Topological Analysis and Measurements of an Online Chinese Student Social Network

Duoyong Sun[1], Jiang Wu[2], Shenghua Zheng[3], Bin Hu[2], and Kathleen M. Carley[4]

[1] College of Information System and Management, National University of Defense Technology, Changsha, Hunan, China, 410073
`duoyongsun@gmail.com`
[2] Huazhong University of Science and Technology
1037, Luoyu Road, Wuhan, China, 430074
`jiangwu.john@gmail.com, bin_hu@mail.hust.edu.cn`
[3] Zhejiang University of Technology
18 Chaowang Road, Hangzhou, China, 310032
`zheng0210@163.com`
[4] CASOS, ISRI, Carnegie Mellon University
5000, Forbes Avenue, Pittsburgh, USA, 15213
`kathleen.carley@cs.cmu.edu`

Abstract. Online social network attracts more researchers now. In this paper, we topologically analyze an online Chinese student social network--Xiaonei.com. We use Python language to crawl two datasets of Xiaonei in January and February, 2008. The degree distribution and small world phenomena are testified. We also use a social network analysis tool to analyze these two datasets from the viewpoint of social network structure. Seventeen measurements such as Fragmentation, Component Count, Strong/Weak are summarized to identify the exogenous attributes of Xiaonei.com. Additionally, two latent applications of online social network service are proposed in the discussion section.

Keywords: Online Social Network, Topological Analysis, SNA, Complex Network Analysis.

1 Introduction

Nowadays, online social networks (OSN) such as MySpace, Facebook, Friendster, LinkedIn and Orkut have attracted millions of users, many of whom have integrated these sites into their daily lives. These sites have various objectives including connecting those with shared interests such as music or politics (e.g., MySpace.com), focusing on the college student population (e.g., Facebook.com), dating through one's own friends to create a romantic relationship (e.g., Friendster.com), creating networks of co-workers and business associates (e.g., LinkedIn.com) and linking with Google site developers (e.g., Orkut). Besides these pure online social networks, there are also some online community sites such as the Flickr photo-sharing site, the Youtube

video-sharing site and the SinaBlog blog-sharing site, all of which inside maintain a latent social network.

Out of all these social network services, Facebook is different because one must provide a real campus email address or a valid student identification if one wants to attend a certain network such as the CMU network. Thus the users in Facebook usually use their real names, real photos and have made highly identifiable profiles [1]. Facebook is a social networking site that reinforces and expands real-world social connections. In Facebook, there are more than 65 million active users and over 6 million active user groups. Facebook has 85% market share of U.S. university students over half of whom return daily and spend an average of 20mins per day on it. In the view of the huge potential of Facebook, some cloned sites are emerging now. Xiaonei (www.xiaonei.com) is the Chinese version of Facebook.

A social network represents relationships among friends and its structure has attracted a lot of interest from scholars. The evolution of the structure within the larger online social network of Flickr and Yahoo! 360 has been studied, characterizing users as either passive members of the network; inviters who encourage offline friends and acquaintances to migrate online; and linkers who fully participate in the social evolution of the network [2]. From the viewpoint of complex network theory, Alan Mislove et.al analyze Flickr, YouTube, LiveJournal and Orkut at the same time to confirm the power-law, small-world and scale-free properties of online social networks [3]. In this paper, we focus on researching college student-oriented network services because they more closely approximate a social network in the reality. In fact, these online users like to search to create links with friends with whom they have offline relationships [4]. About 396,836 nodes and 7,097,144 edges [5] in the Xiaonei network and about 4,200,000 users [6] in Facebook network have all manifested their small-world and power-law phenomena. Besides, personal privacy and visualization [7] in these types of network services have attracted more and more scholarly in-depth studies [8].

In this study, in addition to analyzing online social networks from the viewpoint of complex system theory, we analyze them also from the viewpoint of topological structure using a social network analysis method. Social network analysis methods include centrality measures, subgroup identification, role analysis, elementary graph theory, and permutation-based statistical analysis. From this study, we answer sociological questions such as who is the most important in the network when information diffuses, what is the impact on the entire network when one is isolated from the network and how do informal groups form to influence the information diffusion. Our contributions are listed as follows: 1) we analyze the degree distribution and small world phenomena of an entire Chinese student online social network; 2) we use a social network analysis tool—ORA[9] to determine shortest paths and to analyze connected components for online social network; 3) we introduce the new measurements from the viewpoint of social network analysis.

This paper is organized as follows. In section 2, Data sets and crawling method are introduced. In Section 3, we use topological analysis to analyze the structure of online social networks and give some measurements for SNA. In Section 4, we discuss the practical applications of this study. Finally, conclusions and suggestions for further research are presented.

2   Data Sets

Because of the privacy policy of Xiaonei.com, we use the network of Huazhong University of Science and Technology (HUST). In the HUST network, there were 44,419 nodes and 803,987 links in the beginning of January, 2008. In the end of February, 2008, there were 47,546 nodes and 876,983 links. We used these two datasets to analyze the topological structure and evolution of social networks among friends online. Compared to an ego-network, HUST network is a complete social network within a famous Chinese college online social network. It is entirely complete and closed, thus it is suitable for analysis as a sample of online social networks [10].

2.1   An Overview of Xiaonei.com

Xiaonei.com was created in December of 2005 and was an absolutely dominant college social network service in China. Currently, more than 1000 colleges outside China, 3000 Chinese colleges, 8000 Chinese high schools and 7000 companies have opened their network service in Xiaonei.com. Users can find their old friends, make new friends, share photos, music, movies, and share their blogs and personal news. College-oriented social networking sites such as Facebook and Xiaonei provide opportunities to combine online and face-to-face interactions within an ostensibly bounded domain. This makes them different from traditional networking sites: they are communities based "on a shared real space"[11]. Since the majority of these sites require a college's email account for a participant to be admitted to the online social network of that college, expectations of the validity of certain personal information provided by others on the network may increase. Together with the apparent sharing of a physical environment with other members of the network, that expectation may increase the sense of trust and intimacy across the online community.

2.2   Crawling Methods

We used Python to design a script program to analyze the webpage in order to find user profiles and relationships with friends, and then to automatically download user profiles to store into the Mysql database. Due to personal privacy, we were only able to access the HUST network. Furthermore, there were also 5% users who didn't open their profiles for non-friends in the HUST network. First, we used "browse users" function to obtain a fraction of the user ids randomly till the number of selected nodes didn't increase dominantly. Second, we used these user ids to do a snowball sampling by breadth-first search (BFS) in the HUST network. It is known that the power-law nature in the degree distribution is well conserved under snowball sampling since the snowball sampling method easily picks up hubs. This property reduces the degree exponent and produces a heavier tail[12]. It was also annoyed that the verification code was required to input every time when we downloaded 100 profiles. Therefore, we use multi accounts and multi threads to download the profiles.

2.3   Demographics

The majority of users of the Xiaonei at HUST are undergraduate students (83.3% of all profiles). Furthermore, the majority of users are male (65.4% vs. 34.6%). The

average age is 22.14 years. Figure 1 shows the distribution for different users according to their entrance year that represent the year of using Xiaonei because Xiaonei only permits IP addresses from college campuses to register on the HUST network. Every September is the entrance month for students, thus students with entrance year of 2006 have used Xiaonei for one and a half years at most, and the students who entered into college in 2007 have just used Xiaonei for half a year. From their online times and last online dates of two sample data sets, we performed statistical analysis,, finding 71.4% of users login to Xiaonei more than 2 times one week. These users are active users who contribute more to the creation of a virtual community than non-active users.
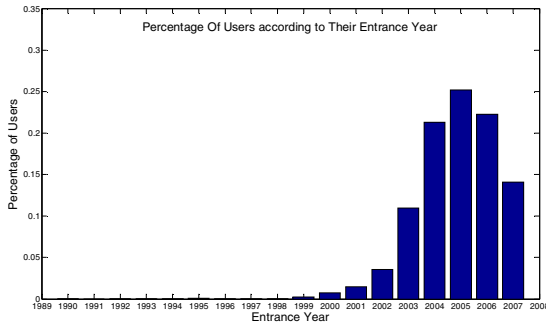

Fig. 1. Percentage of Users according to Their Entrance year

3   Topological Analysis

A social network is a social structure made of nodes (which are generally individuals or organizations) that are tied together by one or more specific types of interdependency, such as values, visions, ideas, financial exchange, friends, kinship, dislikes, conflicts, trades, web links, sexual relations, disease transmission (epidemiology), or airline routes. The resulting structures are often very complex. The Social network analysis method originated in the sociology field and has been widely used in the analysis of online social networks [10]. We pay more attention to the topological structure of online social networks and use ORA [9] plus Matlab to analyze them.

In the Xiaonei network, relationships between friends are reciprocal. The fact that one is directed to certain links can be useful for locating content in information networks. The links in the social networks we studied are regarded as undirected and users may link to each other. This property of symmetry is consistent with that of offline social networks [13].

3.1   Topological Analysis Tool

UCINET is a very popular tool for social network analysis (SNS). It can only handle a maximum of 32,767 nodes (with some exceptions) although practically speaking many procedures get too slow at around 5,000 - 10,000 nodes. We tried to convert our

data into the UCINET data type and input it into UCINET, however the data file was very huge (2.3G) and the process speed was very slow. Therefore, UCINET is not suitable to analyze our data set because the Xiaonei network has 47,546 nodes that exceed the process ability of UCINET.

In this paper, we used ORA [9] that was developed by the Center for Computational Analysis of Social and Organizational Systems (CASOS) of Carnegie Mellon University. Beside being a SNS tool, ORA is also a risk assessment tool for locating individuals or groups that are potential risks given social, knowledge and task network information. ORA is based on JAVA and uses XML file to store data, thus it is compatible with other tools in various operation systems. The converted XML file is just 79.6M large and importing it into ORA is very fast.

3.2   Degrees Distribution

For the college students' online network, the entire Facebook and Xiaonei networks have been proven to satisfy the Power-law [14] rule in degree distribution [5, 6]. The partial samples of the virtual online networks Orkut and MySpace also satisfy Power-law. Here, we also observed the degree distribution of Xiaonei HUST network. In Figure 2, we report the cumulative distribution of degree P (>k), which indicates the probability that a randomly selected node has more than k links. It can satisfy the Power-law (see the right log-log sub-figure of Figure 2), and it also can fit exponential function f(x) = a*exp(b*x) where a=4.247e+004 and b = -0.05131 in data1 and a=4.363e+004 and b = -0.04898 in data2. Here, data1 and data2 were obtained at the beginning of January, 2008 and at the end of February, 2008 respectively.

Furthermore, we observed the distribution of "online times" in the Xiaonei HUST network. The ratio of users according to their online times is shown in Figure 3. We used SPSS to compute the correlation of Degree and Online Times, and a strong correlation (Correlation Coefficient is 0.883) between them was found. That is to say, if one person had more online times, he/she would have more time to search for and make friends. Also he/she could spend more time to construct his/her own webpage to attract more users.
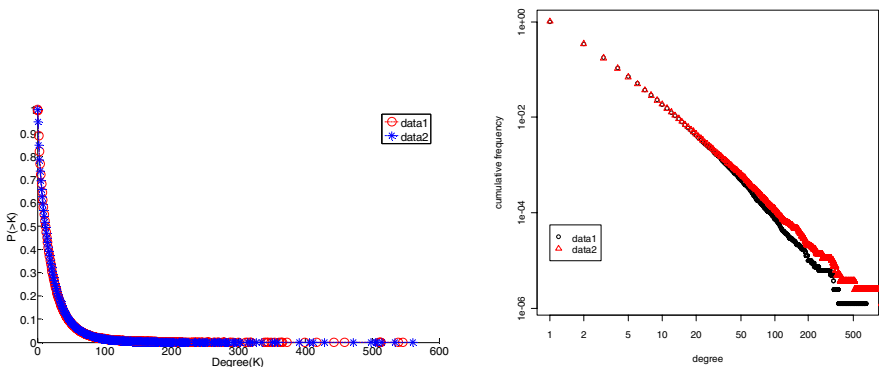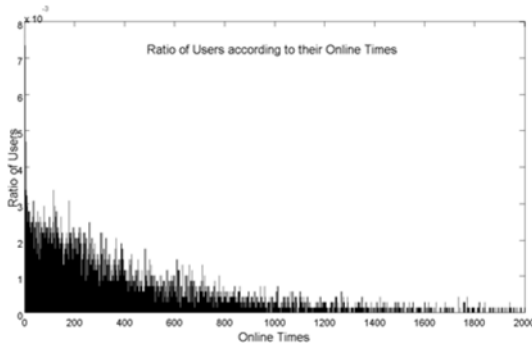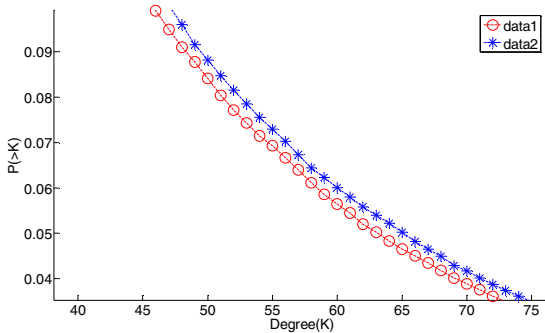

Fig. 2. Degree Distribution of Xiaonei HUST network

Fig. 3. Ratio of Users according to their Online Times

Fig. 4. Degree (45~75) Distribution of HUST network

The users in Xiaonei increased very quickly, from the beginning of January to the end of February 2008. More than 3000 users register Xiaonei account in two month. These users need time to develop their friendships. Also, the CTO of Xiaonei Inc. reported that some users just register to attain a certain objective and then are never active to make friends. A large number of passive members [2] results in exponential decay in the first part of distribution curve. We amplified the degree distribution in Figure 2 and focused on the part between 45 and 75 degree shown in Figure 4. We can find the successive degree distribution curves move upwards. The probability that a randomly selected node has more than k links increases at the same "degree" value. There is a tendency to close to Power-law, but it evolves very slowly because a large number of new passive members register for Xiaonei HUST network.

3.3 Small World Phenomena

The small world experiment was comprised of several experiments conducted by Stanley Milgram examining the average path length for social networks of people in the United States. The research was groundbreaking in that it revealed that human society is a small world type network characterized by shorter-than-expected path

Table 1. Topological Measurements of Online Social Network

MEASURE	TYPE	Data1	Data2
Average Distance[16]	Graph	4.15207	4.16321
Clustering Coefficient, Watts-Strogatz[16]	Graph	0.198627	0.187391
Component Count, Strong[10]	Graph	*NA*	*NA*
Component Count, Weak[10]	Graph	173	764
Connectedness, Krackhardt[17]	Graph	0.991732	1.03602
Density[10]	Graph	0.000412741	0.000426163
Diameter[10]	Graph	44419	47546
Edge Count Ratio, Lateral[18]	Graph	0.563342	0.556119
Efficiency, Global[19]	Graph	0.253951	0.249533
Efficiency, Local[19]	Graph	0.303109	0.304126
Fragmentation[20]	Graph	0.0082677	0.0360216
Span of Control[21]	Graph	*NA*	*NA*
Speed, Average[18]	Graph	0.240843	0.2402
Upper Boundedness, Krackhardt[17]	Graph	0.999871	0.997303
Network Centralization/Total Degree[22]	Graph	0.0102928	0.00994435
Hierarchy, Krackhardt[17]	Graph	0.0082217	0.0080065
Centrality, Closeness[22]	Graph	8.88937e-05	8.81563e-05

lengths. The experiments are often associated with the term six degrees of separation [14], which was a small world phenomenon. We check the average shortest path length between entities. As shown in Table 1, Average Distance on two data sets is 4.15207 and 4.16321 respectively. Values less than six represent small world phenomena. Therefore, it can be proven that there exist small world phenomena in the Xiaonei HUST network.

3.4 Shortest Path Finder

Given a pair of two selected entities, we computed the shortest path between two entities according to Newman's algorithm[15]. We used Java to implement Newman's algorithm and compared it with ORA shortest path finder. The average consuming time of Newman's algorithm is 25 second; however ORA just take 20 second. For example, we computed the shortest path from agent 0-100013893 to agent 17868-223752718 and determined the number of shortest paths. There were 12 shortest paths and the shortest path length was 5, as shown in Figure 5. The Sphere of Influence of each entity represents a series of relationships with friends who have a direct relationship with this entity. It can also be computed using ORA. As shown in Figure 6 and 7, where the degree of entity 0-100013893 is 3 and the degree of entity 17868-223752718 is 12, the sphere of influence of entity 0-10001389 is obviously larger than that of entity 17868-223752718. The computation of the shortest path and

Fig. 5. Shortest paths from agent 0-100013893 to entity 17868-223752718

Fig. 6. Sphere of influence for 0-100013893 and 17868-223752718

the sphere of influence can provide a latent application for online social network, which will be discussed in Section 4.

3.5 Connected Component Analysis

The nodes in a disconnected graph may be partitioned into two or more subsets in which there are no paths between the nodes in different subsets. The connected sub-graphs in a graph are called components [10]. A component in an online social network is a sub-network in which there are links between all pairs of users in the sub-network, and no link between a user in the network and any user not in the network. These disconnected components represent the informal groups in a network organization. Although online network services provide interest groups for users, it cannot be ignored that informal groups also take an important role in the construction of online social networks, which will be discussed more in Section 4. There are two types of connected components: strongly connected components and weakly connected components [9]. Because our Xiaonei HUST network is undirected, we chose the weakly connected component method to compute the number of connected

components. As shown in Table 1, through two months, the number of components increased from 173 to 764. More informal groups were born with the interaction of users in online social networks.

3.6 Measurements on SNA

For the purpose of analyzing online social networks, we had to filter some measures from more than 100 measurements in ORA. We paid more attention to the entire topological structure of online social networks, thus we filtered on the base of the Graph type of the measurements. In Table 1, we choose 17 measurements. The relative references are also listed in the Table 1. We briefly analyzed the results according to these measurements as follows.

1) Density: The number of edges divided by the number of possible edges including self-reference.
2) Diameter: The maximum shortest path length between any two nodes in a network.
3) Connectedness: Measures the degree to which a square network's underlying (undirected) network is connected.
4) Edge Count Ratio Lateral: Fixing a root entity x, a lateral edge (i,j) is one in which the distance from x to i is the same as the distance from x to j.
5) Efficiency, Global: Measures the closeness of the entities in the network.
6) Efficiency, Local: Measures the closeness of the entities in each ego network in the network.
7) *Fragmentation:* The proportion of entities in a network that are disconnected. It is related to Component Count. In two month, the number of fragmentations in Xiaonei HUST network increases.
8) Span of Control: The average number of out edges per node with non-zero out degrees. It is only used in the directed network. Therefore, for Xiaonei HUST undirected network, we can not obtain this result.
9) Speed, Average: The average inverse geodesic distance between all entity pairs. The highest score is achieved for a clique, and the lowest for all isolates.
10) Upper Boundedness: The degree to which pairs of agents have a common ancestor.
11) Network Centralization, Total Degree: A centralization of a square network based on total degree centrality of each entity.
12) Hierarchy: The degree to which a square network N exhibits a pure hierarchical structure.
13) Clustering Coefficient, Watts-Strogatz: Measures the degree of clustering in a network by averaging the clustering coefficient of each entity. The clustering coefficient of a entity is the density of its ego network which is the sub graph induced by its immediate neighbors.
14) *Component Count, Strong/Weak:* The number of strongly/ weekly connected components in a network.
15) Centrality, Closeness: The average closeness of an entity to the other entities in a network. Loosely, Closeness is the inverse of the average distance in the network between the entity and all other entities. This is defined for directed networks.

4   Discussions

Facebook has opened its API so that anyone can develop his/her own application, and then upload the application for using it online. Unfortunately, Xiaonei has not opened its API yet. Therefore, the latent applications demonstrated here don't specify the application development of Xiaonei network. Furthermore, offline social networks also benefit from online social network services.
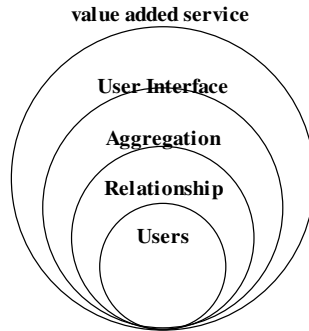

Fig. 7. Applications of Online Social Network

In Figure 7, we classify applications of online social networks five categories. The basic application allows users to edit their own information freely. Second, for the purpose of creating friendships, we need to provide user-friendly application to display and maintain friend lists. Third, aggregation lets online social network services push all the information to users and lets users efficiently use that information. Fourth, the user interface on the website is very important for attracting new users and keeping old users. Finally, a value added service such as mobile SNA can allow users to enjoy a full virtual life. From the study on which paper is based, two latent applications emerge for online social networks at the level of relationships and aggregation.

Shortest path finder. The ultimate objective of using online social networks is for people to extend their circle of friends. Since users even want to meet new people online, it is very useful to find the shortest path to do so. Even though, users can find this certain people according to his/her name, work place or other information and send an invitation to him/her. Generally, users do not like to add strangers as new friends[4]. Rather they prefer to add users whom they have known offline. Also, some people would like to become initially acquainted with each other online but actually interact in the real world. SNA is just a tool to make offline friends by first meeting them online. Therefore, shortest path finder can help users find efficient ways to make offline friends by search shortest path online.

Informal Online Group Mining. The objective of this application is to help users to jump out of his/her own social group. By judging if the ego-network is disconnected components or not or by finding shortest path for jumping out of this disconnected

component, people can expend their social circles. Also, by mining informal groups that are disconnected components, online social network services can provide a lot of references to help users take part in certain informal groups.

5 Conclusions

In this paper, we used a topological analysis tool to analyze an online Chinese student social network—Xiaonei.com. We used Python to crawl two datasets of Xiaonei from January to February 2008. By these two datasets, degree distribution and small world phenomena were analyzed using Matlab and SPSS. It was found that degree distribution has a tendency to close to Power-law, however it evolves very slowly because a large number of new passive members routinely register for the Xiaonei HUST network. Xiaonei HUST network also has small world attribute. Then we used a social network tool, ORA, to analyze our datasets from the viewpoint of social network analysis. We summarized 17 measurements such as Fragmentation, Component Count, Strong/Weak. Also, 2 latent applications of online social network services were proposed in the discussion section. This paper also shows that ORA[1] have the advantages in finding shortest path and analyzing connected component.

This is only a preliminary report for our work-in-progress research. In the future, we will contact Xiaonei Inc. to obtain their cooperation in order to analyze the entire Xiaonei network dataset. Further study for applying classical social network analysis to the analysis of online social networks will be carried out.

References

1. Ralph Gross, A.A.: Information Revelation and Privacy in Online Social Networks (The Facebook case). In: ACM Workshop on Privacy in the Electronic Society (WPES) (2005)
2. Ravi Kumar, J.N., Tomkins, A.: Structure and Evolution of Online Social Networks. In: Proceedings of 12th International Conference on Knowledge Discovery in Data Mining, ACM Press, New York (2006)
3. Alan Mislove, M.M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. 2007, San Diego, California, USA. ACM, New York (2007)
4. Lampe, C., Ellison, N., Steinfeld, C.: A familiar Face (book): Profile elements as signals in an online social network. In: Proceedings of Conference on Human Factors in Computing Systems. ACM Press, New York (2007)
5. Fu, F., Long Wang, L.L.: Empirical analysis of online social networks in the age of Web 2.0. Physica A 387, 675–684 (2008)

---

[1] Please visit http://www.casos.cs.cmu.edu/projects/ora/

6. Scott, A., Golder, D.W., Huberman, B.A.: Rhythms of social interaction: messaging within a massive online network. In: 3rd International Conference on Communities and Technologies (CT 2007), East Lansing, MI, June 28-30 (2007)
7. Jeffrey Heer, D.B.: Vizster: Visualizing Online Social Networks. In: Proceedings of the 2005 IEEE Symposium on Information Visualization (INFOVIS 2005) (2005)
8. Acquisti, A., Gross, R.: Imagined communities: Awareness, information sharing, and privacy on the Facebook. In: Proceedings of 6th Workshop on Privacy Enhancing Technologies. Robinson College, Cambridge (2006)
9. Carley Kathleen, C.D., Matt, D., Jeffrey, R., Moon, I.-C.: ORA User's Guide, 2007, Carnegie Mellon University, School of Computer Science, Institute for Software Research, p. 388 (2007)
10. Wasserman Stanley, K.F.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
11. Sege, I.: Where everybody knows your name (April 27, 2005) Boston.com
12. YongYeol Ahn, S.H., Kwak, H.: Analysis of Topological Characteristics of Huge Online Social Networking Services. In: International World Wide Web Conference, Banff, Alberta, Canada (2007)
13. Thomas Lento, H.T.W., Gu, L.: The Ties that Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop Weblogging System (2007)
14. Newman, M.: The Structure and Function of Complex Networks. SIAM Review 56, 167–256 (2003)
15. Newman, M.E.J.: Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E 64 (016132) (2001)
16. Watts, D.J., Strogatz, S.: Collective dynamics of 'small-world' networks. Nature 393(6684), 440–442 (1998)
17. David, K.: Graph Theoretical Dimensions of Informal Organizations. In: Carley, M.J.P.H., Kathleen, M. (eds.) Computational Organization Theory. Lawrence Erlbaum Associates, NJ (1994)
18. Carley, K.: Summary of Key Network Measures for Characterizing Organizational Architectures, Carnegie Mellon University (2002)
19. Latora, V., Marchiori, M.: Efficient Behavior of Small-World Networks. Phys. Rev. Lett. 87(19) (2001)
20. Borgatti, S.P.: The Key Player Problem. Dynamic Social Network Modeling and Analysis. In: Committee on Human Factors, National Research Council, Workshop Summary and Papers (2003)
21. David, K., Tie, S.: Super Strong and Sticky. Power and Influence in Organizations. In: Roderick Kramer, M.N. (ed.), pp. 21–38. Sage, Thousand Oaks (1998)
22. Freeman, L.C.: Centrality in Social Networks I: Conceptual Clarification. Social Networks 1, 215–239 (1979)