First Monday, Volume 16, Number 4 - 4 April 2011

f i ® s t m ¤ ñ d @ ¥
PEER-REVIEWED JOURNAL ON THE INTERNET

The impact of group attributes on communication activity and shared language in online communities

by David A. Huffaker

Abstract

This study examines how group size, network density, participation equality and member turnover are related to communication activity and shared language in online communities. It examines the social interactions of 16 randomly selected discussion groups (33,450 participants and 632,622 messages) covering topics in politics, health, hobbies, and science and technology over a 20−month period. It relies on social network analysis, computational linguistics and hierarchical linear modeling to uncover the extent to which group attributes impact message replies, conversational thread length and shared word choices. The results show that group size and network density are positively related to communication activity and shared language, while participation equality and member turnover are not significant predictors.

Contents

As online communities continue to dominate the Internet, understanding how they impact the creation and flow of information remains of paramount concern to scholars and technologists. While scholars have explored social interaction in online communities for some time (Baym, 2000; Preece, 2000; Rheingold, 2000), we still know little about the extent to which group attributes impact an individual's communication activity and social influence. This study investigates the extent to which group attributes such as member size, interconnectedness, equality of participation and turnover affect the ability for individuals to trigger feedback from others, enhance user engagement, or diffuse information.

Scholars have been increasingly able to test existing social science theories within large−scale online networks, offering new insights into the ways in which individuals and groups interact to create and share information. This study examines the social behaviors of over 30,000 participants, who contributed over 600,000 messages to a variety of online discussion groups over a 20−month period. The study of online discussion forums is especially relevant because they are often utilized for expertise−sharing and collaboration, allowing users to ask and answer questions, post important events or news, or to openly discuss or debate a variety of topics. Understanding how group attributes affect these social behaviors provides important insights for the design of deployment of these technologies.

This research makes several contributions. First, it extends existing theoretical frameworks (Heinz and Rice, 2009), by providing empirical support using a large dataset from one the

FM
Volume 16, Number 4 - 4
April 2011

Table of Contents

Reading Tools

The impact of gro...

Huffaker

- Abstract
- Review policy
- About the author
- How to cite item
- Indexing metadata
- Print version
- Notify colleague*
- Email the author*

Related items
- Author's work
- Government policy
- Book
- Book reviews
- Dissertations
- Online forums
- Quotations
- Resources
- Media reports
- Web search

Search journal

All

Search

Close

* *Requires registration*

most popular and active online settings. Second, this research demonstrates how individual–level outcomes can be predicted using group–level variables, and provides several other novel approaches for measuring the attributes of online groups using social network analysis and computational linguistics. And most importantly, it shows that: (a) group size is not a hindrance to online communication in the ways previously noted; (b) that the interconnectedness of the group remains extremely important for engagement and influence; and, (c) that participation equality and member attrition may not affect user feedback as one might expect. These are important insights for information and communication scholars, as well as technology companies and designers.

Background

Scholars have noted the important role that groups play in information exchange in computer–mediated systems early on (Rice, 1982), including the impact that group norms and other constraints have on individual members (Rice, 1987). These include structural aspects of the group, such as its size and stability. They also include social aspects, such as the level of commitment, shared sociocultural properties or language, and the norms of openness and trust among its members. Each of these attributes are important in online groups, as they can impact the production and exchange of information and knowledge (Heinz and Rice, 2009; Rice, 1987).

For example, while more group members can increase the overall number of links between members, the amount of information that can be exchanged, or the diversity of roles in the group, Rice (1987) outlines two problems as a group's size increases. First, there is an upper limit to the amount of information that individuals in the group can process and the amount of reciprocal relations they can maintain [1]. In other words, information processing does not scale well with group size. Second, as group size increases, they become "loosely coupled" and it becomes more difficult for individuals to integrate into the group or understand its conventions [2]. Together, these constraints can break down a group's structure, leading to problems for communication and information exchange.

In addition to size, Heinz and Rice (2009) propose a set of group–level attributes that impact participation and contribution in online groups. They argue that interaction frequency, shared language, commitment, openness, and trust positively impact participation in online group systems. Interaction frequency refers to the interconnections between members of the group, or the overall network density. The frequency of interaction can impact the strength of relational ties between individuals, or increase trust and social influence [3]. Shared language within a group represents a common linguistic background, in which individuals converge on a set of words including acronyms or jargon [4]. Commitment represents a strong sense of belonging to the collective or a group identity, which increases the amount of participation and knowledge sharing [5]. Openness refers to norms where cooperation, sharing and participation are encouraged [6]. Similarly, trust represents "a belief in the competence, openness, good intentions and reliability of others" [7].

While shared language is a potential explanatory variable for increases in information contribution and exchange (Heinz and Rice, 2009), it is also a potential outcome variable because it represents a type of social influence. For example, in dyadic interactions, social psychologists and sociolinguists show that speakers tend to influence each other's word choices as a way to come upon shared meaning, and that some individuals influence this convergence better than others (Brennan and Clark, 1996; Chambers, 2001; Garrod and Pickering, 2004). Labov (2001) finds that language convergence happens at a large–scale in local communities, and that when two groups are in constant communication, convergence is expected; when groups are separated, divergence occurs (Labov, 2002). Studies in computer–mediated communication (CMC) also suggest that users tend to converge upon shared topics in addition to their usual general chatter (Gruhl, *et al.*, 2004). Applications such as Google Trends and Twitter's "trending" features demonstrate how Internet users begin to talk about the same items using the same word choices. Because the primary goal of this study is to understand how group properties can affect communication, shared language will be considered an outcome variable.

The theoretical framework of Heinz and Rice (2009) is adapted to include the propositions of group size by Rice (1987). As shown in Figure 1, it is proposed that group size, network density, commitment, and openness and trust are all correlated with communication activity and shared language. As described next, several studies in CMC lend support to this framework.

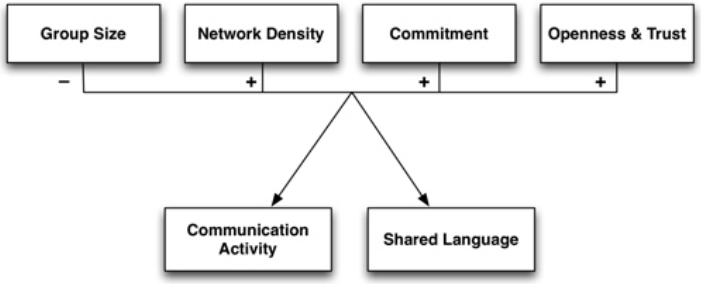

Figure 1: Theoretical model of the impact of group attributes on communication activity and shared language.
Note: Framework adapted from Rice (1987) and Heinz and Rice (2009).

First, there has been some empirical work in CMC settings on the impact of membership size on communication activity. In a study of over 1,000 listservs, Butler (2001) finds that size impacts group stability, and contributes both the retention of new members and the loss of others. Further, he finds that this effect is mediated by communication activity. In a study of 500 newsgroups, Jones, *et al.* (2004) finds that membership size impacts the content of online communication; as it increases, messages become shorter and less. Given the results of these studies, it is expected that there is a negative relationship between group size, communication activity and shared language.

Some CMC research has examined the impact of interconnectedness and network density. A study of Sourceforge.net, an open source software community, reveals that "network embeddedness" — which represents the interconnectedness and network density of the project groups — has a positive impact on both the technical and commercial success of a group's product (Grewal, *et al.*, 2006). In a study of Yahoo! Answers, strongly interconnected groups demonstrate more discussion, more reciprocity and a sense of group identity (Adamic, *et al.*, 2008). More generally, when new users connect with the group via message responses, they are more likely to continue participating in the group (Arguello, *et al.*, 2006; Joyce and Kraut, 2006). Based on these findings, it is expected that the interconnectedness of online groups should positively impact communication activity and shared language.

In terms of commitment, it is widely held that membership turnover can have a negative impact on group performance and member loyalty (Van Vugt, *et al.*, 2004). In a recent study of open source software communities, Howison, *et al.* (2006) show the dynamic nature of online groups and a large amount of attrition, find that a few participants remain steady members and facilitate its communication. Butler (2001) finds a negative association between membership turnover and communication activity on a listserv community. Intuitively, member turnover will likely impact the structure and content production of online groups. If users come and go quickly, they have less chance to make a major contribution to the group or influence other members. Therefore, a negative relationship is expected in terms of communication activity and shared language.

As Joyce and Kraut (2006) show, when newcomers receive feedback from group members, they feel welcome and are more likely to contribute and become committed to the group. Similarly, when new users receive positive feedback and opportunities to learn the norms of the group, they are more likely to continue participation (Lampe and Johnston, 2005). Thus, a group that displays more openness and equality is more likely to have more members interacting and participating. As described in more detail below participation equality is used as a proxy for trust and openness, as it shows that all users are participating equally and openly, a common goal for online communities (Koh, *et al.*, 2007; Nonnecke, *et al.*, 2006).

Methodology

Sample

The sample in this study consists of 16 randomly selected discussion groups found on

Google Groups between from 21 June 2003 and 31 January 2005 (comprised of 33,540 users and 632,622 messages). Google Groups is an evolved form Usenet, which allows users to post messages to a variety of online forums, and reply to the posts of other users, forming a conversational thread between two or more authors. Authors are able to post new messages (with a unique subject header), or reply to any individual author in the thread. Note that this is distinguished from early Usenet interactions, in which authors always reply to the overall group.

All social interactions that took place on Usenet (who replied to whom and when) were captured by the Netscan project (Smith, 2007), along with all message content exchanged between users during the 20−month period (Kraut, *et al.*, under review). From this data set matching social interactions with actual message content (*i.e.*, 2.2 million messages from 99 discussion groups), a strata of four genres of discussion groups were randomly sampled: (a) Politics; (b) Health and Support (c) Recreation and Hobby; and, (d) Science and Technology, which researchers argue are among the most popular types of online community topics (Horrigan, 2001). Table 1 lists the names and descriptions of the randomly selected discussion groups.

Table 1: Summary of Google Group names, categories and descriptions.
Note: Descriptions and categories were identified from information from Google Groups. In some cases, the description was ascertained after reviewing several discussion group messages. These are marked with brackets ("[ ]").

Name	Category	Description
alt.politics.economics	Politics	War == Poverty, & other discussions
alt.politics.radical−left	Politics	Who remains after the radicals left?
alt.politics.usa.constitution.gun-rights	Politics	Constitutional ramifications of gun rights
talk.politics	Politics	[General political discussions]
alt.support.cancer.breast	Health and Support	Support for those diagnosed with breast cancer and their families
alt.support.depression	Health and Support	Depression and mood disorders
alt.support.diabetes	Health and Support	Support for dealing with diabetes and related topics
alt.support.hepatitis−c	Health and Support	[Support for dealing with hepatitis−c and related topics]
rec.arts.manga	Recreation and Hobbies	All aspects of the Japanese storytelling art form
rec.crafts.textiles.quilting	Recreation and Hobbies	All about quilts and other quilted items
rec.music.blueNote:blues	Recreation and Hobbies	The Blues in all forms and all aspects
rec.food.veg.cooking	Recreation and Hobbies	Vegetarian recipes, cooking, nutrition
sci.op−research	Science and Technology	Research, teaching & application of operations research
sci.chemistry	Science and Technology	Chemistry and related sciences
sci.lang	Science and Technology	Natural languages, communication, etc.
microsoft.public.security	Science and Technology	Deals with security issues for Microsoft products

Procedure

Information regarding the posting and reply behaviors of all authors in each group, such as how often they posted, which messages were posts or replies, and when they first appeared and left the topic group, are captured as user log files on a mySQL database. These user logs are used to extract general information about the size and communication activity of each group, and to create the social graphs for each discussion group based on who replied to whom. In addition to the user log and social network analysis, this study utilizes text analysis to examine the content of each message. Message content was extracted for each individual message, converted to text files and preprocessed to remove all headers, subject lines, signatures and quoted text. All stop words (*e.g.,* a, an, the) were also removed in order to focus on content−bearing words.

Dependent measures

In order to identify the amount of communication activity and shared language, three measures are used. This includes the amount of message replies that individuals receive, the amount and length of conversational threads, and the frequency of share words between any two connected messages. These are discussed in more detail below.

Communication activity

Message replies. The first measure of communication activity represents the feedback between members of the group. Online groups that demonstrate more feedback and reciprocity tend to retain more members and increase overall participation (Arguello, *et al.*, 2006; Butler, 2001), therefore this particular pattern of communication activity is important. It is calculated as the total number of replies that each individual in the group receives from other members.

Conversational thread length. Successful online interactions include those in which users post a message or reply that sparks a long dialogue between other users (Matsumura and Sasaki, 2007). Whereas message replies provides a sense of how responsive the group is to individual members, conversational thread length represents the quality and depth of the communication. Longer threads suggest a deeper interaction or level of engagement among users. Therefore, conversational thread length is used in conjunction with message replies to establish both feedback levels and user engagement.

Conversational thread length is calculated as the number of messages that proceed from a user′s initial post. As illustrated in Figure 2, an initial post has the potential to branch out into a series of threads. *Author A* posts a message about a new health product and receives three direct replies. *Reply 1* and *2* spur additional replies, and *Reply 2a1* is a third−tier reply. Because *Author A* initiated the discussion or question, the number of proceeding replies and sub−replies is calculated (in this case, she receives a score of "7″). Likewise, *Author B* receives a score of "2″ for receiving two direct replies in the chain, and *Author C* receives a score of "2″ for receiving a direct reply and starting a new sub−chain. Note that an author does not receive an additional score if she replies within their own thread (this avoids an inflated score if a single author does all the talking in the thread). A total conversational thread score is calculated for each individual user.

Shared language

Using a similar approach to calculating the conversational thread length, the number of shared word choices is calculated for each author. Matsumura, *et al.* (2002) refer to this approach as the *influence diffusion model*, and argue that it is an appropriate proxy of social influence in text−based communication. For example, as Figure 2 depicts, if *Author A* includes the term "nike″ in a post, and *Author B* also says "nike″ in his reply, they are converging upon a shared set of lexical items. Likewise, the entire group may begin using the same word choices, having been influenced by the initiators of the conversation. For each individual poster, the total number of words that are repeated in subsequent messages down the conversational thread are calculated.

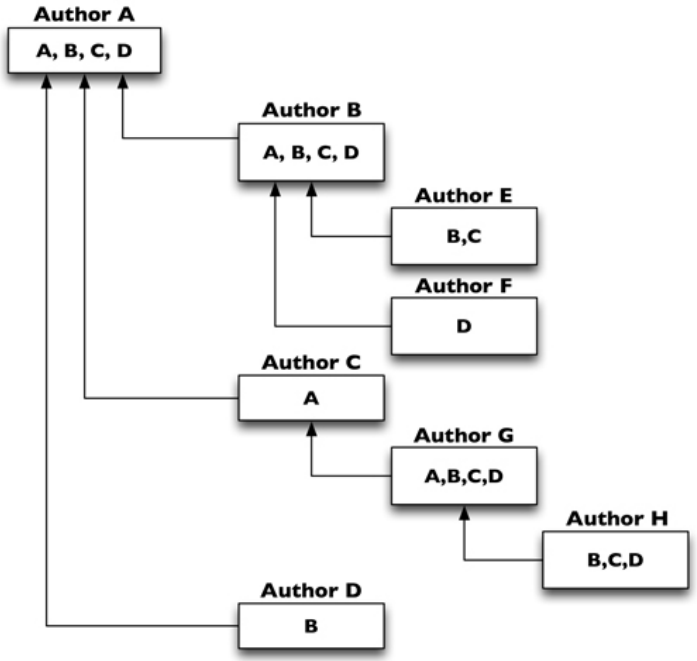

Figure 2: Example of conversational threads and shared language in an online discussion.

In order to capture the shared words, *Text::Similarity* (Pederson, *et al.*, 2008), an open source Perl module is used. *Text::Similarity* counts the frequency of overlapping words or phrases between two text files. The frequency of shared words or phrases is normalized by the length of the each file. This ensures that longer messages do not have a greater chance of diffusing words since there are more opportunities.

Independent measures

By aggregating the user log data, unique characteristics of each topic group, including its membership size and the dynamics of overall participation are extracted. This creates a nested structure, providing several group−level predictors described below, which are later used to measure their correlation with the individual−level communication activity and shared language dependent variables. Except for membership turnover, all group variables were log-transformed to reduce the positive skew in the data distribution.

Group size

Group size was calculated as the average number of authors that contribute to the topic group within a three−month period (*i.e.*, quarterly) for the 20−month period, and includes both message posts and replies. This was chosen rather than the total number of authors over the entire period, since the rate of attrition is high and users should only be influenced by the size of the community by which they are surrounded.

Network density

In this analysis, network ties exist when one user replies to the message of another user. Network density represents the proportion of ties between authors divided by the total potential connections available. The more that users reply to other users, the more interconnected the group. In directed graphs such as this one, density can be also interpreted as the average tie strength of users (Hanneman and Riddle, 2005). In this case, density is calculated as:

$$\Delta = \frac{L}{g(g-1)} \quad (1)$$

In (1) *L* is the number of arcs, or ordered pairs of nodes, and *g(g − 1)* is the possible number of arcs in the network (Wasserman and Faust, 1994). UCINET is utilized to calculate this measure (Borgatti, *et al.*, 1999).

Member turnover

In order to understand the level of commitment in an online group, the relative level of stability is assessed. This is measured in terms of member turnover as the absolute value of the average percentage change of contributing authors every three months over the 20−month period.

$$\sum \frac{\left| t_{n+1} - t_n \right|}{t_n} \quad (2)$$

In (2), *t* is the average number of authors, and *n* is a particular time period. Then the absolute value of the average across all time periods to create a final percent change value.

Participation equality

Participation inequality within each group represents the distribution of the proportion of participation, whether a message post or a reply, by all members of a particular topic group. Smith (1999) proposed a poster−to−post ratio to measure interaction quality in each group; however, this measure does not capture reply structures. Therefore, a different measure is proposed to capture the openness of the group, which is called participation equality.

Participation equality is calculated using a Lorenz curve and the inverse of the Gini coefficient. The Lorenz curve, typically used to graphically show the inequality of an income distribution (Kakwani, 1977), demonstrates the proportion of total income given a percentage of a population. The exemplar is Pareto's argument that 80 percent of the wealth resides with 20 percent of the Italian population (Rosen and Resnick, 1980).

This concept is applied here to represent a cumulative distribution of participation within each topic group. As shown in Figure 3, the x−axis represents the cumulative proportion of participants ranked by their participation level, and the y−axis represents the cumulative proportion of participation for a given proportion of the user population. For a technical description of the Lorenz curve calculation, see Gastwirth (1972). To measure participation, the total posts and replies for each participant is identified.

After a Lorenz curve is calculated, the degree of inequality can be measured using the Gini coefficient (Gastwirth, 1972). First, a line of perfect equality is created where y = x (*i.e.*, a 45−degree line). Second, the Gini coefficient is calculated as a ratio of: (a) the area between the line of perfect equality and the Lorenz curve, and (b) the area beneath the Lorenz curve and the axes (*i.e.*, a line of perfect inequality). The higher this coefficient is, the more *unequal* the participation distribution remains. For a technical description of the Gini coefficient calculation, see Atkinson (1970). Because the Gini coefficient represents inequality, the inverse of the Gini coefficient was calculated in order to represent participation equality.

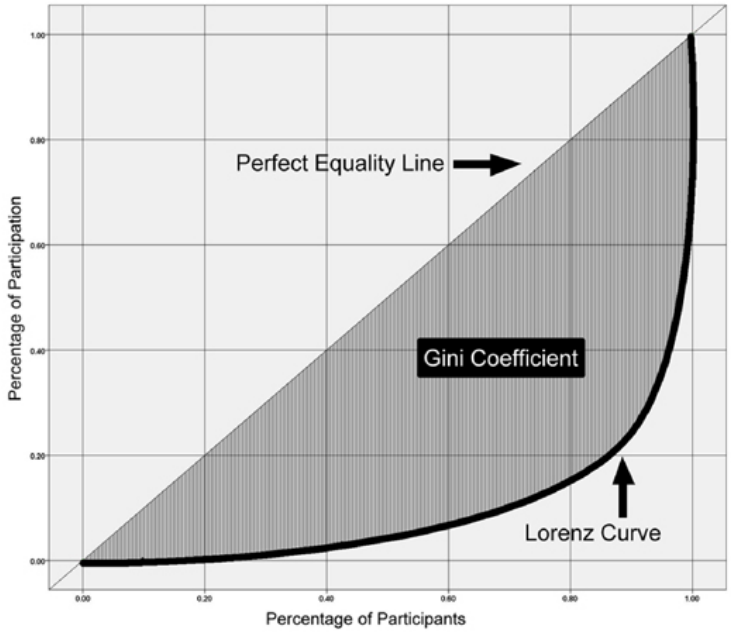

Figure 3: Example of a Lorenz curve, perfect equality line and Gini coefficient for a random topic group.

Results

In order to examine the relationship between group−level attributes and the communication activity and shared language of the individual members, hierarchical linear modeling (HLM) is employed. HLM is appropriate for data structures where individual−level variables are nested within group−level variables. All three dependent variables (*i.e.*, message replies, conversational thread length and shared language) represent event count data, which is integer−based and non−negative. Therefore, the Poisson regression model, a special non−linear generalized modeling approach (Cameron and Trivedi, 1998) is used to fit the data. Note that Poisson regression relies on a log−transformation of the dependent variable, and requires an antilog transformation of the coefficients of each predictor in the regression model (Gelman and Hill, 2007). Throughout these analyses, the *exp(B)* — the antilog — is included when interpreting each variable. Finally, all predictors were grand−mean centered, which is a common and recommended practice in HLM models (Gelman and Hill, 2007).

Descriptive statistics

The means, standard deviations and range of each group variable are listed in Table 2. Group size represents the average number of authors participating in the group during each quarter of the 20−month time period (M=362.58, SD=353.49). The other group variables, participation equality and network density, suggest that these groups are fairly sparse and unequal in the amount of participation among all authors. Member turnover, which represents the percent change of authors during all eight quarters of the 20−month period, suggests a general increase in authors over time (M=.20, SD=.93). As shown in Table 3, there were no significant correlations between the group-level predictors.

Table 2: Summary of means, standard deviation and range for variables of interest.
Note: N=16.

Variables	M	SD	Min	Max
Dependent variables				
Message replies	14.05	93.24	0	7,369
Conversational thread length	12.08	86.04	0	5,945.36
Shared language	1.05	10.14	0	731.21
Independent variables				
Group size	362.58	353.49	0	7,369
Network density	.008	.0009	.0002	.034
Member turnover	.55	.99	.12	4.23
Participation equality	.21	.10	.07	.42

There are also no significant differences between group type (*i.e.*, politics vs. health vs. hobbies vs. science) and the group−level variables. A chi−square analysis shows no significant differences for size (p=.35), participation equality (p=.52), density (p=.43) or turnover rate (p=.35). So while there is variance across all 16 groups, no topic area shows a significantly different size, complexity or stability.

Table 3: Correlation matrix for group size, complexity and turnover.
Note: Bivariate correlations, two−tailed tests. N=16.

Variables	1	2	3
1. Size	—		
2. Network density	−.331	—	
3. Member turnover	−.060	.050	—
4. Participation equality	−.493	−.307	−.174

There is a high attrition rate associated with many online communities (Andrews, *et al.*, 2003), and these discussion groups are no exception. In this sample, only 56 individuals (less than one percent) contribute over the entire 20 months. Of these, 21 individuals are from alt.support.depression, while 16 reside in talk.politics. Six of the 16 groups have no authors that remain through the entirety. In the groups focused on hobbies and recreation, no one contributed over the entire 20 months, while the majority of long−term contributors came from groups focused on health and support (*i.e.*, 25 participants) and politics (*i.e.*, 22 participants), with science and technology a distance third (*i.e.*, nine participants).

In fact, the size and contribution rates of this random sample of discussion groups vary. At least four groups show higher frequencies of posting new messages and replying to other messages. It is also important to note that these prolific groups are not always the ones with the most authors. For example, alt.support.diabetes and rec.crafts.textiles.quilting have fewer authors and still contribute at the highest levels. By contrast, microsoft.public.security has many authors, but not many contributions; see Figures 4 and 5.

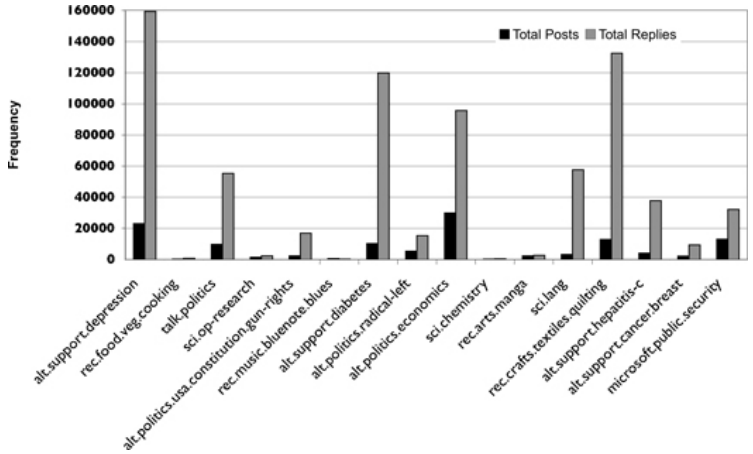

Figure 4: Post and reply behavior of 16 randomly selected discussion groups.

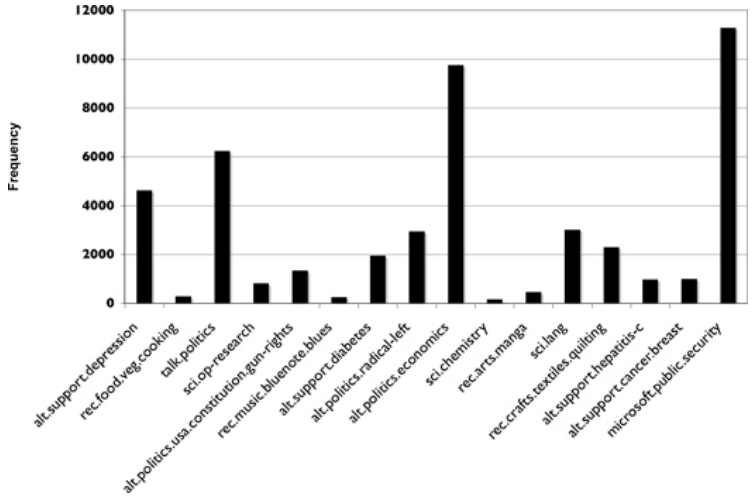

Figure 5: The impact of group attributes on communication activity and shared language.

It is typical to begin with a baseline or null model in which none of the predictors are included, allowing an investigator to examine the independence of the group−level units on the outcome variable using the chi−square difference test (Hoxx, 1995). In this case, the baseline model suggests that differences in discussion group size, equality of participation, network density and turnover rate are different than zero across groups (χ^2 (15) =1147.85, ρ<.001). After assessing the baseline model, the group−level predictors are added to create a means−as−outcomes HLM model, which measures the group−level effect on the individual outcome variable. The model is a good fit, χ^2 (11) =39.62, ρ=.001.

It hypothesized that group size (H1) and member turnover (H3) would negatively impact message replies, conversational thread length and shared language (H1), while network density (H2) and participation equality (H4) would positively impact message replies, conversational thread length and shared language. The results show support for the network density hypothesis, but find that group size has a positive impact, and that

member turnover and participation equality show no effect. A separate model and description are provided for each dependent variable.

Message replies

As shown in Table 4, larger discussion groups, measured in terms of the average number of participants every three months for a 20−month period, are positively related to message replies, and each additional member can result in roughly six more replies (B=1.79, SE=.30, $exp(B)$=5.99, t(11) =6.02, ρ<.001). Discussion groups with more network density, or more connections between members in terms of message replies, are positively related to message replies, and each additional tie can result in over five more replies (B=1.70, SE=.11, $exp(B)$=5.49, t(11)=15.20, ρ<.001). Contrary to prediction, discussion groups with more participation equality are not significantly related to message replies (ρ=.91). Nor are groups with a high member turnover related to message replies (ρ=.30).

Table 4: Summary of hierarchical regression analysis for group variables predicting the number of message replies.
Note: The restricted maximum likelihood method is used for estimation. Predictor variables are estimates of the fixed effects, γs, with robust standard errors, and adjusted for overdispersion. N=16. *ρ<.05. **ρ<.01. ***ρ<.001.

Variable	*B*	*SE*	*Exp(B)*
Group size	1.78***	.30	5.99
Network density	1.70***	.11	5.49
Turnover	.03	.03	1.03
Participation equality	.06	.53	1.06

Conversational thread length

The model is a good fit, χ^2 (11)=71.80, ρ<.001, and shows the same results as the model for message replies. As shown Table 5, larger discussion groups are positively related to conversational thread length, and each additional member can result in roughly four more threaded messages (B=1.46, SE=.39, $exp(B)$=4.30, t(11) =3.71, ρ<.005). Discussion groups with more network density are positively related to conversational thread length, and each additional tie can result in roughly four more threaded messages (B=1.46, SE=.21, $exp(B)$=4.32, t(11) =6.96, ρ<.001). Again, contrary to prediction, discussion groups with more participation equality are not significantly related to conversational thread length (ρ=.49). Nor are groups with a high member turnover related to conversational thread length (ρ=.84).

Table 5: Summary of hierarchical regression analysis for group variables predicting conversational thread length.
Note: The restricted maximum likelihood method is used for estimation. Predictor variables are estimates of the fixed effects, γs, with robust standard errors, and adjusted for overdispersion. N=16. *ρ<.05. **ρ<.01. ***ρ<.001.

Variable	*B*	*SE*	*Exp(B)*
Group size	1.46***	.39	4.30
Network density	1.46***	.21	4.32
Turnover	.01	.05	1.01
Participation equality	-.54	.75	.58

Shared language

The model is a good fit, χ^2 (11) =34.61, ρ<.001, and reflects the same previous findings. As shown in Table 6, larger discussion groups are positively related to shared language, and

each additional member can result in roughly six more mirrored words (*B*=1.34, *SE*=.26, *exp (B)*=3.81, *t*(11) =5.09, ρ<.005). Discussion groups with more network density, or more connections between members in terms of message replies, are positively related to shared language, and each additional tie can result in four more mirrored words (*B*=1.54, *SE*=.26, *exp(B)*=4.67, *t*(11) =8.84, ρ<.001). Contrary to prediction, discussion groups with more participation equality are not significantly related to shared language (ρ=.99). Nor are groups with a high member turnover related to shared language (ρ=.43).

Table 6: Summary of hierarchical regression analysis for group variables predicting shared language.
Note: The restricted maximum likelihood method is used for estimation. Predictor variables are estimates of the fixed effects, γs, with robust standard errors, and adjusted for overdispersion. N=16. *ρ<.05. **ρ<.01. ***ρ<.001.

Variable	*B*	*SE*	*Exp(B)*
Group size	1.34***	.26	3.81
Network density	1.54***	.17	4.67
Turnover	-.02	.04	.97
Participation equality	-.01	.78	.99

In sum, group size and density are positively related to online communication activity and shared language, while participation equality and member turnover show no effects. In the next section, these findings are discussed in relation to previous work, along with implications for the design of technology.

Discussion

This study examines how group attributes impact communication activity and shared language in online discussion groups. Based on an adapted theoretical model from Rice (1987) and Heinz and Rice (2009) and previous empirical work, several key group variables are identified including: (a) group size, which is calculated as the average number of authors who contributed to the group every three months during a 20−month period; (b) network density, which stands for the proportion of linkages between members of the group to the potential number of linkages; (c) member turnover, which is captured by the percentage change in membership every three months for a 20−month period and, (d) participation equality, which measures the proportion of participation from all members of the group, and is used a proxy for openness and egalitarianism.

Communication activity was measured in two ways: (1) the number of message replies that individuals receive from other members of the group, which represents the level of feedback from group members; and (2) the conversational thread length, which is calculated as the amount and length of message threads spurred by individuals, which represents the amount of user engagement and depth in the communication. Shared language was measured as the number of words or phrases repeated down the conversational thread.

In order to examine the relationship between these group−level and individual−level variables, hierarchical linear modeling (HLM) was employed. The findings show that network density positively impacts communication activity and shared language, while participation equality and member turnover are not significantly related. Contrary to expectation, group size shows a positive impact on communication activity and shared language.

This effect is particularly interesting in light of both face−to−face and CMC literature. For example, in face−to−face environments, large organizations can become cumbersome, hinder communication, limit the connections between members, and create more social distance (Bass, 1990). Even in small groups, as the size increases, participants have fewer opportunities to talk or engage in leadership activities (Hare, 1976), probably due to norming behaviors or fear of rejection (Brown, 2006). Research in CMC argues that while group size can increase the connections and roles of members, it suffers from an upper bound, and ultimately creates a negative impact on group communication (Rice, 1987). This is because large online groups can quickly become too complex, causing information overload and a lack of group cohesion (Rice, 1987).

However, group size appears to have an opposite effect in these online discussion groups. The larger the group, the more feedback, engagement and linguistic convergence is witnessed. It is as if more members mean more opportunities to contribute, respond, discuss and spread ideas. There are a several possible explanations for this. First, the nature of online discussion groups may reduce the information overload that Rice (1987) notes. These conversations are persistent, archived and searchable. The user interface makes it easy for users to engage in relevant threads and ignore others, or to search for a particular conversation they are interested in. These interactions are likely a mix of topical 'spikes' and more stable patterns of general discussion (Gruhl, *et al.*, 2004), and all users can converge on a particular news topic, or debate a political issue (Kelly, *et al.*, 2005), while others spend more time answering questions and sharing knowledge (Constant, *et al.*, 1996). In other words, more participants result in more collective benefits in terms of communication activity and social influence.

While Rice (1987) notes that larger groups suffer from segmentation and loss of group cohesion, this may not be as strong an issue with voluntary discussion groups. The dynamic nature of these groups likely makes cohesion difficult to achieve; the results suggest that only a handful of authors stick around for the entire time period, and that many users come and go in a three−month period. It could be that although there is a lot of attrition, a subset of users represents a cohesive clique that maintains the group norms and institutional knowledge despite the overall size of the group. This might also explain why participation equality and member turnover do not matter, as will be discussed in more detail below.

A third point on group size is that these groups may not be reaching the upper bounds that Rice (1987) identifies. The average group size is roughly 350 members during a three−month period, and the largest group is roughly 1,000 members. This group size might be more manageable in online settings and not exceed a threshold that makes communication and cohesion more difficult. Future work should attempt to uncover if a size threshold does indeed exist, and if there are curvilinear effects on group dynamics because of it.

Network density is also significant predictor of communication activity and shared language. This suggests that as users provide more and more feedback to one another, their level of engagement increases, as well as the ability for individual members to influence the language and topics of the group as a whole. This resonates with the concept of scale−free networks, in which major communication hubs are connected by many smaller hubs, resulting in close connections between all participants (Barabási and Crandall, 2003). Second, it also implies that highly connected groups represent a layer of social support, and that interconnections represent some level of relationship building and maintenance behaviors. As users become more connected, they begin to engage in deeper conversations, and share the same sociocultural and linguistics behaviors that Heinz and Rice (2009) suggest. Many CMC scholars have pointed out that even large−scale online networks exhibit social support and strong ties between community members (Cummings, *et al.*, 2002; Wellman and Gulia, 1999), and this research shows that group connectedness leads to more participation, feedback, engagement and social influence.

On the other hand, participation equality, which includes the proportion of message posts that all members of the group made, is not a significant predictor of communication activity and shared language. This suggests that disproportionate participation does not necessarily hinder feedback, engagement or social influence. Admittedly, participation equality does not encapsulate the openness and trust that Heinz and Rice (2009) describe. Their conceptualization surrounds a cultural milieu that is difficult to assess from behavioral data. However, it still hints at the kind of openness that would encourage all members of the group to participate equally, a reason why scholars have hailed online communities (Rheingold, 2000), and a reason why it is a useful variable to study.

The findings show that groups have low participation equality on average, suggesting that most of the contributions come from a subset of users. This could be a feature of online discussion groups, where many users might come in to ask a single question and never come back (Joyce and Kraut, 2006), or that the majority of the feedback, conversation and shared language occurs among a subset of users. Previous work suggests that leaders of online groups tend to contribute more messages than typical users, and engage in both technical and social tasks in order to keep the group thriving (Butler, *et al.*, 2008). If this is the case, then it may be naïve to expect that all users must participate equally in order to encourage feedback, user engagement and social influence.

Membership turnover, which is measured in terms of percentage change in membership, is not significantly associated with message replies, conversational thread length or shared language. While groups in this study maintain a *status quo* in membership size, there is a dynamic nature of online groups that includes high levels of attrition. Only 50 or so authors of over 30,000 contributed through the entire period examined. However, it appears that many authors participate for months at a time and serve as the foundation for interaction and social influence regardless of whether newcomers arrive or some veterans depart. And although commitment to the group can lead to its success, the findings here show that high member turnover does not negatively affect the ability of an individual to trigger a response, spark conversation or diffuse language. In light of the group size and participation equality observation, it could be that a subset of users might be making steady

impact on the other users, even if they come and go. Future work should investigate whether this subset exists, and the extent to which they impact the behavior of the group–at–large.

In effect, these findings show the importance of reaching a critical mass when developing online communities. Larger groups or communities will result in more feedback, more conversations and more diffusion of information. But what is intrinsic to this success is the ability to connect members to one another, and to facilitate interactivity among as many members as possible. These design features seem well integrated into the most successful online communities, including Facebook, Twitter, YouTube, all of which provide plenty of opportunities to contribute content, respond to other contributions, and have constant access to new members in order to extend one's social network. Recommendation systems, community managers, and tangible or intangible rewards for interacting with newcomers and old timers would likely foster increased connectedness, keeping the application growing and thriving.

While these findings have raised important insights on the role of group size and density — and the lack of role for participation equality and member turnover — for increasing participation and social influence, there are some limitations. First, the generalizability of Google Groups and its Usenet predecessor is questionable. As others have shown, online discussion groups develop their own cultural practices (Baym, 2000; Tepper, 1997), which may not always translate to other online communities or social media platforms. Second, some unique user behavior such as *trolling*, in which a user purposely provokes others to respond (Herring, *et al.*, 2002), or *flaming*, in which users begin engaging in heated exchanges, passing insults or harassing one another (Alonzo and Aiken, 2004) might inflate the community activity measures. Finally, the shared language variable could be refined in ways to distinguish words that all members of a particular group use (*e.g.*, politics, republicans, democrats) from new words introduced by individuals and mirrored later (*e.g.*, health care reform, tea parties, bipartisanship). Even so, the study of online discussion groups and the current measures still provide important insights for computer-mediated communication and information science.

In conclusion, the results show that group sizes reaching 1,000 members, coupled with the interconnections between these members, increase the communication activity and shared language of online groups. They also show that user attrition and participation equality may not hinder feedback, engagement and social influence in ways previous thought. For technology designers, this suggests that customer attraction is more important than retention, and that encouraging subsets of 'power users' will likely keep the group thriving. For scholars, this work contributes to our understanding of group dynamics, and the importance of including group–level measures in the study of online user behavior.

About the author

David Huffaker (Ph.D., Northwestern University) is a researcher at Google, Inc.
Web: http://www.davehuffaker.com/.

Notes

1. Rice, 1987, p. 112.

2. Rice, 1987, p. 113.

3. Heinz and Rice, 2009, p. 150.

4. Heinz and Rice, 2009, p. 147.

5. Heinz and Rice, 2009, p. 149.

6. Heinz and Rice, 2009, pp. 148–149.

7. Heinz and Rice, 2009, p. 148.

References

L. Adamic, J. Zhang, E. Bakshy, and M. Ackerman, 2008. "Knowledge sharing and Yaho!

Answers: Everyone knows something," paper presented at the WWW '08 (Beijing, China), at http://www2008.org/papers/fp840.html, accessed 28 March 2011.

M. Alonzo and M. Aiken, 2004. "Flaming in electronic communication," *Decision Support Systems*, volume 36, number 3, pp. 205–213.

D. Andrews, B. Nonnecke, and J. Preece, 2003. "Electronic survey methodology: A case study in reaching hard–to–involve Internet users," *International Journal of Human–Computer Interaction*, volume 16, number 2, pp. 185–210.

J. Arguello, B.S. Butler, E. Joyce, R.E. Kraut, K.S. Ling, C. Rosé, and X. Wang, 2006. "Talk to me: Foundations for successful individual–group interactions in online communities," *CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 959–968.

A.B. Atkinson, 1970. "On the measurement of inequality," *Journal of Economic Theory*, volume 2, number 3, pp. 244–263.

A–L. Barabási and R.E. Crandall, 2003. "Linked: The new science of networks," *American Journal of Physics*, volume 71, number 4, p. 409.

B.M. Bass, 1990. *Bass & Stogdill's handbook of leadership: Theory, research, and managerial applications*. Third edition. New York: Free Press.

N.K. Baym, 2000. *Tune in, log on: Soaps, fandom, and online community*. Thousand Oaks, Calif.: Sage.

S.P. Borgatti, M.G. Everett, and L. Freeman, 1999. *UCINET V: Software for social network analysis*. Natick, Mass.: Analytic Technology, at http://www.analytictech.com/ucinet/, accessed 28 March 2011.

S.E. Brennan and H.H. Clark, 1996. "Conceptual pacts and lexical choice in conversation," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, volume 22, number 6, pp. 1,482–1,493.

C. Brown, 2006. *Social psychology*. London: Sage.

B.S. Butler, 2001. "Membership size, communication activity, and sustainability: A resource–based model of online social structures," *Information Systems Research*, volume 12, number 4, pp. 346–362.

B.S. Butler, L. Sproull, S. Kiesler, and R. Kraut, 2008. "Community building in online communities: Who does the work and why?" In: S. Weisband and L. Atwater (editors). *Leadership at a distance: Research in technologically–supported work*. New York: Erlbaum, pp. 171–194.

A.C. Cameron and P.K. Trivedi, 1998. *Regression analysis of count data*. Cambridge: Cambridge University Press.

J.K. Chambers, 2001. "Dynamics of dialect convergence," *Journal of Sociolinguistics*, volume 6, number 1, pp. 117–130.

D. Constant, L. Sproull, and S. Kiesler, 1996. "The kindness of strangers: The usefulness of electronic weak ties for technical advice," *Organization Science*, volume 7, number 2, pp. 119–135.

J. Cummings, B. Butler, and R. Kraut, 2002. "The quality of online social relationships," *Communications of the ACM*, volume 45, number 7, pp. 103–108.

S. Garrod and M.J. Pickering, 2004. "Why is conversation so easy? Syntactic priming in language production," *Trends in Cognitive Sciences*, volume 8, number 1, pp. 8–11.

J.L. Gastwirth, 1972. "The estimation of the Lorenz curve and Gini index," *Review of Economics and Statistics*, volume 54, number 3, pp. 306–316.

A. Gelman and J. Hill, 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

R. Grewal, G. Lilien, and G. Mallapragada, 2006. "Location, location, location: How network embeddedness affects project success in open source systems," *Management Science*, volume 52, number 7, pp. 1043–1056.

D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, 2004. "Information diffusion through blogspace," *Proceedings of the 13th International Conference on the World Wide Web*, pp. 491-501.

R.A. Hanneman and M. Riddle, 2005. *Introduction to social network methods*. Riverside, Calif.: University of California Riverside, at http://faculty.ucr.edu/~hanneman/, accessed 28 March 2011.

A.P. Hare, 1976. *Handbook of small group research*. Second edition. New York: Free Press.

M. Heinz and R.E. Rice, 2009. "An integrated model of knowledge sharing in contemporary communication environments," *Communication Yearbook*, volume 33, pp. 134-175.

S. Herring, K. Job-Sluder, R. Scheckler, and S. Barab, 2002. "Searching for safety online: Managing 'trolling' in a feminist forum," *Information Society*, volume 18, number 5, pp. 371-384.

J. Horrigan, 2001. *Online communities: Networks that nurture long-distance relationships and local ties*. Washington D.C.: Pew Internet & Family Life Project, at http://www.pewinternet.org/Reports/2001/Online-Communities.aspx, accessed 28 March 2011.

J. Howison, K. Inoue, and K. Crowston, 2006. "Social dynamics of free and open source team communications," *Proceedings of the International Federation for Information Processing (IFIP) Second International Conference on Open Source Software* (Lake Como, Italy), pp. 319-330.

J.J. Hoxx, 1995. *Applied multilevel analysis*. Amsterdam: TT-Publikaties.

Q. Jones, G. Ravid, and S. Rafaeli, 2004. "Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration," *Information Systems Research*, volume 15, number 2, pp. 194-210.

E. Joyce and R. Kraut, 2006. "Predicting continued participation in newsgroups," *Journal of Computer-Mediated Communication*, volume 11, number 3, pp. 723-747, and at http://jcmc.indiana.edu/vol11/issue3/joyce.html, accessed 28 March 2011.

N.C. Kakwani, 1977. "Applications of Lorenz curves in economic analysis," *Econometrica*, volume 45, number 3, pp. 719-728.

J. Kelly, D. Fisher, and M. Smith, 2005. "Debate, division, and diversity: Political discourse networks in USENET newsgroups," paper prepared for Online Deliberation Conference 2005, Stanford University (24 May), at http://www.coi.columbia.edu/pdf/kelly_fisher_smith_ddd.pdf, accessed 28 March 2011.

J. Koh, Y.-G. Kim, B. Butler, and G.-W. Bock, 2007. "Encouraging participation in virtual communities," *Communications of the ACM*, volume 50, number 2, pp. 68-73.

R. Kraut, X. Wang, B. Butler, E. Joyce, and M. Burke, under review. "Beyond information: Developing the relationship between the individual and the group in online communities," *Information Systems Research*; version at http://www.cs.cmu.edu/~kraut/RKraut.site.files/articles/wang08-isr-relationship-rev2-submitted.pdf, accessed 28 March 2011.

W. Labov, 2002. "Driving forces in linguistic change," *2002 International Conference on Korean Linguistics* (2 August, Seoul National University), at http://www.ling.upenn.edu/~wlabov/Papers/DFLC.htm, accessed 28 March 2011.

W. Labov, 2001. *Principles of linguistic change*, volume 2: *Social factors*. Oxford: Blackwell.

C. Lampe and E. Johnston, 2005. "Follow the (slash) dot: Effects of feedback on new members in an online community," *Proceedings of GROUP '05: Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*, pp. 11-20.

N. Matsumura and Y. Sasaki, 2007. "Human influence network for understanding leadership behavior," *International Journal of Knowledge-Based and Intelligent Engineering Systems*, volume 11, number 5, pp. 291-300.

N. Matsumura, Y. Ohsawa, and M. Ishizuka, 2002. "Influence diffusion model in text-based communication," *International World Wide Web Conference* (Hawaii), at http://www2002.

org/CDROM/poster/109/, accessed 28 March 2011.

B. Nonnecke, D. Andrews, and J. Preece, 2006. "Non−public and public online community participation: Needs, attitudes and behavior," *Electronic Commerce Research*, volume 6, number 1, pp. 7−20.

T. Pederson, S. Patwardhan, S. Banerjee, and J. Michelizzi, 2008. *Text::Similarity*. Duluth: University of Minnesota, at http://www.d.umn.edu/~tpederse/text-similarity.html, accessed 28 March 2011.

J. Preece, 2000. *Online communities: Designing usability, supporting sociability*. Chichester, U. K.: Wiley.

H. Rheingold, 2000. *The virtual community: Homesteading on the electronic frontier*. Cambridge, Mass.: MIT Press.

R.E. Rice, 1987. "New patterns of social structure in an information society," In: J.R. Schement and L.A. Lievrouw (editors). *Competing visions, complex realities: Social aspects of the information society*. Norwood, N.J.: Ablex, pp. 107−120.

R.E. Rice, 1982. "Communication networking in computer conferencing systems: A longitudinal study of group roles and system structure," *Communication Yearbook*, volume 6, pp. 925−944.

K.T. Rosen and M. Resnick, 1980. "The size distribution of cities: An examination of the Pareto law and primacy," *Journal of Urban Economics*, volume 8, number 2, pp. 165−186.

M.A. Smith, 2007. *Netscan: A social accounting search engine*. Redmond, Wash.: Microsoft Research Community Technologies.

M.A. Smith, 1999. "Invisible crowds in cyberspace: Mapping the social structure of Usenet," In: M. A. Smith and P. Kollock (editors). *Communities in cyberspace*. London: Routledge, pp. 194−218.

M. Tepper, 1997. "Usenet communities and the cultural politics of information," In: D. Porter (editor). *Internet culture*. New York: Routledge, pp. 39−54.

M. Van Vugt, S.F. Jepson, C.M. Hart, and D. De Cremer, 2004. "Autocratic leadership in social dilemmas: A threat to group stability," *Journal of Experimental Social Psychology*, volume 40, number 1, pp. 1−13.

S. Wasserman and K. Faust, 1994. *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

B. Wellman and M. Gulia, 1999. "Net surfers don't ride alone: Virtual communities as communities," In: M.A. Smith and P. Kollock (editors). *Communities in cyberspace*. London: Routledge, pp. 167−194.

Editorial history

Received 27 February 2011; accepted 28 February 2011.

The impact of group attributes on communication activity and shared language in online communities
by David A. Huffaker.
First Monday, Volume 16, Number 4 - 4 April 2011
http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/3450/2856