

DO SOCIAL NETWORKING SITES HAVE SMALL-WORLD EFFECTS?

David Riphagen BSc
Faculty of Technology, Policy and Management, Delft University of Technology
Jaffalaan 5, 2628BX Delft, Netherlands
(Final version: January 7, 2008)

Keywords: social networking sites, small-world effects, average shortest path, clustering coefficient.

Social networking sites (SNS) are websites that have as primary function facilitating contact between their users. Their amount and the amount of users of these sites has increased significantly. This means that more Personal Identifiable Information is online and privacy threats for users increase. The way information disseminates through a SNS affects privacy threats. In this paper we examine if SNS have small-world effects, meaning that the average shortest path is low and the clustering coefficient of vertices is high. We find that social networking sites have a small average shortest path, but cannot conclude on their clustering coefficient. However, the average degree of SNS goes from 7 and up, increasing the clustering coefficient. We conclude that SNS have small-world effects, but that more empirical research is needed.

1. Introduction

The worldwide web (WWW) is rapidly becoming a social web (Fu et al.: 2007), characterized by websites that facilitate uploading, collaboration and meeting people. A rich user interface, the presence of network effects and an emphasis on decentralization complete this picture. This new web is called Web 2.0 (O'Reilly: 2005) and a major boomer is the Social Networking Site (SNS). These social networking sites differ from other websites because:

- Users of SNS can conveniently create and join communities that share the same interests and activities (Fu et al.: 2007).
- Users of SNS connect to one another by bilateral agreement (Fu et al.: 2007).
- SNS are communities characterized by sustained social interaction (Rothaermel: 2001 based on Lawrence: 1995 and Karp et al.: 1977).
- SNS are communities with their own community standards (Rothaermel: 2001 based on Lawrence: 1995 and Karp et al.: 1977).
- SNS are communities with membership rules (Rothaermel: 2001 based on Lawrence: 1995 and Karp et al.: 1977).

We define social networking sites as websites that have the primary function to facilitate social contact between their users. Users place Personal Identifiable Information (PII) in their profiles on SNS. This leads to privacy threats (Gross: 2005) and the dissemination of personal information through the network. Dissemination of personal information is one of the four categories of privacy threats identified by Solove (2007).

Users with malicious intends can gather PII to create profiles (aggregation) and sell these. Furthermore, the intends of private companies exploiting SNS are not fully clear, making the security and integrity of users' data on SNS even more unclear.

The speed of information dissemination is affected by the topology of the network. Especially, Newman (2003) mentions that the small-world effect implies that the spread of information will be fast on most real-world networks.

This paper provides a survey of the technological complexities in social networking sites from a perspective of graph theory. Social networking sites can be characterized as graphs with vertices (people) and edges (connections between people). In this paper, the terms networks and graphs are used interchangeably. The topology of a network can be characterized by the number of users, which influences the average shortest path between any nodes, and the number of possible links. The number of possible links influences the clustering coefficient. If networks have a small average shortest path and a high clustering coefficient, they show small-world effects, which means people can connect to each other within a few steps and information disseminates faster through the network.

The goal of this paper is *to assess whether social networking sites have small-world effects*. It is a first step in understanding the privacy threats for users of SNS.

In paragraph two we will give a summary of relevant graph theories for SNS. Paragraph three deals with the analysis of literature, based on empirical research of social networking sites. On basis of this literature, we present an analysis and findings in

paragraph four. Paragraph five discusses conclusions and recommendations for further research.

2. Graph theory

Social networking sites are networks with complex topologies. Because of the enormous amount of users (vertices), the number of connections between people (edges) can be enormous. Apart from the number of vertices, other measures are important for modeling the characteristics of graphs.

The average degree is the average number of edges a vertice has (Fu et a: 2007). In a directed graph, in which an edge between two vertices doesn't have to be reciprocal, the in-degree and out-degree of a node can differ significantly. Most SNS have reciprocal edges because of the consent that users need to give before others can add them. However, this is not a standard procedure in all networking sites.

All things equal, the higher the average degree is, the lower the average shortest path, defined as the mean of the geodesic distance between any pairs that have at least a path (directed chain) connecting them (Fu et al.: 2007). Barabasi (2003) defines it as the logarithm of the number of nodes divided by the logarithm of the average degree. In a famous experiment in 1967, Stanley Milgram showed that there exist 'six degrees of separation' (the average shortest path) between any two persons on earth when using only local information to contact one another (Barabasi: 2003). The diameter of a network is the maximum value of a set of shortest paths between nodes (Fu et al.: 2007).

Watts and Strogatz found another important indicator for small-world networks, a high clustering coefficient (Strogatz: 2001). By rewiring a regular

lattice graph they found that small-world networks have shortest paths between any two nodes.

The clustering coefficient measures how well the neighbors of any node in a network are locally connected (Wang et al.: 2006). Albert-Laszlo and Barabasi (2003) define the clustering coefficient as the actual numbers of edges a vertice has divided by the number of edges the same vertice could maximally have. The relations between average shortest path and clustering coefficient are shown in figure one.

The node degree is an important measurement for defining the assortativity of a graph. Newman (2002) shows that some graphs tend to have vertices with high degree that preferentially connect to other vertices with high degree. He states that if this is the case, the network is assortative mixed. If not, it is disassortative mixed.

This preferential attachment forms, together with the expansion of networks, the basis for the scale-free networks theorem of Barabasi and Albert-Laszlo (1999). They note that although networks have traditionally been described with the random graph theory of Erdos and Renyi, in reality not many networks are characterized by an equal probability of connectivity. When analyzing the collaboration of movie actors, the WWW and American power grids, Barabasi and Albert-Laszlo found that the degree distribution follows a power law tail, explaining the existence of a few vertices with a very high degree and many vertices with a small degree, shown in figure two. The random graph model has a Poisson distribution for its nodes degrees. By modeling a network according to their scale-free model and successively deleting the preferential attachment and the growth of the network, they show that these two aspects are the most important factors determining the power law tail.

Causal map of small-world effects and privacy information dissemination

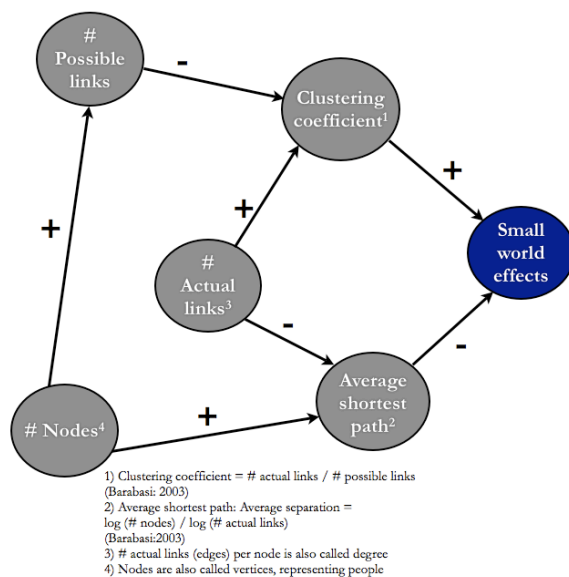


Figure 1: relationships between the indicators that influence small-world effects

We have dealt with random graph models, small-world models and scale-free models for networks. Other interesting models for modeling social networking sites are the grown graph model by Callaway et al. (2001) and the generalized random graph theory with bipartite nodes from Newman, Strogatz and Watts (2000).

The model by Callaway describes the characteristics of a growing network. This is relevant because social networking sites are growing networks. At different time steps edges are added to two vertices that are chosen uniformly at random. They conclude that a growing network tends to be connected earlier (the giant component size, a measure for connectivity, increases) than random graphs. A fully connected network has edges between all vertices. However, the giant component size takes into account that paths between vertices can travel over other vertices. Callaway mentions the preferential attachment also mentioned by Barabasi as an important factor for the growth of the network.

The generalized random graph theory with bipartite nodes concentrates on networks in which the vertices are not of the same kind: connections between directors and the boards of directors they are seated in (Newman:2001). A model is used to explain the clustering between different kind of vertices, such as the movie actors and the movies they participate in and biomedical scientists and the papers they coauthored. They find that the random graph model underestimates the clustering coefficient of such networks, because it does not identify the strong connections between different types of vertices. This could prove valuable when looking at the clustering coefficient of social networking sites.

A high clustering coefficient and a small average shortest path are indicators for small-world effects. In these small-world networks, a node can connect within a small amount of steps to any other node. In the next paragraph we will look at empirical research on different social networking sites and their findings, to examine if social networking sites show such small-world behavior.

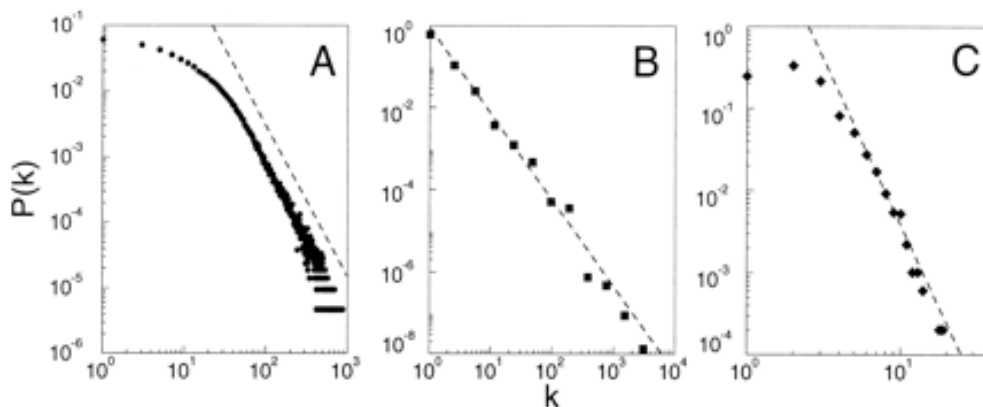


Figure 2: The distribution function of connectivities for various large networks. (A) Actor collaboration graph with $N = 212,250$ vertices and average connectivity $\langle k \rangle = 28.78$. (B) WWW, $N = 325,729$, $\langle k \rangle = 5.46$ (6). (C) Power grid data, $N = 4941$, $\langle k \rangle = 2.67$. The dashed lines have slopes (A) $\gamma_{actor} = 2.3$, (B) $\gamma_{www} = 2.1$ and (C) $\gamma_{power} = 4$ (from Barabasi and Albert: 1999).

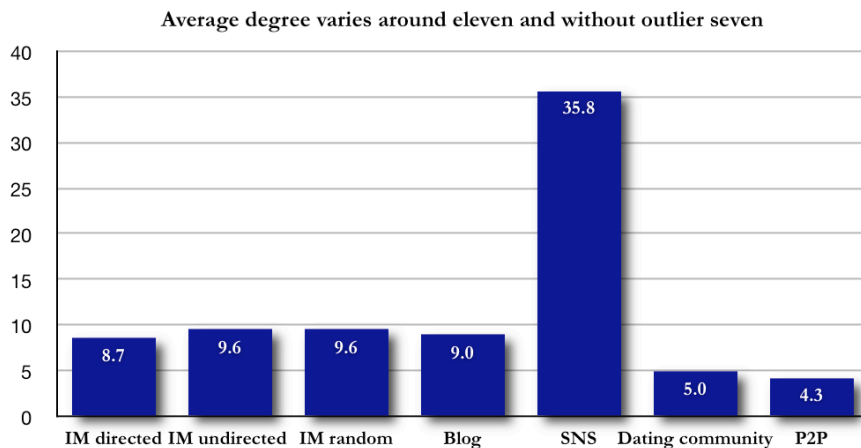


Figure 3: average degree of examined social networking sites

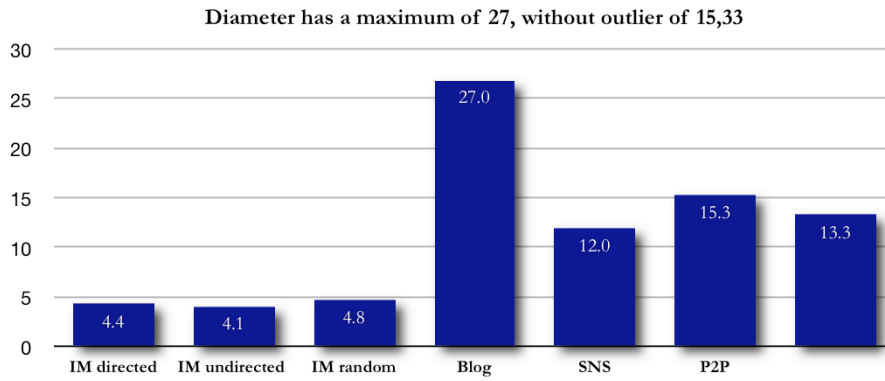


Figure 4: diameter of examined social networking sites

Subject	# vertices	Average degree (k)	Diameter (D)	Average shortest path (l)	Clustering coefficient	Assortativity
Instant messaging (IM) directed (Smith)	50158	8.651 ¹	4.35	N/A	N/A ⁷	N/A
IM undirected (Smith)	50158	9.6	4.1	N/A	0.33	N/A
IM random graph (Smith)	50158	9.6	4.79	N/A	1.9x10 ⁻⁴	N/A
Online social networks blog (Fu et al.)	50158	9	27	6.84	0.149	Disassortative
Online social networks SNS (Fu et al.)	50158	35.8	12	3.72	0.16	Assortative
Internet dating community (Holme et al.)	± 29341 ^{4,5}	± 5.0 ^{5,6}	N/A	4.4 ⁵	0.006 ⁵	Dissortative
Peer-to-peer social networks (Wang et al.)	42186 - 191679 ²	4.3	11, 16, 19 ³	6.6	0.02	Dissortative
P2P social networks major nodes (Wang et al.)	221 - 960 ²	N/A	9, 14, 17 ³	4.6	0.1	N/A

Table 1: statistical measurements for SNS from various literature sources.

- (1): calculated as average from in-degree and out-degree from Smith (2002), links are bi-directional?
- (2): three networks were constructed in three trials, with a different amount of nodes (Wang et al.: 2006)
- (3): three networks were constructed in three trials, with different diameters (Wang et al.: 2006)
- (4): amount of users increased during 500 days of research (Holme et al.: 2004)
- (5): values only used of all contact connections (Holme et al.: 2004)
- (6): average degree converges towards 5 in 500 days, no cutoff to degree distribution (Holme et al.: 2004)
- (7): not calculated, see Smith: 2006

3. Analysis of literature on characteristics of Social Networking Sites

In this paragraph we provide a survey of literature on the topology of social networking sites and specifically look for indicators of small-world effects. We focus on literature that is based on empirical research of SNS, in other words research on data gathered from real-life networks. A summary of our finding is displayed in table one.

We start at the most recent publication known to the author on this subject from Fu, Liu & Wang (2007). They conclude that the studied networks possess small-world and scale-free features. They provide an empirical analysis of two online social networks: the Chinese weblog Sina blog and the Chinese Xiaonei social networking site. By looking at their references, the empirical research of Holme et al. (2004) on the Swedish Internet dating community pussokram.com was found and analyzed. Holme used logged data of 512 days to study the evolution of the Internet dating community.

Analysis of their argumentation pointed to the paper of Rothaermel and Sugiyama(2001), which looks at a specific virtual Internet community not from a topological, but from a user's point of view. Because Rothaermel doesn't focus on topology, his findings have not been taken into account for the quantitative analysis in this paper.

Holme (2004) also mentions the work of Smith (2002) on instant messaging as a scale-free network. Smith has used data of the nioki.com French instant messaging network that uses the Jabber protocol. By analyzing the raw data from the nioki.com database, he finds that the networks in-degree, out-degree and undirected edges degree follow a power law.

The paper of Fu, Liu & Wang also mentions the work of Wang, Moreno and Sun (2006). They have examined the P2P social network of the Gnutella file exchange network, by gathering data with a Gnutella client and analyzing the gathered data. They have conducted three experiments of data gathering, resulting in three different networks and extracted one additional network per network based on the major nodes (vertices with a very high degree) of these networks. They have gathered a substantial amount of data to base their analysis on. They conclude that the examined P2P social networks are scale-free, small-world networks that show disassortative mixed behavior.

For relevant literature on graph theory, the book by Barabasi (2003) and the article of Fu, Liu & Wang (2007) have been examined.

Table one on summarizes the findings from the different empirical research examined for this paper.

In the next paragraph, we will look at the data gathered from the different literature and present findings.

4. Findings

After analyzing the data shown in table in the previous paragraph, the following findings can be presented:

- The number of vertices found in social networking sites is large. If the subgraph of major nodes constructed by Wang et al. is left out, the minimum amount of vertices is 29341 and the maximum 105147. Because the subgraph of major nodes constructed by Wang is a subset with only the vertices with a high degree, it is not representative for the amount of vertices in SNS and therefore left out.
- The average degree varies around nine for instant messaging and blogs, shown in figure 3. For the Internet dating community and the peer to peer social networks the average degree is much lower, around 4 or 5. The online social networking site examined by Fu et al. proves to be a significant outlier, with an average degree of 35,8. With the social networking site included in the data, the average of the average degrees is 11,71 and when excluding the outlier it is 7,69. The social networking site examined by Fu et al. comes close to our definition of a social networking site, together with the Internet dating community examined by Holme. It is recommended to take a closer look at the differences between these networks and find explanations for the great difference in their average degree.
- The diameter, defined as the maximum of a set of shortest paths in a network, has a minimum of 4,1 and a maximum of 27, shown in figure 4. However, as shown in the graph, the diameter of the blogs is a significant outlier. Because the diameter is the maximum of the shortest paths, the most important measure will be the maximum of the diameters shown. Without the blog's diameter, this is 15,33, with a standard deviation of 5,1.
- The average shortest path of the social networking sites examined varies around 5,38 with a standard deviation of 1,3. This shows that the social networking sites resemble the small-world model in the respect of having small average shortest paths.
- However a high clustering coefficient is significant to determine if the network has small-world effects. If the clustering coefficient of a network is significantly higher than the clustering coefficient of the same network modeled as a

random graph, the precondition for a small-world model is satisfied (Wikipedia: 2007 and Barabasi: 2003). It is difficult to draw conclusions on the clustering coefficient of social networking sites because the standard deviation of the indicators is fairly high compared with the range and the only network for which the random graph clustering coefficient is calculated is the instant messaging network. The latter does have a significant higher clustering coefficient than its random graph representation. The Chinese blog and social networking site tend to have a higher clustering coefficient. The P2P network has a low clustering coefficient and the dating community doesn't show any clustering at all. It is recommended to create the appropriate random graphs for the networks mentioned and compare the clustering coefficients to determine if the small-world model fits.

- Two researchers conclude explicitly that the networks they have examined have small-world effects: Fu et al. and Wang et al. The undirected Instant Messaging network from Smith resembles SNS with bilateral links and shows a small average shortest path and high clustering coefficient, pointing towards small-world effects. This are three significant indications of small-world effects in social networking sites.
- Nothing can be concluded about the assortativity of social networking sites. Three of the investigated networks score disassortive, however this is not significant on a total of four networks examined, because an error on one of these analysis changes this picture dramatically. Interesting in this respect is the article of Newman (2002). In this article, Newman states that social networks are assortatively mixed, while technological and biological networks tend to be disassortatively mixed. Fu et al. (2007) reason that although Internet dating communities, blogs and P2P networks are social networks, they are embedded in technological networks and therefore show disassortatively mixed behavior. It is recommended to look deeper into the assortativity of social networking sites, because it can give significant insight in how these networks grow.

The graph theory in paragraph two is the framework to assess our findings with. Until now, it seems that social networking sites possess small-world effects. This means that most pairs of vertices are connected by a short path through the social networking site (Newman: 2003). This affects the speed of information dissemination through the network and increases privacy threats. In the next paragraph we come up with conclusions and recommendations for further research.

5. Conclusions and recommendations

We have looked at the topology of several online social networking sites to assess whether social networking sites have small world-effects. These small-world effects affect information dissemination and per definition of Solove (2007) the privacy of users. We assume that small-world effects increase the probability of privacy breaches by accelerating the flow of information through the network.

Our research is based on a survey of relevant literature and simple statistical analysis of their findings. From this analysis we can conclude the following:

Social networking sites have large numbers of users. The average connections that these users has goes from 7 and up. The maximum of shortest paths in the social networking sites examined is 15,33.

If we look at the average shortest path of the social networking sites, they tend to show small-world effects, with an average shortest path of 5,38 and a standard deviation of 1,3, almost similar to the 6 degrees of freedom of Milgram.

However, to verify the small world effects the networks also need to have a higher clustering coefficient than the same network modeled as a random graph. This holds only for the instant messaging network of Smith. Fu et al. and Wang et al. conclude specifically that the networks under their investigation have small-world effects.

We can conclude that social networking sites have small-world effects, meaning that the most pair of vertices in the network seem to be connected by a short path through the social networking site (Newman: 2003)

More research is needed on the behavior of social networking sites to come up with conclusions about privacy threats for users of social networking sites. Therefore we make the following recommendations:

- Our definition of social networking sites, given in paragraph one, comes closest to those examined by Fu et al. and by Holme. However, the average degree of the SNS they examined varies significantly and we recommend to examine what is the cause of this.
- In the literature that is surveyed there is no real distinction between users that actively use their profile and those that don't.
 - The research of Wang (2006) shows that active users have a much shorter average shortest path.
 - Furthermore, Amaral (2000) shows that networks with aging nodes (nodes that become inactive after a while) provides a cutoff in the power-law regime for the degree distribution. This means that nodes with a high degree become more sparse. We recommend to do

further research on the differences in topology of networks with and without less active users.

- The generalized random graph theory with bipartite nodes of Newman (2001) could explain this.
- Fu et al. (2007) reason that although Internet dating communities, blogs and P2P networks are social networks, they are embedded in technological networks and therefore show disassortatively mixed behavior. It is recommended to look deeper into the assortativity of social networking sites, because it can give significant insight in how these networks grow.
- To examine the small world effect in SNS, it is recommended to use random graphs with the same amount of vertices and edges as the networks examined here and compare their clustering coefficient with the original networks' clustering coefficient.
- Holme et al. do not find indicators of small-world effects in the Internet dating community they examined. However, he does find a small shortest path, indicating the probability of small-world effects. Explanations for the low clustering coefficient are a different way of calculating the clustering coefficient, measurement of different types of contact than reciprocal agreement and the sparsity of a starting dating community. However, Holme et al. state that the clustering coefficient converges to a finite value in 500 days. We recommend to execute further research on this project.
- Analysis of the topology is a first step in understanding the privacy threats for users of social networking sites. Small-world effects provide insight in the speed of information dissemination. We recommend further research on social and economic aspects of SNS that influence the privacy of users.

The research on online social networking sites, their topology and the influence they have on society is an emerging scientific field. With the conclusion that social networking sites have small world effects we've made a first step for understanding the impact that these networks have on our privacy and future.

References

- ALBERT, R. & BARABÁSI, A.-L. (2002) Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74, 47.
- AMARAL, L. A. N., SCALA, A., BARTHÉLÉMY, M. & STANLEY, H. E. (2005) Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 102, 10421-10426.

BARABASI, A. L. (2003) *Linked: How Everything Is Connected to Everything Else and What It Means*.

BARABASI, A. L. & ALBERT, R. (1999) Emergence of scaling in random networks. *Science*, 286, 509-512.

CALLAWAY, D. S., HOPCROFT, J. E., KLEINBERG, J. M., NEWMAN, M. E. J. & STROGATZ, S. H. (2001) Are randomly grown graphs really random? *Physical Review E*, 64, 041902.

FU, F., LIU, L. & WANG, L. (2007) Empirical analysis of online social networks in the age of Web 2.0. *Physica A: Statistical Mechanics and its Applications*, In Press, Corrected Proof.

GROSS, R., ACQUISTI, A. & HEINZ III, H. J. (2005) Information revelation and privacy in online social networks. *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, 71-80.

HEER, J., BOYD, D. (2005) Vizster: Visualizing Online Social Networks.

HOLME, P., EDLING, C. R. & LILJEROS, F. (2004) Structure and time evolution of an Internet dating community. *Social Networks*, 26, 155-174.

KARP, D. A., STONE, G. P. & YOELS, W. C. (1977) *Being Urban: A Social Psychological View of City Life*, Heath.

LAWRENCE, T. B. (1995) Power and resources in an organizational community. *Academy of Management Best Papers Proceedings*, 251-255.

NEWMAN, M. E. J. (2001) Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64, 16131.

NEWMAN, M. E. J. (2002) Assortative Mixing in Networks. *Physical Review Letters*, 89, 208701.

NEWMAN, M. E. J. (2003) The Structure and Function of Complex Networks. *SIAM Review*, 45, 167-256.

NEWMAN, M. E. J., STROGATZ, S. H. & WATTS, D. J. (2001) Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64, 026118.

O'REILLY, T. (2005) What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. Retrieved March, 1, 2007.

ROTHAERMEL, F. T. & SUGIYAMA, S. (2001) Virtual internet communities and commercial success: individual and community-level theory grounded in the atypical case of TimeZone.com. *Journal of Management*, 27, 297-312.

SMITH, R. D. (2002) Instant Messaging as a Scale-Free Network. Unpublished.

SOLOVE, D. J. (2007) 'I've Got Nothing to Hide' and Other Misunderstandings of Privacy. *San Diego Law Review*, Vol 44.

STROGATZ, S. H. (2001) Exploring complex networks. *Nature*, 410, 268-276.

WANG, F., MORENO, Y. & SUN, Y. (2006) Structure of peer-to-peer social networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 73, 036123-7.

WATTS, D. J. & STROGATZ, S. H. (1998) Collective dynamics of 'small-world' networks. *Nature*, 393, 409-10.

WIKIPEDIA, C. Social network service Retrieved 10 November 2007 12:49 UTC-from http://en.wikipedia.org/w/index.php?title=Social_network_service&oldid=170426240

WIKIPEDIA, C. List of social networking websites Retrieved 10 November 2007 13:06 UTC-from http://en.wikipedia.org/w/index.php?title=List_of_social_networking_websites&oldid=170450717

WIKIPEDIA, C. Small-world network Retrieved 22 November 2007 11:52 UTC-from http://en.wikipedia.org/w/index.php?title=Small-world_network&oldid=171947076

WIKIPEDIA, C. Clustering coefficient Retrieved 30 November 2007 15:20 UTC-from http://en.wikipedia.org/w/index.php?title=Clustering_coefficient&oldid=172795865