Structure and Reciprocity in Technology-Centered Q&A Communities

Ming Jiang, Tao Dong, Yung-Ju Chang
School of Information
University of Michigan
Ann Arbor, MI 48109
{mjng, dongtao, yuchang}@umich.edu

Abstract

In this paper we examine the network structure of the MythTV mailing list, an online technology Q&A user community, and we use time-series analysis techniques to study users' reciprocity behavior in this community. We find that the amount of help users provide is strongly correlated to the amount of help they receive. Further, by conducting the Granger Causality test on the time series data of active users' activity, we find that the amount of help given is actually the reason why one gets a lot of help. This finding corresponds to the concept of directed reciprocity in social networks and provides insights into social dynamics in technology-centered online communities.

Introduction

Online question-answering (Q&A) communities have emerged as a popular venue for knowledge sharing (Adamic et al. 2008; Nam, Ackerman, and Adamic 2009). Technology-centered online communities in particular, are the most important resource that provides support for users of open source software or programming languages (Zhang, Ackerman, and Adamic et al. 2007). Various topics about Q&A communities are being discussed, including answer quality (Harper et al. 2008), user motivation (Rafaeli, Raban, and Ravid 2005), effects of user participation (Shah, Oh, and Oh 2008), competition (Yang, Ackerman, and Adamic 2008), and structural characteristics of user knowledge (Adamic et al. 2008).

How is this kind of online communities sustainable? Why do people answer others' questions voluntarily, since they do not get paid and they do not even know each other in real life? One particularly interesting reason is reciprocity, which is defined by sociologists as "a pattern of mutually contingent exchange of gratifications" (Gouldner 1960). Not only is reciprocity applicable in the real life, but in online virtual communities as well (Blanchard and Horan 1998).

Reciprocity in social networks however, is not as simple as it seems. Researchers have further distinguished two kinds of reciprocity: *directed* reciprocity--reciprocity toward some specific people--and *generalized* reciprocity -- reciprocity toward a community as a whole. Leider et al. (2009), based on an online field experiment on Facebook, find that people direct their reciprocity to specific friends in social networks. Jian and MacKie-Mason (2008), by studying P2P networks, find instead that people do not direct their reciprocity toward any specific individual but toward the entire community.

There are several studies on reciprocity in online communities. For example, Sadlon et al. (2008) study user behavior on the website Digg and find that those who submit stories that become popular also actively read and vote for each other's stories. Teng, Lauterbach, and Adamic (2010) study reciprocity behavior in online reputation systems, such as those in Amazon and Epinions, and find that reciprocity plays an important role in user reputation ratings. Those studies, however, do not identify the causal relation behind reciprocity, and therefore fail to distinguish directed reciprocity from generalized reciprocity.

In this paper, we study a technology-centered online Q&A community, the MythTV user mailing list, and try to understand network structure and user reciprocity in this community. Rather than identifying reciprocity based on correlation, we try to infer causality based on time series data. Specifically, we employ the Granger causality test from time-series econometrics to test whether reciprocity in this social network is directed or generalized.

We find that users of the MythTV network engaged in both question-asking and question-answering actively, which indicates strong reciprocity; in addition, we find that reciprocity in this network is directed, which answers one

aspect of motivation of user contribution in this community: I help people today because I might need help from people in the future.

The Dataset

MythTV users communicate mainly through the mythtv-users mailing list[1]. According to Huh, Newman, and Ackerman (2011), the mailing list was devoted to answering technical questions about troubleshooting and tailoring the MythTV system, and because we observed that there was very little casual chatting and general discussions, it is our assumption that an initial message is a question and replies to that message are potential answers.

We crawled the online archive of the mailing list consisting all of the 32326 messages posted between October 13, 2009 and November 4, 2010, and stored the data in a database. Table 1 shows the data fields each message was parsed into.

msg_id	The identifier of the message
reply_id	The identifier of the initial message that this message replies to. If this message is an initial message, then the reply_id is the same as the msg_id.
author_name	The name of the message author
subject	The subject of the message
body	The content of the message
timestamp	The date and time the message is posted

Table 1: Data fields of a parsed message

The dataset contains 1650 unique usernames, and 5298 threads of email messages. The daily traffic of the mailing list in our dataset is 83.5 messages and 13.7 threads in average.

Following the concept of the community expertise network proposed by Zhang, Ackerman, and Adamic (2007), we constructed a directed network of users. Specifically, User A receives an *indegree* from User B when User A replies to User B's initial message (assumed as a question). This type of network not only captures the interactions between users but also allows higher-level analysis of community structure and reciprocity behavior.

In order to do time-series analysis, we took a cross-sectional sample of our data every two weeks in a cumulative manner, and then generated 28 networks for each time period. In our data analysis, the Bowtie Structure analysis was based on the network of sampled in the latest time period, while the reciprocity analysis examined all 28 networks as a time series.

[1] http://www.gossamer-threads.com/lists/mythtv/users/

Data Analysis

Network Characteristics

The Mythtv-users mailing list is functionally similar to other technical forums such as the Java Forum studied in Zhang, Ackerman, and Adamic (2007). However, our bowtie analysis of the Mythtv expertise network revealed an interesting structural difference from the Java Forum where most of the participants only posted questions, while a much smaller fraction of participants were responsible for answers.

Bowtie structure analysis, first used to study the structure of the Web (Broder, et al. 2000), is useful to illustrate the structure and connectivity of a directed network. A bowtie structure divides a network into 6 components: 1) the strongly connected core (SCC) in which every node can reach every other node, 2) the *IN* component in which nodes only have one-way links to the nodes in the core, 3) the *OUT* component in which nodes only have one-way links from the core, 4) tendrils which are nodes connected to either nodes in the *IN* component or the *OUT* component, but not to those in the core, 5) tubes which are nodes connecting the *IN* component and the *OUT* component but do not belong to the core, and 6) disconnected nodes.

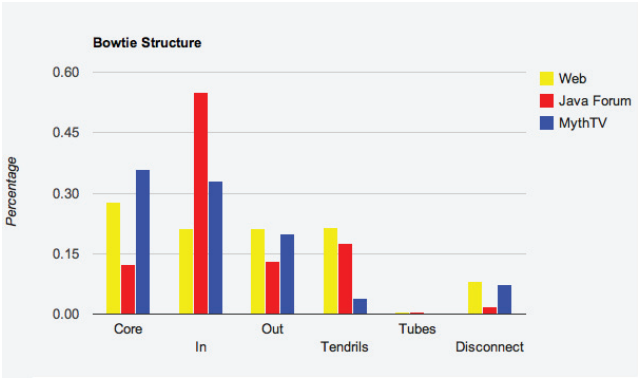

Figure 1: Bowtie structure of MythTV network, compared with Java Forum and the Internet.

Figure 1 compares the bowtie structure of the MythTV network with that of the Java Forum network (as reported in Zhang, Ackerman, and Adamic 2007) and of the Web (as reported in Broder, et al. 2000). We have several interesting observations from these results. First, the MythTV network has a much larger Strongly Connected Core (35.8% of the nodes) than that of the Java Forum (12.3%) as well as the Web (27.7%). This result indicates that, as compared to the Java Forum, there are remarkably more users in the MythTV network who both ask and answer questions. Secondly, although the category of "users who only ask questions (nodes in the IN component)" is about 1.7 times more than the category of "those who only an-

swer questions (nodes in the OUT component)" in the MythTV network, the difference is actually a lot smaller than that in the Java Forum. Moreover, the IN component of the MythTV network is much smaller than that of the Java Forum, suggesting that the proportion of pure question-askers is much smaller than that in the Java Forum.

The above results of bowtie structure analysis imply that, in comparison to functionally similar technical Q&A forums, MythTV users were more engaged with each other in the community, and notably fewer users managed or chose to only receive help from the community without making any contribution. In the next section, we further investigate this phenomenon from the perspective of reciprocity.

Reciprocity

In our MythTV network, *indegree* and *outdegree* of the nodes in the network are defined as the number of unique people one person helps (i.e., replies/answers) in their email threads, and the number of unique people from whom one person receives help in all of his or her threads by our dataset construction.

In this network, we find that *indegree* and *outdegree* of users are strongly correlated (Pearson correlation, $r = 0.5036$, $p < 0.001$). We interpret this to mean that users who answer more questions get more answers, based on the Q&A-exclusive nature of this network.

We do not know, however, the direction at which one factor causes the other yet from this correlation. There are two possible explanations. If high *indegree* causes high *outdegree*, this corresponds to directed reciprocity (Leider et al. 2009): I have helped many people in the past, they have either received my help, or observed my effort, and now they are directing their efforts toward helping me when I need assistance (or from an alternative, but equivalent perspective: some people helped me before so in response, I am helping those specific people); on the other hand, if high *outdegree* causes high *indegree*, this corresponds to generalized reciprocity (Jian and MacKie-Mason 2008): many people helped me before, so I'm now helping a lot of other, non-specific, people.

In order to find out the causal relation between *indegree* and *outdegree*, we employed a one time-series statistical method widely used in empirical macroeconomics, called Granger causality (Granger 1969). It has also been used in areas outside of economics, such as neural science (Kamiński et al. 2001), and psychology (Bressler et al. 2008). The Granger causality test is a statistical hypothesis test for determining whether one time series is able to forecast another. A time series X is said to Granger-cause Y if it can be shown, usually through a series of t-tests and F-tests on lagged values of X, with lagged values of Y also included, that those X values provide statistically significant information about future values of Y. And consider here that X and Y refer to *indegree* and *outdegree* in our network.

Mathematically, to test against the null hypothesis that X does not Granger-cause Y, we first construct a univariate vector auto-regression (VAR) of variable Y (the regression of Y based on its lagged value):

$$y_t = a_0 + a_1 y_{t-1} + \ldots + a_m y_{t-m} + \varepsilon_t$$

whereas only y with significant t-statistics is retained in this regression, m is the maximum time lag, and ε is the random error.

Next, we extend this equation with lagged value of X:

$$y_t = a_0 + a_1 y_{t-1} + \ldots + a_m y_{t-m} + b_p x_{t-p} + \ldots + b_q x_{t-q} + \varepsilon_t$$

whereas only y with significant t-statistics is retained in this regression, provided that collectively they add explanatory power to the regression according to an F-test. p is the shortest, and q is the longest lag length for which the lagged value of x is significant. The null hypothesis is accepted if and only if, no lagged values of x are retained in the regression.

To utilize this method, we used the time series data generated from the MythTV network. We took a cross-sectional sample of our network every two weeks, and thus we had 28 networks for each time period. Then we calculated the *indegree* and *outdegree* for each user in each network. For users who had not yet registered the email list by each time point, we treated their degrees as zeros. Thus we had a panel data for all users over 28 time periods.

In our time series model we used the *indegree* and *outdegree* data of the first 50 most active users (with rankings based on the sum of their *indegree* and *outdegree*), since, according to the degree distribution, the top 10% most active users contributed more than 90% of the total content. We then took the average of those users' *indegrees* and *outdegrees*, since the Granger test required time series *vectors*, and we were more interested in the community aggregate than in individual behavior. In addition, we assumed that the response time for reciprocity behavior was a month; therefore we set the time lag parameter to 2.

Equation	Excluded	χ^2	p-value
Indegree	*Outdegree*	1.4182	0.234
Outdegree	*Indegree*	7.8153	0.005

Table 2: Granger Causality test for *indegree* and *outdegree*

We found through the Granger Causality test results (see Table 2) that the null hypothesis that Indegree does not Granger-causes Outdegree is rejected ($p = 0.005$), while the null hypothesis that Outdegree does not cause Granger-causes Indegree cannot be rejected ($p = 0.234$). Therefore we drew the conclusion that the reciprocity in the MythTV

network is directed: generosity towards others earns generosity from them.

Conclusion

In this paper we study the network characteristics of a technology-centered online Q&A community (the MythTV mailing list). We find that the MythTV network is more connected and reciprocal than some technology-centered communities (for example, Java Forum). Also, we show, by looking at the correlation of *indegree* and *outdegree,* that reciprocity behavior is prevalent in the network. Finally, by using time-series analysis techniques, we are able to show that reciprocity in this community is actually more directed: people tend to direct their help to those who helped them before.

Our study has important implications for understanding social dynamics in online communities, especially technology-centered ones. Reciprocity is indeed the driving force behind user contribution and cooperation. The reason why users answer questions initially is that they can expect to get help from other people later on. This explains the existence and sustainability of such communities.

Since our research is a preliminary study of applying time-series analysis techniques in economics to social networks, it has a few limitations. First, we only look at the reciprocity behavior in aggregate since we take the average *indegree* and *outdegree* for most active users; we do not study the behavior patterns for individual users. Second, drawing conclusions about reciprocity based on the Granger test, though better than simple correlation or regression, can still be questioned. If both *indegree* and *outdegree* are driven by a common third process with a different time lag, one might still accept the alternative hypothesis of Granger causality. Moreover, due to the community we chose, we were only able to access data from a limited period of time (1 and half years); different conclusion might be drawn if the data set spanned 3 or more years.

Further work could be done to extend the study of user behavior to other kinds of online Q&A forums and mailing lists to see if the reciprocity pattern is robust across different communities, and identify the community traits that foster reciprocity. In addition, text analysis of email transaction content could be done to provide qualitative evidence for the validity of our conclusion.

References

Adamic, L. A.; Zhang, J.; Bakshy, E.; and Ackerman, M. S. 2008. Knowledge Sharing and Yahoo Answers: Everyone Knows Something. In *Proceeding of the 17th International Conference on World Wide Web*, 665–674. New York, NY, USA: ACM.

Blanchard, A. L.; and Horan, T. 1998. Social Capital and Virtual Communities. *Social Science Computer Review* 16(3): 293-307

Bressler, S., Tang, W., Sylvester, C., Shulman, G., and Corbetta, M. Top-Down Control of Human Visual Cortex by Frontal and

Parietal Cortex in Anticipatory Visual Spatial Attention. Journal of Neuroscience, 2008; 28 (40): 10056-10061.

Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A.; et al. 2000. Graph Structure in the Web. *Computer Networks* 33(1-6): 309-320.

Gouldner, A. W. 1960. The Norm of Reciprocity: A Preliminary Statement. *American Sociological Review* 25(2):161-178.

Granger, C. W. 1969. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica* 37 (3) : 424–438.

Harper, F. M.; Raban, D.; Rafaeli, S.; and Konstan, J. A. 2008. Predictors of Answer Quality in Online Q&A Sites. In *CHI '08: Proceeding of the Twenty-sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, 865–874. New York, NY, USA: ACM.

Huh, J.; Newman, M. W.; and Ackerman, M. S. 2011. Supporting Collaborative Help for Individualized Use. To appear in *Proceedings of the Twenty-ninth Annual SIGCHI Conference on Human Factors in Computing Systems*. Vancouver, Canada: ACM.

Jian, L.; and MacKie-Mason, J. K. 2008. Why Share in Peer-to-Peer Networks?. In *Proceedings of 10th International Conference on Electronic Commerce (ICEC) '08,* 4:1--4:8. Innsbruck, Austria: ACM.

Kamiński, M., Ding, M., Truccolo, W. A., and Bressler, S. L. 2001. Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological Cybernetics* 85(2): 145-157.

Leider, S., Möbius, M. M., Rosenblat, T., and Do, Q. 2009. Directed Altruism and Enforced Reciprocity in Social Networks. *Quarterly Journal of Economics*, 124(4):1815-1851

Nam, K., Ackerman, M. S., and Adamic, L. 2009. Questions in, knowledge in?: a study of naver's question answering community. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, pp. 779-788.

Rafaeli, S., Raban, D. R., and Ravid, G. 2005. Social and economic incentives in Google Answers. In ACM Group 2005 Conference, ACM.

Sadlon, E., Sakamoto, Y., Dever, H. J., Nickerson, J. V. 2008. The Karma of Digg: Reciprocity in Online Social Networks. In *Proceedings of the 18th Annual Workshop on Information Technologies and Systems*.

Shah, C., Oh, J. S. & Oh, S. 2008. Exploring characteristics and effects of user participation in online Q&A sites. *First Monday*, 13 (9).

Teng, C., Lauterbach, D., and Adamic L. 2010. I rate you. You rate me. Should we do so publicly? 3rd Workshop on Online Social Networks, Boston, MA.

Yang, J., Adamic, L. and Ackerman, M. S. 2008. Competing to Share Expertise: The Taskcn Knowledge Sharing Community. In *Proceedings of The Second International Conference on Weblogs and Social Media,* 161-168. Menlo Park, Calif.: AAAI Press.

Zhang, J., Ackerman, M. S., and Adamic, L. 2007. Expertise Networks in Online Communities: Structure and Algorithms. In *Proceedings of the 16th International Conference on World Wide Web*, 221-230. New York, NY, USA: ACM.