Relational and Attributive Activity in Virtual Communities

-Working Paper-

October 2004

Thomas Schoberth (University of Bayreuth),

Armin Heinzl (University of Mannheim),

Sheizaf Rafaeli (University of Haifa)

Relational and Attributive Activity in Virtual Communities

Thomas Schoberth (University of Bayreuth),

Armin Heinzl (University of Mannheim),

Sheizaf Rafaeli (University of Haifa)

Abstract.

Although Virtual Communities (ViCs) have been considered as an important e-commerce instrument, little is known about their evolution and the changes in the communication activity of its users. A frequent finding is that a small number of the participants is responsible for the majority of messages and, in contrast, a large portion of the users only write one or a few message(s). This discrepancy deserves more attention. In this paper, the heterogeneity of the communication activity is examined longitudinally on the basis of the graph-theoretical models "Random Network Theory" and "Scale-free Networks". The fusion of these two models allows operationalization of the heterogeneity of the relational as well as the attributive communication activity in ViCs. The adjustment of the empirical distribution functions of the examined ViCs to this model shows a predominance of preferential over uniform binding. This "rich get richer phenomenon" proves temporally stable and leads to the emergence of heterogeneity of the member's activities. We conclude that instead of stimulating all participants, it appears to be more promising to address the lead users as a main target. Their attachments to other users may be utilized in a positive and amplifying way in order increase a community's communication activity.

References: virtual communities, social network analysis, graph theory, heterogeneity of communication activity, "Degree"-distribution, preferential binding, uniform binding, "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 that visitors of websites, who appeared 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, number of spams, Usenet content, user logs etc. These artifacts are available for scrutiny and research [Jones and Rafaelt 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 exist scant theoretically and empirically supported explanations of this communication activity. Most extant empirical studies are static cross-sectional analyses. This paper therefore aims to empirically examine the communication activity of members of virtual communities and their determinants in the context of a comparative longitudinal profile study [Schoberth et al. 2003].

We begin by operationalizing the communication activity of two ViCs from a longitudinal perspective. It is often pointed out in the 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 the participants are responsible for the majority of messages in ViCs while most write only one or few messages. Although such a strong imbalance might be quite important, this effect has been given little formal and empirical expression in the literature. The second goal of this paper is to quantify and observe the heterogeneity in participation in communities over time. A model by Pennock et al. [2002] is utilized, which was initially developed for the relationship of websites to each other, however, as will be shown in the following, this model may explain the extremely heterogeneous activity of the participants in ViCs.

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 et al. [Albert and Barabási 1999; 2001; Albert et al. 1999], *the relational activity can be operationalized with the help of the average number of communication partners per active user and measured week*. The *network density* will be deployed as an alternative measure.

Individual characteristics of the users will be represented by attributive activity. These characteristics refer to the level of the individual; however, they are typically aggregated over all active users in a specific time interval. According to Whittaker et al. [1998], the *average number of messages per active user and week* will be utilized in the following.

1.1.3 Virtual Communities examined. The two virtual communities examined in this paper used web-based forums as their 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 hoped to stimulate stock trade volume and achieve a higher rate of customer retention by the use of this forum. The available data archives cover 1.03 million postings in a period of nearly three years (140 weeks).

The second Virtual Community ("ViC B") dealt with stocks and securities as well. However, this community was part of a website whose operator acted 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.

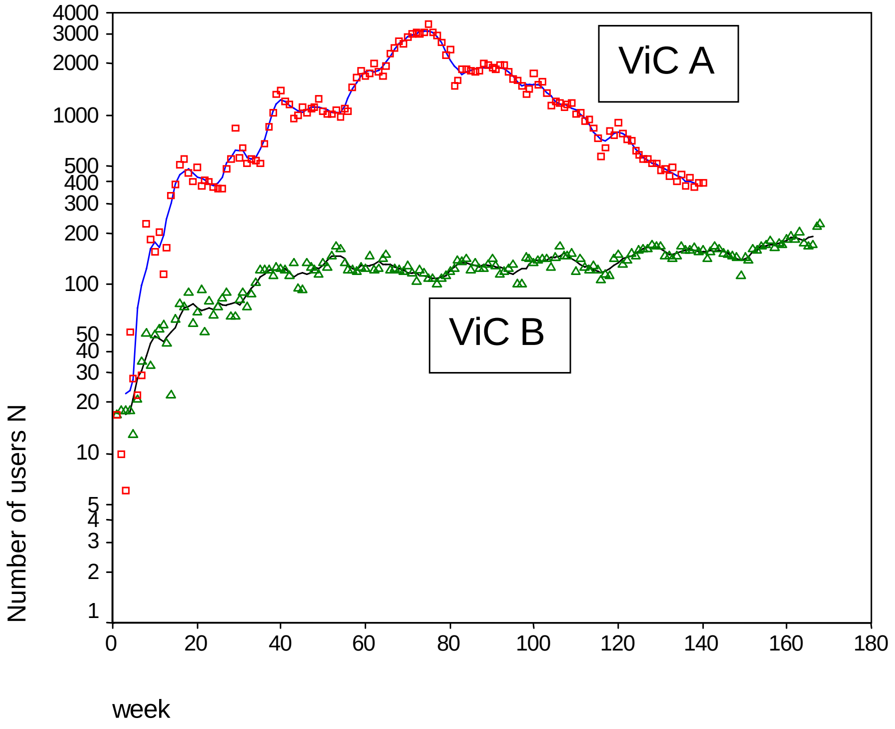

Illustration 1: Course of the number of active users N of the two ViCs A and B.

Illustration 1 shows the course of the number of active (writing) users N per monitored week. 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, 4, 7, and 9).

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 3405 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 are the relational and attributive communication activities of ViCs operationalizable?
2. How can the heterogeneity of the user’s activity be quantified?
3. How do the indicators of the two preceding research questions change over time?

By investigating these questions, we expect to gain (1) useful insights for supporting existing virtual communities as well as for stimulating their members. The development of quantitative measures allows (2) the comparison and evaluation of different communities. A suitable model for the quantitative description of the heterogeneity is expected to (3) provide first evidence for its causes and possible starting points for influencing them.

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) and attributive communication activity (section 4), three graph-theoretical models are introduced. The models's suitability for the description of heterogeneity in virtual communities is tested. We then explain the emergence of heterogeneity with the help of two examples. Then, in section 5, the development of the heterogeneity of the two ViCs is examined over time.

2 VIRTUAL COMMUNITIES AS SOCIAL NETWORKS

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

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 ongoing reciprocity of sending and receiving messages. However, relations can be asymmetrical. This occurs when communication is initiated mostly by one side. 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 in the fourth week after the community’s establishment. The connections between these nodes, e.g. the edges, are based on discussion threads in the community. These threads indicate the sequence and relation of messages in a tree-like graph.

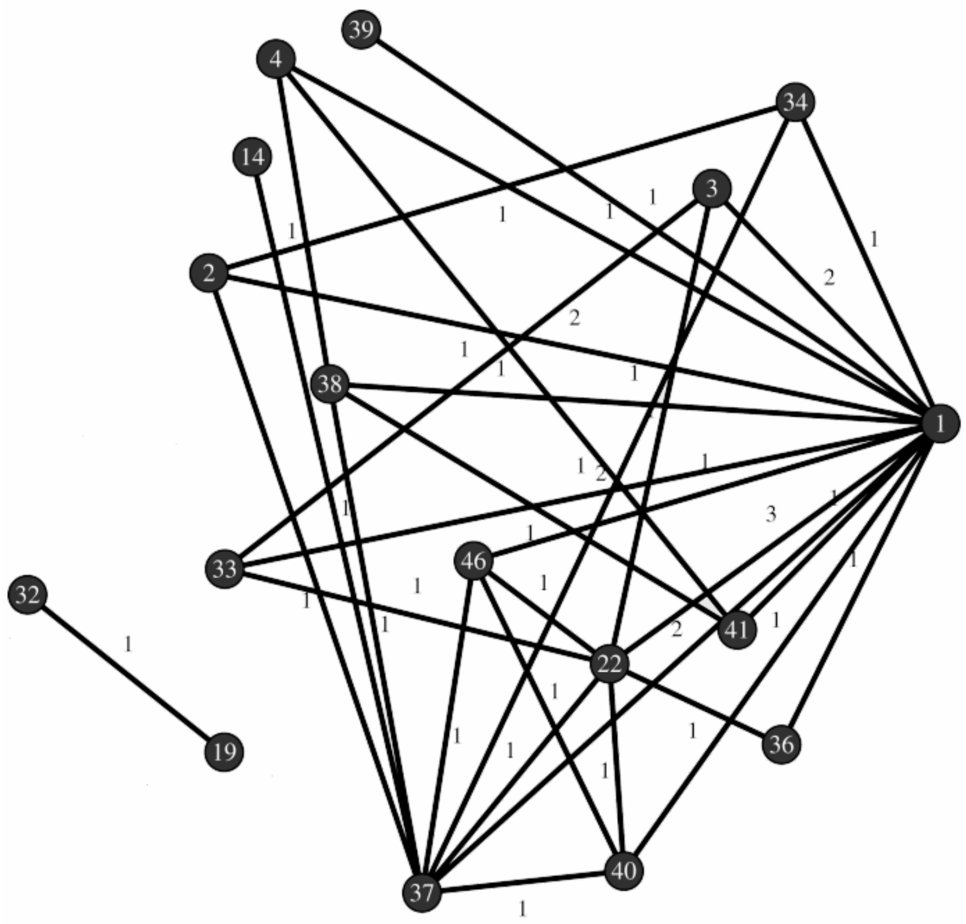

Illustration 2: Visualization of the social network of ViC A in the fourth week after its establishment (created with UCINET VI). The nodes represent active members, the digits on the edges represent the number of discussion threads in which connected members were communicating which each other.

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; Wasserman 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 (e.g. [Stegbauer 2001]), possess a limited suitability for the investigation of temporal changes since the complexity and multi-layered ness is hard to measure with the help of metric scales. Thus, a measure needs to be developed in this section which could capture the relational communication activity of large Virtual Communities over time as indicated with the help of case studies A and B.

We begin with an examination of the frequently used measure of network density (e.g. [Rafaeli et al. 2004]) for this purpose. Then, the measure of the average degree of a network (average number of edges per node) 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 which reflects the heterogeneity of the relational communication activity in virtual communities and which explains their causes.

3.1 Network Density

Within a network of N nodes there are maximally *N(N-1)/2* connections (edges) possible between these nodes. The network density *d* represents the relation between the observable edges *K* and the maximum possible number of edges (see equation 1).

Equation 1: Network density

$$d = \frac{2K}{N(N-1)}$$

The density of a network expresses the extent of interconnectedness. A value of *d = 0* for ViCs denotes that no connections exist between the members respectively that no relational interactivity takes place. On the other hand, *d = 1* indicates that all possible connections are realized and that each participant communicates with all others members of the forum. From the perspective of the individual, *d* denotes the probability that a member is connected with any other member in the community [Albert and Barabási 2001].

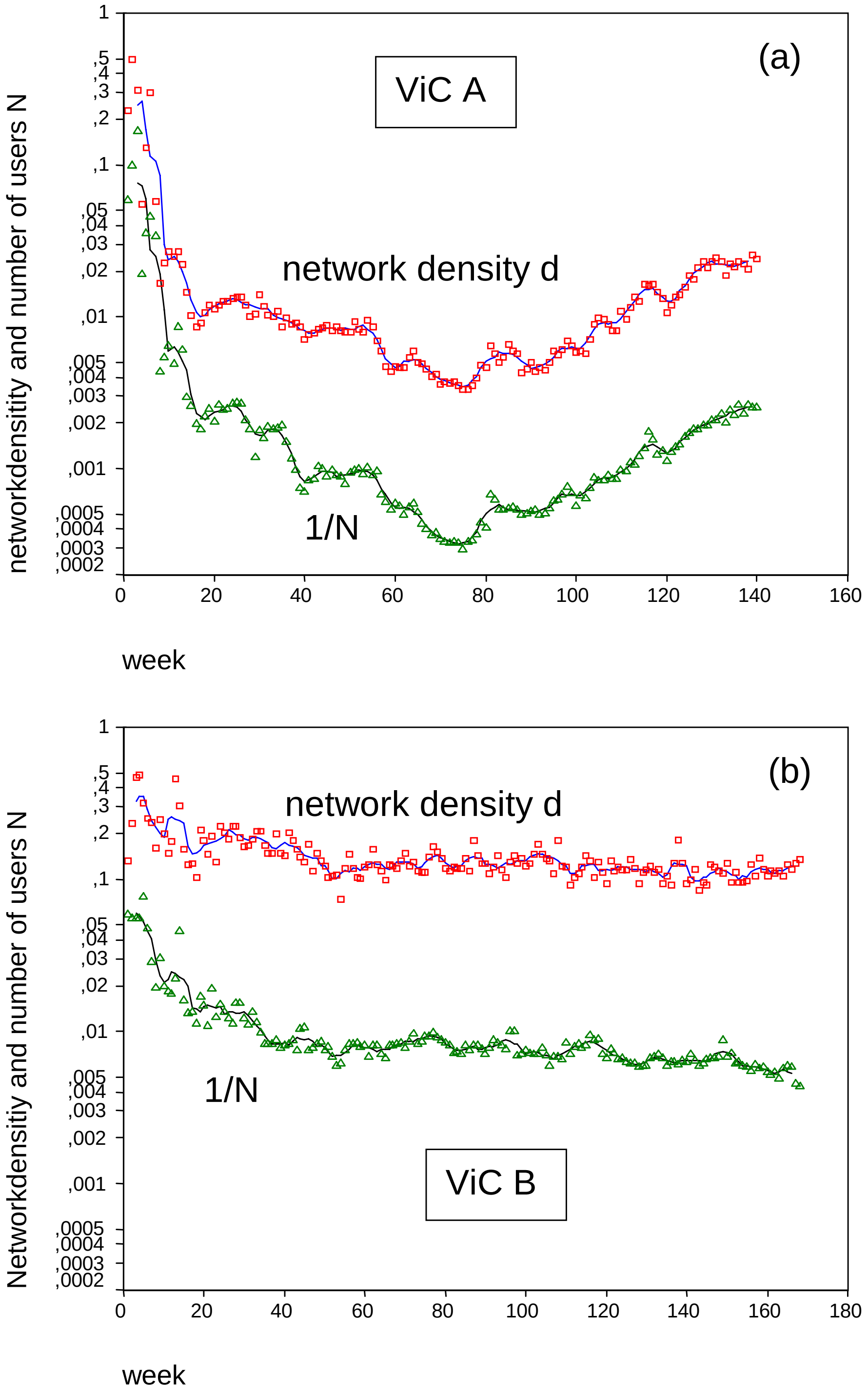

Illustration 3: Course of the network density d and reciprocal value of the number of active users 1/N of the two ViCs A and B.

Illustrations 3a and b show the evolution over time of the network density for the two ViCs charted on a logarithmic scale. Albert and Barabási [2001] state that the network density *d* behaves reciprocally proportional to the number of nodes *N*. In order to examine this phenomenon further, the reciprocal value of the number of members *1/N* was added into Illustration 3. A comparison of the plots *d* and *1/N* indicates a linear relationship $\ln d \sim \ln(1/N)$. Thus, we can conclude in accordance with Albert and Barabási that $d \sim 1/N$.

As a consequence, a larger community like ViC A possesses a lower network density *d* than a smaller one like ViC B. The measure of the network density is therefore only meaningful for structural analysis if the size of two networks is similar over time. Of course such similarity may be a rare occurance, happening in only a few instances.

3.2 Average Degree

An alternative measure for structural analysis is the average degree <*k*>. It represents the degree of connectedness and denotes the number of edges *k* that a node in the network possesses. In the context of an online community, *k* symbolizes the number of members that communicate directly with each other. The average degree can be calculated from the total number of non-directional connections *n* and from the size of the network (number of the active users *N*) as shown in equation 2.

Equation 2: average degree

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

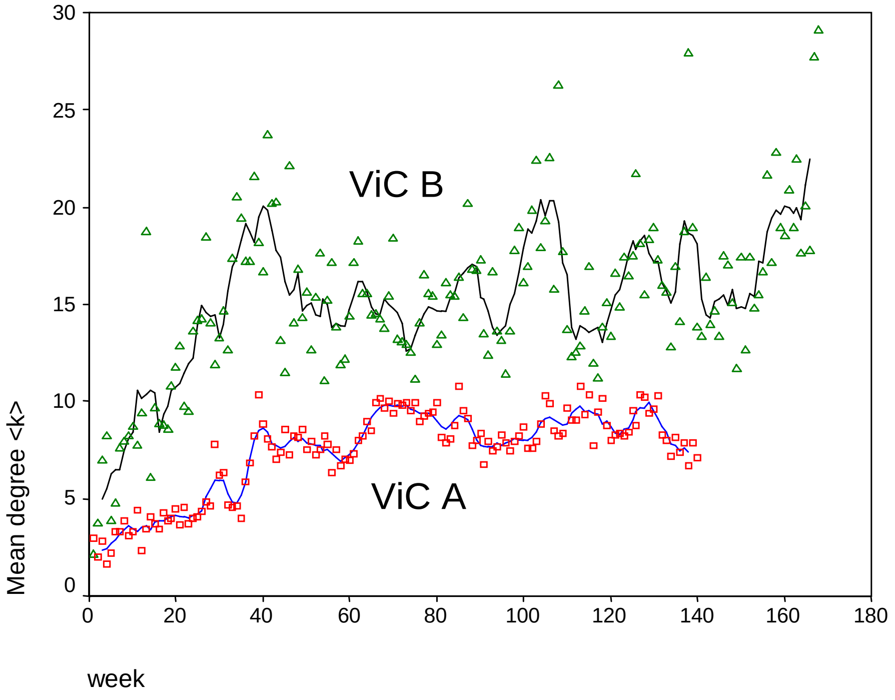

Illustration 4: Plot of the average degree <k> of the two ViCs A and B.

Equation 2 yields the relationship $<k> = (N-1)d \approx Nd$ between the average number of edges and the network density. This relationship and the comparison of the plots of $<k>$ in Illustration 4 and N in Illustration 1 it becomes evident that $<k>$ is obviously not depending on N in contrast to d. Thus, the average degree $<k>$ appears to be a suitable measure for the comparative analysis of the relational structure of virtual communities.

The comparison of the two time series indicates that the degree of interconnectedness and, thus, the relational communication activity of ViC B is about twice as large as in ViC A (based on the temporal means, $<k_B>/<k_A> = 2{,}13 \pm 0{,}07$). Based on the temporal means, the participants of ViC A have an average degree $<k_A> = 7{,}4 \pm 0{,}2$, ViC B has in contrast a degree $<k_B> = 15{,}2 \pm 0{,}3$. In the course of both communities, $<k>$ first rises and seems to stabilize starting from approximately week 40.

3.3 Degree Distribution

In Illustrations 5a and b, the degree distribution of the two communities is represented by the example of weeks 129 (ViC A) and 145 (ViC B). For each degree *k* found, the value of the empirical probability function [Bronstein and Semendjajew 1989, S.678 ff] $P(k) = N(k)/N$ is plotted. $N(k)$ is the number of participants with a number of edges *k*. *N* is the total number of the participants (N_A = 147 and N_B = 447) in these two weeks. 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 clearly that the distributions are extremely skewed and that they are not normally distributed. 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, S.679]. The terms distribution and probability function are used synonymously in this context.

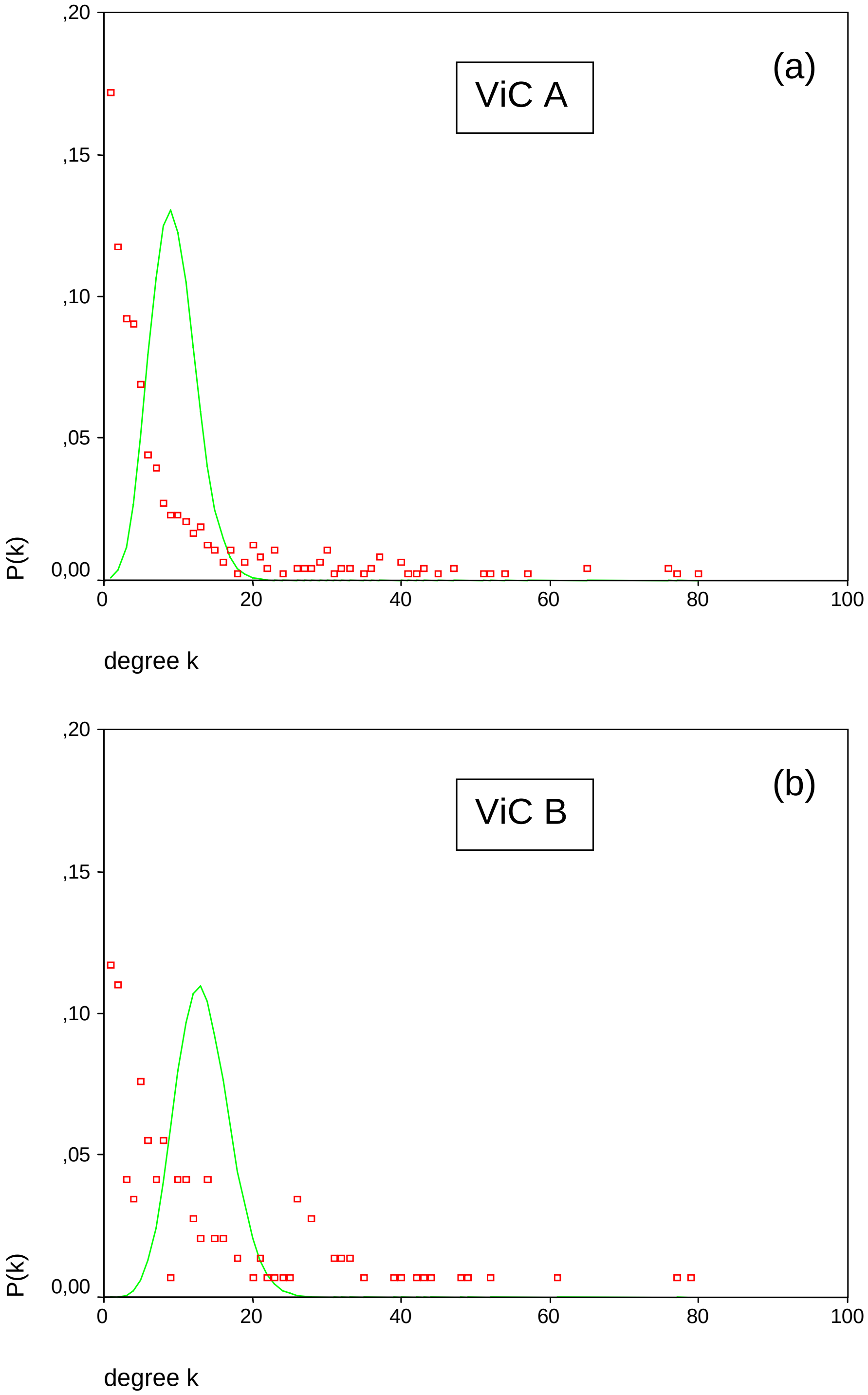

Illustration 5: Histograms the degree distributions in the 129th week of ViC A and the 145th week of ViC B. P(k) = N(k)/N with N_A = 147 and N_B = 447. The solid lines represent the Poisson distributions (equation 3) according to the empirical average values of the two distributions ($<k_A>$ = 9.4 and $<k_B>$ = 13,3).

3.3.1 "Random Graph Theory". The Random Graph Theory assumes a network which has developed 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 2001]. For large N, this would lead to the Poisson distribution as shown in Equation 3 with the mean of <k>. Such a distribution can be regarded as homogeneous, since its values are randomly (homogeneously) distributed around the mean value.

Equation 3: Poisson distribution

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

In Illustration 5, Poisson distributions are plotted as solid lines according to Equation 3 with the empirical average degrees of *<k_A> = 9,4* and *<k_B> = 13*. The clearly visible deviations of the average values to the empirical distributions 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.

3.3.2 "Scale-free Networks". An approach which considers a non-uniform attachment of the number of edges to the numbers of nodes, e.g. a scheme that assumes heterogeneity, comes from Albert and Barabási [2001] who developed the "Scale-free Networks". The origin of this typology begins with n0 nodes. In each step t, a node with m edges is added until n0 + t = N. The probability Π(ki) that the new node is connected with an already existing node i, depends on its degree ki according to equation 4. This scheme is called preferential attachment.

Equation 4: 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>>n₀* (i.e. for large *N*) assymptotically to probability functions which adhere to power laws (equation 5) and, thus, to distributions of member activities which are highly heterogeneous.

Equation 5: Probability function degree k for scale-free networks

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

Albert and Barabási [2001] as well as Ravid and Rafaeli [2004] have been able to observe this generalizable pattern with regard to 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 Poisson or the Gauss distribution – the mean value (or scale) does not appear to be meaningful for characterizing the network.

In the double logarithmic representation of Illustration 6, 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. Thus, the assumption of mere preferential attachment cannot be maintained just as the assumption of mere uniform attachment.

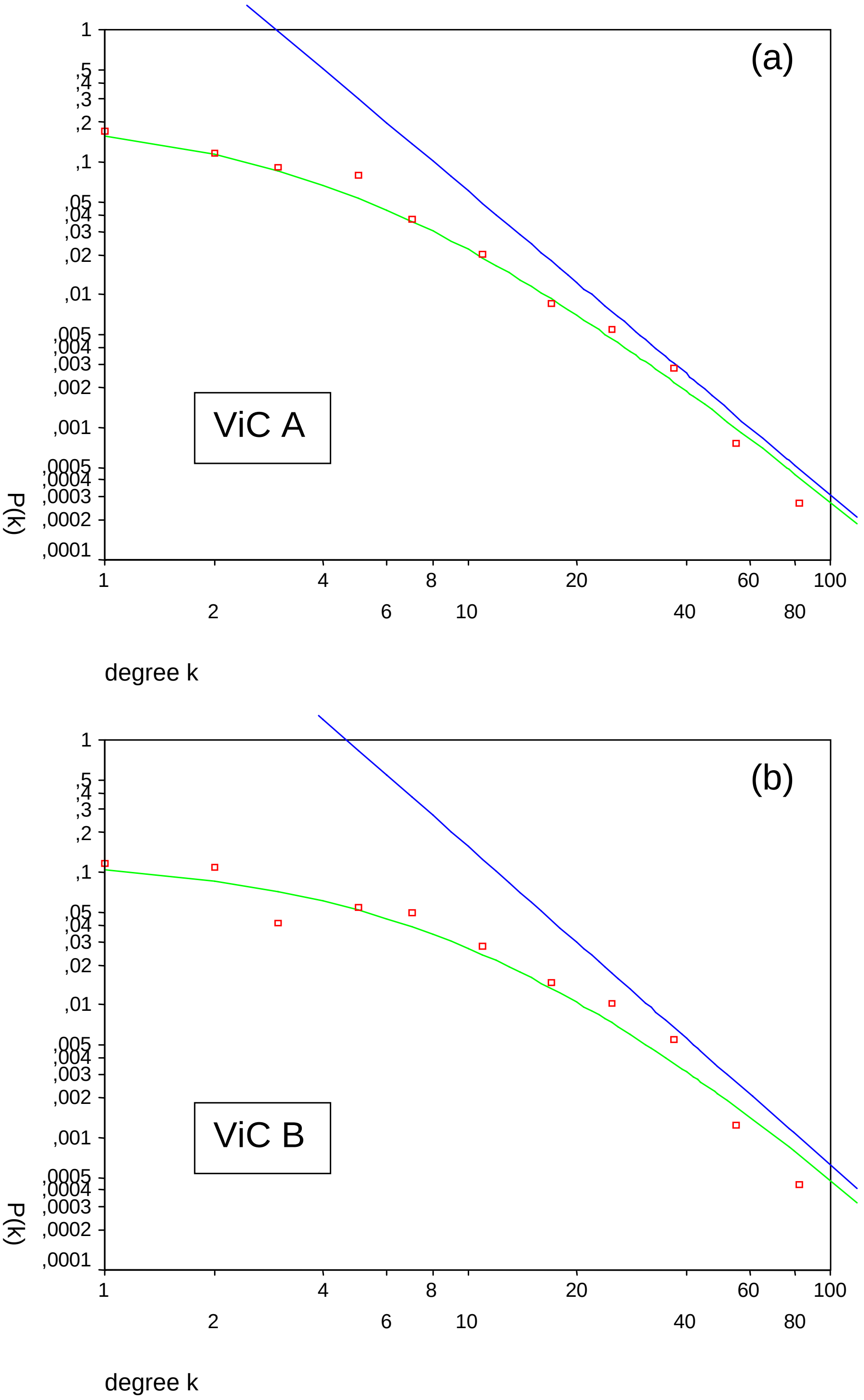

Illustration 6: Histograms of illustration 5 in double logarithmic representation. The solid drawn curves (created with DataFit 6.0) represent distribution functions according to Pennock et al. (equation 7)

with <k$_A$>=9,4, α$_A$=0,77 and <k$_B$>=13,3, α$_B$=0,71. The straight lines correspond to distributions according to Albert and Barabási [AlBa01] (equation 5) with λ=1+1/ α.

3.3.3 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 therefore is:

Equation 6: 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 6 can be interpreted thusly: 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). The mixing factor *α* models the combination of the two kinds of attachments and can be regarded as a measure for the heterogeneity. For instance, if *α* = 0, exclusively uniform attachments take place (maximum homogeneity). If *α* = 1, only preferential attachments take place (maximum heterogeneity). For $t>>n_0$ (i.e. large *N*), equation 6 leads to the probability function of equation 7 (<*k*> is again the average degree).

Equation 7: Probability function degree k according to Pennock et al.

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

Although relationships between web pages have not been analyzed as in Pennock at al., but rather relationships between members of virtual communities, the application of mixed model leads to compelling results (see Illustration 6). The curve adjustment lines in Illustration 6 provide first evidence that the model fits nicely with the data of our two communities using the empirically determined <*k*>. Moreover, Table 1 indicates the statistical significance of the mixed model with the help of the mixing factor *α*. The model of Pennock et al. [2001] appears to be extremely suitable to explain the origin of heterogeneity of relational communication activity in virtual communities as well as to parameterize the heterogeneity by means of the mixing factor *α*.

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

Table 1: Fit of the heterogeneity measure α from equation 7 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%).

4 ATTRIBUTIVE ACTIVITY

The attributive communication activity focuses on individual characteristics of community members. The relationships between participants are irrelevant in this context. In order to analyze the attributive activity of our communities, the average number of messages per participant will be examined in a manner reminiscent to the use of the average degree <*k*> over the course of time. Subsequently, a modified Pennock et al. model was interpreted, and the number of messages in each ViC was investigated for one example week.

4.1 Average Number of Messages

The average number of messages <*s*> can be computed from the total number of messages *S* and the number of active users *N* according to Equation 8.

Equation 8: Average number of messages

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

Illustration 7 shows different time series for the two communities. In both communities, values of up to approximately 15 messages per participant are reached, however, ViC B indicates a steep rise of the average message number in the beginning (until about week 38) which is then followed by a slight decline of the number of messages. In contrast, ViC A shows an almost linear increase of <*s*>.

However, the changes in the average number of messages do not correspond with the number of active users over time (see Illustration 1). Thus, we assume – as has been the case with the average degree –that <*s*> represents a suitable measure to describe the attributive activity in virtual communities.

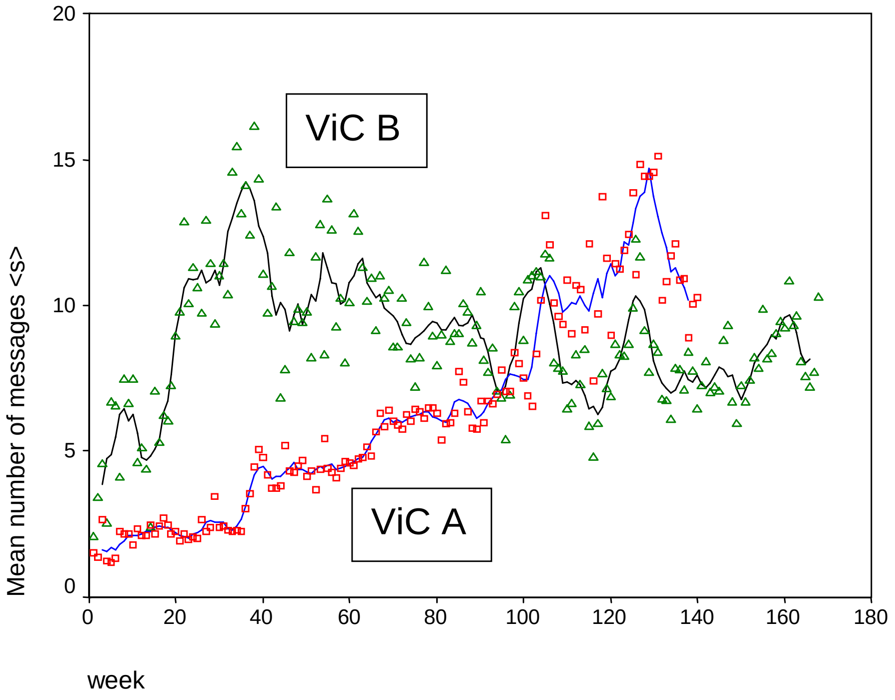

Illustration 7: Plots of the average number of messages <s> for ViC A and B.

4.2 Message Distribution

As indicated by Illustrations 8a and b, the empirical distribution function of the number of messages is extremely distorted. 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 important 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 8 with the empirical distribution function of the average degree in the Illustration 6, it is straightforward to apply the model of Pennock et al. [2001] again. In Equation 6 and Equation 7, the number of nodes k and the average number of nodes $<k>$ are replaced by the number of messages s and the average number of messages $<s>$.

The adjustment of the heterogeneity measure *α* on the basis of the empirical <*s*> leads to statistically and visually satisfying results, too (see Table 2 and curve adjustment lines in Illustration 8).

The applicability of the model by Pennock et al. [2001] may seem surprising since the communication activity observed allows no construction of a relational network due to the intentional exclusion of the relational level which is a substantial basis of this model. But the model of Pennock et al. [2001] (see Equation 6 again) may be interpreted in a different way. At each point of time *t*, a new node with *m* new edges is not added automatically, but rather *m* new messages are composed which will be attached to other participants either preferentially or uniformly. The first term in Equation 6 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 new messages independently of the past messages (uniform binding).

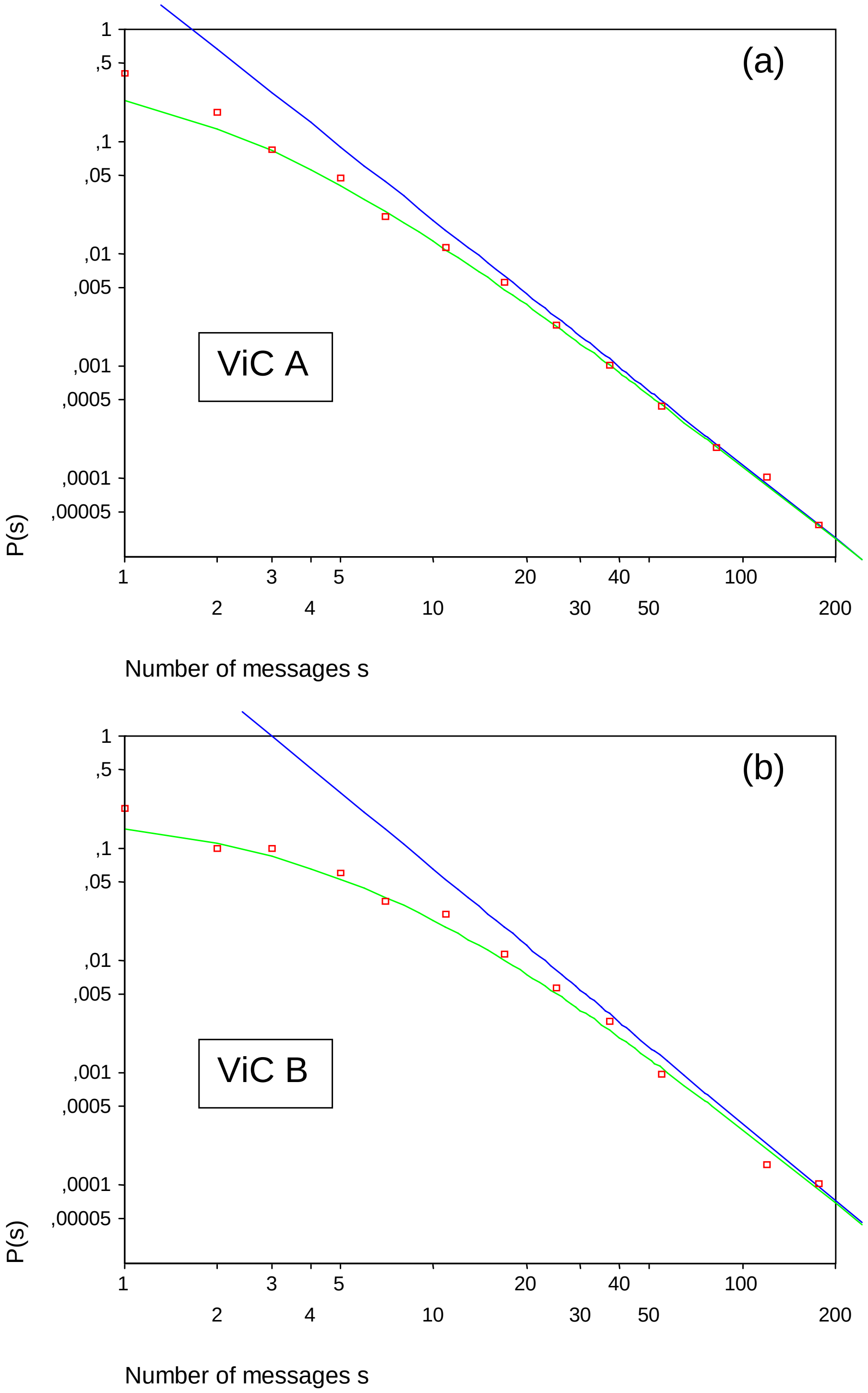

Illustration 8: Histograms of the number of messages in ViC A (84th week) and ViC B (82nd week). P(k)=N(k)/N with N_A=1865 and N_B=139 (double logarithmic representation). The adjusted curves (created with DataFit 6.0) represent distribution functions according to Pennock et al. (equation 7) with <s_A>=6,3, α_A=0,85 and <s_B>=11,2, α_B=0,79. The

straight lines correspond to distributions according to Albert and Barabási [AlBa01] (equation 5) with λ=1+1/ α.

Parameter	Fit of α
ViC A; Week 84: <k> = 6,3	*α = 0,85±0,01; (T = 59; R^2_{adj} = 0,995)*
ViC B; Week 82: <k> = 11,2	*α = 0,79±0,02; (T = 35; R^2_{adj} = 0,992)*

Table 2: Fit of the heterogeneity measure α from Equation 7 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%).

5 HETEROGENEITY EVOLUTION

If we take a closer look at Equation 7 with constant <*k*>, the following may be recognized: with declining values of the mixing factor *α*, the probability function for small *k* gets more flattened and the asymptotically decline of the probability function for large *k* gets steeper (see the slope of lines in Illustration 6 and Illustration 8). For the entire distribution, a smaller value for *α* thereby actually means a smaller heterogeneity of the participants communication activity. Thus, the use of the mixing factor *α* as measure of heterogeneity appears appropriate. For this reason, the evolution of the heterogeneity measures of the two ViCs determined from Equation 7 will be examined in this section with the help of Illustration 9.

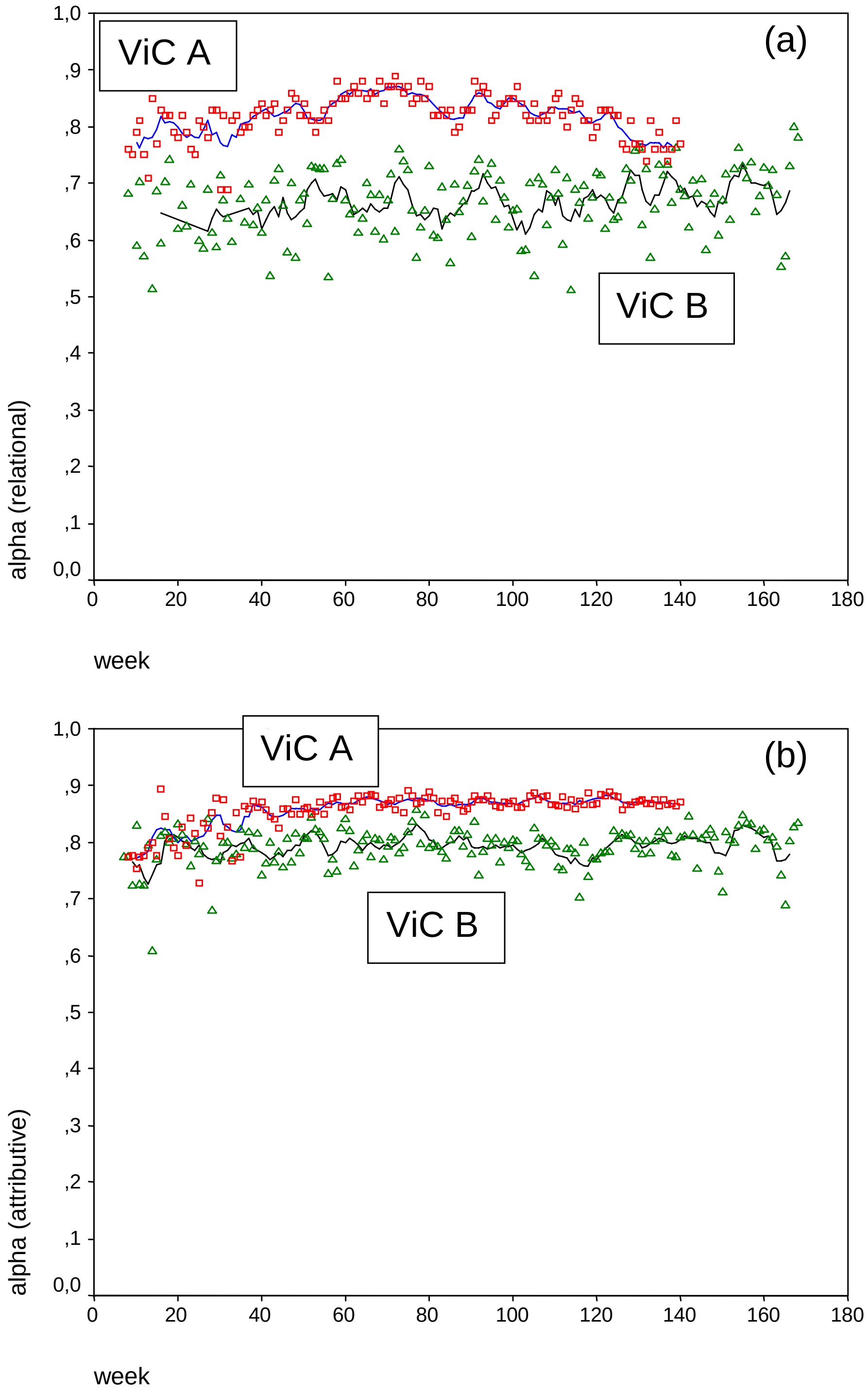

Illustration 9: Evolution of the heterogeneity measure α_k for the relational acitivity (a) and α_s for the attributive activity (B) of ViC A and B.

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 6 of the Pennock et al. model, this can be interpreted as indicating that participants communicate preferentially with such partners who already have many communication partners. 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. Similar to our investigation, both authors find, that subcategories like web pages of scientists, universities or companies distributions, cannot be explained by a preferential attachment alone and that the mixing parameter *α* is therefore obviously smaller than one. 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 6.

From an attributive communication activity perspective (Illustration 9b), it can be argued that participants possess a higher probability of writing new messages, the more postings they have already made. The probability that less active users will be addressed is smaller due to the low influence of uniform binding.

ViC A	ViC B	*Average 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 the plots in Illustration 8 indicate, the relational and the attributive communication activity were more heterogeneous in ViC A compared to ViC B, since the 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 size of the 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 relationships of participants among themselves is more homogeneous in ViC B than was their attributive communication activity.

The weak temporal changes of the heterogeneity measure α_k and α_s and their independence from fluctuations of the average degree (Illustration 4), the average message count (Illustration 8) and the size of the communities (Illustration 1), point out the fact that the heterogeneity measure α represents a characteristic and useful measure for analyzing the communication behavior in different 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. But nevertheless it also indicates that “rich” users are of importance for stimulating the communication acitivity in a ViC. Thus, key users should be the target group for commercial or non-commercial marketing activities in order to increase the communication activity.

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 et al. 2004] proves to be extremely useful for the description of the relational communication activity of its members. The data and analysis reported here demonstrates that the frequently used measure of the network density [Rafaeli et al. 2004] behaves indirectly proportional to the number of nodes. This strong dependence on the size of the network (i.e. the community) leads to the fact that the network density measure is only applicable with caution for the comparison of different communities. The closely related measure of average degree (average edge count of the nodes or average number of connections per participant) does not show this dependence and, therefore, appears to be well suitable as a relational measure. In other words, the average degree provides a good basis for comparing different communities since it is a relative metric which does not get distorted by the size of the community. In addition, changes with regard to the average number of messages of the users for the attributive communication activity have been examined on the basis of two ViCs.

The literature frequently refers to a strong heterogeneous distribution of the number of messages of the individual members of virtual communities [Jones and Rafaeli 1999; Light and Rogers 1999; Nonnecke and Preece 2000a; Schoberth 2002; Schoberth et al. 03; Stegbauer 2001; Whittaker et al. 1998]. This heterogeneity was, however, found to be significant not only for the attributive but also for the relational communication activity. 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 Barabási 1999; Albert et al. 1999; Albert and Barabási 2001] have not been able to deliver satisfying results. Only the model of Pennock et al. [2002], which permits the combination of both types of attachment with the help of a mixing factor α, is able to extrapolate the empirical distribution functions. With slight modifications, the model may also be applied for modeling the distribution of the number of messages sent/posted by the users, although this attributive measure does not permit the formation of a network graph.

In this paper, only active – writing – members have been observed. In order to also include the passive – reading – community users, Rafaeli et al. [2004] define a connection between two users if both jointly read a message. Since the number of passive users often exceeds the number of active users by a factor of ten [Brunold et al. 2000], this promising model should be considered for future investigations in order to capture all users of a community.

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 the heterogeneity. At $\alpha = 0$, only uniform attachments takes place (maximum homogeneity), whereas at $\alpha = 1$, only preferential attachments (maximum heterogeneity) occur. Both ViCs showed heterogeneity measures close to one (see Table 3 and Illustration 9), which corresponds to a stronger tendency to preferential than to uniform attachment. This indicates that the true probability that a participant will write further messages or develop relationshipsgrows as a function of the number of messages or connections she 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 (which represents the relational activity) and α_s (which represents the attributive activity) of the two ViCs prove to be quite stable over time. This fact denotes inherently stable characteristic of the respective communities. Since this stability is of interest, but cannot be sufficiently validated on the basis of two case studies, there is a challenge to explore additional communities in future research. Such data, if collected, may provide additional insight into the range of heterogeneity in virtual communities. Also interesting for future investigations would be the relationship of the communication heterogeneity of communities and the content, protocols and software technologies used to implement different online interaction spaces i.e. chat rooms, bulletin board and the like (see [Jones et al. 2004]).

Besides these commonalities, interesting differences between the two ViCs could be found. The smaller ViC B revealed a lower heterogeneity during the whole observation period ($<\alpha_k>= 0,667±0,005$ and $<\alpha_s>= 0,792±0,003$) than ViC A ($<\alpha_k>= 0,819±0,003$ and $<\alpha_s>=$

0,856±0,003). This difference is particularly evident in the heterogeneity α_k of the relational communication activity. An explanation for this phenomenon could be the high relational activity of ViC B, reaching almost twice the levels of ViC A. ViC B represents a tighter network of the 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. Operators of virtual communities may conclude that encouraging the construction of relations between members could aid reaching more homogeneous participation and, thus, promote stability and richness of the community in the long run.

This investigation provides further empirical support to the notion that there is no equal communication in virtual communities. This is in accord with postulations in the literature [Figallo 1998; Hagel and Armstrong 1997; Preece 2000; Rheingold 1993]. Instead of homogeneity, a dominance of a few communicating participants who face many less active participants can be assumed (see also Nonnecke and Preece [2000a; 2000b] and Stegbauer [2001]). This inequality affects both the ViCs themselves as well as their commercial exploitability. Heterogeneous distribution of the behavior of the users should be assumed by anyone who tries to support and influence the members of a ViC. For example, in the 129th week of the community A, a total number 447 members wrote messages. The strong heterogeneity may be exemplified by the fact that most active 5% of the users (24 participants) had interactions with 64% of all members (285). Instead of launching marketing activities for all participants, it appears to be more effective and efficient to address particularly the active users (e.g. the lead users) as a main target in order to leverage their attachments in a positive and amplifying way. Moreover, it is seems less appropriate to concentrate marketing efforts on a randomly chosen sample of community members. Due to the heterogeneous activity distribution, the randomly selected members may have few contacts with other members which leads to a high probability that this marketing effort might be ineffective.

References

Albert, R. and Barabási, A.-L. (1999). Emergence of scaling in random networks. *Science*. 286 (1999), pp. 509 – 512.

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*. 74 (2002), pp. 47 – 97.

Albert, R., Jeong, H., and Barabási, A.-L. (1999). Diameter of the World-Wide Web. *Nature*. 401 (1999), pp. 130 – 131.

Aviv, R., Erlich, Z., and Ravid, G. (2003). Cohesion and roles: Network analysis of CSCL communities. In *Proceedings of the IEEE ICALT-2003*. Athens, 2003.

Brown, S.L.; Tilton, A.; Woodside, D.M.: The case for online communities. The McKinsey Quarterly (2002) Nr. 1, http://www.healthyplace.com/Advertise/story_communities.asp, Requested at 2004-10-05.

Bronstein, I.N. and Semendjajew, K.A. (1989). *Taschenbuch der Mathematik. 24th Ed*. Verlag Harri Deutsch, Frankfurt/Main, 1989.

Brunold, J., Merz, H., and Wagner, J. (2000). *www.cyber-communities.de: Virtual Communities: Strategie, Umsetzung, Erfolgsfaktoren*. Verlag Moderne Indusrie, Landsberg, 2000.

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

Garton, L., Haythornthwaite, C., and Wellman, B. (1997). Studying online social networks. *Journal of Computer Mediated Communication*. 3 (1997).

Hagel, J. and Armstrong, A. (1997). *Net Gain: Expanding Markets Through Virtual Communities*. Harvard Business School Press, Boston, 1997.

Jansen, D. (1999). *Einführung in die Netzwerkanalyse - Grundlagen, Methoden, Anwendungen*. Leske & Budrich, Opladen, 1999.

Jones, Q. and Rafaeli, S. (2003). User population and user contributions to virtual publics: A systems model. In *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work*. Phoenix, 2003, pp. 239 – 248.

Jones, Q., Ravid, G., and Rafaeli, S. (2004). Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical explanation. *Information Systems Research*. 15 (2004), pp. 194 – 210.

Licklider, J. R. C. and Taylor, R.W. (1968). The computer as a communication device. *Science and Technology*. 76 (1968), pp. 21-31.

Light, A. and Rogers, Y. (1999). Conversation as publishing: The role of news forums on the Web. In *Proceedings of the 32nd Hawaii International Conference on System Sciences*. Hawaii, 1999.

Nonnecke, B. and Preece, J. (2000a). Lurker demographics: Counting the silent. In *Proceedings of CHI 2000*. The Hague, 2000, pp. 73 – 80.

Nonnecke, B. and Preece, J. (2000b). Persistence and lurkers in discussion lists: A pilot study. In *Proceedings of the 33rd Hawaii International Conference on System Sciences*. Hawaii, 2000.

Parsimony. (2004). Top 1000 Parsimony-Foren. http://parsimony.net/top/top1000.htm, requested 2004-07-22.

Pennock, D.M., Flake, G.W., Lawrence, S., Glover, E.J., and Giles, C.L. (2002). Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences*. 99 (2002), pp. 5207 - 5211.

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

Rafaeli, S., Ravid, G., and Soroka, V. (2004). De-lurking in virtual communities: A social communication network approach to measuring the effects of social capital. In *Proceedings of the 34th Hawaii International Conference on System Sciences*. Hawaii, 2004.

Ravid, G. and Rafaeli, S. (2004). Asynchronous discussion groups as Small World and Scale Free Networks. *First Monday*. 9 (2004).

Rheingold, H. (1993). *The Virtual Community: Homesteading on the Electronic Frontier*. MIT Press, Reading, 1993.

Schoberth, T. (2002). DiViCom - Eine Längsschnittstudie der Kommunikationsaktivität in Virtual Communities. In *Kopplung von Anwendungssystemen – FORWIN-Tagung 2002*, Bartmann, D., Ed. Shaker Verlag, Aachen, 2002.

Schoberth, T., Preece, J., and Heinzl, A. (2003). Online communities: A longitudinal analysis of communication activities. In *Proceedings of the 33rd Hawaii International Conference on System Sciences*. Hawaii, 2000.

Schoberth, T. and Schrott, G. (2001). Virtual Communities - WI-Schlagwort. *Wirtschaftsinformatik*. 5 (2001), pp. 517 – 519.

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

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, 1994.

Wellman, B. (1997). An electronic group is virtually a social network. In *Culture of the Internet*, Kiesler, S., Ed. Lawrence Erlbaum Associates, Mahwa, 1997.

Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia M., and Haythornthwaite, C. (1996). Computer networks as social networks: Collaborative work, telework and virtual community. *Annual Review of Sociology*. 22 (1996), pp. 213 – 238.

Whittaker, S., Terveen, L., Hill, H., and Cherny, L. (1998). The dynamics of mass interaction. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work*. Seattle, 1998, pp. 257 – 264.

Yoshioka, T, Herman, G., Yates, J., and Orlikowski, W. (2001). Genre taxonomy: A knowledge repository of communicative actions. *ACM Transactions on Information Systems*. 19 (2001), pp. 431 – 456.

Zakon, R. H. (2003). Hobbes' Internet Timeline v5.3. Uppsala University, http://www.zakon.org/robert/internet/timeline/, requested 2003-08-11.