T. Koschmann, D. Suthers, & T.W. Chan (Eds.), *Computer Supported Collaborative Learning 2005: The Next 10 Years!*
Mahwah, NJ: Lawrence Erlbaum Associates.

Macroscopic Study of the Social Networks Formed in Web-based Discussion Forums

Yau-Yuen Yeung
Department of Science
Hong Kong Institute of Education
yyyeung@ied.edu.hk

Abstract. It is proposed and outlined in this paper on how to investigate the macroscopic features of those large-size social networks formed in web-based discussion forums. Some preliminary results on the pattern of the distribution of replies for individual topics, views for individual topics and co-discussants of individual participants will be presented. The present results will be compared with those found in other areas of large-size social networks and the significances and future work on those macroscopic properties of social networks for better understanding of computer-supported collaborative learning will be discussed.

Keywords: Social Network Analysis, Web-based Discussion Forum, Computer-Mediated Communication

INTRODUCTION

Since mid1990s, social network theory (see, e.g. Wasserman and Faust, 1994; Degenne and Forse, 1999; and Batten, Casti and Thord, 1995) has been employed by a number of researchers to analyze students' interaction and learning in various computer-supported collaborative learning (CSCL) environment as facilitated by different kinds of computer-mediated communication systems in several university programs. Although most of those prior network analyses (see, e.g. Haythornthwaite, 1998; Palonen and Hakkarainen 2000; Cho, Stefanone and Gay, 2002; Martinez et al., 2003; and Aviv, Erlich and Ravid, 2003) were very intensive, were established in varied settings and were often coupled with qualitative data as collected from face-to-face interviews for triangulation of their results, yet one serious limitation of their work comes from the fact that the size of their networks under investigation is usually very small (ranging from a dozen to a hundred only). This shortcoming will not only cast doubt on the accuracy and general validity of their results but also prevent them from studying many significantly important macroscopic features of social networks such as identification of the network type, distributions of various network properties, formation of giant cluster and percolation phenomena etc which have attracted extensive research interest by many researchers on different types of social networks (e.g. the Internet, the World Wide Web, phone call networks, citation networks, research collaboration networks, country roadmaps, airline routing networks, electricity transmission networks, spread of AIDS and the food web) in the last few years (see, e.g. Albert and Barabasi, 2002; Buchanan, 2002 and Watts, 2003).

There are several key channels such as web-based/online discussion forum, email, ICQ/chat room, NetMeeting (or equivalent GnomeMeeting in Linux/Unix systems), phone call, scheduled/unscheduled face-to-face discussion which are usually found or adopted in many CSCL environments. Online discussion forum is specifically chosen for the present study because by default it can record almost all the participants' communication information and the messages themselves can readily be retrieved for content analysis without additional efforts for hardware or software modification. Furthermore, Haythornthwaite's (1998) study on the growth of community among distance learners revealed that web-board was the most popular (with nearly 100% usage) channel of communication adopted by those distance learners enrolled for a master degree in his university.

IDEAS AND METHODOLOGY

First of all, we can construct two kinds of social network from a web-based discussion forum. It is known that the mathematical description of a network is a graph (Wasserman and Faust, 1994; and Yeung, Liu and Ng, 2005) which is denoted by $G(\mathbf{N}, \mathbf{R})$, where **N** is the set of nodes or actors and **R** is the set of relationships or links between these nodes. In a given discussion forum, there is a set of participants **P** who post the set of messages **M** in it. Hence, an obvious network for the discussion forum can be formed by taking **P** and **M**

altogether as the node set, i.e. **N** = {**P**, **M**}. The relationship set **S** is imply "who submits/posts that message" and this links up individual elements in **P** with one or more elements in **M**, i.e. a one-to-many mapping from P to M because one participant can post many messages in the discussion forum. It is noted that we exclude multi-authorship by treating the one who posts the message as the sole node in our network but unlike research publications, multi-authored messages are rather rare in discussion forum. Let us call it the *basic network* G_B({**P**, **M**}, **S**) and it is obviously a kind of bipartite graph (Wasserman and Faust, 1994) in which there is no link or relationship given to relate elements within the set **P** or within the set **M**. Can we have an objective way to provide the linkages amongst certain elements within the set **P**? And what are the significances or implications of studying those linkages for the understanding of CSCL?

To answer the questions raised in the last paragraph, we shall borrow the idea of research collaboration from Newman (2001) who has constructed several very large-size (up to 1.5 million nodes) collaboration networks of researchers in various major fields of science by identifying two researchers as "socially linked" if they have published at least one paper together. In an online discussion forum, there is a topic starter who raises a question, announces a message or expresses his/her view on a certain issue and this forum message may be subsequently replied by one or more forum participants. Those followers and the original topic starter can all be treated as "socially linked" because they have mutually exchanged ideas, shared information or learnt from the peers (there are of course some rare cases that the topic starter never returns to view the follow-up messages). Therefore, a second kind of network called the *collaborative learning network* G_C(**P**, $\mathbf{M}_S$) can be constructed by taking the forum participants **P** as the set of nodes and the relationship set consists of all the submission of messages $\mathbf{M}_S$ to individual topics for indicating the co-discussion of a particular topic by various participants.

Based on the afore-mentioned conceptual framework and the usual social network theory, the following procedures have been adopted to construct both the basic network G_B and the collaborative learning network G_C for groups of online discussion forums:

1. Retrieve the forum participant's name (maybe nickname) from every message of a chosen forum and put all names for a particular topic in the same line (separated by a certain delimiter) to implicitly denote their relationships. A special Windows-based program called the "HAS Centre Browser" (available from the author) has been developed to provide the capability to complete this task automatically while the researcher uses it to browse the online forum. Some other relevant information such as the number of views for a given topic can also be retrieved by this program.
2. Participants' names from all topics of one or a group of forums are combined into a single file, sorted and duplicates eliminated and re-labeled with unique sequential numbers. This step can be done by using the MS Excel program.
3. Participant names in the original computer files for Step 1 above are then converted into the unique number labels as given in the Step 2. A small program has been developed to accomplish this task so that the data will be given in a format suitable for input by other social network analysis computer programs.
4. Two powerful and well-known shareware/freeware programs called UCINET (Borgatti, Everett and Freeman, 2002) and PAJEK (Batagelj and Mrvar 1999) for social network analysis can be used to extract various network statistics and draw the corresponding network graphs for providing a macroscopic view of the complex networks.

As an initial study, a public website called Linux Forum (http://www.linuxform.com/forums/) is chosen because the forums inside are very well-documented and are provided with many useful statistics. Unlike the Microsoft Windows, Linux is an open-source operating system which has rather little official technical supports and so peer support and collaborative learning are very essential in building up the knowledge base. On the other hand, authors need to register with a unique login name before they are allowed to post messages in any forum and this can effectively eliminate the problem of misidentification of forum participants. There are some moderators present to keep the forum discussions evolving in a proper manner and the content analysis of some randomly selected messages inside many forums of this website reveals that most forum participants are very professional in attitude and most of their messages do contain meaningful and useful knowledge and experiences for peer sharing or collaborative learning.

RESULTS AND DISCUSSION

For the Linux Forum website mentioned above, the HAS Centre Browser was used to retrieve messages from 36 forums (grouped under 6 main areas) during the period of the first two weeks in Nov., 2004. As extracted from all those messages, there were in total 24,384 topics which were followed by 51,724 replies and 53,070 counts of participant names. Four areas of forums, namely Linux Forum, Miscellaneous, Linux News Discussions and Rants were excluded from our study because their themes are either not directly related to the collaborative learning or sharing of knowledge on the Linux systems or they are read-only archives copied from other Usenet groups.

Figure 1 shows a semi-log pilot for the frequency distribution of the number of replies for individual topics posted in all the 36 forums under study. The greatest number of follow-up replies for a particular topic is 148. Since most topics were followed by 30 or less replies, the sparsely distributed data for higher number of replies were truncated from Figure 1. As most data nearly fall along a straight line, this set of data is very likely to follow the *exponential distribution* which is commonly found in the medical and industrial engineering field. For instance, the life span of a person or of an engineering product follows this kind of "memoryless" right-skewed distribution whose probability density function has the *form* $p(k) = \lambda \exp(-\lambda k)$ where the parameter λ = 0.266 yields the best fitted straight line in Figure 1. This result implies that a given topic started in any forum is most likely unreplied at all and on average there are $1/\lambda = 3.76$ replies. In research collaboration networks, the degree of co-authorship (number of authors per paper) and the degree collaboration (number of collaborators per author) also follow this exponential distribution (Yeung, Liu and Ng, 2005).

Since there are many people who simply want to find answers (or passively learn without sharing or mutual communication) from other forum participants' conversation, Figure 2 shows the frequency distribution of views for all messages posted around a given topic. After some critical examination, it is discovered that those data actually follow a *lognormal distribution* with the form $p(k) = \frac{1}{k \cdot S\sqrt{2\pi}} \exp(-\frac{(\ln \cdot k - M)^2}{2S^2})$ where M = 4.3 and S = 0.92 yield the best fitted "bell-shape" curve. This is a new kind of network distribution which has not been identified by any previous researchers in most well-known large-size networks even though lognormal distribution is quite commonly used to describe various biological, social and economic phenomena. This result means that for each group of messages posted around a given topic of this Linux Forum website, there were in total most likely viewed ($\exp(M+S^2/2)$ =) 113 times or in average ($\exp(M-S^2)$ =) 32 times by other people. However, we must be aware that the number of views per topic will naturally grow with time and be cautious in our interpretation that one or more search engines (e.g. Google) are scanning through all messages at regular time intervals and their activities are recorded in the view count of this forum, yielding a same rate of growth for the view count in every messages posted. Further investigation will be required to uncover and explain for the occurrence of this kind of distribution.

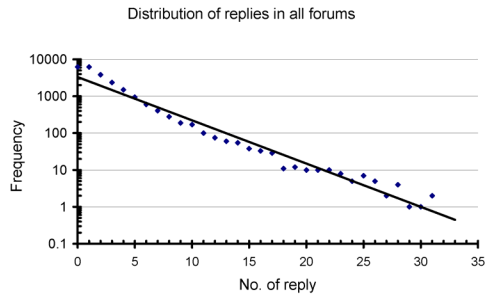

Figure 1: Frequency distribution of follow-up replies for a given topic posted.

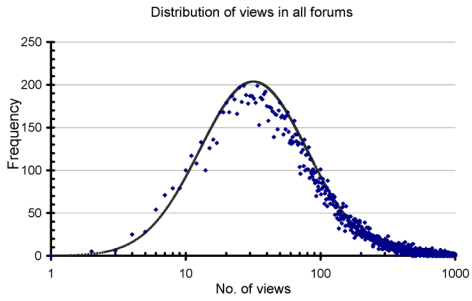

Figure 2: Frequency distribution of views for messages posted around a given topic.

To study the interaction between forum participants, we need to employ the UCINET and PAJEK software to carry out the very time-consuming and memory intensive computation. As an initial analysis, a smaller collaborative learning network G_C was constructed for 4 forums in the General Linux Issues area which contains 9,327 topics with 3,214 different participants in total. Figure 3 shows a log-log plot for the frequency distribution of co-discussants of individual participants. Co-discussants are defined to be those who have ever posted messages for a certain topic. These data could reveal many useful characteristics such as centrality, social roles, cohesion, and cluster formation of a social network. For this short paper, we just present one key result – the data follows a power-law form distribution $p(k) \sim k^{-n}$ where the power-law exponent n = 1.20 for the best fitted straight line given in Figure 3. Power-law form distribution is a characteristic form of the so-called "scale-free" network (Albert and Barabasi, 2002) which is commonly found in many other kinds of social networks and it is postulated to come from the "rich get richer" phenomenon or a combined effect of "growth" and "preferential attachment". For examples, Newman (2001) and Yeung, Liu and Ng (2005) also got the power-law form for the productivity (or number of papers per authors) in the research collaboration networks and got the values of the exponent n = 2.1, 3.41 and 2.86 for the physics teaching, computer science and medicine networks, respectively. In the present network, the most active participant has 633 co-discussants while 23% of

participants have no co-discussant (as all topics started by them were not replied by anyone else at all and they did not participate in other topic starters' discussion). Another 29% of the participants have just only one co-discussant. These results are comparable with the result for physics teaching networks in which 32% of the authors have no collaborator at all (Yeung Liu and Ng, 2005). Figure 4 reveals a global picture of the network concerned in which each participant is represented by a dot and links are used to join up co-discussants. Isolated dots have been intentionally moved to the circumference of the figure for the ease of identification.

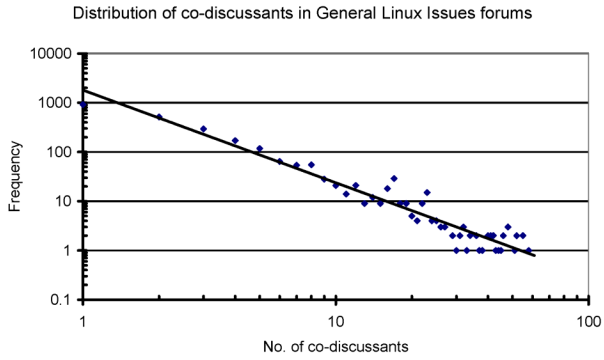

Figure 3: Frequency distribution of co-discussants

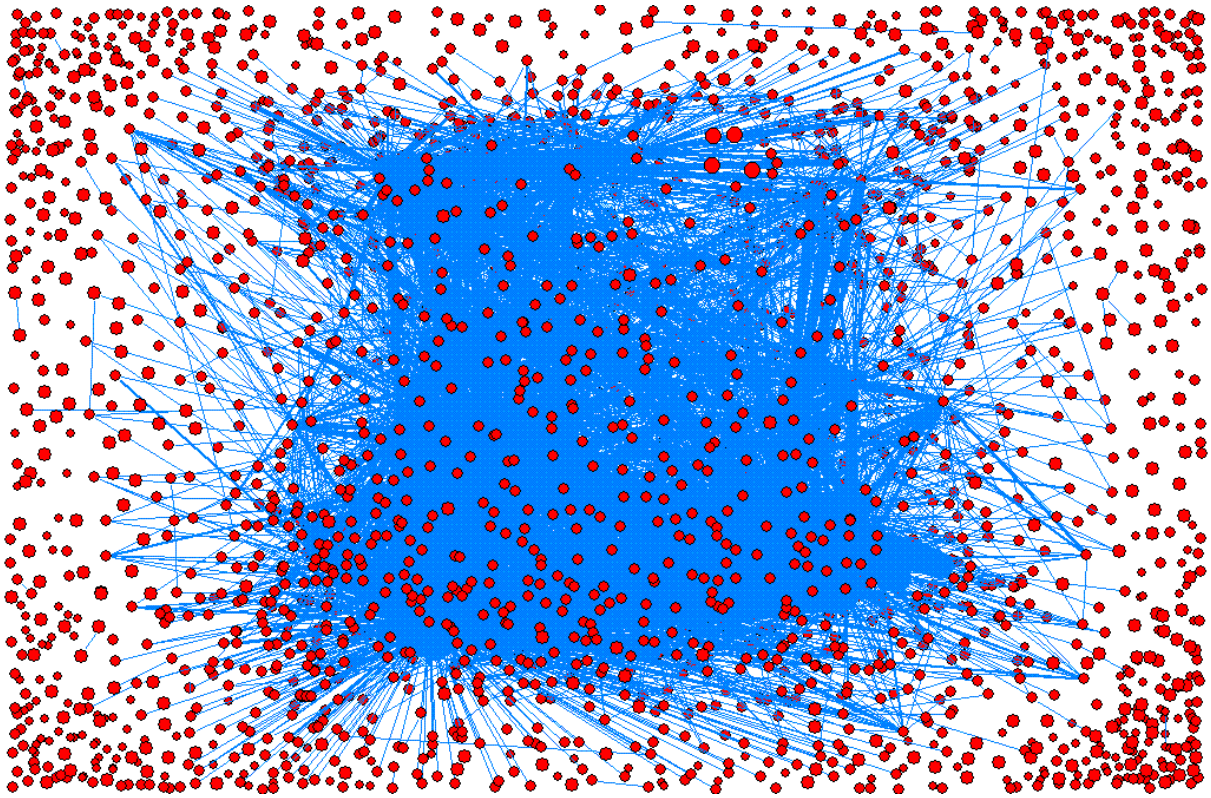

Figure 4: The collaborative learning network of 3,214 participants (denoted by dots with straight line for co-discussion linkage) in the 4 forums of the General Linux Issues area

CONCLUSIONS

The rationales and importance of studying large-size CSCL networks have been introduced and discussed. A workable framework was outlined together with a concrete example on how to analyze some online discussion forums in a chosen public website. Some new and interesting results for large-size CSCL networks were obtained. In particular, it was found that the number of replies for a given topic follows an exponential distribution, the number of views follows a log normal distribution and the number of co-discussants follows a power-law distribution, revealing that it is a kind of "scale free" network. All results deviate significantly from the Poisson distribution which is a typical characteristic of a *random network* (Albert and Barabasi, 2002).

While this short paper opens a new direction of research on CSCL networks, there are still much more work to be carried out in future. In particular, we need to develop computer programs to efficiently extract important network features from very large-size networks. We could also study the time evolution of those networks by retrieving forum data at regular time intervals (say, every 3 months). Finally, we need to study forums in many other types of websites to confirm if those network characteristics are actually universal in nature. Those macroscopic results for large-size CSCL networks will certainly help us obtain a global understanding of the sorts of interaction and sharing between learners and could potentially be applied to design and implement some small/medium size CSCL networks in a better way.

ACKNOWLEDGMENTS

Financial support from The Hong Kong Institute of Education is gratefully acknowledged.

REFERENCES

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

Aviv, R., Erlich, Z., and Ravid, G. (2003) Network Analysis of Cooperative Learning. *Proc. Information Communication Technologies in Education (ICICTE) 2003, Samos, Greece, July 2003.*

Batagelj, V. and Mrvar, A. (1999) *Pajek – Program for Large Network Analysis*. [Online] http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

Batten, D., Casti, J., and Thord, R. (1995) *Networks in Action*. Berlin: Springer-Verlag.

Borgatti, S.P., Everett, M.G. and Freeman, L.C. (2002) *Ucinet for Windows: Software for Social Network Analysis*. Harvard: Analytic Technologies.

Buchanan, M. (2002) *Nexus : small worlds and the groundbreaking science of networks***.** New York : W.W. Norton.

Cho, H., Stefanone, M. and Gay, G. (2002). Social network analysis of information sharing networks in a CSCL community. In *Proceedings of the Computer-Support for Collaborative Learning* (CSCL) 2002 Conference, G. Stahl (Eds.), Jan. 7-11, Boulder Colorado, NJ: Lawrence Erlbaum Associates, 43-50.

Degenne, A, and M. Forse, M. (1999) *Introducing Social Networks.* (translated by A.Borges) London: SAGE Publications.

Haythornthwaite, C. (1998). A social network study of the growth of community among distance learners, *Information Research,* **4** (1). http://informationr.net/ir/4-1/paper49.html.

Martinez, A., Dimitriadis, Y., Rubia, B., Goomez, E., and de la Fuente, P. (2003) Combining qualitative evaluation and social network analysis for the study of classroom social interactions. *Computers & Education,* **41**, 353–368.

Newman, M.E.J. (2001) Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review* **E64**, 016131-8.

Palonen, T., and Hakkarainen, K. (2000) Patterns of Interaction in Computer-supported Learning: A Social Network Analysis. In B. Fishman and S. O'Connor-Divelbiss (Eds.), *Proceedings of the Fourth International Conference of the Learning Sciences* (pp. 334-339). Mahwah, NJ: Erlbaum.

Watts, D.J. (2003) *Six Degrees: The science of a connected age.* New York: W.W. Norton.

Wasserman, S. & Faust, K. (1994). *Social Network Analysis: methods and applications.* Cambridge: Cambridge University Press.

Yeung, Y.Y., Liu, T. C.Y. & Ng, P.H. (2005). A social network analysis of research collaboration in physics education. *American Journal of Physics,* **73**(2)*,* 145-150.