This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE Globecom 2011 proceedings.

Analysis of Privacy in Online Social Networks from the Graph Theory Perspective

Leucio Antonio Cutillo, Refik Molva, Melek Önen
EURECOM
Sophia-Antipolis, France

***Abstract*—The extremely widespread adoption of Online Social Networks (OSNs) raises many questions on privacy and access control. Regardless of the particular centralized or de-centralized nature of the OSN, the achievable security and privacy degree strongly depends on the graph-theoretical properties of the *social graph* representing the real friendship relations between the users. In this paper, we analyze the relationship between the social network graph topology and the achievable privacy. We observe three metrics, namely degree distribution, clustering coefficient and mixing time, and show that they give fundamental insights on the privacy degree of the OSN. We propose how to exploit these insight for the design of future privacy-friendly OSN.**

I. Introduction

Online Social Networks (OSNs) can be seen as means to share and discover content generated by users and intended to be consumed by friends. Their extremely widespread adoption raises many questions about access and diclosure policies; news about privacy issues such as companies checking out job candidates[1], hackers blackmailing Social Network Services (SNS) providers[2], insurances cutting benefits to customers[3] are extremely frequent in the media, nowadays.

Unfortunately, privacy protection solutions offered by any existing OSN applications are revealed to be unsatisfactory no matter their robustness. Decentralized OSNs (e.g. [1], [2], [3]) attempt to remedy to this problem by avoiding the adoption of any omniscient entity that can directly manage and misuse the user data and propose an infrastructure for user data management and storage that is distributed (often based on a peer-to-peer architecture). Such solutions still present some weaknesses in terms of privacy.

In this paper, we show that the privacy degree of an OSN application, be it centralized or de-centralized, strongly depends on the topological properties of the *Social Graph* which represents friendships (trust relationships) between actual OSN users. Social graph analysis has already proved its importance for studies such as sociology [4], [5] and network performance[6].

In this paper a similar analysis is driven with respect to the impact of social network topology on privacy. We show that there exists a strong relationship between a set of metrics and privacy properties. More specifically, three graph metrics, namely the node degree, the clustering coefficient and the mixing time, give fundamental insights on the privacy degree of the resulting OSN.

In Section II, we discuss critical privacy and security requirements for online social networks; in Section III, we draw a link between these properties and several metrics of the social network graph and we analyze these metrics in various large-scale social network dumps, giving several hints for any design of future privacy-friendly OSNs. Finally, we give an overview of the related work in section IV.

II. Problem Statement

Due to the huge amount of sensitive data that can easily be gathered, stored, replicated and correlated[4] [7], [8], [9], privacy protection becomes one of the main objectives for services provided by an OSN platform [10], [11].

Privacy is a relatively new concept, being shaped by the capability of new technologies to share information. Conceived as 'the right to be left alone'[12] during the period of newspapers and photographs growth, privacy now refers to the ability of an individual to control and selectively disclose information about him, and its importance is so relevant to have been reported in the Universal Declaration of Human Rights (art.12):

> *"No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such interference or attacks."*

The privacy problem can first be analyzed with respect to whom users are protected from. Indeed, as in current solutions, users can choose the privacy degree they would like to achieve with respect to some categorized users. Their friends have more privileges in terms of access control than their co-workers for example. In this case, the main attack considered here against privacy is the ***disclosure of sensitive data***. Users do not want to reveal their confidential information to other unknown (or even known) users. Recently, researchers also considered that the OSN provider can be a potential threat since it has control over all users' data; taking an approach that is radically different from the one of commercial OSN applications, researchers recently proposed to design the OSN

[1] http://www.firedfornow.com/job-loss-and-the-economy/can-facebook-hurt-your-job-prospects/

[2] http://eu.techcrunch.com/2009/10/21/hacker-arrested-for-blackmailing-studivz-and-other-social-networks/

[3] http://www.cbc.ca/canada/montreal/story/2009/11/19/quebec-facebook-sick-leave-benefits.html

[4] http://www.sophos.com/pressoffice/news/articles/2007/08/facebook.html

978-1-4244-9268-8/11/$26.00 ©2011 IEEE

application based on a peer-to-peer architecture [13] in order to avoid centralized control over users' data. In such networks, since the data storage application is distributed, the disclosure probability may become more important. Furthermore some solutions such as [13] leverages social links to construct the OSN: users' friends store their data in order to ensure data availability. Therefore, data confidentiality and access control are one of the main privacy goals that OSN solutions, be centralized or decentralized, have to be taken into consideration starting from their very design. An additional aspect that also needs to be considered is taking a step ahead from access control. Indeed, even with a very strong and specific access control policy where the user that uploads or "***posts***" a content can prevent unauthorized access, he unfortunately looses the control on it after its very first publication. This issue defined as *usage control* [14] can have a serious impact on users' privacy as well.

Moreover, in addition to basic data confidentiality, a curious adversary should not be able to gather any information on the history of actions performed by arbitrary users in the system. This calls for the property of untraceability[15], and is a mandatory requirement for privacy preserving OSNs. Therefore, usage control, data confidentiality and untraceability are the main properties that OSN solutions should ensure.

Furthermore, recently ***Sybil attacks*** have been considered to be another type of serious threats against users' privacy[5,6]. The purpose of Sybil attacks is to generate a very large number of fake identities (profiles) and try to establish friendships with legitimate users in order to gather sensitive information from them or distribute spam messages[7] or malware[8]. Sybil attacks can also have severe impact on decentralized solutions since malicious nodes can perform many types of Denial of Service attacks such as Eclipse [16] or other kinds of attacks on the underlying routing protocol against a selected victim.

Finally, another aspect that can be considered as a major privacy problem is identity and friendship privacy. Indeed, such information became very valuable for centralized OSN solutions. Furthermore, some new decentralized solutions also leverage the social links to construct the network itself. Therefore, the network itself can inherently reveal information on users' friends.

In the next section, we analyze the main problems described above with respect to the underlying social graphs: indeed, the impact of malicious attacks can be evaluated based on the underlying social graph's characteristics. Furthermore, since some of the privacy preserving solutions leverage the social links among users, such architectures can be a potential threat against privacy. Social graph analysis can inherently be a useful tool to prevent such problems.

[5]http://www.nature.com/news/2009/090423/full/news.2009.398.html

[6]http://www.sophos.com/pressoffice/news/articles/2009/12/facebook.html

[7]http://www.pcworld.com/businesscenter/article/191847/facebook_users_targeted_in_massive_spam_run.html

[8]http://content.usatoday.com/communities/technologylive/post/2009/12/koobface-compels-facebook-victims-to-help-spread-worm-/1

III. Impact of Social Graph Topology on Privacy

A. Social graph topology

An Online Social Network can be represented as an undirected ***social graph*** $G(V, E)$ comprising a set V of users and a set E of edges representing social ties, such as friendship, kinship, trust and the like. Graph theory was very useful for many interesting fields such as networking and sociology: graphs are used to represent communication networks or social networks and the analysis of some of their basic properties can help on evaluating and improving the performance of networking solutions.

In this paper, three different characteristics are analyzed, namely, the ***node degree***, the ***clustering coefficient***, and the ***mixing time***. The impact of the evolution of these parameters is also evaluated based on existing social graphs: in September 2005, Facebook published anonymous social graphs of 5 universities in the United States[9]: California Institute of Technology (Caltech), Princeton University (Princeton), Georgetown University (Georgetown), University of North Carolina (UNC), Oklahoma University (Oklahoma). Each graph is represented by an adjacency matrix A whose non diagonal elements a_{ij} are set to one if user $\nu_i \in V$ is a friend of user $\nu_j \in V$, or zero otherwise. As each adjacency matrix is symmetric, the represented social graph is undirected.

B. Node degree

In graph theory, the degree of a vertex, denoted by $deg(\nu)$ is defined as the number of edges incident to the vertex. Since in a social graph $G(V, E)$, a vertex represents a user and the edges represent friendship links, a user's degree defines the number of friends a user has. This degree shows a straightforward relationship with privacy since when ν establishes a relationship with a new friend, with the increase of the degree, the probability of connecting to a misbehaving user increases.

Different studies have shown that participants clearly represent a weak link for security in OSNs and are vulnerable to a series of social engineering attacks [7], [8], [9], often caused by a lack of awareness regarding the consequence of simple actions like accepting contact requests.

Assume p_{mal} denotes the probability a new friend η of ν is a malicious user, and assume the events of befriending a malicious user are independent. The number of malicious friends $F_{mal}(\nu)$ of ν then follows a binomial distribution:

$$\mathcal{F}_{mal}(\nu) \sim B(p_{mal}, deg(\nu))$$

In particular, the probability p_ν of having at least one misbehaving friend is:

$$p_\nu = 1 - p_{mal}^{deg(\nu)} \quad (1)$$

Once a malicious η gets access to ν's sensitive data, η can disclose them out of band, or inside the social network itself. In this latter case, the disclosure targets, among all η's friends, the common friends between η and ν, and can turn out to severely damage ν.

[9]http://people.maths.ox.ac.uk/ porterm/data/facebook5.zip

Therefore, the outdegree of a node is directly related with usage control. The more a node has friends, the larger the probability of having a malicious friend which can disclose sensitive personal data. Furthermore, if the correct execution of OSN applications such as in [2] depends on the structure of the social graph, the probability of discovering nodes' friends increases with the malicious behavior as well.

Figure 1 shows the distribution of the node degree for the five Facebook datasets. In this figure, the degree of the Caltech social network is much lower than the degree of the other four social networks. This is probably due to the fact that the Caltech dataset is significantly smaller than the others, and, as a consequence, the opportunities to add friends are lower. If we assume p_{mal} as constant for all the graphs, by applying eq.1, we observe that the probability of having at least a misbehaving contact is lower in the Caltech network. In Caltech, in fact, when p_{mal} is set to 0.01, p_ν is on average as high as 0.35, while in the other networks this value ranges from 0.59 to 0.64.

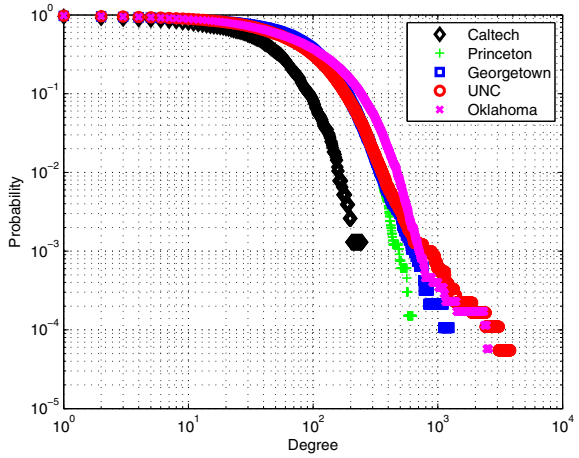

Fig. 1. Log-log plot of the degree complementary cumulative distribution of real-life social networks, from [17].

C. Clustering Coefficient

In an undirected graph, the clustering coefficient $c(\nu)$ of a node ν having $deg(\nu)$ edges is defined as the number of existing links between these nodes, denoted as $e_{deg(\nu)}$, divided by the number of possible links that could exist $(\frac{deg(\nu)(deg(\nu)-1)}{2})$. We therefore have:

$$c(\nu) = \frac{2e_{deg(\nu)}}{deg(\nu)(deg(\nu)-1)} \tag{2}$$

The clustering coefficient of the overall graph, denoted as $C(G)$, is the average clustering coefficient of all its nodes, hence:

$$C(G) = \frac{\sum_{\nu \in V} c(\nu)}{\|V\|} \tag{3}$$

Knowing or estimating the clustering coefficient of a graph can give an idea on the impact of a malicious node whenever it has information on nodes friendship and can further disclose it. Once a malicious node, η, is added in the contact list of ν, η can access ν's sensitive data, and disclose it indiscriminately using the social network facilities like wall posting, picture publishing and the like. In particular, if η sets up a profile corresponding to a real identity ν strongly trusts, η could access a portion of data outside the visibility of the other friends of ν. In this case, the disclosure of ν's sensitive data to the common contacts between ν and η can strongly be dangerous.

We can then evaluate in a first approximation the average ratio $\mathcal{Q}_\nu$ of ν's friends that can obtain sensitive information disclosed by a malicious η as

$$\mathcal{Q}_\nu = p_\nu c(\nu) \tag{4}$$

Therefore, similarly to the outdegree, the clustering coefficient has a direct effect on usage control. The tighter the friendset, the broader the disclosure of sensitive data to the user's contacts.

Figure 2 shows the distribution of the clustering coefficient for the different social networks that were previously introduced. Similarly to the previous analysis, the clustering coefficient of the Caltech social network strongly differs from those of other networks, as it is almost twice in size. This is probably again due to the small size of the Caltech dataset. A smaller community is in fact more likely to be tightly knit.

We then observe that in the case a friend misbehaves, the victim in Caltech exposes his sensitive data to a ratio of friends two times higher compared with the one of a victim in the other networks. Nevertheless, due to the lower p_ν, the average ratio $\mathcal{Q}_\nu$ does not strongly vary in the different networks, ranging from 0.11 to 0.14.

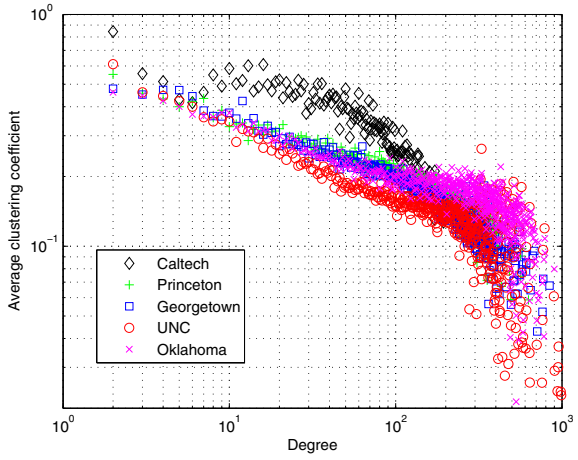

Fig. 2. Average clustering coefficient of real-life social networks with respect to node degree, from [17].

D. Mixing time

Random walks [18] in a graph have an important property: when the random walk approximates its steady state distribution after a sufficient number of hops, the startpoint and

endpoint of the walk are uncorrelated. This number of hops is called ***mixing time***, and the smaller it is, the faster the abovementioned property is met.
We will introduce the mixing time starting from the steady state distribution.
The ***steady state distribution*** for a node θ represents the probability this random walk reaches θ after a sufficient number of hops, and does not depend on the node where this random walk originated from:

$$ssd(\theta) = \frac{deg(\theta)}{2\|E\|} \tag{5}$$

The mixing time [18] $\tau_x(\epsilon)$ is then computed as:

$$\tau_x(\epsilon) = min\{h : \Delta_x(h) \leq \epsilon\} \tag{6}$$

where $\Delta_x(h)$ is the variation distance between the random walk distribution $R^h(x)$ after h hops, and the steady state distribution $ssd(x)$:

$$\Delta_x(h) = \|R^h - ssd\| = \frac{1}{2}\sum_{x \in V} \|R^h(x) - ssd(x)\| \tag{7}$$

For the whole network, the mixing time is:

$$\tau(\epsilon) = \max_{x \in V} \tau_x(\epsilon) \tag{8}$$

When a network is fast mixing, $\tau(\epsilon)$ is $O(log\|V\|)$.

In social networks, mixing time is varying widely: in [19] authors found that mixing time is much higher in social networks where links represent face-to-face interactions. Recently, further measurements [20] confirmed this concept.

Two solutions in the literature have already been proposed to exploit the topology properties of the social network itself to increase users' privacy and security and both of them rely on random walks through the social network graph. The two applications are using social links: the first one, Safebook [2], to forward data, analogously to what is done through a mix network [21], but with the advantage of exploiting trust between actors; the second one, Sybilguard [22], to defend against Sybil attacks. In both cases, a small mixing time is required to increase the security and privacy performance. Therefore, the mixing time of a social network graph is directly related with both profile integrity and communication untraceability. Figure 3 plots the mixing time $\tau(\epsilon)$ of each of the five Facebook social graphs for different values of a predefined maximum variation distance ϵ. As the Caltech network presents a faster mixing time, solutions like Safebook or Sybilguard would perform better if applied on this social network rather than in the Georgetown one, whose mixing time is always roughly five times higher.

E. Summary

Table I summarizes the main topological properties of the examined social network dumps, where the probability p_{mal} of befriending a misbehaving user is set to 0.01. In this scenario, even if the average p_ν is high (ranging from 0.35 to 0.64), the UNC network ensures the best privacy protection (in terms of anonymity and usage control) with respect to the other networks because it shows the lowest average value for Q_ν. In terms of communication untraceability and profile integrity, the Caltech network provides the best protection due to the faster mixing time.

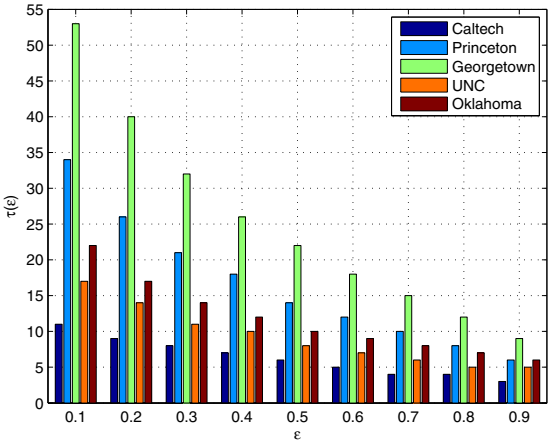

Fig. 3. Mixing time of real-life social networks.

	$\|V\|$	$\overline{deg(\nu)}$	$C(G)$	$\overline{p_\nu}$	$\overline{Q_\nu}$	$\tau(0.1)$
Caltech	769	43.32	0.41	0.35	0.14	11
Princeton	6596	88.93	0.24	0.59	0.14	34
Georgetown	9414	90.43	0.22	0.60	0.13	53
UNC	18163	84.44	0.20	0.57	0.11	17
Oklahoma	17425	102.44	0.22	0.64	0.14	22

TABLE I
MAIN CHARACTERISTICS OF FIVE SOCIAL GRAPHS FROM FACEBOOK ($\overline{p_\nu}$ COMPUTED ASSUMING p_{mal}=0.01).

IV. RELATED WORK

Many of the properties of online social network graphs have been studied by both sociologists and computer scientists [4], [5]. Recent studies [6] confirmed that social network graphs are ***scale-free*** [23], i.e. their degree follows a power-law distribution: the probability that a node has degree k is proportional to $k^{-\gamma}$ and high-degree nodes tend to be connected to other high-degree nodes. Another study shows that social networks also exhibit ***small-world*** [24] behavior: the maximum shortest path between any pair of nodes in the network is on the order of the logarithm of the number of nodes in the network. Furthermore, social networks are nowadays used as a powerful tool to disseminate information. The study of topological properties helps researchers to design and implement new dissemination or lookup protocols [25]. Moreover, topological properties of social graphs have also been studied in the context of trust/security. For example, in [26], friend-of-friend relationships among users are exploited to accept email from authorized senders. Authors in [27] observe a severe impact of worm propagation in mobile phone

networks with the help of a simulator modelling the network determined by the cell phone address books. Sybilguard [22] limits the influence of sybil attacks by bounding the number and size of sybil groups, detected by the high number of connections between sybils compared to the low number of connections between sybil and genuine nodes.

Compared with these solutions, to the best of our knowledge, our work can be considered as the first one that specifically analyses the privacy problem together with the social network topology. We believe that the undertaken study would help scientists on the design of new privacy preserving online social networks.

V. Conclusion and Future Work

This paper investigates the strong relationship between the topological properties of the social network graph and the achievable users' privacy in centralized or decentralized OSN. We observe that metrics such as the degree and the clustering coefficient of nodes severely affect users' privacy with respect to identity/friendship privacy and usage control, while the mixing time of random walks in the social network graph plays an essential role in preserving the users' communication untraceability.

An analysis on real social network dumps reveals the probability of befriending at least a misbehaving contact is not negligible. In this case, the number of nodes the stolen sensitive data can reach depends on the number of common friends between the victim and the attacker.

Privacy preserving OSN architectures should address this problem by discouraging the indiscriminate action of adding friends. Moreover, when providing communication obfuscation and identifiers integrity through random walk on the social network graph, the OSN should guarantee the fast mixing property to the network. This can be done by ensuring the small wolrd property of the social network graph, and encouraging "long links" connecting different clusters together, otherwise most of the random walk would be confined to the originating cluster.

As a future work, we intend to study the impact of additional graph properties, such as the assortativity and the betweenness, on user privacy.

Acknowledgment

This work has been supported by the SOCIALNETS project, grant agreement number 217141, funded by the EC seventh framework programme theme FP7-ICT-2007-8.2 for Pervasive Adaptation. See http://www.social-nets.eu/ for further details.

References

[1] S. Buchegger, D. Schiöberg, L. H. Vu, and A. Datta, "PeerSoN: P2P Social Networking," in *Social Network Systems*, 2009.

[2] L. A. Cutillo, R. Molva, and T. Strufe, "Safebook : a privacy preserving online social network leveraging on real-life trust," *"IEEE Communications Magazine", Consumer Communications and Networking*, 2009.

[3] A. Shakimov, A. Varshavsky, L. P. Cox, and R. Cáceres, "Privacy, cost, and availability tradeoffs in decentralized osns," in *Proceedings of the 2nd ACM workshop on Online social networks*, ser. WOSN '09, 2009.

[4] S. Milgram, "The Small World Problem," *Psychology Today*, vol. 2, pp. 60–67, 1967.

[5] M. S. Granovetter, "The Strength of Weak Ties," *The American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[6] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks," in *Proceeding of the 5ht ACM/USENIX Internet Measurement Conference (IMC'07)*, October 2007.

[7] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, "All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks," 2008, wWW 2009, Madrid.

[8] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, "Social phishing," *Communications of the ACM*, vol. 50, no. 10, pp. 94–100, 2007.

[9] S. Moyer and N. Hamiel, "Satan is on My Friends List: Attacking Social Networks," http://www.blackhat.com/html/bh-usa-08/bh-usa-08-archive.html, 2008.

[10] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *ACM Workshop on Privacy in the Electronic Society*, 2005, pp. 71 – 80.

[11] danah m. boyd , "Facebook's privacy trainwreck," *Convergence: The International Journal of Research into New Media Technologies*, vol. 14(1), pp. 13 – 20, 2008.

[12] S. D. Warren and L. D. Brandeis, "The right to privacy," *Harward Law Review*, vol. 4, no. 5, pp. 193–220, December 1890.

[13] L.-A. Cutillo, R. Molva, and T. Strufe, "Privacy preserving social networking through decentralization," in *IEEE WONS*, 2009.

[14] J. Park and R. Sandhu, "Towards usage control models: beyond traditional access control," in *SACMAT '02: Proceedings of the seventh ACM symposium on Access control models and technologies*. New York, NY, USA: ACM, 2002, pp. 57–64.

[15] L. A. Cutillo, M. Manulis, and T. Strufe, "Security and Privacy in Online Social Networks." Chapter book of "Handbook of Social Network, Technologies and Applications", Springer, October 2010, ISBN: 978-1-4419-7141-8.

[16] *Eclipse Attacks on Overlay Networks: Threats and Defenses*, 2006. [Online]. Available: http://dx.doi.org/10.1109/INFOCOM.2006.231

[17] L. A. Cutillo, R. Molva, and M. Önen, "Performance and Privacy Tradeoff in Peer-to-Peer On-line Social Networks," 2010, technical Report RR10244.

[18] M. Mitzenmacher and E. Upfal, *Probability and Computing : Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, January 2005.

[19] M. Dell amico and Y. Roudier, "A measurement of mixing time in social networks," in *STM 2009, 5th International Workshop on Security and Trust Management, September 24-25, 2009, Saint Malo, France*.

[20] A. Mohaisen, A. Yun, and Y. Kim, "Measuring the mixing time of social graphs," *10th USENIX/ACM SIGCOMM Internet Measurement Conference (IMC'10)*, 2010.

[21] D. Chaum, C. O. T. Acm, R. Rivest, and D. L. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Communications of the ACM*, vol. 24, pp. 84–88, 1981.

[22] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman, "Sybilguard: Defending against sybil attacks via social networks," in *In ACM SIGCOMM 06*. ACM Press, 2006, pp. 267–278.

[23] L. Li, D. Alderson, R. Tanaka, J. C. Doyle, and W. Willinger, "Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications (Extended Version)," Oct. 2005.

[24] J. Kleinberg, "The small-world phenomenon: an algorithm perspective," in *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, ser. STOC '00. New York, NY, USA: ACM, 2000.

[25] A. Mislove, "Online social networks: Measurement, analysis, and applications to distributed information systems," Ph.D. dissertation, Rice University, Department of Computer Science, May 2009.

[26] S. Garriss, M. Kaminsky, M. J. Freedman, B. Karp, D. Mazires, and H. Yu, "Re: Reliable email," in *In Proc. NSDI*, 2006, pp. 297–310.

[27] C. Fleizach, M. Liljenstam, P. Johansson, G. M. Voelker, and A. Mehes, "Can you infect me now?: malware propagation in mobile phone networks," in *Proceedings of the 2007 ACM workshop on Recurring malcode*, ser. WORM '07. New York, NY, USA: ACM, 2007.