The Ties that Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop Weblogging System

Thomas Lento
Cornell University
323 Uris Hall
Ithaca, NY 14850
thomas.lento@cornell.edu

Howard T. Welser
University of Washington
202 Savery Hall, Box 353340
Seattle WA, 98195
htwiii@gmail.com

Lei Gu
Microsoft News
Redmond, WA
Shanghai, China
legu@microsoft.com

Marc Smith
Microsoft Research
Redmond, WA
masmith@microsoft.com

ABSTRACT

Are people who remain active as webloggers more socially connected to other users? How are the number and nature of social ties related to people's willingness to continue contributing content to a weblog? This study uses longitudinal data taken from Wallop, a weblogging system developed by Microsoft Research, to explore patterns of user activity. In its year long operation Wallop hosted a naturally occurring opportunity for cultural comparison, as it developed a majority Chinese language using population (despite the English language focus of the system). This allows us to consider whether or not language communities have different social network characteristics that vary along different activity levels. Logistic regression models and network visualizations reveal two key findings. The first is that not all ties are equal. Although a count of incoming comments appears to be a significant predictor of retention, it loses its predictive strength when strong ties created by repeated, reciprocal interaction and ties from other dedicated webloggers are considered. Second, the higher rate of retention among the Chinese language users is partly explained by that population's greater ability to draw in participants with pre-existing social ties. We conclude with considerations for weblogs and directions for future research.

Keywords

Weblogs, Social Ties, Social Networks, Retention, Wallop

1. INTRODUCTION

Why do people maintain weblogs? Although existing research [7, 14] provides excellent insight into individual motivations for creating weblogs, such work rarely focuses on changes in behavior over time. Work on time-series data tends to focus on linking patterns between weblogs [10] and the flow of information through these link structures [2]. However, the question of why some people stop updating their weblogs after a short period of time while others remain active for months or even years is not well researched. Is this a phenomenon dependent upon social factors, the behavior of others as much as individual qualities? Do people who receive responses and interact with others remain active for longer than those who do not? Or is it more personal, where some individuals are more internally motivated to continue posting content than others? Comparing social tie patterns with activity over time can answer this question.

Technical developments and the growth of the user base have made weblogs a common medium for social interaction [4], and understanding this increasingly important element of network usage is critical to effective analysis of the medium. Recognizing the importance of social interaction, recent research ranging from studies of political opinions [1] to analyses of the conversational nature of weblogs [8] has focused on links and comments as indicators of interaction between users. However, such research is necessarily limited to analysis of individuals who actively post content, and generalizing these results to the full user population is difficult since the bulk of the content produced in Internet discussion groups [15] and some weblogging systems [6] is contributed by a very small percentage of the user population. A systematic study of the factors contributing to continued user activity will provide an important backdrop to future research on social behavior in weblogs, and could allow researchers to examine questions which were previously intractable due to the difficulty of generalizing results to a broader population. Furthermore, understanding the factors which contribute to continued user activity could allow the creators of new weblogging systems to make more informed design decisions.

This study draws on rich longitudinal data taken from Wallop, a personal publishing and social networking system designed by Microsoft Research, to compare users who remain active with those who do not. The analysis focuses on the relationship between social ties and continued activity as expressed through various features of the system, and tests if people who remain active are more socially connected to other users within the system. An additional dimension of the Wallop user base adds a cultural element to the analysis. Over the life of the Wallop system 65% of the active users were Chinese language users who, as we have shown elsewhere, contribute more content and remain active in the system longer than non-Chinese language users [6]. This suggests some difference between the two groups, which contributes to differences in their levels of continued activity in the system. Is this difference cultural? Is it social? How much of the variance between the two groups is explained by differences in their respective social networks?

2. WEBLOGS AND SOCIAL NETWORKS

As the popularity of weblogs has increased, interaction between authors has become more important to users, developers, and researchers. The popularity of services targeted at promoting user interaction, such as LiveJournal, has led to the creation of alternative weblogging systems which emphasize sharing content with others. Weblogging features have been added to popular social networking sites like MySpace and Friendster, and photo sharing tools like Flickr are increasing in popularity and beginning to draw competition from similar products.[1] Meanwhile, researchers have been busily analyzing the link structure taken from collections of weblogs, searching for clues to the social relationships between weblog authors. Analyses of these network structures have led to excellent studies of deliberation within the medium [1], and the creation of metrics for authority and influence [11]. Another study of the social structure of weblogs indicates that in some cases weblog postings are conversational in nature [8], which may suggest a strong social connection between the participants.

Research examining the nature of social links contained within the weblog medium is rarely focused on explaining individual participation. Yet understanding why people upload content to their weblogs, and why some continue to publish content months or even years later, may be a central element needed for understanding the nature of the social phenomena present within the weblogging medium. Existing research on individual motivations, based on content analysis [7] and interviews [14], suggests that people publish weblogs for many different reasons, including a desire to publish a diary, share knowledge, express a viewpoint, or become part of a community based on a particular shared interest. However, this research only touches on the social connections between the users and how they affect individual decisions to participate.

Outside of the weblog literature, research suggests that social ties are critically important for participation. Recruitment to political and social movements is often dependent on pre-existing social ties [3, 13]. In e-learning settings, continued participation is dependent on social interaction with the professors and other students, even more so than the course contents and materials [12]. Studies of data taken from other online groups show that closely connected groups are more supportive of members [16], and individuals with a strong sense of attachment to a group are more likely to participate [9].

Despite the apparently strong connection between social connections and participation in online and offline groups, empirical research on the relationship between social network structure and individual activity is limited. This lack of empirical research is largely a result of the difficulty of collecting network data. Even in the case of weblog research, where data on the links connecting one weblog to another is both archived and publicly available, the sheer number of weblogs, the variety of different links, and the tendency of users to change hosts makes comprehensive data collection difficult. Furthermore, defining different levels of participation while controlling for the enormous variation in the different types of weblogs is a massive undertaking, making any study of social network attributes and continued participation either exceedingly difficult or dramatically simplified.

This study seeks to address these concerns by using a unique dataset to examine the relationship between social network ties and continued participation in a weblogging system. The dataset, which is taken from the database of activity within the Wallop weblogging and social networking system, includes full information about each users' invitation and entry into the system, activity in the system, and interactions with other users. These activity logs can be used to generate data on the full network of social interactions within the Wallop system, and by combining the activity network data with changes in activity over time it is possible to examine the relationship between social ties and continued participation.

It is important to remember that Wallop is a partially closed unit of weblogging activity bounded by a snowball sample invitation mechanism. Although application of these findings may be applicable to weblogs and other content creation platforms even when they lack Wallop's membership features, the features that enable some aspects of this research simultaneously set the Wallop dataset apart from the wider "blogosphere." Therefore, these findings may be most directly comparable to other similarly bounded social network systems such as Orkut.

2.1 Wallop

Wallop is an online personal publishing and social networking system. An individual gains access to the system when she receives an invitation from an existing Wallop user. It is not possible to navigate to the site and create an account without receiving an invitation.

Figure 1. The Wallop interface, where users can manage their Wallop network.

Once she logs in, the user is presented with an interface which allows her to publish content in the form of text, images, or music posted to her personal weblog. She can interact with other users by making comments on their content, or by viewing, navigating, and adjusting the network of other Wallop users connected to her.

[1] http://www.zooomr.com/

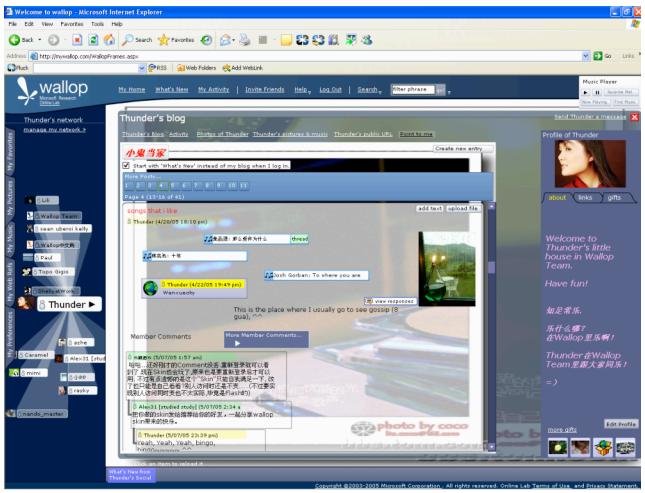

Figure 2. The Wallop interface for uploading content including comments, images and music.

2.2 Research Questions

An interesting phenomenon occurred after Wallop was released to the public in October 2004. Despite having an English-only user interface, restricting access to invited users, and starting with an English-speaking population of Microsoft employees, Wallop attracted an enormous contingent of Chinese language users. The Chinese language users became the dominant group in the system after a period of explosive growth in mid-November 2004, and continued to grow at a rapid rate, outpacing the growth of the non-Chinese language group.

Earlier work [6] shows higher levels of activity among Wallop's Chinese language user population than the non-Chinese language users. Even after adjusting for differences in population size, the Chinese language users send more invitations, have a higher acceptance rate for their invitations, and contribute more content than the non-Chinese language users. Most importantly, much of the variance is explained by the Chinese language users' tendency to remain active in the system longer.

The fact that the Chinese language users remain active in the Wallop system for a longer period of time than the non-Chinese language users raises some interesting questions about the reasons for the variance between the two groups. This study, therefore, focuses on two general research questions:

1. How are social network statistics related to a user's tendency to remain active over a relatively long period of time?

2. How much of the variance between the Chinese and non-Chinese language groups can be explained by differences in social network structure?

3. DATA AND METHODS

Data were taken from the Wallop back-end database, which includes a record of all invitations and uploads for each user, along with date and time information. The database also includes data associating responses or comments with the original object. This makes it possible to reconstruct the network of interactions created by users sending comments to each other. It also makes it possible to reconstruct the invitation tree created by existing Wallop users inviting potential new users to join the system.

Three types of networks, a comment network (4,285 nodes, 14,884 edges), an invitation network (4,514, nodes, 4,255 edges), and a combination network (3,119 nodes, 4,323 edges), were created for this study. Ties in the comment network, which is the primary network used in the analysis, are defined as comments sent from one user to another. The comment network was constructed from data taken in a five week period spanning the entire month of November 2004. This time period was chosen because it was an interesting growth period in the Wallop lifecycle, occurring shortly after the system went public. It also has the advantage of producing networks which, while large, are substantially smaller than networks taken from similar periods later in the life cycle. A five week period was chosen as short enough to capture a relatively brief moment in time, while still being long enough to provide rich data. The invitation network contains all nodes included in the comment network, plus any additional inviters. An invitation (which did not need to be accepted) defines an edge, and each user could only be invited once. The combination network is a subset of the comment network, where edges are defined for any two nodes connected by both a comment and an invitation.

The network data generated from the raw Wallop data was used to generate variables, which were then included in a logistic regression model. The regression model also included individual level variables, such as activity measures and language group affiliation. Finally, the network data was used to generate visualizations, which were used to illustrate general trends in user behavior and show some distinctions between those who remained active and those who did not.

3.1 Network Variables

Except where noted, network variables for the analysis were constructed from the comment network.

In-degree is the number of others who commented on a person's weblog. This variable captures interaction received from an individual. Out-degree was not included because, as a measure of messages sent, it may not be representative of social interaction from the sender's perspective. Receipt of a message is a slightly stronger indication to the recipient that there is a social connection. Reciprocal ties would be the ideal measure, but due to difficulties with the database they were not available at the time of this analysis.

Density is the ratio of the observed number of ties divided by the maximum possible ties [(n-1)*(n-2)]. Density is calculated for each node as the density of a subset of the comment network, consisting of all nodes and edges within a two degree radius of ego. This is a measure of the connectedness of an individual's local network. The more connected an individual's network, the more likely it is that she will feel she is part of a social group, as opposed to an isolated voice.

Betweeness centrality measures the degree to which a node in a graph is a necessary conduit on the shortest path between other nodes. In these data the raw measure for betweeness is highly right skewed. Therefore the natural log of the values is used, which resolves some of the skew. For a formula of betweeness centrality see Brandes [5].

Total degree in combined network measures the number of strong inward and outward ties that a weblogger has, where a strong tie is defined as an edge between a dyad where both an invitation and a comment has occurred. Such ties indicate that webloggers interacted after invitation, suggesting a pre-existing and therefore more

substantial relationship between the nodes. While Wallop users did invite random strangers to join the system, we argue that those that then *also* engaged in commentary with their inviter have stronger ties than those that did not.

Time one alters active at time two reports a count of connections to others in time one who are also active during time two. Ties to dedicated bloggers (those that will stick around) will likely have an impact on an individual's decision to remain active. If social ties are important to activity, and all of the people with whom a person communicates leave the system, then we would expect that person to leave the system as well.

Percent Time One Alters Active at Time Two is a normalized version of time one alters active at time two. The original measure is imperfect, because an individual with a few friends might remain active if all of those friends also remain active. Conversely, an individual with many friends might have a low percentage of her friends remain active, but still have many active alters and therefore stay active herself. Therefore, both measures are necessary.

3.2 Individual Variables

Active at Time 2 is the dependent variable, and is coded as a binary for active or not active. Time 2 represents a 5 week period from January of 2005. Active users are defined as those who posted comments during this time period. Content uploads would be a better measure, but unfortunately this data was not available at the time of this work.

Language Group is a binary variable coded 1 for Chinese Language Group, 0 otherwise. Any user with more than 5% of her comments, or with 3 or more total comments, containing characters from the CJK (Chinese, Japanese, and Korean) Unicode ranges was classified as a Chinese language user. All other users were placed in the non-Chinese language user group. Close examination of the comments posted by a random sample of 200 users from each group suggests that the error rates for these classification systems are quite low. Less than 1% of the Chinese language sample did not include Chinese comments, and all of these were Japanese or Korean.[2] Similarly, less than 3% of the non-Chinese language group were using Chinese characters, and most were communicating in English.

4. ANALYSIS

Table 1 compares the means of structural variables for Active T2 and Non Active T2 webloggers. The network attributes of long lived webloggers are strikingly and significantly different from those who quit. On average, they have more than twice as many people who send them comments, they are more central in the comment network, they have one more person in their strong tie network, and they are disproportionately connected to other bloggers who themselves are long lived. These differences suggest that social ties are positively related to activity within the system, particularly where those ties link to other active users.[3]

Table 1. Means for Structural Variables, Grouped by Retention

	Active T2 *N = 781*	Not Active *N = 3327*	Difference
In Degree	6.2	2.4	3.8***
Betweeness (Logged)	5.1	2.9	2.2***
Density	.16	.23	-.07***
Language Group	.84	.62	.22***
Combined Total Degree	2.6	1.6	1.0***
T1 Alters Active in T2	3.9	1.1	3.9***
% T1 Alters Active in T2	46%	29%	17%***

t-tests ***p < .001

Two of the most striking differences stem from ties from dedicated bloggers. On average, those active at T2 had almost four times as many ties to dedicated bloggers, and within their current ties, dedicated bloggers represented almost half of their ties.

Table 2. Means for Structural Variables, Grouped by Retention

	Chinese *N=2734*	Non Chi. *N=1376*	Difference
In Degree	3.5	2.2	1.3***
Betweeness (Logged)	3.9	2.3	1.5***
Density	.188	.262	-.074***
Combined Total Degree	1.83	1.59	.24**
T1 Alters Active in T2	2.11	.80	1.31***
% T1 Alters Active in T2	37%	24%	13%***
% of T1 Active at T2	*24%*	*9%*	*15%***

t-test **p < .01 ***p < .001

Table 2 reports the same variables while dividing the sample by language group rather than retention. Table 1 showed that 84% of those active at T2 were in the Chinese language group. Therefore variables associated with retention are expected to be higher for the Chinese language group.

Chinese language users send comments to more unique alters (they have higher degree), are more central (higher betweeness), and have more strong ties. They also have more ties to dedicated bloggers and a higher percentage of their ties are to more dedicated bloggers. Finally, their retention rate is more than double the Non-Chinese language group.

Overall, these comparisons of means suggest both that social structure should enhance retention, and that these structural features are more prevalent among the Chinese language users.

[2] We recognize the cultural differences between Japanese, Koreans, and Chinese, but due to the small number of Japanese and Korean users who did not use Chinese we felt justified in accepting these as a minimal classification error.

[3] The one contradictory variable is density, which is *higher* for those who quit. However, this is an artifact of how density is conventionally standardized in network analysis, by essentially taking the square of the number of nodes. This unrealistically penalizes larger networks. We include it here as a warning against using the standard definition of density to study online social networks, especially because many online networks are large.

4.1 Logistic Regression Predicting Retention

The primary analyses employ a logistic regression model to predict the probability of continued participation in the Wallop weblogging system. Table 3 reports three models which allow comparison of how the effects on retention of these social structural variables differ between the full sample, the Non-Chinese, and the Chinese language groups. The model coefficients have been exponentiated and thus reveal how a one unit change in that variable would affect the odds of retention over the odds of quitting. Values greater than one indicate a change that increases the odds of retention, values less than one indicate a decrease in the odds of retention, and values equal to one indicate no change in the odds of retention. In interpreting multivariate regression results it is important to remember that the coefficient for each variable has been calculated after conditioning on the other variables.

The general trends in the three models, combined with the high significance levels for several of the variables, support the notion that social ties can positively affect continued participation.

4.1.1 Model for Combined Sample

The model for the combined sample suggests that bloggers will be more likely to remain active if they are more central (betweeness), have larger networks of strong ties, and are tied to larger numbers and higher percentages of bloggers who themselves remain active.

The coefficient for percent T1 alters active is an especially powerful predictor of continued activity. All other things being equal, the odds of a user remaining active in the system increase 2.6 times when all alters in her network remain active.

The most interesting result of this model is that in-degree is not significant once other variables are taken into account. Although it seems intuitive that the more connections an individual has, the longer she will remain active in the system, it turns out that not all ties are equal. Combined network ties, which consist of both comment and invitation links, represent stronger ties and a higher likelihood of a pre-existing relationship between the users. These types of strong ties, along with ties to other active users, are much more important predictors of retention than simply the number of comments one receives (In Degree).

Given earlier work studying the Wallop system [6], it is not surprising that language group has a strong effect on retention. Even after controlling for other factors, a Chinese language user is more than twice as likely to remain active than a Non-Chinese language user.

4.1.2 Comparison of Models for Chinese and Non Chinese Language User Samples

Comparing the model for Chinese and non-Chinese language users can provide some insight into the differences between the two groups. In general, the effects of the network variables included in this model are similar for both groups. The two striking differences are in the effects of the Total Degree for Combined Network and Active T1 Alters variables.

Table 3. Means of network variables for Wallop webloggers

Variable	Combined	Non Chinese	Chinese
Intercept	.047***	.034***	0.123***
	(1.164)	*(1.339)*	*(1.151)*
In Degree	1.013	.958	0.978
	(1.017)	*(1.049)*	*(1.019)*
Betweeness (Logged)	1.052***	1.041*	1.088*
	(1.014)	*(1.004)*	*(1.016)*
Density	1.59†	1.51135	1.800†
	(1.326)	*(1.809)*	*(1.387)*
Language Group	2.330***	-	-
	(1.114)	-	-
Total Degree,	1.081***	.978	1.105***
Combined Network	*(1.023)*	*(1.054)*	*(1.026)*
T1 Alters Active in T2	1.110***	1.58***	1.113***
	(1.029)	*(1.113)*	*(1.031)*
Percent T1 Alters	2.627***	2.616**	2.233***
Active in T2	*(1.145)*	*(1.41)*	*(1.162)*

†p < .1 *p < .05 **p < .01 ***p < .001

For non-Chinese language users, a larger total degree for combined network is related to a lower level of activity at T2. This result, while surprising, is not significant. The effect of this variable for Chinese language users, by contrast, is strongly significant in the opposite direction. This suggests that the pre-existing relationships and strong connections to other users are more important to the Chinese language users than the non-Chinese language users.

Having a larger number, and a higher percentage, of alters active at T2 is related to a higher level of retention, and is strongly significant for both groups. However, the effects are substantially stronger for non-Chinese language users than Chinese language users. This suggests that connections to dedictated bloggers are more important to the non-Chinese language users, whereas for the Chinese language users these factors are less important, as they are sustained both by these connections, and by their strong ties.

4.1.3 Evaluating model fit

An intuitive way to evaluate a fitted logistic regression model is to compare the predicted probability of the outcome to the observed outcome for all cases. The following plot does just this, by depicting predicted probability against the observed presence of subsequent participation for the combined sample. All cases above the midline were active at time 2 and all cases below the midline were not. Vertical variation within quadrants is simply the result of random jittering introduced to reduce overlap of cases.

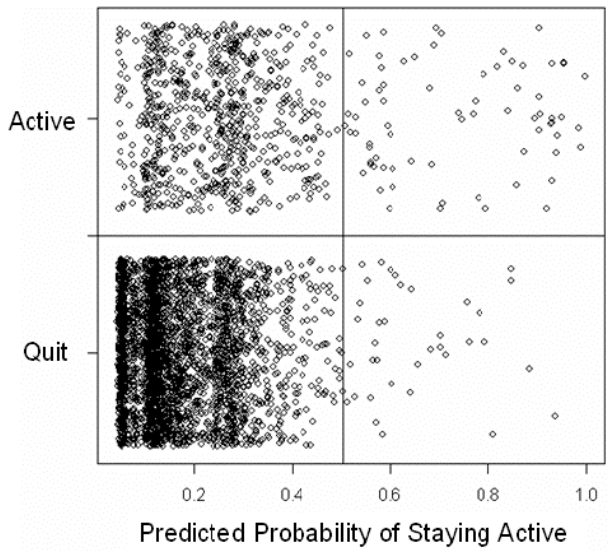

Figure 3. A visual cross-classification of actual and predicted retention

How well did the model used in this analysis predict actual retention? The first notable result is that the vast majority of cases are in the lower left quadrant, meaning they both dropped out, and had social networks unlikely to sustain continued activity. Thus there appears to be a clear relationship between low values in the measured social structural attributes and loss of activity. Furthermore, the webloggers who had high levels of these social structural variables were more likely to stay active. This can be seen by comparing the right hand quadrants to each other. Cases are clearly more numerous in the upper right than the lower right, and as predicted probability increases, the frequency of false positives decreases.

However, the upper left quadrant can be thought of as cases of 'false negatives', i.e. they are results which the model predicted would drop out of the system, yet they actually remained active. This suggests that a large amount of what keeps people involved in weblogging is not included in this model. Therefore (1) we can do a better job of measuring social structural variables, and (2) we can try to measure more individualistic characteristics that might also explain sustained participation. Furthermore, although this analysis provides some insight into how social structure might sustain continued involvement in weblogging systems, there remains a great deal of variation to explain. The next section presents network visualizations, which help provide insight into how structural features might be related to continued participation, and which point to some additional future directions for research.

4.2 Network Visualizations

Network visualizations are used to explore structural relationships in the Wallop system and to generate new intuitions. Here are three visualizations, which highlight the major language group division in the system.

These visualizations are based on three network datasets constructed from three definitions of ties: *invitation*, where ego is tied to alter when one invited the other to join Wallop; *comment*, where ego or alter commented on the other's weblog; and *combined*, where within an invitation relationship there was also the presence of a comment tie. The data includes ties from November 2004, the initial period of rapid increase in the Chinese language users.

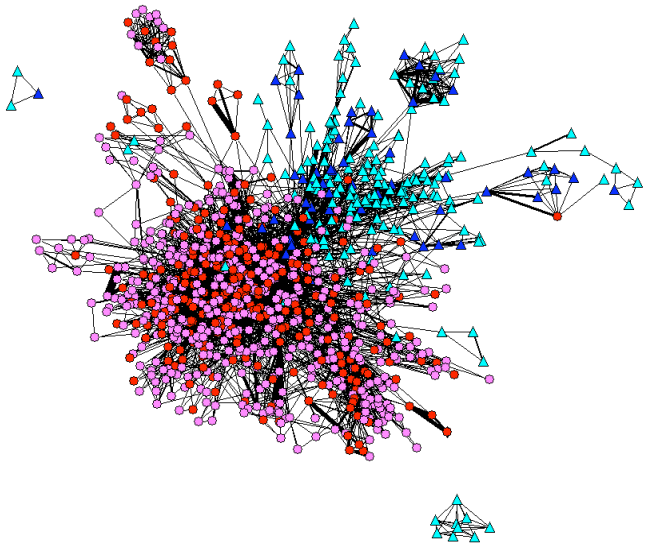

Figure 4. Core of the comment network.

The visualizations in Figures 4, 5, and 6 report on these three datasets, while coding language through color and shape of the node: English (Triangle, Blue), Chinese (Circle, Red), and Uncategorizable[4] (Square, White). Continued activity in Wallop is indicated by depth of shading, with darker hues corresponding to participants in T1(November) who posted a comment during T2 (January).

The comment network, which represents the primary mode of interaction in Wallop, is represented in Figure 4. Since the comment network is large (over 4,000 nodes) this visualization presents a reduced graph that includes only those nodes with at least 5 neighbors. This graph has several interesting attributes: (1) Commenting is highly, but not completely segregated by language group. A notable exception is on the far right. (2) Intense ties (indicating multiple comments within the dyad) are prevalent within the core of the population. This is true for both language groups, although the intense Chinese language core seems larger. (3) There is some evidence that that those who remain active during T2 are centrally located in their local network during T1. This supports the effects of betweenness centrality observed in the regression models.

Figure 5 illustrates a reduced graph of the invitation network (nodes with at least 4 invitees).

[4] Users who did not upload any comments could not be classified as Chinese or non-Chinese language users.

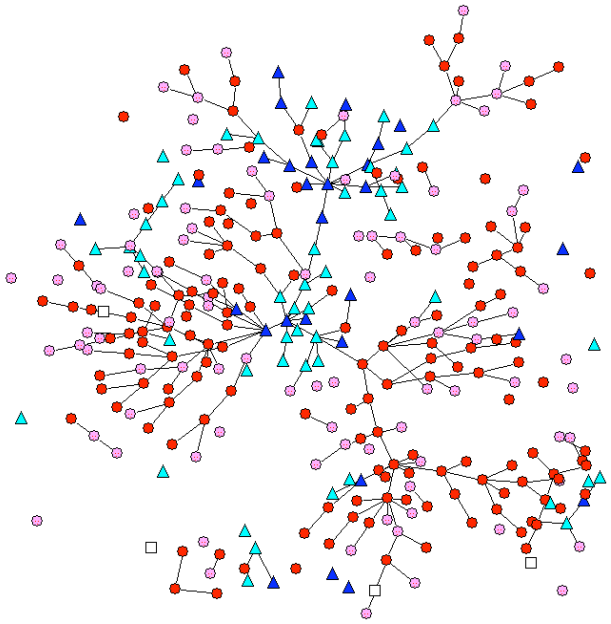

Figure 5. Core of the invitation network.

There are several interesting elements in this graph: (1) There is a similar general tendency towards segregation as in the comment network, although there are several notable exceptions. (2) While there is one major branch that is initially prevalent with Non-Chinese language users, it and all other branches are primarily Chinese language users. (3) People who are branching points within this graph are disproportionately active during T2. (4) The high degree invitation network is composed largely of a single large component.

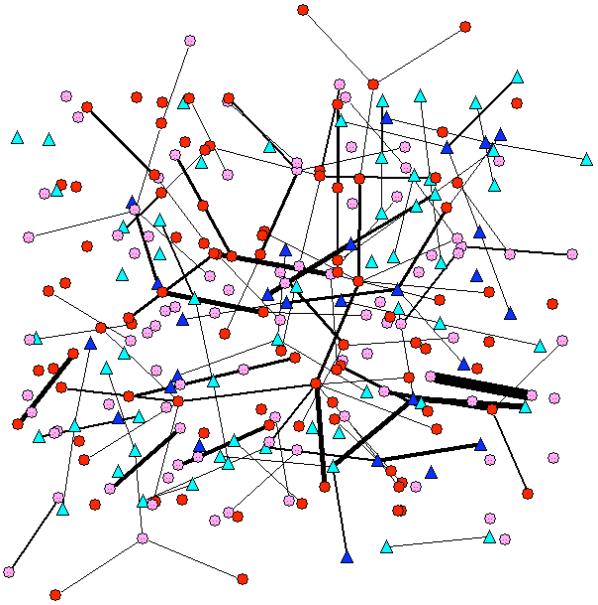

Figure 6. Absence of a core in the combined network.

What does an identical reduction of the combined invitation and comment network show? The graph above, which includes nodes with at least 4 invitees who also commented on or were commented upon by their child, has two interesting characteristics: (1) There is no clear central component. While the reduced combined graph contains nearly the same number of nodes (332 to 267) as the invitation graph, the combined graph is highly fragmented. This result fits with qualitative data that suggests that among Chinese users, invitations would often be made available online to strangers, who would then invite a group of their friends. This would result in fragmented groupings. (2) Branching is less prevalent and ties are more fragile. It seems that the invite and comment relationship was seldom both intense and numerous.

5. DISCUSSION

This study provides some initial data on the relationship between social interaction and an individual's propensity to continue uploading content. Interacting with active alters is a strong and significant predictor of continued user activity. Users who maintain a connection with the person who invited them are also more likely to stay active, which suggests that pre-existing networks are likely an important part of retention. Chinese users had more of these relationships, they were retained more, and the coefficient is strong and significant for them. However, non-Chinese had fewer of these relationships, they were not retained nearly as well, and this coefficient is weak, negative and non-significant for them.

One important point to take away from this study is that the amount of ties does not matter as much as the strength of those ties, and the number of ties to other dedicated bloggers. That is, after controlling for combined ties and ties to bloggers who themselves remain active, in degree had no affect on retention. This has implications for more general studies of weblogs: measuring ties by simply counting them is not enough because not all ties are equally effective indicators of a strong connection, which may itself be an indicator of community. More important than measuring ties is to measure ties that have some underlying reason to affect behavior, in this case, ties that were likely to represent pre-existing relationships and ties to dedicated participants.

A second major point to take away from this study concerns the idea of pre-existing social networks. Organizations that can draw existing social networks into them will build stronger community than those that do not. This is demonstrated by the Chinese Language user group, which was composed of much more durable and active participants. Part of their success, we argue, is that they had a higher prevalence of "strong ties" that combined invitation and commenting, which can be thought of as a proxy indicator of pre-existing social relationships. More important then their prevalence, though, was their influential nature. In the group with greater cohesion, the presence of those strong ties had a strong positive effect on retention. This variable was an important and significant predictor of retention for the Chinese Language group, and not for the non-Chinese language group. The positive effect of pre-existing relationships was suggested by theory and prior research [13], and it is instructive to find evidence of this same mechanism in the context of Weblogs as well. This mechanism can also explain the meteoric growth of MySpace and FaceBook in the US, which are populated predominantly by students who can easily draw in their extensive and intense school-based social networks into the system. Further, weblogs seeking growth need to look to populations within organizations for recruitment, bringing in users who may then pull in their already existing social connections, in favor of individuals outside of institutions who may have less developed connections.

Although this study represents a promising start, it has only scratched the surface of the relationship between network structure and retention in weblogs. More sophisticated network statistics, such as triangle census measures, measures based on reciprocal ties, and a more thorough treatment of data available from analysis of localized egocentric network data for each node might help explain more of the variance in the model. Creating continuous variables which measure decline in user activity would also be beneficial. For example, a typical user might begin with a flurry of activity, which then declines until the user is no longer active in the system. Comparing the social structures around these users to users with dramatically different behavioral patterns might provide further insight into the importance of social ties.

Furthermore, controls for individual behavior are noticeably absent from this analysis. Login, weblog posting, and image uploading behavior might shed some light on user activity patterns. People who rarely log in, even if they are positioned in an active social network, are not as likely to remain active. If possible, controls for demographic features would also enhance this analysis. Unfortunately, due to problems with searching against the database, the data for these variables was not available at the time of this analysis, but future research along these lines should include this data wherever possible.

Despite the limitations of this study, it has brought some interesting findings and intriguing theoretical questions to light. Social interaction does appear to be important to user activity in weblogging systems. Even simple network measures can explain some of the variance between the heavily active Chinese language user group and the rest of the Wallop user population. Further refinement of these models might provide some insight into how people from different cultural backgrounds use weblogging systems. Do the Chinese language users remain active because they are more willing to make social connections through this medium? Or is the higher activity level a result of differences in the demographic distribution of the two populations? Finally, further exploration of the combination network, and the effects of those combination ties, might shed some light on the importance of pre-existing relationships. Do users who actively contribute to weblogging systems do so because they have offline friends there? What other reasons might motivate an individual to contribute? Finally, the visualizations presented above prompt additional questions for further consideration. What is the significance of the centrality of active users? Is this an artifact of their higher activity levels – i.e. more comments and invitations sent tend to move an individual towards the center of the graph – or is it indicative of something else? Could this be related to status within the Wallop system?

6. ACKNOWLEDGMENTS

Our thanks to Wallop Users, the Wallop Development Team, MSR Social Computing and Community Technologies Groups, Shelly Farnham, and Danyel Fisher.

7. REFERENCES

[1] Adamic, L. A. and N. Glance. 2005. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. *Proceedings of the WWW2005 Conference's 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis, and Dynamics*. http://www.hpl.hp.com/research/idl/papers/politicalblogs/AdamicGlanceBlogWWW.pdf

[2] Adar, E. and L. A. Adamic. 2005. Tracking Information Epidemics in Blogspace. *Web Intelligence 2005*.

[3] Anheier, H. 2003. Movement Development and Organizational Networks: The Role of 'Single Members' in the German Nazi Party, 1925-30. *Social Movements and Networks: Relational Approaches to Collective Action*. D. McAdam and M. Diani. New York, Oxford University Press: 49-76.

[4] Blood, R. 2004. How Blogging Software Reshapes the Online Community. *Communications of the ACM*. 47: 53-55.

[5] Brandes, U. 2001. A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology*. 25(2):163-177.

[6] Gu, L., P. Johns, T. M. Lento and M. A. Smith. 2006. How Do Blog Gardens Grow? Language Community Correlates with Network Diffusion and Adoption of Blogging Systems. *Proceedings of the AAAI Symposium on Computational Approaches to Analyzing Weblogs*.

[7] Herring, S., L. A. Scheidt, S. Bonus and E. Wright. 2004. Bridging the Gap: A Genre Analysis of Weblogs. *Hawai'i International Conference on Systems Science*.

[8] Herring, S., I. Kouper, J. C. Paolillo, L. A. Scheidt, M. Tyworth, P. Welsch, E. Wright and N. Yu. 2005. Conversations in the Blogosphere: An Analysis "From the Bottom Up." *Hawai'i International Conference on Systems Sciences*.

[9] Kollock, P. and M. Smith. 1996. Managing the Virtual Commons: Cooperation and Conflict in Computer Communities. *Computer-Mediated Communication*. S. Herring. Amsterdam, John Benjamins.

[10] Kumar, R., J. Novak, P. Raghavan and A. Tomkins. 2003. On the Bursty Evolution of Blogspace. *WWW2003*.

[11] Marlow, C. 2004. Audience, Structure and Authority in the Weblog Community. Presented at the International Communication Association Conference.

[12] Martinez, M. 2003. High Attrition Rates in E-Learning: Challenges, Predictors, and Solutions. *The E-Learning Developer's Journal*. July 14: 9 pages.

[13] McAdam, D. 1988. *Freedom Summer*. New York, Oxford University Press.

[14] Nardi, B. A., D. J. Schiano, M. Gumbrecht and L. Swartz. 2004. Why We Blog. *Communications of the ACM*. 47: 41-49.

[15] Smith, M. 2004. Measures and Maps of Usenet. *From Usenet to Cowebs*. C. Lueg and D. Fisher, Springer.

[16] Wellman, B. 1996. An Electronic Group Is Virtually a Social Network. *Culture of the Internet*. S. Kiesler. Hillsdale, NJ, Lawrence Erlbaum: 179-208.