Quantifying the Skewed Distribution of Activity in Virtual Communities within a longitudinal study

Thomas Schoberth, Armin Heinzl, Sheizaf Rafaeli

Working Paper 03 / 2009
May 2009

- Currently under revision -

Working Papers in Business Administration and Information Systems

University of Mannheim
Department of General Management and Information Systems
D-68131 Mannheim/Germany
Phone +49 621 1811691, Fax +49 621 1811692
E-Mail: wifo1@uni-mannheim.de
Internet: http://wifo1.bwl.uni-mannheim.de/

Quantifying the Skewed Distribution of Activity in Virtual Communities within a longitudinal study

Thomas Schoberth (University of Bayreuth),
Armin Heinzl (University of Mannheim),
Sheizaf Rafaeli (University of Haifa)

Abstract

Virtual Communities (ViCs) are subject of interest for quite some time now [Hagel and Armstrong, 1997, Garton, et al., 1997, Rheingold, 1993]. Recently, ViCs in the form of Social Network Services like MySpace or StudiVz received a lot of attention. Though, fairly little is known about the temporal evolution of virtual communities and the changes in the communication activity of its users. The research described here is an explorative study examining the communication activity of members of two virtual communities on a longitudinal basis.

For that purpose, a graph theoretical model by Pennock et al. [2002] is used which unites "Random Network Theory" and "Scale-free Networks". This model allows the operationalization of the empirical distribution functions of the communication activity in ViCs with only one free parameter. That parameter – the mixing factor α – represents the ratio between the antagonists heterogeneity and homogeneity.

The nonlinear curve fitting of the empirical distribution functions shows a predominance of preferential over uniform binding in both communities. Participants prefer to communicate with community members having already a lot of communication partners, while members with low activity are less attractive. This phenomenon is less strong in the smaller ViC B. The members of ViC B have almost twice as many connections as those of ViC A. ViC B represent a tighter network which might lead to the more homogeneous distribution of its activity. In both communities the mixing factor α and therefore the level of heterogeneity shows quite stable over time.

References: virtual communities, social network analysis, degree, communication activity, heterogeneity, preferential binding, uniform binding, random graph theory, scale-free networks.

1 Introduction

The term Virtual Community (ViC) was introduced as early as 1968 by the Internet pioneers J.C.R. Licklider and Robert W. Taylor [1968]. ViCs reach back to the emergence of the Internet. ViCs appeared as mailing lists in 1975 and as newsgroups in 1979 [Zakon, 2003]. They were used at first by scientists for thought and information exchange, a fact that shaped the understanding of ViCs as social communities without commercial focus [Rheingold, 1993]. With the diffusion of the Internet and its accompanied commercialization, ViCs were also discovered for economic interests [Hagel and Armstrong, 1997]. Brown et al. [2001] showed for example that visitors of websites, appearing as active members of these communities, visited these sites nine times more often and bought nearly twice as much there, compared to those, who did not use these communities.

Virtual communities are overt phenomena. They assemble by communication, and leave behind multiple artifacts such as listserv postings, web site structures, Usenet content, user logs etc. These artifacts are available for scrutiny and research [Jones and Rafaeli, 1999], and should be viewed as a challenge for researchers. Despite the accumulation of artifacts, surprisingly little is known about how community activity develops over time. There exists scant theoretically and empirically supported explanation of this communication activity. Most extant empirical studies are static cross-sectional analyses. This paper therefore aims to give impulses to theory by empirically examining the communication activity of two exemplary virtual communities in the context of a comparative longitudinal profile study [Schoberth, et al., 2006].

It is often pointed out in literature [Jones and Rafaeli, 1999, Jones et al., 2004, Light and Rogers, 1999, Nonnecke and Preece, 2000a, Schoberth, 2002, Schoberth, et al., 2003, Stegbauer, 2001, Whittaker, et al., 1998] that a small number of participants are responsible for the majority of messages in ViCs, while most write only one or few messages. Although such a strong imbalance seems to be important, this effect has been given little formal and empirical expression in the literature. Therefore, the goal of this paper is to quantify this skewed distribution in participation and observe it over time.

1.1 Research topic

1.1.1 Virtual Communities

ViCs are the main research object in this article. Despite a multiplicity of attempts [Rheingold, 1993, Hagel and Armstrong, 1997, Figallo, 1998, Schoberth and Schrott, 2001], there exists no

generally accepted definition for ”community”, much less for “virtual community” [Preece, 2000]. Here, the term ViC is used to represent ongoing communication gatherings and social interaction of groups and larger aggregates of individuals in the Internet that use tools such as web-based forums, list servers, newsgroups and chats.

The literature provides a set of categorization attempts for ViCs. Hagel and Armstrong [1997] differentiate between “communities of interest“, “communities of relationship“, “communities of fantasy“ and “communities of transaction”. It is assumed that communities of different types and different objectives differ strongly [Preece, 2000, Rheingold, 1993]. However, there are also references for common characteristics of ViCs [Whittaker, et al., 1998, Stegbauer, 2001, Preece, 2000, Brunold, et al., 2000].

1.1.2 Attributive vs. Relational Communication Activity

Since ViCs consist of humans who use electronic platforms as means for communication and meeting and not as ends or goals [Preece, 2000], the communication activity of these users should be the main focus of an examination of ViCs. According to Stegbauer [2001], we can differentiate between attributive and relational characteristics of users.

Relational activity focuses on the interaction among the participants and may also be called interactivity. The interactions of users in ViCs assume the form of discussion threads. These specified “threads“ are a tree-like visualization of the discussion topics and represent the sequence and dependencies of the messages. Social Network Analysis is used as a tool for the analysis of these relations [Wellman, et al., 1996, Wellman, 1997]. Analog to Pennock et al. [2002] as well as Albert and Barabási [Albert and Barabási 1999, Albert and Barabási 2002, Albert, et al., 1999] and Holme et al. [2004], the relational activity can be operationalized with the help of the *average degree*, that is in the context of ViCs the *average number of communication partners per active user*.

Individual characteristics of the users will be represented by attributive activity. These characteristics refer to the level of the individual. However, to research a lot of users at a time, they are typically aggregated over all actors. According to Whittaker et al. [1998], the *average number of messages per active user* will be utilized in the following.

1.1.3 Virtual Communities examined

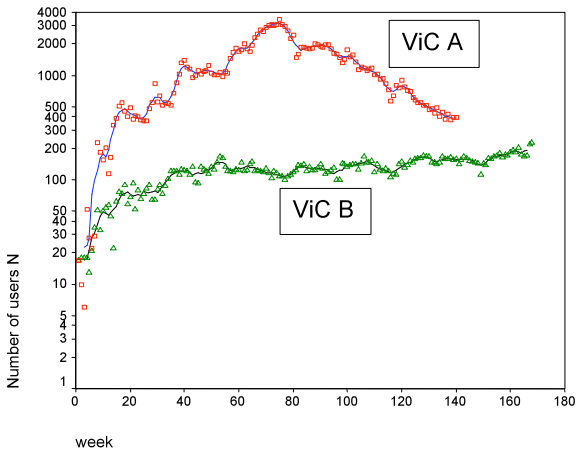

Illustration 1: Course of the number of active users N of the two ViCs A and B.

The two virtual communities examined in this paper use web-based forums as communication platform. In contrast to other basic types of asynchronous platforms (mainly email-based list servers and newsgroups), web-based forums are located on central servers and their data is usually archived in a coherent form for a longer period of time. Despite this advantage, web-based forums have rarely been subject of empirical research; presumably, because they are run and owned by enterprises or organizations. For this reason, it is more difficult to access and obtain their data compared to public newsgroups and list servers that are easily traceable via subscription. Web-based forums have not been investigated as frequently as other public, open interaction spaces [Jones et al., 2004], but they are widespread and popular. For example, parsimony.net hosts more than a thousand forums and more than two hundred of them have at least a thousand page views per day [Parsimony, 2004]. We consider web-based forums to be a useful target for researching the longitudinal behavior of ViC participants.

The first forum ("ViC A") is operated by a German financial service provider. The financial service provider hopes to stimulate stock trade volume and achieve a higher rate of customer retention by the use of this forum. The available data covers 1.03 million postings from 33,536 active members in a period of nearly three years (140 weeks).

The second Virtual Community ("ViC B") discusses stocks and securities as well. However, this community is part of a website whose operator performs as a financial expert, selling information about the occurrences in the German and international financial marketplace. The website's archive, accessible via the World Wide Web, covers more than three years (169 weeks) of data and contains 188,000 messages from 1,456 posters.

Illustration 1 shows the course of the number of active users *N* per monitored week. Nonnecke and Preece [2000b] state that users, who are not posting but do merely reading, may understand

themselves as part of the community. Though, it is not possible here to include those users because of the lack of traceable data. Therefore, a user is seen as active, if he is writing a message and thereby leaves an artifact, which can be analyzed.

Time axes of both ViCs were superimposed on the same figure. Since the two communities started their operations at different points of time (ViC A: 10/1998 and ViC B: 2/2000), "week 100" for example refers to the temporal distance of a hundred weeks after the foundation of the ViC but not to the same absolute point of time. A centered five-week average illustrates the time series more transparently as an added thin solid line (Illustrations 1, 3, 6, and 8).

Considering the logarithmic scale of Illustration 1, it becomes evident that ViC A is substantially larger than ViC B (in the temporal means, $N_A/N_B = 9.6 \pm 0.5$). In ViC A, up to 3,405 users per week are active, whereas in ViC B there were at most 228 users. Up to week 75, the number of participants within ViC A rose strongly and dropped just as sharply thereafter. ViC B in contrast displayed a clearly weaker rise which continued over the entire observation period.

1.2 Research Questions and Course of Action

The following research questions are raised:

1. How can the distribution of the user's activity be quantified?
2. How do the obtained indicators change over time?

By investigating these questions, we expect to give useful impulses for the theory building on ViCs. The development of quantitative measures allows the comparison and evaluation of different communities. A suitable model for the quantitative description of the heterogeneity in ViCs is expected to provide first evidence for its causes. The investigation of the longitudinal perspective might give further clues.

Virtual communities will be mapped as social networks to assess their relational level (section 2). After the operationalization of the relational communication activity (section 3), three graph-theoretical models are introduced. The models' suitability for the description of heterogeneity in virtual communities is compared. In section 4 the results are adopted to the attributive communication activity. Then, in section 5, the development of the heterogeneity of the two ViCs is examined over time. Last, a summary is given.

2 Virtual Communities as Social Networks

According to Wellman et al. [1996] and Wellman [1997] ViCs can be considered as social networks. In a graphic visualization, the members are represented by nodes; the edges symbolize relations between the members (Illustration 2). This view of virtual communities as networks allows the utilization of Social Network Analysis (SNA) methods. The SNA methodology provides from a macro-perspective insights into the internal structure of virtual communities. It supports the analysis of social structures as a whole, where an investigation of the behavior and surroundings of community members, e.g. the micro-perspective, is not feasible within large communities.

For network analysis, the investigation of relations between individuals is fundamental. Relations can be characterized by means of contents, direction and strength [Garton, et al., 1997, Jansen, 1999, Wellman, et al., 1996, Wellman, 1997, Yoshioka, et al., 2001]. In the context of virtual communities, different kinds of information are exchanged, like administrative, private, professional as well as social information. Letters or emails represent directional messages which can be precisely assigned to a sender and a receiver. In contrast, the directionality of relations between persons is frequently difficult to determine through the reciprocity of sending and receiving messages. The strength of relations can be determined according to Jansen [1999, p.53] by its frequency, its importance for the individual, and according to the amount of resources transferred. A relationship between two participants will be called binary, if it is modeled by the two states “existing“ or “non-existing“.

In this paper, relations between participants are regarded as binary and non-directional. For simplicity’s sake, they are also observed as non-cumulative. The contents of relations [Yoshioka, et al., 2001] were not considered due to the vast number of messages. To enable examination of temporal changes in the social network, data were collected weekly.

In Illustration 2 the entire social network of ViC A at the fourth week after its establishment is visualized as a graph. At that time ViC A was still small. At later times with some hundreds or thousands participants it would be hard to represent the network graphically. The nodes represent the active (writing) members. The connections between these nodes, e.g. the edges, are based on the members’ meeting in one or more discussion threads.

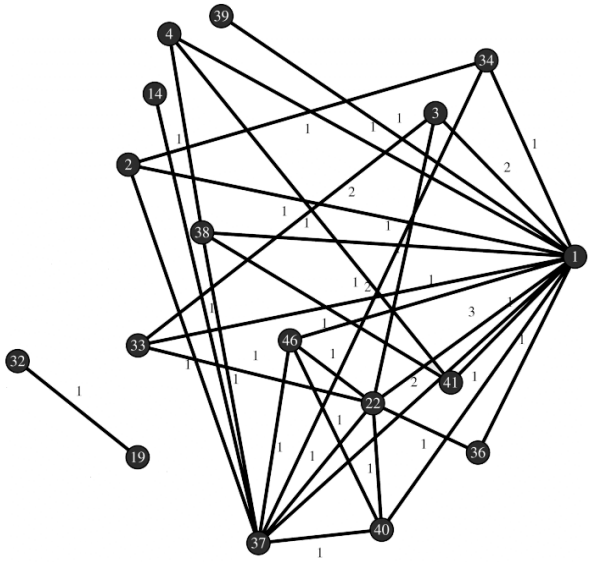

Illustration 2: Visualization of the social network of ViC A in the fourth week after its establishment (created with UCINET VI). The nodes represent active members; the digits on the edges represent the number of discussion threads in which connected members were communicating which each other.

3 Relational Activity

The Social Network Analysis approach provides a set of metrics that describe the relational structure of networks [Garton, et al., 1997, Jansen, 1999, Wassermann and Faust, 1994]. However, these instruments are mainly suitable for the description of small networks (orders of magnitude of approximately 10 persons), since they focus on the relations of individual members (see e.g. Aviv et al. [2003]). Other methods, like Block Model Analysis (see e.g. Stegbauer [2001]), possess a limited suitability for the investigation of temporal changes since its complexity and multilayerdness. For a longitudinal study a suitable indicator should be measurable on a metric scale and should be useful as indicator for the individual as well as for the community as whole.

Here the degree centrality *k* is used as an individual's measure of relational activity [Wassermann and Faust, 1994]. The degree *k* of a network node is defined as the number of edges *k* that a node possesses. In the context of an online community, *k* represents the number of actors a member communicates with via messages in one or more threads.

In order to transfer this individual indicator into indicators for the community as whole, first, the average degree is introduced and its changes over time will be investigated. Subsequently, three

graph-theoretical approaches are examined in order to identify a suitable analytical model which is able to represent the extremely skewed distribution of the degree of network.

3.1 Average Degree

Aggregating the degree k leads to the average degree $\langle k \rangle$ representing the level of connectedness in the network or the community as whole. The average degree is calculated adding the individual's degree k_i or simpler, by dividing the total sum of non-directional connections K by the size of the network N (number of the active users) as shown in Equation 1.

Equation 1: average degree

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2K}{N}$$

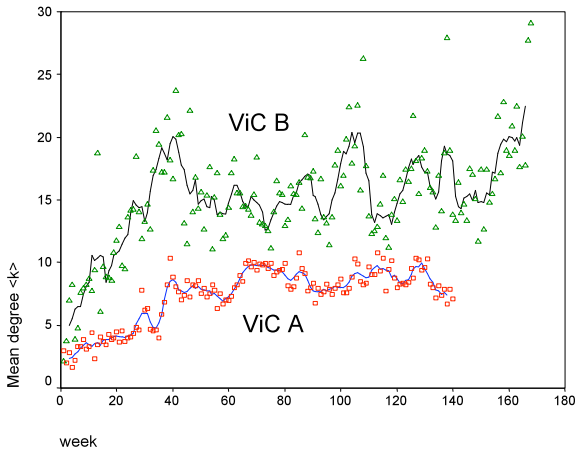

Illustration 3: Plot of the average degree <k> of the ViCs A and B.

The comparison of the two time series in Illustration 3 indicates that the level of interconnectedness and, thus, the relational communication activity of ViC B, are clearly higher than in ViC A. The participants of ViC A have on temporal means an average degree $\langle k_A \rangle = 7.4 \pm 0.2$. In contrast, ViC B has an average degree $\langle k_B \rangle = 15.2 \pm 0.3$. On temporal means $\langle k_B \rangle / \langle k_A \rangle = 2.13 \pm 0.07$ and so the average degree in ViC B is about twice as large as in ViC A. In the course of both communities, $\langle k \rangle$ first rises and seems to stabilize after approximately week 40.

3.2 Degree Distribution

In Illustrations 4a and b, the degree distribution of the two communities is represented by one exemplary week for ViC A (week 129) and ViC B (week 145). For each degree k found, the value of the empirical probability function $P(k) = N(k)/N$ [Bronstein and Semendjajew, 1989,

p.678ff] is plotted. *N(k)* is the number of participants which have *k* edges. *N* is the total number of the participants in these two weeks (N_A = 147 and N_B = 447). In other words: *P(k)* is the empirical probability that any member of the community communicates with *k* other members during one week.

One recognizes that the distributions are extremely skewed. A small portion of members possess a large number of connections. At the same time, the majority of the participants hold only very few connections. In the following, graph-theoretical approaches will be examined in order to identify a suitable analytical model able to fit these distributions as well as to explain their causes. This will be conducted on the basis of the fundamental theorem of the mathematical statistics which denotes that for large samples, the empirical function converges towards the actual distribution function [Bronstein and Semendjajew, 1989, p.79]. The terms distribution and probability function are used synonymously in this context.

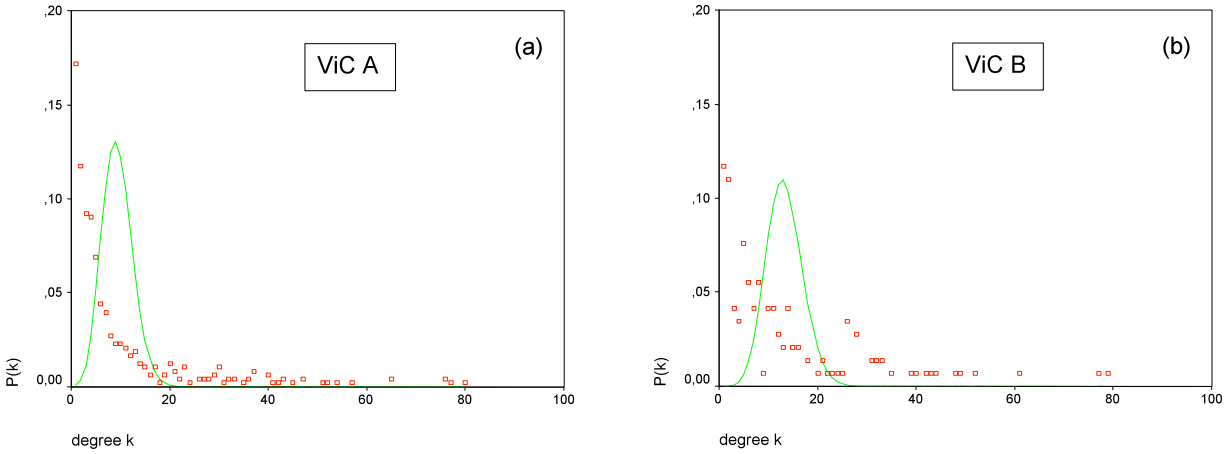

Illustration 4: Histograms of the degree distributions in the 129th week of ViC A and the 145th week of ViC B. P(k) = N(k)/N with N_A = 147 and N_B = 447. The solid lines represent distributions of Poisson (Equation 2) according to the empirical average values of the two distributions (<k_A> = 9.4 and <k_B> = 13.3).

3.2.1 "Random Graph Theory"

The Random Graph Theory assumes a network developing completely randomly. In such a case, all nodes of the graph would have the same probability to obtain *k* connections ("uniform attachment") [Albert and Barabási, 2002]. For large *N,* this would lead to the distribution of Poisson as defined in Equation 2. Such a distribution can be regarded as homogeneous, since its values are randomly (homogeneously) distributed around the mean value <*k*>.

Equation 2: Poisson distribution

$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$

In Illustration 4, distributions of Poisson are plotted as solid lines according to Equation 2 with the empirical average degrees of $<k_A> = 9.4$ and $<k_B> = 13$. The clearly visible deviations of the ideal distribution to the empirical ones indicate that the assumption of a uniform attachment cannot be maintained. In other words, the probability of community members to be connected with a large number of communication partners is neither equally distributed nor homogeneous.

3.2.2 "Scale-free Networks"

An approach which bases on non-uniform attachment of edges to the nodes, e.g. a scheme that assumes heterogeneity, originates from Albert and Barabási [2002], who developed the "Scale-free Networks". This typology starts with n_0 nodes. In each step t, a node with m edges is added until $n_0 + t = N$. The probability $\Pi(k_i)$ that the new node is connected with an already existing node i, depends on its degree k_i according to Equation 3. This scheme is called preferential attachment.

Equation 3: Connection probability with preferential attachment

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} = \frac{k_i}{2mt}$$

As a consequence, community members who already maintain many communication links possess a high probability to attain further ones. This converges for $t >> n_0$ (i.e. for large N) asymptotically to probability functions which adhere to power laws (Equation 4) and, thus, to distributions of member activities which are highly heterogeneous.

Equation 4: Probability function degree k for scale-free networks

$$P(k) \propto k^{-\lambda}$$

Albert and Barabási [2002] as well as Ravid and Rafaeli [2004] have been able to observe this pattern within different kind of networks, like the World Wide Web (websites connected through links), the topology of the Internet (physical connections between computers and other network devices), a network of movie actors (connected by common movie appearances), networks of scientists (connected by common publications), ecological networks (connection between hunters and bounty), a network of dating partners (connected by dates), and online discussion groups. The attribute "scale-free" has been chosen, because due to the form of the distribution function – in contrast to the distribution of Poisson or Gauss – the mean value (or scale) $<k>$ does not appear to be meaningful for characterizing the network.

In the double logarithmic representation of Illustration 5, it becomes evident that the typical straight lines of "Scale-free Networks" only evolve asymptotically for large *k*. For small *k*, however, the plot of the empirical probability distributions is obviously flatter. This observation is according to the findings of Holme et al. [2004] in a dating community. Thus, the assumption of mere preferential attachment cannot be maintained here, as well.

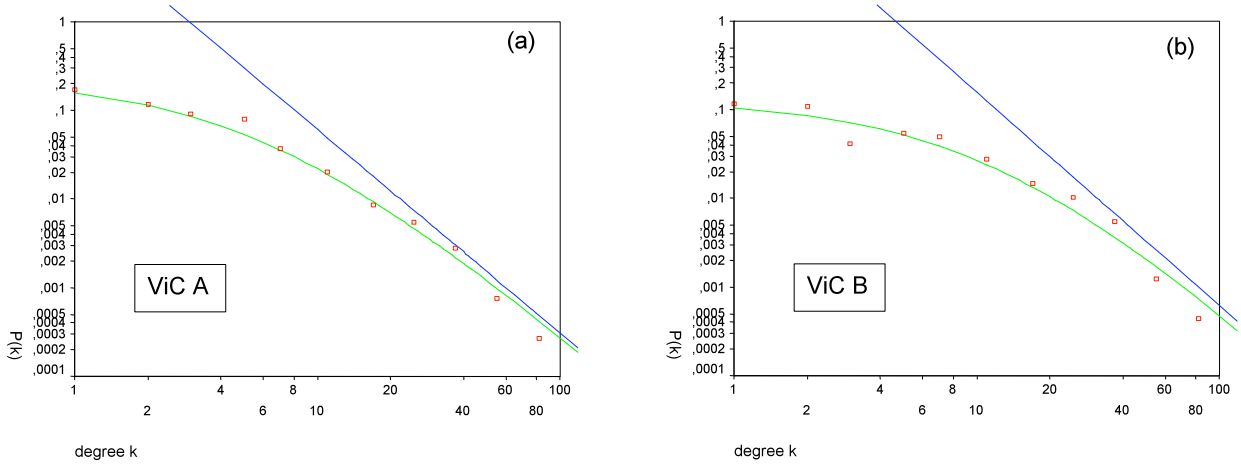

Illustration 5: Histograms of Illustration 4 in double logarithmic representation. The solid drawn curves (created with DataFit 6.0) represent distribution functions according to Pennock et al. (Equation 6) with $<k_A>=9.4$, $\alpha_A=0.77$ and $<k_B>=13.3$, $\alpha_B=0.71$. The straight lines correspond to distributions according to Albert and Barabási [2002] (Equation 4) with $\lambda=1+1/\alpha$.

3.2.3 Mixed Model of Pennock et al.

Consequently, the third model of Pennock et al. [2002] is a combination of the two approaches discussed earlier. Here the network emerges from a mixture of preferential and uniform attachment. The probability that a new node connects with an existing node *i* is:

Equation 5: Probability of connection with mixed preferential and uniform attachment

$$\Pi(k_i) = \alpha \frac{k_i}{2mt} + (1-\alpha)\frac{1}{n_0 + t}$$

The first term in Equation 5 can be interpreted, that new participants prefer to attach themselves to popular participants with many communication partners (preferential binding). The second term is independent of the popularity of participants and corresponds to individual reasons of the new participants for choosing communication partners, which – from a macroscopic view – can be regarded as randomly (uniform binding). So, the first term represents connection probability of the scale-free networks (Equation 3), while the second term represents the constant probability of a random network.

The mixing factor α controls the ratio of the two kinds of binding forms. The factor α can be regarded as a measure for the heterogeneity in the network. For instance, if $\alpha = 0$, exclusively uniform attachment takes place (maximum homogeneity). If $\alpha = 1$, only preferential attachment takes place (maximum heterogeneity). For $t >> n_0$ (i.e. large N), Equation 5 leads to the probability function of Equation 6, where $<k>$ is the average degree, again.

Equation 6: Probability function of the degree k according to Pennock et al.

$$P(k) = \frac{[2\langle k\rangle(1-\alpha)]^{\frac{1}{\alpha}}}{[\alpha k + 2\langle k\rangle(1-\alpha)]^{1+\frac{1}{\alpha}}}$$

In this paper relationships between members of virtual communities are analyzed and not relationships between web pages as in Pennock at al. [2002]. Though, the application of their model leads to compelling results. The curve adjustments shown in Illustration 6 provide first evidence that the model fits nicely with the data of our two communities. Moreover, Table 1 indicates the statistical significance of this model by the significance of its single free parameter the mixing factor α.

Parameter	Fit of α
ViC A; Week 129: <k> = 9,4	*$\alpha = 0{,}77 \pm 0{,}03$; ($T = 25$; $R^2_{adj} = 0{,}986$)*
ViC B; Week 145: <k> = 13,3	*$\alpha = 0{,}71 \pm 0{,}05$; ($T = 13$; $R^2_{adj} = 0{,}962$)*

Table 1: Fit of the heterogeneity measure α from Equation 6 to the degree distribution from Illustration 6 using the empirical average degree <k>. T is the value of the t-Test on significance of α (➔ level of significance < 0.1%).

The model of Pennock et al. is able to describe the distribution of the degree k with only two parameters, the mixing factor α and the average degree $<k>$. The average degree can be calculated directly from the network and so, only the mixing factor has to be fitted. Therefore, the model is very parsimonious and suited for longitudinal observations. Furthermore, it provides an indicator for the level of heterogeneity by means of the mixing factor α.

4 Attributive Activity

The attributive communication activity focuses on individual characteristics of community members. The relationships between participants are not considered in this context. In order to analyze the attributive activity of the two communities, the average number of messages $<s>$ will be examined in a manner reminiscent to the use of the average degree $<k>$. Subsequently, a reinterpreted Pennock et al. model will be applied.

4.1 Average Number of Messages

The average number of messages <*s*> can be calculated from the sum of the individuals' messages or the division of total sum of messages *S* by the number of active users *N* according to Equation 7.

Equation 7: Average number of messages

$$\langle s \rangle = \frac{1}{N} \sum_{i=1}^{N} s_i = \frac{S}{N}$$

Illustration 6 shows the time series for the two communities. In both communities, values of up to approximately 15 average messages per participant are reached. However, ViC B indicates a steep rise of the average message number until approximately week 38 which is then followed by a decline. In contrast, ViC A shows an almost linear increase of <*s*>. However, the range of the average number of messages does not differ between the two communities as clearly as they differed in their average degree in Illustration 3.

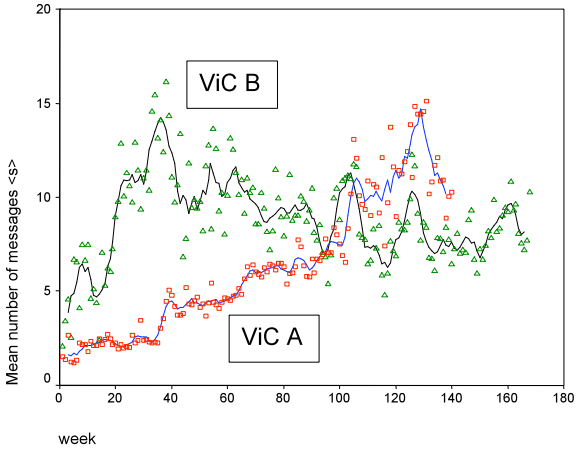

Illustration 6: Plot of the average number of messages <s> for the ViCs A and B.

4.2 Message Distribution

As indicated by Illustrations 7a and b, the empirical distribution function of the number of messages is extremely distorted, again. A minority of participants writes the majority of the messages, whereas the majority of participants show little attributive activity. This disparity has been frequently reported as a typical and remarkable behavior in virtual communities [Jones and Rafaeli, 1999, Light and Rogers, 1999, Nonnecke and Preece, 2000a, Stegbauer, 2001, Whittaker, et al., 1998]. However a suitable formal quantification for this phenomenon is not yet available.

Due to the similarity of Illustration 7 with the empirical distribution function of the average degree in the Illustration 5, it is straightforward to apply the model of Pennock et al., again. In Equation 5 and Equation 6, the number of edges k_i of a node i and the average number of edges $<k>$ are substituted by the number of messages s_i of a user i and the average number of messages $<s>$. Table 2 and the curve adjustment lines in Illustration 7 show, that the adjustment of the heterogeneity measure α on the basis of the empirical $<s>$ leads to statistically and visually satisfying results, again.

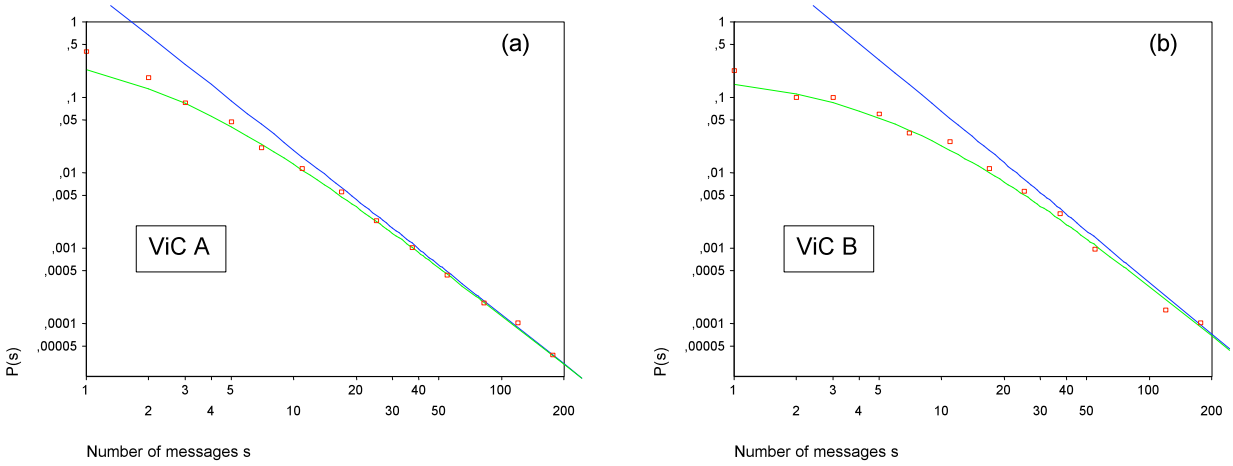

Illustration 7: Histograms of the number of messages in ViC A (84th week) and ViC B (82nd week). P(k)=N(k)/N with N_A=1,865 and N_B=139 (double logarithmic representation). The adjusted curves (created with DataFit 6.0) represent distribution functions according to Pennock et al. (Equation 6) with $<s_A>$=6.3, α_A=0.85 and $<s_B>$=11.2, α_B=0.79. The straight lines correspond to distributions according to Albert and Barabási [2002] (Equation 4) with $\lambda=1+1/\alpha$.

Parameter	Fit of α
ViC A; Week 84: <k> = 6.3	*$\alpha = 0.85 \pm 0.01$; ($T = 59$; $R^2_{adj} = 0.995$)*
ViC B; Week 82: <k> = 11.2	*$\alpha = 0.79 \pm 0.02$; ($T = 35$; $R^2_{adj} = 0.992$)*

Table 2: Fit of the heterogeneity measure α from Equation 6 to the empirical distribution of the number of messages from Illustration 8. The parameter <s> is the empirical average message count of the users. T is the value of the t-Test on significance of α (➔ level of significance < 0.1%).

The applicability of the model of Pennock et al. in this context may surprise at first sight, since the attributive communication activity does not allow constructing a network of relations, which

is the base of their model. Instead, the definition of attributive activity explicitly excludes the relation between the actors while looking at the individual's characteristics.

But, the model of Pennock et al. may be reinterpreted in a different way. Again, see Equation 5: At each point of time t, not a new node with m edges is added to the network, but now m new messages are composed, which will be attached to users in either preferentially or uniformly way. The first term in Equation 5 now represents the probability that a user i, who already "possesses" s_i messages, writes another one (preferential binding). The second term is the basic probability that user i will write a new message independently of the number of messages he "possesses" (uniform binding). Again, the mixing factor α is the ratio of the two forms of binding and therefore is an indicator of heterogeneity.

5 Heterogeneity Evolution

If in Equation 6 the mixing factor α is varied, while the average degree $<k>$ is kept constant, the following may be recognized: With declining values of the mixing factor α, the probability function gets more flattened for small k and for large k the asymptotically decline of the probability function gets steeper (see the slope of lines in Illustration 5 and Illustration 7). Thereby, a smaller value of α means a more homogeneous communication activity of the users vice versa a less heterogeneous communication activity.

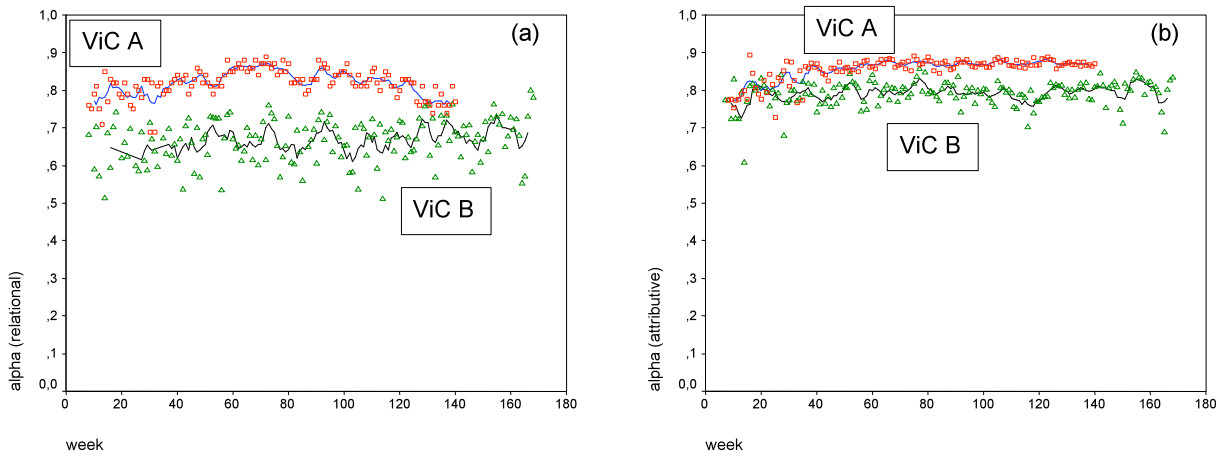

Illustration 8: Evolution of the heterogeneity measure α_k for the relational acitivity (a) and α_s for the attributive activity (B) of ViC A and B.

In order to examine the evolution of the heterogeneity measures of the two ViCs, we calculated for every week of observation the empirical averages and adjusted for each weekly dataset the mixing parameters to Equation 6. The results are shown in Illustration 8.

First, it can be stated that the heterogeneity measure for both the relational activity (α_k) and the attributive activity (α_s) is associated substantially closer to $\alpha = 1$ than to $\alpha = 0$. Thus, the heterogeneity in both communities is remarkable (see also Table 3).

According to Equation 5, this indicates from the relational perspective that participants mostly communicate with such actors who already have many communication partners (preferential binding). Pennock et al. term this phenomenon the "rich get richer". In agreement with Albert and Barabási, they observe a purely preferential binding of "Scale-free Networks" within the whole World Wide Web. But, similar to our investigation, Holme et al. [2004] find, that the degree distribution in a dating community cannot be explained by preferential attachment solely. Illustration 8a shows that the mixing parameter α, albeit near 1, is obviously smaller than 1. Therefore, uniform ("random") attachment plays a small, but nevertheless clearly measurable role, which manifests itself in the flattening of the distribution for small k in relation to the asymptotical straight line in Illustration 5.

From the perspective of the attributive communication activity (Illustration 8b), it can be argued that participants possess a higher probability of writing new messages, the more postings they have already made. The probability that previously less active users will be writing is smaller due to the low influence of uniform binding. Though, the influence of uniform attachment is also seen here.

ViC A	ViC B	*Mean Difference*
$<\alpha_k>= 0.819\pm0.003$	$<\alpha_k>= 0.667\pm0.005$	$<\Delta>=0.158\pm0.006$
$<\alpha_s>= 0.856\pm0.003$	$<\alpha_s>= 0.792\pm0.003$	$<\Delta>=0.066\pm0.004$

Table 3: Temporal mean values of the heterogeneity measure α_k for the relational activity and α_s for the attributive communication activity and their temporal mean differences.

As Table 3 and Illustration 8 indicate, the relational and the attributive communication activity were more heterogeneous in ViC A compared to ViC B, since its values of α were closer to 1. This means that the uniform or homogeneous portion of ViC B was larger than in ViC A due to the smaller preferential or heterogeneous portion. In addition one recognizes – contrary to ViC A – within ViC B a larger discrepancy between $<\alpha_k>$ and $<\alpha_s>$. The heterogeneity of the relational activity is obviously smaller than the attributive one. Thus, the relationship of participants among each other is more homogeneous in ViC B than their attributive communication activity.

The temporal stability of the heterogeneity measure α_k and α_s and their independence from fluctuations of the average degree (Illustration 3), the average message count (Illustration 6) and the size of the communities (Illustration 1), point out that the heterogeneity measure α may represent a characteristic constant for specific ViCs. In any case, it is a useful measure for analyzing the communication behavior in virtual communities. However, we are fully aware that the validation of this finding requires additional testing efforts since our two communities served primarily as a basis for exploration.

6 Summary and Overview

The analysis of virtual communities as social networks according to Wellman et al. [Garton, et al., 1997, Wellman, et al., 1996, Wellman, 1997, Rafaeli, 2004] proves to be extremely useful for a rich quantitative description of the communication activity of the members of ViCs. The literature frequently refers to a strong heterogeneous distribution of the number of messages of the individual members [Jones and Rafaeli, 1999, Light and Rogers, 1999, Nonnecke and Preece, 2000a, Schoberth, 2002, Schoberth, et al., 2006, Stegbauer, 2001, Whittaker, et al., 1998]. This heterogeneity was found to be significant not only for the attributive, but also for the relational communication activity. This investigation provides further empirical support to the notion communication in virtual communities is not evenly distributed. Instead of homogeneity, a dominance of a few communicating participants who face many less active participants can be assumed [Nonnecke and Preece, 2000a, Nonnecke and Preece, 2000a, Stegbauer, 2001].

Three graph-theoretical models have been examined for the description of these skewed distributions. Purely uniform attachment ("Random Graph Theory") and purely preferential attachment ("Scale-free Networks") [Albert and Bararbasi, 1999, Albert, et al., 1999, Albert and Barabási, 2002] have not been able to deliver satisfying results. The model of Pennock et al. [2002], combining both types of attachment, was found to be able to describe the empirical distribution functions. By reinterpreting, the model could also be applied for modeling the distribution of the number of messages sent/posted by the users. The model of Pennock et al. is very parsimonious as it needs just two parameters. The first, the average degree or the average number of messages, can be calculated directly from the community. The second, the mixing factor α, has to be obtained by nonlinear curve fitting.

Pennock et al.'s model offers a basis for the explanation of the strong imbalance of the communication activity in ViCs. The mixing factor α represents a suitable measure for the quantification of this heterogeneity. At $\alpha = 0$, merely uniform attachment takes place (maximum homogeneity),

whereas at $\alpha = 1$, merely preferential attachment occurs (maximum heterogeneity). Both ViCs show mixing factors close to 1 (see Table 3 and Illustration 8), which corresponds to a stronger tendency to preferential than to uniform attachment. This indicates that the probability that a participant will write further messages or develop new relationships grows with the number of messages or connections he already has. This "rich get richer" phenomenon [Pennock, et al., 2002] leads in the long run to the emergence of the strong heterogeneity of the activity of members of the communities observed here.

The heterogeneity measures α_k belonging to the relational activity and α_s belonging to the attributive activity prove to be quite stable over time for the two exemplary ViCs. This fact denotes inherently stable characteristic of the respective communities. Since such stability is of interest, but cannot be sufficiently validated on the basis of two examples, there is a challenge to explore additional communities in future research. Such data may provide insight into the range of heterogeneity in virtual communities. Also interesting for future investigations would be the influence of different community interests and different community platforms might show on the distribution of activity.

Besides these commonalities, interesting differences between the two ViCs could be found. During the whole observation period, the smaller ViC B revealed less heterogenic ($<\alpha_k>= 0.667 \pm 0.005$ and $<\alpha_s>= 0.792 \pm 0.003$) than ViC A ($<\alpha_k>= 0.819 \pm 0.003$ and $<\alpha_s>= 0.856 \pm 0.003$). This difference is particularly prominent in the heterogeneity α_k of the relational communication activity. An explanation for this finding could be the high relational activity of ViC B, reaching almost twice the levels of ViC A. ViC B represents a tighter network of relations between participants which seems to lead to a more equal distribution of the activity. However, this contingency cannot be finally verified on the basis of the available data.

References

ALBERT, R.; BARABÁSI, A.-L. (1999). Emergence of Scaling in Random Networks. *Science, 286*, pp. 509-512.

ALBERT, R.; BARABÁSI, A.-L. (2002). Statistical Mechanics of Complex Networks. *Reviews of Modern Physics, Vol.7, 1*, pp. 47-97.

ALBERT, R.; JEONG, H.; BARABÁSI, A.-L. (1999). Diameter of the World-Wide Web. *Nature, 401*, pp. 130-131.

AVIV, R.; ERLICH, Z.; RAVID, G. (2003). Cohesion and Roles: Network Analysis of CSCL Communities. *Proceedings of the 3rd IEEE International Conference on Advanced Learning Technologies*, Athens, Greece.

BROWN, S.L.; TILTON, A.; WOODSIDE, D.M. (2002). The case for online communities. *The McKinsey Quarterly, 1,* Retrieved October 20, 2005, from http://www.mckinseyquarterly.com/article_page.aspx?ar=1143&tk=273538:1143:24&L2=24&L3=45&pagenum=1.

BRONSTEIN, I.N.; SEMENDJAJEW, K.A. (1989). *Taschenbuch der Mathematik* (24th ed.). Frankfurt/Main: Harri Deutsch, Thun

BRUNOLD, J.; MERZ, H.; WAGNER, J. (2000). *www.cyber-communities.de: Virtual Communities: Strategie, Umsetzung, Erfolgsfaktoren.* Landsberg, Lech: Moderne Industrie.

FIGALLO, C. (1998). *Hosting Web Communities: Building Relationships, Increasing Customer Loyalty, and Maintaining A Competitive Edge.* New York: John Wiley & Sons, Inc.

GARTON, L.; HAYTHORNTHWAITE, C.; WELLMAN, B. (1997). Studying Online Social Networks. *Journal of Computer Mediated Communication,* Vol. 3, 1, Retrieved July 27, 2008, from http://jcmc.indiana.edu/vol3/issue1/garton.html.

HAGEL, J.; ARMSTRONG, A. (1997). *Net Gain: Expanding markets through virtual communities.* Boston, MA: Harvard Business School Press.

HOLME, P.; EDLING, C.; LIJEROS, F. (2004). Structure and time evolution of an Internet dating community. *Social Networks*, Vol. 26, pp. 155-174.

JANSEN, D. (1999). *Einführung in die Netzwerkanalyse - Grundlagen, Methoden, Anwendungen.* Opladen: Leske + Budrich.

JONES, Q.; RAFAELI, S. (1999). User Population and User Contributions to Virtual Publics: A Systems Model. *Proceedings of the International ACM SIGGGROUP Conference on Supporting Group Work*, Phoenix, Arizona: ACM Press.

JONES, Q.; RAVID, G.; RAFAELI, S. (2004). Information Overload and the Message Dynamics of Online Interaction Spaces: A theoretical model and empirical explanation. *Information Systems Research.* Vol. 15, 2, pp. 194-210.

LICKLIDER, R.W.; TAYLOR, R.W. (1968). The Computer as a communication device. *Science and Technology: For the Technical Man in Management*, pp. 21-31.

LIGHT, A.; ROGERS, Y. (1999). Conversation as Publishing: The Role of News Forums on the Web. *Proceedings of the 32nd Hawaii International Conference on System Sciences*, Big Island, Hawaii: IEEE Computer Society.

NONNECKE, B.; PREECE, J. (2000a). Lurker demographics: Counting the silent. *Proceedings of CHI*. The Hague, The Netherlands.

NONNECKE, B.; PREECE, J. (2000b). Persistence and Lurkers in Discussion Lists: A Pilot Study. *Proceedings of the 33rd Hawaii International Conference on System Sciences*. Big Island, Hawaii: IEEE Computer Society.

PARSIMONY (2004). Top 1000 Parsimony-Foren, Retrieved July 22, 2004, from http://parsimony.net/top/top1000.htm.

PENNOCK, D.M.; FLAKE, G.W.; LAWRENCE, S.; GLOVER, E.J.; GILES, C.L. (2002). Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences,* Vol. 99, 8, pp. 5207-5211.

PREECE, J. (2000). *Online Communities: Designing Usability, Supporting Sociability*. New York: John Wiley & Sons.

RAFAELI, S.; RAVID, G.; SOROKA, V. (2004). De-lurking in virtual communities: a social communication network approach to measuring the effects of social capital. *Proceedings of 37th Hawaii International Conference on System Science*, Hawaii, Big Island: IEEE Computer Society.

RAVID, G.; RAFAELI, S. (2004). Asynchronous discussion groups as Small World and Scale Free Networks. *First Monday*, Vol. 9, 9. Retrieved July 27, 2008, from http://firstmonday.org/issues/issue9_9/ravid/index.html.

RHEINGOLD, H. (1993). *The virtual community: homesteading on the electronic frontier.* Reading, Massachusetts: MIT Press

SCHOBERTH, T. (2002). DiViCom - Eine Längsschnittstudie der Kommunikationsaktivität in Virtual Communities*. Kopplung von Anwendungssystemen – FORWIN-Tagung 2002* (Bartmann, D.). Aachen: Shaker Verlag, pp. 125-147.

SCHOBERTH, T.; HEINZL, A.; PREECE, J. (2006). Exploring Communication Activities in Online Communities: A Longitudinal Analysis in the Financial Services Industry. *Journal of Organizational Computing and Electronic Commerce*, Vol. 16, 3&4, Philadelphia: Taylor & Francis, pp. 245-263.

SCHOBERTH, T.; PREECE, J.; HEINZL, A. (2003). Online Communities: A Longitudinal Analysis of Communication Activities. *Proceedings of the 36th Hawaii International Conference on System Science*, Hawaii, Big Island: IEEE Computer Society.

SCHOBERTH, T.; SCHROTT, G. (2001). Virtual Communities – WI-Schlagwort. *Wirtschaftsinformatik*, Vol. 43, 5, pp. 517-519.

STEGBAUER, C. (2001). *Grenzen Virtueller Gemeinschaft – Strukturen Internetbasierter Kommunikationsforen*. Wiesbaden: Westdeutscher Verlag.

WASSERMAN, S.; FAUST, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

WELLMAN, B.; SALAFF, J.; DIMITROVA, D.; GARTON, L.; GULIA M.; HAYTHORNTHWAITE, C. (1996). Computer Networks as Social Networks: Collaborative Work, Telework and Virtual Community. *Annual Review of Sociology,* Vol. 22, pp. 213-238.

WELLMAN, B. (1997). An Electronic Group is Virtually a Social Network. *Culture of the Internet (Kiesler*). Mahwa, New Jersey: Lawrence Erlbaum Associates, pp. 179-205.

WHITTAKER, S.; TERVEEN, L.; HILL, H; CHERNY, L. (1998). The dynamics of mass interaction. *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work*, Seattle Washington: Lawrence Erlbaum Associates: pp. 276-283.

YOSHIOKA, T; HERMAN, G.; YATES, J.; ORLIKOWSKI, W. (2001). Genre taxonomy: A knowledge repository of communicative actions. *ACM Transactions on Information Systems,* Vol. 19, 4, pp. 431-456.

ZAKON, R.H.(2003). Hobbes' Internet Timeline v8.2. Uppsala University. Retrieved July 27, 2008, from http://www.zakon.org/robert/internet/timeline.