associations which do not have many structurally interactive discussions are nonetheless perceived by their members as interactive, and if so, why. Low levels of sporadic communication prompts questions about the processes that might lead to identification, norm formation, and norm maintenance in long duration, low activity social settings. The flow of people into and out of these structures highlights the dynamic nature of membership in online social structures and leads to questions about the mechanisms and impact of membership changes. Characterizing online social structures as voluntary associations encourages researchers to critically assess the assumptions that have been made in prior work, and from that assessment develop our understanding of "normal" online social structures.

However, while these results illustrate how a dominant metaphor can affect the study of a phenomenon such as on-line social structure, combining metaphors of small groups and voluntary associations is likely to be the most effective strategy for understanding the social environment in emerging communication infrastructures. On one hand, structures based on synchronous communication technology often seem to be 'group-like'. While this may be true within a single on-line session it is often the case that a changing set of individuals interacts over many sessions. Thus, while the operation of particular sessions might be best examined through the lens of the small group metaphor, the dynamics of repeated synchronous on-line social structures could be studied from the perspective of voluntary associations. Likewise, while the structural dynamics of asynchronous on-line social structures, such as listservs, are likely to have commonality with organizations and other macro-social structures, the behaviors of individuals within these structures can be considered from the within the small group framework. Furthermore, like traditional structures, which make use of both synchronous technologies (e.g. meetings) and asynchronous technologies (e.g. print), on-line social structures are not inherently limited to one type of communication. In addition to considering the nature of on-line structures in different infrastructures, work needs to be done to better model the structural consequences of hybrid infrastructures. Therefore, models that combine features of small group and association metaphors are likely to provide significant insight into the processes and structures that underlie the use and impact of emerging communications infrastructures.

References

Bakeman, R. and Beck S. (1974) The size of informal groups in public. Environment and Behaviour. Vol. 6, pp. 378-390.

Bales, R.F., Strodtbeck,F.L., Mills,T.M., and Rosenborough,M.E. (1951) Channels of communication in small groups. American Sociological Review, Vol. 16, pp. 461-468.

Baym, N. (1993) Interpreting soap operas and creating community: Inside a computer-mediate fan culture. Journal of Folklore Research, Vol. 30, pp. 143-176. [Also appears in S. Kiesler (Eds.), Culture of the Internet]

Berge, Z. L. (1994) Electronic discussion groups. Communication Education. Vo.. 43, No. 2, pp. 102-111.

Berge, Z.L. (1995) Computer-mediated scholarly discussion groups. Computers in Education, Vol. 24, No. 3, pp. 183-189.

Bikson, Tora and Eveland, J.D. (The Interplay of work group structures and computer support. In J. Galegher, R.E. Kraut, and C. Egido (Eds.) Intellectual Teamwork: Social and Technological Foundations of Cooperative Work (pp.245-290). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bonito, Joseph A. (1996) Topical Contributions to Group Discussions: Assessing the Contribution of Topic Knowledge to Participation. Presented at International Communication Association annual Meeting, Chicago.

Bonito, joseph A. and Hollingshead, Andrea B. (1998) Participation in Small Groups. Communication Yearbook, 20. pp. 227-261.

Burgess, J.W. (1984) Do humans show a "species-typical" group size? Age sex and environmental differences in the size and composition of naturally occurring casual groups. Ethology and Sociobiology, Vol. 5, pp. 51-57.

Coleman, James S., and James, John. (1961) The equilibrium size of freely-forming groups. Sociometry, Vol. 24, pp. 36-45.

Collins, L. (1973) Industrial Size Distributions and Stochastic Processes. Progress in Geography, Vol. 5, pp. 121-165.

Collins, Mauri P., and Berge, Zane L. (1997) Electronic Discussion Group Lists in Adult Learning. Unpublished Manuscript. Northern Arizona University.

Daft, Richard L. and Lengel, Robert H. (1986) Organizational Information Requirements, Media Richness and Structural Design. Management Science, Vol. 32, No. 5. pp. 554-571.

Daft, Richard L. and Lewin, Arie Y. (1993) Where are the theories for the "New" organizational forms? An editorial essay. Organization Science, Vol 4, No. 4. pp. i-vi.

DeSanctis, G. and Gallupe, R.B. (1987) A foundation for the study of group decision support systems Management Science, Vol. 33, No. 5, pp 589-609.

Desportes, J.P. and Lemaine, J. M. (1988) The size of human groups: An analysis of their distributions. In D. Canter, J.C. Jesuino, L.Soczka, and G.M. Stephenson (Eds.), Environmental social psychology, (pp. 57-65). Dordrecht, The Netherlands: Kluwer.

Dubrovsky, V. J. , Kiesler S., and Sethna, B. N. (1991) The equalization phenomena: Status effects in computer-mediated and face-to-face decision making groups. Human Computer Interaction, Vol. 6, pp. 119-146.

Dunbar, R.I.M. (1993) Coevolution of neocortical size, group size, and language in humans. Behavioral and Brain Sciences, Vol. 16, pp. 681-735.

Faraj, S., and Sproull, L. (1994) Interaction dynamics in electronic groups. Unpublished manuscript, Boston University, Boston.

Finholt, T., and Sproull, L. (1990) Electronic Groups at Work. Organization Science, 1, 41-64.

Forsyth, Danelson R. (1990) Group Dynamics (2nd. Ed.) Pacific Grove, CA: Brooks/Cole.

Freeman, Linton C. (1984) The Impact of Computer-Based Communication on the Social Structure of an Emerging Scientific Specialty. Social Networks, Vol. 6, pp. 201-221.

Garramone, Gina M., Harris, Allen C., Anderson, Ronald (1986) Uses of Political Computer Bulletin Boards. Journal of Broadcasting and Electronic Media, Vol. 30, No. 3, pp. 325-339.

Garramone, Gina M., Harris, Allen, and Pizante, Gary (1986) Predictors of Motivation to Use Computer-Mediated Political Communication Systems. Journal of Broadcasting and Electronic Media, Vol. 30, No. 4, pp. 445-457.

Gallupe, R.B., Dennis, A.R., Cooper, W.H., Valacich, J.S., Bastianutti, L.M., and Nunamaker, J.F. (1992) Electronic brainstorming and group size. Academy of Management Journal, Vol. 35, pp. 350-369.

Ha, Louisa (1995) Subscriber's Behavior in Electronic Discussion Groups: A Comparison Between Academics and Practitioners. COTIM-95: Proceedings of the Conference on Telecommunications and Information Markets, Newport, RI.

Hagel, John III, and Armstrong, Arthur G. (1997) net.gain: Expanding markets through virtual communities. Boston, MA: Harvard Business School Press.

Hare, A. Paul, Blumberg, Herbert H., Davies, Martin F., and Kent, M. Valerie (1994) Small Group Research: A Handbook. Norwood, NJ: Ablex.

Hiltz, Starr Roxanne (1985) Online communities: A case study of the office of the future. Norwood, NJ: Ablex Publishing.

Hof, Robert D., Browder, Seanna, and Elstrom, Peter (1997) Internet Communities, Business Week, May 5, 1997.

James, John. (1953) The distribution of free-forming small group size. American Sociological Review, Vol. 18, pp. 569-570.

Kollock, Peter and Smith, Marc (1996) Managing the Virtual Commons: Cooperation and Conflict in Computer Communities. In S.C. Herring (Eds.) Computer Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives (pp.226-242), Philadelphia: John Benjamins Publishing.

Kollock, Peter (1997) Design Principles for Online Communities. In Internet and Society: Harvard Conference Proceedings. Cambridge, MA: O'Reilly and Associates (CD-ROM).

Korenman, Joan, and Wyatt, Nancy (1996) Group Dynamics in an E-mail Forum. In S.C. Herring (Eds.) Computer Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives (pp.226-242), Philadelphia: John Benjamin Publishing.

Lally, Laura (1995) Exploring the adoption of bulletin board services. The Information Society, Vol. 11, pp. 145-155.

Lyon, Eleanor (1974) Work and Play: Resource Constraints in a Small Theater. Urban Life and Culture, Vol. 3, No. 1, pp. 71-97.

McGrath, J. E. (1984) Groups: Interaction and Performance. Englewood Cliffs, NJ: Prentice-Hall.

McGrath, Joseph E. and Hollingshead, Andrea B. (1994) Groups Interacting with Technology: Ideas, Evidence, Issues and an Agenda. Thousand Oaks, CA: Sage.

McGuire, T.W., Kiesler, S., and Siegel, J. (1987) Groups and computer-mediated discussion effects in risk decision making. Journal of Personality and Social Psychology, Vol. 52, pp. 917-930.

McPherson, Miller (1983a) The Size of Voluntary Organizations. Social Forces, Vol. 61, No. 4, pp. 1045-1064.

McPherson, Miller (1983b) An Ecology of Affiliation. American Sociological Review, Vol 48. pp. 519-532.

McPherson, Miller J. (1990) Evolution in Communities of Voluntary Organizations, In. J.V. Singh (Ed.). Organizational Evolution New Directions.(pp. 224-245). Newbury Park, CA: Sage.

McPherson, Miller J. and Smith-Lovin, Lynn (1988) The comparative ecology of voluntary associations in five nations. In G.R.Carrol (Ed.). Ecological Models of Organizations (pp. 85-109). Cambridge, MA: Ballinger.

McPherson, Miller J. and Rotolo, Thomas (1996) Testing a Dynamic Model of Social Comparison: Diversity and Change in Voluntary Groups. American Sociological Review, Vol. 61. pp. 179-202.

Meyers David (1987) "Anonymity is Part of the Magic": Individual Manipulation of Computer Mediated Communication Contexts. Qualitative Sociology, Vol. 10, No. 3, pp. 251-266.

Moreland, Richard, L. Hogg, Michael A., and Hains, Sarah C. (1994) Back to the Future: Social Psychological Research on Groups. Journal of Experimental Social Psychology, Vol. 30, pp. 527-555.

Ogan, Christine (1993) Listserver Communication during the Gulf War: What Kind of Medium is the Bulletin Board? Journal of Broadcasting and Electronic Media, Vol. 37, No. 2, pp. 177-196.

Quandt, J. (1966) On the size distribution of firms. American Economic Review, Vol. 36 (August), pp. 624-637.

Rafaeli, S. (1986) The electronic bulletin board: A computer driven mass medium, Computers and the Social Sciences, Vol. 2, No. 3, pp. 123-136.

Rafaeli,S. and LaRose,R.J. (1993) Electronic bulletin boards and 'public goods' explanations of collaborative mass media, Communications Research, Vol. 20, No. 2, pp. 177-197.

Rheingold, H. (1993) The virtual community. Reading, MA: Addison-Wesley.

Rice, R.E. (1982) Communication Networking in Computer-Conferencing Systems: A Longitudinal Study of Group Roles and System Structure. Communication Yearbook, Vol. 6, Beverly Hill, CA: Sage, pp. 925-944.

Rice, R.E., Grant, August E., Schmitz, Joseph, and Torobin, Jack (1990) Individual and Network Influences on the Adoption and Perceived Outcomes of Electronic Messaging. Social Networks, Vol. 12, p. 27-55.

Rice, Ronald E. and Love, Gail (1987) Electronic Emotion: Socioemotional Content in a Computer-Mediated Communication Network. Communication Research, Vol. 14, No. 1, pp. 85-108.

Roberts, Teresa L. (1998) Are newsgroups virtual communities? Proceedings of CHI'98, Los Angeles, CA.

Rojo, Alejandra (1995) Participation in scholarly electronic forums. Unpublished Phd. Disseration, University of Toronto, Ontario, Canada [http://www.oise.edu/~arojo/]

Saunders C, Robey, D., and Vaverek, K. (1994) The persistence of status differentials in computer conferencing. Human Communication Research, Vol. 20, pp. 443-472.

Simon, H. A. (1957) Models of man (chapter 9), New York: Wiley.

Simon, H.A. and B. Bonini (1958) On the size distribution of business firms. American Economic Review, pp. 607-617.

Skvoretz, John (1988) Models of Participation in Status-Differentiated Groups. Social Psychology Quarterly, Vol. 51, No. 1, pp. 43-57.

Smith, M.A. (1997) Measuring the Social Structure of the Usenet, Presented at Sunbelt XVI, San Diego, CA.

Smolensky, Eugene (1994) Lorenz Curve. In D. Greenwald (Ed.), The McGraw-Hill Encyclopedia of Economics (2nd Edition). (pp. 651-653) New York: McGraw-Hill.

Sproull, Lee., and Faraj, Samer (1997) Atheism, Sex, and Databases: The net as a social technology. In S. Kiesler (Ed.), Culture of the Internet. Mahwah, NJ: Lawrence Erlbaum.

Sproull, Lee, and Kiesler, S. (1986) Reducing Social Context Cues: Electronic Mail in Organizational Communication. Management Science, Vol. 32, No. 11, pp. 1492-1512.

Sproull, L., and Kiesler, S. (1990) Connections: New ways of working in the networked organization. Boston, MA: MIT Press.

Sudweeks, Fay (1995) Communication Patterns and Developmental Processes of Computer-Mediated and Collaborative Groups., Unpublished Manuscript. [Online: http://www.arch.usyd.edu.au/~fay/phd/proposal.html].

Thraser, Fredric M. (1927) The Gang. University of Chicago.

Tucker, J. and Friedman, S. T. (1972) Population density and group size. American Journal of Sociology, Vol. 77, pp. 742-749.

Warner, Keith W. (1965) Attendance and Division of Labor in Voluntary Associations. Rural Sociology, Vol. 29, 396-407.

Warner, W. Keith. and Hilander, James S. (1964) The Relationship between Size of Organization and Membership Participation. Rural Sociology, Vol 29. pp. 31-39.

Wilderom, Celeste P. M. and Miner, John B. (1991) Defining Voluntary Groups and Agencies within Organization Science, Vol. 2, No. 4. pp. 366-378.

Weisband, Suzanne P., Schneider, Sherry K., and Connolly, Terry (1995) Computer Mediated Communication and Social Information: Status Salience and Status Differences. Academy of Management Journal, Vol. 38, No. 4, pp. 1124-1151.

Wellman, Barry (1997) An electronic group is virtually a social network. . In S. Kiesler (Ed.), Culture of the Internet. Mahwah, NJ: Lawrence Erlbaum.

Whitaker, Steve (1996) Talking to strangers: an evaluation of factors affecting electronic collaboration., .Proceedings of CSCW'96, Boston, MA.

Zernhausern, Robert and Wong, Florence (1997) Virtual Personality of a List: A Preliminary Examination of the Demography of Interent Lists. In Sudsweeks, McLaughlin, and Rafaeli (Eds.) Network and Netplay: Virtual Groups on the Internet. Cambridge, MA: AAAI/MIT Press.

Membership Size, Communication Activity, and Sustainability: The Internal Dynamics of Networked Social Structures

Brian S. Butler

April 21, 1999

Membership Size, Communication Activity, and Sustainability: The Internal Dynamics of Networked Social Structures

Abstract

Effective communication infrastructures provide both the technical and social structures necessary to support social interaction. As networked communication infrastructures become more common, there is increasing interest in the factors underlying the development of online social structures. It has been proposed that networked social structures are new forms of organizing which are fundamentally different from traditional social structures, and thus, are not subject to the same constraints as traditional structures. However, from anecdotal evidence and case studies it is difficult to evaluate whether networked social structures are subject to the same problems as traditional social structures. Drawing from prior studies of traditional social structures and empirical analysis of longitudinal data from a sample of e-mail based Internet social structures, this work addresses the question of whether the role of size and communication activity in a structure's sustainability is fundamentally altered by the use of networked communication technologies.

A resource-based theory of sustainable social structures is presented. Members contribute time, energy, and other resources, enabling a social structure to provide benefits for individuals. These benefits, which include information, influence, and social support, are the basis for a social structure's ability to attract and retain members. This model implies that internal features, such as current size and communication activity, will play complex roles in the ongoing sustainability of a social structure. While traditional social structures often must rely on internal structure, such as editorial control or member screening, to manage the problems that arise due to increased size and communication, it has been proposed that networked communication technologies also significantly reduce these problems.

However, analysis of longitudinal data from a random sample of e-mail based Internet social structures (listservs) indicates that size and communication activity continue to have significant positive and negative effects. These results suggest that while the use of networked communication technologies may alter the form of communication used within a social structure, balancing the positive and negative impacts of membership size and communication activity remains a fundamental problem underlying the development of sustainable social structures. Since the adoption of networked technology does not eliminate these problems, both infrastructure developers and future research need to consider how internal structures could be applied to better support the emergence of long-term social structures.

Networked social environments are an increasingly common part of everyday life. Businesses are investing in Lotus Notes, Intranets, and other infrastructures with the goal of facilitating communication and learning. Governments are investing in the development of extensive information infrastructures in the hopes that they will promote the flow of information and the spread of knowledge. Many educational institutions, both public and private, are considering how the capabilities of new communication technologies can extend their reach. Early empirical and anecdotal evidence indicates that public data communications networks, such as the Internet, can create new opportunities for people to interact with one another (Kraut, Scherlis, Mukhopadhyay, Manning, and Kiesler, 1996; Baym, 1993, Rheingold, 1993). Whether public or private, networked environments are increasingly the site of social activity. USENET groups, e-mail distribution lists, Multi-User Dungeons (MUDs), Internet Chat spaces, Lotus notes databases, and Web-based interactive forums are just a few of the infrastructures that have been developed to support social activity. Although the functionality provided by these infrastructures varies, they all provide facilities that enable multi-person social communication. Each one allows individuals to engage in constrained many-to-many communication by broadcasting and receiving messages within a collection of other people. As a result, each of these systems provides the basic communication capabilities needed to support significant social activity.

However, as has been increasingly noted in discussions among developers, these networked systems only provide technical infrastructures in which social activity may take place (e.g. Hof, Browder, and Elstrom, 1997; Hagel and Armstrong, 1998). Like a park, dormitory commons area, or conference room, these systems simply provide a context in which people can interact. Just as providing a meeting area at academic conferences can encourage but not ensure social interaction, providing electronic infrastructures can support, but does not guarantee, the emergence of social activity.

To provide many of the expected benefits, such as improved information sharing and better coordination, communication infrastructures must do more that simply provide the facilities for communication. They must also be the site of sustained social structures. In traditional organizations, social structures affect how information flows and knowledge is shared by influencing which individuals communicate and what they discuss (Allen, 1977). For a communication infrastructure to significantly impact an organization it must become the site of social structures which support ongoing activity (Sproull and Kiesler, 1990; Finholt and Sproull, 1990). The availability of a technical infrastructure does not guarantee that individuals will be willing to join and participate in the collective activities that underlie social structures. Likewise, efforts to attract new members are likely to be wasted if online social structures fail to keep members over a longer term. Although definitions of success will vary from case to case, the ability of networked infrastructures to support useful social activity is significantly affected by the system's ability to encourage the emergence of sustainable online social structures; that is, structures that are able maintain their membership over the long term.

Drawing from studies of traditional small groups, voluntary associations, and organizations, this paper presents a model of the internal dynamics of membership sustainability in networked social structures. The core process in the model is a feedback loop that links membership size and features of the emergent communication activity with a networked social structure's ability to attract and retain its members. The model is then examined with longitudinal data collected from a sample of e-mail based Internet listservs. A set of log-linear, time-series/cross section regression models are estimated to test the relationships proposed in the model. Implications for researchers and practitioners interested in technology supported social activity are discussed and areas for future research are described.

A Resource-Based Model of Sustainable Social Structures

Traditional social structures arise because they are able to provide benefits that outweigh the costs of membership (Moreland and Levine, 1982). Social collectives that can provide positive net benefits are better able to attract and retain members, and hence survive over the long term. Traditional social structures provide many benefits for their members (Forsyth, 1990). They provide opportunities for affiliation or companionship that may be desired by individuals (McClelland, 1985; Rubenstein and Shaver, 1980; Russel, Peplau, & Cutrona, 1980). They provide opportunities to influence people, which are valued by individuals who have a need for power and control over others (Winter, 1973). Indirectly, social groups provide benefits by supporting the development of personal relationships. Belonging to a social group with another person increases the chances that you will have a social tie with them (Feld, 1982) and hence it is more likely that you will get information or assistance from them through personal communication (e.g. Granovetter, 1972). Traditional social collectives also provide information because they enable members to observe one another's behavior (Festinger, 1954). These structures also provide emotional and social support in the form of personal advice, small favors, and positive feedback. (Barrera, 1986; Cutrona, 1986; Sarason, Sherarin, Pierce, and Sarason, 1987). Social support, (Wellman and Whortley, 1989, 1990; Wellman, Carrington and Hall, 1988; and Wellman, 1990), personal relations, companionship (Roberts, 1998), access to information, the ability to ideas rapidly, (Kaufer and Carley, 1993), and collective action (Ostrom, 1990), have also been cited as important benefits provided by traditional communities. Collectives and organizations provide the foundation for accomplishing activities that would be difficult, or impossible for isolated individuals to perform. These are just a few of the benefits that enable traditional social structures to attract and retain members.

In networked social environments, membership sustainability is also linked with a structure's ability to provide benefits for individuals. Like traditional social structures, networked social structures provide a variety of benefits (Wellman, 1996, 1997; Wellman, Salaff, Dimitrova, Garton, Gulia, Haythornwaite, 1996; Roberts, 1998). Early studies suggested that computer-mediated communication limited the social cues that are the basis for power and status structures (Sproull and Keisler, 1991; Daft and Lengel, 1986). However, subsequent work suggests that the acquisition and use of power and influence can be an important component of online social activity (Saunders, Robey, and Vaverek, 1994; Walther, 1994; Baym, 1993). There are many networked social collectives that support discussion and knowledge sharing (Kling, 1996; Wellman, 1995; Abbot, 1988; Kraut & Attewell, 1993). Within organizations, online social structures allow individuals to access information and quickly disseminate their ideas (Finholt and Sproull, 1990; Sproull and Keisler, 1991; Constant, Sproull, and Keisler, 1996; Whittaker, 1996). Networked collectives can also support learning (Collins and Berge, 1996) and provide social and emotional support (Rice and Love, 1987; McCormick and McCormick, 1992; Haythronwaite, Wellman, and Mantei 1995; Walther, 1996; Wellman and Gulia, 1996; King, 1994). They also can support the development of interpersonal relationships, feelings of companionship, and perceptions of affiliation (Furlong, 1995; Hiltz, 1985; Walther, 1994; Rheingold, 1993; Meyer, 1989; Kraut, et. al. 1998). In addition, online associations and virtual organizations provide benefits by enabling a variety of collective activities, including software development and political action (Ogen, 1993). Like traditional social structures, networked social structures provide a variety of benefits for individuals, enabling them to attract and retain members.

Resources are the foundation of a social structure's ability to provide benefits. The availability of knowledge, time, energy, along with the more concrete financial and material resources, underlies the provision of benefits for individuals. To encourage information sharing, a networked social structure must have members who are knowledgeable about relevant topics. If a support group is to provide emotional encouragement and/or counseling it must have members who have the time and energy to be supportive. Providing members with the opportunity to participate in dramatic productions requires that a community theatre company must have financial and materials resources. Whether traditional and networked, social structures are sustainable only if they have access to resources that allow them to provide benefits for their members.

Although resources may come from outside a structure, current members are typically a crucial source. Social structures may be created explicitly to combine available resources (Fine & Stoecker, 1985) or they can emerge without a formal 'creation' action, arising simply because it is more effective for a collection of individuals to pool their resources, time, and energy. Whether it is managed or emergent, the availability of a resource pool is essential if a social structure is to be sustainable. Without resources, it is impossible to provide benefits, and without providing benefits, it is impossible to attract and retain members. Hence, the creation and maintenance of social structures inherently involves collective action, with resource contributions of individual members being aggregated to provide the benefits that sustain a structure's membership.

However, having many sources of resources available is not sufficient to sustain a social structure. It is also necessary for the pool of potential resources to be transformed, through social activity, into benefits for individuals. A networked social collective with many informed, knowledgeable members will not be sustainable unless it also supports discussion. A support group with many caring members who are willing and able to provide help will not be sustainable if it does not also support the communication that is necessary to turn an individual's time into the supportive contact that troubled people need. The amount of money and the quality of the facilities owned by a theatrical company become irrelevant if the organization is unable to support the coordination and management activities needed to execute a series of performances. Whether traditional or networked, to be sustainable social structures must support the social processes of providing benefits.

The resource-based model of sustainable social structure works from the premise that a structure's ability to attract and retain members is a consequence of the available resources (Figure 1).

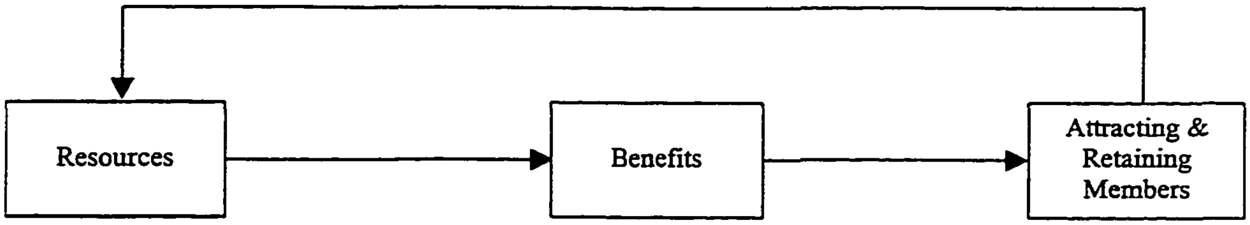


Figure 1: Resource-Based Model of Sustainable Social Structures

To be sustainable, social structures must have access to resources, typically through the contributions of the current members. In addition, the structures must also support the social processes that underlie the efficient transformation of those resources into benefits for individuals. Only then will the structures be able to attract and retain members, and hence be sustainable over the long term.

The Internal Dynamics of Sustainable Social Structures

The resource-based model presented above implies that there is a feedback process at the core of the internal dynamics of social structure sustainability. The current members of a structure act as key providers of resources. Though various communication dependent social processes, those resources are aggregated and transformed into benefits for individuals. Those benefits, in turn, enable the social structure to attract and retain members, hence developing and sustaining its membership. By affecting the available resources and the processes of providing benefits, current membership and communication activity within a structure interact in this feedback loop to affect the sustainability of a social structure.

Membership Size and Sustainability

Social structures maintain their membership by providing benefits. These benefits are derived from the resources, time, energy, and information provided by the current members. As a result, features of the current membership are an important factor in the development of sustainable social structures. Membership size is one feature of social structures that has received much attention in prior studies of organizations, associations, and other traditional social structures.

Membership size affects sustainability because it affects the resources available within a social structure. Larger voluntary associations are more likely to have access to greater economic resources (McPherson, 1983). By aggregating their members' knowledge, larger decision-making groups potentially have access to more information and might be expected to make better decisions (Wittenbaum and Stasser , 1996). Larger social structures also increase the chances that there is a member who knows the information needed to answer a question, has the time and interest to provide social support, or has some other needed resource. In all of these cases, membership size is significant because it is positively related to the potential availability of more time, energy, information and/or materials resources.

The sustainability of social structures is also positively affected by membership size because, in addition to contributing resources through their actions, members also contribute by being exposed to the messages sent by others. One of the benefits a social structure can provide is the ability to communicate with others. When making an announcement or distributing information, larger audiences will be preferred over smaller audiences with similar members. Likewise, individuals who seek visibility within a community or simply an audience for their ideas will be more likely to benefit from a social structure with more members. A group, organization, or association's ability to provide these benefits is dependent, at least in part, on the size of the audience that can be reached through it (Markus, 1990; Rafaeli and LaRose, 1992). By choosing to be exposed to the communication activity of a social structure, members are implicitly contributing resources, such as their time, energy, and attention, to the construction of the social structure's 'audience resource'. In this way, social structures are subject to positive externalities (Cornes, 1996). Although members may directly benefit from being exposed to communication, they also provide spillover benefits for others as a result of that choice. Whether because of the increased audience size or more explicit contributions, larger social structures will tend to have access to more resources than smaller structures. These resources allow them to provide more benefits, and hence attract and retain members more effectively than smaller structures.

However, while increasing size provides access to more resources, in traditional social structures it can also have significant negative effects. In larger face-to-face social structures, individuals have fewer opportunities to participate and less time to talk (Krech & Crutchfield, 1948). As a result, while larger membership leads to greater audience resources, it also makes it more difficult for individuals to benefit from that resource. The number of possible interaction partners increases non-linearly with size making it substantially more difficult to know the rest of the members (Bossard, 1945). This, in turn, reduces the chances that individuals will form relationships and receive benefits such as social support or opportunities for indirect influence. It also decreases the likelihood that individuals will know the entire membership well, increasing the chances that they will not be able to access the resources that are present. As traditional social structures increase in size, they are subject to increasing logistical problems (Hare, 1976). These problems can significantly hinder the processes by which resources are transformed into benefits, ultimately affecting a social structure's ability to attract and retain members.

In addition to the logistical problems, size may have a negative impact on the provision of benefits because it affects individuals' perceptions and attitudes (Slater, 1958; Milgram, Bickman, & Beckowitz, 1969). Larger social structures are more likely to be subject to free-riding and social loafing. Individuals will tend to contribute less time, energy, and resources because they expect that other members will provide enough to provide the desired benefits (Petty, Harkins, Williams, & Latone, 1977). Thus, while larger structures may have more potential resource providers, the amount of contributions per person (and overall) may be lower than in smaller social collectives (Olsen, 1965; Oliver and Marwell, 1993). If adequate resources are not contributed by the current membership, then the social structure will not be able to provide the benefits necessary to continue to attract and retain members. The undersupply of benefits in larger structures is reflected in the general finding that individuals in larger structures tend to be less committed and less satisfied (Slater, 1958; Cartwright, 1968; Indik, 1965) and hence less likely to join or remain members (Baumgartel & Sobol, 1959; Cleland, 1955; Porter & Lawler, 1965).

In traditional unstructured groups the negative impact of size typically outweighs the positive benefits which arise from the potential availability of additional resources Logistical constraints and free-riding behaviors combine to have a negative effect that overwhelms the positive consequences of size in all but the smallest of structures. For example, proponents of brainstorming argued that, under the right conditions, teams working together would be able to generate more ideas than individuals working alone

(Osborn, 1957). However, decades of research examining the operation of brainstorming groups failed to support this assertion, finding instead that the negative logistical and psychological effects of size consistently outweigh the expected benefits (Diehl & Strobe, 1987).

Adopting internal structures or alternative communication technologies are the primary approaches that are used to overcome the negative consequences of size. Internal structure addresses the logistical problems by constraining interaction within a social structure. Rather than allowing members to freely enter, interact, or leave, social collectives relying on internal structure may screen potential members, limit the timing, structure and content of interactions, or force uncooperative members to leave. The establishment of formal roles, communication structures, and membership requirements are other forms of internal structure. All of these structures can be seen as attempts by organizations and groups to shift the balance from the negative to the positive consequences of increased size.

However, while internal structure can reduce the negative consequences of size, it is not costless. Maintaining an internal structure requires resources. Screening potential members, enforcing constraints on interaction, and determining when to force a current member to leave are all activities that require a substantial expenditure of time, energy, and attention. In addition, because they constrain individuals' choices, internal structure is also costly for a group or organization because it reduces (or eliminates) many individuals' access to the benefits. Social collectives that have formal communication structures reduce the logistical problems that come with size by limiting individual's ability to directly communicate with one another. When internal structure is used within a social collective it must be applied carefully. Without internal structure, large traditional social structures are typically unsustainable. At the other extreme, a costly structure may significantly reduce the negative effects of size, but at the same time consume more resources that the increased membership can provide.

Another way that structures may reduce negative effects of increased size is through the use of alternative communication technologies. Researchers interested in technology supported groups have considered various effects that networked environments might have on the relationships between group size and process. Networked communication infrastructures provide features such as asynchronous communication buffering and archiving which have the potential to drastically reduce the the logistical problems that occur in traditional social structures (Nunamaker, Dennis, Valacich, Vogel, & George, 1991). Other features, such as member anonymity and the general invisibility of individuals (Finholt and Sproull, 1990) may lower the salience of a networked collective's membership, and hence reduce the negative psychological effects of size.

It has been argued that, unlike in traditional social structures, in networked collectives the negative impacts of size will be significantly reduced, potentially shifting the balance from the negative to the positive, even in the absence of formal structures. For example, in studies of online brainstorming groups it has been found that, unlike in traditional small groups, larger online brainstorming groups are able to generate more ideas than smaller teams (Connolly, 1997). Thus, it might also be expected that networked social structures would be better able than traditional structures to take advantage of the increased resources that can be provided by a larger membership while avoiding the problems. As a result, it may be the case that larger online social structures will find it easier to attract and retain members than comparable smaller structures. However, it remains uncertain whether, in the absences of internal structure, the use of alternative communication technologies reduces the negative impacts enough to shift the balance to the positive impacts of size.

Communication Activity and Sustainability

Communication activity is an important feature of sustainable social structures. No matter what resources are available within a structure, without communication activity those resources will remain dormant, and no benefits will be provided for individuals. The importance of communication activity is reflected in theoretical definitions of small groups and communities which highlight the importance of interaction (Bonner, 1959; Homans, 1950; Stogdill, 1959; Hare, 1976; Shaw, 1981) . Communication activity is an important feature of sustainable social structures because it plays a key role in the processes by which resources are converted into benefits for individuals. Without some form of communication activity influence, social support, coordination, or information sharing cannot occur. Thus, in the absence of communication activity a social structure will fail to provide benefits for individuals and ultimately be unsustainable because it is unable to attract or retain members.

To the degree that communication activity is at the core of the social processes underlying provision of benefits for individuals, there arises a direct, and positive relationship between the volume of communication activity and the amount of benefit provided. At the extreme, a social structure in which there is no communication at all cannot provide benefits for its members. Even nominal, or minimal, structures rely on some basic communication activity to support the formation of an identity among their members. More communication activity is expected to enable more information sharing, development of strong relationships among members, and coordination of more complex activities – all of which correspond to provision of greater benefits for individuals, allowing to the structure to attract and retain members more effectively.

On the other hand, benefits are not valued equally by all individuals. Communication activity seen by one individual as providing valuable benefits may be seen by another as noise. Information that is useful to one member may be distracting to another. Interaction that provides social support for one individual may be perceived as unimportant by others. It is rarely possible to provide 'benefits' that are valued equally by all members. Different types of communication activity provide different benefits, that are, in turn, valued by different subsets of a social structure's members and potential members.

However, while communication activity is an important factor in the provision of benefits it is also a major source of costs for the members of a social structure. Individuals incur costs when they contribute resources to a social structure. When individuals choose to actively participate in the communication, they are explicitly deciding to contribute their time, energy, attention, and knowledge. However, members also implicitly contribute resources to a social collective when they choose to remain a member, and hence remain part of the audience that is exposed to the communication activity. Attendees at conferences and meetings incur costs, in terms of time, energy, and possibly even financial resources, whether or not they choose to explicitly contribute to the communication activity. By being part of an audience, individuals contribute resources to a social structure, and in the process incur costs.

While more communication activity is likely to be associated with more benefits, it also imposes higher costs. Longer meetings, more issues of a newsletter, or more electronic mail messages all have the potential to support higher levels of information sharing, social support, and other benefits. However, they also require that individuals contribute more resources the structure's audience resource. Longer meetings require that everyone spend more time and energy. Higher volumes of print or e-mail force audience members to expend more time and attention in order to process the communication, even if they do not benefit from it personally. For an individual, more (and more diverse) communication activity is an improvement only if the benefits provided by that communication outweigh the costs of being exposed to it. For a social structure overall, higher volume and diversity of communication activity enhances its sustainability only if the number of members who are attracted or retained because of the additional benefits outnumber those who are lost due to the increased cost.

As with size, there are two approaches to managing the positive and negative impacts of communication on the sustainability of a social structure: internal structure or use of alternative communication technologies. Together these two components, internal structure and communication technology, comprise the communication infrastructure of a social collective. Both internal structure and technology facilitate sustainable social collectives by altering the costs of communication or constraining the content of communication activity. The use of specialized jargon, language, or symbols within a social structure reduces the costs of communication activity, making it possible for members to communicate complex ideas more efficiently. Formal summaries, meeting agendas, and structured presentations enable members to more selectively participate in the audience, which reduces the costs of being part of a social structure's audience. Editorial control of content also facilitates sustainable social structures by attempting to screen out communication activity that imposes costs on the membership of a social structure that are not consistent with the benefits provided.

However, as with internal structures for managing the effects of size, internal structures for managing the effects of communication are costly. Use of specialized language or symbols increases the costs associated with joining a social structure. So, while this type of internal structure can help a social structure retain members by lowering communication costs, it may also make it more difficult to attract new members. Formal summaries or editorial control can increase the overall benefit/cost ratio for a structure's communication activity, but it does so by redirecting resources that otherwise might have been used to directly provide benefits for members.

Using different technologies can also moderate the impact of communication activity on the sustainability of a social structure. In discussions of small groups and communities, it has often been assumed that face-to-face communication must be the basis for these social collectives' communication infrastructure. Underlying this assumption is the observation that physical meeting spaces can be an effective 'technology' for supporting communication activity which contributes to the sustainability of a social structure. However, it is an oversimplification to assume that only face-to-face interaction can support sustainable social structures. Kaufer and Carley (1993) argue that the application of print provided many social structures with an alternative to physical space/face-to-face communication infrastructures. Print-enhanced infrastructures, they argue, supported the development of large-scale, long-term social structures, such as professions and academic disciplines. Similarly, as networked communications technologies have developed, there has been increased interest in the idea of "virtual" social structure. It has been expected that virtual organizations, groups, and associations will, through use of new communication technologies, be able to overcome the problems and costs of using internal structure to manage the impact of communication activity on sustainability. However, at this time it remains unclear whether the use of new networked communication technology is itself enough to overcome the negative consequences (i.e. cost) of communication activity.

Summary

The resource model of sustainable social structure focuses on a feedback loop involving resources, benefit provision, and a collective's ability to attract and retain members (Figure 2).

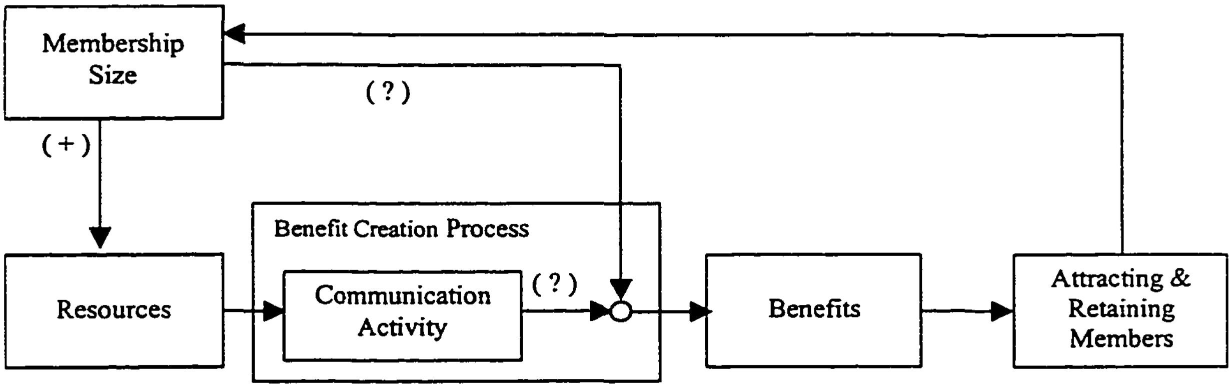


Figure 2: Impacts of Communication Activity on Sustainability

Within that process, membership size and communication activity are highlighted as emergent features of a social collective which have a significant role the sustainability of that structure. In traditional contexts, the negative consequences of membership size and communication activity tend to outweigh the positive effects, unless they are managed. Traditional social collectives often must adopt internal structures that moderate the negative effects of size and activity on the structure's sustainability by constraining membership and limiting communication.

However, the use of networked communication technology has the potential to significantly change the costs and form of communication, resulting in online social structures that are subject to a different balance of positive and negative effects. It is expected that features of networked communication infrastructures, such as reduced costs and support for asynchronous communication, will mitigate the negative effects of size and communication activity, even in the absence of internal structures. In the analysis that follows, data from a random sample of e-mail based social structures (listservs) will be used to test the hypotheses that in networked environments, even in the absence of internal structure, the net effects of size and communication activity on a structure's ability to attract and retain members will be positive.

Data, Measures, and Methods

The premise that networked communication technologies shift the impacts from the negative to positive was examined using data from a sample of electronic mail (e-mail) based Internet listservs. These online social structures use Internet-based electronic mail and a server (i.e. a list server or listserv) to centrally maintain a mailing list that enables individuals to broadcast text messages to the other members. Electronic mail based social structures were chosen for this study because of their prevalence, availability, and ability to support the necessary data collection. E-mail based social structures are known to be prevalent in both private and public networked environments

(Finholt & Sproull, 1990). As a result, e-mail based Internet social collectives are both representative of a large class of naturally occurring online social structures and available for study. Also, unlike other decentralized networked communications infrastructures, the centralized architecture of a listserv based social collective supports measurement of structural features such as membership size and change.

A stratified random sample was selected to ensure that the included listservs represented a variety of topical focuses and member populations. From a census of approximately 70,000 structures, a stratified base sample of 1066 listservs was created. One third of the initially selected collectives focused on work-related topics. One third focused on personal topics (hobbies, lifestyles, etc.). The remaining listservs considered topics that mixed work-related and personal interests (e.g geographic locations). A multiple stage confirmation process was then used to construct an analysis sample in which the listservs had comparable technology infrastructures and minimal internal structure. Online structures that integrate different network technologies such as e-mail, WWW conferencing tools, or USENET newsgroups were eliminated, focusing the sample on listservs that relied solely on electronic mail for supporting online communication. Social collectives that made use of identifiable internal structures, such as moderated listservs, newsletters, or formal new member screening, were also removed from the sample. Each listserv was checked to ensure that it was mechanically functional, able to provide the needed data, and available for inclusion in the study. The result of this process was a set of 284 unstructured listservs that relied on e-mail as a basis for online communication. Membership and communication activity records were collected for these listservs from August 1, 1997 to November 30, 1997. As data was collected and measures were constructed, the analysis sample was reduced to 206 as listservs ceased operation, changed structure, or restricted access to membership data (for additional information about the listserv selection process and the resulting sample see Butler, 1999).

Measures of size, communication activity, and membership change serve as the basis for examining the following analytical model (Figure 3).

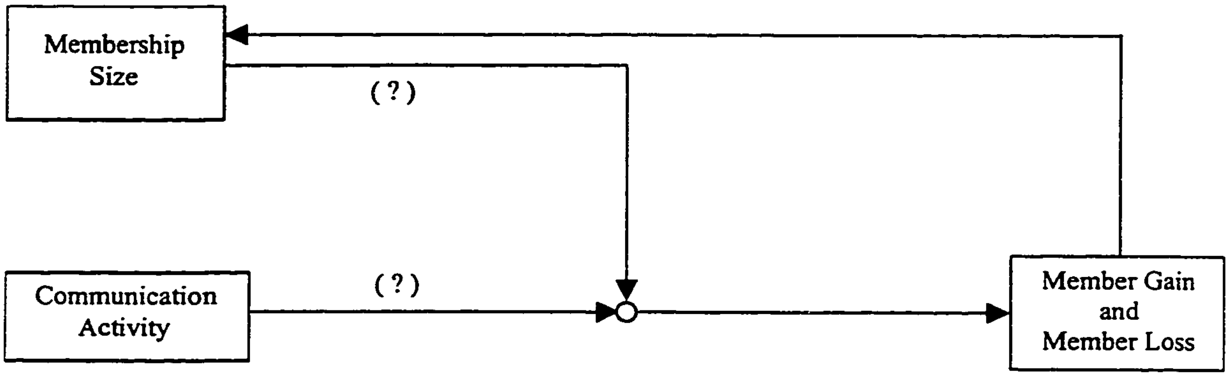


Figure 3: Analytical Model of Size, Communication Activity, and Sustainability

The analytical model is a reduced form of the theoretical model (Figure 2). The sample focuses on minimally structured social collectives that use electronic mail. This eliminates the need to explicitly include the moderating effects of internal structure and communication technology in the analytical model. Resources and benefits are also not included in the analytical model. Although they are expected to play a central role in the dynamics of social structure, they are not observable constructs. Furthermore, the individual nature of benefits and the valuation of benefits make it infeasible to explicitly assess the over time provision of benefits across a large sample of social structures. The constructs of size and communication activity are explicitly included. Communication activity is characterized in terms of volume and topic variation. A social structure's ability to attract and retain members is measured by member gain and loss. The positive and negative impacts of size and communication activity are assessed by estimating the relationship between these features of an online structure and changes in the structure's membership.

Membership size was measured by counting the number of individuals in a listserv's mailing list at the beginning of each month during the observation period. This measure is based on the premise that a social structure's members are those people who are exposed to the structure's internal communication activity. The impact of a social structure on an individual is limited if she is not exposed to the communication activity. Likewise, the impact of an individual is limited if she is not exposed to or contributing to the communication. Thus, in a listserv the relevant members are those individuals who receive the broadcast messages. While this measure may overestimate size by including individuals who receive messages but do not read them, it is conceptually equivalent to counting the number of meeting attendees, a common measure of size in studies of traditional groups and organizations. To simplify interpretation of the empirical results the size measure was divided by 100.

A listserv's communication activity consists of the text messages that are distributed to all members. Communication activity is characterized in terms of volume and topical variation. Communication volume is significant because it is linked with the costs of membership. In traditional social structures, more communication activity might result in more, or longer, meetings – which require more time from the members. Likewise, in online social structures, more communication results in more messages that in turn require more time to process (i.e. read or delete). The higher the volume of communication activity in a social structures, the higher the costs associated with membership. Higher membership costs force a structure to provide higher levels of benefits in order to attract and retain members.

While there is expected to be a relatively strong relationship between communication volume and costs of membership the relationship between activity volume and benefits is less clear. Differences in reading speed and time valuation aside, individuals incur approximately similar costs due to communication activity. However, because of the individualized nature of benefits volume is only probabilistically associated with benefits. Although all members incur the costs of each instance of communication activity, it is likely that only a subset of members benefit. In addition, as the subject of the communication activity varies, different subsets of the membership will be interested in, and benefit from the communication activity. Thus, communication volume and variation are related to the benefits that arise from that communication activity. Volume and variation interact to determine, for a given membership, the overall net benefits provided by the structure.

The volume of communication activity is measured by counting number of messages distributed with the mailing list during each month. Topic variation does not indicate how different topics are from one another; rather, it refers to the relative variation in the content of the communication. Low topic variation indicates that the messages in a listserv have focused on a small number of topics, while high topic variation means that many topics were considered.

Topic variation was inferred from the dialog structure of messages. In e-mail based social settings, participants link related messages by labeling later messages as replies to earlier messages. This linking creates sequences of related messages known as discussion threads. Topic variation in a listserv's communication is reflected by the concentration of messages within the discussion threads. Discussion threads were identified by removing the "re:" marker, a subject line tag commonly used to label messages as replies, and grouping messages based on the first 40 characters of the remaining subject line text. Concentration of messages across the discussion threads was characterized by computing a normalized Herifindal –Hirschman index [HHI][1] (Hirschman, 1964) for each day's communication activity. The HHI was chosen as the basis for the topic variation measure because it captures both concentration within a set of categories (i.e. discussion threads) and the number of categories (Davies, 1988). The HHI, which is a measure of concentration, was reversed (1-HHI) to create a measure of variation. In addition, the topic variation measure was set to 0 in cases where there were no message, indicating that there is no topic variation when there is no communication activity. The mean value for each month was then used to characterize the topic variation of a listserv's communication activity.

This measure of topic variation provides an indication of a listserv's participant's assessment of variation in the topics of structure's communication activity. Messages are identified as being similar or different based on the labels assigned by the participants themselves. Individuals label messages as replies when they expect their contribution to be of interest to the same individuals who read the earlier messages. Thus, this measure of topic variations, is roughly the equivalent of using knowledgeable coders to cluster

[1] The topic variation measure is computed according to the following formula:

$$\text{Topic variation} = (1 - \text{HHI}) = (1 - ([S^2_1 + S^2_2 + \ldots + S^2_n] / \text{ MsgCount}))$$

where S_i is the percentage of messages [0..1] which are part of discussion thread i and MsgCount is the total number of messages distributed that day.

messages based on the subset of the listserv who would be interested in them. While it is subject to noise arising from flawed labeling and alternative uses of the 're:' tag, this measure provides a general indication of whether the communication activity has the potential to provide benefits to few or many subsets of the listserv's membership.

An online social structure's ability to attract and retain members is reflected in the inflow of new members and the loss of current members over time. Individuals enter and leave listservs by requesting to be added to or removed from the mailing list. Hence, membership gain and loss rates can be determined by recording changes in the mailing list records. Member gain is measured by counting the individuals whose names are added to a listserv's mailing list each month. This count also includes individuals who are coming back to a listserv after a formal absence. However, since across the entire sample returning individuals represent fewer than 10% of all entering members, no adjustment was made. Member loss is measured by counting the people who are removed from a listserv's mailing list during each month. Calculation of monthly member gain and loss is based on aggregation of daily data to ensure that individuals who are members of a listserv for less than a month are also included.

The measures of size, communication activity, member loss, and member gain are the basis for testing the proposition that technology supported social structures are subject to a different balance of size and communication activity impacts (Table 1).

Construct	Label	Measure	Units
Membership Size	$Size_t$	The number of people on a listserv's mailing list at the beginning of month t.	100's of People
Communication Volume	$Volume_t$	The number of messages distributed to individuals on a listserv's mailing list during month t	Messages/Month
Topic Variation	$Variation_t$	The mean value of a reversed, normalized HHI index which reflects the number of topics and the distribution of messages among those topics each day during month t.	(Unitless)
Member Gain	$MemberGain_t$	The number of people who were added to a listserv's mailing list during month t.	People/Month
Member Loss	$MemberLoss_t$	The number of people who are removed from a listserv's mailing list during month t	People/Month

Table 1: Construct and Measure Summary

These measures were constructed for each of the listservs in the sample, resulting in 206 (listservs) x 4 (months) x 5 (variables) panel data set.

The analysis sample contains listservs covering a range of sizes and levels of communication activity (Table 2).

	Mean	Std Deviation	Minminm	Maximum
Membership Size	1.6841	2.8666	0.0300	23.9200
Communication Volume	33.2900	94.7733	0.0000	1084.0000
Topic Variation	0.0820	0.1797	0.0000	0.9008
Member Gain	10.2047	26.3640	0.0000	277.0000
Member Loss	7.4470	20.8389	0.0000	208.3333

Table 2: Descriptive Statistics for Listserv Data

The descriptive statistics for the dataset also indicate that there is significant variation in the member gain and loss measures (For additional descriptive analysis of the data set see Butler, 1999).

Analysis and Results

A set of time series, cross section, random effects, log-linear regression models was used to estimate the analysis model (Figure 3). The error structure for each model included a time-period dependent component, a cross-section (i.e. listserv) dependent component, and a component that was assumed to be normally distributed and independent of the time-period and listserv. Two-way random-effects models were used because both the sampled listservs and the sampled time-periods were representative of a large 'population' (Greene, 1993). Log-linear models were selected because they focus on interaction of size and communication activity, while minimizing problems with non-normal data. Conceptually, it is the interaction of communication activity and size that is expected to play a role in the sustainability of online social structures. Furthermore, distributions of the various measures are skewed. Applying the log transformation reduces the impact of this non-normality. Since there were a significant number of cases where the measures had zero values, a small constant (0.000001) was added to allow the log (base 10) transformation to be applied. The constant was chosen based on the smallest non-zero value in the data set. This minimized the transformation's impact of the ordering of the data.

Using these procedures, the following models[2] were estimated with the TSCSREG procedure in SAS (v6.12):

(a) $\mathrm{LOG(MemberGain_t)} = B_0 + B_1\mathrm{LOG(Size_t)} + B_2\mathrm{LOG(Volume_t)} + B_3\mathrm{LOG(Variation_t)}$

(b) $\mathrm{LOG(MemberLoss_t)} = B_0 + B_1\mathrm{LOG(Size_t)} + B_2\mathrm{LOG(Volume_t)} + B_3\mathrm{LOG(Variation_t)}$

[2] These log-linear models correspond to the following multiplicative models:

(a) $\mathrm{MemberGain_t} = 10^{B0}\ \mathrm{Size_t}^{B1}\ \mathrm{Volume_t}^{B2}\ \mathrm{Variation_t}^{B3}$

(b) $\mathrm{MemberLoss_t} = 10^{B0}\ \mathrm{Size_t}^{B1}\ \mathrm{Volume_t}^{B2}\ \mathrm{Variation_t}^{B3}$

Hence, the log-linear models capture the solitary interaction that is described in Figure 3.

Model (a) examines the impact of size, communication volume, and topic variation on a online social structure's ability to attract new members. Model (b) considers the impact of these features on a listserv's ability to retain members. The remaining relationship in the analytical model (Figure 3), the trivial link between member gain and loss, and subsequent changes in size, is not explicitly modeled.

The regression results suggest that size and communication activity have a significant impact on the ability of online social structures to attract and retain members (Table 3).

	MemberGain$_t$ (Log)	*MemberLoss$_t$ (Log)*
Intercept	-0.4009	-0.5436**
Size$_t$ (Log)	2.4859***	1.7849***
Volume$_t$ (Log)	0.2169***	0.3107***
Variation$_t$ (Log)	0.0739	0.2707***
R^2	0.3096	0.4367

* : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$
N = 206 * 4

Table 3: Impact of Size and Communication Activity on Member Gain and Loss

The significant coefficients for size and communication activity in the model predicting member gain suggests that these features positively impact a listserv's ability to attract members. Larger and more active listservs see higher rates of member gain. However, the significant coefficients for size, communication activity, and topic variation in the member loss model implies that these features also have a negative effect on a listserv's ability to retain members. Larger listservs and those with more and more varied communication activity have more member loss. Thus, size and communication activity have both positive and negative impacts on the sustainability of the online social structures.

It is possible that the negative relationship between size and a listserv's ability to retain members is an artifact of the measure of member loss. Member loss is assessed in terms of an absolute count of number of people who leave a listserv in a given month. This measure is subject to a variable ceiling. The number of individuals who could potentially leave is limited to the number of individuals who are present (size) and the number who have entered (member gain). Thus, it is possible that relationship between size and member loss is a trivial consequence of larger networked social structures having the potential for more people to leave.

To determine whether the effect of size on membership loss extended beyond the trivial ceiling effect, the member loss model was estimated using proportional member loss as an alternative measure of a listserv's ability to retain members. Proportionate member loss was assessed by dividing the absolute member loss by the listserv's size. This measure captures a social structure's ability to retain members, relative to its size. The results of this model indicate that the impact of size on a listserv's ability to retain members is not simply due to the variable ceiling for absolute membership loss (Table 4).

	Proportional Member Loss (Log)
Intercept	-0.7410***
Size_t (Log)	1.3443***
Volume_t (Log)	0.3263***
Variation_t (Log)	0.2535***
R^2	0.3907

* : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$
N = 206 * 4

Table 4: Proportional Member Loss Model Results

The coefficient of size remains statistically significant and positive. These results suggest that larger listservs will tend to have a higher relative loss rate. Not only do larger online structures lose more members, they also lose a larger percentage of their membership than smaller structures.

In the analytical model considered above (Figure 3) implies that a structure's membership size and communication activity are independent. However, this is inconsistent with the resource-based model of social structure (Figure 2). Size affects the volume of resources that are available to a structure. More members also mean that there are more individuals attempt to use the structure's resources. Communication activity is a key part of the process of converting resources into benefits for individuals. Thus, more resources and needs should be associated with more (and more varied) communication activity (Figure 4).

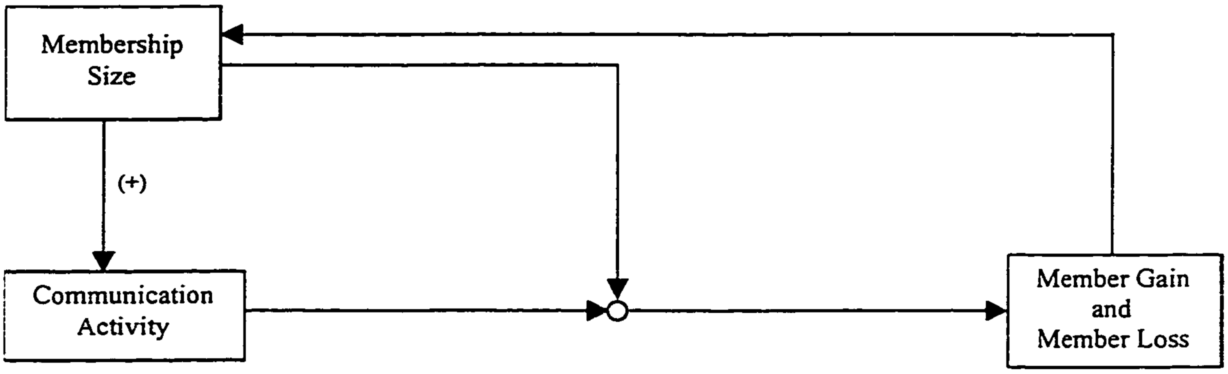


Figure 4: Size – Communication Mediation Model

In this model, communication activity acts as a mediator between size and a listserv's ability to attract and retain members. Under this model, some or all of the influence of size is the result of the effect of size on communication activity. Size impacts communication activity, which in turn influences member gain and loss.

The presence of mediation can be tested by a estimating a series of regression models and comparing the results (Baron and Kinney, 1986). First models of size and the membership change variables (gain and loss) were estimated. Then models were estimated to assess the relationship between size and communication volume and variation. These results indicate that there is a significant, positive relationship between size and the features of a listserv's communication activity. Finally, a model that includes size, communication volume, and variation, is estimated. This set of models was estimated for both member gain and member loss.

	Size & Communication Activity		**Member Attraction**		**Member Retention**	
Dependent Variable	(a) Volume (Log)	(b) Variation (Log)	(c) Member Gain (Log)	(d) Member Gain (Log)	(e) Member Loss (Log)	(f) Member Loss (Log)
Intercept	-1.5229***	-3.7427***	-1.0096***	-0.4009	-2.032***	-0.5436**
Size$_t$ (Log)	2.9411***	1.9474***	3.2541***	2.4859***	3.2141***	1.7849***
Volume$_t$ (Log)				0.2169***		0.3107***
Variation$_t$ (Log)				0.0739		0.2707***
R^2	0.1013	0.0943	0.2380	0.3096	0.2046	0.4367

* : p < 0.05, ** : p < 0.01, *** : p < 0.001
N = 206 x 4

Table 5: Mediated Model Results

The difference between the size coefficients in the direct models (c and e) and the comparable coefficient n in the full models (d and f) indicates that for both member gain and loss, there is a mediation structure involving size and communication activity. For member gain, there is a reduction of 0.7682 in the size coefficient when comparing the size model (c) with the full model (d) (one-tailed test: t = 1.8475, p < 0.05). This reduction implies that 23% of the impact of size on member attraction is accounted for by the indirect link between size and membership gain that is mediated by communication activity. In the case of member loss, the size coefficient drops by 1.4292 (or 44%) when communication volume and variation are included (one-tailed test: t = 3.5090, p < 0.001). Overall these results indicate that some portion of the impact of size on online social structure's ability to attract and retain members is mediated by communication activity.

However, in both cases there remains a significant direct relationship between size and the membership change variables. Thus, while the link between size and communication activity may play a role in the sustainability of online social structures, size has other, more direct, consequences as well.

Discussion

The resource-based model of sustainable social structure states that size and communication activity can each have positive and negative effects on a social structure's ability to attract and retain members. Thus, it is the balance of these effects that plays a key role in the ultimate sustainability of a social structure. In traditional contexts, the logistical and free-riding problems which arise from increased size and communication activity often outweigh the positive impacts. To manage these problems, traditional social structures adopt internal structures that limit membership, participation, or communication content. However, while internal structure can enable groups and organizations to maintain larger size or more communication activity, these structures also require resources and constrain the provision of benefits.

It has been proposed that networked communication technology may support new forms of social structure by significantly reducing the negative consequences of increased size and communication activity. Theoretically, reducing the logistical and free-riding problems that arise as a result of larger size and higher volume of communication activity would allow online social structures to remain sustainable without the need for extensive internal structure.

However, this analysis suggests that, at least for e-mail based Internet social structures, the use of networked communication technology does not fundamentally change the problems underlying the development of sustainable social structures. Size and communication activity have both positive and negative effects on the sustainability of online social structures. Larger listservs are better able to attract members, but they are less able to keep them. Likewise, listservs with more communication activity are more able to attract members, but less able to retain them. This result suggests that, as with traditional social structures, developing and maintaining sustainable online social structures requires a balance of the positive and negative impacts of size and activity. When the impact of size on member attraction and retention is considered alone (models c and e in Table 5), there is no significant difference between the magnitude of the positive and negative impacts. While larger structures are subject to more turnover, size does not seem to have a net impact on the sustainability of these online social structures. This is to be expected, since unsustainable social structures would have ceased operation and hence not be included in the sample.

When the interaction of size and communication activity is also considered, a more complex set of interrelationships emerges. The coefficient of size in the member gain model (model d in Table 5) is larger than the coefficient of size in the member loss model (model e) (one-tailed t-test: $t = 1.76$; $p < 0.05$). This result suggests that overall the direct effect of current size on a listserv's sustainability is positive. However, comparison of the communication activity effects suggests that the impact of communication activity is in the opposite direction. Although the difference is only significant for the variation (one-tailed t-test: $t = 1.717$; $p < 0.05$), the negative impact of more and more varied communication activity on the retention of members appears to be greater than the positive impact on member attraction. Thus, while the direct impact of size on a listserv's ability to attract and retain members is generally positive, the net impact of communication is negative. This result implies that within these social structures, size and communication may act to offset one another. This is further complicated by the presence of an indirect, or mediated, relationship involving size and communication activity. Size is positively associated with communication volume and variation, which in turn negatively impact sustainability.

Overall, these results suggest that although the use of networked technologies may change the mechanics of communication, it does not seem to change the fundamental problems associated with maintaining sustainable social structures. As has been seen in traditional social structures, in online social structures, size and communication activity have mixed effects on a structure's ability to attract and retain members.

As with any research, this work is subject to limitations. The online social structures considered here make use of one type of networked communication technology (centralized e-mail servers), have minimal internal structure, and exist in a public network (the Internet). Future research should consider how the resource-based model might be applied to online social structures that use other technical infrastructures or more complex internal structures. Infrastructure designers would benefit from additional work examining how social structures that use 'pull' technologies, such as WWW conferencing, or hybrid infrastructures, which combine traditional and network communication infrastructure, differ from the structures considered here which use a 'push' technology (e-mail). Assessment of the impact of internal structures, such as moderation and member screening, on a structure's ability to attract and retain members would also be useful when developing the social and managerial infrastructure for support online communication. Also, consideration of social structures operating within other, larger contexts, such as within an organization or a well-defined community, would provide additional insight into the factors which underlie the relationships between size, communication activity, and sustainability.

Conclusion

Although much has been written about the idea of virtual teams, organizations, and communities, it remains unclear whether or not these social structures are truly "new forms". Although networked communication technology alters the mechanics of communication, there is no significant evidence that it alters the fundamental problems which underlie the development of long-term social structures. Size and communication activity, which are themselves interrelated, influence a structure's ability to attract and retain members, which in turn affects the sustainability of a structure in the long term. Hence, both researchers and practitioners interested in online social structure would benefit from considering not just the differences introduced by technology but also the fundamental problems that underlie the development and maintenance of any social structure.

References

Abbot, A. (1988) The system of professions: An essay on the division of expert labor. Chicago: University of Chicago Press.

Allen, Thomas. J. (1977) Managing the Flow of Technology: Technology transfer and the dissemination of technological information within the R&D organization. MIT Press: Cambridge MA.

Barrera, M. Jr. (1986) Distinctions between social support concepts, measures, and models. American Journal of Community Psychology, Vol. 14, 413-422.

Baumgartel, H., & Sobol, R. (1959) Background and organizational factors in absenteeism. Personnel Psychology, Vol. 12, pp. 431-443.

Baym, N. (1993) Interpreting soap operas and creating community: Inside a computer-mediate fan culture. Journal of Folklore Research, Vol. 30, pp. 143-176. [Also appears in S. Kiesler (Eds.), Culture of the Internet]

Bonner, H. (1959) Group dynamics: Principles and applications. New York: Ronald.

Baron, Rueben M. and Kenny, David A. (1986) The Moderator-Mediator Variable Distinction in Social Psychology Research: Conceptual, Strategic, and Statistical Considerations. Journal of Personality and Social Psychology, Vol. 51, No. 6, 117-1182.

Bossard, J.H. (1945) Law of family interaction. American Journal of Sociology, Vol. 50, pp. 292-294.

Butler, B. S. (1999) When is a group not a group: An empirical examination of metaphors for online social structure. Unpublished Working Paper, University of Pittsburgh.

Cartwright, Dorwin (1968) The Nature of Group Cohesivenss. In D. Cartwright & A. Zander (Eds.) Group Dynamics: Research and Theory (3rd Edition) New York: Harper & Row.

Cornes, Richard and Sandler, Todd. (1996) The theory of externalities, public goods, and club goods. Cambridge University Press: New York.

Cleland,S. (1955) Influence of plant size on industrial relations. Princeton University Press.

Collins, Mauri P., and Berge, Zane L. (1997) Electronic Discussion Group Lists in Adult Learning. Unpublished Manuscript. Northern Arizona University.

Connolly, T. (1997) Electronic brainstorming: Science meets technology in the Group Meeting Room. In S. Kiesler (Ed.) Culture of the Internet. Mahwah, NJ: Lawrence Erlbaum.

Constant, D., Sproull, L, and Kiesler, S. (1996) The kindness of strangers: The usefulness of electronic weak ties for technical advice. Organizational Science, Vol. 7, pp. 119-135.

Cutrona, C.E. (1986) Objective determinants of perceived social support. Journal of Personality and Social Psychology, Vol. 50, 349-355.

Daft, Richard L. and Lengel, Robert H. (1986) Organizational Information Requirements, Media Richness and Structural Design. Management Science, Vol. 32, No. 5. pp. 554-571.

Davies, Stephen (1988) Choosing between concentration indices: the Iso-concentration curve. Economica, 46, 67-75.

Diehel, M. & Strobe, W. (1987) Productivity loss in brainstorming groups: Towards the solution of a riddle. Journal of personality and social psychology, Vol. 53, pp. 497-509.

Feld, Scott (1982) Structural determinants of similarity among associates. American Sociological Review, Vol. 47, 797-801.

Festinger, L. (1953) A theory of social comparison processes. Human Relations, Vol. 7. 117-140.

Fine, G.A. and Stoecker, R. (1985) Can the circle be unbroken? Small groups and social movements. In E. Lawler (Ed.) Advances in Group Processes, Vol. 1 (pp. 1-28). Greenwich: JAI Press.

Finholt, T., & Sproull, L. (1990) Electronic Groups at Work. Organization Science, 1, 41-64.

Forsyth, Danelson R. (1990) Group Dynamics (2nd. Ed.) Pacific Grove, CA: Brooks/Cole.

Furlong, M.S. (1995) An electronic community for older adults: the SeniorNet network. Journal of Communication, 39(3), 145-153.

Grannovetter, M. S. (1973) The strength of weak ties. American Journal of Sociology, Vol. 78, 1360-1380.

Greene, William H. (1993) Econometric Analysis (2nd Edition). Macmillan Publishing: New York.

Hagel, John III, & Armstrong, Arthur G. (1997) net.gain: Expanding markets through virtual communities. Boston, MA: Harvard Business School Press.

Hare, A. P. (1976) Handbook of small group research (2nd Ed.). New York: Free Press.

Hathornthwaite, Caroline, Wellman, Barry, and Mantei, Marylin (1995) Work Relationships and Media Use: A social network analysis. Group Decision and Negotion, 4(3), 193-211.

Hiltz, Starr Roxanne (1985) Online communities: A case study of the office of the future. Norwood, NJ: Ablex Publishing.

Hirschman, Albert O. (1964) The paternity of an index. The American Economic Review, Vol. 54, No. 5., p. 761.

Hof, Robert D., Browder, Seanna, & Elstrom, Peter (1997) Internet Communities, Business Week, May 5, 1997.

Homans, G.C. (1950) The human group. New York: Harcourt, Brace, and World.

Indik, B.P. (1965) Organization Size and Member Participation: Some Empirical test of alternative explanations. Human Relations, Vol. 18, pp. 339-350.

Kaufer, David S. and Carley, Kathleen M. (1993) Communication at a Distance: The Influence of Print on Sociocultural Organization and Change. Hillsdale, NJ: Lawrence Erlbaum.

King, Storm (1994) Analysis of electronic support groups for recovering addicts. Interpersonal Computer Technology, 2(3), pp. 47-56.

Kling, R. (1996) Social relationships in electronic forums: Hangouts, salons, workplaces, and communities. In R. Kling (Ed.) Computerization and Controversy (2nd Ed.) San Diego: Academic Press.

Kraut, R.E. and Attewell, P (1993) Electronic Mail and organizational knowledge. Working Paper, Carnegie Mellon University.

Kraut, Robert, Scherlis, William, Mukhopadhyay, Tridas, Manning, Jane, and Kiesler, Sara (1996) The HomeNet field trial of Residential Internet Services. Communications of the ACM, Vol. 39, No. 12. pp 55-63.

Krech, D. & Crutchfield, R.S. (1948) Theory and problems of social psychology New York: McGraw Hill.

Markus, M. L. (1990) Toward a "Critical Mass" theory of interactive media. In J. Fulk & C. Steinfield (Eds.) Organizations and Communication Technology (pp. 194-218). Newbury Park, CA: Sage.

McClelland, D.C. (1985) How motives, skills, and values determine what people do. American Psychologist, Vol. 40, 812-825.

McCormick, N and McCormick, J. (1992) Computer friends and foes: content of undergraduates' electronic mail. Computers in Human Behavior, 8: 379-405.

McPherson, Miller (1983) The Size of Voluntary Organizations. Social Forces, Vol. 61, No. 4, pp. 1045-1064.

Marwell, Gerald and Oliver, Pamela (1993) The critical mass in collective action: a micro-social theory. Cambridge University Press: New York.

Meyer, G.R. (1989) The social organization of the computer underground. Master's Thesis. Northern Illinois University.

Milgram, S., Bickman, L. & Berkowitz, L. (1969) Note on the drawing power of crowds of different sizes. Journal of Personality and Social Psychology, Vol. 13, pp. 79-82.

Moreland, R. L. & Levine, J. M. (1982) Socialization in Small Groups: Temporal Changes in Individual-Group Interactions. Social Psychology, Vol. 15, New York: Academic Press.

Nunamaker, J.F., Dennis, A.R., Valacich, J.S., Vogel, D.R., George, J.F. (1991) Electronic meeting systems to support group work. Communication of the ACM, Vol. 34, p. 40-61.

Ogan, Christine (1993) Listserver Communication during the Gulf War: What Kind of Medium is the Bulletin Board? Journal of Broadcasting and Electronic Media, Vol. 37, No. 2, pp. 177-196.

Olson, Mancur (1965) The logic of collective action: public goods and the theory of groups. Harvard University Press: Cambridge, MA.

Osborn, A.F. (1957) Applied imagination. New York: Scribner.

Ostrom, Elinor (1990) Governing the Commons: The evolution of institutions for collective action. New York: Cambridge University Press.

Petty, Richard E., Harkins, Stephen, G., Williams, Kipling, D., & Latane, Bibb (1977) The effects of group size on cognitive effort and evaluation. Personality and Social Psychology Bulletin, Vol. 3, pp. 579-582.

Porter, L.W. & Lawler, E.E. III (1965) Propoerties of organization structure in relation to job attitudes of job behavior. Psychological Bulletin, Vol. 64, pp. 23-51.

Rafaeli,S. and LaRose,R.J. (1993) Electronic bulletin boards and 'public goods' explanations of collaborative mass media, Communications Research, Vol. 20, No. 2, pp. 177-197.

Rheingold, H. (1993) The virtual community. Reading, MA: Addison-Wesley.

Rice, Ronald E. and Love, Gail (1987) Electronic Emotion: Socioemotional Content in a Computer-Mediated Communication Network. Communication Research, Vol. 14, No. 1, pp. 85-108.

Roberts, Teresa L. (1998) Are newsgroups virtual communities? Proceedings of CHI'98, Los Angeles, CA.

Rubenstein, C. M., & Shaver, P. (1980) Loneliness in two northeastern cities. In J. Hartog & R. Audy (Eds.), The anatomy of loneliness. New York: International Universities Press.

Russell, D., Peplau, L.A., & Catrona, C.E. (1980) The revised UCLA loneliness scale: Concurrent and discriminant validity evidence. Journal of Personality and Social Psychology, 39, 472-480.

Sarason, B.R., Shearin,E.N., Pierce, G.R. and Sarason, I.G. (1987) Interrelations of social support measures: Theoretical and practical implications. Journal of Personality and Social Psychology, Vol. 52., 813-832.

Saunders, C., Robey, D., and Vaverek, K. (1994) The persistence of status differentials in computer conferencing. Human Communication Research, Vol. 20, pp. 443-472.

Shaw, Marvin E. (1981) Group Dynamics: The Psychology of Small Group Behavior (3rd. Edition). New York: McGraw-Hill.

Slater, P.E. (1958) Contrasting correlates of group size. Sociometry, Vol. 21, pp. 129-139.

Sproull, L., & Kiesler, S. (1990) Connections: New ways of working in the networked organization. Boston, MA: MIT Press.

Stogdill, R.M. (1959) Individual behavior and group acheivement. New York: Oxford.

Walther, J. B. (1994) Anticipated ongoing interaction vs. channel effects on relationshal communication in computer-mediated interaction. Human Communication Research, 20(4), pp. 473-501.

Walther, J.B. (1996) Computer-mediated communication: impersonal, interpersonal, and hyperpersonal interaction. Communication Research, 23(1), 3-43.

Wellman, Barry (1990) The place of kinfolk in community networks. Marriage and Family Review, 15, pp. 195-228.

Wellman, Beverly (1995) Lay Referral Networks: Using Convential Medicine and Alternative Therapies for Low Back Pain. Sociology of Health Care, 12: 213-238.

Wellman, Barry (1996) Are personal communities local? A dumptarian reconsideration. Social Networks, 18. 347-354.

Wellman, Barry (1997) An Electronic Group is virtually a social network. In S. Kiesler (Ed.) Culture of the Internet. (pp. 179-205). Lawrence Erlbaum: Mahwah, NJ.

Wellman, Barry, Carrington, Peter, and Hall, Alan (1988) Networks as Personal Communities (pp. 130-184) in B. Wellman and S.D. Berkowitz (Eds.) Social Structures: A network approach. Cambridge University Press: New York.

Wellman, Barry and Gulia, Milena (1998) Net Surfers Don't Ride Alone: Virtual Communities as Communities. In P. Kollock and M. Smith (Eds.) Communities in Cyberspace. University of California Press: Berkeley.

Wellman, Barry, Salaff, Janet, Dimitrova, Dimitrina, Garton, Laura, Gulia, Milena, Haythornthwaite, Caroline (1996) Computer Networks as Social Networks: Collaboratibe Work, Telework, and Virtual Community. In Annual Review of Sociology, 22, pp. 213-38.

Wellman, Barry and Whortley, Scot (1989) Brothers' Keepers: Situating Kinship Relations in Broader Networked of Social Support. Sociological Perspectives, 32: 273-306.

Wellman, Barry and Whortley, Scot (1990) Different Strokes from Different Folks: Community Ties and Social Support. American Journal of Sociology, 96: 558-88.

Winter, D.G. (1973) The power motive. New York: Free Press.

Whitaker, Steve (1996) Talking to strangers: an evaluation of factors affecting electronic collaboration., .Proceedings of CSCW'96, Boston, MA.

Wittenbaum, Gwen M. and Stasser, Garold (1996) Management of information in small groups. In J. L. Nye and A. M. Brower (Eds.) What's social about social cognition? : Research on socially shared cognition in small groups. (pp. 261-282). Thousand Oaks, CA : Sage.

Communication Cost, Attitude Change and Membership Maintenance: A Model of Technology and Social Structure Development

Brian S. Butler

April 21, 1999

Communication Cost, Attitude Change and Membership Maintenance: A Model of Technology and Social Structure Development

Abstract

Voluntary electronic collectives, in which a networked infrastructure supports bounded many-to-many communication, are one of the most common social structures in online contexts. Like other natural groups, features of these collectives, such as member commitment and identity, develop over time. Prior studies of group development provide some indication how voluntary collectives might operate. Psychological studies of group dynamics have considered how members' attitudes change, while structural models have examined the role of member movement in the development of voluntary collectives. However, because existing development models typically do not consider particular communication mechanisms and processes, they provide little insight into how different communication infrastructures will affect the development of voluntary collectives.

This paper integrates the processes of individual belief change and member movement in a dynamic model of voluntary collective development. Contributed messages create a composite signal, providing members with noisy and incomplete information about the collective. This information changes members' beliefs; those beliefs, in turn, are used as the basis for deciding whether or not to continue as a member of the collective. Communication costs, a feature of the communication infrastructure, affect a collective's development by moderating the process of member belief change. The processes of communication, individual belief change, and membership maintenance form a cycle that underlies the development of the collective.

To develop this theory of voluntary collective development, a dynamic, multi-agent computational model was developed, validated, and analyzed. During development, the model was calibrated based on a subset of the empirical data collected from a random sample of e-mail based Internet listservs. Using the remaining data, the model was validated, focusing on the ability of the model to accurately represent a type of structural change in social collectives. A set of virtual experiments was conducted to determine the model's predictions regarding the impact of alternative technologies on collective development. The results imply that reduced communication costs, as are expected in networked environments, slow down the development process, resulting in voluntary collectives which have more (and more diverse) members while also being less stable than traditional face-to-face associations.

Voluntary collectives[1], in which membership and participation is a matter of personal choice, are a common social structure in online environments. However, unlike the small groups considered in prior studies of computer-mediated communication systems, voluntary collectives are dynamic structures (McGrath and Hollingshead, 1994). Some, including many that might be thought of as "required", see high rates of membership loss as individuals stop regularly looking at shared messages (Butler, 1999a; Whittaker, 1996). Some have high message volumes (Rojo, 1995; Sproull and Faraj, 1997), while many see little or no activity (Butler, 1999a). Their structural characteristics, including size, membership composition, communication volume, and topical diversity, vary significantly between (Butler, 1999a) and within particular collectives (Baym, 1993; Zernhausern & Wong, 1997). Thus while prior research on technology supported small groups may provide some insight, it does not directly address the question of how the use of new technologies affects voluntary collectives.

Face-to-face meetings, print, telephones, and electronic mail are all technologies that can support communication among members of a voluntary collective. Each has associated with it different costs and can support different communication structures. Yet communication is only part of the operation of social collectives: members join and leave, individuals' perceptions and evaluations change, and from these processes the focus and activity of the collective develop. Thus while networked environments may support more efficient communication, it is unclear how altering the communication costs and structures will affect the development of voluntary collectives.

[1] The term 'collective' is used instead of group when referring to the current research to avoid referencing the metaphor of 'small groups' which is prevalent in existing studies of computer-mediated social behavior. As we have argued elsewhere (Butler, 1999a) many online social structures are not structurally similar to the prototypical small group, and misapplication of this small group model can result in inappropriate conclusions regarding the nature of these structures.

Existing streams of group development research independently consider the role of member attitude change, at the individual level, and membership movement, at the group level, on the development of voluntary collectives (e.g. Tuckman, 1965, 1977; McPherson, 1983a,1983b,1990). The individual approach to studying social collectives is characterized by a concern with the perceptions, attitudes, and behavior of individuals within small groups. From this perspective, studies of development focus on how members of a collective change over time. (c.f. Gersick, 1988). Social collectives are seen as a context in which to consider individuals' mental state and processes (Moreland, Hogg, & Hains, 1994). Although this approach is concerned primarily with changes in individual members, some theorists have proposed models that characterize the social processes of small groups in terms of developmental stages (e.g. Hare, 1976; LaCoursiere, 1980; McGrath, 1984). The most prominent example, presented by Tuckman (1965; Tuckman and Jensen, 1977), outlines a set of phases: forming, storming, norming, performing, and adjourning.

However, sequential stage models are limited in that they are essentially descriptive models, providing a set of snapshots of a prototypical group (Tuckman ,1965; Hare, 1976; Poole, 1983). They have little to say about the mechanisms which underlie the observed changes (Gersick, 1988) and the sequential structure is a poor representation of the development process (Fisher, 1970; Scheidel & Crowell, 1964; Poole, 1981, 1983). As a result, recent work has begun developing models of the social processes which link individual attitudes and behaviors to developing features of a social collective (Gersick, 1988, 1989; Moreland and Levine, 1982; Worschel, 1996). However, because they have historically focused on small groups that, by definition, consist of a few members meeting face-to-face, these models have had little to say about the role of communication and communication technologies in the development processes of social collectives.

Under the structural approach, social collectives are seen not as environments for member change, but as dynamic social entities that arise within larger social systems (c.f. Carley, 1990). Structural theorists characterize this system in a variety of terms, including institutional (Blau, 1967; Etzioni, 1964; Simmel, 1955), ritual (Goffman, 1959), and competitive (Hannen and Freeman, 1977; McPherson, 1983b; McPherson and Rotolo, 1996) models. Development studies focus on the social processes, such as membership movement and participation, which underlie the formation and continued existence of collectives as entities. The models typically do not describe the mechanisms that link individual behaviors and the development of social collectives (Carley, 1991). Consequently, with the exception of recent work with computational models of emergent social structures (Kaufer and Carley, 1993; Carley and Wendt, 1991; Carley, 1995a, 1995b) these models, like those considered within the individual approach, do not consider the role of communication or communication technology in the development and maintenance of social collectives.

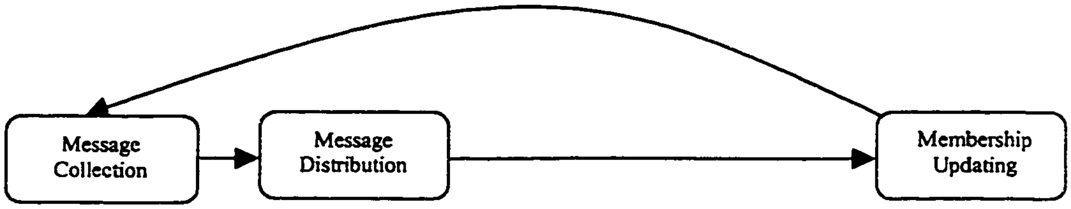
Although traditional studies of social collectives have typically not considered the impact of alternative communication infrastructures, there is a growing literature about technology-supported small groups and social collectives. Studies of group decision support systems (GDSS) and electronic meeting systems (EMS) have generally adopted the individual approach, considering how the use of new communication technologies change individuals' attitudes, perceptions, and behaviors in small groups (for review see McGrath and Hollingshead, 1994). This work provides insight into the social behaviors of individuals in networked environments. However, as with traditional studies taking this approach, it often unclear how (or if) changes in individual members will be reflected in the development of the collective as a whole (McGrath and Hollingshead, 1994).

In contrast, field studies of networked social collectives have generally adopted a structural perceptive, focusing not on individuals but on the processes and structure of social collectives (for review see Butler, 1999a). Through rich description, these studies provide insights into the features of collectives in networked environments. However, much of this work is focused on the project of demonstrating that "real" social activity can occur in online contexts. As a result, it fails to systematically consider how alternative technologies might differentially impact the development process in social collectives.

Drawing from both the individual and structural streams, we propose a model of voluntary collectives that considers the role played by the communication technology. Individual change, in the form of member attitude shifts, and structural change, in the form of membership movement into and out of the collective, are integrated to describe the development processes of voluntary social collectives. Communication activity within a collective acts as a signal, providing information about the emergent features of the membership of that social structure. Communication technology is seen as setting the communication costs incurred by members, costs which impact a collective's development by affecting the way individuals receive and process information about the collective. Through the communication infrastructure, that signal affects individuals' perceptions; those perceptions, in turn, underlie their decision to continue or end membership in the collective. Thus, the process of individual attitude change underlies the movement of members out of the collective, and together these changes result in the emergent features of collective composition, size, and focus.

A Model of Voluntary Collectives

The social system considered here consists of a set of N individuals and a collective (C). Unlike recent theoretical studies of social structure, in which groups and collectives are conceptualized as emergent results (Carley, 1991,1995a,1995b; Epstein and Axtell, 1996), a social collective exists as an independent social entity which is known to the individuals within the system. The collective has internal processes that are not emergent (Figure 1). People associate with, interact with, and evaluate a collective, not as a collection of distinct, known individuals, but as a composite social 'agent'.



w: Message impact
c: Noise cost (per message)
m: Message volume threshold

Members = {Members(1),...,Members(t)} : The set of individual agents which are members of the group during each time period.
Messages = {Messages(1),...,Messages(t)} : The set of messages collected and distributed during each time period

Figure 1: Collective Structure

One of the common features of social groups and associations is that there is some form of communication among the members[2] (Forsyth, 1990). A voluntary social collective is a socio-technical structure that supports bounded many-to-many communication among a collection of people who individually choose to contribute and be exposed to the communication activity. A collective's membership is the set of individuals (members) who are exposed to and

[2] In contrast, nominal groups and psychological groups are defined in terms of individual perceptions and, in theory, may exist without inter-member communication (Forsyth, 1990). However, even for these types of social structure the underlying perceptions are the result of some sort of communication activity.

have access to the collective's communication infrastructure[3]. Over time, as individuals leave the collective, the membership changes. The membership list is implemented as a binary matrix in which Members_{it} is 1 if individual *i* is a member during time period t and 0 otherwise.

Communication activity within the collective occurs as messages, discrete units of communication each of which has a single topic and is distinguishable from other messages only in terms of that topic. Collective members create messages. After all of the members have made a decision about participating, the set of messages is distributed, exposing each member to the collectives communication activity[4]. The message list is implemented as a matrix in which the entries (Message_{it}) are set as [0,1] values indicating a message topic if individual *i* sent a message during time period *t* and –1 otherwise[5]. The set of non-negative values in a column of the Message matrix (Message_t) describes the communication activity within the collective during time period *t*. After all of the members have processed the messages, updated their evaluation of the collective, and made a decision about maintaining their membership, the collective's membership list is updated. Members are only removed from the membership list if they explicitly decide to terminate their membership. Consequently, once an individual agent is added to a group's membership it remains a member, and sees all group messages, until it takes explicit action to terminate its connection with the group[6]. This processes of message gathering,

[3] Although this definition of membership may seem limited to collectives which make use of computer-mediated communication systems, it can also describe more traditional settings. For example, one might define membership in a hobby group in terms of attendance of a minimal number of meetings. This is the face-to-face equivalent of maintaining exposure and access to the collective's communication infrastructure.

[4] Although this process is described in terms of 'distributing' messages it is implemented by having all member agents process each entry in the message list. Thus, while there is no transfer of data to members, all members are exposed to each of the messages.

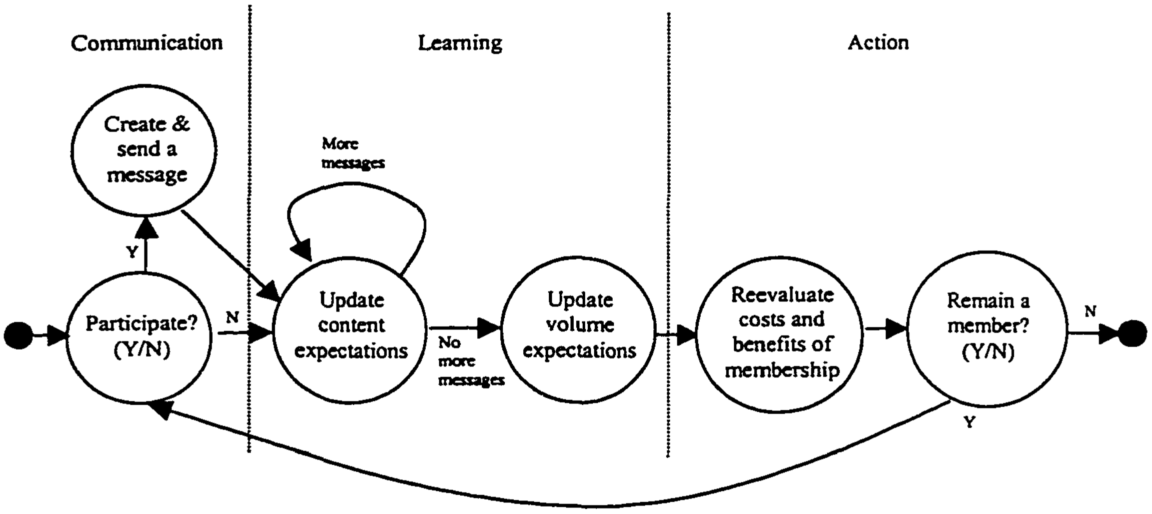
[5] For modeling convenience individuals are restricted to sending at most one message per period.

[6] Future models might also consider the dynamics of collectives in contexts, such as Lotus notes or USENET, which make use of pull technologies. In those cases, maintaining membership is not a passive activity; exposure to communication activity occurs only as the result of repeated active decisions by the individual.

distribution, and membership adjustment (Figure 1) forms a cycle which iterates once per time period.

Collectives make use of various communication infrastructures. Traditional social collectives rely on face-to-face meetings, tightly knit networks of interpersonal contacts, or paper-based print media. Audio, video, and real-time text conferencing, e-mail distribution lists, and groupware (e.g. Lotus Notes and USENET) are all technologies which can support voluntary social collectives. The technologies used within a collective determine the communication costs incurred by its members. Different technologies allow for different forms of communication, resulting in different costs and structures. A collective's *noise cost* (c), *message impact* (w), and *message threshold* (m) reflect the features of the technological infrastructure. Noise cost (c) is the cost incurred by a member as a result of processing a message that is outside the individual's interests (i.e. noise). The value of this parameter is modeled relative to normalized signal benefit, a value that represents the maximum net benefit the individual receives from processing an interesting message. In this implementation, the normalized signal benefit is fixed at 1, and noise cost is set in terms of the percentage of this value (e.g.. noise cost = X implies that the costs incurred processing noise are X% of the normalized signal benefit). Message impact (w), or the weight given to single units of communication as individuals learn about the group, is set as [0,1] value. Message threshold (m) is the number of messages of interest a member can receive in a day before the benefits of those messages are outweighed by the costs of processing them. Within different infrastructures, the features of communication vary, and as result different collectives have different cost and impact structures.

Although they are identifiable social entities, voluntary collectives also have emergent features. Communication is not the result of a unified activity, but rather the action of individual members. Membership is maintained not by a coordinated process, but by the choices of independent individuals. People involved with voluntary collectives contribute messages, process other messages, and, based on their perceptions and expectations, decide whether to continue or maintain their membership (Figure 2). This three stage cyclic model of communication, learning, and action is similar to those proposed by Turner (1988) and Carley (1991) as underlying the development of emergent social structures.



CE_{it}: Expectations at time t regarding the content of group messages
VE_{it}: Expectations at time t regarding the daily volume of group communication activity

pp_i: Participation probability
$[INT_{iL}, INT_{iH}]$: The range of an agent's content interests

Figure 2: Individual Structure

Individuals participate in a collective's communication by constructing messages which are then sent to the other members. Following prior work on participation in social groups (Skvoritz, 1988) and information sharing in decision-making teams (Wittenbaum and Stasser, 1996), each individual's message contribution behavior is modeled as the result of an independent stochastic process. An individual chooses to contribute a single message in a time period with a given probability (pp_i), which may vary between individuals but does not change over time. Upon choosing to participate, an individual creates a message, selecting a topic from his interests. The message is then passed to the collective for distribution to the members.

In this implementation of the model, an individual's interests are represented here as a range of values ($[INT_L, INT_H]$) which define an arc within a collective's circular [0,1] topic space. A message is constructed by selecting a random value from a uniform distribution over this range.

An individual's relationship with a social collective develops over time (Moreland and Levine, 1982). As they are exposed to a collective's activities, members' expectations about the focus (CE_{it}) and volume of activity (VE_{it}) change. Content expectations are implemented here as a matrix of [0,1] values in which CE_{it} is individual *i*'s expectation in time period *t* regarding the probability that future messages in the collective will be of interest to him. Volume expectations (VE_{it}) are positive values representing individual *i*'s expectations at time *t* about a collective's future daily message volume.

In addition to their primary function of supporting the collective's activities, the messages also serve a secondary purpose of providing information about the collective's focus. As a member of a collective, an individual is exposed to messages. Based on this exposure, individuals then change their expectations about the collective's focus and level of activity. This change in beliefs about the collective is implemented as a reinforcement process (Hunter, Danes, and Cohen, 1984), where the change in an individual's content expectations (ΔCE_{it}) is determined by:

$$\Delta CE_{it} = rw[CE_{i(t-1)}][\, 1 - CE_{i(t-1)}]$$

where *w* is the message impact parameter, which is inherited from the collective, and r denotes the individual's reaction to the message; $r = 1$ if the message is of interest, and $r = -1$ otherwise. A message is deemed interesting when its topic (MsgTopic) falls within the arc defined by the individual's interest parameters (INT_{iL} and INT_{iH}). When multiple messages are distributed in a time period, the change in content expectation is computed separately for each message. After all messages have been received, volume expectations for the time period (VE_{it}) are set to the observed mean message volume for all previous time periods.

Individuals' expectations about the future activities of a collective underlie their assessment of the rewardingness of membership (Moreland and Levine, 1982). After adjusting their beliefs in response to being exposed to a set of messages, individuals assess the expected costs and benefits of continued membership. This assessment takes into account the expected benefit from messages that are of interest to the individual, a benefit that is subject to limits. Individuals have strict limits on the time available to them. As more time and attention is 'spent' processing messages, the remaining time is more valuable. Thus as the amount of interesting materials increases, the incremental net benefit of an additional interesting message in the same time period is lower. The assessment also considers the costs incurred as a result of processing uninteresting, or noise, messages. Expectations of the content and volume of activity (CE_{it}, VE_{it}) and the known costs of processing messages within the collective's communication infrastructure (m,c) are combined to assess expected costs and benefits of continued membership. This assessment is implemented with the following formula:

$$E_{it} = E(CE_{it}, VE_{it}; c, m) = (-1/2m)(CE_{it}VE_{it})^2 + (CE_{it}VE_{it}) - c((1 - CE_{it})VE_{it})$$

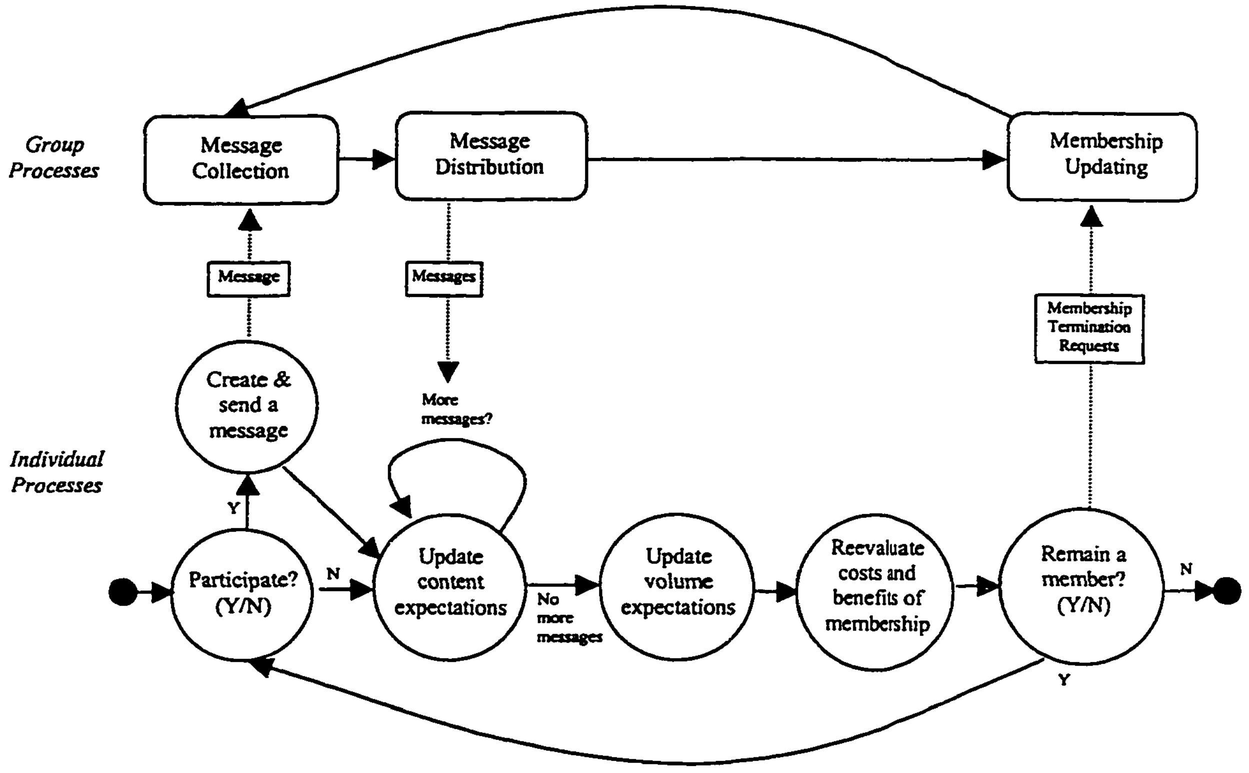
The first two terms $[(-1/2m)(CE_{it}VE_{it})^2 + (CE_{it}VE_{it})]$ indicate the total expected net benefit due to messages which are expected to be of interest. The final term $[- c((1 - CE_{it})VE_{it})]$ is the expected costs due to noise messages.

Individuals' assessment of the costs and benefits determines their willingness to maintain membership in a collective (Moreland and Levine, 1982). If the expected net benefit is positive, the individual will choose to remain a member; otherwise she will choose to terminate membership and leave the group. Membership termination is modeled as the individual choosing to send a signal to the collective, requesting removal from the membership list.

Within a social system individuals and collectives interact, and from those interactions arise the emergent features of collectives. Individuals pass messages to a collective, the collective then passes those messages on to its members (including the sender). After responding, members then may request that the collective remove them from the membership list, an action which results in the individual not being exposed to future messages.

The individual (Figure 2) and collective (Figure 1) processes are different components of a social system containing a single collective with an initial membership of N (Figure 3). This system can be described in terms of a communication infrastructure, represented by the noise cost (c), message impact (w), and message threshold (m) parameters; a participation structure, represented by the distribution of participation probabilities (pp_i) in the population of individuals; and an interest structure, represented by the population distribution of individual interests[7].

[7] In the current implementation, a uniform interest structure is assumed. A maximum interest range is specified. For each individual a interest range length is selected randomly from a uniform distribution bounded by 0 and the specified maximum. A interests 'base point' is also selected, from a [0,1] uniform distribution. The range and base point are then used to specify the individual's interest range $[INT_{iL} = \text{BasePoint}, INT_{iH} = \text{BasePoint} + \text{Range}]$



N: Initial group size (i.e. number of individual agents)
c,w,m : Group agent communication parameters
ppDistribution: The distribution of participation probabilities among the individual agents
IntDistribution: The distribution of interests among the population of individual agents

Figure 3: System Structure

The current implementation of this composite system is lockstepped. All individuals perform an action before anyone moves to the next step in the process. Within a given step, the individual agents are executed in a fixed sequence. After the individuals have completed a step, the collective performs any relevant activities. All individuals make a choice about contributing messages, create messages, and send them to the collective before the messages are distributed to the membership. Likewise, all individuals decide about continuing membership before the collective updates the membership list and starts accepting new messages.

Development of Voluntary Social Collectives

As members are exposed to communication within a collective, their beliefs about the collective change. Developing beliefs alter individuals' assessments of the benefits of being part of the collective, changing their level of commitment and possibly causing them to end their membership (Moreland and Levine, 1982). Member movement changes the composition and size of a collective's membership, which is reflected in the development of its aggregate interests and the focus of its communication activity. These changes, in turn, affect the development of remaining member's beliefs.

The simultaneous development of member commitment and the structural features of a collective's membership is a result of the cycle of individual and structural change processes. As the following example illustrates, these development processes can be seen in the proposed model of voluntary social collectives[8]. The mean commitment level among members converges to a stable value (Figure 4).

[8] The example was the result of a single run of the model (as coded in Appendix A) with the following parameters: N = 75 ; PartProb = 0.0749 ; w = 0.05 ; c = 0.1 ; m = 5 ; INTRange = 0.6132. These results were chosen as representative after running a virtual experiment which systematically varied w,c,and m and randomly varied N, PartProb, and INTRange.

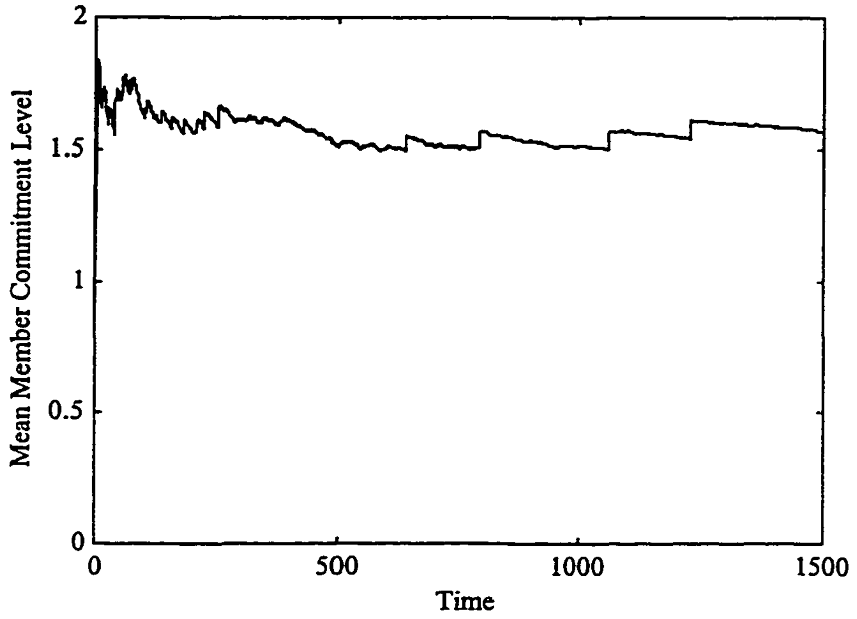


Figure 4: Mean Member Commitment

Changes in the mean level of member commitment with a voluntary social collective arise from two processes. The smooth progression of the mean toward a stable point is the result of core members commitment converging on a common value, while the abrupt shifts are the result of peripheral members terminating their membership.

The trajectories of individual's commitment development illustrate how these two processes play out within a social collective (Figure 5).

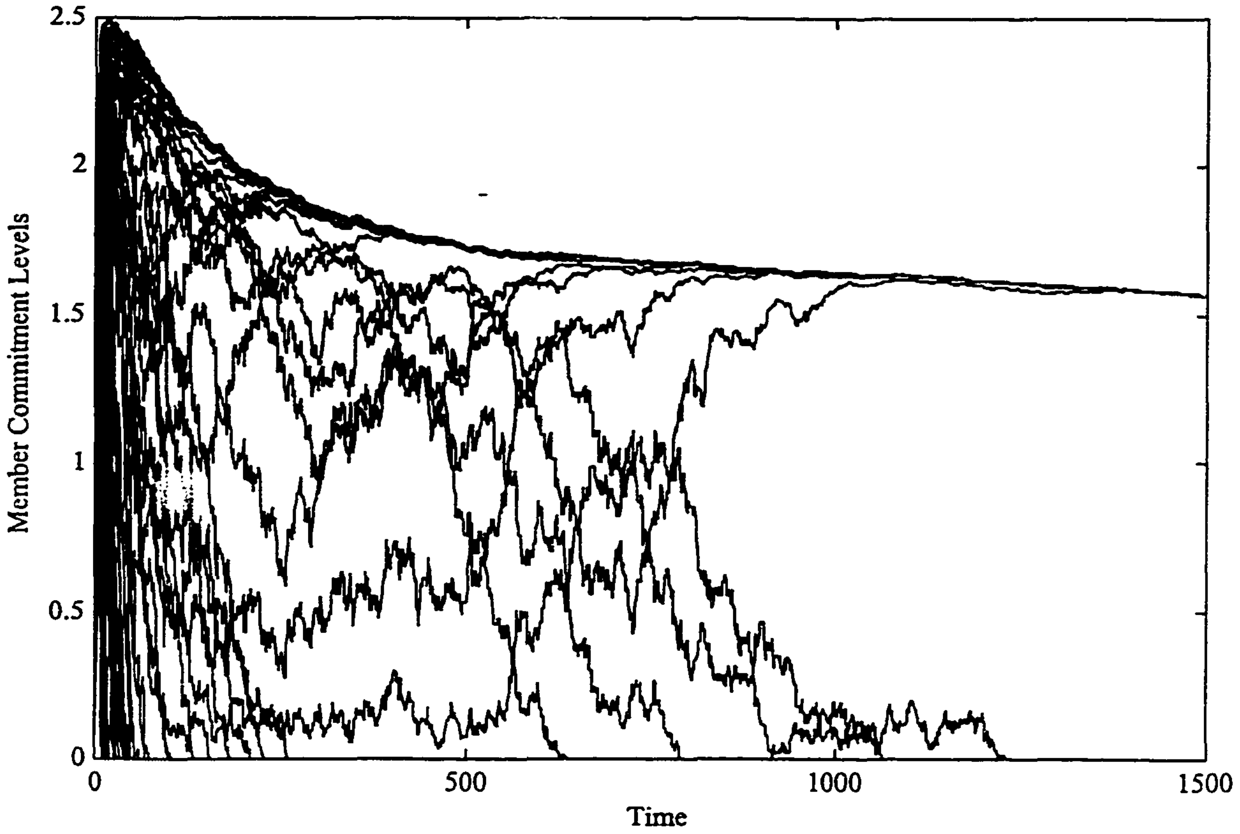


Figure 5: Individual Commitment Development Trajectories

Jumps in the mean level of member commitment occur as individuals who have lower commitment choose to end their membership. These abrupt shifts occur in tandem with the convergence of remaining members' commitment to a common level, as a result of their development of strong beliefs about the content of future communication activity.

The combination of individual and structural change also results in several emergent phenomena that are consistent with prior conceptual models of social collective development. For example, the minimum level of member commitment, another indication of a collective's development, is expected to suddenly shift (Figure 6).

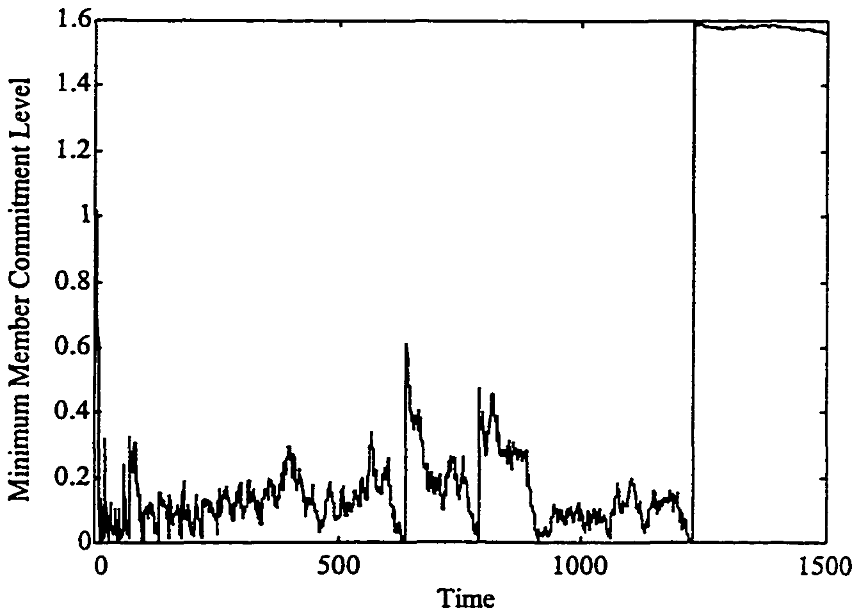
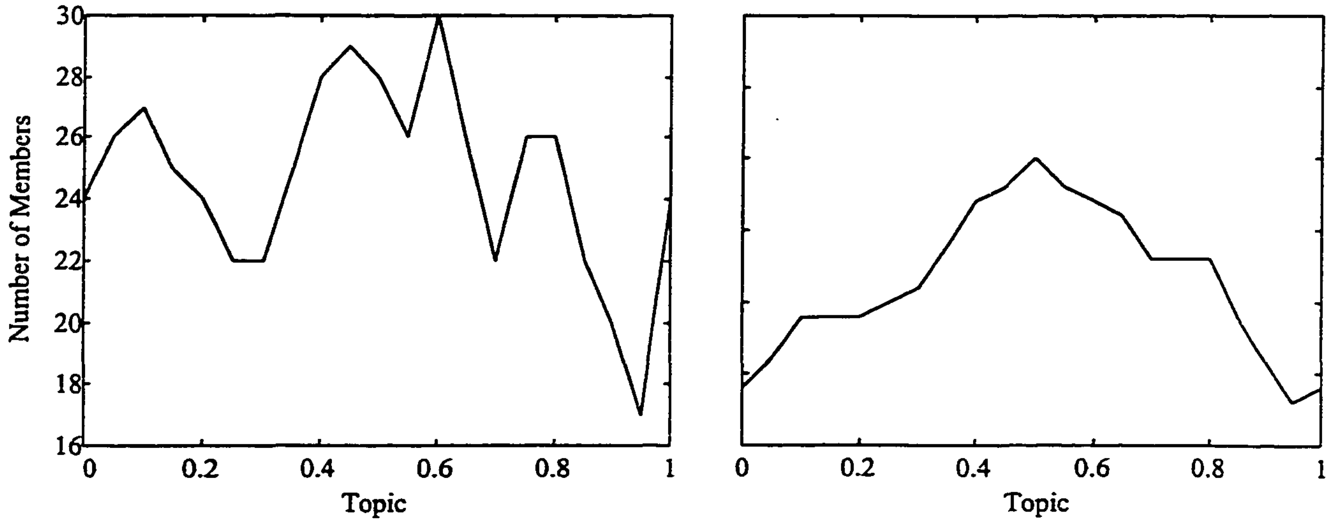


Figure 6: Minimum Commitment Level with the Collective

This shift in the minimum level of commitment may underlie significant changes in the operation of the collective, not unlike the development phases proposed by Tuckman (1965) or the transitions observed by Gersick (1988, 1989).

Another feature of social structure development seen in this example is the emergence of an interest focus. Although it is common to refer to a social collective's 'interests', especially in discussions of online collectives (Baym, 1993; Kollock and Smith 1997), even in formally managed settings, collective communication activity is actually the aggregate result of members' individual choices. The true focus or interests of a collective are not a monolithic construct, but an emergent one which arises from the actions of its members. As a collective's membership develops, the distribution of interests among the members changes (Figure 7).



Time (t) = 0

Time (t) = 150

Figure 7: Development of a Collective's Interest Distribution

In the extreme, a collective's membership becomes stable. When stability is reached the distribution of interests[9] is roughly normally distributed (Figure 7). Thus, although there is no formally defined interest specification, a collective still develops an interest focus as a result of member and structural change processes.

Model Calibration

Calibration is the process of adjusting a computational model to reflect the features of empirical data. This process provides a conceptual reference point for validation and analysis of the model. Calibration of the voluntary collectives model is based on empirical data from a sample of e-mail based Internet[10] listservs. These collectives utilize Internet-based electronic mail and a centralized mailing list to enable individuals to broadcast text-messages to other members. Although there may be an individual who is responsible for maintaining the technical infrastructure (i.e. the listowner) the sampled collectives are unmanaged, voluntary collectives.

[9] A modeled collective's interest distribution is characterized by determining the number of members with interests at twenty equally spaced points within the collective's [0,1] topic space.

Individuals are free to enter, leave, or send messages as they please. Listowners take no formal steps to restrict membership or message content.

From a population of approximately 70,000, an initial set of 1066 listservs was created. This initial sample was stratified, to ensure that it spanned a range of topic and member communities. One third focused on work-related topics. One third focused on personal topic (hobbies, lifestyles, etc.). The remaining collectives were associated with topics that mixed work-related and personal interests (e.g geographic locations).

The initial sample was filtered through a multiple stage confirmation process that screened out managed collectives, i.e. moderated listservs and those with formal new member screening. This selection process also verified that each listserv was mechanically functional, able to provide the needed data, and available for inclusion in the study (for more details on sample selection process see Butler, 1999a). The result of this process was a sample of 217 listservs. 192 of these collectives provided data that could be used to construct measures of membership change and communication activity.

For a 130 day period, between July 23, 1997 and November 30, 1997, communication and membership data was collected for each listserv. The communication data consisted of all messages, which were aggregated to create collective-specific archives of all communication activity that occurred during the observation period. During the data collection period a copy of each collective's membership list was requested daily. These lists were archived to create a record of the changes in collective membership during the observation period (for more details on data collection procedures see Butler, 1999a and Butler, 1999b).

[10] The relevance of the Internet here is that it is a public network environment. This is to be contrasted with an organizational network in which participation in a collective is limited to a relatively small number of organization members.

The message and membership archives characterize the structural features of the listservs. Measures of these structural features serve as the basis for calibrating the model. The empirical data was used to construct measures of daily communication volume and percentage membership loss. Daily communication volume was calculated by determining the total number of messages distributed to the members during the observation period and dividing that by the number of days (i.e. 130). Percentage membership loss was determined by counting the number of people who left the listserv during the observation period and normalizing it by the number of members present on the first day of the observation period (i.e. the collective's initial size)[11]. These measures were used to assess two aspects of the sampled online collectives: the distribution of communication volumes and membership loss among the collectives. The relationship between communication activity and membership loss was also assessed (see Butler, 1999b for more analysis of the relationship between structural features of online social collectives).

From the full sample of listservs[12], 100 were randomly selected to be the calibration sample and the remaining 92 listservs were used as the validation sample. Calibration was performed with a series of sessions. Each session involved simulating 100 collectives and comparing the resulting distributions and relationships with data from a randomly selected subset of 100 listservs from the empirical sample. In each calibration run, one or more of the model settings were modified to better reflect the features of the observed collectives. Also, to better represent the state of the empirically observed collectives, all of which were known to be at least four months old, each calibration run included two phases. The first phase, a 100 time period

[11] Membership loss could not be calculated by simple subtraction (intial size – final size) because these collectives also had members entering during the observation period, a process that is not considered in the current computational model.

[12] In several cases there were major problems creating the structural measures due to the presence of membership management activities (i.e. manipulation of the membership by the listowner) or the occasional presence of non-standard message formats. The cases were dropped, leaving a total sample of 192 listservs.

initialization phase, was run to represent a collective's prior history. The second phase, with 130 time periods, was then run and the results compared with data from the online collectives.

The initial calibration run was performed with the following model settings:

Model Setting	Value
N	100
c	0.33
w	0.05
m	15
ppDistribution	Fixed @ 0.005
Interest range distributions	Uniformly distributed between 0 and MaximumInterestRange, a value which is chosen for each group from a uniform distribution between 0.25 and 0.75
Initial CE(0) distribution	Uniformly distributed between 0.5 and 1
Initial VE(0) distribution	Fixed @ 1

Table 1: Model Settings for Calibration Run #1

The distribution of initial content expectations (CE(0) distribution) was set based on the premise that voluntary members would, at least early on, have positive expectations regarding the collective's communication content. Individuals are therefore assumed to initially expect at least a majority of the communication activity to be of interest (i.e. CE(0) would be distributed between 0.5 and 1 for the initial members). Similarly, it is assumed that individuals would not voluntarily join which they expect to have no communication activity. Consequently, initial volume expectations (VE(0)) were set at 1 to indicate a uniform initial expectation of one message per day.

As described above (and in Butler, 1999a), the sample of e-mail based Internet listservs was selected to ensure that the sampled collectives varied in terms of the member populations that they drew from. This variation is most clearly reflected in the range of topics represented in the sample. However, it is also likely to result in interest ranges varying within the collectives, with some being attractive to individuals with narrow, well-defined interests and others appealing to individuals with a wider range of interests. This diversity within the sample is reflected in the setting of the interest range distribution. The largest interest range within a collective is randomly selected from a uniform distribution between 0.25 and 0.75. Then within each run the individual's interests ranges are set by selecting one end point from the topic space (uniform distribution) and selecting an interest range size from a uniform distribution between 0 and the modeled collective's maximum interest range. These settings represent variety in interest structures at both the individual and population level.

Overall, the empirical results from the online collective are not well represented by this instantiation of the model (Figure 8 & Table 2).

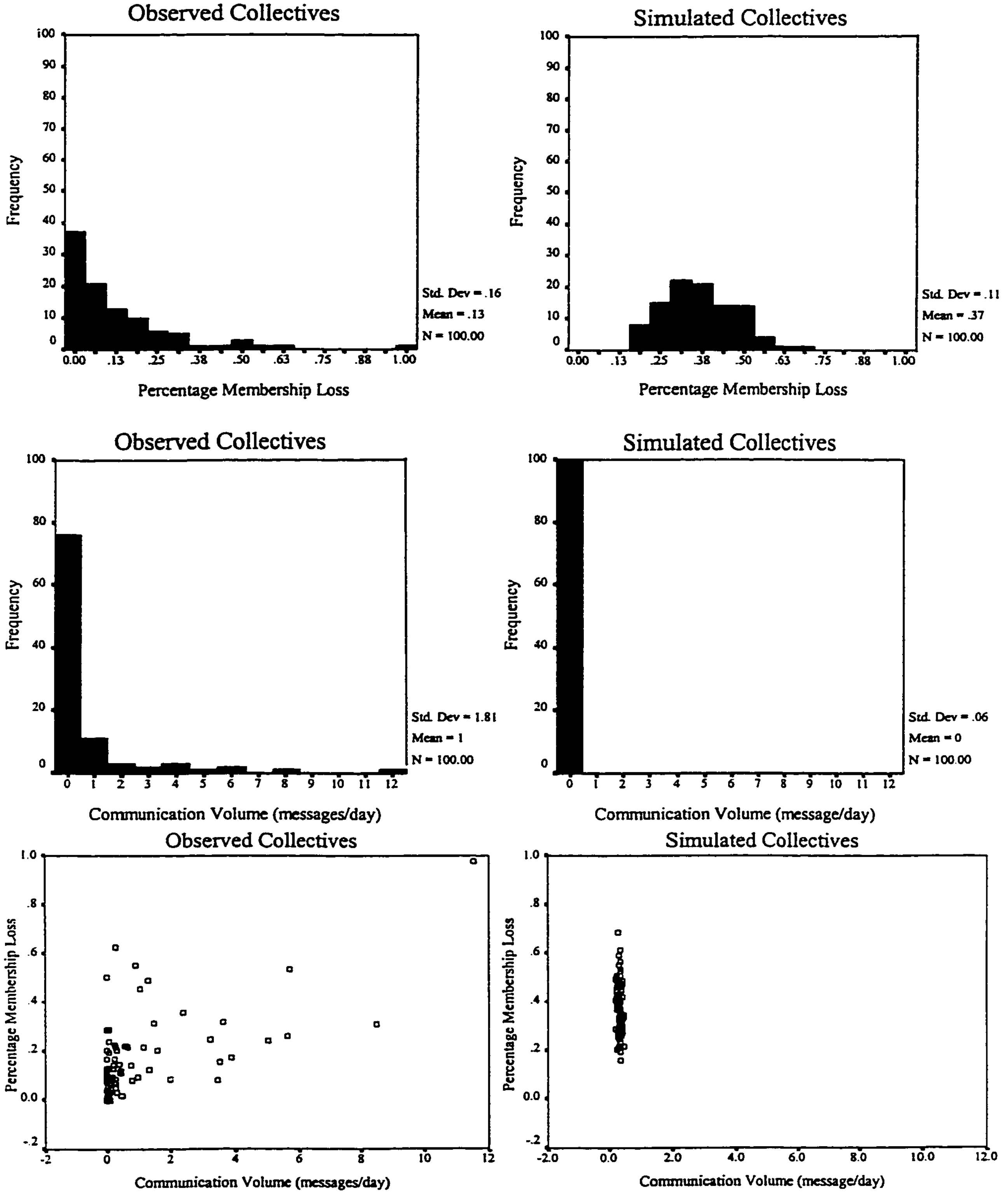


Figure 8: Results of Calibration Run #1

		Observed	*Collectives*	*Simulated Collectives*	
		Percentage Membership Loss	Communication Volume (messages/day)	Percentage Membership Loss	Communication Volume (message/day)
Mean		.125	.770	.368	.330
Median		.078	.062	.366	.328
Mode		.000	.000	.250[a]	.328
Std. Deviation		.164	1.806	.105	.059
Skewness		2.342	3.679	.360	.011
Std. Error of Skewness		.241	.241	.241	.241
Kurtosis		7.533	15.808	-.230	-.249
Std. Error of Kurtosis		.478	.478	.478	.478
Percentiles	25	.000	.000	.287	.292
	50	.078	.062	.366	.328
	75	.198	.456	.455	.366

a Multiple modes exist. The smallest value is shown

N = 100

Table 2: Distribution Statistics for Calibration Run # 1

	Observed	Simulated
Pearson's Correlation	0.629*	-0.203*
Spearman's Rho	0.688*	-0.198*

N = 100

Table 3: Relationship between Communication Activity and Membership Loss

Unlike the observed data, in the computational collectives both percentage membership loss and communication volume are normally distributed (Figure 8). In addition, the predicted relationship between these measures is negative, not positive as in the empirical data (Table 3).

Calibration Run #2: Size Distribution

One of the most unrealistic aspects of the model settings used in the first calibration run is the use of a single value for membership size among the computational collectives (N = 100). In the empirical sample the initial collective sizes vary according to a log-normal distribution (Figure 9).

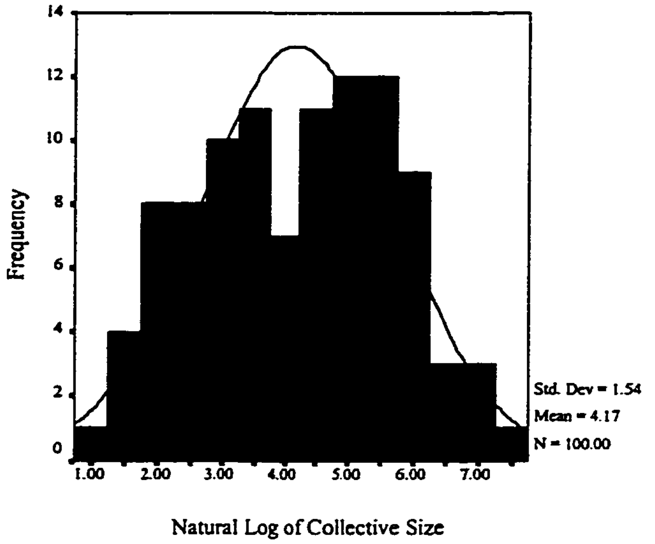


Figure 9: Estimation of Collective Size Distribution Parameters

To more accurately reflect this aspect of social collectives, the second calibration run was performed with initial membership sizes (N) drawn from a log-normal distribution with a mean of 4.17 and standard deviation of 1.54. All other model settings were unchanged (see Table 1).

Although the result of this instantiation of the computational model are a closer match with certain aspects of the empirical data, there remain several significant differences (Figure 10, Table 4, and Table 5).

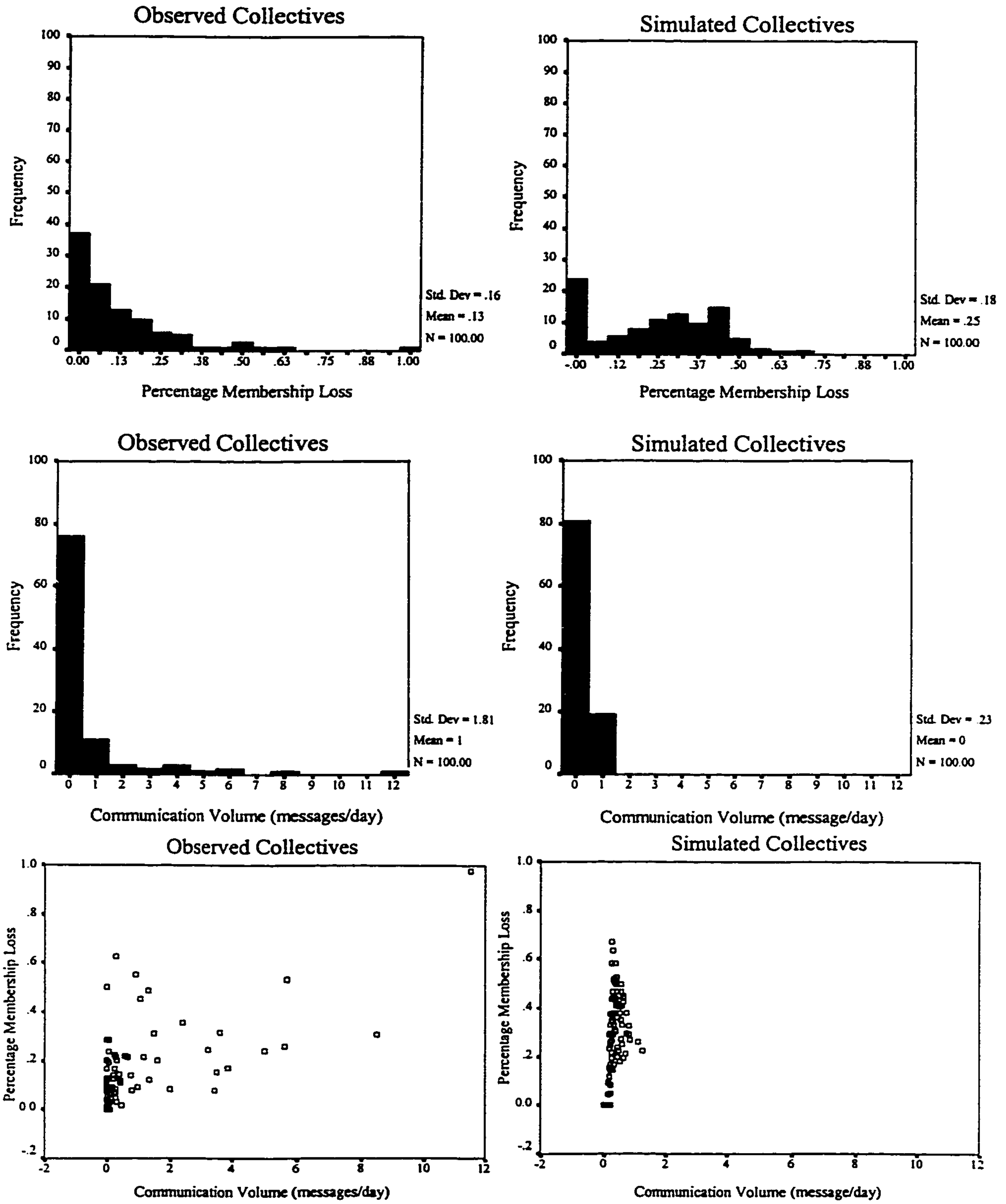


Figure 10: Results of Calibration Run #2

		Observed	*Collectives*	*Simulated Collectives*	
		Percentage Membership Loss	Communication Volume (messages/day)	Percentage Membership Loss	Communication Volume (message/day)
Mean		.125	.770	.246	.316
Median		.078	.062	.262	.267
Mode		.000	.000	.000	.183
Std. Deviation		.164	1.806	.184	.230
Skewness		2.342	3.679	.061	1.179
Std. Error of Skewness		.241	.241	.241	.241
Kurtosis		7.533	15.808	-1.021	2.051
Std. Error of Kurtosis		.478	.478	.478	.478
Percentiles	25	.000	.000	.043	.141
	50	.078	.062	.262	.267
	75	.198	.456	.399	.433

a Multiple modes exist. The smallest value is shown

N = 100

Table 4: Distribution Statistics for Calibration Run # 2

	Observed	Simulated
Pearson's Correlation	0.629*	0.519*
Spearman's Rho	0.688*	0.682*

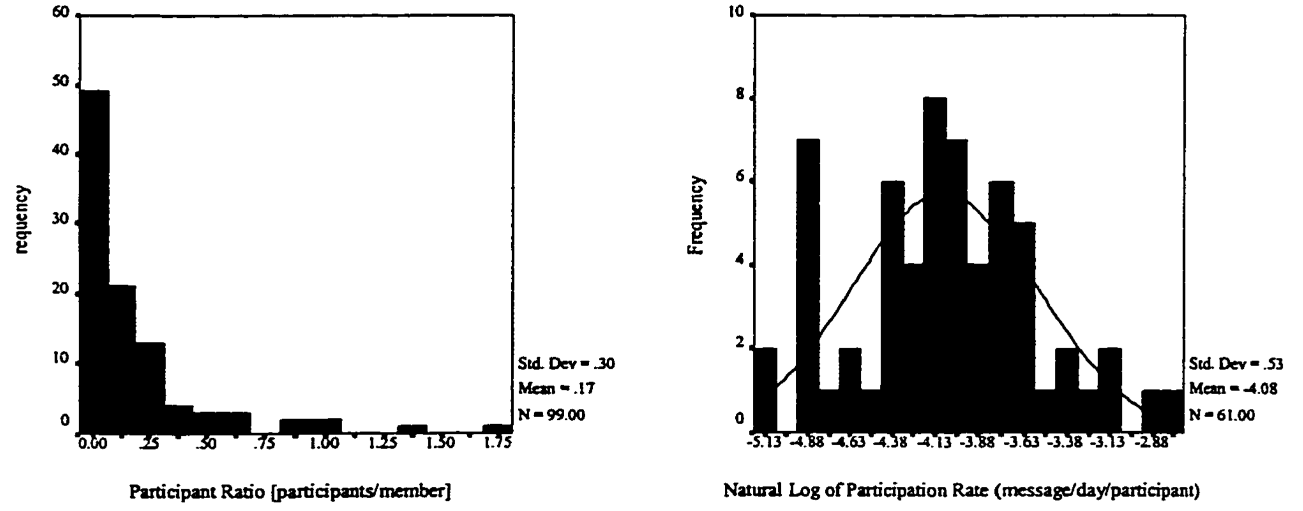
N = 100

Table 5: Relationship between Communication Activity and Membership Loss

Use of an empirically derived distribution of initial collective sizes resulted in a predicted relationship between percentage membership loss and communication that is similar to the observed relationship (Table 5). However, the membership loss and communication volume distributions in the population of computational collectives are structurally different that the results seen in the empirical data (Figure 10).

Calibration Run #3: Participation Distribution

The settings used in the first two calibration runs assume that participation is uniformly distributed among all members of the collective. However, in the observed listservs, participation is concentrated among a subset of the membership (Figure 11).



(a) (b)

Figure 11: Empirically Observed Distribution of Participation Features

In the third calibration run, the participation structure for a collective is constructed in two steps. In the first step, a participation ratio is selected from an exponential distribution with a mean of 0.17[13] (Figure 11a). The participation ratio, which is the proportion of individuals who will contribute any messages to the group, is used to probabilistically label individual agents as non-participants (participation probability (pp) = 0) or participants (pp > 0). The individual agent's participation probability is then set as a fixed value which is drawn from a log-normal distribution with a mean of –4.08 and standard deviation of 0.53 (Figure 11b). Otherwise, the model parameters were identical to those used in calibration run 2.

The results of the third calibration run (Figure 12, Table 6, and Table 7) indicate that this instantiation of the computational model is incrementally more similar to the empirical data.

[13] As Figure 11clearly indicates, there is an outlier in the distribution of participation ratios (11.57). This collective was not included in the calculation of the mean participation ratio used to calibrate the model.

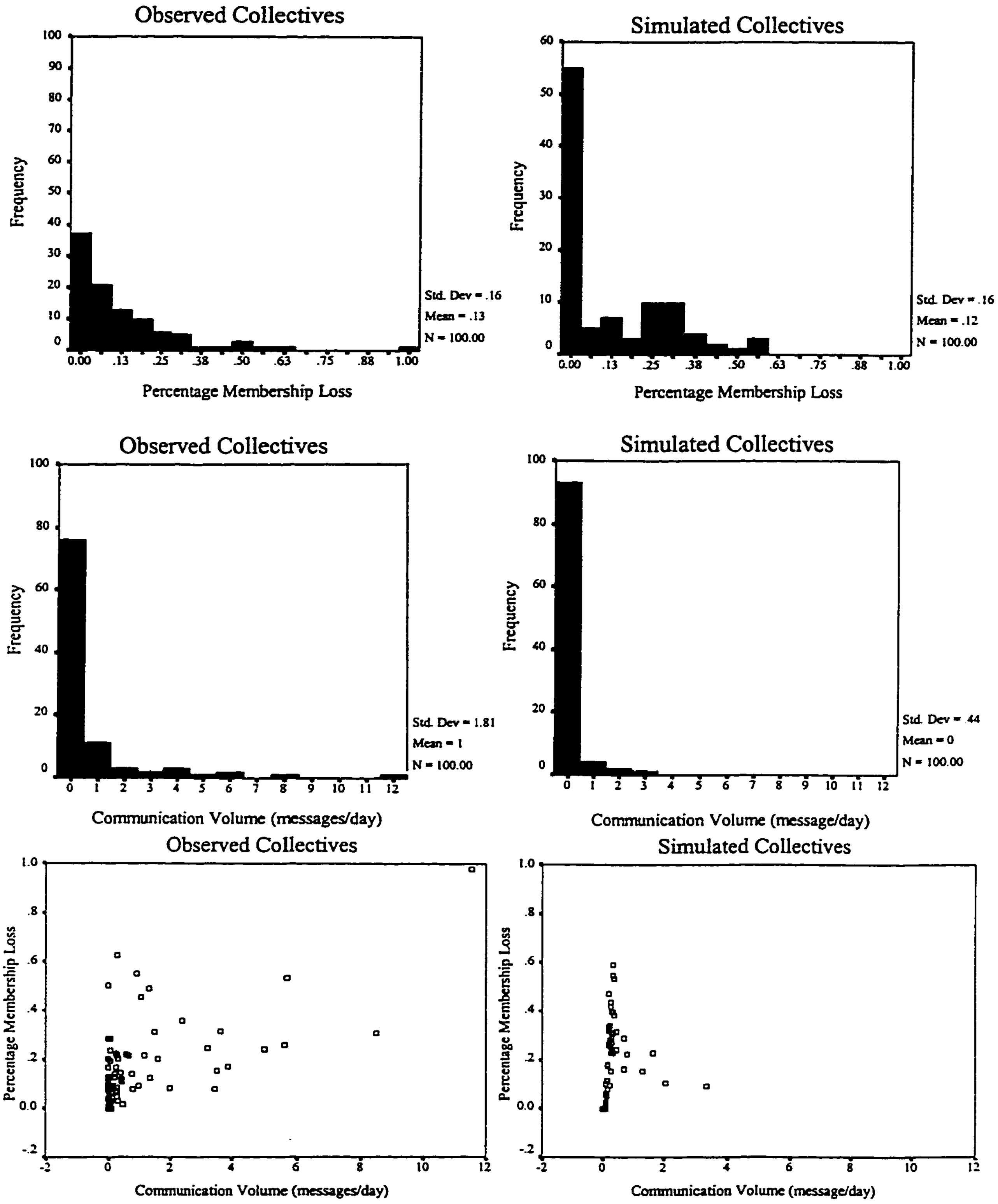


Figure 12: Results of Calibration Run #3

		Observed Collectives		*Simulated Collectives*	
		Percentage Membership Loss	Communication Volume (messages/day)	Percentage Membership Loss	Communication Volume (message/day)
Mean		.125	.770	.119	.226
Median	-	.078	.062	.000	.099
Mode		.000	.000	.000	.000
Std. Deviation		.164	1.806	.160	.444
Skewness		2.342	3.679	1.115	4.735
Std. Error of Skewness		.241	.241	.241	.241
Kurtosis		7.533	15.808	.116	27.177
Std. Error of Kurtosis		.478	.478	.478	.478
Percentiles	25	.000	.000	.000	.015
	50	.078	.062	.000	.099
	75	.198	.456	.250	.275

a Multiple modes exist. The smallest value is shown

N = 100

Table 6: Distribution Statistics for Calibration Run # 3

	Observed	Simulated
Pearson's Correlation	0.629*	0.290*
Spearman's Rho	0.688*	0.851**

N = 100

Table 7: Relationship between Communication Activity and Membership Loss

Although the relationship between the volume of communication activity and membership loss was slightly affected it remains positive (Table 7). The distribution of communication volumes was slightly more right-skewed, while the percentage member loss distribution shifted to the left (Figure 12).

Calibration Run #4: Communication parameters

Unlike initial size and participation structure, the communication parameters (noise cost, message threshold, and message impact) are unobservable. In the initial model calibration runs, these settings were chosen arbitrarily. However, in the fourth, and final, stage of model calibration these model settings were varied. Values incrementally higher and lower then the settings used in prior calibration runs (see Table 1) were used to select the parameter values which resulted in distributions of percent member loss and daily volume which most accurately reflect the empirical observations. Model threshold values of 2, 5, and 8 were used to represent a range of message thresholds. These values were chosen because cover a low to high level of activity relative to the levels of message activity seen in the listservs (Mean daily activity ≈ 0.2 messages/day; Maximum mean ≈ 3 messages/day). Likewise, the noise cost is varied to reflect situations in which the cost of noise is low (0.1) and those in which the cost of unwanted messages is directly comparable to the benefit received from messages of interest. Finally, the message impact parameter is set at 0.01, 0.05, and 0.09 to model situations in which messages are seen as providing weak, medium, or strong indications of the content of future communication activity.

Each of the 27 (3x3x3) communication parameters combinations were used for 100 groups. Each condition was then compared, both with one another and with the empirical data to characterize the impact of the communication parameters, and select the condition which best fit the data.

						W				
			0.01	0.05	0.09	0.01	0.05	0.09	0.01	0.05	0.09
C	0.10	Mean	0.03	0.12	0.12	0.02	0.12	0.16	0.05	0.11	0.16
		Median	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
	0.33	Mean	0.05	0.12	0.19	0.06	0.20	0.17	0.05	0.12	0.19
		Median	0.00	0.00	0.19	0.00	0.21	0.12	0.00	0.00	0.18
	1.00	Mean	0.08	0.18	0.18	0.08	0.17	0.21	0.08	0.20	0.19
		Median	0.03	0.18	0.16	0.05	0.15	0.20	0.03	0.16	0.20
				M = 2.00			M = 5.00			M = 8.00	

Table 8: Mean and Median Values for the Distribution of Percentage Membership Loss in Computational Collectives

By comparing the percentage membership loss distribution location measures (mean and median) for a variety of model settings (Table 8) with the empirically observed data (mean = 0.125; median = 0.078), the model settings of c = 1.0, m = 5.0, and w between 0.01 and 0.05 were selected. To further refine these settings, a supplementary calibration run was performed with the settings of c = 1.0, m = 5.0, and w = [0.01, 0.02, 0.03, 0.04, 0.05]. All other model settings were identical to those used in calibration run 3. For each of the 5 combinations 100 collectives were simulated.

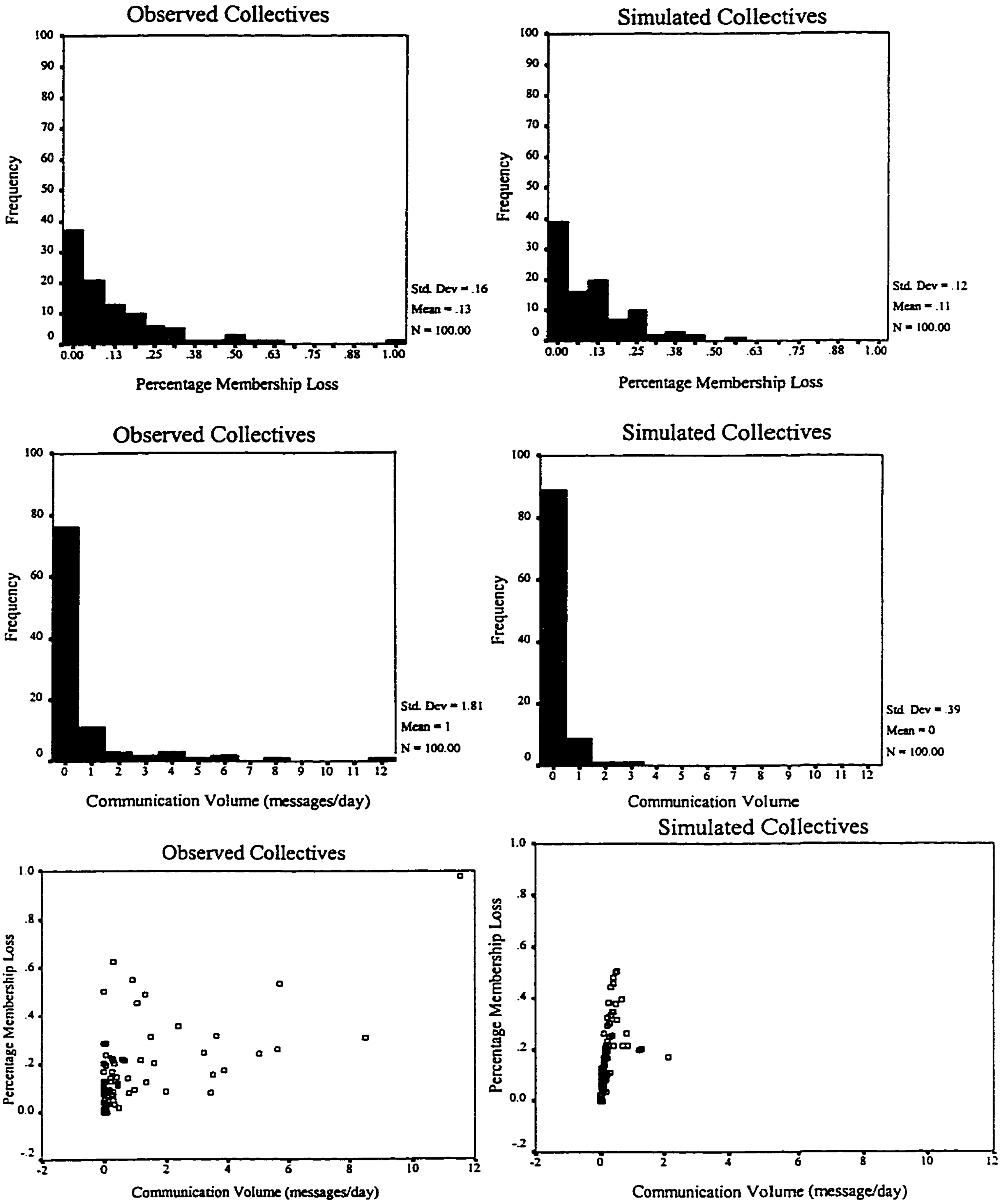
	w				
	0.01	0.02	0.03	0.04	0.05
Mean	0.09	0.11	0.13	0.17	0.17
Median	0.05	0.07	0.09	0.13	0.15

c = 1.0; m = 5.0

Table 9: Percentage Membership Loss Results for Calibration Run 4b

Based on the results of the secondary calibration run (Table 9), the message impact parameter (w) was set at 0.02.

The structure and location of the distribution of percentage membership loss results from the computational collectives is comparable to that observed in the empirical data (Figure 13, Table 10, and Table 11).


Observed Collectives

Simulated Collectives

Observed Collectives

Simulated Collectives

Observed Collectives

Simulated Collectives

Figure 13: Results of Calibration Run #4

		Observed	*Collectives*	*Simulated Collectives*	
		Percentage Membership Loss	Communication Volume (messages/day)	Percentage Membership Loss	Communication Volume (message/day)
Mean		.125	.770	.107	.227
Median		.078	.062	.068	.080
Mode		.000	.000	.000	.000
Std. Deviation		.164	1.806	.124	.394
Skewness		2.342	3.679	1.392	3.705
Std. Error of Skewness		.241	.241	.241	.241
Kurtosis		7.533	15.808	1.761	18.049
Std. Error of Kurtosis		.478	.478	.478	.478
Percentiles	25	.000	.000	.000	.015
	50	.078	.062	.068	.080
	75	.198	.456	.170	.256

a Multiple modes exist. The smallest value is shown

N = 100

Table 10: Distribution Statistics for Calibration Run # 4

	Observed	Simulated
Pearson's Correlation	0.629*	0.506*
Spearman's Rho	0.688*	0.891**

N = 100

Table 11: Relationship between Communication Activity and Membership Loss

On average, the computational model understates the volume of communication activity (Table 10). This discrepancy is a result of changes that occur during the model initialization stage. The relationship between membership movement and communication volume is in positive in both cases. However, overall the distribution of communication volume, membership loss, and the relationship between these two structural aspects of social collectives (Figure 13 and Table 11) is reasonably represented by the resulting instantiation of the computational model (Table 12).

Model Setting	Value
N	Log normally distributed [LN(4.17,1.34)]
c	1.0
w	0.02
m	5
ppDistribution	Ratio of Participants (i.e. pp > 0) to Members is exponential with = 0.17; All participants have the same participation probability (drawn for the group from a log normal distrubition [LN(-4.03,0.53)]
Interest range distributions	Uniformly distributed between 0 and MaximumInterestRange, a value which is chosen for each group from a uniform distribution between 0.25 and 0.75
Initial CE_0 distribution	Uniformly distributed between 0.5 and 1
Initial VE_0 distribution	Fixed @ 1

Table 12: Model Settings for E-mail Based Internet Social Collectives

Calibration used empirical data about structural change and communication activity in e-mail based Internet listservs to identify a baseline set of parameters for the computational model of social collective development. These settings provide a reference point for the subsequent validation and analysis of the model.

Model Validation

Validation is the process of comparing the results of a computational model with empirical data for the purposes of testing the model. Unlike calibration, in which the parameters of the model are adjusted to identify the most appropriate settings, during validation the proposed parameters and processes are taken as given, the model is run, and the outcomes are compared with empirical data. The results of this process provide information about how well, and under what conditions, the computational model accurately represents the intended phenomena.

The computational model of social collective development was validated using structural change data from the 92 listservs that were not used for calibrating the model. The model was validated using matched analysis (Law & Kelton, 1991). The empirically observable model settings, initial collective size (N) and the participation structure (as described by the participation ratio and participation probability), were set based on the data from a single listserv in the validation sample. The communication parameters (noise cost (c), message weight (w), and message threshold (m)) and initial conditions (CE_0 and VE_0) were set based on the results of model calibration. Based on these model settings, 10 model runs were performed. For each run, the remaining unobservable model setting, the variation in the collective's member interest ranges (INTRange), was randomly selected from a uniform distribution between [0.25,0.75]. The mean percentage membership loss and mean daily communication volume for the set of 10 computational collectives were then recorded as the predicted structural outcomes. This process was repeated for each listserv in the validation sample.

Comparison of the predicted and observed outcomes with paired t-tests indicates that, overall, there is not a statistically significant difference between the model's membership loss predictions and the empirical data (H_0: $\mu_{predicted} = \mu_{observed}$; $p = 0.263$). There is a significant difference between the predicated and observed values for communication activity (H_0: $\mu_{predicted} = \mu_{observed}$; $p < 0.01$). As was observed during the calibration process, the model under-predicts the communication activity ($\mu_{predicted} = 0.408$ and $\mu_{observed} = 0.795$).

OLS regression analysis of the membership loss error (Table 13) indicates that the model is most accurate for collectives with low membership loss. The model also tends to overstate the membership loss when applied to larger and more active collectives.

	Dependent Variable			
	Membership Loss Error		Message Volume Error	
	Unstandardized	Standardized	Unstandardized	Standardized
Intercept	0.051***		0.0251	
Initial Size	0.00012**	0.209**	0.0013***	0.242***
Observed Message Volume	0.025***	0.301***	-0.7585***	-1.011***
Observed Membership Loss	-0.818***	-0.882***	-	-
Adjusted R^2	0.572		0.955	

$p < 0.05$: * $p< 0.01$: ** $p < 0.001$ ***

Table 13: OLS Regression Analysis of Computational Model Error

The results of error analysis, along with graphical inspection of the relationship between the error and empirically observed levels of activity, imply that model error is greatest when it is applied to collectives with high levels of communication activity (> 3 messages per day). However, because the distribution of activity levels in online voluntary collectives has been seen to be highly skewed (Butler 1999a) this error should not significantly bias conclusions when considering the majority of these social structures.

The validation and error analysis results indicate that the model provides reasonable predictions for membership loss. Furthermore, they imply that the model is most accurate for collectives in the range of sizes and activity levels that are most common. Thus, the computational model can be seen as providing an accurate representation of the structural development of these social collectives.

The Impact of Communication Cost on Collective Development

One of the most often discussed features of networked communication environments is their ability to reduce the costs of communication (Sproull and Keisler, 1990). By reducing the costs of message transmission, networks can enable messages to be sent that in more costly traditional settings, such as group meetings or print publications, would have gone undistributed. Computer-mediated systems also change the way people process communication, affecting both the incremental costs of receiving a desirable message and the costs incurred as a result of noise.

Networked environments also reduce the importance of many economies of scale, allowing communication to take place in smaller units. Rather than having a two-hour meeting or sending a multi-page newsletter, collective communication can take the form of shorter messages.

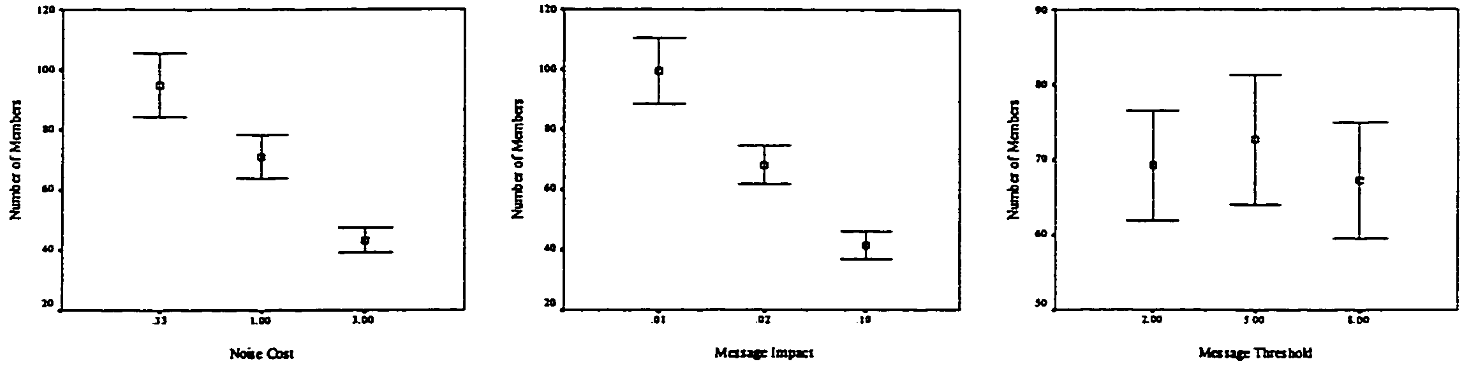
Different infrastructures lead to different communication costs for the members of voluntary social collectives. Different costs affect the development of the collective in several ways. Communication costs and structures affect the development of individuals' perceptions by altering the impact of individual units of communication activity on the individuals' beliefs about the collective. If communication activity is large-grained and expensive, the impact of any given unit on a member's beliefs will be greater. Lower processing costs also affect development by altering individuals' assessments of expected net benefits of continued membership. This, in turn, may alter when, or if, individuals choose to end their membership. Thus, changing communication costs have the potential to affect collective development.

A virtual experiment was performed with the computational model to assess the expected impact of different communication infrastructure on the development of social collectives. The experimental conditions were created by systematically varying the communication parameters, noise cost (c), message weight (w), and message threshold (m). Noise cost was set at 0.33, 1.0, or 3.0. Message weight was set at 0.005, 0.02, or 0.1. Message threshold was set at 2, 5, or 8. The values were chosen to represent a range of communication infrastructures. To anchor the analysis, the experimental settings included the values identified during calibration as most appropriate for e-mail based Internet groups. The remaining model settings and the initial conditions were set as determined during calibration (c.f. Table 12).

For each condition one hundred collectives were modeled. To simulate a year, runs were 365 time periods long, not counting the initialization stage of 100 time periods. However, rather than comparing conditions in terms of time to membership stability, an outcome that takes years for most of the modeled collectives, measures of membership size and stability were considered. Size, measured in terms of the number of members, was considered, because it is an important structural feature of a voluntary social collective. Larger collectives provide members with larger audiences and, potentially, more sources of information and support.

Stability is the likelihood of individuals leaving the collective. In voluntary collectives, members are free to leave whenever they choose. Individuals end their membership when they expect the costs of continued membership to outweigh the benefits. Collectives regularly experience shocks in the form of events, both internal and external, that have the potential to alter the membership of the collective. A collective can be seen as more stable if its membership is less likely to leave in the face of these shocks. Examination of preliminary results indicated that differences in individual members' evaluation of the group were due primarily to differences in their expectations about the content of activity (and not expectations about volume) (Figure 2). The lower a member expectations about the probability of future message usefulness, the lower her assessment of membership benefits would be, and the higher the chances that he would leave the group if a shock occurs. Stability can thus be measured in terms of the minimum content expectations among a collective's members.

Measures of size and stability were recorded for each of the 100 computational collectives in each condition. The model results were then analyzed in a series of ANOVA models.



(a) (b) (c)

[95% confidence intervals for collective size after 365 time periods]

Figure 14: Effects of Communication Features on Collective Size

Lower costs and message weights result in larger collectives. Voluntary social collectives operating in contexts with lower relative noise costs see less membership loss, and as a result they are larger than those in which the relative cost of processing noise is higher (F = 42.514; p < 0.001) (Figure 14a). The effect of relative noise cost on collective size is a consequence of altering the minimum acceptable signal to noise ratio. When faced with lower relative costs of processing noise, individuals are willing to tolerate a higher proportion of unwanted messages. Thus the threshold at which individuals choose to end their membership is lower. This slows down the rate at which members filter out, resulting in larger collectives.

Collectives in which the impact of individual messages on member beliefs is lower are also larger (F = 54.220; p < 0.001) (Figure 14b). Lower message impacts reduce the rate at which individuals' beliefs about a collective change. This, in turn, slows down the rate at which members either leave the collective or become fully committed. In addition, there is a significant interaction between noise cost and message impact (Figure 15). In contexts with lower noise costs, the impact of decreasing message weights on collective size is greater.

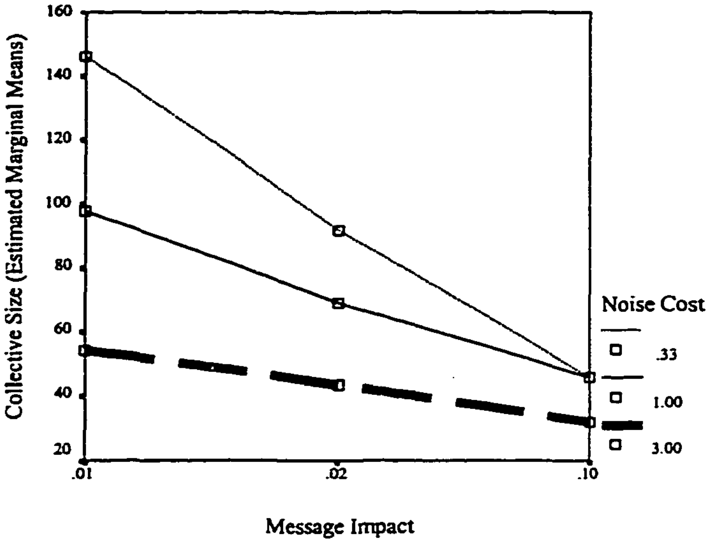
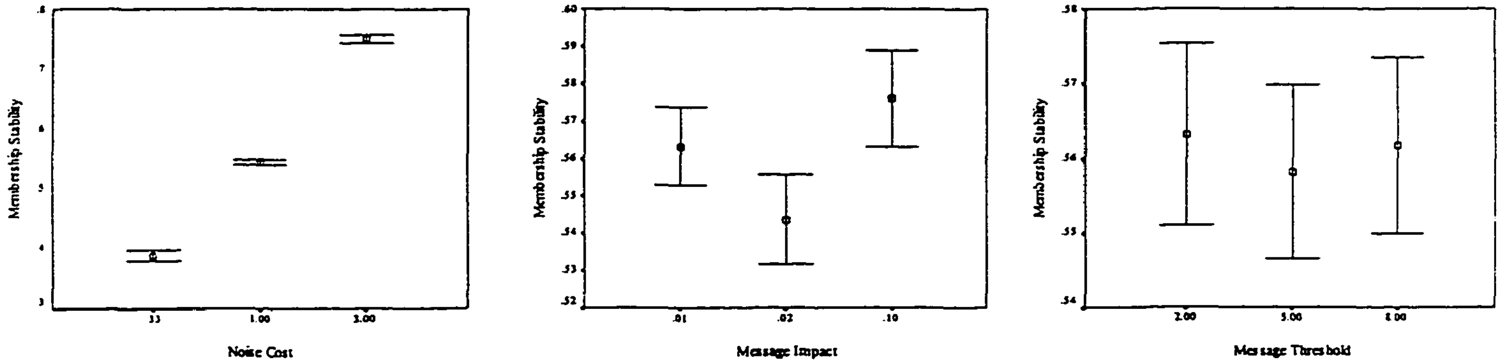


Figure 15: Interaction of Message Impact and Noise Cost

However, the model does not predict that message threshold will have a significant impact of a collective's size. (F = 0.471; p = 0.625) (Figure 14c).

Lower costs and impacts also result in less stability. Voluntary collectives operating in infrastructures with lower noise costs are less stable (F = 2761.603; p < 0.001) (Figure 16a).



(a) (b) (c)

[95% confidence intervals for collective stability after 365 time periods]

Figure 16: Effects of Communication Features on Collective Stability

As with size, message threshold does not have a significant effect on collective stability (F = 0.181; p = 0.835) (Figure 16c). However, the effects of message weight on collective stability are more complex (Figure 16b). Based on the trajectory of their belief development, individuals can be classified in one of three categories: leavers, committed core, and peripheral members. Leavers, who are not actually interested in the collective, are characterized by a strictly declining content evaluation trajectory. Given enough time, these individuals all leave the collective. The committed core consists of those members who have consistently high and increasing evaluations of the communication content. Their content evaluation trajectories are strictly increasing. Peripheral members are characterized by evaluation trajectories that are first decreasing, then increasing. Early on, peripheral members are only interested in a subset of the collective's activity. As a result, their content evaluation decreases over time. However, as the collective's topic focus develops, peripheral members' evaluations improve. Given enough time, peripheral members' evaluations increase to the level of that of the committed core.

Early in the development of a collective, the members all have relatively high expectations. As time passes, leavers and peripheral members lower their expectations of the content and committed core members increase theirs. As the leavers work their way out, their commitment drops and with it the overall stability of a collective. Thus, early on leavers undermine a collective's stability. Furthermore, after these individuals leave, the peripheral members keep stability low. Then as the peripheral members' evaluations of the collective recover, their expectations increase and with them the stability of the collective. Reduced message impact decreases the rate at which individuals' beliefs change, slowing down the rate at which leavers exit a collective, the committed core forms, and peripheral members' evaluations fall and recover. As a result, reducing message impact alters the rate at which a collective's stability develops.

The non-linear pattern seen in Figure 16b is the result of "catching" collectives in different stages of development. In the low message impact condition, a reduced rate of belief change results in many leavers having not yet worked their way out of a collective. In the high impact condition, stability is higher because the leavers have end their membership and the peripheral members have recovered. In the moderate condition, most of the computational collectives have lost the leavers, but because the rate of belief change is reduced, there has not yet been time for the peripheral members to recover. As a result, stability in these collectives is lower than the low impact condition, because the peripheral members have had time to 'reach bottom', and lower than the high impact condition, because they have not had enough time to recover.

Communication features, such as message impact and noise cost, significantly affect both member and structural development processes in voluntary social collectives. As the impact of a message increases, individual beliefs form more rapidly. As noise processing costs increase, members demand more focus from collectives. Consequently, in contexts with higher message impact and noise costs, member and structural development proceeds more quickly, resulting in smaller, more stable collectives. In contrast, in the presence of reduced message weight and noise costs, as are expected in networked environments, member and structural development processes take more time, resulting in larger, less stable collectives. Thus while networked environments may make communication more efficient by reducing communication costs, they may slow down the development of voluntary collectives, and as result be the site of significantly different social structures.

Discussion

The theory presented here draws from research on belief change and social structure to model the intertwined processes of member and structural development in voluntary social collectives. In some ways these collectives may seem different than the task groups which have been the focus of past research. In the structures considered here membership, specifically exposure to communication activity, is the result of individual choice. In addition, activity was assumed to be interest-based. This is in contrast to the task or decision oriented formal groups, such as production and management teams, that have been the focus of prior research.

In spite of these apparent differences, the model presented here is applicable for management researchers. On one hand, the process model of voluntary social collective development is useful because it tells us something about a type of social structure which plays an important role in the flow of information within and between organizations (e.g. Goodman and Darr, 1998; Van Hippel, 1988). As organizations, both public and private, spend more on information technology with the goal of facilitating information flow, it becomes increasingly important to understand how features of the technology and the social processes of collective development interact. This theory is also important for researchers interested in more traditional task and decision-oriented groups. Although formal membership in traditional teams may be the result of managerial action, it is often the case that exposure to communication activity is the result of individual choice. In addition, the process of developing beliefs about a formal team's activities and goals is likely to be driven by exposure to communication activity, much the same way belief about content develop in voluntary social collectives. Hence, future work should consider how the proposed theory might be applied for describing development of traditional teams and groups in dynamic organizational environments.

Although the model provides a basic description of core elements of collective development, there are several areas that warrant further attention. One such area is the participation model. The participation model is similar to those used in prior studies of small group participation and information sharing. An individual's probability of contributing to the communication stream is fixed and determined exogenously. Individual's contributions are limited to one per time period. Message topics are randomly chosen through a process that is independent of the prior activity within the group. Development of a more detailed model of communication activity and participation would provide potentially useful insights into the link between member beliefs, communication activity, and membership movement and their role in the development of social collectives.

Another aspect of the model that would benefit from additional development and analysis is the introduction of members. The current model focuses on membership loss. While this is likely to be an important mechanism for determining the composition and focus of voluntary social collectives, it is, in some sense, only half the story. A more complete model of membership movement would also consider the role that the inflow of new members plays in the structural development processes considered here.

Finally, it must also be recognized that voluntary collectives do not exist in a social vacuum. The development processes considered here take place within a social system in which there are other collectives. However, as with studies of the development of single collectives, discussions of larger social systems provide little insight into how a changing communication infrastructure might alter the development of these systems (for notable exceptions see Carley, 1995a, 1995b, Carley and Wendt, 1991). Combination of this and other work which considers the role of communication technology, with explicit models of social system dynamics, such a those proposed by McPherson (1983b) and McPherson and Rotolo (1996), would allow researchers to consider the role that technology plays in altering the dynamics of complex organizations and societies.

Conclusion

Over time, the processes of individual attitude change and membership movement combine to shape both a collective's true interests and member perceptions of that focus. Although networked environments are seen as speeding up communication, reducing the costs of communication may actually slow down these processes which underlie important aspects of voluntary collective development. Lower communication costs and small units of communication reduce the pressure on individuals to reach confident conclusions about continuing membership. Members who would have left quickly under conditions of high cost, instead remain in the collective. As a result, networked collectives are expected to be larger, more diverse, and less stable, having lower minimum and average member evaluations, than collectives that rely on face-to-face communication.

New communication technologies change the ways members of groups, associations, and organizations communicate with one another. Yet having the capability for communication does not necessarily mean that it will occur. Based on their perceptions of the social context, individuals decide which collectives and people they will interact with. However, in addition to providing new means for communication, these technologies also alter the way individuals receive and process information about the social structures in which they are members. This, in turn, affects perceptions of the social context and alters how communication patterns change over time. Thus while new technologies may increase the mechanical efficiency of communication, it is the secondary effects on perception and evaluation of collectives that may ultimately underlie the development of the structures in which social communication actually occurs.

As network infrastructures such as the Internet become widely available there is an increasing tendency to equate providing the ability to communicate with supporting and encouraging interaction. In support of this perspective, an extensive stream of computer-mediated communication research has focused on demonstrating the many ways that individuals *can* interact effectively in online environments. This work has served to define a universe of possibilities for designers and developers of networked infrastructures. However, much of this early work has failed to recognize that, while individuals can behave in many ways, how they *will* behave is significantly affected by the social structures that arise both within and around a networked environment. Furthermore, while there is some chance of effective social structures developing spontaneously, as the description of a population of listservs presented in Butler, 1999a indicates, the emergence of online social structures is far from a deterministic outcome of providing a technological infrastructure that allows communication.

While it is important to understand how individuals can behave in online social settings, for both practical and theoretical reasons, it is also crucial to consider how those social contexts develop and evolve. Online social structures, such as listservs, are subject to structural dilemmas similar to those that have been observed in traditional social structures (Butler, 1999b). While new technologies may alter the form of communication, membership and communication activity continue to have both positive and negative consequences which must be balanced in order to maintain ongoing online social structures. Furthermore, while the first-order effects of new technologies on mechanistic efficiency may be the most visible it is incorrect to assume that they are the most important consequences of introducing new communication infrastructures.

However, without dynamic models which take into account individual, structural, and technological features of social environments our ability to understand, and perhaps ultimately predict the impact of new communication technologies is likely to be limited. By combining empirical data analysis and computation modeling this work provides a foundation for future research that promises to provide insight into how technology has, and will continue, to affect the social structures around us.

All of our social actions in on-line environments take place in the context of larger social structures. Those social structures play significant role in determining who you interacts with, who you can influence, and who can influence your beliefs, attitudes, knowledge, and actions. Consequently, it is important for researcher to consider not just the behavior of individuals, but the nature online social contexts as well. Towards this end it is important that theories be developed that combine communication, technology, and social structure development. Achieving this requires that we move beyond the 'novelty' of computer-mediated communication technologies. The idea that computer-mediated communication environments should operate in a fashion unlike traditional structures leads to biased descriptions of online social contexts (Butler, 1999a). Rather than developing models that explain the structures that are *likely* to arise in networked infrastructures, research focuses on highlighting "new" forms of structure. All social structures make use of some type of communication infrastructure. Online social structures are subject to many of the same structural constraints as other traditional social structures (Butler, 1999b). Consequently, Modeling the communication aspects of social structure operations is complicated by the dynamic nature of these systems. Not only the nature of the social structures around you affect who you influence (and can influence), who you interact with, and how you interact with them, plays an important, but often subtle role, in the development of these context.

References

Baym, N. (1993) Interpreting soap operas and creating community: Inside a computer-mediate fan culture. Journal of Folklore Research, Vol. 30, pp. 143-176. [Also appears in S. Kiesler (Eds.), Culture of the Internet]

Blau, Peter M. (1967) Exchange and Power in Social Life. New York: Wiley.

Butler, B. S. (1999a) When is a group not a group: An empirical examination of metaphors for online social structure. Unpublished Working Paper, University of Pittsburgh.

Butler, B. S. (1999b) Membership Size, Communication Activity, and Sustainability: The Internal Dynamics of Networked Social Structures. Unpublished Working Paper, University of Pittsburgh.

Carley, Kathleen M. (1991) A Theory of Group Stability. American Sociological Review, Vol. 5-6, pp. 331-354.

Carley, Kathleen M. (1995a) Communicating New Ideas: The Potential Impact of Information and Communication Technology. Technology in Society, Vol. 18, No. 1. pp. 1-12.

Carley, Kathleen M. (1995b) Communication Technologies and Their Effect on Cultural Homogeneity, Consensus, and the Diffusion of Ideas. Sociological Perspectives, Vol. 38, No. 4, pp. 547-571.

Carley, Kathleen M. and Wendt, Kira (1991) Electronic Mail and Scientific Communication. Knowledge: Creation, Diffusion, and Utilization, Vol. 12, No. 4. pp. 406-440.

Etzioni, Anitai (1964) Modern Organizations. Englewood Cliffs, NJ. Prentice Hall.

Epstein, Joshua M., and Axtell, Robert (1996) Growing Artificial Societies. Brookings Instituion Press. Washington, DC.

Forsyth, Danelson R. (1990) Group Dynamics (2nd. Ed.) Pacific Grove, CA: Brooks/Cole.

Fischer, B.A. (1970) Decision emergence: Phases in group decision-making Speech Monographs, 37: 53-66.

Gersick, Connie J.G. (1988) Time and Transition in Work Teams: Toward a New Model of Group Development. Academy of Management Journal, 31(1): 9-41.

Gersick, Connie J.G. (1989) Marking Time: Predictable Transitions in Task Groups. Academy of Management Journal, 32(2): 275-309.

Goodman, Paul S. and Darr, Eric D. (1998) Computer-Aided Systems and Communities: Mechanisms for Organizational Learning in Distributed Environments. Unpublished Manuscript, Carnegie Mellon University, Pittsburgh, PA.

Goeffman, Erving (1959) Presentation of Self in Everyday Life. New York: Doubleday/Anchor.

Hannan, Michael T. & Freeman, John (1977) The Population Ecology of Organizations. American Journal of Sociology, Vol. 82, pp. 929-940.

Hare, A.P. (1976) Handbook of Small Group Research (2nd Ed.) New York: Free Press.

Hunter, John E., Danes, Jeffery E., and Cohen, Stanley H. (1984) Mathematical Models of Attitude Change (Volume 1). . New York: Academic Press.

Kaufer, David S. and Carley, Kathleen M. (1993) Communication at a Distance: The Influence of Print on Sociocultural Organization and Change. Hillsdale, NJ: Lawrence Erlbaum

Kollock, Peter & Smith, Marc (1996) Managing the Virtual Commons: Cooperation and Conflict in Computer Communities. In S.C. Herring (Eds.) Computer Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives (pp.226-242), Philadelphia: John Benjamins Publishing.

LaCoursiere, R.B. (1980) The life cycle of groups: Group development stage theory. New York: Human Sciences Press.

Law, Averill M. and Kelton, W. David (1991) Simulation modeling and analysis (2nd ed.) New york : McGraw-Hill.

McGrath, J. E. (1984) Groups: Interaction and Performance. Englewood Cliffs, NJ: Prentice-Hall.

McGrath, Joseph E. and Hollingshead, Andrea B. (1994) Groups Interacting with Technology: Ideas, Evidence, Issues and an Agenda. Thousand Oaks, CA: Sage.

McPherson, Miller (1983a) The Size of Voluntary Organizations. Social Forces, Vol. 61, No. 4, pp. 1045-1064.

McPherson, Miller (1983b) An Ecology of Affiliation. American Sociological Review, Vol 48. pp. 519-532.

McPherson, Miller J. (1990) Evolution in Communities of Voluntary Organizations, In. J.V. Singh (Ed.). Organizational Evolution New Directions.(pp. 224-245). Newbury Park, CA: Sage.

McPherson, Miller J. and Rotolo, Thomas (1996) Testing a Dynamic Model of Social Comparison: Diversity and Change in Voluntary Groups. American Sociological Review, Vol. 61. pp. 179-202.

Moreland, R. L. & Levine, J. M. (1982) Socialization in Small Groups: Temporal Changes in Individual-Group Interactions. Social Psychology, Vol. 15, New York: Academic Press.

Moreland, R.L., Hogg, M.A., and Hains, S.C. (1994) Back to the Future: Social Psychological Research on Groups. Journal of Experimental Social Psychology, 30: 527-555.

Poole, M.S. (1981) Decision development in small groups I: A comparison of two models. Communication Monographs, 48: 1-24.

Poole, M.S. (1983) Decision development in small groups II: A study of multiple sequences of decision making. Communication Monographs, 50: 206-232.

Rojo, Alejandra (1995) Participation in scholarly electronic forums. Unpublished Phd. Disseration, University of Toronto, Ontario, Canada [http://www.oise.edu/~arojo/]

Scheidel, T. and Crowell, L. (1964) Idea development in small discussion groups. Quarterly Journal of Speech, 50: 140-145.

Simmel, Georg [1908] (1955) Conflict and the Web of Group Affiliations. Translated by K. Wolff and R. Bendix. Glencoe, IL: Free Press.

Skvoretz, John (1988) Models of Participation in Status-Differentiated Groups. Social Psychology Quarterly, Vol. 51, No. 1, pp. 43-57.

Sproull, Lee., & Faraj, Samer (1997) Atheism, Sex, and Databases: The net as a social technology. In S. Kiesler (Ed.), Culture of the Internet. Mahwah, NJ: Lawrence Erlbaum.

Sproull, L., & Kiesler, S. (1990) Connections: New ways of working in the networked organization. Boston, MA: MIT Press.

Tuckman, B. (1965) Developmental sequence in small groups Psychological Bulletin, 63: 384-399.

Tuckman, B. and Jensen, M. (1977) Stages of small-group development. Groups and Organizational Studies, 2: 419-427.

Turner, Jonathan H. (1988) A Theory of Social Interaction. Stanford, CA: Stanford University Press.

Von Hippel, Eric (1988) The Sources of Innovation. New York, Oxford University Press.

Whitaker, Steve (1996) Talking to strangers: an evaluation of factors affecting electronic collaboration., .Proceedings of CSCW'96, Boston, MA.

Wittenbaum, Gwen M. and Stasser, Garold (1996) Management of information in small groups. In J. L. Nye and A. M. Brower (Eds.) What's social about social cognition? : Research on socially shared cognition in small groups. (pp. 261-282). Thousand Oaks, CA : Sage.

Worschel, Stephen (1996) Emphasizing the Social Nature of Groups in a Developmental Framework. In J. L. Nye and A. M. Brower (Eds.) What's social about social cognition? : Research on socially shared cognition in small groups. (pp. 261-282). Thousand Oaks, CA : Sage.

Zernhausern, Robert & Wong, Florence (1997) Virtual Personality of a List: A Preliminary Examination of the Demography of Interent Lists. In Sudsweeks, McLaughlin, and Rafaeli (Eds.) Network and Netplay: Virtual Groups on the Internet. Cambridge, MA: AAAI/MIT Press.

Paper One Appendices

When is a Group not a Group:
An Empirical Examination of Metaphors for Online Social Structure

Appendix A: Group Selection Procedures

This appendix contains details of the procedure used to select the sample of listservs used in this work. This process consisted of three major phases: topic selection, preliminar sample construction, and listserv screening. Each of these phases is described here. When possible intermediate 'results' of the procedures, such as topic lists, are also included.

Topic Selection

To ensure that the sample spanned a range of topics and member populations the process began with the selection of a stratified set of topics. Topics were randomly selected from the categories used to index listservs in an on-line directory entitled Publicly Accessible E-mail Lists (PAML)[1]. To address ethical concerns about public observation of certain social settings, personally sensitive topics (such as mental disorders or sexual lifestyles) were removed. To increase the chances that the resulting sample covered a range of member populations the topic set was chosen to equally represent the following general categories: professional/academic, personal/non-work, and mixed.

The following procedure was used to construct this collection of topics:

1. Two hundred topics were randomly chosen from the PAML topic list
2. Two raters categorized each topic was categorized by type and sensitivity.
3. Topics coded as sensitive by either rater were eliminated (Table A-1).

codependence	postpartum
diabetes	support
theism	midwiving
substance-abuse	child-abduction

Table A-1: Sensitive Topics

[1] The directory used for topic selection was PAML (Publicly Accessible Mailing Lists) which is located on the WWW. This directory in was chosen because it provided a reasonably extensive subject index (1200+ categories) for a set of accessible e-mail based Internet collectives (2500+ listservs).

4. Topics were randomly selected from the remaining topics until one hundred topics were chosen that represented the three general types. [The topics were selected sequentially and when new topic's type was already adequately represented in the final topic list it was discarded.].

This procedure resulted in a collection of one hundred topics split between subjects that were likely to be of interest for professional/academic reasons, personal/non-work reasons, or a combination (Table A-2).

Professional/ Academic	**Personal/ Non-Work**	**Mixed**
botany	olympics	gender
teachers	trips	feminism
manual-therapy	vacations	geology
typography	porcelain	oceania
consultanting	needlework	law-enforcement
manufacturing	hebrew	cooking
document-delivery	astrology	cdrom
narrative	rock-bands	architecture
neuroscience	fireworks	stage
public-servant	dance	youth
medical-services	dyeing	mexico
ethnogrophy	florida	spanish-language
editing	glassware	freedom
advertising	connecticut	poetry
agriculture	atari	paramedics
employment	genealogy	archeology
translation	knives	ethics
accounting	walking	new-york
energy	handball	fiction
imaging	lakes	greek-language
metabolism	fiberart	colorado
oceanography	nostalgia	piano
broadcasting	theme-parks	folktales
	running	family
	power-boating	athletics
		afrikaans
		german-language
		uruguay
		investment
		money
		usenet
		philosophy
		canada
		taxation
		acting
		healthcare
		media
		pop-culture
		electronics
		arts
		scotland

Table A-2: Final Topic List

Preliminary Sample Selection

The following iterative selection process was used to construct the preliminary sample of listservs. A topic was randomly selected from each of the general categories in the topic list (Table A-2). All listservs indexed in the PAML (Publicly Accessible Mailing List) directory under these topics were selected. In addition, the topic labels were used as keywords to search for listservs in LISZT (http://www.liszt.com/) and the List of Lists, two other directories with a combined coverage of 70,000+ e-mail based groups[2]. The listservs identified from these three sources were then added to the preliminary sample and duplicates were eliminated.

This process was repeated until a preliminary sample of 1066 listservs was created (21 rounds). The preliminary sample size was selected based on estimated elimination rates for the screening process in order to result in a final sample of 100 listservs.

Listserv Screening

The final sample was refined in a multi-stage process by removing inaccessible or unsuitable listservs from the preliminary sample. At each stage some portion of the preliminary sample was eliminated (Table A-3). Occasionally was not clear until a later stage of the process that a sampled listserv should be eliminated for a given reason. In those cases, the listserv was eliminated in the stage when it became apparent that it did not meet the overall criteria for inclusion in the sample. Thus, the criteria described in the stages here should be seen as cumulative. See the discussion of the sample in the text of the paper for a breakdown of the eliminated lists categorized by reason for removal.

Phase 1 involved screening out highly managed listservs, non-automated lists, non-English lists, and listservs that dealt with sensitive topics. Broadcast listservs, such as

[2] One of the consequences of using the topic labels are keywords in the directory searches is that, because the directory search engines were not particularly sophisticated, the topics only loosely characterize the selected groups.

newsletters and announcement lists, were not included because they are typically highly edited, with one individual generating content for consumption by a passive audience. Listservs which had significant formal management, either in the form of moderation (i.e. content review or control) or explicit member screening (i.e. member review or control), were also eliminated. Listserv which are linked directly with other online social structures (e.g. lists which mirror the contents of a USENET newsgroup) were eliminated because it would not be possible to collected data about the membership of the other structure. Listservs that existed solely to support interaction among student in a single academic course were removed because of the formal management provided by the instructor. Listservs that were not technically accessible, did not make use of an automatic infrastructure, used languages other than English, did not provide access to membership information, or dealt with sensitive topics were also eliminated from the sample.

In phase 2, listservs located on inoperative, inaccessible, or unreliable list servers were removed from the sample. Listservs that resided on accessible list servers, but were themselves inaccessible, were also eliminated. Inaccessibility was typically a result of the site being inoperable, though there were several cases in which access to a listserv was restricted such that only particular individuals could become members.

During phase 3, listservs whose membership was likely to be sensitive to observation were removed from the sample. Two raters coded each listserv as either sensitive or non-sensitive based the short (typically one line) descriptions provided by the on-line directories (see Appendix B for descriptions of the listservs in the final sample). Listservs identified as sensitive, by either rater, were eliminated.

For this reason the topic labels is used only in the group selection process and not in any direct analysis of the groups.

To further protect the interests of group members, in phase 4 the owners of the remaining lists were sent e-mail messages describing the study and informing them that their lists had been selected for inclusion in the sample. If the message to the owner could not be delivered the listserv was eliminated. If the owner replied stating that they did not want to be included in the study the listserv was removed from the sample and a confirmation message was sent to the listsowner. For the remaining listservs the an addition step was taken to ensure that the listowners had the opportunity to op out of the study. The project e-mail account, which was labelled "Egroup Dynamics Study", was subscribed to each of the listservs remaining in the sample. After 14 days, in which no messages were save and no additional action was taken, continued membership was verified. If the project e-mail account had been removed from the listserv then it was dropped from the sample.

Phase 5 involved a final verification that the listserv was publicly accessible, that the listserv membership information was available, and that it met all the criteria for inclusion in the sample (i.e. no formal management, not a broadcast group, etc.). This final assessment was conducted by the researcher based on the contents of the introductory messages that each listserv provided for new subscribers.

At each stage varying number of groups were removed from the sample resulting in the construction of a working sample of listservs that met the basic criteria for inclusion in the study (Table A-3). During the data collection period listservs were removed from the sample as they became inaccessible or inoperable. Finally, during measure construction and data analysis several listservs were eliminated when it was discovered that due to highly non-standard message or membership list formats it was not feasible to process the archived records.

	Number Eliminated	% of Preliminary Sample	Remaining Sample Size
Initial Sample (1066)			
Phase 1: Group Type Verification	329	30.9%	737
Phase 2: Availability	50	4.7%	687
Phase 3: Sensitivity	14	1.3%	673
Phase 4: Informing List Owners	343	32.2%	330
Phase 5: Verification of Data Availability	46	4.3%	284
During Data Collection	80	7.5%	217
During Measure Construction and Analysis	13	1.2%	204
Final Sample (204)			

Table A-3: Elimination of Listservs During Screening

To check that the sample covered the intended diversity of topics an additional coding based on the short descriptions was performed. Topic type (professional, academic, and/or personal) measures were constructed by combining two rater's evaluations of each listservs. Coders were asked to indicate on a scale from 1 (low) to 5 (high) the likelihood that a significant portion of listservs membership being involved for professional, personal and academic reasons (three measures for each group). Inter-rater reliability was found to be acceptable, with the Cronbach alphas of 0.79, 0.78, and 0.88 the three measures. The rater's evaluations were averaged to create three measures for characterizing the topic of each listserv (Table A-4).

	Mean	Median	Std Dev.	Alpha
Professional/Work-Related Membership	3.42	3.5	1.36	0.79
Academic Membership	2.49	2	1.30	0.78
Personal Membership	1.83	1	1.39	0.88

Table A-4: Listserv Topic Coding Summary

Although the sample is not completely balanced (being slightly biased towards professional and academic listservs), the three topic measures indicated that the sample included a reasonable number of listservs with each focus.

Appendix B: Sampled Listserv Descriptions

This appendix contains a list of the short (1 line) descriptions provided for each listserv in the final sample. These descriptions were used to create the coding of topic type (Appendix A) and collective type (Pure vs. Hybrid). In addition, these publically available descriptions are presented here to provide an indication of the composition of the final sample without compromising the privacy of the lists and listowners. If the listserv is identified by name in this list it is because the listowner chose to include the name in the description of the list which was made available to the general public through the listserv directories.

1) Discussions on the plant kingdom ethnobotany issues
2) Women and Gender in the Ancient World
3) AARE - Women Researchers on Gender Equity Mail list
4) an interdisciplinary forum to foster dialogue on issues of race ethnicity gender class and sexuality and discuss the future of ethnic studies at the undergraduate and graduate level.
5) Discussion list about gender economics
6) Digest for gender economics list
7) Magyarorszagi Gender Studies lista
8) Race/Gender Resource Center/Bucknell University
9) Women and Gender Issues Discussion List
10) Concordia Outdoors Club and member organized trips
11) List for teacher trainer to use telet
12) A list for the Discussion of Trips in the Voyageur Wilderness Program
13) Discussion list for Feminism in Geography
14) ACT Teachers Professional Development Mailing List
15) American Federation of Teachers with Oregon
16) I*EARN faculty lounge area
17) For Teachers of the Advanced Placement Statistics Course.
18) American String Teachers Association List
19) Teachers participating in Book Rap Project
20) Teachers of Celtic Languages
21) list for Central-European teachers
22) Action Research for teachers list
23) Discussion list for Eng 384-Composition for Teachers at WIU
24) Computer Studies Teachers' discussion list
25) Dutch language teachers' listserv
26) professional development of teachers and caregivers of young children
27) Discussions on mentoring in the professional development of teachers
28) EMPATHY <Teachers of Interpersonal Communication>
29) Teachers involved in the Endeavour Project
30) College Experience 199 class teachers discussion
31) Univ. System Teachers in GA

32) Computer Studies Teachers' discussion list
33) Teachers of Hindi Languages
34) International Business Class Teachers
35) Israeli English Teachers Network
36) Int Collab - teachers & users of Internet in classroom
37) Discussion by students and teachers about education
38) For Legal Assistant Teachers
39) Less Commonly Taught Language teachers
40) List for the Massachusetts Assoc. of Biology Teachers
41) Computer Teachers and Computer Ed in Michigan"
42) Discussion between Maryland Assoc. of Science Teachers
43) Missouri Business Education Teachers Discussion List
44) Music Teachers National Association Mailing List
45) Nebraska Association of Teachers of Mathematics
46) Teachers of Nordic Languages
47) PA Assn. of Scholar Teachers List
48) Project Harmony for Teachers
49) Teachers of Polish Languages
50) Preservice and Student Teachers Online
51) -a Discussion group for teachers school administrators and educational
52) Special Needs Education Network For Teachers
53) Investing in teachers of color
54) A discussion list for teachers of American Sign Language
55) Western Cape Teachers' Mailing List
56) TESLK-12: Teachers of English as a second language to children
57) Vermont Council of Teachers Educators
58) Western North Carolina Teachers list
59) Way Cool Software Reviews by Children Teachers and Parents
60) Pre-physical Therapy Club
61) Geology T.A. List
62) Geology and Earth Science Education Discussion Forum
63) The SCIENCE.ORG geology enthusiast mailing list (http://geology.science.org/)
64) The SCIENCE.ORG geology professional mailing list (http://geology.science.org/)
65) The SCIENCE.ORG geology scientist mailing list (http://geology.science.org/)
66) The SCIENCE.ORG geology student mailing list (http://geology.science.org/)
67) -"Geology Graduate Students"
68) The UPJ Geology department mailing list
69) Alcala Consulting Group
70) Alcala Consulting Group Staff
71) Discussion List for the Business Aspects of Consulting to Nonprofits
72) CJUST-L: Criminal Justice Discussion List
73) ICCON design drafting assembly sheet metal harness
74) ICCON simulation test plastics solver TMG/ESC ...
75) ICCON manufacturing generative NC CAMAX ...
76) ICCON data mgmt system adm plotting data trans ...
77) ICCON misc info conferences general tips ...
78) Integration In Manufacturing And Beyond
79) MMM: Members of the Masters of Manufacturing Management Program
80) Discussion of Hebrew Grammar and Etymology.

81) For those working with the beta CD-ROM project.
82) CD-ROM Beta Project Digest
83) NARRATE - International Conference on Narrative
84) Association for Computer-Aided Design in Architecture
85) ASI Software Architecture & Standards Committee
86) ASI Software Architecture & Standards Committee
87) European Landscape Architecture Network - e.LAN
88) European Landscape Architecture Students Association - ELASA
89) Intelligent Tutoring Systems Architectures mailing list
90) Communication architectures for ITS Components mailing list
91) Exploring Industry Standard Architectures mailing list
92) Department of Landscape Architecture
93) Landscape Architecture Staff
94) Mail ARCHitecture Task Force
95) Announcement/discussion of Dept. of Landscape Architecture's Network of Environmental Management Interests
96) The STI Architecture Framework Group
97) Women in Architecture and Allied Arts
98) Info about the Ominous Seapods rock band.
99) Undergraduate Society of Neuroscience List
100) NEUROSCI- Zlotowski Center for Neuroscience
101) NUIN: Northwestern University Institute for Neuroscience
102) Macromedia Backstage discussion list
103) Beall Hall Stage Crew schedule and information
104) Public Sector Management List
105) 4-H Center for Youth Developemnt
106) Creative Youth Ministry Idea Sharing List
107) CTC Youth Discussion List
108) European Youth Mailing List
109) Volunteers helping Eritrean youth
110) Food Stamp Nutrition Education for youth
111) Gustavus Youth Outreach Mailing list and Dis
112) LIYSF - 37. London International Youth Science
113) A discussion for Presb. of New Hope Pr. Youth Connection
114) Presb. Youth Connection for the National Capital
115) summer sports camps & clinics for disadvantaged youths
116) Children's and Youth Services List
117) Discussion group on youth hockey
118) A list for attendance data for the HealthNet project
119) Central Illinois English Country Dancers' Mailing List
120) Urbana Country Dancers' Mailing List
121) Bilkent Dance Group Mailing List
122) -UF International Folk Dancers
123) Kentucky Guidance Counselor Discussion List
124) A list for those interested in Chinese Lion Dancing.
125) Guidance Counsellors Discussion Group
126) The PhD program in Dramatic arts and Dance
127) modern western square dance caller discussion.
128) Univ of Md's Mexico Exchange Program
129) New Mexico Council on Technology in Education

130) Immersion dyeing and surface application of dyes to fabr
131) Immersion dyeing andsurface application of dyes to fabri
132) ANet Accounting Ethnography List
133) Editorial board for Spanish Language Magazine
134) -Florida - China Linkage Institute Discussion List
135) Florida F-Body Mailing List
136) Florida School-to-Work Discussion List
137) FLorida Artificial Intelligence Research Symposium
138) Florida Society of Geographers mailing list
139) Presbytery of Florida PC(USA) Online Discussion
140) The South Florida Science Fistion Soceity Discussion Li
141) -University of Central Florida Educator's List
142) Copy Editors and Editing
143) Doom Editing
144) Doom Editing
145) California Library Association's Intellectual Freedom Roundtable
146) Albany New York SAS User's Group
147) Cornell Center for the Environment discussion on environmental issues as they relate to Ithaca - New York State - the US - and the Earth.
148) The Hillel of New York/IJC Mailing List
149) New York/New Jersey Regional Chapter/Medical Library Association List
150) New York City / Africa science education projects
151) NYRHN-L New York Rural Health Network
152) The New York State Council of World Trade Associations
153) New York State Library Assistants' Association: Discussion List
154) Western New York Music Forum
155) Connecticut Library Technology List Server
156) Connecticut Planning Listserv
157) University of Connecticut Chinese Student Association
158) Agriculture Discussion
159) Women in Agriculture Mailing List
160) Precision Agriculture Mailing list
161) SLA-FAN Special Libraries Association--Food Agriculture and Nutrition
162) I*EARN student poetry prose and art
163) Medieval Lyric Poetry: 1995 NEH Summer Institute
164) Discussion of poetry in English 1900-1945
165) Discuss and share original poetry verse and prose.
166) Atari 2600 emulator group
167) Correspondence with pg students re: temporary posts
168) Employment Opportunites List
169) Student Employment Discussion List
170) Invitation/employment of international scholars
171) Ringwald Geneology List
172) Manual Japanese translation Project
173) Russian Literature in Translation
174) Discussion of Translation Theory and Practice
175) Vincent of Beauvais Translation Project
176) VN Translation List
177) ANet Accounting Information Systems List

178) ANet Accounting Educational Programs List
179) Accounting Issues Discussion List
180) ACCOUNTING-WG
181) ANet Financial Accounting List
182) ANet Accounting History List
183) ANet International Accounting List
184) ANet Management Accounting List
185) ANet Oil & Gas/Extractive Industries Accounting
186) ANet Social Accounting List
187) ANet Accounting Education List
188) ANet - Accounting and Technology list
189) Med School Departmental Accounting System task force members
190) -Southeast Conference on College Cost Accounting
191) Announcements/questions related to campus On-line Purchasing/Accounting Link (PAL) system.
192) ANet Ethics List
193) K-12 acceptable use policy discussions
194) Cyber-Ethics Discussion (TIS)
195) legal ethics list
196) The Social Ethics Discussion List
197) SPJ Ethics Mailing List
198) Technology Ethics Dementia
199) Leicester University Fell-Walking Society List
200) MSU Healthy Walking Club
201) NCSU Energy Management Continuing Education Program
202) Discuss Energy Scheming energy analysis software
203) IASEE-L Discussion List on Solar Energy Education
204) Research into fuel cells and new energy techniques
205) Photoinduced Charge and Energy Transfer List
206) World Information Service on Energy (WISE)
207) Nanotechnology in science fiction mailing list.
208) Humorous Shared-World Superhero Fiction
209) Forum on Designing Fictional Settings / Worlds
210) Infocom's Z-machine for interactive fiction
211) To provide a public forum for the discussion of issues related to Greekand Latin Language and Lexicography.
212) Great Lakes F-Body Mailing List
213) Great Lakes Curriculum resources
214) -"Florida School District Council on Comprehensive MIS"
215) Great Lakes Pollution Prevention Roundtable
216) Harvard Medical Area Ballroom Dance Club
217) Carnivorous Plants

Table B-1: Final Sample Listserv Short Descriptions

Appendix C: Data Collection and Processing Procedures

The data collection process had four stages: daily data collection, listserv data archive creation, datafile creation, and measure construction. This appendix contains the PERL and SAS scripts that were used in each stage to collect and process the data.

A project account was created and subscribed to each of the selected listservs. This resulted in a stream of messages being delivered to this account. A PERL script was created which processed the messages on a regular schedule (Section C-1). This script ran on a DEC workstation for the duration of the observation period. Once a day this script executed a subprocess that requested the membership list from each listservs (Section C-2). In addition, every size hours the scheduling script executed another subprocess which filed the messages (Section C-3). This subprocess is performed in several stages. The content messages are filed in one directory (Section C-4) and the messages containing the membership records are filed in another (Section C-5). Then the content messages are processed (headers are summarized and sender information is encrypted) and combined into a single file (Section C-6). The membership lists are also processed (headers are summarized and the membership records are encrypted) and combined into a single file (Section C-7). This results in two archives, one for messages and one for membership data, being created for each day.

The daily membership and message archives were then downloaded to a Macintosh where they were collected for the observation period. After the set of daily archives had been compiled (a total of 260 files, 130 message archives and 130 membership archives), they were then converted into listserv specific archives. Two PERL scripts were run to convert the daily membership records (Section C-8) and message archives (Section C-9) into files that contained only the records for a single listserv.

From the listserv raw data archives a set of comma-delimited datafiles was created for use with statistics analysis software. First, raw membership data was used to create an intermediate file that contains the daily size of each listserv (Section C-10). This file is then coverted into a comma-delimited file which includes daily size, membership growth, and loss measures for each day in the observation period (Section C-11). A similar process was done, with an intermediate file created for daily message volume data (Section C-12). From this a comma-delimited file containing the message activity was created (Section C-13). Additional message and participation activity data was computed and stored in secondary data files (Section C-14).

The comma-delimited datafiles were then combined into a single data set using SAS (Section C-15). This script also used the processed data to construct the measures of membership size, participation structure, communication activity, and membership change.

Daily data collection

C-1: Processing Cycle

```
#!/usr/local/bin/perl
# -----------------------------------------------------------------
# This perl script is the control for the first phase of the
# electronic group dynamics data stream.  It executes various
# data processing routines on a timed schedule.
# -----------------------------------------------------------------
# Brian Butler
# Created: 6/97
# -----------------------------------------------------------------

# Include the definitions of directories, log files, and addresses
require "filenames.pl";

# Send a message indicating the the processing loop was started
$startDate = `date`;
chop $startDate;
$subject = "EG_Processing_Start";
$message = "EGROUP Processing Loop has been started ($date)";
system("printf \"$message\" | mail -s $subject $monitorAddress");

# Loop continuously
#  Once and hour determine if there is a scheduled activity.
do
{
     # Determine the amount of time to the next hour
     ($sec, $min, $hour, $mday, $mon, $year, $wday, $yday, $isdat) =
       localtime(time);
     $sec_left = 3600 - ($sec + (60 * $min));

     # Wait until 1 minute into the next hour
     sleep($sec_left + 60);

     # Determine the current hour
     ($sec, $min, $hour, $mday, $mon, $year, $wday, $yday, $isdat) =
       localtime(time);

     # Based on the current hour executeany scheduled activities
     # ------------------------
     # Schedule
     # Midnight, 6AM, Noon, 6PM: File and process messages
     # 8PM: Request member lists from primary groups
     # 11PM: Send daily session log to monitorAddress
     # ------------------------
     if($hour == 20)
     {
         # Run the memberlist request script
         system("./list-request.pl&");

       # Send a message indicating that memberlist were requests
         $date = `date`;
       chop $date;
       $subject = "EG_List_Request";
       $message = "Member lists were requested ($date)";
       system("printf \"$message\" | mail -s $subject $monitorAddress");
     }
     elsif(($hour % 6) == 0)
     {
       # Run the memberlist request script
       system("./msg-filer.pl&");

      # Send a message indicating that msgs were filed and processed
       $date = `date`;
      chop $date;
      $subject = "EG_Filing_and_Processing";
      $message = "Incoming messages were filed and processed ($date)";
      system("printf \"$message\" | mail -s $subject $monitorAddress");    }
    elsif($hour == 23)
    {
       # Send a status message to my personal e-mail account
       $subject = "EG_Log";
      system("mail -s $subject $monitorAddress < $session_log_filename");

      # Clear the session log
       open(LOGFILE,">".$session_log_filename);
      close(LOGFILE);
    }

    # Revalidate the login
    system("klog egroup thesis");

}
until(0);
```

C-2: Membership List Request

```
#!/usr/local/bin/perl
# ---------------------------------------------
# This perl script
# requests the memberlist from the groups in the
# sample.
# ---------------------------------------------
# Created: 6/26/97
# Brian Butler
# ---------------------------------------------

# Include the definitions of filenames, logs, etc.
require "filenames.pl";

# Set the group information filename
$group_filename = $primary_list_filename;

# Open the log file
open(LOGFILE,">>".$session_log_filename);

# Write the entry for memberlist request
($sec, $min, $hour, $monthday,$month,$year,$weekday,$yearday,$dst) =
   localtime(time);
$month++;
print LOGFILE "MEMBER LIST REQUEST ($month\/$monthday $hour:$min): ";

# Print a status message
print "Requesting memberlists...";

# Open the group information file
open(INFO_FILE,$group_filename) || die "Can't open: $group_filename\n";

# Read the group info and generate appropriate memberlist request
$list_counter = 0;
while(<INFO_FILE>)
{
   # Increment the list counter
   $list_counter++;

   # Read the group information
   chop;
   ($group_num,$group_name,$server_address,$server_type)
        = split(/\,/,$_);

   # Construct the memberlist request command
   if($server_type eq "Majordomo")
   {
       $request_command = "who $group_name";
   }
   else
   {
       $request_command = "review $group_name";
   }

   # Send the request message
   system("printf \"$request_command\" | mail -s $group_name $server_address");
}

# Close the group info file
close INFO_FILE;
```

```
# Create the log output
print LOGFILE $list_counter," memberslists requested.\n";
close LOGFILE;

# Display status information on the screen
print "...finished ($list_counter memberlists requested)\n";
```

C-3: Message Processing

```
#!/usr/local/bin/perl
# ----------------------------------------------
# This script executes the sub-scripts which
# filter and archive the group and list messages
# ----------------------------------------------
# Created: 5/24/97
# Brian Butler
# ----------------------------------------------

# Include the description of the directories and log files
require "filenames.pl";

# Directories
$incomingDir = $mailboxDir;

# Display a status message to the screen
$date = `date`;
chop $date;
$| = 1;
print "--- Filtering & Processing Messages \@ $date\n";

# Open the log file
open(LOGFILE,">>".$session_log_filename);

# Write the entry for primary processing
($sec, $min, $hour, $monthday,$month,$year,$weekday,$yearday,$dst) =
   localtime(time);
$month++;
print LOGFILE "FILING AND PROCESSING SESSION ($month\/$monthday $hour:$min)\n";
close LOGFILE;

# Setup the summary entry for this session
open(SUMMARYFILE,">>".$session_summary_filename);
print SUMMARYFILE "$date,";
close SUMMARYFILE;

# Process the primary sample messages
system("./primary-msg-filer.pl");

# Process the secondary sample messages
system("./secondary-msg-filer.pl");

# Process the secondary sample messages
system("./list-filer.pl");

# Process the primary sample messages
system("./primary-msg-processor.pl");

# Process the secondary sample messages
system("./secondary-msg-processor.pl");

# Process the secondary sample messages
system("./list-processor.pl");

# Count the unprocessed messages
opendir(MSGDIR,$incomingDir) || die "Can't open $incomingDir\n";
@filelist = readdir(MSGDIR);
$uncounted_msgs = @filelist - 2;
closedir(MSGDIR);
```

```
# Add the count of unfiled messages to the session log
open(LOGFILE, ">>".$session_log_filename);
print LOGFILE "FILING AND PROCESSING SESSION ($month\/$monthday $hour:$min):
$uncounted_msgs unfiled msgs.\n\n";
close LOGFILE;

# End the filing session summary file entry
open(SUMMARYFILE, ">>".$session_summary_filename);
print SUMMARYFILE "\n";
close SUMMARYFILE;

# Display final status message
print "--- Filing and Procesing completed ($uncounted_msgs msgs remain)\n\n";
```

C-4: Message Filing

```
#!/usr/local/bin/perl
# ---------------------------------------------
# This perl script
# filters out group messages coming from
# groups in the primary sample and
# places them in the 'incoming-msgs' for
# further processing.
# ---------------------------------------------
# Created: 6/23/97
# Brian Butler
# ---------------------------------------------
# Include the description of the directories and log files
require "filenames.pl";

# Filenames
$group_filename = $primary_list_filename;

# Identify the directories
$incomingDir = $mailboxDir;
$storageDir = $primaryMsgStorageDir;

# Display a status indicator to the screen
$| = 1;
print "Primary message filing started....";

# Open the log file
open(LOGFILE,">>".$session_log_filename);
($sec, $min, $hour, $monthday,$month,$year,$weekday,$yearday,$dst) =
    localtime(time);
$month++;
print LOGFILE "PRIMARY MSG FILING ($month\/$monthday $hour:$min): ";

# Read in the group list
open(INFILE,$group_filename) || die "Can't open: $group_filename\n";
$list_counter = 1;
while(<INFILE>)
{
    # Read the group information
    chop;
    ($group_num,$group_name,$server_address,$server_type)
         = split(/\,/,$_);

    # Convert into lower case
    $group_name =~ tr/A-Z/a-z/;

    # Store group name and initialize the message counter
    $group_list{$group_name} = 0;

    # Increment the list counter
    $list_counter++;
}

# Process the messages
opendir(MSGDIR,$incomingDir) || die "Can't open $incomingDir\n";
```

```
@filelist = readdir(MSGDIR);
closedir(MSGDIR);

# Process each of the message files
$msgcounter = 0;
$proc_msgs = 0;
$direct_msgs = 0;
foreach $filename (@filelist)
{
    # Skip the . and .. entries in the directory
    if(($filename eq ".") || ($filename eq ".."))
    {
        next;
    }

    # Count the message
    $msgcounter++;

    # Open the message file
    open(MSGFILE,$incomingDir.$filename) || die "Can't Open: $filename\n";

    # Read and process the message
    $StudyAddressFound = 0;
    $GroupAddressFound = 0;
    $LineHeader = "";
    $LineContent = "";
    while((<MSGFILE>))
    {
       # If this is the end of the header then stop
       if(/^\n/)
       {
            last;
       }
       # If the current line begins with a blank then assume it is not
       # a header line
       elsif(/^\s/)
       {
           # Save the remainder of the line for additional processing
           $LineContent = $_;
       }
       else      # If the current line is a header field then identify it
       {
           # Save the line header
           $LineHeader = $_;
           $LineHeader =~ /^(\S*):/;
           $LineHeader = $1;
           $LineHeader =~ tr/A-Z/a-z/;  # Make it lowercase

           # Save the remainder of the line for additional processing
           $LineContent = $_;
           $LineContent =~ s/^\S*:/ /;
       }

       # Clean up the current line
       $LineContent =~ s/\s//;       # Remove the spaces
       $LineContent =~ tr/A-Z/a-z/;  # Make everything lowercase
```

```
   # Process the line type based on the type of header it is
   # If it is a from, to, or cc process the addresses
   if($LineHeader eq "to" ||
      $LineHeader eq "cc" ||
      $LineHeader eq "sender" ||
      $LineHeader eq "from")
   {
       # Store the addresses in the address list
       @address_list = split(/,/,$LineContent);

     # Find each address and encrypt it
       foreach $Address (@address_list)
       {
            # Extract the address
            $Address =~ /(\b[a-z0-9\.\-\+\_]*@[a-z0-9\-\_\.]*\b)/;
          $Address = $1;

            # Extract the group name
            $Address =~ /(\b[a-z0-9\.\-\+\_]*)@([a-z0-9\-\_\.]*\b)/;
            $AddressName = $1;
            $AddressDomain = $2;

        # Remover the owner markers
          if($LineHeader  eq "from" || $LineHeader eq "sender")
          {
              $AddressName =~ s/owner-//i;
            $AddressName =~ s/-owner//i;
        }

            # Process exceptions

          # Convert pod-net-digest (A non-existant group) to
          # straight pod net
          $AddressName =~ s/pod-net-digest/pod-net/;

        # Remove the bitnet info present in the TEACHASL name
          $AddressName =~ s/humber\.bitnet/teachasl/;

            # Determine if this the name of a group in the study
            $Temp = $group_list{$AddressName};
            if($Temp ne "")
          {
               $GroupAddressFound = 1;
             $GroupName = $AddressName;
            }
          # Otherwise check and see if it is the study address
          elsif($AddressName =~ /egroup/i)
            {
              $StudyAddressFound = 1;
            }
      }
    }
}

# Close Message file
close MSGFILE;
```

```
    # File the message appropriately
    if(($GroupAddressFound == 1) && ($StudyAddressFound == 0))
    {
   # Store the group address in the file
   system("printf \"EGroup:$GroupName\n\" | cat - $incomingDir$filename >
$incomingDir$temp_filename");

         # Move the classified messages
         system("mv $incomingDir$temp_filename $storageDir$filename");
         system("rm $incomingDir$filename");

         # Count the processed messages
         $proc_msgs++;
    }
    elsif($StudyAddressFound == 1)
    {
        # Count the direct messages
        $direct_msgs++;
    }
}

# Create the log output
print LOGFILE "$proc_msgs stored msgs, $direct_msgs direct msgs\n";
close LOGFILE;

# Write the summary entry for secondary processing
open(SUMMARYFILE,">>".$session_summary_filename);
print SUMMARYFILE "$proc_msgs,";
close SUMMARYFILE;

# Display status information on the screen
print "...finished ($proc_msgs msgs filed)\n";
```

C-5: Membership List Message Filing

```
#!/usr/local/bin/perl
# ----------------------------------------------
# This perl script
# filters out group membership-lists coming from
# groups in the primary sample and
# places them in an intermediate directory for
# further processing.
# ----------------------------------------------
# Created: 6/23/97
# Brian Butler
# ----------------------------------------------

# Include the definitions of filenames, logs, etc.
require "filenames.pl";

# Set the group information filename
$group_filename = $primary_list_filename;

# Setup the directories
$incomingDir = $mailboxDir;
$storageDir = $memberListStorageDir;

# Display status indicators on the screen
$|=1;
print "Memberlist filing started......";

# Start the logfile entry
open(LOGFILE,">>".$session_log_filename);
($sec, $min, $hour, $monthday,$month,$year,$weekday,$yearday,$dst) =
   localtime(time);
$month++;
print LOGFILE "MEMBER LIST FILING ($month\/$monthday $hour:$min): ";

#  ------ Load the group information ------

# Open the group information file
open(INFO_FILE,$group_filename) || die "Can't open: $group_filename?\n";

# Read the group info and store it appropriately
$list_counter = 0;
while(<INFO_FILE>)
{
   # Increment the list counter
   $list_counter++;

   # Read the group information
   chop;
   ($group_num,$group_name,$server_address,$server_type)
        = split(/\,/,$_);

   # Contruct the group address (group_name@server_domain)
   $server_address =~ /(\b[a-z0-9\.\-\+\_]*)@([a-z0-9\-\_\.]*\b)/;
   $server_domain = $2;
   $group_address = $group_name."\@".$server_domain;
   # Convert the group name and address intc lower case
   $group_name =~ tr/A-Z/a-z/;
   $group_address =~ tr/A-Z/a-z/;
   $server_address =~ tr/A-Z/a-z/;

   # Store group name and the server address
   $group_address_list{$group_address} = $server_address;

   # Store the server address
   $server_address_list{$server_address} = $group_address;

}

# Close the group info file
close INFO_FILE;


# ----- File the group memberlists ------

# Get a list of the message files
opendir(MSGDIR,$incomingDir) || die "Can't open $incomingDir\n";
@filelist = readdir(MSGDIR);
closedir(MSGDIR);

# Set the counters used for status checking
$filed_memberlists=0;  # Number of member lists files in the session

# Loop through the message files
foreach $filename (@filelist)
{
    # Skip the . and .. entries in the directory
    if(($filename eq ".") || ($filename eq ".."))
    {
        next;
    }

    # Open the next message file
    open(MSGFILE,$incomingDir.$filename) || die "Can't Open: $filename\n";

    # Find the header information needed to process the message
    while((<MSGFILE>))
    {
       # Stop if this is the break between the header and the message
       last if /^\n/;

       # Remove the end of line marker
       chop;

       # Check for the 'From:' header
       if(/^From:/i)
       {
           # Start with the baseline
           $FromLine = $_;

           # Collect the entire To Line
           while(<MSGFILE>)
        {
            # If the header is finished stop
            last if /^\n/;
            chop;

            # Stop if another header is encountered
            last if /^.*:/;

            # Add the new line to the previous line
          $FromLine = $FromLine.$_;
        }

        # Extract the list server addresses
        $FromLine =~ s/^From://;     # Remove the To:
        $FromLine =~ s/\s//;         # Remove the spaces
        $FromLine =~ tr/A-Z/a-z/;    # Make everything lowercase


        # Store the addresses in the address list
        @source_address_list = split(/,/,$FromLine);
    }
}

# Close the message file
close MSGFILE;

# Process each entry in the Destination address list
$StudyAddressFound = 0;
$GroupAddressFound = 0;
$ServerAddressFound = 0;

# Process each entry in the source address list
foreach $Address (@source_address_list)
{
    # Extract the address
    $Address =~ /(\b[a-z0-9\.\-\+\_]*@[a-z0-9\-\_\.]*\b)/;
    $Address = $1;

    # Determine if this the name of a group in the study
    $Server_Address = $group_address_list{$Address};
    if($Server_Address ne "")
    {
        $GroupAddressFound = 1;
    }

    # Determine if this is the study address
    if($Address =~ /egroup/i)
    {
        $StudyAddressFound = 1;
    }

    # Determine if the source was a known list server
    $Temp = $server_address_list{$Address};
  if($Temp ne "")
    {
       $ServerAddressFound = 1;
    }
```

```
        }

        # If this is a member list message then file it
        if($ServerAddressFound == 1)
        {
             # Move the identified memberlist
             system("mv $incomingDir$filename $storageDir$filename");

             # Count the list as filed
             $filed_memberlists++;
        }
    }

    # Create the log output
    print LOGFILE $filed_memberlists," msgs stored in $storageDir\n";
    close LOGFILE;

    # Write the summary entry for primary memberlist processing
    open(SUMMARYFILE,">>".$session_summary_filename);
    print SUMMARYFILE "$filed_memberlists";
    close SUMMARYFILE;

    # Display status information on the screen
    print "...finished ($filed_memberlists msgs filed)\n";
```

C-6: Message Processing (Daily Archive Creation)

```
#!/usr/local/bin/perl
# ---------------------------------------------
# This perl script
# processes the filtered messages from the
# secondary sample and files then in the
# outgoing secondary messages file for
# transfer to the archives.
# ---------------------------------------------
# Created: 6/23/97
# Brian Butler
# ---------------------------------------------

# Include the description of the directories and log files
require "filenames.pl";

# Filenames
$group_filename = $primary_list_filename;

# Archive file information
$archive_file_extension = $primary_archive_extension;

# Identify the directories
$incomingDir = $primaryMsgStorageDir;
$storageDir = $primaryMsgArchiveDir;

# Display a status indicator to the screen
$| = 1;
print "Primary message processing started....";

# Open the log file
open(LOGFILE,">>".$session_log_filename);
($sec, $min, $hour, $monthday,$month,$year,$weekday,$yearday,$dst) =
   localtime(time);
$month++;
print LOGFILE "PRIMARY MSGS PROC ($month\/$monthday $hour:$min): ";

# Read in the group list
open(INFILE,$group_filename) || die "Can't open: $group_filename\n";
$list_counter = 1;
while(<INFILE>)
{
   # Read the group information
   chop;
   ($group_num,$group_name,$server_address,$server_type)
        = split(/\,/,$_);

   # Convert into lower case
   $group_name =~ tr/A-Z/a-z/;

   # Store group name and initialize the message counter
   $group_list{$group_name} = 0;

   # Increment the list counter
   $list_counter++;
```

```
}

# Process the messages
opendir(MSGDIR,$incomingDir) || die "Can't open $incomingDir\n";
@filelist = readdir(MSGDIR);
closedir(MSGDIR);

# Process each of the message files
$msgcounter = 0;
foreach $filename (@filelist)
{
    # Skip the . and .. entries in the directory
    if(($filename eq ".") || ($filename eq ".."))
    {
        next;
    }

    # Count the message
    $msgcounter++;

    # Construct the archival filename
    @statistics = stat($incomingDir.$filename);
    ($sec, $min, $hour, $monthday,$month,$year,$weekday,$yearday,$dst) =
         localtime(@statistics[9]);
    $month = $month + 1;
    $archive_filename = sprintf("%.2d%.2d%.2d",$month,$monthday,$year);
    $archive_filename = $archive_filename.$archive_file_extension;

    # Record the date that the message was received
    $date_received = "$month,$monthday,$year,@statistics[9]";

    # Open the relevent archive file
    open(MSGOUTFILE,">>".$storageDir.$archive_filename)
       || die "Can't Open: $archive_filename\n";

    # Write the date received to the output file
    print MSGOUTFILE "received-date: $date_received\n";

    # Open the message file
    open(MSGFILE,$incomingDir."/".$filename) || die "Can't Open:
$filename\n";

    # Read and process the message
    while((<MSGFILE>))
    {
       # If this is the end of the header then stop
       if(/^\n/)
       {
            print MSGOUTFILE "\n";
            last;
       }
       # If the current line begins with a blank then assume it is not
       # a header line
       elsif(/^\s/)
       {
           # Save the remainder of the line for additional processing
           $LineContent = $_;
```

```
}
else      # If the current line is a header field then identify it
{
    # Save the line header
    $LineHeader = $_;
    $LineHeader =~ /^(\S*):/;
    $LineHeader = $1;
    $LineHeader =~ tr/A-Z/a-z/;  # Make it lowercase

    # Save the remainder of the line for additional processing
    $LineContent = $_;
    $LineContent =~ s/^\S*:/ /;
}

# Clean up the current line
$LineContent =~ s/\s//;       # Remove the spaces
$LineContent =~ tr/A-Z/a-z/;  # Make everything lowercase

# Process the line type based on the type of header it is
# If it is a from, to, or cc process the addresses
if($LineHeader eq "to" ||
   $LineHeader eq "from" ||
   $LineHeader eq "cc")
{
    # Store the addresses in the address list
    @address_list = split(/,/,$LineContent);

  # Find each address and encrypt it
    $Encrypted_content = "";
  $first_address = 1;
    foreach $Address (@address_list)
    {
         # Extract the address
         $OrigAddress = $Address;
         $Address =~ /(\b[a-z0-9\.\-\+\_]*@[a-z0-9\-\_\.]*\b)/;
       $Address = $1;

         # Extract the group name
         $Address =~ /(\b[a-z0-9\.\-\+\_]*)@([a-z0-9\-\_\.]*\b)/;
         $AddressName = $1;
         $AddressDomain = $2;

         # Determine if this the name of a group in the study
         $Temp = $group_list{$AddressName};
         if($Temp ne "")
         {
           # Add commas as necessary
             if($first_address == 1)
             {
                 $first_address = 0;
         }
         else
         {
               $Encrypted_content = $Encrypted_content.",";
           }

           # Store the address
```

```
            $Encrypted_content = $Encrypted_content.$Address;

            # Count the message
            $group_list{$AddressName}++;
            }
          else  # Otherwise encrypt the address
            {
                # Encrypt the name (actually just the first 8 chars)
                $Ename = crypt($AddressName,$encrypt_salt);

                # Encrypt the domain (actually just the first 8 chars)
            $Edomain = crypt($AddressDomain,$encrypt_salt);

              # Add commas as necessary
                if($first_address == 1)
                {
                    $first_address = 0;
            }
            else
            {
                  $Encrypted_content = $Encrypted_content.",";
              }

                # Store the encrypted info
            $Encrypted_content = $Encrypted_content."$Ename$Edomain";

               # Write the address and encrypted address to the
             # people list change file
             $people_list{$AddressName."\@".$AddressDomain} =
                   $Ename.$Edomain.",".$OrigAddress;
            }
        }

      # Write the modified Line to the Output file
        print MSGOUTFILE $LineHeader,": ",$Encrypted_content,"\n";
    }
    elsif($LineHeader eq "subject" ||
          $LineHeader eq "date" ||
        $LineHeader eq "egroup")
    {
        print MSGOUTFILE $LineHeader,": ",$LineContent;
    }
}

# After the header is processed then write the rest of the message to
#  output file.
while((<MSGFILE>))
{
    print MSGOUTFILE $_;
}

# End the archival message by writing the message separator
print MSGOUTFILE $archive_msg_separator,"\n";

# Close Message file and the MsgOutput File
close MSGFILE;
close MSGOUTFILE;
```

```
    # Delete the original message file
    system("rm $incomingDir$filename");

}

# Save the people list from this session
open(NAMEFILE,">>".$name_change_filename) ||
     die "Can't open: $name_change_filename\n";
foreach $key (keys(%people_list))
{
     print NAMEFILE $key,",",$people_list{$key},"\n";
}
close NAMEFILE;

# Create the log output
print LOGFILE "$msgcounter archived msgs\n";
close LOGFILE;

# Display status information on the screen
print "...finished ($msgcounter msgs processed)\n";
```

C-7: Membership List Processing (Daily Archive Creation)

```
#!/usr/local/bin/perl
# ----------------------------------------------
# This perl script
# processes the filtered messages from the
# which are believed to be memberlists.
# Processor status messages are destroyed
# Memberlist messages are archived.
# ----------------------------------------------
# Created: 6/23/97
# Brian Butler
# ----------------------------------------------

# Include the definitions of filenames, logs, etc.
require "filenames.pl";

# Set the group information filename
$group_filename = $primary_list_filename;

# Archive setup
$archive_file_extension = $list_archive_extension;

# Identify the directories
$incomingDir = $memberListStorageDir;
$storageDir = $listArchiveDir;
$problemDir = $listProblemDir;
$statusDir = $statusMsgDir;

# Open the group summary file
open(GROUPFILE,">>".$group_size_filename);

# Open the log file
open(LOGFILE,">>".$session_log_filename);
($sec, $min, $hour, $monthday,$month,$year,$weekday,$yearday,$dst) =
   localtime(time);
$month++;
print LOGFILE "MEMBER LIST PROC ($month\/$monthday $hour:$min): ";

# Display a status indicator to the screen
$| = 1;
print "Member list processing started....";

# ----- Read in the group list ------
open(INFILE,$group_filename) || die "Can't open: $group_filename\n?";
$list_counter = 1;
while(<INFILE>)
{
   # Read the group information
   chop;
   ($group_num,$group_name,$server_address,$server_type)
        = split(/\,/,$_);

   # Convert into lower case
   $group_name =~ tr/A-Z/a-z/;
   $server_address=~ tr/A-Z/a-z/;
   # Store group name and initialize the size counter
   $group_list{$group_name} = 0;

   # Store the server name
   $server_list{$server_address} = $server_type;

   # Increment the list counter
   $list_counter++;
}

# Process the lists
opendir(LISTDIR,$incomingDir) || die "Can't open $incomingDir??\n";
@filelist = readdir(LISTDIR);
closedir(LISTDIR);

# Process each of the message files
$msgcounter = 0;
$processed_lists = 0;
$status_msgs = 0;
$problem_msgs = 0;
foreach $filename (@filelist)
{
    # Skip the . and .. entries in the directory
    if(($filename eq ".") || ($filename eq ".."))
    {
        next;
    }

    # Count the message
    $msgcounter++;

    # Construct the archival filename
    @statistics = stat($incomingDir.$filename);
    ($sec, $min, $hour, $monthday,$month,$year,$weekday,$yearday,$dst) =
         localtime(@statistics[9]);
    $month = $month + 1;
    $archive_filename = sprintf("%.2d%.2d%.2d",$month,$monthday,$year);
    $archive_filename = $archive_filename.$archive_file_extension;

    # Store the date that the message was received
    $date_received = "$month,$monthday,$year,@statistics[9]";

    # Open the relevent archive file
    open(LISTOUTFILE,">>".$storageDir.$archive_filename)
       || die "Can't Open: $archive_filename\n";

    # Open the message file
    open(LISTFILE,$incomingDir.$filename) || die "Can't Open:
$filename???\n";

    # Read and process the message
    while((<LISTFILE>))
    {
       # If this is the end of the header then stop
       if(/^\n/)
       {
        last;
   }
   elsif(/^From:/i)
   {
       $FromLine = $_;
   }
   elsif(/^Date:/i)
   {
       $DateLine = $_;
   }
   elsif(/^Subject:/i)
   {
       $SubjectLine = $_;
   }
   # If the current line is not a header line
   elsif(/^\s/)
   {
       # Store the current line based on the prior line
       if($PriorLine =~ /^From:/i)
       {
           $FromLine = $FromLine.$_;
       }
       elsif($PriorLine =~ /^Subject:/i)
       {
           $SubjectLine = $SubjectLine.$_;
       }
   }
   $PriorLine = $_;
}

# First use the from line to make sure it is a known server
$FromLine =~ s/^From:/ /i; # Remove the header marker
$FromLine =~ s/\s//;       # Remove the spaces
$FromLine =~ tr/A-Z/a-z/;  # Make everything lowercase

# Extract the address
$FromLine =~ /(\b[a-z0-9\.\-\+\_]*@[a-z0-9\-\_\.]*\b)/;
$ServerAddress = $1;

# If the message is from known list server then process it
$ServerType = $server_list{$ServerAddress};
if($ServerType ne "")
{
   # Clean up the subject line
   $SubjectLine =~ s/^Subject:/ /i; # Remove the header marker
   $SubjectLine =~ tr/A-Z/a-z/;  # Make everything lowercase

   # Process the subject line
   $SubjectLine =~ s/file://;
   $SubjectLine =~ s/review//;
   $SubjectLine =~ s/majordomo results//;
   $SubjectLine =~ s/ list//;
   $SubjectLine =~ s/\s//g;
   $SubjectLine =~ s/\"//g;
   $SubjectLine =~ s/^\://;

   # If this is as status message the quit
   if(($SubjectLine =~ /output/i) || ($SubjectLine =~ /re:/i))
   {
        $MessageType = "status";
   }
   else
   {
        $MessageType = "list";
        $group_name = $SubjectLine;
   }
}
else
{
    $MessageType = "unknown-server";
}


# If the group name is unknown then find it in the message
if($group_name eq "" && $MessageType eq "list")
{
     while((<LISTFILE>))
     {
          # Look for a line with 'who' in it
          # [This is based on the assumption that majordomo is
          #  is the only listserver type which does not always
          #  include the group_name in the subject line
          if(/who/)
          {
              $group_name = $_;
          $group_name =~ s/>//g;
          $group_name =~ s/who//;
          $group_name =~ s/\s//g;

          # Drop out of the reading loop
          last;
         }
  }
}

# Deal with the WAYCOOL-L list exception
if($group_name eq "wycool-l")
{
     $group_name = "waycool-l";
}

# Based on the message type process the rest of the file
$Temp2 = $group_list{$group_name};


if(($MessageType eq "list") && ($Temp2 ne ""))
{

  # Clean up the date line
  $DateLine =~ s/Date://i;

  # Write the group information
  print LISTOUTFILE "received-date: $date_received\n";
  print LISTOUTFILE "date: $DateLine";
```

```
      print LISTOUTFILE "group: $group_name\n";

      # If this is a listproc the skip the header information
      if($ServerType eq "Listproc")
      {
          while((<LISTFILE>))
          {
             # Find the list label
           if(/list of non-concealed subscribers/)
           {
               last;
           }
        }
      }

      # Process the address information  - filtering ouut extra comments
      $member_count = 0;
      while((<LISTFILE>))
      {
          # If the line contains an address and does not contain
          # the word 'is' and does not start with a '*' or a ' '
          if((/\@/) && (/^[^\*]/) && !(/ is /) && !(/^   /))
          {
              print LISTOUTFILE $_;

            # Count the member of the list
            $member_count++;
          }
          elsif(/total/i) # Extract the total information
          {
              print LISTOUTFILE "TOTALINFO: ",$_;
          }
      }

      # Save the group information & Separator string
      print LISTOUTFILE $archive_msg_separator,"\n";

      # Store the group information in the group_info_summary file
      print GROUPFILE "$group_name,$month,$monthday,$year,$member_count\n";

      # Count the file as a processed list
      $processed_lists++;

    }
    # If it is an unknown group then store it as a problem-msg
    elsif(($MessageType eq "list") && ($Temp2 eq ""))
    {
         # Copy the file to the problem-file directory
         system("cp $incomingDir$filename $problemDir$filename");
       $problem_msgs++;
    }
    # If it is a problem message save it for manual handling
    elsif($MessageType eq "unknown-server")
    {
         # Copy the file to the problem-file directory
         system("cp $incomingDir$filename $problemDir$filename");
       $problem_msgs++;
    }
    # If it is a status message don't do anything
    elsif($MessageType eq "status")
    {
         # Copy the file to the problem-file directory
         system("cp $incomingDir$filename $statusMsgDir$filename");
       $status_msgs++;
    }

    # Close Message file and the MsgOutput File
    close LISTFILE;
    close LISTOUTFILE;

    # Delete the original message file
    system("rm $incomingDir$filename");
}

# Close the group summary file
close GROUPFILE;

# Create the log output
print LOGFILE
   "($processed_lists list msgs, $status_msgs status msgs, $problem_msgs
problem msgs)\n";
close LOGFILE;

# Display status information on the screen
print "...finished ($msgcounter msgs processed)\n";
```

C-8: Membership Archive Creation

```
#!/usr/local/bin/perl
# ---------------------------------------------
# This perl script converts a daily membership
# archive file into the group archive file.
# ---------------------------------------------
# Created: 8/17/97
# Brian Butler
# ---------------------------------------------

# ----------- Define Filenames & Setup Parameters ----------------

# Include the filename and encryption info definitions
require "filenames.pl";

# Define the name of the filenames
$input_file_directory = ":Daily Archives:";
$group_file_directory = ":Group Archives:";
$people_db_filename = ":people.db";
$group_db_filename = ":groups.db";
$people_update_filename = "new_people.txt";
$member_data_filename = "new_memberdata.txt";
$group_archive_extension = ".lar";   # List ARchive

# Redefine the end-of-line character to deal with the non-Mac files
$/ = "\r";

# -------------- Get and Sort the Input Filelist ------------------

# Process the messages
opendir(ARCHIVEDIR,$input_file_directory) || die "Can't open
$input_file_directory\n";
@filelist = readdir(ARCHIVEDIR);
closedir(ARCHIVEDIR);
@archive_files = sort {$a cmp $b} @filelist;

# Display the initial filelist & check to continue
print "Known archive files: ";
foreach $file (@archive_files) { print $file,"\n" };
print "<< Enter q to cancel (anything else to continue) >>\n";
if(getc() eq "q"){ exit };

# -------------------------- Open Files ---------------------------

# Open the people DB file
dbmopen(%PEOPLE_LIST,$people_db_filename,0666);

# Open the group DB file
dbmopen(%GROUP_LIST,$group_db_filename,0666);

# If necessary initialize the person counter
if(!(defined $PEOPLE_LIST{"NEXTID"}))
```

```
{
   $PEOPLE_LIST{"NEXTID"} = 1;
}

# Open the file for recording new individuals
open(PEOPLEFILE,">>".$people_update_filename) || die "Can't open:
$people_update_filename\n";

# ------------------- For Each Archive File... ------------------
foreach $filename (@archive_files)
{
     # ---------------- Create an archive profile ----------------
     # This is a list of group_name, received time, location
     # triplets.  Store this as a list of strings.
     # ------------------------------------------------------------
     open(INFILE,"<".$input_file_directory.$filename);
     while(<INFILE>)
          {
          # --- If this is the received-date: line (and hence the first
          # --- line of a group entry record it in the archive_profile
          if(/^received-date: /)
          {
                # Get the received date and time
                $position = tell(INFILE) - length($_);
                $received_date = $_;
                chop $received_date;

                $received_date =~ s/received-date: //;
                ($day,$month,$year,$time_value) = split(/\,/,$received_date);


                  # Get the sent date
                $sent_date = <INFILE>;

                # Get the group name
                $group_name = <INFILE>;
                chop $group_name;
                $group_name =~ s/group: //;

                # Construct the profile entry
                $profile_entry = $group_name.",".$time_value.",".$position;


                  # Add the profile entry to the profile
                push @archive_profile, ($profile_entry);
          }
     }

     # Sort the profile entries
     @archive_profile = sort @archive_profile;

     # Display the archive profile & check to continue
     print "Finished Creating the Profile for $filename: \n";
     print "It has: $#archive_profile entries\n";
     # print "<< Enter q to cancel (anything else to continue) >>\n";
     # if(getc() eq "q"){ exit };
     # ---------- For each entry in the profile.... ------------
     $entry_counter = 0;
     $total_entries = $#archive_profile + 1;
     foreach $entry (@archive_profile)
     {
          # Increment the entry counter and display a status msg
          $entry_counter++;
          print "Entry $entry_counter (out of $total_entries in $filename)\n"
                if $entry_counter % 10 == 0;

          # Retreive the group_entry information
          ($group_name,$time_value,$position) = split(/\,/,$entry);

          # Return to the group entry
          seek(INFILE,$position,0);

          # Read the group entry header information
          $received_dateline = <INFILE>;
          chop  $received_dateline;
          $sent_dateline = <INFILE>;
          $group_nameline = <INFILE>;
          chop  $group_nameline;

          # Read the membership data in from the archive file
          undef @raw_memberdata;
          $current_line = <INFILE>;
          while(!($current_line =~ /^1\-9=2\-8/))
          {
               # Store the current line and read the next one
               chop $current_line;
               push @raw_memberdata, ($current_line);
               $current_line = <INFILE>;
          }

          # Processes the memberdata
          undef @final_memberdata;
          @raw_memberdata = reverse @raw_memberdata;
          $current_line = pop @raw_memberdata;
          while(defined $current_line)
          {
               # If it is a total line add it to the output
               if($current_line =~ /^TOTALINFO/)
               {
                    push @final_memberdata, ($current_line."\n");
               }
               else  # Otherwise process the member info
               {
                    # Extract the e-mail address from the record
                    $current_line =~ tr/A-Z/a-z/;   # Make it lowercase.
                   $current_line =~ /(\b[a-z0-9\.\-\+\_]*@[a-z0-9\-
\_\.]*\b)/;
                         $Adcress = $1;
                    $Address =~ /(\b[a-z0-9\.\-\+\_]*)@([a-z0-9\-\_\.]*\b)/;
                    $AddressName = $1;
                    $AddressDomain = $2;

                    # Encrypt the e-mail address (first 8 chars of id + 8
                    #  chars of domain)
                    $EncryptedID =
      crypt($AddressName,$encrypt_salt).crypt($AddressDomain,$encrypt_salt);

                   # Using the memberDB determine whether this
                   # is a known individual
                   # Based on the group name detemine the copy number
                   $person_id = $PEOPLE_LIST{$EncryptedID};
                    if(!(defined $person_id))
                    {
                        # Update the people list DB
                        $person_id = $PEOPLE_LIST{"NEXTID"};
                        $PEOPLE_LIST{$EncryptedID} = $person_id;

                        $PEOPLE_LIST{"NEXTID"}++;

                        # Add the person data to the new_person data file
                        print PEOPLEFILE
                             "$person_id,$EncryptedID,$AddressDomain\n";
                     }

        # At this point all the information necessary for the member-date-
file
        # is available

               # Store the memberdata
               push @final_memberdata, ($person_id."\n");
               }

               # Get the next line
               $current_line = pop @raw_memberdata;
          }

          # Construct group filename and Open the group archive file
          $group_archive = $GROUP_LIST{$group_name}.$group_archive_extension;
          open(GROUPFILE,">>".$group_file_directory.$group_archive)
              || die "Can't open: $group_file_directory$group_archive\n";

          print $group_name," -- ",$group_archive,"\n";

          # Write the group entry header info
          print GROUPFILE $received_dateline,"\n";
          print GROUPFILE $group_nameline,"\n";

          # Write the group memberdata & total info
          foreach $entry (@final_memberdata) { print GROUPFILE $entry };

          # Write the end of entry marker to the group file
          print GROUPFILE "$archive_msg_separator\n";

          # Close the group file
          close GROUPFILE;
     }

     # Close the archive file
     close INFILE;
     # Remove the archive profile
     undef @archive_profile;

     # Display the archive profile & check to continue
     print "Finished Processing $filename: \n";
     # print "<< Enter q to cancel (anything else to continue) >>\n";
     # if(getc() eq "q"){ exit };
}

# ---------------------- Close Files -------------------------
close PEOPLEFILE;
dbmclose(%PEOPLE_LIST);
dbmclose(%GROUP_LIST);
```

C-9: Message Archive Creation

```
#!/usr/local/bin/perl
# ---------------------------------------------
# This perl script converts a daily message
# archive file into the group archive file.
# ---------------------------------------------
# Created: 8/17/97
# Brian Butler
# ---------------------------------------------

# ------------ Define Filenames & Setup Parameters ---------------

# Include the filename and encryption info definitions
require "filenames.pl";

# Define the name of the filenames
$input_file_directory = ":Daily Archives:";
$group_file_directory = ":Group Message Archives:";
$group_db_filename = ":groups.db";
$group_archive_extension = ".mar";   # Message ARchive

# Redefine the end-of-line character to deal with the non-Mac files
# $/ = "\r";

# -------------- Get and Sort the Input Filelist -----------------

# Process the messages
opendir(ARCHIVEDIR,$input_file_directory) || die "Can't open
$input_file_directory\n";
@filelist = readdir(ARCHIVEDIR);
closedir(ARCHIVEDIR);
@archive_files = sort {$a cmp $b} @filelist;

# Display the initial filelist & check to continue
print "Known archive files: \n";
foreach $file (@archive_files) { print $file,"\n" };
print "<< Enter q to cancel (anything else to continue) >>\n";
if(getc() eq "q"){ exit };

# ------------------------- Open Files --------------------------

# Open the group DB file
```

```
dbmopen(%GROUP_LIST,$group_db_filename,0666);

# ------------------- For Each Archive File... ------------------
foreach $filename (@archive_files)
{
     # ----------------- Create an archive profile ----------------
     # This is a list of group_name, received time, location
     # triplets.  Store this as a list of strings.
     # -----------------------------------------------------------

     # Start with an empty profile
     undef @archive_profile;

     open(INFILE,"<".$input_file_directory.$filename);
     while(<INFILE>)
     {
       # --- If this is the received-date: line (and hence the first
          # --- line of a group entry) process the message and
          # --- record it in the archive_profile
          if(/^received-date: /)
          {
                # Get the received date and time
                $position = tell(INFILE) - length($_);
                $received_date = $_;
                chop $received_date;

                $received_date =~ s/received-date: //;
                ($day,$month,$year,$time_value) = split(/\,/,$received_date);

                # Sift through the header info and find the group info
                undef %group_names;
                $current_line = <INFILE>;
                while($current_line ne $/)
                {
                     # If this is an "egroup: " line then get the group
                     # name from it
                     if($current_line =~ /^egroup: /)
                     {
                          # Remove the header & store group name
                          $current_line =~ s/^egroup: //;
                          chop $current_line;
                          $group_names{$current_line} = 1;
                     }
                     # Check addresses for the group name
                     elsif (($current_line =~ /^from: /) ||
                            ($current_line =~ /^to: /) ||
                            ($current_line =~ /^cc: /))
                     {
                          # Remove the header
                          $current_line =~ s/^from: //;
                          $current_line =~ s/^to: //;
                          $current_line =~ s/^cc: //;

                          # Search the address for group names
                          # !! This routine assumes that messages
                          # !! have been processed so that only
                          # !! the addresses of groups in the study
```

```
                           # !! are unencrypted.
                           @addresses = split("\,",$current_line);
                           foreach $entry (@addresses)
                           {
                                # If the address is one of groups then get
                                # the group name
                                if($entry =~ /\@/)
                                {
                                    # Get the group name
                                    $entry =~ /^(.*)\@/;
                                    $group_names{$1} = 1;
                                }
                           }
                      }
                      $current_line = <INFILE>;
                 }

                 # Based on the list of group_names construct the profile
                 # entry.  Specifically, if there are is more than one
                 # group name flag the primary group and indicate a possible
                 # crosspost (0 = Normal message, 1 = first name in cross
post,
                 # 2 = secondary names encountered in crosspost)
                 @names = keys(%group_names);
                 if($#names == 0)
                 {

                       # Construct and store the profile entry
                       $profile_entry =
                            $names[0].",".$time_value.",".$position.",0";

                     push @archive_profile, ($profile_entry);
                 }
                 elsif(@names == ())     # No group names....
                 {
                      # Display an error message
                      print "Unknown group: $time_value\n";

                      # Construct and record a entry for unknown-group
                      $profile_entry =
                         "unknown-group,".$time_value.",".$position.",0";
                      push @archive_profile, ($profile_entry);
                 }
                 else # Multiple group names
                 {
                      # Display a message
                      print "Possible cross-post: $time_value\n";

                    # Construct and store the profile entry for the first name
                 $profile_entry =
$names[0].",".$time_value.",".$position.",1";
                      push @archive_profile, ($profile_entry);

                      # Process the other names
                      for ($i = 1; $i <= $#names; $i++)
                      {
```

```
                    # Construct and store the profile entry for the first
name
                         $profile_entry =
                              $names[i].",".$time_value.",".$position.",2";
                          push @archive_profile, ($profile_entry);
                     }
                }
           }
      }

     # Sort the profile entries
     @archive_profile = sort @archive_profile;

     # Display the archive profile & check to continue
     print "Finished Creating the Profile for $filename: \n";
     print "It has: $#archive_profile entries\n";
     # foreach $entry (@archive_profile) { print $entry,"\n" };
     # print "<< Enter q to cancel (anything else to continue) >>\n";
     # if(getc() eq "q"){ exit };

     # ---------- For each entry in the profile.... ------------
     $total_entries = $#archive_profile + 1;
     $last_groupname = "";
     $i = 0;

     # As long as there are entries left
     for ($i = 0; $i <= $#archive_profile; $i++)
     {
          # Display a status msg
          print "Entry ",$i+1,"
             (out of $total_entries in $filename)\n" if ($i + 1) % 10 == 0;

          # Retrieve the message entry information
          ($group_name,$time_value,$position,$msg_type) =
                  split(/\,/,$archive_profile[$i]);

          # If the current group different from the last group then
          # open the new group file
          if($group_name ne $last_groupname)
          {
               # Close the prior group archive file (if it is open)
               close GROUPFILE if GROUPFILE;

               # Construct group filename and Open the group archive file
               $group_archive =
                  $GROUP_LIST{$group_name}.$group_archive_extension;
               open(GROUPFILE,">>".$group_file_directory.$group_archive)
                    || die "Can't
open:$group_file_directory$group_archive\n";
          }


          # Return to the message entry
          seek(INFILE,$position,0);

          # Read & store the msg header information
          undef @header;
```

```
$current_line = <INFILE>;
while($current_line ne $/)
{
     push @header,($current_line);
     $current_line = <INFILE>;
}

# Read & store the msg body information
undef @msg_body;
$current_line = <INFILE>;
while($current_line ne $archive_msg_separator.$/)
{
     push @msg_body,($current_line);
     $current_line = <INFILE>;
}

# Process the header
undef @final_header;
foreach $line (@header)
{
     # Prepare to write the header in the following order:
     # Received-date
     # Egroup (name)   [Generated - not copied]
     # Date
     # Subject
     # From
     # To
     # CC
     # Alert Notes (normal, possible cross-post, etc)
     if($line =~ /^received-date: /)
        { push @final_header,("01".$line); }
     if($line =~ /^date: /) { push @final_header,("03".$line); }
     if($line =~ /^subject: /) { push @final_header,("04".$line); }
     if($line =~ /^from: /) { push @final_header,("05".$line); }
     if($line =~ /^to: /) { push @final_header,("06".$line); }
     if($line =~ /^cc: /) { push @final_header,("07".$line); }
}

# Add the alert: entry
push @final_header, ("08alert: normal\n") if $msg_type == 0;
push @final_header, ("08alert: crosspost\n") if $msg_type > 0;

# Add the egroup: entry and arrange the header
push @final_header,("02egroup: $group_name\n");
@final_header = sort @final_header;

# Write the header the header and message to the group file
foreach $line (@final_header)
{
     # Clean up the header line
     $line =~ s/..//;
     chop $line;

     # Write the header line to the archive file

     print GROUPFILE $line,"\n";
}
```

```
           # Write the header ending to the archive file
           print GROUPFILE "\n";

           # Write the message body
           foreach $line (@msg_body)
           {
               # Clean up the line
               chop $line;

               # Write the msg line to the archive file
               print GROUPFILE $line,"\n";
           }

           # Write the end of entry marker to the group file
           print GROUPFILE "$archive_msg_separator\n";

     }

     # Close the last group file
     close GROUPFILE;

     # Close the archive file
     close INFILE;

     # Display the archive profile & check to continue
     print "Finished Processing $filename: \n";
     # print "<< Enter q to cancel (anything else to continue) >>\n";
     # if(getc() eq "q"){ exit };
}
# ------------------------ Close Files ---------------------------
dbmclose(%GROUP_LIST);
```

Datafile creation

C-10: Membership datafile creation (Intermediate file creation)

```
#!/usr/local/bin/perl
# ----------------------------------------------
# This perl script creates a membership total
# series of total for each group.
# ----------------------------------------------
# Created: 8/17/97
# Brian Butler
# ----------------------------------------------

# ------------ Define Filenames & Setup Parameters ----------------

# Define the name of the filenames
$groupsum_filename = "group-totals.txt";
$group_file_directory = ":Group Archives:";
$people_db_filename = ":people.db";
$group_db_filename = ":groups.db";
$group_archive_extension = ".lar";   # List Archive
```

```
# -------------- Get and Sort the Input Filelist ------------------

# Get the list of group archive files
opendir(ARCHIVEDIR,$group_file_directory) || die "Can't open
$group_file_directory\n";
@filelist = readdir(ARCHIVEDIR);
closedir(ARCHIVEDIR);
@archive_files = sort {$a cmp $b} @filelist;

# ------------------- For Each Archive File... -------------------
$lastdate = "";
foreach $filename (@archive_files)
{
     # For each group create a summary of the number of entries per
     # day.  Record this in the group summary file as a list
     # of group name, date, time entries
     open(INFILE,"<".$group_file_directory.$filename);
                              while(<INFILE>)
                              {

           # Get the received date and time
           $received_date = $_;
           chop $received_date;
           $received_date =~ s/received-date: //;
           ($month,$day,$year,$time_value) = split(/\,/,$received_date);

           # Get the group name
           $group_name = <INFILE>;
           chop $group_name;
           $group_name =~ s/group: //;

           # Determine the copy number
           if($lastdate eq $month.$day.$year) { $copy_number++ }
           else { $copy_number = 1 };
           $lastdate = $month.$day.$year;

           # Determine the number of member known
           $member_count = 0;
           $current_line = <INFILE>;
           while(!($current_line =~ /^1\-9=2\-8/) && !($current_line =~
/^TOTALINFO/))
           {
               $member_count++;
               $current_line = <INFILE>;
           }
           while(!($current_line =~ /^1\-9=2\-8/))
           {
               $member_count++;
               $current_line = <INFILE>;
           }

           # Construct the profile entry
           $group_entry =


      $group_name.",".$member_count.",".$month."\/".$day."\/".$year.",".$time
_value.",".$copy_number;
```

```
           # Add the profile entry to the profile
           push @archive_profile, ($group_entry);

     }

     # Close the archive file
     close INFILE;

     # Display a status message
     print "Processing: $filename\n";

}

# --------------------- Open the summary file ---------------------

# Open the file for recording the group summary
open(GROUPSUMFILE,">".$groupsum_filename) || die "Can't open:
$groupsum_filename\n";

# -------------- Write the data summary to a file -----------------

foreach $entry (@archive_profile)
{
    print GROUPSUMFILE "$entry\n";
}

# ----------------------- Close Files --------------------------
close GROUPSUMFILE;
```

C-11: Membership datafile creation (Final datafile creation)

```
#!/usr/local/bin/perl
# -------------------------------------
# This script creates a time series
# data file from the group-totals.txt
# file.
# Series are create as fixed length
# sequences of with missing values "."
# where appropriate.
# -------------------------------------

# -------- Filenames and Process Parameters ----------
$input_filename = "group-totals.txt";
$output_filename = "member-data.txt";
$data_points = 58;

# -------- Create date index -----------
$month = 7;
$year = 97;
for($day = 4, $i = 0; $day <= 31; $day++, $i++)

{
    $date_key = sprintf "%d%.2d%.2d",$year,$month,$day;
    $date_index{"$month\/$day\/$year"} = $i;
    $label_index{$i} = "$month\/$day\/$year";
}

$month = 8;
for($day = 1; $day <= 31; $day++, $i++)

{
    $date_key = sprintf "%d%.2d%.2d",$year,$month,$day;
    $date_index{"$month\/$day\/$year"} = $i;
    $label_index{$i} = "$month\/$day\/$year";
}

# -------- Read and process the data file -----------
open(INFILE,"<".$input_filename);
open(OUTFILE,">".$output_filename);
$last_groupname = "Group Name";
while(<INFILE>)
{
     # Parse the line
     chop;
     ($group_name, $people_count, $date_entry, $time_value, $copy) =
        split("\,",$_);

     # If this is a new group then store the old group info
     if($group_name ne $last_groupname)
     {
          # Determine whether this group should be ommited
          # (These groups are known to contain seriously
          #  flawed data).
          if(($last_groupname ne "internetstikaipar0.60ls!") &&
             ($last_groupname ne "ethics-ldearn") &&
              ($last_groupname ne "ethics-luga"))
           {
               # Determine whether this is a fixed group or variable
               $fixed = 0;
               $max = 0;
               $min = 99999;
               foreach $data (@entry)
               {
                    if(($data > $max) && ($data ne ".")) { $max = $data; }
                    if(($data < $min) && ($data ne ".")) { $min = $data; }
               }
               if($max == $min) { $fixed = 1; }

               # Write the prior group info to the outfile
               $last_groupname =~ s/\"//g;
               foreach $date (0..$#entry)
               {
                  # Write the data to the data file
                  print OUTFILE
"$last_groupname$date,$last_groupname,$date,$entry[$date],$fixed\n";
               }
           }

           # Reset the entry info
           for($i = 0; $i <= $data_points; $i++) { $entry[$i] = "."; }
      }

      # Store the current line's info
      if($copy == 1)
      {
           $entry[$date_index{$date_entry}] = $people_count;
      }

      # Record the current group
      $last_groupname = $group_name;
}

close OUTFILE;
close INFILE;
```

C-12: Message datafile creation (Intermediate datafile creation)

```
#!/usr/local/bin/perl
# ----------------------------------------------
# This perl script creates a total
# series of total for each group.
# ----------------------------------------------
# Created: 8/17/97
# Brian Butler
# ----------------------------------------------

# ------------ Define Filenames & Setup Parameters ----------------

# Define the name of the filenames
$groupsum_filename = "group-msg-totals.txt";
$group_file_directory = ":Group Message Archives:";
$people_db_filename = ":people.db";
$group_db_filename = ":groups.db";
$group_archive_extension = ".lar";   # List ARchive

# -------------- Get and Sort the Input Filelist ------------------

# Get the list of group archive files
opendir(ARCHIVEDIR,$group_file_directory) || die "Can't open
$group_file_directory\n";
@filelist = readdir(ARCHIVEDIR);
closedir(ARCHIVEDIR);
@archive_files = sort {$a cmp $b} @filelist;

# --------------------- Open the summary file ---------------------

# Open the file for recording the group summary
open(GROUPSUMFILE,">".$groupsum_filename) || die "Can't open:
$groupsum_filename\n";


# ------------------- For Each Archive File... -------------------
$lastdate = "";
foreach $filename (@archive_files)
{
     # For each group create a summary of the number of entries per
     # day.  Record this in the group summary file as a list
     # of group name, date, time entries
     open(INFILE,"<".$group_file_directory.$filename);
     $total_sender_count = 0;

      while(<INFILE>)
      {
          if(/received-date: /)
          {
               # Get the received date and time
               $received_date = $_;
               chop $received_date;
               $received_date =~ s/received-date: //;
               ($month,$day,$year,$time_value) = split(/\,/,$received_date);
          # Get the group name
          $group_name = <INFILE>;
          chop $group_name;
          $group_name =~ s/egroup: //;

          # Remove the date
          $date_sent = <INFILE>;

          # Get the message subject
          $subject = <INFILE>;

          # Rough marker of message type
          if($subject =~ /re:/i) { $type = 1; }
          else { $type = 0; }

          # Process the rest of the header info for the message
          $current_line = <INFILE>;
          while(!($current_line =~ /^\n/))
          {
              # Process the sender information
              if($current_line =~ /^from:/i)
              {
                    $sender_id = $current_line;
                    chop $sender_id;
                    $sender_id =~ s/from://i;
                    $sender_id =~ s/abmF1QH4PEr.EabmF1QH4PEr.E,?//;
                     # Eliminate @ from sender field

                    # Store the sender identifier and count the message
                    if(exists $sender_list{$sender_id})
                    {
                          # Get the sender number
                          $sender_number = $sender_list{$sender_id};
                    }
                    else
                    {
                          # Get the sender number
                          $sender_number = $total_sender_count;
                          $sender_list{$sender_id} = $sender_number;
                          $total_sender_count++;
                    }
              }
              $current_line = <INFILE>;
          }

          # Construct the profile entry
          $msg_entry =
             $group_name.",".$type.",".$month."\/".$day."\/".$year.",";
          $msg_entry = $msg_entry.$time_value.",".$sender_number;

          # Add the profile entry to the profile
          push @archive_profile, ($msg_entry);
    }
}

# Close the archive file
close INFILE;
```

```
     # Display a status message
     print "Processing: $filename\n";


     # -------------- Write the data summary to a file -----------------
     foreach $entry (@archive_profile)
     {
         print GROUPSUMFILE "$entry\n";
     }


     # Clean up the group record structures
     undef @sender_list;
     undef @archive_profile;

}

# ------------------------ Close Files --------------------------
close GROUPSUMFILE;
```

C-13: Message datafile creation (Final datafile creation)

```
#!/usr/local/bin/perl
# --------------------------------------
# This script creates a time series
# data file from the group-totals.txt
# file.
# Series are create as fixed length
# sequences of with 0's where there are
# missing values.
# --------------------------------------

# -------- Filenames and Process Parameters ----------
$input_filename = "group-msg-totals.txt";
$output_filename = "msg-data.txt";
$data_points = 0;

# -------- Create date index -----------
$month = 7;
$year = 97;
for($day = 4, $i = 0; $day <= 31; $day++, $i++)

{
    $date_key = sprintf "%d%.2d%.2d",$year,$month,$day;
    $date_index{"$month\/$day\/$year"} = $i;
    $label_index{$i} = "$month\/$day\/$year";

    # Count the data points
    $data_points++;
}

$month = 8;
for($day = 1; $day <= 31; $day++, $i++)

{
    $date_key = sprintf "%d%.2d%.2d",$year,$month,$day;
    $date_index{"$month\/$day\/$year"} = $i;
    $label_index{$i} = "$month\/$day\/$year";

    # Count the data points
    $data_points++;
}

$month = 9;
for($day = 1; $day <= 30; $day++, $i++)

{
    $date_key = sprintf "%d%.2d%.2d",$year,$month,$day;
    $date_index{"$month\/$day\/$year"} = $i;
    $label_index{$i} = "$month\/$day\/$year";

    # Count the data points
    $data_points++;
}

$month = 10;
```

```
for($day = 1; $day <= 31; $day++, $i++)

{
    $date_key = sprintf "%d%.2d%.2d",$year,$month,$day;
    $date_index{"$month\/$day\/$year"} = $i;
    $label_index{$i} = "$month\/$day\/$year";

    # Count the data points
    $data_points++;
}

$month = 11;
for($day = 1; $day <= 30; $day++, $i++)
{
    $date_key = sprintf "%d%.2d%.2d",$year,$month,$day;
    $date_index{"$month\/$day\/$year"} = $i;
    $label_index{$i} = "$month\/$day\/$year";

    # Count the data points
    $data_points++;
}

print "N = $data_points\n";

# -------- Read and process the data file ----------
open(INFILE,"<".$input_filename);
open(OUTFILE,">".$output_filename);
$last_groupname = "Group Name";

# Set the counters
$msg_count = 0;
$reply_count = 0;
$new_count = 0;

# Initialize data record
for($i = 0; $i <= $data_points; $i++)
{
    $entry[$i] = "0,0,0,0,0";
}

while(<INFILE>)
{
     # Parse the line
     chop;
     ($group_name, $msg_type, $date_entry, $time, $sender_id) =
            split("\,",$_);

     # If this is a new group then store the old group info
     if($group_name ne $last_groupname)
     {
          # Compute the sender count for the day
          @sender_ids = keys %sender_list;
          $sender_count = $#sender_ids + 1;

          # Compute a Gini coefficient for concentration of contribution
          #  activity
          $gini = 0;
          foreach $sender (keys %sender_list)
          {
              $gini += abs(($sender_list{$sender} / $msg_count) -
                         (1/$sender_count));
          }
          if($sender_count <= 1)
          {
              $gini = ".";
          }
          else
          {
              $gini = (1 / (2 * (1 - (1/$sender_count)))) * $gini;
          }

          # Store the daily stats
          $entry[$date_index{$last_date_entry}] =

$msg_count.",".$reply_count.",".$new_count.",".$sender_count.",".$gini;

          # Reset the daily records
          $msg_count = 0;
          $reply_count = 0;
          $new_count = 0;
          undef %sender_list;

          # Determine whether this group should be ommited
          # (These groups are known to contain seriously
          #  flawed data).
          if(($last_groupname ne "unknown-group") &&
             ($last_groupname ne "Group Name") &&
             ($last_groupname ne "ethics-l"))
          {
               # Write the prior group info to the outfile
               $last_groupname =~ s/\"//g;
               foreach $date (1..$#entry)
               {
                  print OUTFILE
                  "$last_groupname$date,$last_groupname,$date,$entry[$date]\n
                  ";
               }
          }

          # Reset the entry info
          for($i = 0; $i <= $data_points; $i++) { $entry[$i] = "0,0,0,0,0"; }
      }
      elsif($last_date_entry ne $date_entry)
      {
          # Compute the sender count for the day

          @sender_ids = keys %sender_list;
          $sender_count = $#sender_ids + 1;

          # Compute a Gini coefficient for concentration of contribution
          #activity
          $gini = 0;
          foreach $sender (keys %sender_list)
          {
              $gini += abs(($sender_list{$sender} / $msg_count) -
(1/$sender_count));
          }
          if($sender_count <= 1)
          {
              $gini = ".";
          }
          else
          {
              $gini = (1 / (2 * (1 - (1/$sender_count)))) * $gini;
          }

          # Store the daily stats
          $entry[$date_index{$last_date_entry}] =

$msg_count.",".$reply_count.",".$new_count.",".$sender_count.",".$gini;

          # print "$last_date_entry $date_index{$last_date_entry} --
".$msg_count.",".$reply_count.",".$new_count;

          # Reset the summary data structures
          $msg_count = 0;
          $reply_count = 0;
          $new_count = 0;
          undef %sender_list;
      }

      # Store the current line's info
       # Message Count
       $msg_count++;

       # New and Reply Count
       if($msg_type == 1)  { $reply_count++; }
       else { $new_count++; }

       # Daily sender profile
            if(exists $sender_list{$sender_id})
       {
            # Count the message as part of this senders messages
            $sender_list{$sender_id}++;
        }
        else
        {
            # Count the message as part of its thread
            $sender_list{$sender_id} = 1;
        }

      # Record the current group
      $last_groupname = $group_name;
      $last_date_entry = $date_entry;
}

close OUTFILE;
close INFILE;
```

C-14: Secondary membership and message datafile creation

```
#!/usr/local/bin/perl
# ----------------------------------------------
# This perl script profiles messages
# ----------------------------------------------
# Created: 8/17/97
# Brian Butler
# ----------------------------------------------

# -------------------- Main Body of the code starts here --------------------

# ------------ Define Filenames & Setup Parameters ----------------

# Define the name of the filenames
$msgsum_filename = "msg-features.txt";
$thread_filename = "thread-features.txt";
$group_filename = "group-features.txt";
$group_file_directory = ":Group Message Archives:";
$people_db_filename = ":people.db";
$group_db_filename = ":groups.db";
$group_archive_extension = ".lar";   # List ARchive

# The frequency of message status message
$status_freq = 10;


# -------------- Get and Sort the Input Filelist -----------------

# Get the list of group archive files
opendir(ARCHIVEDIR,$group_file_directory) || die "Can't open
$group_file_directory\n";
@filelist = readdir(ARCHIVEDIR);
closedir(ARCHIVEDIR);
@archive_files = sort {$a cmp $b} @filelist;

# --------------------- Open the summary file ---------------------

# Open the file for recording the message summary
open(MSGFILE,">".$msgsum_filename) || die "Can't open: $msgsum_filename\n";

# Open the file for recording the thread summary
open(THREADFILE,">".$thread_filename) || die "Can't open:
$thread_filename\n";

# Open the file for recording the group summary
open(GROUPFILE,">".$group_filename) || die "Can't open: $group_filename\n";

# ------------------- For Each Archive File... ------------------
$lastdate = "";
$group_count = 0;
foreach $filename (@archive_files)
{
     # Count the group
     $group_count++;
     print "Processing group #",$group_count," (",$filename,")\n";
     # For each group create a profile of the group messages
     open(INFILE,"<".$group_file_directory.$filename);
                        $current_line = <INFILE>;
     $msg_count = 0;
     $thread_count = 0;
     $sender_count = 0;

     # This loop is setup to process one message in each iteration
     while($current_line)
                               {
          if($current_line =~ /received-date: /)
          {
              # Count the message
              $msg_count++;

              # Get the received date and time
              $received_date = $current_line;

              # Get the group name
              $current_line = <INFILE>;
              $group_name = $current_line;
              chop $group_name;
              $group_name =~ s/egroup: //;

              # Remove the date
              $current_line = <INFILE>;
              $date_sent = $current_line;

              # Get the message subject
              $current_line = <INFILE>;
              $subject = $current_line;
              $subject =~ s/^subject://;
              $subject =~ tr/ / /s;
              chop $subject;

              # Determine the message type (Reply or start)
              if($subject =~ /re:/i) { $msg_type = 1; }
              else { $msg_type = 0; }

              # Create the thread identifier
              $thread_string = $subject;
              $thread_string =~ s/\s//g;
              $thread_string =~ s/re://i;
              $thread_string = substr($thread_string,0,40);

              # Store the thread identifier and count the message in its
topic
              if(exists $thread_list{$thread_string})
              {
                   # Get the thread id
                   $thread_id = $thread_list{$thread_string};

                   # Count the message as part of its thread
                   $thread_length_list{$thread_id}++;
              }
              else
```

```
               {
                    # Get the thread id
                    $thread_id = $thread_count;
                    $thread_list{$thread_string} = $thread_id;
                    $thread_count++;

                    # Count the message as part of its thread
                    $thread_length_list{$thread_id} = 1;
               }

               # Process the header info for the message
               while(!($current_line =~ /^\n/))
               {
                    # Process the sender information
                    if($current_line =~ /^from:/i)
                    {
                          $sender_id = $current_line;
                          chop $sender_id;
                          $sender_id =~ s/from://i;
                          $sender_id =~ s/abmFlQH4PEr.EabmFlQH4PEr.E,?//;
                          # Eliminate @ from sender field

                          # Store the sender identifier and count the message
                          if(exists $sender_list{$sender_id})
                          {
                                # Get the sender number
                                $sender_number = $sender_list{$sender_id};

                                # Count the message as part of this senders
messages
                                $sender_volume_list{$sender_number}++;

                          }
                          else
                          {
                                # Get the sender number
                                $sender_number = $sender_count;
                                $sender_list{$sender_id} = $sender_number;
                                $sender_count++;

                                # Count the message as part of its thread
                                $sender_volume_list{$sender_number} = 1;
                          }

                          # Count the senders replys and starts
                          if($msg_type == 1)   # If it is a reply
                          {
                                if(exists $reply_list{$sender_number})
{$reply_list{$sender_number}++;}
                                else {   $reply_list{$sender_number} = 1;   }
                          }
                          else
                          {     if(exists $new_list{$sender_number})
{$new_list{$sender_number}++;}
                                else {   $new_list{$sender_number} = 1;   }
                          }
                    }
```

```
                   $current_line = <INFILE>;
              }

              # Process the message content
              while(!($current_line =~ /1-9=2-8=3-7=4-6=5/))
              {
                  # Read the next line
                  $current_line = <INFILE>;
              }
          }
          $current_line = <INFILE>;
     }

     # Store the total message count
     $message_total = $msg_count;

     # Close the archive file
     close INFILE;

     # Store the thread info
     foreach $key (keys %thread_length_list)
         {   print THREADFILE "$key,$group_name,$thread_length_list{$key}\n";
}

     # Compute and store the group information
     # Determine the GINI coefficient for the group
     @sender_keys = keys %sender_volume_list;
     $sender_count = $#sender_keys + 1;
     $gini = 0;
     $gini2 = 0;
     foreach $key (keys %sender_volume_list)
     {
         $gini += abs(($sender_volume_list{$key} / $msg_count) -
(1/$sender_count));

         if($msg_count != $sender_count)
         {
             $gini2 +=
               abs((($sender_volume_list{$key} - 1) / ($msg_count -
$sender_count)) - (1/$sender_count));
         }
     }
     if($sender_count <= 1)
     {
         $gini = ".";
         $gini2 = ".";
     }
     else
     {
         $gini = (1 / (2 * (1 - (1/$sender_count)))) * $gini;
         $gini2 = (1 / (2 * (1 - (1/$sender_count)))) * $gini2;
     }

     # Determine the % of activity account for by top participants
     @sorted_senders = (sort {$b <=> $a} (values %sender_volume_list));
     $two_count = ".";
     $five_count = ".";
     if($#sorted_senders >= 1)
     {
          $two_count = ($sorted_senders[0] + $sorted_senders[1]) /
$msg_count;
          print "---------- $sorted_senders[0] + $sorted_senders[1] --------
\n";
     }

     if($#sorted_senders >= 4)
     {
          $five_count = ($sorted_senders[0] + $sorted_senders[1] +
$sorted_senders[2] +
                               $sorted_senders[3] + $sorted_senders[4]) /
$msg_count;
     }

     # Count the reply and new message senders
     @reply_keys = keys %reply_list;
     $reply_count = $#reply_keys + 1;
     @new_keys = keys %new_list;
     $new_count = $#new_keys + 1;

     print "@sorted_senders\n";
     printf "Msg Count: %d, Top two: %.2f, Top Five:
%.2f\n",$msg_count,$two_count,$five_count;
     printf "GINI: %.2f\n",$gini;
     printf "GINI2: %.2f\n",$gini2;
     printf "Reply Senders: %d         New Senders:
%d\n",$reply_count,$new_count;


     print GROUPFILE

"$group_name,$sender_count,$gini,$gini2,$two_count,$five_count,$reply_count,$
new_count\n";

     # Compute the group diversity
     # $tsum = 0;
     # foreach $key (keys %thread_length_list)
     #     {   $tsum += ( $thread_length_list{$key} / $msg_count)**2;   }

     # $ssum = 0;
     # foreach $key (keys %sender_volume_list)
     #    {   $ssum += ($sender_volume_list{$key} / $msg_count)**2;    }

     # $tdiversity = 1 - $tsum;
     # $sdiversity = 1 - $ssum;

# -------------------------------------------------------------

     # For each group create a profile of the group messages
     open(INFILE,"<".$group_file_directory.$filename);
                        $current_line = <INFILE>;
     $msg_count = 0;
     $thread_count = 0;
     # This loop is setup to process one message in each iteration
     while($current_line)
                               {
          # Initialize Message Line Count
          $msg_length = 0;
          $lines = 0;

          if($current_line =~ /received-date: /)
          {
              # Count the message
              $msg_count++;

              # Get the received date and time
              $received_date = $current_line;
              chop $received_date;
                                                       $received_date =~
s/received-date: //;
              ($month,$day,$year,$time_value) = split(/\,/,$received_date);
              $date_string = $month."/".$day."/".$year;

              # Get the group name

                                        $current_line = <INFILE>;
              $group_name = $current_line;
              chop $group_name;
              $group_name =~ s/egroup: //;

              # Remove the date
              $current_line = <INFILE>;
              $date_sent = $current_line;
              $date_sent =~ s/^date://;
              $date_sent =~ tr/ //s;
              chop $date_sent;

              # Get the message subject & construct a marker of message type
              $current_line = <INFILE>;
              $subject = $current_line;
              $subject =~ s/^subject://;
              $subject =~ tr/ / /s;
              chop $subject;

              # Determine the thread identifier
              $thread_string = $subject;
              $thread_string =~ s/\s//g;
              $thread_string =~ s/re://i;
              $thread_string = substr($thread_string,0,40);
              $thread_id = $thread_list{$thread_string};

              # Determine the age of the thread_message
              if(exists $thread_age_list{$thread_id})
              {
                   $thread_age_list{$thread_id}++;
                   $thread_age = $thread_age_list{$thread_id};
              }
              else
              {
                   $thread_age_list{$thread_id} = 0;
                   $thread_age = $thread_age_list{$thread_id};
              }

              # Determine the message type (Reply or start)
              if($subject =~ /re:/i) { $msg_type = 1; }
              else { $msg_type = 0; }

              # Determine if the message is a digest message
              if($subject =~ /digest\s?/) { $digest_msg = 1; }
              else { $digest_msg = 0; }

              # Process the header info for the message
              while(!($current_line =~ /^\n/))
              {
                   # Process the sender
                   if($current_line =~ /^from:/i)
                   {
                         # Find the sender number
                         $sender_id = $current_line;
                         chop $sender_id;
                         $sender_id =~ s/from://i;
                         $sender_id =~ s/abmF1QH4PEr.EabmF1QH4PEr.E,//; #
Eliminate @ from sender field
                         $sender_number = $sender_list{$sender_id};
                   }
                   $current_line = <INFILE>;
              }

              # Process the message content
              while(!($current_line =~ /1-9=2-8=3-7=4-6=5/))
              {
                  # Count (word in the) the current line
                  # If there is only one non-space string on the count it as
a word
                  $_ = $current_line;
                  if(s/\S+\s//g == 1)
                  {
                       $msg_length += 1;
                  }
                  else
                  {
                       # Remove punctuation
                       $_ = $current_line;

                       # Count all the words
                       $msg_length += (s/\w+\W//g);
                  }
                  undef @token_list;

                  # Count the lines
                  $lines++;

                  # Read the next line
                  $current_line = <INFILE>;
              }

              # $wordsxlines = $msg_length / $lines;
              # print "Msg Length (Words): $msg_length in $lines lines
($wordsxlines)\n";

              # Create thread volume percentage measure
              $thread_typicality = $thread_length_list{$thread_id} /
$message_total;

              # Create the person volume percentage measure
              $sender_typicality = $sender_volume_list{$sender_number} /
$message_total;

              # Construct the entry msg record and store it
              $msg_record = $group_count."-
".$msg_count.",".$group_name.",".$msg_length;
              $msg_record =
$msg_record.",".$thread_length_list{$thread_id}."\n";
              push @group_records,$msg_record;
          }
          $current_line = <INFILE>;
     }

     # Close the archive file
     close INFILE;

     # Clean up the thread lists
     undef %sender_volume_list;
     undef %thread_length_list;
     undef %sender_list;
     undef %thread_list;
     undef %thread_age_list;
     undef %new_list;
     undef %reply_list;

     # Write the records to the summary file
     while($#group_records >= 0)
     {
          $current_record = shift @group_records;
          printf MSGFILE $current_record;
     }

     # print "Type q to quit ";
     # if(getc() eq "q"){ exit };
}

# ------------------------ Close Files ---------------------------
close MSGFILE;
close THREADFILE;
close GROUPFILE;
```

C-15: Measure and SAS datafile creation

```
/* ------------------------------- */
/* This script combines the raw    */
/* data files for group size and   */
/* group message volume to create  */
/* a time series data set.         */
/* ------------------------------- */

/* Read the group membership data  */
DATA RawMmbr(compress=YES);

/* Create variables for the data entry index values and group name */
LENGTH MINDEX $ 30 GNAME $ 25;

/* Read the data */
INFILE '/afs/andrew.cmu.edu/gsia/nets/stx/egroup/raw-data/member-data.txt'
delimiter=',' stopover;

INPUT MINDEX $ GNAME $ DAYNO SIZE ENTRY EXIT FIXED STABLE;

/* Convert record key string values to uppercase */
GNAME = UPCASE(GNAME);
MINDEX = UPCASE(MINDEX);

* Remove the variables that are never used;
DROP FIXED STABLE;

/* Read the group message data */
DATA RawMsg(COMPRESS=YES);

/* Create variables for the data entry index values and group name */
LENGTH MINDEX $ 30 GNAME $ 25;

/* Read the data */
INFILE '/afs/andrew.cmu.edu/gsia/nets/stx/egroup/raw-data/msg-data.txt'
delimiter=',' stopover;

INPUT MINDEX $ GNAME $ DAYNO MSGCNT RECNT NEWCNT SNDRCNT SNDRCNC;

if SNDRCNT = 0 then
   SNDRCNC = 0;

if SNDRCNT = 1 then
   SNDRCNC = 0;

/* Convert record key string values to uppercase */
GNAME = UPCASE(GNAME);
MINDEX = UPCASE(MINDEX);

/* Read the group message data */
DATA RawLen(COMPRESS=YES);
```

```
/* Create variables for the data entry index values and group name */
LENGTH MINDEX $ 30 GNAME $ 25;

/* Read the data */
INFILE '/afs/andrew.cmu.edu/gsia/nets/stx/egroup/raw-data/length-data.txt'
delimiter=',' stopover;

INPUT MINDEX $ GNAME $ DAYNO TMSGCNT TMSGLEN;

/* Convert record key string values to uppercase */
GNAME = UPCASE(GNAME);
MINDEX = UPCASE(MINDEX);

/* Create measure of average message length */
if TMSGCNT = 0 THEN AMSGLEN = 0;
ELSE AMSGLEN = TMSGLEN / TMSGCNT;

/* Get rid of the total message count (duplicated) */
DROP TMSGCNT;

/* Read the group characteristics data */
DATA RawGrp;

/* Create variables for the data entry index values and group name */
LENGTH GNAME $ 25 SUBLBL $ 20 SRVADD $ 30 SRVTYPE $ 15 DESC $ 10 NOTES $ 10;

/* Read the data */
INFILE '/afs/andrew.cmu.edu/gsia/nets/stx/egroup/raw-data/sample-data.txt'
delimiter=',' stopover;

INPUT GID GNAME $ SUBID SUBLBL $ SUBTYPE SRVID SRVADD $ SRVTYPE $ DESC $
MSGONLY CHANGED REJECT NOTES $;

/* Convert record key string values to uppercase */
GNAME = UPCASE(GNAME);

* Drop Unused variables
DROP SUBLBL GID SUBTYPE SRVID SRVADD SRVTYPE DESC NOTES CHANGED REJECT
SRVTYPE;

DATA RawCnt(compress=YES);

/* Create variables for the data entry index values and group name */
LENGTH MINDEX $ 30 GNAME $ 25;

/* Read the data */
INFILE '/afs/andrew.cmu.edu/gsia/nets/stx/egroup/raw-data/thread-data.txt'
delimiter=',' stopover;

INPUT MINDEX $ GNAME $ DAYNO MC RC NC TC NTC RTC TA SC THD TGD;
drop mc rc nc;

/* Convert record key string values to uppercase */
GNAME = UPCASE(GNAME);
MINDEX = UPCASE(MINDEX);

proc sort data=RawCnt;
```

```
    by MINDEX;

/* Sort the member and message data sets for merging */
proc sort data=RawMmbr;
by MINDEX;

proc sort data=RawMsg;
by MINDEX;

proc sort data=RawLen;
by MINDEX;

/* Create the complete dataset by merging the rawdatasets, dropping     */
/* the data entry index, and placing 0's for create missing msg values  */
data RawTemp(COMPRESS=YES);
    merge RawLen RawMsg RawMmbr RawCnt;
    by MINDEX;
    drop MINDEX;

    IF MSGCNT = .  then MSGCNT = 0;
    IF RECNT = .  then RECNT = 0;
    IF NEWCNT = .  then NEWCNT = 0;
    if SNDRCNT = . then SNDRCNT = 0;
    if SNDRCNC = . then SNDRCNC = 0;

    IF TMSGCNT = . then TMSGCNT = 0;
    IF TMSGLEN = . then TMSGLEN = 0;
    IF AMSGLEN = . then AMSGLEN = 0;

    IF TC = . then TC = 0;
    IF TA = . then TA = 0;
    IF RTC = . then RTC = 0;
    IF NTC = . then NTC = 0;

    IF SC = . then SC = 0;
    IF THD = . then THD = 0;
    IF TGD = . then TGD = 0;

    /* Repair disjointed group data gliches */
    if GNAME='MAJORDOMORESULTS' and DAYNO<4 THEN GNAME = 'AGWOMEN-L';
    if GNAME='AGWOMEN-L' and DAYNO<4 THEN DELETE;
    if GNAME='MAJORDOMORESULTS' THEN DELETE;

    if GNAME='WYCOOL-L' and DAYNO<4 THEN GNAME = 'WAYCOOL-L';
    if GNAME='WAYCOOL-L' and DAYNO<4 THEN DELETE;
    if GNAME='WYCOOL-L' THEN DELETE;

    /* Fill in missing size data for early data collection error */
    if GNAME='GEOLIST' and DAYNO<4 THEN SIZE=99;
    if GNAME='GEOLIST' and DAYNO=4 THEN ENTRY=0;

    if GNAME='MODPOETRY' and DAYNO<4 THEN SIZE=53;
    if GNAME='MODPOETRY' and DAYNO=4 THEN ENTRY=0;

    if GNAME='VT-CTE' and DAYNO<4 THEN SIZE=33;
    if GNAME='VT-CTE' and DAYNO=4 THEN ENTRY=0;
    /* Remove the Harvard domain name change impact */
    if GNAME='HUSN-LIST' and DAYNO = 55 THEN EXIT = 0;
    if GNAME='HUSN-LIST' and DAYNO = 55 THEN ENTRY = 0;
    if GNAME='HMA-BDC-LIST' and DAYNO = 55 THEN EXIT = 0;
    if GNAME='HMA-BDC-LIST' and DAYNO = 55 THEN ENTRY = 0;

    /* Correct for a system change (mass entry and/or exodus)*/
    if GNAME='APSTAT-L' and DAYNO = 24 THEN SIZE = 429;
    if GNAME='APSTAT-L' and DAYNO = 24 THEN EXIT = 0;
    if GNAME='APSTAT-L' and DAYNO = 25 THEN ENTRY = 0;

    /* Remove groups with major data problems */
    IF GNAME='CETEFL-L' THEN DELETE;       /* 0 Group size */
    if GNAME='FELL-WALKERS' then DELETE;  /* 0 Size group */
    if GNAME='ONTAG' then DELETE;  /* Sporatic data availability */
    if GNAME='ASSODEAN-L' then DELETE; /* Known message collection problems
*/

    /* Remove groups with for which data stops and does not return */
    if GNAME = 'DDD' THEN DELETE;
    if GNAME = 'DUKE-FEMILIST' THEN DELETE;
    if GNAME = 'MOMENTUM' THEN DELETE;
    if GNAME = 'E2-FANFIC' THEN DELETE;
    if GNAME = 'JSCOPE' THEN DELETE;
    if GNAME = 'MIDDLESCHOOL-LIST' THEN DELETE;
    if GNAME = 'PEA' THEN DELETE;
    if GNAME = 'WEBADVER' THEN DELETE;
    if GNAME = 'GENDER-TEACHING' THEN DELETE;
    if GNAME = 'WISE-L' THEN DELETE;             /* !!!! Final 3 days !!!! */
    if GNAME = 'FIG-TEACHERS' THEN DELETE;
    if GNAME = 'HINDI-T' THEN DELETE;
    if GNAME = 'ACGSTAFF-L' THEN DELETE;
    if GNAME = 'ACCOUNTING-DISCUSS' THEN DELETE;
    if GNAME = 'AESP-NET' THEN DELETE;
    if GNAME = 'CENLA-YOUTH' THEN DELETE;
    if GNAME = 'CTESL-L' THEN DELETE;
    if GNAME = 'CYBER-FREEDOM' THEN DELETE;
    if GNAME = 'GYO-L' THEN DELETE;
    if GNAME = 'KN-NEWBIEHELP' THEN DELETE;
    if GNAME = 'KN-POETRY' THEN DELETE;
    if GNAME = 'KN-YOUTH' THEN DELETE;
    if GNAME = 'LIYSF' THEN DELETE;
    if GNAME = 'OCCTA' THEN DELETE;
    if GNAME = 'PEACEMAKERS' THEN DELETE;
    if GNAME = 'PRACSPAN' THEN DELETE;
    if GNAME = 'WOFL-B5' THEN DELETE;
    if GNAME = 'WTOG-B5' THEN DELETE;
    if GNAME = 'SUPERGUY' THEN DELETE;

    /* Remove groups with a significant internal series missing */
    if GNAME = 'NATM' THEN DELETE;
    if GNAME = 'NEEDLEWORK' THEN DELETE;
    if GNAME = 'NEEDLEWORK-DIGEST' THEN DELETE;

    if GNAME = 'CP' then DELETE;
```

```
/* Sort the group attribute file and the combined (temporary) raw data file
*/
proc sort data=RawTemp;
    by GNAME;

proc sort data=RawGrp;
    by GNAME;

/* Combine the temporary and group attribute file into the raw data set */
data RawData(compress=yes);
    merge RawTemp RawGrp;
    by GNAME;

    /* Remove the record if there is no day # (i.e. no member or msg data) */
    if DAYNO ~= .;

    /* Drop unuseful data values */
    DROP SRVADD DESC NOTES CHANGED STABLE MSGONLY REJECT SRVTYPE SUBLBL;

/* Sort the complete dataset by group and date */
proc sort data=RawData;
by GNAME DAYNO;

/* Fill the missing size, exit, and entry values with extrapolated values */
proc expand data=RawData out=ExtTemp(compress=yes) method=JOIN;
BY GNAME;
ID DAYNO;
convert SIZE;
convert exit = exitld1 / method = none transform=(lead);
convert entry = entryld1 / method = none transform=(lead);

proc sort data=ExtTemp;
   BY GNAME DAYNO;

/* Provide values for missing entry and exit data */
data ExtData(COMPRESS=YES);    /* Membership, Message, and Group Dataset */
set ExtTemp;

/* Fill in Missing Values for one or two periods at end of data run */
/* Extrapolation by EXPAND won't do this                             */
SIZE1 = lag(SIZE);
SIZE2 = lag2(SIZE);
if DAYNO = 147 and SIZE = . THEN SIZE = SIZE1;
if DAYNO = 148 and SIZE = . THEN
    DO;
       IF SIZE1 ~= . THEN SIZE = SIZE1;
       ELSE SIZE = SIZE2;
    END;
if DAYNO = 149 and SIZE = . THEN
    DO;
       IF SIZE1 ~= . THEN SIZE = SIZE1;
       ELSE IF SIZE2 ~= . THEN SIZE = SIZE2;
    END;

/* Determine the needed working values */
DSIZE = dif(SIZE);   /* Size change based on extrapolated size values */
ENTRY1 = lag1(ENTRY);
```

```
EXIT1 = lag1(EXIT);

/* If information is available about next value (after a missing value) */
/* Then spread the entries or exits across the two values.              */
IF EXIT = . and EXITLD1 ~= . THEN EXIT = FLOOR(EXITLD1 / 2);
IF EXIT1 = . AND EXIT ~= . THEN EXIT = CEIL(EXIT/2);

/* If the exit values is missing construct an EXIT value based on */
/* the change in the SIZE variable.                               */
If EXIT = . and DSIZE < 0 THEN EXIT = ABS(DSIZE);
ELSE IF EXIT = . THEN EXIT = 0;

/* If information is available about next value (after a missing value) */
/* Then spread the entries or exits across the two values.              */
IF ENTRY = . and ENTRYLD1 ~= . THEN ENTRY = FLOOR(ENTRYLD1 / 2);
IF ENTRY1 = . AND ENTRY ~= . THEN ENTRY = CEIL(ENTRY/2);

/* If the exit values is missing construct an ENTRY value based on */
/* the change in the SIZE variable.                                */
If ENTRY = . and DSIZE > 0 THEN ENTRY = ABS(DSIZE);
ELSE IF ENTRY = . THEN ENTRY = 0;

/* Remove the Working Values */
DROP DSIZE ENTRY1 EXIT1 ENTRYLD1 EXITLD1 SIZE1 SIZE2;

/* NOTES ON THE EXTRAPOLATION PROCEDURE */
/* First: The EXPAND proc is used to create values for internal */
/* missing size data. Linear extrapolation is used.             */
/* Second: End missing values for size are created by extending */
/*   the last known size out with no change. (Only done for     */
/*   values within 2 days of the end of the period).            */
/* Third: Missing exit and entry values are created by averging */
/*   the first known value after a missing value and placing the */
/*   floor value in the missing spot andthe ceiling result in the */
/*   following spot. (Only done for the last missing in a sequence */
/*   [ primarily to deal with single missing values ]            */
/* Fourth: Any remaining missing values for entry and exit are  */
/*   created based on extrapolated size data.

/* In preliminary data set only 24 data points (0.2%) are created */
/*  with the fourth step.                                         */

/* Remove the front portion of the data because it contains a     */
/* significant sequence of missing memebership data in all groups */
if DAYNO >= 19;

/* Create a scaled size variable */
RealSize = SIZE;
SIZE = SIZE/100;

/* Create a measure of entries and exits as a percentage of size */
PENTRY = (ENTRY/SIZE);
PEXIT = (EXIT/SIZE);
PPART = (MSGCNT/SIZE);

data MMGData1(COMPRESS=YES);
set ExtData;
```

```
by GNAME;

     /* Remove the mean of the sample */
     if GNAME = ' ' THEN DELETE;

     /* Create lagged and difference values for absolute values */
     DSIZE = dif(size) * 100;
     SIZE1 = lag1(size);
     MSGCNT1 = lag1(msgcnt);
     SNDRCNT1 = lag1(sndrcnt);
     THD1 = lag(THD);
     TGD1 = lag(TGD);
     SNDRCNC1 = lag1(sndrcnc);
     RECNT1 = lag1(recnt);
     NEWCNT1 = lag1(newcnt);
     EXIT1 = lag1(exit);
     ENTRY1 = lag1(entry);

     /* Clear the first value (which is pulled from the prior group) */
     if first.GNAME THEN
     DO;
        DSIZE=.;
        SIZE1=.;
        EXIT1=.;
        ENTRY1=.;
        MSGCNT1=.;
        RECNT1=.;
        NEWCNT1=.;
     END;

     /* Two moving average for messages */
     MSGMN1 = ((MSGCNT + MSGCNT1)/2);
     THDMN1 = ((THD + THD1) / 2);
     TGDMN1 = ((TGD + TGD1) / 2);

     /* Create interaction variables */
     sm1 = size1 * msgcnt1;
     sm = size1 * msgcnt;
     srm1 = size1 * recnt1;
     snm1 = size1 * newcnt1;
     smm1 = size1 * msgmn1;

     /* Drop the first time period in each group (which is missing data */
     /* for the lagged variables.)                                       */
     if FIRST.GNAME THEN DELETE;

     /* Drop the last time period due to data problems */
     if DAYNO = 150 THEN DELETE;

     /* Remove working values */
     DROP OBS _TYPE_ _FREQ_;


/*  ------ Create a data set from the embedded group data file ----- */
DATA RawEmb(compress=YES);

/* Create variables for the data entry index values and group name */
```

```
LENGTH GNAME $ 25;

/* Read the data */
INFILE '/afs/andrew.cmu.edu/gsia/nets/stx/egroup/raw-data/emb-group.txt'
delimiter=',' stopover;

INPUT GID EMBMARK GNAME $;

/* Convert record key string values to uppercase */
GNAME = UPCASE(GNAME);

/* Convert the markers for questionable groups to embedded groups */
if EMBMARK = 2 then EMBMARK = 1;

/* Sort the data file by group name */
proc sort data=RAWEMB;
   BY GNAME;

/* Sort the data file by group name */
proc sort data=mmgdata1;
   BY GNAME;

/* Create data set which includes a subset of the group data and the */
/* embedded group markers.                                           */
data net.MMGDATA(compress=YES);
    merge RawEmb mmgdata1;
    by GNAME;

    /* Remove the record if there is no day # (i.e. no member or msg data) */
    if DAYNO ~= .;
    if EMBMARK ~= .;

    /* Keep only a subset of the variables to reduce storage space */
    keep DAYNO EMBMARK ENTRY EXIT GNAME MSGCNT MSGCNT1 MSGMN1
         PENTRY PEXIT PPART SIZE SIZE1 RECNT RECNT1 NEWCNT NEWCNT1
         TC NTC RTC TA SC THD TGD TMSGLEN AMSGLEN THD1 THDMN1 TGD1 TGDMN1
         EXIT1 ENTRY1 SNDRCNT SNDRCNC SNDRCNT1 SNDRCNC1;

/* Display the list of variables in the main datafile */
proc contents short data=net.MMGData;
```

Appendix D: Human Subjects Review Documentation

This appendix contains documentation related to the Institutional Review Board (IRB) review of the Electronic Group Dynamics Study:

- The original IRB proposal
- Excerpts from E-mail correspondence summarizing IRB concerns
- An addendum to the proposal addressing the IRB concerns
- Excerpts from E-mail correspondence with an IRB reviewer expressing concerns
- A second addendum addressing the reviewers concerns

Carnegie Mellon University
Human Subjects Clearance Request

Date: 1/25/96 CMU Protocol No: ____________

New Request: [X] Renewal: [ ]

Principal Investigator(s): Brian Butler [Advisors: Robert Kraut (CS) and Kathleen Carley (SDS)]

P.I. Title/Degree: Phd Student in Information Systems Department: GSIA

Phone: 268-8740 E-mail: bb26@andrew.cmu.edu

Project Dates: (2/1/97 - 1/31/98)

Project Title: Electronic Group Dynamics Study

Name of the Experimenter(s): [Data collection and analysis will be conducted by the PI]

Brief Description of Research:

This study combines observation of public groups and simulation modeling to consider how groups form, grow, and die. Membership records and message archives, which serve as the basis for group observation, will be collected using publicly available facilities in Internet-based electronic mail groups.

1. How many subjects will be used in this study?

The sample will consist of 100 groups.

2. From what source do you plan to obtain subjects?

The groups sampled in this study are public Internet-based electronic mail groups. For more information see the attachment entitled 'Group Selection Methods and Criteria'.

3. Is there any benefit gained by the subject for participating? No

4. Will the subjects include any of the following? No [ ] Yes [X]

[ ] Fetuses [ ] Mentally Retarded
[ ] Hospitalized Patients
[ ] Institutionalized Patients
[ ] Mentally Disabled
[ ] Minors
[ ] Pregnant Women
[ ] Prisoners

Individuals from these populations may be members of the sampled groups. However, none of these populations are targeted in this study. Groups which focus solely on these populations will eliminated from the sample and no attempt will be made to distinguish individuals in these populations from other members of the sampled groups.

5. Degree of Physical Risk: [X] Negligible [ ] Mild [ ] Moderate [ ] High
6. Degree of Psychological Risk: [X] Negligible [ ] Mild [ ] Moderate [ ] High

---

Please submit each of the following with this Clearance Request Form:

1. A draft of the proposal or abstract
2. A clear definition of how the subjects will be utilized or how the experimental treatment will be administered
3. A copy of the "informed" consent form(s) which the subjects will be required to sign
4. An indication of how confidentiality/anonymity will be protected
5. The name(s) and address(es) of official(s) authorizing access to any subjects in cooperating institutions not under the direct control of Carnegie Mellon University
6. Risk/Benefit analysis

The Dynamics of Cyberspace: A Model of Public Goods Development in Electronic Groups

Brian Butler
bb26@andrew.cmu.edu
Graduate School of Industrial Administration
Carnegie Mellon University

"The net isn't 30 million people, it's tens of thousands of overlapping groups ranging from a few people to perhaps a couple of hundred thousand at the largest"

(O'Reilly, 1996)

Electronic groups are a prominent social structure in cyberspace. Firms use these groups to monitor public opinion and support customers. In both academic and professional communities electronic groups can support existing relationships, foster communication, and help individuals make new contacts. Investigating the development of electronic groups is an important part of understanding how the capabilities of the evolving information infrastructure are used. In addition, these groups also provide a valuable opportunity to study the general process of group development. This research begins with a public goods model of electronic groups that characterizes group development in terms of group membership and messages. Group membership is described as a connective public good that supports communication within a community. Group messages serve as a communal public good which is derived directly from the information contributed by individual group members. These two features in electronic groups are likely to evolve interdependently. Membership data and messages from a sample of naturally occurring electronic groups will be collected. Time series analysis will be used to test aspects of the interdependent public goods model. In addition, the dataset will serve as the basis for simulation modeling of group dynamics. The goal of this study is to combine empirical analysis and simulation to refine and test the interdependent public goods model of electronic group development.

Group Selection Methods and Criteria

For this study two different samples of electronic groups will be selected. The first sample consists of randomly selected existing groups which are either professionally or recreationally oriented. The second sample consists of new groups which are each paired with a topically comparable existing group. Here we will describe the methods and criteria used to select groups for inclusion in these samples.

Sample 1: Professional vs. Recreational Groups

This sample is constructed to provide a cross-section of the population of two common types of electronic groups. It will also enable us to consider the hypothesis that members of professional groups value time and information differently than members of recreational groups. The one hundred groups in this sample will be selected using the following procedure:

1. Topic selection

Using the list of topics provided by the on-line directory of Publicly Accessible Mailing Lists (PAML) five professional and five recreational topics will be randomly selected subject to the following criteria:

a. Each topic must have at least five groups listed in the PAML directory

b. Each topic should be generally identifiable as either professional or recreational [more focused selection related to this criteria will occur when individual groups are selected].

c. Each topic must not be a subset of a previously selected topic. If a topic shares more that 75% of its listed groups with a previously selected topic then the topic with the larger number of listed groups will be used.

d. Topics of a sensitive personal or political nature will not be included. Specifically topics relating to the following subjects will be excluded:

1. Erotica and other sexuality related topics

2. Support groups

2. Group Selection

For each of the topics selected above a list of possible groups will be constructed by combining the groups listed under the topic in PAML and the results of searching for the topic in LISZT and the ListofLists, two other on-line directories of publicly available electronic groups. From this list of possibilities groups will be selected subject to the following criteria:

a. Only groups which are clearly focused on either professional or recreational audiences will be considered

b. Newsletters and other broadcast-oriented, one-way mailing lists will not be included.

c. Moderated groups will not be included

d. Only groups with open membership will be selected. If a group has any type of explicit membership screening it will not be included in the sample.

e. Groups which restrict access to the membership information will not be included

f. Groups which have more than 10% of their membership concealed (i.e. included in their membership count, but not visible in their membership listing) will not be included.

g. Groups which are linked directly with a newsgroup (e.g. groups which mirror the contents of a USENET newsgroup) will not be included.

h. Groups which focus on sensitive personal or political subjects will not be included.

If, after this criteria is applied, there are less than 50 professional groups or 50 recreational groups in the sample then more topics will be randomly chosen and more groups selected until the desired sample size is achieved.

Sample 2: New vs. Existing Groups

This sample is constructed as a basis for systematically considering differences between new groups and older, existing groups. The groups in this sample will be selected using the following procedure:

1. New group identification and selection

Public announcements from the NEW-LIST mailing list and other USENET newsgroups will be reviewed to identify groups which meet the group selection criteria outlined above.

2. Matching group identification and selection

When an appropriate new group is identified a search of the above mentioned on-line directories of publicly accessible electronic groups (PAML, LISZT, and ListofLists) will be conducted to identify one or more existing lists which focus on topics similar to the new list. Each topically-comparable existing list will then be evaluated using the group selection criteria described above.

If a single suitable existing group is found then the new group/existing group pair will be added to the sample. If more than one suitable existing group if discovered then the new group and a single, randomly selected existing list will be added to the sample. If not suitable existing group can be located then the new group will not be included in the sample. This procedure will be repeated for each acceptable new group until there are 50 new group/existing group pairs in the sample.

Selection log

For both sample selection procedures a log will maintained documenting the topics and groups which were considered. A group will be listed in this log whether or not it was included in the study. For all unselected groups a reason for non-inclusion will be noted. For groups which are included information about the source and type of group will be recorded to verify that the group is publicly announced and accessible.

Data Collection Procedures

For each of the groups selected as part of this study the following data will be collected: daily membership listings and group messages. Here we will describe the facilities and procedures that will be used to collect this data. A later attachment, entitled "Data Handling and Storage Procedures", will describe the facilities and procedures that will be used to process the data once it has been received.

For this project an special Andrew account (egroup@andrew.cmu.edu; "Electronic Group Dynamics Study") has been created. Access to this account and the files in it is limited to the researcher responsible for data collection. This account will be used to send and receive all electronic mail associated with collecting the data for this study.

For each group that has been selected for this study the following procedure will be used to collect the raw data:

1. Subscribing to the group

Using the project account (egroup@andrew.cmu.edu) and the project name ("Electronic Group Dynamics Study" or "Egroup Study") a subscription to the group will be requested. This is done by sending a command, such as "SUBSCRIBE <Listname>", in an electronic mail message to the mail server that acts as the infrastructure for the group. This subscription results in several things occurring. First, the electronic mail address of the sender of the request (in this case the project account) is added to the group membership list. Second, in most groups a set of administrative materials which describe the group and its facilities are sent to the subscriber. For the purposes of this study these documents will be reviewed to ensure that no group policies are violated by the data collection. After this review, if the group remains part of the sample, the administrative documents will be archived for future reference.

2. Collecting group message data

Subscribing the project account (egroup) to a group results in the addition of the project electronic mail address to the group membership list. When this is done the project account will begin to receive message that are distributed to the group. These messages are the raw data, which will be processed and archived for analysis (for more detail on processing of the raw messages see the attachment entitled "Data Handling and Storage Procedures").

3. Collecting group membership data

The facilities for retrieving membership lists are commonly available within electronic groups. Requesting this information requires sending a command, such as "REVIEW <listname>" or "WHO <listname>" to the server that provides the infrastructure for the group. This facility, and the relevant commands are described clearly in the help documentation for the group, which in many cases is sent along with the administrative documents to new subscribers. At the same time, is it possible for group administrators (i.e. "list owners") to disable this capability. In this study, groups which have disabled this public access to their membership information are excluded. In addition, individuals members may often choose to conceal their identifying information (electronic mail address and name) so that it does not appear on the group's publicly accessible membership list. To do this requires that an individual send a single command, in an e-mail message, to the group's mail server. Thus, it is technically possible for individuals to remove the identifying information from the groups membership list [NOTE: Concealing oneself in this way does not remove identifying information from group messages - just from the group membership list]. For both practical and privacy related reasons groups which have more than 10% of their members concealed will be excluded from this study.

The command requesting membership information will be sent once a day from the project account during late night, off-peak hours. When the membership list is received, it will be held as the raw data that will be processed and stored for further analysis (for more detail on processing of the raw messages see the attachment entitled "Data Handling and Storage Procedures"). This procedure will be repeated daily for at least 100 days. After that time data collection will continue contingent on the availability of facilities for storing the data.

Addressing Inquiries Regarding the Study

It is common for group owners to monitor both new subscription requests and requests for information about the group membership. It is for this reason that the project account identifying information was chosen to clearly indicate the non-individual nature of the address (egroup@andrew.cmu.edu and "Electronic Group Dynamics Study"). Because of the unusual nature of the address we are prepared for three possible actions by group administrators (1) removal of the project account from the group membership with no additional correspondence, (2) further restriction of the group membership information, or (3) an inquiry regarding the nature of the study. Here we will describe the action that will be taken in the event of each of these actions by the group administrators.

Project account removal

If the administrator removes the project account from the group list with no further correspondence a message will be sent which describes the study, states that no further data will be collected, and asks for explicit permission to use any data that has been collected previously. If the administrator responds positively then previously collected data will be used. If the administrator responds negatively or does not respond after several requests then previously collected data will be destroyed.

Restriction of membership information

If the administrator alters the accessibility of the group membership information a message will be sent which describes the study, states that no further data will be collected, and asks for explicit permission to use any data that has been collected previously. If the administrator responds positively then previously collected data will be used. If the administrator responds negatively or does not respond after several requests then previously collected data will be destroyed.

Inquiry regarding the nature of the study

If the administrator sends an inquiry regarding the nature of the study a message will be sent which describes the study and clearly states that if the administrator wishes to be removed from the study that all they need to do it reply to the message (A positive response is assumed in this case because the administrator initiated the interaction and he or she has control - i.e. they can easily remove the study from the list). If the administrator does not respond or responds positively it will be assumed that he or she is willing to allow the group to remain in the study. If the administrator responds negatively then previously collected data will be destroyed.

Inquiry Response Message

To: <Group Owner>

From: egroup@andrew.cmu.edu <"Electronic Group Dynamics Study">

Subject: Re: <Original message title> - Description of the Electronic Group Dynamics Study

The Electronic Group Dynamics Study is an ongoing research project focused on understanding the factors and processes which underlie the development of successful on-line groups.

For the first phase of this project we are collecting membership lists and group message data from approximately two hundred different professional and recreationally oriented publicly accessible groups. The data collected for this project will be used for research purposes only. Individuals' e-mail addresses and names are encrypted in the research dataset such that it is impossible for the dataset to be used to contact or identify particular individuals. The data collection is non-intrusive. No messages will be sent to your group or to individual group members.

<Insert custom paragraph here>

Thank you for your interest and attention,

The Electronic Group Dynamics Study
Carnegie Mellon University
Pittsburgh, PA

-=-=-=-=-=-=-=-=-=-=-

(Custom paragraph for situation 1)

Our records indicate that our project account has been removed from your group. We interpret this as an signal that you and your group are unwilling to continue to provide data for this research. If you do not contact us within 4 weeks and provide explicit permission we will destroy all data that we have previously collected regarding your group. If we may continue to collecting data from your group (or at least use the data that we have already collected) contact us at egroup@andrew.cmu.edu.

-=-=-=-=-=-=-=-=-

(Custom paragraph for situation 2)

Our records indicate that recently have changed your policy regarding the accessibility of the membership list of your group . We interpret this as an signal that you and your group are unwilling to continue to provide data for this research. If you do not contact us within 4 weeks and provide explicit permission we will destroy all data that we have previously collected regarding your group. If we may continue to collecting data from your group (or at least use the data that we have already collected) please contact us at egroup@andrew.cmu.edu.

-=-=-=-=-=-=-=-=-

(Custom paragraph for situation 3)

If you feel that this project is not appropriate for your group please contact us at egroup@andrew.cmu.edu and we will stop collecting data for your group and destroy any data that we have collected previously from your group. If you have any questions feel free to contact us.

Data Handling and Storage Procedures

The raw data for this study, which includes membership lists and group messages, will all be received as plain text, electronic mail messages in the project account. Here we will describe the procedures and facilities that will be used to process the raw data into analyzable form. Generally the procedures and facilities fall into three phases: daily procedures, weekly processing, and long-term archives.

Daily procedures

Group messages will arrive continually and membership list messages will be received daily. The mail directories of the project account, which is accessible only to the researcher responsible for data collection, acts as the holding area for this data. The following procedures will be conducted at least once a day. If the storage capacity of the project account is reached then these procedures will be perform several times per day.

1. Classifying incoming messages

Using the Flames capabilities of the Andrew Mail System, the incoming messages will be filed in the following folders (one set for each group):

<groupname>.lists : Membership data

<groupname>.msgs : Group message data

<groupname>.admin : Administrative messages and correspondence regarding the group

2. Transferring data to temporary archive

The folder structure described above is mirrored in an AFS project volume, to which access is restricted to the researcher responsible for data collection and processing. Furthermore, in order to protect the temporary archive should the project account be compromised, there will be no explicit link between the project account and the AFS project volume. On a regular basis (i.e. at least once a day) the content of the project account mail folders will be transferred to the project AFS volume directories. This project volume serves as temporary archive which will hold at most one week of collected data.

Weekly processing

The following procedures will be conducted at least once a week. As with the daily procedures the following processes may be conducted more frequently if storage constraints require it.

<u>Membership list processing</u>

For each membership list message the raw datafile will be converted into a standard membership data file with the following format:

Line 1: <Groupname>

Line 2: <Data & time data sent>

Line 3: <Total number of members>

Line 4: <Number of concealed members [-1 if unknown]>

Lines 5 to the End: Encrypted identifiers for group members (one per line)

Identifiers will be encrypted using the following procedures:.

1. The e-mail address for the membership entry will be identified. Because it is unique this address will be used as the basis for creating the encrypted identifier.

2. Based on the e-mail address the encrypted identifier will be constructed [The particular encryption algorithm has not yet been chosen. The requirement is that this algorithm completely obscure the identifying information while remaining unique for each individual address].

3. The encrypted identifier is checked against the list of previously encountered identifiers. If the identifier has not been previously encountered then both the encrypted identifier and the associated identifying information is written to the new identifier screening file. This temporary file will then be reviewed manually by the researcher

responsible for collecting the data to identify unusual "members", including administrative accounts and distribution accounts, which if left undetected would significantly hinder reliable interpretation of the data and subsequent analysis. After this manual review has been completed the identifying information will be reduced to the domain information (i.e. cmu.edu) and a type indicator that identifies various categories of non-individual members. This data is then written to the new identifier update file.

4. Finally, the encrypted identifier will be written to the new membership list file.

This processing results in the creation of the following data files:

1. Raw membership messages [The files will be transferred to data cartridge where they will be archived until the data processing utilities are verified to be working properly - at which time these files will be destroyed]

2. Encrypted membership data files [These files will be removed from the AFS volume and transferred to a removable data storage media (i.e. data cartridge) for further processing]

3. A new identifier update file [These files will be removed from the AFS volume and transferred to a removable data storage device for further processing]

Group Message Processing

Each group message received will be in a separate text file. Each of these files will be processed to create a file with the following format:

Line 1: <Groupname>

Line 2: <Date and time message sent>

Line 3: <Message Subject>

Line 4: <Encrypted identifier of the message sender>

Line 5: <Encrypted identifiers of other recipients of the message [comma-separated]

Lines 6 to end: Contents of the message

The sender and other recipients (i.e. the contents of the cc: field) information will be processed using the same procedure described above for encrypting the membership list entries. This will result in the following data files:

1. Raw group message files [The files will be transferred to data cartridge where they will be archived until the data processing utilities are verified to be working properly - at which time these files will be destroyed]

2. Encrypted group message files [These files will be removed from the AFS volume and transferred to a removable data storage media (i.e. data cartridge) for further processing]

3. New identifier screening file [These files will be removed from the AFS volume and transferred to a removable data storage device for further processing]

After processing of the membership and group message data is complete the only data remaining in the AFS project volume will be a list of previously encountered encrypted identifiers for each group. This file will contain no unencrypted data.

Long-term archive

The long term archive for this project will be maintained on data cartridges which will not be accessible on-line, including though CMU's campus network, except when transferring new data updates from the AFS project volume.

The incoming data files will be used to construct group data files which record the daily volume of messages, number of people entering the group, and other daily group level statistics. After the group level statistics have been compiled, the data files will be compressed and archived for later use. The archives for this project will have the following components:

1. Original, unencrypted data file archives [This set of compressed files will be maintained until the data processing utilities are verified to be working reliably and then it will be destroyed.]

2. Identifier information database [This database will combine the new identifier update files]

3. Encrypted membership data file archives

4. Encrypted group message file archives

These archives will serve as the basis for the creation of working dataset, such as the group level summary dataset, which will in turn serve as the basis for the analysis and results of the study.

Use and Presentation of the Data

The data collected for this study can be analyzed at several different levels. Here we will discuss alternative analysis approaches and the procedures that will be used when presenting the various types of results.

Group level analysis

This analysis strategy focuses on modeling the changes in the group as a whole. As a result, the data which is used begins by aggregating individual numbers into statistics which characterize each group, such as size, daily number of entering members, or message volume. In this case the group is the unit of analysis and as a result individual behavior is not identifiable in the final results.

Individual level analysis

Another strategy that can be used to analyze this data is to consider questions in which the individual is the unit of analysis. For example, this approach might consider whether a particular communication behavior, such as voluntarily introducing oneself to the group is significantly related to the likelihood of the individual leaving the group. Because the dataset contains almost no data about individuals outside the context of the group this approach will focus on understanding how general behaviors and conditions within the group affect important individual behaviors within the group context.

Use of message contents

The dataset described above contains a significant archives of group message contents. This data will be used in several ways in the analysis and presentation of results. First, these archives will serve as the raw data for quantitative content analyses. These analyses will be used to identify significant communication behaviors, such as group conflict or individual introductions, which in turn will be analyzed using the general strategies described above. Second, the archives will serve as the basis for constructing group histories. Finally, individual messages and series of messages may be used to illustrate relevant behaviors and group phenomena.

Under no circumstances will any information that can identify individuals be included in publication or presentation of the results or data. In any presentation of message content steps will be taken to hide the identity of the individuals associated with the message. Identifying information in the content of the message, such as 'signatures' and other references to individuals, will be removed or replaced with general labels (i.e. YourTown or PersonName). Under no circumstance will any attempt be made to locate or present personal information about an individual which might serve to identify the individual.

Risk and Benefit Assessment

The risks of this study to individual group members or the group as a whole are minimal. The study is entirely observation oriented. As described above, groups which regularly deal with personally sensitive topics will be avoided when selecting the sample. No interventions will be attempted. The data will be encrypted to obscure identifying information and results will be presented in such a way that it is not possible to identify individuals. Thus, it is unlikely that individuals will be affected in any significant fashion by this work.

This research will help develop a better understanding of how social structures develop in computer-mediated environments, enabling those interested in promoting the development of these groups to better apply these new technologies in a range of contexts. More generally, this work will inform the creation of more general models of group development. Therefore, the results of this research will contribute to the study of natural group evolution which is an area of interests to researchers in a variety of disciplines.

Attachment 1: Excerpts from E-mail correspondence summarizing IRB concerns

As I mentioned to you earlier this week, I sent out your revised protocol for a final review. Several additional concerns have been raised, so I am sending these comments on to you. Once you have addressed these concerns, we should be able to approve this.

I know this has been a bit of trying process, and I think you have done a good job at addressing many of the eithical issues such as, (1) excluding from the study "erotica and other sexually related topics" and"support groups", (2) developing in advance responses to groupowners/administrators actions and inquiries, (3) promising encryption of group member identifiers, (4) promises of confidentiality of the identify of individuals in publication or presentation of the results or data, and (5) the assurance that no interventions will be undertaken in this research.

This attention is good as far as it goes. There are, however, some additional concerns about this proposed research:

Electronic groups are prickly, highly sensitive to perceived assaults on their rights of free speech. The medium is also one in which information and misinformation can be rapidly disseminated and with few inherent reality checks. Carnegie Mellon has had one bad experience with a research effort in this area. Furthermore, the electronic community will be alert to any research study coming from Carnegie Mellon that monitors behavior in cyberspace, and can be expected to examine it with prejudice for any perceived ethical violation. With this background suggesting caution, the external reviewers want to highlight some specific concerns:

1. Most importantly, the members of the electronic group are given no opportunity to opt out of the surveillance proposed in this study. They are not asked for informed consent, nor are they notified about the study. Given the difficulties that we presume would be involved in obtaining informed consent, the PI can likely argue that this route is not practical. It is harder, we think, to argue on practicality grounds that the group administrator should not be notified about the study and asked for consent. We have to be suspicious that the reason for not asking for the group administrator's consent is simply that in too many cases consent would be denied.

2. Of course you do not disguise your entry into the group
(egroup@andrew.cmu.edu)
so this does permit the group administrator to pick up on this and object. From an ethical perspective, this approach is suspect. It places the burden on the administrator to pick up on this name and be suspicious of it.

3. If and when a group administrator does object, you propose to send one of several possible messages in response, depending on the action taken by the administrator. Wechave two concerns here: (1) the

phrase "unwilling to continue to provide data for this research" suggests that previously they had been willing (but they were never asked), and (2) the promise to destroy all data provides no mechanism for the administrator to verify that in fact it has been done.

4. There is nothing in the proposal about how you would handle the "electronic moral outrage" that might erupt if a group reacted negatively to the uninvited surveillance.

Attachment 2: An addendum to the proposal addressing the IRB concerns

The following concerns were raised regarding the proposed research: individual consent, interaction with group administrators, and dealing with "electronic moral outrage". In this document we will address each of these issues.

- Individual Participation in the Group and Consent

In the review comments it is implied that the primary argument against seeking the consent of each individual in the studied groups is one of practicality. While, practicality is a concern, there are important theoretical reasons why seeking individual consent is not a reasonable procedure.

Participation in an Internet-based electronic group is a semi-public activity. It is public because the individual is engaging in actions that are visible to an unspecified group of others. Specifically, the groups considered in this research all have open-membership policies and many have public archives of group communication. On the other hand, individuals often engage in activities that are publicly visible but not salient. The presence of many others who are involved in the same activity or the 'normalcy' of an action leads individuals to believe, very reasonably, that though the activity is publicly visible it is not likely to be useful for identification purposes. For example, individuals will often walk down a busy city street and implicitly assume that they are "blending in with the crowd".

However, it is important to note that the degree to which individuals take advantage of these features of semi-public spaces varies. Individuals who choose to cross a street at a busy downtown intersection are likely to perceive their actions as being more private than individuals who choose to stand at the street corner and preach, shout obscenities, or sell newspapers. Likewise, in electronic groups individuals who remain "lurkers", not contributing to the group communication, are likely to view their action as more private than individuals who contribute.

To ask individuals to provide explicit consent would undermine the choice made by the individuals who choose to participate in a non-salient way. As lurkers, they remain non-salient, blending in with the crowd. Sending these individuals messages regarding consent would result in a destruction of the "crowd", by explicitly identifying each individual. This is likely to significantly affect the individual's behavior and the operation of the group as a viable social system.

It is for this reason that protocol of this study mirrors the semi-public structure of the electronic groups. Through encryption of identifying information individuals who choose to remain lurkers will remain completely anonymous - there will be no way to identify them or associate participation in the group with them. This will maintain the "crowd-like" character of the semi-public space. Individuals who choose to contribute to group communication will also shielded to the degree that they allow. With the data, as with the group itself, individuals who contribute more, or participate in unusual ways, will likely be more visible. Thus, by mirroring the semi-public nature of the electronic groups under consideration, this research protocol provides a data source that is of value for research seeking to understand electronic groups while not compromising the group, or individuals involved.

- Interaction with Group Administrators

The second set of concerns mentioned related to interaction with group administrators. These concerns included:

* Requesting group administrators' consent
* Placing the "burden on the administrator"
* Dealing with administrators' responses

Again, while there are some practicality concerns, there are also important theoretical bases for the protocol.

In many research contexts it is the case that the researcher has a social "upper hand". That is, individuals perceive the activity of the researcher as being societal or institutionally sanctioned and as a result are more likely to comply, even though they might not otherwise choose to do so. This compliance arises, in part, because the researcher is in control of the situation (laboratory, survey ,etc.), the activity is unusual (experimental tasks or survey questions), and there exists a general norm in many of these contexts (university or corporation) encouraging compliance with authority. However, it is important to note that when dealing with individuals on the Internet, in particular group administrators, these conditions are rarely met.

First, in the context of an electronic group the group administrator - not the researcher - has complete control. Administrators often remove individuals which they perceive as harming the group. Unless the members of the group object the administrator has complete technical and social control.

Second, the actions undertaken as part of the research protocol for this study are not unusual on the Internet. There are many marketing firms which routinely identify electronic groups and collect their membership information or monitor their group communications. In addition, there are many automated tools which undertake similar activities, constructing searchable archives of group messages and directories of individual e-mail addresses. As a result, the data collection activities described in this protocol are relatively common. Groups (or more specifically group administrators) who believe that these activities are harmful routinely restrict access to this information and monitor the use of these facilities.

Finally, because there is no shared organizational context, and little shared institutional context on the Internet the generalized norms which might influence group administrator participation are not present. If anything, it can be argued that the prevailing culture of the Internet is one of rebellion or non-compliance.

As a result, it is likely that group administrators who are concerned about external uses of group information are already be watching for the activities described in the research protocol. On the other hand, because of time constraints which lead many individuals to take a 'triage' approach to public communication, it is also likely that group administrators who are ambivalent or open to external use of group information will fail to respond to explicit requests to participate. Based on discussions with several group administrators it is expected that the normal response to such a request would be no response - an uninterpretable outcome.

The problem of interpreting non-responses and the difficulties of dealing with multiple group administrators for each group (a common situation) motivated the construction of a protocol which combines a visible study identity, a redefined response, and a data collection process which minimizes a interference. This provides group administrators with the ability to remove their groups from the study with the least possible effort.

There were also concerns with the wording of the response message which used the phrase "unwilling to continue to provide data for this research". This phrasing was chosen to alert the group administrator that data had been collected prior to the group administrator's query or action regarding the study.

Finally, there were concerns that the promise to destroy all data "provided no mechanism for the administrator to verify that it in fact has been done." This is a fundamental characteristic of any electronic data store. Providing evidence that data has been destroyed would require demonstrating that there are no copies in existence - a logically intractable problem. However, a final confirmation message will be sent to the group administrator informing them that the data has been destroyed.

- Dealing with "Electronic Moral Outrage"

Public response to this research is likely to take several forms. For discussion purposes we will categorize the responses in terms of two dimensions: location and method. These two dimensions form the following four categories:

* (Within a studied group)/(Discussion-based responses)
* (Within a studied group)/(Attack-based responses)
* (Outside a studied group)/(Discussion-base responses)
* (Outside a studied group)/(Attack-based response)

If a studied group responses to the research by beginning a discussion of the issues regarding external access to and use of group information, the group administrator will be contacted as described above. As a result, it will become the group administrators' decision whether the group remains in the study or not. If the administrators wish to distribute further information about the study it is their choice.

If individuals outside the studied groups wish to engage in a discussion of the issues raised by this study a summary of the human subject review protocol description will be provided and their concerns will be considered.

However, if individuals choose to engage in a attack-based response, either in the form of attacking the project account or broadcasting hostile messages there is little that can be done in response. The project account e-mail identity will be changed to limit damage to the data collection from spurious mailings. And public requests for information regarding the study will be responded to with information about the protocol and research project.

- Conclusion

The goal of this research is to develop an understanding of electronic social structures that can support the development of future on-line communities. The challenge is to do this with minimum impact on the existing electronic community. As with all social science research it is impossible to construct such a study without affects someone. However, we feel that the protocol described in the current proposal provides an effective compromise between research demands, institutional concerns, and the needs of the current electronic community.

Attachment 3: Excerpts from E-mail correspondence with an IRB reviewer expressing concerns

Brian Butler addresses the three key issues I raised about his proposed research:

1. Individual Participation in the Group and Consent

I find Butler's discussion mostly persuasive on this score. Implicit to his argument however is the notion that surveillance of semi-private arenas is ethically OK. In general this is not completely accepted~note concerns that have been raised about video cameras scanning street corners. Butler makes a distinction between lurkers and participants. The analogy might be that it is OK to videotape street corner preachers, but not passersby (more than incidentally). I would be happier about this if Butler would cite precedents for this kind of research. An appropriately modified version of this discussion should be included in any final version of his proposal.

2. Interaction with Group Administrators

Butler argues "the actions undertaken as part of the research protocol for this study are not unusual on the Internet. There are many marketing firms which routinely identify electronic groups and collect their membership information or monitor their group communications". I suggest giving examples of such firms and providing evidence of what their experience has been with reactions of groups to their activities. Further, if indeed, groups "routinely restrict access to such information", then that is a clear indication that groups find such surveillance inappropriate.

I am concerned about the argument in the paragraph beginning "Finally, ~". Accepting the argument suggests to me that the proposed research is a violation of the prevailing culture of the Internet. This should be a warning signal.

The paragraph regarding the phrase "unwilling to ~" misses my point. The issue is not whether the data had already been collected. Clearly it has been. The issue is, rather, that the phrase suggests that they had been willing to provide it in the past. Since they had not been asked, they had never expressed such willingness. From an ethical point of view I think a key factor here is the duration of the activity. If I have been taking pears from your pear tree for a week and you have not complained, I can well be held liable if you now find fault. On the other hand, if I have been doing this for the past twenty years, you may well now stop me but can hardly hold me liable for my twenty years of transgressions. Actually I think there is something similar to this in real estate law. In the case of the proposed research the duration is presumably short and so the administrator~s "acquiescence" is most likely due to having failed to note the activity previously.

3. Dealing with "Electronic Moral Outrage"

Butler does not seem to take this issue seriously, first by minimizing its likelihood and then by suggesting that the only response needed would be a decision by the group administrator. Further, the intent of the paragraph on individuals in the group starting a discussion of the issues regarding external access is not clear and does not seem fully thought out. One issue is that if such discussion takes place it has an impact on the collected data. Thus the surveillance has become an active intervention and the study design is not that of simple observation. Overall, I don~t find this section to be responsive to my concerns. What is really needed is some contingency planning for what actions would be taken if there is a major flap on the Internet about the surveillance that is required by this research.

Attachment 4: A second addendum addressing the reviewers concerns

To address the most recent comments the following items considered below:

1. A modification of the group administrator interaction protocol

2. A discussion of data collection in semi-public, electronic groups

3. A protocol for managing "electronic moral outrage"

(Also included are several appendices with supporting materials and discussions)

Thanks for your effort,
Brian

** Interaction with Group Administrators **

Concerns have been raised regarding the proposed protocol for interacting with group administrators. After consideration of these comments we have developed the following modified protocol.

Before data is collected a short description of the research (included in Appendix A) will be sent to each group administrator(s). The message length is limited to increase the likelihood of administrators reading it. Included in this description is:

- The e-mail address of the project account which will collect the data
- Instructions for removing the group from the study

Data will not be collected until 7 days after the informing message is sent.

If an administrator requests more information, a more detailed description of the study will be sent. Data will be collected from a group until 5 days after the last information request is accommodated.

If an administrator requests that the group not be included in the study then a message will be sent thanking them for their response and confirming that the group will NOT be included in the study (included in Appendix A).

Explicitly informing administrators and minimizing the cost of opting out of the study minimizes the chances of a group being included in the data collection against the wishes of the available group representative.

-- Comments on other specific concerns

The above protocol addresses the general concerns regarding interaction with group administrators. Other specific concerns are addressed below.

> Butler argues "the actions undertaken as part of the research protocol
> for this study are not unusual on the Internet. There are many marketing

> firms which routinely identify electronic groups and collect their
> membership information or monitor their group communications". I
> suggest giving examples of such firms and providing evidence of what
> their experience has been with reactions of groups to their activities.

One example this type of service is DejaNews (http://www.dejanews.com/). Their policy is that all Usenet postings are "published" (i.e. public) and hence privacy is not an issue (http://www.dejanews.com/pr/dndn.html). While this is just one approach to using a repackaging and distributing individual level information on the Internet it is not an uncommon one. However, in this research the data will be stored off-line and will not be redistributed.

> Further, if indeed, groups "routinely restrict access to such
> information", then that is a clear indication that groups find such
> surveillance inappropriate.

The existence of groups which restrict access to membership information is an indication that some groups find such observation inappropriate. Conversely, the existence of groups which do not restrict access to membership information, even when it is possible to do so with minimal effort, is an indication that other groups do not consider external use of this information inappropriate. Data collection in this study is limited to these groups.

> I am concerned about the argument in the paragraph beginning "Finally,
> ~". Accepting the argument suggests to me that the proposed research is
> a violation of the prevailing culture of the Internet. This should be a
> warning signal.

In a prior note the phrase "Internet culture" was used to emphasize that group administrators are not all members of an organization, corporation, or institution in which the norm of compliance with a particular authority is implicitly or explicitly promoted. However, conceptualizing "Internet culture" as a unified community is problematic. The Internet consists of many groups, including corporate executives, rebellious teenagers, militia groups, and knitting enthusiasts - each of whom use the technology to construct their own spaces with different norms and "cultures" (O~Reilly, 1996).

> The paragraph regarding the phrase "unwilling to ~" misses my point.
> ...
> and so the administrator~s "acquiescence" is most likely due to having
> failed to note the activity previously.

This is addressed by the modified protocol which explicitly informs administrators prior to the collection of any data.

** Individual Participation in the Group and Consent **

Professor Duncan comments that "[i]mplicit to [Butler's] argument however is the notion that surveillance of semi-private arenas is ethically OK". However, this is not quite accurate. Implicit to the prior arguments is the notion that observation of behavior in semi-public arenas is ethically acceptable. While the distinction may seem to be a minor one, it has significant implications for the ethics of this research. [For more details on this distinction see the discussion in Appendix B]

When impact on the individual is low, the cost of participation if low, and there is minimal risk to the individual then the use of unobtrusive observation in semi-public arenas is considered to be ethically justifiable (Diener and Crandall, 1978: pp. 38-41). Thus, I believe the proposed research protocol is ethically defensible because steps are taken to minimize impact and risk to individuals and to limit observation to public and semi-public groups.

** Dealing with "Electronic Moral Outrage" **

The concerns with "electronic moral outrage" fall into two categories: discussion within a sampled group and "outrage" on the Internet in general.

-- Discussion of issues within a sampled group

There are two concerns which arise when dealing with discussions within a sampled group: the implications of these discussion for continued data collection and the role of the researcher, if any, in these discussions.

There are difficulties which arise when dealing with the concept of "group consent" (Diener and Crandall, 1978; Beauchamp, Faden, Wallace, and Walters, 1982). If the extreme view is applied, that a collective should not be studied if even one member objects, then most studies of organizations are suspect. In the proposed protocol the issue of group consent is addressed by interacting with a representative of the group, the group administrator. Thus, if a group administrator chooses to remove the group from the study - with or without internal discussion of the matter - he or she is easily able to do so.

Another possiblity is that groups which engage in discussion of these issues should be removed because the data will been "corrupted". However, this assumes that these discussions are unusual in on-line groups. Casual observation suggests that groups often engage in administrative discussions which consider various group procedures and policies. Also common are "off-topic" discussions, which are then informally or formally managed within the group. Thus, discussion of external uses of group data can be considered a normal part of group operations and hence its presence does not have an corrupting effect on the collected data.

What, then, should the role of the researcher be in the discussions within sampled groups? An on-line group is dynamic social system which adapts to meet the needs and of members. If a researcher contributes directly he or she is likely to impact the direction and duration the discussion. It then becomes more likely that the researcher's goals, and not the groups members' interests are being served by the discussion. Direct contribution to discussions is an explicit intervention. It is for this reason that the proposed protocol involves the researcher NOT contributing directly to the group discussion, but interacting with the group administrator(s) instead.

Discussion of the external use of group records is within the range of normal group activities. The proposed protocol takes steps to allow groups and group administrators to discuss these issues (if they choose to) and make their choices with minimal interference from the researcher.

-- External "Electronic Moral Outrage"

One form of extra-group "electronic moral outrage" is Internet-based attacks on the research project itself. To minimize the likelihood of this the data will be stored off-line and the project e-mail account will be monitored to ensure that the impact of 'spamming' is minimized.

Another form of external "outrage" is the dissemination of derogatory messages regarding the research (i.e. flaming). Prudent execution of the research and responsible reporting of the result should minimize the likely of this occurring. However, as with any media, there is little that can be done to prevent public attention if it should arise. If this research should attract attention of individuals on the Internet an "electronic press release" which describes the goals, methods, and limitations of study will be created and distributed in response to any inquiries.

If in the University's view it becomes necessary, I will work with the CMU Public Relations department to prepare a traditional press release to accurately represent the goals, methods, and limitations of this work. However, it is my intention to focus on academic outlets in the publication of the results of this research.

References

Diener, Edward and Crandall, Rick (1978) Ethics in Social and Behavioral Research. The University of Chicago Press: Chicago.

Beauchamp, Tom L., Faden, Ruth R., Wallace, R. Jay., and Walters, LeRoy (1982) Ethical Issues in Social Science Research. John Hopkins University Press: Baltimore, MD.

O~Reilly, Tim (1996) Publishing Models for Internet Commerce, Communications of the ACM, 39(6), pp79-86.

Appendix A: E-Mail Messages for Interaction with Group Administrators

[Informing message for Group Administrators]

To: Group Administrator
From: Electronic group dynamics study
Subject: Electronic group dynamics study

Hello,

We are conducting a study of the growth and dynamics of electronic groups and have selected <- Insert group name here -> as part of a representative sample of on-line groups. The goal of our study is simply to observe how electronic groups develop and change over time. The results of this study will inform the development of new communication technologies and on-line communities. It is important for us to include this group in the sample to accurately represent the diversity of groups which exist on the Internet.

This research involves only observation. We will collect membership data (using the 'review' or 'who' commands) and group messages. To protect the identity of individual members, e-mail addresses will be encrypted and the dataset will not be made publicly available. No messages will be sent to the group as part of this study and individuals e-mail addresses will NOT redistributed.

If you feel that it would be inappropriate for <-- Insert group name here --> to be included in this study, respond to this message within 5 days and we will not include it in the sample.

If you would like more information about the study please contact us at: egroup@andrew.cmu.edu.

Thank you for your time and attention,

Brian Butler
Electronic Group Dynamics Study
Carnegie Mellon University
Pittsburgh, PA

[Response for administrators who request removal of their group from the study]

To: Group Administrator
From: EGroup Study
Subject: Confirmation of your request to be removed from the study

This message is to confirm that <- Insert group name here -> will not be included in the electronic group dynamics study. No membership data or group messages will be collected.

Thank you,

Brian Butler
Electronic Group Dynamics Study
Carnegie Mellon University
Pittsburgh, PA

-=-=-=-=-=-

Appendix B: Comments on the Observation of Semi-Public Arenas

Public arenas are spaces in which an individual's actions are potentially visible to an undefined set of others. In contrast, a private arena is a space in which an individual's actions are potentially visible to a well-defined, known set of others. One aspect of privacy is the ability to control who knows what about you. Hence, it can be argued that when an individual acts in a purely public arena they are relinquishing control of who may observe that activity. For example, a television news anchor is literally open to the world for the time that he or she is on the air. On the other hand, when an individual acts within a purely private arena they maintain that control - and hence maintain their privacy with regard to that action. In contrast, in their personal bedrooms or offices, individuals have detailed knowledge of the individuals who have access to these spaces - and hence the ability to directly control the inherent visibility of their actions.

However, as noted earlier, saliency or unusualness of an action can also influence its visibility. In semi-public and semi-private arenas variation in the salience of actions may affect the degree to which an action is visible. Semi-private arenas are private spaces which in which individuals engaged in unusual or salient actions often feel that their actions are highly visible or memorable. For example, individuals attending a small private holiday party may find that their unusual actions are "publicly visible" because they are memorable and likely to be repeated in future descriptions of the party. Semi-private arenas often arise due to secondary effects of word-of-mouth which result in actions being "visible" far beyond the boundaries of the private space.

In contrast, semi-public arenas are public spaces in which actions are inherently visible to an undefined set of others, but low salience or normal actions can be treated as if it were not publicly visible (e.g. "blending in with the crowd"). For example, an individual walking down a city street can reasonably assume that though their action (walking) is visible to an undefined set of others it is unlikely that it will be noted or remembered by anyone. As I argued in prior comments the Internet groups which are the focus of this study are semi-public spaces [1].

However, there remain concerns about the ethics of data collection in semi-public arenas. To address these issues we must distinguish between surveillance of individuals and observation of behavior. Surveillance which involves the matching of identified individuals with recorded behaviors is typically undertaken in order to learn about particular individuals. In contrast, observation of behavior, which involved identifying individuals only for logistical purposes, is undertaken to learn about the frequency or nature of certain behaviors, independent of the particular individuals who are involved.

The key concern is that though data is collected for observation purposes it may also be applied for surveillance purposes. In his most recent comments the reviewer notes that "the analogy might be that it is OK to videotape street corner preachers, but not passersby (more than incidentally)." This overlooks a key concern. Collection of video-tape records of individual behavior in a semi-public space inherently supports surveillance because it is difficult to obscure individual identities on video-tape. Even if a person is only taped "incidentally" the nature of visual data (as banks and police know) is that individual are identifiable. Even if the purpose of data collection is merely to observe and record certain behaviors with video-tape based data collection it is almost impossible to do so without supporting surveillance.

There are, however, many procedures for observing behavior which significantly hinder the use of the records for surveillance purposes. Manually or automatically collected counts - as are often done in public libraries or highway traffic studies - keep records of activity in such a way that it is essentially impossible to match the identity of particular individuals with the record of their action. In longitudinal research programs where it is important for logistical reasons to maintain some code identifying individuals it is standard procedure for researchers to maintain strict control over the access to and use of identifying information. The proposed research protocol take this type of precautions to ensure that the collected data cannot effectively be used for surveillance purposes.

Thus, we return to concerns regarding the ethics of observation of behavior in semi-public arenas. There are many studies which involve the observation, and in some cases experimental manipulation of individuals in semi-public spaces including, but not limited to:

- Studies of charitable giving on the street
- Studies of littering behavior
- Marketing studies of consumer behavior in shopping centers

Therefore, I believe that unobtrusive observation of semi-public electronic groups is it is an ethically defensible research strategy.

Footnotes:

[1] This does not imply that all Internet groups are semi-public arenas. The selection criteria described in this research protocol creates a sample of groups which are semi-public spaces.

Paper Three Appendices

Communication Cost, Attitude Change and Membership Maintenance: A Model of Technology and Social Structure Development

Appendix A: Collective Development Example

This appendix contains the MATLAB scripts for the collective development example. The script was executed with MATLAB 5.2 on a Windows PC.

Filename: parameters.m
Description: This file defines the filenames, operational settings, and virtual experiment parameters for the basic model.

```
% ----------------------------------------------------
% This is a template for the parameters of the
% networked collective simulations
% ----------------------------------------------------
% Brian Butler © Copyright 1998
% Created: 7/21/98
% ----------------------------------------------------

% Define the filenames
OutputFilename = 'rundata.out';
InterestFilename = 'intprofile.out';

% Define the simulation operations paramters
InitPeriodLength = 5;
RunLength = 3000;
TotalRunLength = InitPeriodLength + RunLength + 1;
CellSize = 15;

% Define virtual experiment parameters
wParameters = [0.005];
cParameters = [0.33];
mParameters = [5];
irParameters = [0.25,0.75];
```

Filename: model.m
Description: This file contains the MATLAB script that implements the model.

```
% ---------------------------------------------------
% This MATLAB script implements the process model of
% voluntary social collective development
% ---------------------------------------------------
% Brian Butler © Copyright 1998
% ---------------------------------------------------

% Clear all variables currently in use
clear;

% Set the random number generator
rand('state',sum(100*clock));

% Read the analysis parameter file
parameters;

% ---------------------------------------------
% Open the experiment data files
% ---------------------------------------------
OUTFILE = fopen(OutputFilename,'w');
INTFILE = fopen(InterestFilename,'w');

% ---------------------------------------------
% Loop through the cells
%  (Changing the parameters each time)
% ---------------------------------------------
for wIndex = 1:length(wParameters),
for cIndex = 1:length(cParameters),
for mIndex = 1:length(mParameters),

% ---------------------------------------------
% Set cell parameters
% =============================================
% wValue : The impact of messages on content perceptions
% cValue : Average cost of noise messages (to individual)
% mValue : Average max positive marginal benefit signal messages
% ---------------------------------------------
wValue = wParameters(wIndex);
cValue = cParameters(cIndex);
mValue = mParameters(mIndex);

% Create string containing the parameter to simply results recording
CellParameterRecord = sprintf('%.3f,%.3f,%d',wValue,cValue,mValue);

% Display a status message
fprintf('Cell: %s\n',CellParameterRecord);

% ---------------------------------------------
% Loop through the groups in the cell
% ---------------------------------------------
for GrpId = 1:CellSize;

fprintf('------- Group %d ----------\n',GrpId);
```

```
% ----------------------------------------------
% Create group
% (i.e. initialize agent parameters)
% ==============================================
% INT(1) = Low point of an agent's interests
% INT(2) = High point of an agent's interests
% w: Message weight (impact on group assessment)
% ce0: Initial beliefs about content
% ve0: Initial beliefs about volume
% c: Message cost
% m: Maximum positive benefit messages
% ----------------------------------------------

% Set the group size based on a draw from a log-normal distribution
% N = 9999;
% while (N > 5000 | N < 5)
%    N = floor(exp(4.17+1.54*randn));
% end;
% N = floor(rand * 500) + 5;
N = 75;

% Set the participation probabilities
PartRatio = 1.0;
PartProb = rand * 0.1;

PersonalPartProb = (rand(N,1) < PartRatio) * PartProb;

w = ones(N,1) * wValue;
ce0 = (rand(N,1) * 0.5) + 0.5;
ve0 = zeros(N,1);
c = ones(N,1) * cValue;
m = ones(N,1) * mValue;

% Randomly select the interest range parameters for the group
INTRange = (rand * (irParameters(2)-irParameters(1))) + irParameters(1);
INT(1:N,1) = rand(N,1);
INT(1:N,2) = INT(1:N,1) + (rand(N,1) * INTRange);

% Create string containing the parameter to simply results recording
GroupParameterRecord =
sprintf('%d,%d,%.3f,%.3f,%.3f,%.3f,%d,%.3f',InitPeriodLength,N,PartRatio,Part
Prob,wValue,cValue,mValue,INTRange);

% ---------------------------------------------
% Create the operations data structures
% ---------------------------------------------
clear Members Messages;
TotalVolume = 0;
ve = zeros(N,TotalRunLength);
ce = zeros(N,TotalRunLength);
Members = zeros(N,TotalRunLength);

% ---------------------------------------------
% Set Initial Conditions and Run Loop
% ---------------------------------------------
ve(1:N,1) = ve0;
ce(1:N,1) = ce0;
```

```
Members(1:N,1) = ones(N,1);
Messages(1:N,1) = ones(N,1) * -1;
MemPer = ce(find(Members(1:N,1)),1);
S = sum(Members(1:N,1));

Evaluations = zeros(N,TotalRunLength);

t = 2;           % t = 1 are the initial conditions

% Loop until stability is reached or RunLength is reached
while (t <= TotalRunLength),

% Record the group state values at the end of the initialization period
if t == InitPeriodLength
   TrueN = S;
   InitMsgVolume = TotalVolume;
end;

% Construct initial and final interest profiles for the group
% if ((mod(t,500) == 0) | (t == 6))
if 1 == 0
   % Determine member and non-member interests
   InitMemInt = INT(find(Members(1:N,InitPeriodLength-1)),1:2);
   MemInt = INT(find(Members(1:N,t-1)),1:2);

   % Determine the distribution of member interests
   index = 1;
   intlist = [];
   for i = 0.05:0.05:1,

      % Store values for the initial profile
      InitDist(index) = sum(((InitMemInt(1:TrueN,1) < i) &
(InitMemInt(1:TrueN,2) > i)) | ((InitMemInt(1:TrueN,2) > 1) &
(mod(InitMemInt(1:TrueN,2),1) > i)));
      % Store values for the final profile
      IntDist(index) = sum(((MemInt(1:S,1) < i) & (MemInt(1:S,2) > i)) |
((MemInt(1:S,2) > 1) & (mod(MemInt(1:S,2),1) > i)));

      % Add appropriate values to the interest value list (used to test
normality)
      intlist = [intlist;(ones(IntDist(index),1) * i)];

      % Increment the profile index
      index = index + 1;

   end;

   plot(IntDist);
   % Setup axis values
   axis([-Inf Inf 0 Inf]);

   hold on;
   fprintf('%d vs. %d / %.3f (T = %d)\n',max(InitDist) -
min(InitDist),max(IntDist) - min(IntDist),min(MemPer),t);
   drawnow;
end;
```

```
% -----------------------------------------------
% Step 1: Generate Group Communication
% -----------------------------------------------
MsgMarkers = (rand(N,1) - (1 - PersonalPartProb) > 0) .* Members(1:N,t-1);

% ----------------------------------------------------
% Step 2: Update Individuals Perceptions of the Group
% ----------------------------------------------------

% Update Volume expectations
MessageCount = sum(MsgMarkers);
TotalVolume = TotalVolume + sum(MessageCount);
ve(1:N,t) = ones(N,1) * (TotalVolume / t);
CurrentVE = ve(1:N,t);

% Update Content perceptions
CurrentCE = ce(1:N,t-1);

% --- Generate messages and update perceptions appropriately ----
SourceList = find(MsgMarkers);
Messages(1:N,t) = ones(N,1) * -1;
for MsgCnt = 1:length(SourceList);

   % Select a message source
   Source = SourceList(MsgCnt);

   % Generate a message based on the sources interests
   Message = mod((rand * (INT(Source,2) - INT(Source,1))) + INT(Source,1),1);

   % Record the message
   Messages(Source,t) = Message;

   % Assess Reaction to the message
   MsgReaction = (((Message > INT(1:N,1) & (Message < INT(1:N,2))) |
((Message + 1) < INT(1:N,2))) * 2) - 1;

   % Compute the change in attitude due to the message
   CEChange = (MsgReaction .* w) .* (CurrentCE - (CurrentCE.^2));
   CurrentCE = CurrentCE + CEChange;

end;

% Record content perceptions
ce(1:N,t) = CurrentCE;

% -----------------------------------------------
% Step 3: Compute Benefit Expectations
% -----------------------------------------------
EBenefits = (-1./(2*m)) .* ((CurrentCE .* CurrentVE) .^2) + (CurrentCE .*
CurrentVE) - (c .* ((1 - CurrentCE) .* CurrentVE));

% Store the expected benefits
Evaluations(1:N,t) = EBenefits;

% -----------------------------------------------
% Step 4: Update the group membership record
% -----------------------------------------------
```

```
Members(1:N,t) = Members(1:N,t-1) .* (EBenefits >= 0);

% -----------------------------------------------
% update operating values
% -----------------------------------------------
MemPer = ce(find(Members(1:N,t)),t);
S = sum(Members(1:N,t));
t = t + 1;

% Print status message;
if(mod(t,1000) == 0),
   fprintf('..%d',GrpId);
end;


end;    % ----------- End of the Single Group Run -------------------

fprintf('\n');

% ------------------------------------
% Determine group feature measures
% ------------------------------------

% S : Group Size (already computed)

% Compute stability measures
if S > 0
   MeanCE = sum(MemPer) / S;
   StabilityIndex = min(MemPer);
else
   MeanCE = '.';
   StabilityIndex ='.';
end;

% Store the stopping time for easy reference
NowT = t-1;

% Determine member and non-member interests
MemInt = INT(find(Members(1:N,NowT)),1:2);
NonMemInt = INT(find(Members(1:N,NowT) == 0),1:2);

% Determine the mean interest range for members and non-members
if S > 0
   MemRange = mean(MemInt(1:S,2) - MemInt(1:S,1));
else
   MemRange = '.';
end;
if N-S > 0
   NonMemRange = mean(NonMemInt(1:(N - S),2) - NonMemInt(1:(N-S),1));
else
   NonMemRange = '.';
end;

% Determine the true participation ratio and participation average
% (First determine the participant count)
ParticipantCount = sum((sum((Messages(1:N,InitPeriodLength:TotalRunLength) >=
0)') > 0)');
```

```
if TrueN > 0
   TruePartRatio = ParticipantCount/TrueN;
else
   TruePartRatio = 9999;
end;

if ParticipantCount > 0
      TruePartProb = ((TotalVolume - InitMsgVolume)/(TotalRunLength -
InitPeriodLength))/ParticipantCount;
else
   TruePartProb = 9999;
end;

% Construct initial and final interest profiles for the group
index = 1;
intlist = [];
for i = 0.05:0.05:1,

   % Store values for the initial profile
   InitDist(index) = sum(((INT(1:N,1) < i) & (INT(1:N,2) > i)) | ((INT(1:N,2)
> 1) & (mod(INT(1:N,2),1) > i)));

   % Store values for the final profile
   IntDist(index) = sum(((MemInt(1:S,1) < i) & (MemInt(1:S,2) > i)) |
((MemInt(1:S,2) > 1) & (mod(MemInt(1:S,2),1) > i)));

   % Add appropriate values to the interest value list (used to test
normality)
   intlist = [intlist;(ones(IntDist(index),1) * i)];

   % Increment the profile index
   index = index + 1;

end;

InitIntRange(GrpId) = max(InitDist) - min(InitDist);
PInitIntRange(GrpId) = InitIntRange(GrpId) / N;

FinalIntRange(GrpId) = max(IntDist) - min(IntDist);
PFinalIntRange(GrpId) = (max(IntDist) - min(IntDist)) / N;

% fprintf('%d vs. %d\n',InitIntRange(GrpId),FinalIntRange(GrpId));

% Compute topic coverage for the final profile
TopicCoverage = sum((IntDist > 0)) / 20;

% Determine member and non-member participation status
NonMembers = find(Members(1:N,NowT) == 0);
NonMemberCount = length(NonMembers');
PNonMembers = sum((PersonalPartProb(NonMembers) > 0));
if NonMemberCount == 0
   PPDroppers = 9999;
else
   PPDroppers = PNonMembers/NonMemberCount;
end;

% fprintf('%.3f vs. %.3f\n',PNonMembers/NonMemberCount,PartRatio);
% ------------------------------------------
% Record group state at stability
% ------------------------------------------
% Record time to stability and size at stability
fprintf(OUTFILE,'%s,%d,%d,%d,%d,%.3f,%.3f,%.2f,%.3f,%.3f,%d,%d,%.3f,%.3f,%.3f
\n',GroupParameterRecord,GrpId,NowT,S,TotalVolume,StabilityIndex,MeanCE,Topic
Coverage,MemRange,NonMemRange,TrueN,InitMsgVolume,TruePartRatio,TruePartProb,
PPDroppers,InitIntRange,FinalIntRange);
fprintf('%s\n',GroupParameterRecord);

% Record Initial Interest Profile
fprintf(INTFILE,'%s,%d,',GroupParameterRecord,GrpId);
fprintf(INTFILE,'%d,',InitDist);
fprintf(INTFILE,'INITDIST\n');

% Record Final Interest Profile
fprintf(INTFILE,'%s,%d,',GroupParameterRecord,GrpId);
fprintf(INTFILE,'%d,',IntDist);
fprintf(INTFILE,'FNLDIST\n');

% Print status message;
if(mod(GrpId,5) == 0),
   fprintf('Group: %d\n',GrpId);
end;

fprintf('end of group run (%d)\n',GrpId);
pause;

% Clean up the graphs plots
hold off;

end;    % ----------- End of the Cell Cycle -------------

% ------------------------------------------
% End the parameter loops
% ------------------------------------------
end; % m (Maximum message) Loop
end; % c (noise Cost) Loop
end; % w (message impact) Loop

% ------------------------------------------
% Close all input and output files
% ------------------------------------------
fclose('all');
```

Appendix B: Model Calibration Scripts

This appendix contains the MATLAB scripts used to perform the initial and final model calibration runs. The scripts were executed with MATLAB 5.2 on a Windows PC.

```
---------------------- Intial Calibration Run ----------------------------
```

Filename: parameters.m
Description: This file defines the filenames, operational settings, and virtual experiment parameters for the basic model.

```
% -----------------------------------------------------
% This is a template for the parameters of the
% networked collective simulations
% -----------------------------------------------------
% Brian Butler © Copyright 1998
% -----------------------------------------------------

% Define the filenames
OutputFilename = 'rundata.out';
InterestFilename = 'intprofile.out';

% Define the simulation operations paramters
InitPeriodLength = 100;
RunLength = 130;
TotalRunLength = InitPeriodLength + RunLength + 1;
CellSize = 100;

% Define virtual experiment parameters
nParameters = [100];
acParameters = [0.005];
wParameters = [0.05];
cParameters = [0.33];
mParameters = [15];
irParameters = [0.25,0.75];
```

Filename: model.m
Description: This file contains the MATLAB script that implements the model.

```
% ------------------------------------------------------
% This MATLAB script implements the primary
% collective development model.
% ------------------------------------------------------
% Brian Butler © Copyright 1998
% ------------------------------------------------------

% Clear all variables currently in use
clear;

% Set the random number generator
rand('seed',sum(100*clock));

% Read the analysis parameter file
parameters;

% ---------------------------------------------
% Open the experiment data files
% ---------------------------------------------
OUTFILE = fopen(OutputFilename,'w');
INTFILE = fopen(InterestFilename,'w');

% ---------------------------------------------
% Loop through the cells
%  (Changing the parameters each time)
% ---------------------------------------------
for nIndex = 1:length(nParameters),
for acIndex = 1:length(acParameters),     % Average contribution
for wIndex = 1:length(wParameters),
for cIndex = 1:length(cParameters),
for mIndex = 1:length(mParameters),

% ---------------------------------------------
% Set group parameters
% =============================================
% N : Group Size
% AverageContrib: The average number of messages
%    contributed by an agent in a time period
% PartProb: The probability of an agent
%  contributing a message on a given day.
% wValue : The impact of messages on content perceptions
% cValue : Average cost of noise messages (to individual)
% mValue : Average max positive marginal benefit signal messages
% ---------------------------------------------
N = nParameters(nIndex);
AverageContrib = acParameters(acIndex);
PartProb = AverageContrib;  % The per person, per day participation rate
wValue = wParameters(wIndex);
cValue = cParameters(cIndex);
mValue = mParameters(mIndex);

% Create string containing the parameter to simply results recording
```

```
CellParameterRecord =
sprintf('%d,%.3f,%.3f,%.3f,%d',N,AverageContrib,wValue,cValue,mValue);

% Display a status message
fprintf('Cell: %s\n',CellParameterRecord);

% --------------------------------------------------
% Loop through the groups in the cell
% --------------------------------------------------
for GrpId = 1:CellSize;

% --------------------------------------------------
% Create group
% (i.e. initialize agent parameters)
% ==================================================
% INT(1) = Low point of an agent's interests
% INT(2) = High point of an agent's interests
% w: Message weight (impact on group assessment)
% ce0: Initial beliefs about content
% ve0: Initial beliefs about volume
% c: Message cost
% m: Maximum positive benefit messages
% --------------------------------------------------
w = ones(N,1) * wValue;
ce0 = (rand(N,1) * 0.5) + 0.5;
ve0 = zeros(N,1);
c = ones(N,1) * cValue;
m = ones(N,1) * mValue;

% Randomly select the interest range parameters for the group
INTRange = (rand * (irParameters(2)-irParameters(1))) + irParameters(1);
INT(1:N,1) = rand(N,1);
INT(1:N,2) = INT(1:N,1) + (rand(N,1) * INTRange);

% Create string containing the parameter to simply results recording
GroupParameterRecord =
sprintf('%d,%d,%.3f,%.3f,%.3f,%d,%.3f',InitPeriodLength,N,AverageContrib,wVal
ue,cValue,mValue,INTRange);

% ------------------------------------------------
% Create the operations data structures
% ------------------------------------------------
clear Members Messages;
TotalVolume = 0;
ve = zeros(N,TotalRunLength);
ce = zeros(N,TotalRunLength);
Members = zeros(N,TotalRunLength);

% ------------------------------------------------
% Set Initial Conditions and Run Loop
% ------------------------------------------------
ve(1:N,1) = ve0;
ce(1:N,1) = ce0;
Members(1:N,1) = ones(N,1);
Messages(1:N,1) = ones(N,1) * -1;
MemPer = ce(find(Members(1:N,1)),1);
S = sum(Members(1:N,1));
```

```
t = 2;          % t = 1 are the initial conditions

% Loop until stability is reached or RunLength is reached
while (t <= TotalRunLength),

% Record the group state values at the end of the initialization period
if t == InitPeriodLength
   TrueN = S;
   InitMsgVolume = TotalVolume;
end;

% -----------------------------------------------
% Step 1: Generate Group Communication
% -----------------------------------------------
MsgMarkers = (rand(N,1) - (1 - PartProb) > 0) .* Members(1:N,t-1);

% ----------------------------------------------------
% Step 2: Update Individuals Perceptions of the Group
% ----------------------------------------------------

% Update Volume expectations
MessageCount = sum(MsgMarkers);
TotalVolume = TotalVolume + sum(MessageCount);
ve(1:N,t) = ones(N,1) * (TotalVolume / t);
CurrentVE = ve(1:N,t);

% Update Content perceptions
CurrentCE = ce(1:N,t-1);

% --- Generate messages and update perceptions appropriately ----
SourceList = find(MsgMarkers);
for MsgCnt = 1:length(SourceList);

   % Select a message source
   Source = SourceList(MsgCnt);

   % Generate a message based on the sources interests
   Message = mod((rand * (INT(Source,2) - INT(Source,1))) + INT(Source,1),1);

   % Record the message
   Messages(Source,t) = Message;

   % Assess Reaction to the message
   MsgReaction = (((Message > INT(1:N,1) & (Message < INT(1:N,2))) |
((Message + 1) < INT(1:N,2))) * 2) - 1;

   % Compute the change in attitude due to the message
   CEChange = (MsgReaction .* w) .* (CurrentCE - (CurrentCE.^2));
   CurrentCE = CurrentCE + CEChange;

end;

% Record content perceptions
ce(1:N,t) = CurrentCE;

% -----------------------------------------------
% Step 3: Compute Benefit Expectations
% ---------------------------------------------
EBenefits = (-1./(2*m)) .* ((CurrentCE .* CurrentVE) .^2) + (CurrentCE .*
CurrentVE) - (c .* ((1 - CurrentCE) .* CurrentVE));

% ---------------------------------------------
% Step 4: Update the group membership record
% ---------------------------------------------
Members(1:N,t) = Members(1:N,t-1) .* (EBenefits >= 0);


% ---------------------------------------------
% update operating values
% ---------------------------------------------
MemPer = ce(find(Members(1:N,t)),t);
S = sum(Members(1:N,t));
t = t + 1;

end;    % ----------- End of the Single Group Run ------------------

% -----------------------------------
% Determine group feature measures
% -----------------------------------

% S : Group Size (already computed)

% Compute stability measures
MeanCE = sum(MemPer) / S;
StabilityIndex = min(MemPer);

% Store the stopping time for easy reference
NowT = t-1;

% Determine member and non-member interests
MemInt = INT(find(Members(1:N,NowT)),1:2);
NonMemInt = INT(find(Members(1:N,NowT) == 0),1:2);

% Determine the mean interest range for members and non-members
if S > 0
   MemRange = mean(MemInt(1:S,2) - MemInt(1:S,1));
else
   MemRange = '.';
end;
if N-S > 0
   NonMemRange = mean(NonMemInt(1:(N - S),2) - NonMemInt(1:(N-S),1));
else
   NonMemRange = '.';
end;


% Construct initial and final interest profiles for the group
index = 1;
intlist = [];
for i = 0.05:0.05:1,

   % Store values for the initial profile
   InitDist(index) = sum(((INT(1:N,1) < i) & (INT(1:N,2) > i)) | ((INT(1:N,2)
> 1) & (mod(INT(1:N,2),1) > i)));

   % Store values for the final profile
   IntDist(index) = sum(((MemInt(1:S,1) < i) & (MemInt(1:S,2) > i)) |
((MemInt(1:S,2) > 1) & (mod(MemInt(1:S,2),1) > i)));

   % Add appropriate values to the interest value list (used to test
normality)
   intlist = [intlist;(ones(IntDist(index),1) * i)];

   % Increment the profile index
   index = index + 1;

end;

% Compute topic coverage for the final profile
TopicCoverage = sum((IntDist > 0)) / 20;

% ---------------------------------------------
% Record group state at stability
% ---------------------------------------------
% Record time to stability and size at stability
fprintf(OUTFILE,'%s,%d,%d,%d,%d,%.3f,%.3f,%.2f,%.3f,%.3f,%d,%d\n',GroupParame
terRecord,GrpId,NowT,S,TotalVolume,StabilityIndex,MeanCE,TopicCoverage,MemRan
ge,NonMemRange,TrueN,InitMsgVolume);

% Record Initial Interest Profile
fprintf(INTFILE,'%s,%d,',GroupParameterRecord,GrpId);
fprintf(INTFILE,'%d,',InitDist);
fprintf(INTFILE,'INITDIST\n');

% Record Final Interest Profile
fprintf(INTFILE,'%s,%d,',GroupParameterRecord,GrpId);
fprintf(INTFILE,'%d,',IntDist);
fprintf(INTFILE,'FNLDIST\n');

% Print status message;
if(mod(GrpId,10) == 0),
   fprintf('Group: %d\n',GrpId);
end;

end;    % ----------- End of the Cell Cycle -------------

% -----------------------------------------
% End the parameter loops
% -----------------------------------------
end; % m (Maximum message) Loop
end; % c (noise Cost) Loop
end; % w (message impact) Loop
end; % pp (Participation Probability) Loop
end; % n (initial size) Loop

% -----------------------------------------
% Close all input and output files
% -----------------------------------------
fclose('all');
```

```
---------------------- Final Calibration Run ----------------------------
```

Filename: parameters.m
Description: This file defines the filenames, operational settings, and virtual experiment parameters for the basic model.

```
% ------------------------------------------------------
% This is a template for the parameters of the
% networked collective simulations
% ------------------------------------------------------
% Brian Butler © Copyright 1998
% ------------------------------------------------------

% Define the filenames
OutputFilename = 'rundata.out';
InterestFilename = 'intprofile.out';

% Define the simulation operations paramters
InitPeriodLength = 100;
RunLength = 130;
TotalRunLength = InitPeriodLength + RunLength + 1;
CellSize = 100;

% Define virtual experiment parameters
wParameters = [0.01,0.02,0.03,0.04,0.05];
cParameters = [1];
mParameters = [5];
irParameters = [0.25,0.75];
```

Filename: model.m
Description: This file contains the MATLAB script that implements the model.

```
% ----------------------------------------------------
% This MATLAB script implements the primary
% collective development model.
% ----------------------------------------------------
% Brian Butler © Copyright 1998
% ----------------------------------------------------

% Clear all variables currently in use
clear;

% Set the random number generator
rand('state',sum(100*clock));

% Read the analysis parameter file
parameters;

% --------------------------------------------
% Open the experiment data files
% --------------------------------------------
OUTFILE = fopen(OutputFilename,'w');
INTFILE = fopen(InterestFilename,'w');

% --------------------------------------------
% Loop through the cells
%  (Changing the parameters each time)
% --------------------------------------------
for wIndex = 1:length(wParameters),
for cIndex = 1:length(cParameters),
for mIndex = 1:length(mParameters),

% ---------------------------------------------
% Set cell parameters
% =============================================
% wValue : The impact of messages on content perceptions
% cValue : Average cost of noise messages (to individual)
% mValue : Average max positive marginal benefit signal messages
% ---------------------------------------------
wValue = wParameters(wIndex);
cValue = cParameters(cIndex);
mValue = mParameters(mIndex);

% Create string containing the parameter to simply results recording
CellParameterRecord = sprintf('%.3f,%.3f,%d',wValue,cValue,mValue);

% Display a status message
fprintf('Cell: %s\n',CellParameterRecord);

% ------------------------------------------------
% Loop through the groups in the cell
% ------------------------------------------------
for GrpId = 1:CellSize;

% ------------------------------------------------
% Create group
% (i.e. initialize agent parameters)
% ==================================================
% INT(1) = Low point of an agent's interests
% INT(2) = High point of an agent's interests
% w: Message weight (impact on group assessment)
% ce0: Initial beliefs about content
% ve0: Initial beliefs about volume
% c: Message cost
% m: Maximum positive benefit messages
% --------------------------------------------------

% Set the group size based on a draw from a log-normal distribution
N = 9999;
while (N > 5000 | N == 0)
   N = floor(exp(4.17+1.54*randn));
end;

% Set the participation probabilities
% PartRatio = -0.28677 * log(rand);      % Exponentatial Distribution
PartRatio = -0.17 * log(rand);
PartProb = exp(-4.02 + 0.53*randn);     % Log-Normal Distribution

PersonalPartProb = (rand(N,1) < PartRatio) * PartProb;

w = ones(N,1) * wValue;
ce0 = (rand(N,1) * 0.5) + 0.5;
ve0 = zeros(N,1);
c = ones(N,1) * cValue;
m = ones(N,1) * mValue;

% Randomly select the interest range parameters for the group
INTRange = (rand * (irParameters(2)-irParameters(1))) + irParameters(1);
INT(1:N,1) = rand(N,1);
INT(1:N,2) = INT(1:N,1) + (rand(N,1) * INTRange);

% Create string containing the parameter to simply results recording
GroupParameterRecord =
sprintf('%d,%d,%.3f,%.3f,%.3f,%.3f,%d,%.3f',InitPeriodLength,N,PartRatio,Part
Prob,wValue,cValue,mValue,INTRange);

% ------------------------------------------------
% Create the operations data structures
% ------------------------------------------------
clear Members Messages;
TotalVolume = 0;
ve = zeros(N,TotalRunLength);
ce = zeros(N,TotalRunLength);
Members = zeros(N,TotalRunLength);

% ------------------------------------------------
% Set Initial Conditions and Run Loop
% ------------------------------------------------
ve(1:N,1) = ve0;
ce(1:N,1) = ce0;
Members(1:N,1) = ones(N,1);
Messages(1:N,1) = ones(N,1) * -1;
```

```
MemPer = ce(find(Members(1:N,1)),1);
S = sum(Members(1:N,1));

t = 2;          % t = 1 are the initial conditions

% Loop until stability is reached or RunLength is reached
while (t <= TotalRunLength),

% Record the group state values at the end of the initialization period
if t == InitPeriodLength
   TrueN = S;
   InitMsgVolume = TotalVolume;
end;

% ----------------------------------------------
% Step 1: Generate Group Communication
% ----------------------------------------------
MsgMarkers = (rand(N,1) - (1 - PersonalPartProb) > 0) .* Members(1:N,t-1);

% ---------------------------------------------------
% Step 2: Update Individuals Perceptions of the Group
% ---------------------------------------------------

% Update Volume expectations
MessageCount = sum(MsgMarkers);
TotalVolume = TotalVolume + sum(MessageCount);
ve(1:N,t) = ones(N,1) * (TotalVolume / t);
CurrentVE = ve(1:N,t);

% Update Content perceptions
CurrentCE = ce(1:N,t-1);

% --- Generate messages and update perceptions appropriately ----
SourceList = find(MsgMarkers);
Messages(1:N,t) = ones(N,1) * -1;
for MsgCnt = 1:length(SourceList);

   % Select a message source
   Source = SourceList(MsgCnt);

   % Generate a message based on the sources interests
   Message = mod((rand * (INT(Source,2) - INT(Source,1))) + INT(Source,1),1);

   % Record the message
   Messages(Source,t) = Message;

   % Assess Reaction to the message
   MsgReaction = (((Message > INT(1:N,1) & (Message < INT(1:N,2))) |
((Message + 1) < INT(1:N,2))) * 2) - 1;

   % Compute the change in attitude due to the message
   CEChange = (MsgReaction .* w) .* (CurrentCE - (CurrentCE.^2));
   CurrentCE = CurrentCE + CEChange;

end;

% Record content perceptions
```

```
ce(1:N,t) = CurrentCE;

% ----------------------------------------------
% Step 3: Compute Benefit Expectations
% ----------------------------------------------
EBenefits = (-1./(2*m)) .* ((CurrentCE .* CurrentVE) .^2) + (CurrentCE .*
CurrentVE) - (c .* ((1 - CurrentCE) .* CurrentVE));

% ----------------------------------------------
% Step 4: Update the group membership record
% ----------------------------------------------
Members(1:N,t) = Members(1:N,t-1) .* (EBenefits >= 0);


% ----------------------------------------------
% update operating values
% ----------------------------------------------
MemPer = ce(find(Members(1:N,t)),t);
S = sum(Members(1:N,t));
t = t + 1;

end;    % ----------- End of the Single Group Run ------------------

% ------------------------------------
% Determine group feature measures
% ------------------------------------

% S : Group Size (already computed)

% Compute stability measures
if S > 0
   MeanCE = sum(MemPer) / S;
   StabilityIndex = min(MemPer);
else
   MeanCE = '.';
   StabilityIndex ='.';
end;

% Store the stopping time for easy reference
NowT = t-1;

% Determine member and non-member interests
MemInt = INT(find(Members(1:N,NowT)),1:2);
NonMemInt = INT(find(Members(1:N,NowT) == 0),1:2);

% Determine the mean interest range for members and non-members
if S > 0
   MemRange = mean(MemInt(1:S,2) - MemInt(1:S,1));
else
   MemRange = '.';
end;
if N-S > 0
   NonMemRange = mean(NonMemInt(1:(N - S),2) - NonMemInt(1:(N-S),1));
else
   NonMemRange = '.';
end;
```

```
% Determine the true participation ratio and participation average
% (First determine the participant count)
ParticipantCount = sum((sum((Messages(1:N,InitPeriodLength:TotalRunLength) >=
0)') > 0)');
if TrueN > 0
   TruePartRatio = ParticipantCount/TrueN;
else
   TruePartRatio = 9999;
end;

if ParticipantCount > 0
      TruePartProb = ((TotalVolume - InitMsgVolume)/(TotalRunLength -
InitPeriodLength))/ParticipantCount;
else
   TruePartProb = 9999;
end;

% Construct initial and final interest profiles for the group
index = 1;
intlist = [];
for i = 0.05:0.05:1,

   % Store values for the initial profile
   InitDist(index) = sum(((INT(1:N,1) < i) & (INT(1:N,2) > i)) | ((INT(1:N,2)
> 1) & (mod(INT(1:N,2),1) > i)));

   % Store values for the final profile
   IntDist(index) = sum(((MemInt(1:S,1) < i) & (MemInt(1:S,2) > i)) |
((MemInt(1:S,2) > 1) & (mod(MemInt(1:S,2),1) > i)));

   % Add appropriate values to the interest value list (used to test
normality)
   intlist = [intlist;(ones(IntDist(index),1) * i)];

   % Increment the profile index
   index = index + 1;

end;

% Compute topic coverage for the final profile
TopicCoverage = sum((IntDist > 0)) / 20;

% ---------------------------------------------
% Record group state at stability
% ---------------------------------------------
% Record time to stability and size at stability
fprintf(OUTFILE,'%s,%d,%d,%d,%d,%.3f,%.3f,%.2f,%.3f,%.3f,%d,%d,%.3f,%.3f\n',G
roupParameterRecord,GrpId,NowT,S,TotalVolume,StabilityIndex,MeanCE,TopicCover
age,MemRange,NonMemRange,TrueN,InitMsgVolume,TruePartRatio,TruePartProb);

% Record Initial Interest Profile
fprintf(INTFILE,'%s,%d,',GroupParameterRecord,GrpId);
fprintf(INTFILE,'%d,',InitDist);
fprintf(INTFILE,'INITDIST\n');

% Record Final Interest Profile
fprintf(INTFILE,'%s,%d,',GroupParameterRecord,GrpId);
fprintf(INTFILE,'%d,',IntDist);
fprintf(INTFILE,'FNLDIST\n');

% Print status message;
if(mod(GrpId,10) == 0),
   fprintf('Group: %d\n',GrpId);
end;

end;    % ----------- End of the Cell Cycle --------------

% ----------------------------------------
% End the parameter loops
% ----------------------------------------
end; % m (Maximum message) Loop
end; % c (noise Cost) Loop
end; % w (message impact) Loop

% ----------------------------------------
% Close all input and output files
% ----------------------------------------
fclose('all');
```

Appendix C: Model Validation Scripts

This appendix contains the MATLAB scripts used to perform model validation. The scripts were executed with MATLAB 5.2 on a Windows PC.

Filename: parameters.m
Description: This file defines the filenames, operational settings, and virtual experiment parameters for the basic model.

```
% ------------------------------------------------------
% This is a template for the parameters of the
% networked collective simulations
% ------------------------------------------------------
% Brian Butler © Copyright 1998
% Created: 7/21/98
% ------------------------------------------------------

% Define the filenames
OutputFilename = 'rundata.out';
InterestFilename = 'intprofile.out';

% Define the simulation operations paramters
InitPeriodLength = 100;
RunLength = 130;
TotalRunLength = InitPeriodLength + RunLength + 1;
CellSize = 100;

% The number of times that each set of data will be run
ParameterRunTotal = 10;

% Define virtual experiment parameters
wParameters = [0.02];
cParameters = [1];
mParameters = [5];
irParameters = [0.25,0.75];
```

Filename: model.m
Description: This file contains the MATLAB script that implements the model.

```
% -----------------------------------------------------
% This MATLAB script implements the primary
% collective development model.
% -----------------------------------------------------
% Brian Butler © Copyright 1998
% -----------------------------------------------------

% Clear all variables currently in use
clear;

% Set the random number generator
rand('state',sum(100*clock));

% Read the analysis parameter file
parameters;

% -----------------------------------------------------
% Read the empirical data
% -----------------------------------------------------
RecordLength = 5;
DATAFILE = fopen('sizepartdata.dat');

RecordNumber = 1;
[Record,Count] = fscanf(DATAFILE,'%f',RecordLength);
while Count > 0
   Data(RecordNumber,1:RecordLength) = Record';
   RecordNumber = RecordNumber + 1;
      [Record,Count] = fscanf(DATAFILE,'%f',RecordLength);
end;
fclose(DATAFILE);

% Determine the number of groups
GroupCount = RecordNumber - 1;

% ---------------------------------------------
% Open the experiment data files
% ---------------------------------------------
OUTFILE = fopen(OutputFilename,'w');
INTFILE = fopen(InterestFilename,'w');

% ---------------------------------------------
% Loop through the cells
%  (Changing the parameters each time)
% ---------------------------------------------
for wIndex = 1:length(wParameters),
for cIndex = 1:length(cParameters),
for mIndex = 1:length(mParameters),

% ----------------------------------------------
% Set cell parameters
% ==============================================
% wValue : The impact of messages on content perceptions
% cValue : Average cost of noise messages (to individual)
% mValue : Average max positive marginal benefit signal messages
% ----------------------------------------------
wValue = wParameters(wIndex);
cValue = cParameters(cIndex);
mValue = mParameters(mIndex);

% Create string containing the parameter to simply results recording
CellParameterRecord = sprintf('%.3f,%.3f,%d',wValue,cValue,mValue);

% Display a status message
% fprintf('Cell: %s\n',CellParameterRecord);

% -------------------------------------------------
% Loop through the groups in the cell
% -------------------------------------------------
SSE = 0;
for GrpId = 1:GroupCount;

% -------------------------------------------------
% Create group
% (i.e. initialize agent parameters)
% =================================================
% INT(1) = Low point of an agent's interests
% INT(2) = High point of an agent's interests
% w: Message weight (impact on group assessment)
% ce0: Initial beliefs about content
% ve0: Initial beliefs about volume
% c: Message cost
% m: Maximum positive benefit messages
% -------------------------------------------------

% Set the group size based on the empirical data
N = Data(GrpId,1);

% Set the participation probabilities based on the empirical data
PartRatio = Data(GrpId,2);
PartProb = Data(GrpId,3);

% Construct the individual values
PersonalPartProb = (rand(N,1) < PartRatio) * PartProb;

fprintf('Group %d: ',GrpId);

% -------------------------------------------------
% Run perform the model multiple times for each
%  set of empirical parameters
% -------------------------------------------------
for CompGrpId = 1:ParameterRunTotal;

w = ones(N,1) * wValue;
ce0 = (rand(N,1) * 0.5) + 0.5;
ve0 = zeros(N,1);
c = ones(N,1) * cValue;
m = ones(N,1) * mValue;

% Randomly select the interest range parameters for the group
INTRange = (rand * (irParameters(2)-irParameters(1))) + irParameters(1);
INT(1:N,1) = rand(N,1);
INT(1:N,2) = INT(1:N,1) + (rand(N,1) * INTRange);

% Create string containing the parameter to simply results recording
GroupParameterRecord =
sprintf('%d,%d,%.3f,%.3f,%.3f,%.3f,%d,%.3f',InitPeriodLength,N,PartRatio,Part
Prob,wValue,cValue,mValue,INTRange);

% ---------------------------------------------
% Create the operations data structures
% ---------------------------------------------
clear Members Messages;
TotalVolume = 0;
ve = zeros(N,TotalRunLength);
ce = zeros(N,TotalRunLength);
Members = zeros(N,TotalRunLength);

% ---------------------------------------------
% Set Initial Conditions and Run Loop
% ---------------------------------------------
ve(1:N,1) = ve0;
ce(1:N,1) = ce0;
Members(1:N,1) = ones(N,1);
Messages(1:N,1) = ones(N,1) * -1;
MemPer = ce(find(Members(1:N,1)),1);
S = sum(Members(1:N,1));

t = 2;          % t = 1 are the initial conditions

% Loop until stability is reached or RunLength is reached
while (t <= TotalRunLength),

% Record the group state values at the end of the initialization period
if t == InitPeriodLength
   TrueN = S;
   InitMsgVolume = TotalVolume;
end;

% ---------------------------------------------
% Step 1: Generate Group Communication
% ---------------------------------------------
MsgMarkers = (rand(N,1) - (1 - PersonalPartProb) > 0) .* Members(1:N,t-1);

% -------------------------------------------------
% Step 2: Update Individuals Perceptions of the Group
% -------------------------------------------------

% Update Volume expectations
MessageCount = sum(MsgMarkers);
TotalVolume = TotalVolume + sum(MessageCount);
ve(1:N,t) = ones(N,1) * (TotalVolume / t);
CurrentVE = ve(1:N,t);

% Update Content perceptions
CurrentCE = ce(1:N,t-1);

% --- Generate messages and update perceptions appropriately ----
SourceList = find(MsgMarkers);
```

```
Messages(1:N,t) = ones(N,1) * -1;
for MsgCnt = 1:length(SourceList);

   % Select a message source
   Source = SourceList(MsgCnt);

   % Generate a message based on the sources interests
   Message = mod((rand * (INT(Source,2) - INT(Source,1))) + INT(Source,1),1);

   % Record the message
   Messages(Source,t) = Message;

   % Assess Reaction to the message
   MsgReaction = (((Message > INT(1:N,1) & (Message < INT(1:N,2))) |
((Message + 1) < INT(1:N,2))) * 2) - 1;

   % Compute the change in attitude due to the message
   CEChange = (MsgReaction .* w) .* (CurrentCE - (CurrentCE.^2));
   CurrentCE = CurrentCE + CEChange;

end;

% Record content perceptions
ce(1:N,t) = CurrentCE;

% ----------------------------------------------
% Step 3: Compute Benefit Expectations
% ----------------------------------------------
EBenefits = (-1./(2*m)) .* ((CurrentCE .* CurrentVE) .^2) + (CurrentCE .*
CurrentVE) - (c .* ((1 - CurrentCE) .* CurrentVE));

% ----------------------------------------------
% Step 4: Update the group membership record
% ----------------------------------------------
Members(1:N,t) = Members(1:N,t-1) .* (EBenefits >= 0);


% ----------------------------------------------
% update operating values
% ----------------------------------------------
MemPer = ce(find(Members(1:N,t)),t);
S = sum(Members(1:N,t));
t = t + 1;

end;    % ---------- End of the Single Group Run -----------------

% -----------------------------------
% Determine group feature measures
% -----------------------------------

% S : Group Size (already computed)

% Compute stability measures
if S > 0
   MeanCE = sum(MemPer) / S;
   StabilityIndex = min(MemPer);
else
```

```
   MeanCE = '.';
   StabilityIndex ='.';
end;

% Store the stopping time for easy reference
NowT = t-1;

% Determine member and non-member interests
MemInt = INT(find(Members(1:N,NowT)),1:2);
NonMemInt = INT(find(Members(1:N,NowT) == 0),1:2);

% Determine the mean interest range for members and non-members
if S > 0
   MemRange = mean(MemInt(1:S,2) - MemInt(1:S,1));
else
   MemRange = '.';
end;
if N-S > 0
   NonMemRange = mean(NonMemInt(1:(N - S),2) - NonMemInt(1:(N-S),1));
else
   NonMemRange = '.';
end;

% Determine the true participation ratio and participation average
% (First determine the participant count)
ParticipantCount = sum((sum((Messages(1:N,InitPeriodLength:TotalRunLength) >=
0)') > 0)');
if TrueN > 0
   TruePartRatio = ParticipantCount/TrueN;
else
   TruePartRatio = 9999;
end;

if ParticipantCount > 0
      TruePartProb = ((TotalVolume - InitMsgVolume)/(TotalRunLength -
InitPeriodLength))/ParticipantCount;
else
   TruePartProb = 9999;
end;

% Construct initial and final interest profiles for the group
index = 1;
intlist = [];
for i = 0.05:0.05:1,

   % Store values for the initial profile
   InitDist(index) = sum(((INT(1:N,1) < i) & (INT(1:N,2) > i)) | ((INT(1:N,2)
> 1) & (mod(INT(1:N,2),1) > i)));

   % Store values for the final profile
   IntDist(index) = sum(((MemInt(1:S,1) < i) & (MemInt(1:S,2) > i)) |
((MemInt(1:S,2) > 1) & (mod(MemInt(1:S,2),1) > i)));

   % Add appropriate values to the interest value list (used to test
normality)
   intlist = [intlist;(ones(IntDist(index),1) * i)];
   % Increment the profile index
   index = index + 1;

end;

% Compute topic coverage for the final profile
TopicCoverage = sum((IntDist > 0)) / 20;

% Calculate percentage loss and message volume
PLoss(GrpId) = (N - S)/N;
DailyVolume(GrpId) = TotalVolume/NowT;

% Store the Mean Squared Error
SSE = SSE + (Data(GrpId,4) - PLoss(GrpId))^2;

% ----------------------------------------------
% Record group state at stability
% ----------------------------------------------
% Record time to stability and size at stability
fprintf(OUTFILE,'%s,%d,%d,%d,%d,%d,%.3f,%.3f,%.2f,%.3f,%.3f,%d,%d,%.3f,%.3f,'
,GroupParameterRecord,GrpId,CompGrpId,NowT,S,TotalVolume,StabilityIndex,MeanC
E,TopicCoverage,MemRange,NonMemRange,TrueN,InitMsgVolume,TruePartRatio,TruePa
rtProb);
fprintf(OUTFILE,'%.3f,%.3f,%.3f,%.3f\n',PLoss(GrpId),Data(GrpId,4),DailyVolum
e(GrpId),Data(GrpId,5));

% Record Initial Interest Profile
% fprintf(INTFILE,'%s,%d,',GroupParameterRecord,GrpId);
% fprintf(INTFILE,'%d,',InitDist);
% fprintf(INTFILE,'INITDIST\n');

% Record Final Interest Profile
% fprintf(INTFILE,'%s,%d,',GroupParameterRecord,GrpId);
% fprintf(INTFILE,'%d,',IntDist);
% fprintf(INTFILE,'FNLDIST\n');

fprintf('%d..',CompGrpId);

end;    % ----------- End of the data set cycle ----------

fprintf('\n');

% Print status message;
%if(mod(GrpId,10) == 0),
%   fprintf('Group: %d\n',GrpId);
%end;

end;    % ----------- End of the Cell Cycle --------------

% ----------------------------------------
% End the parameter loops
% ----------------------------------------
end; % m (Maximum message) Loop
end; % c (noise Cost) Loop
end; % w (message impact) Loop

MSE = SSE / GroupCount;
```

```
fprintf('MSE: %.3f\n',MSE);

% ----------------------------------------
% Close all input and output files
% ----------------------------------------
fclose('all');
```

Appendix D: Model Analysis and Virtual Experiment Scripts

This appendix contains the MATLAB scripts used to perform the virtual experiment in order to analyze the computational model. The scripts were executed with MATLAB 5.2 on a Windows PC.

Filename: parameters.m
Description: This file defines the filenames, operational settings, and virtual experiment parameters for the basic model.

```
% -------------------------------------------------------
% This is a template for the parameters of the
% networked collective simulations
% -------------------------------------------------------
% Brian Butler © Copyright 1998
% Created: 7/21/98
% -------------------------------------------------------

% Define the filenames
OutputFilename = 'rundata.out';
InterestFilename = 'intprofile.out';

% Define the simulation operations paramters
InitPeriodLength = 100;
RunLength = 365;
TotalRunLength = InitPeriodLength + RunLength + 1;
CellSize = 100;

% Define virtual experiment parameters
wParameters = [0.005,0.02,0.1];
cParameters = [0.33,1,3];
mParameters = [2,5,8];
irParameters = [0.25,0.75];
```

Filename: model.m
Description: This file contains the MATLAB script that implements the model.

```
% ------------------------------------------------------
% This MATLAB script implements the primary
% collective development model.
% ------------------------------------------------------
% Brian Butler © Copyright 1998
% ------------------------------------------------------

% Clear all variables currently in use
clear;

% Set the random number generator
rand('state',sum(100*clock));

% Read the analysis parameter file
parameters;

% ----------------------------------------------
% Open the experiment data files
% ----------------------------------------------
OUTFILE = fopen(OutputFilename,'w');
INTFILE = fopen(InterestFilename,'w');

% ----------------------------------------------
% Loop through the cells
%  (Changing the parameters each time)
% ----------------------------------------------
for wIndex = 1:length(wParameters),
for cIndex = 1:length(cParameters),
for mIndex = 1:length(mParameters),

% ----------------------------------------------
% Set cell parameters
% ==============================================
% wValue : The impact of messages on content perceptions
% cValue : Average cost of noise messages (to individual)
% mValue : Average max positive marginal benefit signal messages
% ----------------------------------------------
wValue = wParameters(wIndex);
cValue = cParameters(cIndex);
mValue = mParameters(mIndex);

% Create string containing the parameter to simply results recording
CellParameterRecord = sprintf('%.3f,%.3f,%d',wValue,cValue,mValue);

% Display a status message
fprintf('Cell: %s\n',CellParameterRecord);

% -------------------------------------------------
% Loop through the groups in the cell
% -------------------------------------------------
for GrpId = 1:CellSize;

% -------------------------------------------------
% Create group
% (i.e. initialize agent parameters)
% =================================================
% INT(1) = Low point of an agent's interests
% INT(2) = High point of an agent's interests
% w: Message weight (impact on group assessment)
% ce0: Initial beliefs about content
% ve0: Initial beliefs about volume
% c: Message cost
% m: Maximum positive benefit messages
% -------------------------------------------------

% Set the group size based on a draw from a log-normal distribution
N = 9999;
while (N > 2000 | N == 0)
   N = floor(exp(4.17+1.54*randn));
end;

% Set the participation probabilities
% PartRatio = -0.28677 * log(rand);       % Exponentatial Distribution
PartRatio = -0.17 * log(rand);
PartProb = exp(-4.02 + 0.53*randn);     % Log-Normal Distribution

PersonalPartProb = (rand(N,1) < PartRatio) * PartProb;

w = ones(N,1) * wValue;
ce0 = (rand(N,1) * 0.5) + 0.5;
ve0 = zeros(N,1);
c = ones(N,1) * cValue;
m = ones(N,1) * mValue;

% Randomly select the interest range parameters for the group
INTRange = (rand * (irParameters(2)-irParameters(1))) + irParameters(1);
INT(1:N,1) = rand(N,1);
INT(1:N,2) = INT(1:N,1) + (rand(N,1) * INTRange);

% Create string containing the parameter to simply results recording
GroupParameterRecord =
sprintf('%d,%d,%.3f,%.3f,%.3f,%.3f,%d,%.3f',InitPeriodLength,N,PartRatio,Part
Prob,wValue,cValue,mValue,INTRange);

% -----------------------------------------------
% Create the operations data structures
% -----------------------------------------------
clear Members Messages;
TotalVolume = 0;
Members = zeros(N,TotalRunLength);

% -----------------------------------------------
% Set Initial Conditions and Run Loop
% -----------------------------------------------
ve = ve0;
ce = ce0;
Members(1:N,1) = ones(N,1);
Messages(1:N,1) = ones(N,1) * -1;
MemPer = ce(find(Members(1:N,1)));
S = sum(Members(1:N,1));
```

```
t = 2;          % t = 1 are the initial conditions

% Loop until stability is reached or RunLength is reached
while (t <= TotalRunLength),

% Record the group state values at the end of the initialization period
if t == InitPeriodLength
   TrueN = S;
   InitMsgVolume = TotalVolume;
end;

% ---------------------------------------------
% Step 1: Generate Group Communication
% ---------------------------------------------
MsgMarkers = (rand(N,1) - (1 - PersonalPartProb) > 0) .* Members(1:N,t-1);

% --------------------------------------------------
% Step 2: Update Individuals Perceptions of the Group
% --------------------------------------------------

% Update Volume expectations
MessageCount = sum(MsgMarkers);
TotalVolume = TotalVolume + sum(MessageCount);
ve = ones(N,1) * (TotalVolume / t);
CurrentVE = ve;

% Update Content perceptions
CurrentCE = ce;

% --- Generate messages and update perceptions appropriately ----
SourceList = find(MsgMarkers);
Messages(1:N,t) = ones(N,1) * -1;
for MsgCnt = 1:length(SourceList);

   % Select a message source
   Source = SourceList(MsgCnt);

   % Generate a message based on the sources interests
   Message = mod((rand * (INT(Source,2) - INT(Source,1))) + INT(Source,1),1);

   % Record the message
   Messages(Source,t) = Message;

   % Assess Reaction to the message
   MsgReaction = (((Message > INT(1:N,1) & (Message < INT(1:N,2))) |
((Message + 1) < INT(1:N,2))) * 2) - 1;

   % Compute the change in attitude due to the message
   CEChange = (MsgReaction .* w) .* (CurrentCE - (CurrentCE.^2));
   CurrentCE = CurrentCE + CEChange;

end;

% Record content perceptions
ce = CurrentCE;

% ---------------------------------------------
% Step 3: Compute Benefit Expectations
% -----------------------------------------------
EBenefits = (-1./(2*m)) .* ((CurrentCE .* CurrentVE) .^2) + (CurrentCE .*
CurrentVE) - (c .* ((1 - CurrentCE) .* CurrentVE));

% -----------------------------------------------
% Step 4: Update the group membership record
% -----------------------------------------------
Members(1:N,t) = Members(1:N,t-1) .* (EBenefits >= 0);


% -----------------------------------------------
% update operating values
% -----------------------------------------------
MemPer = ce(find(Members(1:N,t)));
S = sum(Members(1:N,t));
t = t + 1;

end;    % ----------- End of the Single Group Run -------------------

% ------------------------------------
% Determine group feature measures
% ------------------------------------

% S : Group Size (already computed)

% Compute stability measures
if S > 0
   MeanCE = sum(MemPer) / S;
   StabilityIndex = min(MemPer);
else
   MeanCE = '.';
   StabilityIndex ='.';
end;

% Store the stopping time for easy reference
NowT = t-1;

% Determine member and non-member interests
MemInt = INT(find(Members(1:N,NowT)),1:2);
NonMemInt = INT(find(Members(1:N,NowT) == 0),1:2);

% Determine the mean interest range for members and non-members
if S > 0
   MemRange = mean(MemInt(1:S,2) - MemInt(1:S,1));
else
   MemRange = '.';
end;
if N-S > 0
   NonMemRange = mean(NonMemInt(1:(N - S),2) - NonMemInt(1:(N-S),1));
else
   NonMemRange = '.';
end;

% Determine the true participation ratio and participation average
% (First determine the participant count)
```

```
ParticipantCount = sum((sum((Messages(1:N,InitPeriodLength:TotalRunLength) >=
0)') > 0)');
if TrueN > 0
   TruePartRatio = ParticipantCount/TrueN;
else
   TruePartRatio = 9999;
end;

if ParticipantCount > 0
      TruePartProb = ((TotalVolume - InitMsgVolume)/(TotalRunLength -
InitPeriodLength))/ParticipantCount;
else
   TruePartProb = 9999;
end;

% Construct initial and final interest profiles for the group
index = 1;
intlist = [];
for i = 0.05:0.05:1,

   % Store values for the initial profile
   InitDist(index) = sum(((INT(1:N,1) < i) & (INT(1:N,2) > i)) | ((INT(1:N,2)
> 1) & (mod(INT(1:N,2),1) > i)));

   % Store values for the final profile
   IntDist(index) = sum(((MemInt(1:S,1) < i) & (MemInt(1:S,2) > i)) |
((MemInt(1:S,2) > 1) & (mod(MemInt(1:S,2),1) > i)));

   % Add appropriate values to the interest value list (used to test
normality)
   intlist = [intlist;(ones(IntDist(index),1) * i)];

   % Increment the profile index
   index = index + 1;

end;

% Compute topic coverage for the final profile
TopicCoverage = sum((IntDist > 0)) / 20;

% ---------------------------------------------
% Record group state at stability
% ---------------------------------------------
% Record time to stability and size at stability
fprintf(OUTFILE,'%s,%d,%d,%d,%d,%.3f,%.3f,%.2f,%.3f,%.3f,%d,%d,%.3f,%.3f\n',G
roupParameterRecord,GrpId,NowT,S,TotalVolume,StabilityIndex,MeanCE,TopicCover
age,MemRange,NonMemRange,TrueN,InitMsgVolume,TruePartRatio,TruePartProb);

% Record Initial Interest Profile
fprintf(INTFILE,'%s,%d,',GroupParameterRecord,GrpId);
fprintf(INTFILE,'%d,',InitDist);
fprintf(INTFILE,'INITDIST\n');

% Record Final Interest Profile
fprintf(INTFILE,'%s,%d,',GroupParameterRecord,GrpId);
fprintf(INTFILE,'%d,',IntDist);
fprintf(INTFILE,'FNLDIST\n');
```

```
% Print status message;
if(mod(GrpId,10) == 0),
   fprintf('Group: %d\n',GrpId);
end;

end;    % ----------- End of the Cell Cycle --------------

% ----------------------------------------
% End the parameter loops
% ----------------------------------------
end; % m (Maximum message) Loop
end; % c (noise Cost) Loop
end; % w (message impact) Loop

% ----------------------------------------
% Close all input and output files
% ----------------------------------------
fclose('all');
```