A Study on Social Influence Analysis in Social Networks

Sujatha Yeruva
Department of Computer Science
St. Francis College for Women
Hyderabad-16

B. Sarojini Ilango, Ph.D
Department of Computer Science
Avinashilingam University
Coimbatore – 43

Y. Samatha Reddy
Department of System Engineering
NVIDIA Graphics Pvt Ltd
Hyderabad-46

ABSTRACT
The World Wide Web is one of the most inevitable notions in the life of mankind. In the most recent times of the world it is much more in advance gaining popularity due to its enormous amount of capability in making the life more impact-able. Online social networks are one of the areas of the World Wide Web where people congregate to share and be part of the various virtual communities. Online social networks are more fascinating to many of us now as they look out for similarly inclined people in order to share in reciprocity their findings, ideals, thoughts, opinions and views. Much like the physical human networks, cybernetics too has people who can influence or influenced by the other. Some are leaders who inevitably influence voluminous people, while others look out to get influenced or to induce inspiration from their leader. Identifying people who exercise maximum influence could be useful in targeting them for marketing, knowledge dissemination and other such purposes. In this paper, we present empirical analysis of levels of influence in the online social network. Findings of this paper may be of great help to elucidate in ascertaining leaders of a social network which in turn can be used to sight the leaders in real world. The leaders of a social network are traced by using *Pareto front function.* We believe that this is the first study to use *Pareto front function* to identify the leaders in online social networks. The empirical results prove that the number of leaders in each subsequent level monotonically increase while the number of their followers decreases.

General Terms
World Wide Web, Cybernetics, Social Networks

Keywords
Social Network, measuring influence, Pareto front function

1. INTRODUCTION
Online social networks are gaining increasing importance in influencing consumer behaviour, political, religious thinking and public opinion. How the influence propagates across the network makes for interesting study [1]. A study of social networks like Facebook, Twitter and YouTube reveals that a few users can exert tremendous influence on a large section in areas of marketing, social and religious beliefs and in spreading knowledge. The crux of the problem is to identify users who exert maximum influence on a group, so that these individuals can be targeted for marketing and promoting social awareness [2-4]. There is a need not only to discover the leaders in a community, but also discriminate their levels of leadership in order to direct the progress of the evolution of the community, in any particular dimension, from good to better by appropriately influencing the leaders, since every leader inevitably influences and directs the thinking and activity in the community directly or indirectly through his followers. This is true of the political, religious and particularly the business communities today. The effect of leadership in a business community is ultimately manifested in the marketing sector of society. Discovery and discrimination of the various levels of leadership in a business community, for example, can be used to promote individuals from one office to a higher one or in selecting proper people for further business training. Religious and political leaders too can be designated to act upon crucial role for an advancement of drastic social changes. The importance of discovering and discriminating the leadership levels in a given community motivated us to find a reliable method of discovering the levels of leadership in any given community. We believe that this is the first study to examine the influence with *Pareto front function* in online social networks.

We represent a social network using a directed graph data structure. The nodes denote users and the directed arcs between the nodes denote interactions. The edges that originate from a node constitute the out-degree of the node. Similarly, the edges that terminate at a node form the in-degree of the node. In short, the out-degree of a node is a measure of the influence a node exerts on the other nodes. In a similar fashion, the in-degree of a node is an estimate of the extent to which the node is influenced by the other nodes. We trace a leadership in a network by using *Pareto front function.* Pareto front returns the logical Pareto front of a set of points [5]. The term Pareto front is named after Vilfredo Pareto (1848–1923), an Italian economist who used the concept in his studies of economic efficiency and income distribution [6]. In words, this definition says that $\underset{x}{\rightarrow}$ is Pareto optimal if there exists no feasible vector of decision variables x ϵ F which would decrease some criterion without causing a simultaneous increase in at least one other criterion. This concept almost always gives not a single solution, but rather a set of solutions called the Pareto optimal set. The vectors $\underset{x}{\rightarrow}$ corresponding to the solutions included in the Pareto optimal set are called non-dominated. The plot of the objective functions whose non-dominated vectors are in the Pareto optimal set is called the Pareto front [7]. This research uses the Pareto front function to identify the leaders and their followers in each level. Also, the percentage of leaders and their percentage of influence expressed in terms of number of followers is calculated. The empirical results show that the number of leaders in each level increases while the number of followers decreases.

The rest of the paper is organised as follows: in section 2 we discuss in detail the dataset that we used, related work is discussed in section 3 and in section 4 the methodology followed by empirical results and graphical representation in section 5, finally conclusions are drawn in section 6.

2. DATASET

To analyse social network and their influence, a sample of data of the YouTube social network from Measurement and Analysis of Online Social Networks by Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, Bobby Bhattacharjee [8] is taken up for study. They collected data through crawls of the user graphs, by accessing the public web interface delivered by the sites which provided them an access to bulky data sets from numerous sites. The dataset was made available for the research community for the further study [9]. The data set has been anonymized using "best effort anonymization" to safeguard the privacy of the users which is very challenging. The dataset comprises a list of links, that is, the list of all of the user-to-user links which are included in the crawls. All links are considered as directed. Each line holds two user identifiers separated by a tab, denoting a link exists from the first to the second.

3. LITERATURE SURVEY

Resurgence of interest in online social network made many researchers to review and contribute for the further study. Techweb [10] testified that the Social networking services "attract nearly half of all web users". Becker, J. A. H., & Stamp, G. H. [11], made a study that traditional face-to-face communication has expanded accordingly to diverse regions of computer-mediated communication. Strength of network ties were explored by Harmonie Farrow, Y. Connie Yuan [12] to show how attitudes of the alumni is influenced by social network site, Facebook, to volunteer for and make charitable gifts to their alma mater thus fortify consistency between attitude and behaviour. Matthew S. Weber [13] demonstrated that initial actions in the online news community do have a clear and measurable effect on the news media industry. Zsolt Katona, Peter Pal Zubcsek, and Miklos Sarvary [14], have sought "to identify various factors that predict influential power of individuals". They say that in addition to the traditional factors (for instance, demographics) "local network characteristics are important in predicting influential power". Their findings prove that "social network analysis is a promising tool for marketers to predict and influence consumer behaviour".

The findings of Linyuan Lu, Yi-Cheng Zhang, Chi Ho Yeung, Tao Zhou [15] indicate that online communities can intensify the influence of a small number of significant users for the benefit of all other users. They utilized the leadership topology and recognized significant users to develop an adaptive and parameter-free algorithm, the LeaderRank, to measure user influence. Ulrike Pfeil, Panayiotis Zaphiris, Stephanie Wilson [16], offered an analysis of an online support community for older people, examining a data-set of messages posted over the period of six years. They presented how certain sequences of messages within the online community are connected to the level of activity hence providing prized insight in to the role of message-sequences in sustaining online support communities for older people. By knowing the characteristics of important message-sequences and how they are connected to the level of activity within the online support community, they were able to highlight the origins of prevailing problems or successes. This information could also assist moderators of online support communities to uphold and foster an online support community, making them more successful and thus also more beneficial.

Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, Bobby Bhattacharjee [17], indicate that networks comprise a densely connected core of high-degree nodes which link small groups of strongly clustered, low-degree nodes and discuss the implications of the structural properties for the design of social network based systems. Bruce Hoppe, Claire Reinelt [18], offer a structure for visualizing diverse types of leadership networks and uses case examples to detect outcomes typically associated with each type of network. Statistical frameworks which validate the benefits of forming multiple relations jointly for both substantive and predictive purposes were developed by Asim Ansari, Oded Koenigsberg, and Florian Stahl [19]. They developed an integrated statistical framework for concurrently modelling the connectivity structure of multiple relationships of diverse types on a common set of actors. Their first application uses a sequential network of communications among managers involved in new product development activities and the second uses an online collaborative social network of musicians. They also demonstrated the use of information in one relationship that could be used to predict connectivity in another.

Juan-J. Merelo and Carlos Cotta [20] asserted that the most pertinent nodes (i.e. authors) in a co-authorship network of EC at a microscopic level. Their outcomes of study specify that there are some well-known researchers who appear systematically in top rankings providing some clues on the social behaviour of our community. A model of visible peer networks influence effects in electronic markets was presented by Gal Oestreicher-Singer and Arun Sundararajan [21]. They also delivered broad empirical proof of the magnitude and variation in this influence. More accurately, their results have revealed that visible co-purchase networks intensify the shared purchasing of complementary products, and document how this influence differs along a number of different dimensions. Francesco Bonchi [22], consider a data mining standpoint and what (and how) can be learned from the accessible traces of past proliferations to recognise the influential users, by aiming whom certain needed marketing outcomes can be achieved.

The binary concept method proposed by D. M. Akbar Hussain [23], offers a clear image in finding the role of a leader or follower which is difficult to decide with standard centrality measures. Amit Goyal, Francesco Bonchi, Laks V. S. Lakshmanan [24], present an innovative frequent pattern mining methodology to determine leaders and tribes in social networks. Reihaneh Rabbany Khorasgani, Jiyang Chen, Osmar R. Zaïane [25] presented an algorithm which starts by finding talented leaders in a specified network then iteratively accumulates followers to their nearby leaders to form communities, and later finds new leaders in each group about which to gather followers again until convergence. Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, Xiangyang Lan [26] made a study on the development of informal interdependent groups within a huge organization who can offer insight into the organization's overall decision-making behaviour.

4. METHODOLOGY

In the proposed method the social network is represented using a directed graph data structure [27] with nodes representing users and the directed arcs between the nodes representing interactions. The edges that emanate from a node constitute the out-degree of the node. Likewise, the edges that terminate at a node form the in-degree of the node. Thus an out-degree of a node is considered as an amount of influence that particular node has on other nodes, while an in-degree of

a node is assumed as an amount of influence that a particular node gets by the other nodes.

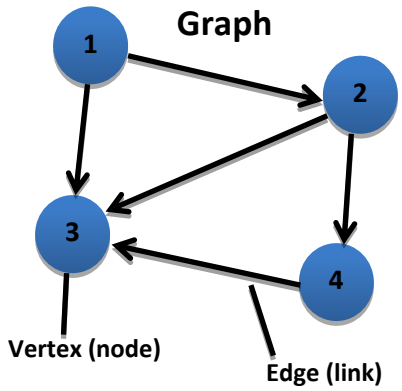

Fig 1: A simple directed graph with nodes and directed links

Figure 1 shows a simple directed graph with four nodes. For instance from the figure 1 the in-degree of node 2 is one and out-degree of node 2 is two. That is node 2 influences or leads two other nodes namely nodes 3 and 4, while it is influenced or followed by only one node that is node 1. In order to study the reachability of each node we have considered Multi-objective optimization. Multi-objective optimization also known as multi-criteria or multi-attribute optimization is the process of simultaneously optimizing two or more conflicting objectives subject to certain constraints [28]. The nodes which have high ratio of out-degree to in-degree have more reachability as they influence many nodes. Also the nodes which have high in-degree too can influence many other nodes as they can incorporate the influence they received from other nodes. We have considered these two as objectives for multi-objective optimization to trace leaders. The nodes which are directly connected to these leaders are their followers in one level. In figure 1 node 2 can be considered as leader and nodes 3 and 4 can be considered as its followers. This is represented in table 1.

Table 1. Influence of the nodes shown in figure 1

Node	OD	ID	R_OD/ID	R_ID	Follower nodes
1	2	0	NA*	NA*	2,3
2	2	1	1	1	3,4
3	0	3	NA*	NA*	None
4	1	1	2	1	3

*NA=Not Applicable

From the sample data, the out-degree and in-degree of every node is calculated and the node that has at least one out-degree is considered for empirical study. 569,913 unique nodes match this criterion and are considered for further investigations. In the first round it was observed that the in-degree of few nodes was zero. The nodes with in-degree zero are the ones who can influence others but are not influenced by others. They are the absolute leaders in the community. The nodes with in-degree as zero and their followers are eliminated from the sample data in order to discriminate and discover the next level of leaders and their followers.

The degree of reachability of a node determines the level of influence a node can have on others. The nodes with high ratio of out-degree to in-degree reach out more and are highly influential. Also the nodes which have high in-degree too can influence many other nodes as they can propagate the influence that they received from other nodes. Figure 2, shows the graph representing the rank of the ratio of out-degree to in-degree along Y-axis and the rank of in-degree along X-axis, with the leader nodes depicted as red colour among the total population which are depicted in blue colour. To trace the leaders in this graph *Pareto front function* is applied. *Pareto front* produces the coherent *Pareto front* of a set of points. Thus, with the help of the *Pareto front function* leaders are picked up as $S= \{(x, y) \mid x<x_i \; \forall \; x_i$ and $y<y_i \; \forall \; y_i$ for i=1 to n$\}$. Thus the first level of non-absolute leaders and their followers are discovered. The first level leaders, their followers and absolute leaders are removed to compute the next level of leaders and their followers. The process is repeated till there are no significant number of leaders and followers left in the total population.

Algorithm:

Step 1: Compute the out-degree, in-degree of entire nodes. Identify the nodes with in-degree zero. They are the born leaders.

Step 2: Compute ratio of out-degree to in-degree and the ranks of in degree and ranks of the ratio of out-degree to in-degree.

Step 3: Apply *Pareto front function* to trace the leaders from the ranks that we computed in step 2.

Step 4: Find out the followers of the leaders found in step 3.

Step 5: Eliminate the leaders and followers that were traced in steps 3 and 4.

Step 6: Repeat Steps from 2 to 5 till there are no significant number of leaders and followers left in the total population.

The graph in Figure 2 depicts the rank of ratio of out-degree to in-degree along Y-axis and the rank of in-degree along X-axis. Nodes in red color denote the leader nodes identified from the entire population of nodes in blue color.

5. EMPIRICAL RESULTS AND GRAPHICAL REPRESENTATION

The experiments were performed using MATLAB R2009a and Java 1.6. Table 2 depicts the summary of the experiment.

The number of leaders and their followers at each level are calculated. At each level leaders along with their followers are eliminated. From the remaining population we calculated the next level of leaders and their followers. The leaders found at each level are accumulated to calculate the influence of the leaders on the total population. At level 1 we observe that 14 leaders influence 71918 from the total population of 569913. That is, the percentage of influence at level 1 is 12.62 of total population. The influence at each level monotonically decreases.

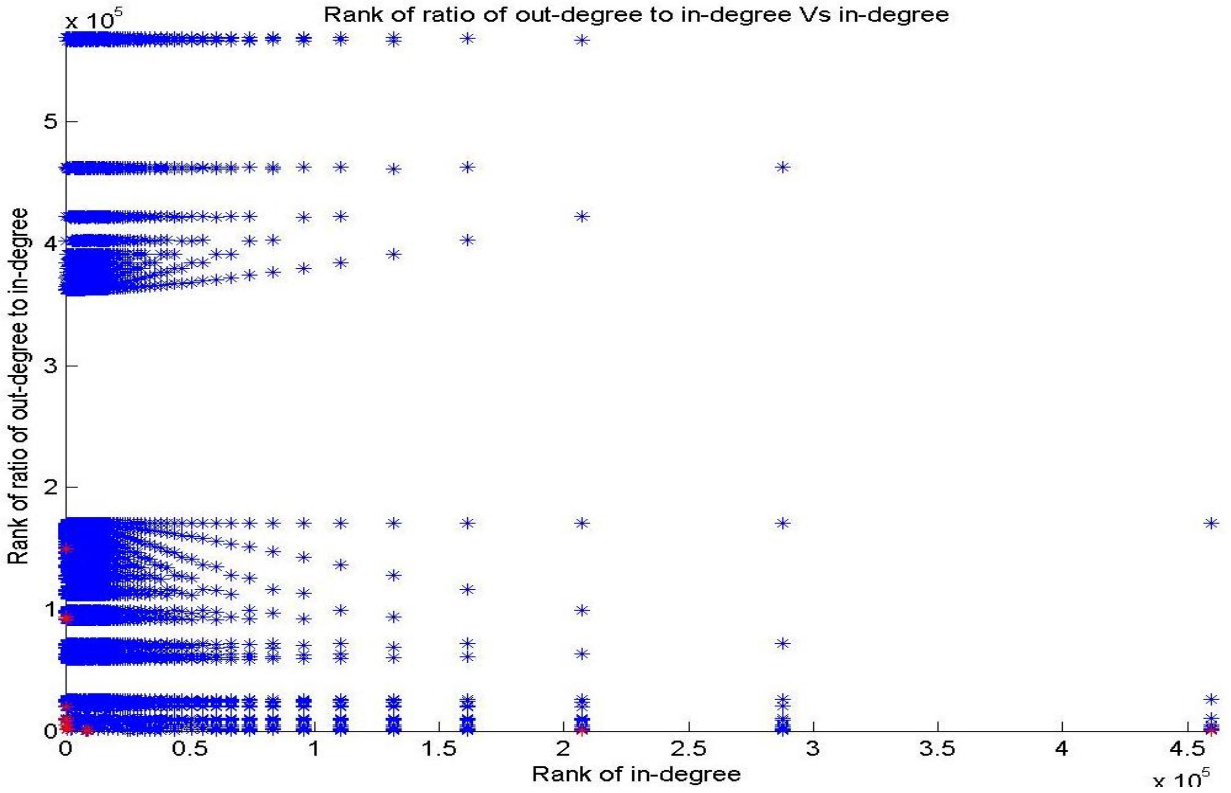

Fig 2: Representation of ratio of rank of out-degree to in-degree and rank of in-degree

Table 2. The computed values of Leaders and the Followers of Leaders at each level

Levels	Total Nodes	l_i Leaders	f_i Followers	% of followers	l_i+f_i Total number of leaders and followers	$\sum l_i = L_i$ Accumulated leaders	$\sum f_i=F_i$ Accumulated Followers	$L_i/N=PL_i$ % of Accumulated leaders	$F_i/N=PF_i$ % of Accumulated Followers
1	569913	14	71918	12.62	71932	14	71918	0.00	12.62
2	515808	18	26988	5.23	27006	32	98906	0.01	17.35
3	500707	20	18852	3.77	18872	52	117758	0.01	20.66
4	492199	29	18875	3.83	18904	81	136633	0.01	23.97
5	483512	23	12480	2.58	12503	104	149113	0.02	26.16
6	478105	26	12221	2.56	12247	130	161334	0.02	28.31
7	473541	31	10532	2.22	10563	161	171866	0.03	30.16
8	469587	32	11592	2.47	11624	193	183458	0.03	32.19
9	465682	32	9309	2.00	9341	225	192767	0.04	33.82
10	462056	36	9034	1.96	9070	261	201801	0.05	35.41
11	458291	38	9327	2.04	9365	299	211128	0.05	37.05
12	454689	36	7423	1.63	7459	335	218551	0.06	38.35
13	451567	46	9902	2.19	9948	381	228453	0.07	40.09

14	447545	44	8048	1.80	8092	425	236501	0.07	41.50
15	444791	46	7715	1.73	7761	471	244216	0.08	42.85
16	441928	54	9372	2.12	9426	525	253588	0.09	44.50
17	438454	51	8672	1.98	8723	576	262260	0.10	46.02
18	435427	50	7251	1.67	7301	626	269511	0.11	47.29
19	432623	51	6262	1.45	6313	677	275773	0.12	48.39
20	430703	50	6417	1.49	6467	727	282190	0.13	49.51
21	428632	50	5826	1.36	5876	777	288016	0.14	50.54
22	426781	46	4946	1.16	4992	823	292962	0.14	51.40
23	425263	52	5859	1.38	5911	875	298821	0.15	52.43
24	423286	49	5467	1.29	5516	924	304288	0.16	53.39
25	421621	55	5969	1.42	6024	979	310257	0.17	54.44
26	419555	53	5305	1.26	5358	1032	315562	0.18	55.37
27	417838	53	5242	1.25	5295	1085	320804	0.19	56.29
28	416062	53	5049	1.21	5102	1138	325853	0.20	57.18
29	414165	48	4287	1.04	4335	1186	330140	0.21	57.93
30	412922	57	5098	1.23	5155	1243	335238	0.22	58.82
31	411244	62	5180	1.26	5242	1305	340418	0.23	59.73
32	409571	60	4709	1.15	4769	1365	345127	0.24	60.56
33	408020	57	4251	1.04	4308	1422	349378	0.25	61.30

The total number of leaders up to level 33 is 1422 and these number of leaders influence 61.30% of the total population. That means in order to influence 61.30% of people; it is enough to target and influence 1422 people who form 0.25% of the total population.

Figure 3 depicts the graphical representation of monotonous increase in the number of leaders and their followers at each level. PL_i is the percentage of cumulative leaders up to level i and PF_i is the percentage of cumulative followers of leaders up to level i.

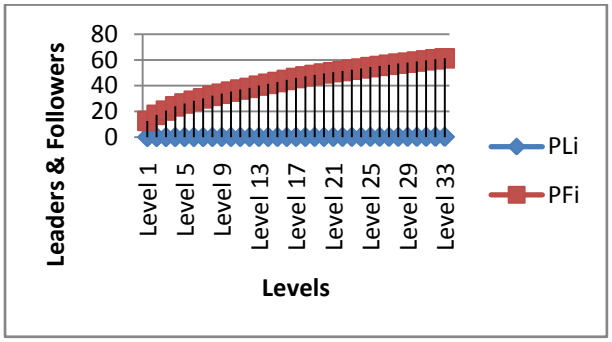

Fig 3: Cumulative number of leaders and followers at each level

In real life scenario more prominent leaders could influence more number of followers compared to the leaders with less prominence. This reflect in our empirical results, at level 1, the most prominent 14 leaders were able to influence 12.62% of the total population, whereas at level 33, 57 leaders were able to influence only 1.04% of the population. This is depicted in Figure 4.

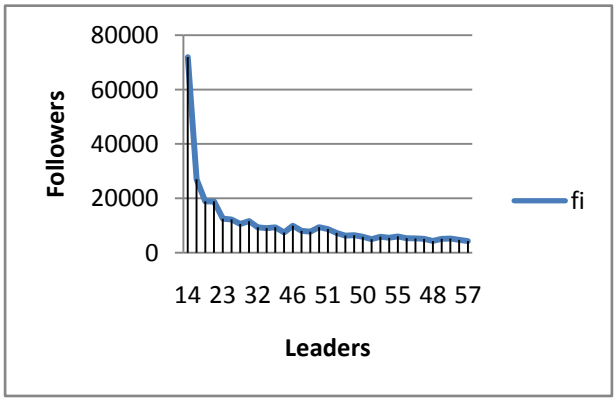

Fig 4: Number of followers influenced by their leaders at each level

6. CONCLUSION

We have presented an analysis of the users of online social networks using data sets publicly made available online for the research community. Our experiments ascertain that at the top most level there are few significant users who make high impact on huge population while with each subsequent level having increasing number of leaders with decreasing number of followers been influenced by them. Thus in order to make a huge impact on the society it is enough to concentrate on the few significant users who in turn influence their followers. Outcomes of this paper may be used to elucidate in discovering prominent leaders of a social network who play a major role in influencing the people. The identification of such groups will be of helpful to Business analysts, knowledge diffusers and social reformers to determine the people who could greatly influence the society. Determination of such promising leaders groups helps in various ways viz., in business the area of marketing can further enhance the growth of a business community; expert leaders of various faculties in spreading the knowledge; social, political and religious leaders in bringing out radical social change. Ample work still remains. We have focused exclusively on the users who make an influence directly on their immediate followers. In our future work we assume to make a study, as, to how many generations the influence of the significant users would reach.

7. ACKNOWLEDGMENTS

We thank the management of St.Francis College for Women for the encouragement and timely assistance in completion of this research paper.

8. REFERENCES

[1] Coleman, J. 1993. S 1990 Foundations of Social Theory. Chs, 8, 12.

[2] Fulk, J., Schemitz, J., & Steinfield, C. 1990. A social influence model of technology use. In:J. Fulk & C. Steinfield (Ed.), Organizations and communication technology (pp. 117–142), Thousand Oaks, CA: Sage.

[3] Li, H., Daugherty, T., & Biocca, F. 2002. Impact of 3D advertising on product knowledge, brand attitude, and purchase intention: The mediating role of presence. Journal of Advertising, 31, 43–57.

[4] Jung, Y. (2011). Understanding the role of sense of presence and perceived autonomy in users' continued use of social virtual worlds. Journal of Computer-Mediated Communication, 16(4), 492-510.

[5] Paretofront :By Yi Cao at Cranfield University, 31 October 2007

[6] *Pareto front function*: Definition, history. Website accessed May 18, 2012 from http://en.wikipedia.org/wiki/Pareto_efficiency

[7] Carlos A. Coello Coello. Basic Concepts

[8] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. 2007, October. Measurement and analysis of online social networks. InProceedings of the 7th ACM SIGCOMM conference on Internet measurement(pp. 29-42). ACM

[9] Online social network research. The Max Plank Institute for Software Systems. Website accessed Feb 12, 2012 from http://socialnetworks.mpi-sws.org/data-imc2007.html

[10] A global business that supplies amenities to the technology industry. Website accessed June 28, 2012 from http://www.techweb.com

[11] Becker, J. A. H., & Stamp, G. H. 2005. Impression management in chat rooms: A grounded theory model. *Communication Studies*, *56*(3), 243–260.

[12] Farrow, H., & Yuan, Y. C. 2011. Building stronger ties with alumni through Facebook to increase volunteerism and charitable giving. Journal of Computer-Mediated Communication, 16(3), 445-464.

[13] Weber, M. S. 2012. Newspapers and the Long-Term Implications of Hyperlinking. Journal of Computer-Mediated Communication, 17(2), 187-201.

[14] Katona, Z., Zubcsek, P. P., & Sarvary, M. 2011. Network effects and personal influences: The diffusion of an online social network. Journal of Marketing Research, 48(3), 425-443.

[15] Lü, L., Zhang, Y. C., Yeung, C. H., & Zhou, T. 2011. Leaders in social networks, the delicious case. PLoS One, 6(6), e21202.

[16] Pfeil, U., Zaphiris, P., & Wilson, S. 2010. The role of message-sequences in the sustainability of an online support community for older people. Journal of Computer-Mediated Communication, 15(2), 336-363.

[17] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. 2007, October. Measurement and analysis of online social networks. InProceedings of the 7th ACM SIGCOMM conference on Internet measurement(pp. 29-42). ACM.

[18] Hoppe, B., & Reinelt, C. 2010. Social network analysis and the evaluation of leadership networks. The Leadership Quarterly, 21(4), 600-619.

[19] Asim Ansari, Oded Koenigsberg, Florian Stahl. Modelling Multiple Relationships in Social Networks. *Journal of Marketing Research* Article Postprint © 2011, American Marketing Association

[20] Juan.J. Merelo and Cotta.C. Who is the best connected EC researcher? Centrality analysis of the complex network of authors in evolutionary computation. February 1, 2008

[21] Oestreicher-Singer, G., & Sundararajan, A. 2008. The visible hand of social networks in electronic markets. *Electronic Commerce Research*.

[22] Bonchi, F. 2011. Influence propagation in social networks: A data mining perspective. *The IEEE Intelligent Informatics Bulletin*, *12*(1).

[23] Hussain, D. M. A.Identifying Leader or Follower using a Binary Approach. Proceedings of the International MultiConference of Engineers and Computer Scientists 2010 Vol I, IMECS 2010, March 17-19, 2010, Hong Kong.

[24] Goyal, A., Bonchi, F., & Lakshmanan, L. V. 2008, October. Discovering leaders from community actions. In Proceedings of the 17th ACM conference on Information and knowledge management (pp. 499-508). ACM.

[25] Khorasgani, R. R., Chen, J., & Zaïane, O. R. 2010, July. Top leaders community detection approach in information networks. In Proceedings of the 4th Workshop on Social Network Mining and Analysis.

[26] Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006, August). Group formation in large social networks: membership, growth, and evolution. InProceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 44-54). ACM.

[27] Social Network Analysis (SNA). A tutorial on concepts and methods. Website accessed July 3, 2012 from http://www.slideshare.net/gcheliotis/social-network-analysis-3273045

[28] Multi-objective optimization: Definition, history. Website accessed July 3, 2012 from http://en.wikipedia.org/wiki/Multi-objective_optimization