Prying Data out of a Social Network

Joseph Bonneau, Jonathan Anderson
Computer Laboratory
University of Cambridge
jcb82@cl.cam.ac.uk,jra40@cl.cam.ac.uk

George Danezis
Microsoft Research
Cambridge, UK
gdane@microsoft.com

***Abstract*—Preventing adversaries from compiling significant amounts of user data is a major challenge for social network operators. We examine the difficulty of collecting profile and graph information from the popular social networking website Facebook and report two major findings. First, we describe several novel ways in which data can be extracted by third parties. Second, we demonstrate the efficiency of these methods on crawled data. Our findings highlight how the current protection of personal data is inconsistent with users' expectations of privacy.**

***Keywords*-social networks; web crawling; privacy;**

I. Introduction

In the past decade, social networking sites have become a mainstream cultural phenomenon [1]. They have proved useful for everything from keeping in touch with friends to dating, research collaboration, and political activism. Despite these uses and their worldwide popularity among youth, they have acquired a poor reputation for privacy and news stories frequently highlight the dangers of online disclosure. These fears include targeted scams, identify theft, cyber-bullying, stalking, and child solicitation [2].

Social networking sites have raised the stakes for privacy protection because of the centralisation of massive amounts of user data, the intimacy of personal information collected, and the availability of up-to-date data which is consistently tagged and formatted. This makes social networking sites an attractive target for a variety of organisations seeking to aggregate large amounts of user data, some for legitimate purposes and some for malicious ones. In most cases, extracting data violates users' expectation of privacy.

We take as a case study the popular website Facebook, which begins its privacy policy with the noble goal that "*You should have control over your personal information*" [3]. Unfortunately, we find this is not the case in practice. Our two main results are a set of techniques for extracting data from the site which most users are unaware of, and a demonstration of their effectiveness using real-world data.

II. Related Work

In the past few years, privacy in online social networks has been studied from many different angles: systems security [4], user psychology [5], public policy [2], and sociology [1]. We focus on the technical question of what information can be crawled from the network. Krishnamurthy and Wills studied social networks from the point of view of data exposure, concluding that lax privacy settings leave much user data able to be crawled [6]. The largest published crawls of social networks were done by Chau et al. [7] and Mislove et al. [8]. Each group collected datasets of roughly 10 million profiles using parallel crawlers.

Korolova et al. studied crawling social networks from a theoretical perspective [9]. They examined many possible strategies for crawling and analysed their mathematical efficiency. Our simulations in Section V use a similar model, and confirm many of their findings, but we focus on real methods for collecting data in the context of a specific network.

Many academic studies, including this one, have utilised data from crawls of social networks to analyse the structure of human social connections [6], [8], [9], [10], [11], [5], [12]. It is often assumed that the use of anonymised data sets is sufficient to protect privacy, although it has been shown that large datasets can usually be de-anonymised even if names are scrubbed from the graph [13], [14].

III. The Facebook Network

We have chosen to focus our study on Facebook because it is the largest, arguably the most feature-rich, and has the most complex privacy model. Facebook was founded in 2004 and originally was available only to university students in the United States. It has since opened to the general public and has over 200 million users, but educational- and corporate-specific sub-networks (which require email addresses from a specific domain to join) still play an important role in access control. Most users belong to at least one of the 531 regional networks, 9,764 university networks, and 129,168 high school networks, to which all of their data is visible. Facebook collects three broad categories of information which we will review next. This huge silo of data is considered one of the company's most valuable assets [2], and hence Facebook should make extracting the data difficult, using access-control mechanisms and anti-spidering techniques.

A. Data of Interest

1) Profile Data: The obvious data of interest is the personal data uploaded onto user profiles. Although users voluntarily upload this information, most do so expecting it to only be shown to their online "friends", thus they consider it a privacy violation if this information is leaked to third parties [5]. Profile data can consist of a user's name, location and contact information, educational and employment history, personal preferences, interests, and photos. Much of this data is tagged by the user with metadata, making it easy to store and analyse. The history of a user's profile is also of interest: previous versions of a user's profile, particularly status messages, provide a chronology of the user's life activities, such as when they have entered romantic relationships or changed jobs.

2) The Social Graph: A broader threat is extraction of the *social graph*, that is, the graph consisting of users (vertices) and their friendship links (edges). It is easier to extract this data than the complete set of user profiles, yet there is a wide variety of interesting uses of the social graph data: detecting communities of users with similar interests [12], identifying well-connected individuals [11], inferring private information from a user's friends [15], [16], and facilitating the efficient collection of user profiles via targeted attacks [10]. While the commonly considered graph is of friendship links, the graph connecting users through common group membership can also be sensitive.

3) Traffic Data: Facebook's privacy policy, like those of most social networks, clearly states that they reserve the right to record users' IP addresses, web browser information, length and frequency of sessions [3] . Privacy policies are typically vague about the collection and use of such data [2], however, it is at the lowest risk of extraction by third parties, so we will not consider it further.

B. Potential Data Aggregators

Social networks contain data which is invaluable to marketers, who use profile data to target their advertisements to particular users, as well as social graph data to advertise to friends of existing customers. Professional data aggregators may attempt to build databases of user profiles and connections for sale to insurance companies, credit-ratings agencies, background-check agencies, or others. Employers and universities can monitor social networking sites either to screen applicants, or to discipline existing employees and students. Controversies have already arisen from law enforcement agencies using social networking data to learn of criminal activity. Social networks are also useful for monitoring the affiliates of known criminals; the social network paradigm has been successfully used to investigate organised crime in the Netherlands [17]. Social network information also aids criminals with many online scams, particularly "social" phishing [18]. Finally, many researchers wish to extract social networking data [6], [8], [9], [10], [11], [5], [12].

IV. Methods of Extracting Data

A. Public Listings

The easiest method of collecting data is crawling public profile listings, which do not require an account to access and are routinely crawled by search engines, as is encouraged by Facebook. Public profiles consist of a user's name, primary photo, network memberships and links to the public profiles of eight friends. Recently, a listing of companies and groups that the user is a "fan" of has been added to the public listing. Public listings are enabled by default, and less than 1% of users opt out [19].

We developed a simple web-crawling script which was able to download 500,000 user profiles per day. This means that, as a rough estimate, a distributed crawl with 500 machines should be able to index every public Facebook listing in one day, easily within the reach of many of the aggregators we have mentioned. Collectively, these listings provide a significant amount of information about the social graph. It has been shown that it is possible to approximate many properties of the complete graph using just the random sample of links provided by public listings [19].

In addition to gathering public listings of users, there are public listings of groups which can be indexed. These provide both personal data, as many groups contain user preferences like "I love Icelandic music," and graph data, as membership in a common group indicates a social tie (especially a small one such as "Churchill Rugby Club"). While only 8 members of a group are publicly listed, all wall postings are publicly listed, which can provide a much larger membership set, as well as indicating the activity level of users within the group.

An interesting inconsistency that we have identified in Facebook's privacy settings is that users who have opted out of public listing for their own profile will still show up in group memberships and group wall postings, clearly a violation of the privacy policy [3]. Listing friendship information in public listings at all is a violation of the privacy policy, as it claims that only a user's name and photo will be displayed.

B. False Profiles

To gather more data than is available in the public view, an aggregator can create false profiles. This is straightforward because profile creation requires only a valid email address; this can be done anonymously using temporary webmail accounts and the anonymity network Tor [20]. When crawling from a valid Facebook account, it is possible to see the name, photo, networks and *all* friends of any user with the default setting of a "searchable" profile.

While Facebook allows users to opt-out of having their profile "searchable," in practice this makes the site difficult to use, as it becomes impossible for legitimate friendship requests to be received.[1] It is possible to remove one's picture or friendship list from search results, but this is also highly impractical as this information is necessary for people with common names to be identified by their friends in search results, due to the lack of unique user identifiers on the site. Experimentally, we have found that just 10% of users remove their profile picture or friendship information from search results.

In order to collect profile data, false profiles can be created in multiple networks, each one giving instant access to the profiles of most users in that network. Regional networks are completely public to join, and comprise the majority of Facebook users. A study found that 70–90% of users expose their profile to their regional network, correlated with the network's size [6]. Some corporate and academic networks require a valid email address within a certain domain, but in many cases these networks are large enough that compromising an email address is possible. Other subnetworks, such as secondary schools, require approval of an existing network member, although this is highly susceptible to social engineering.

Another effective technique is to send friend requests out to highly connected individuals—who are more likely than average to accept a friend request from a stranger—which provides a view of any of their friends who have enabled "Friend of Friend" visibility. These users can be picked out using data from public listings, which have been shown to provide a useful approximation of friendship degree [19]. Previous research has shown that creating a profile with a humorous picture on the profile, and little data, will be accepted as a friend by roughly 80% of users [6]. There is also anecdotal evidence that risqué photos of models will often be accepted as friends.

Unlike crawling public listings, creating false profiles and using them to crawl data is explicitly against Facebook's terms of service and is aggressively combated. For example, accounts can be permanently removed for issuing too many search requests, and accounts which exceed a certain threshold of friendships ($\sim 1,000$) are manually inspected and removed if they are thought to be fraudulent. Still, large organisations are able to utilise many false profiles in an attempt to crawl the network. We will examine the effectiveness of this approach further in Section V.

C. Profile Compromise and Phishing

More effective than creating false profiles is gaining access to legitimate profiles, which are already well-connected to a group of friends. Standard operating system attacks, such as malware and key-loggers, can of course be used to compromise Facebook credentials. Industrialised account harvesting using a large botnet could reveal a significant portion of the entire network, based on our analysis in Section V.

Facebook accounts are also vulnerable to phishing attacks, which have been widely reported in the last year [21]. Phishing is particularly effective against Facebook, as the site routinely sends users email reporting new messages or photos with embedded links to log in to the site. These links typically have complicated URLs such as *www.facebook.com/n/?inbox/readmessage.php&t=10179447* which are difficult for ordinary users to distinguish from phishing URLs. Thus, users are well-trained to click on emailed links and log in to the site, and there are no specific anti-phishing measures in place.[2] The Facebook log-in page is not authenticated via TLS, removing another obstacle for phishing sites.

In addition to regular log-in at Facebook, the recently launched Facebook Connect system allows external websites to provide a link for users to enter their Facebook credentials into a pop-up window, enabling the external website to access Facebook data. Training users to enter their password in this manner will likely lead to phishing attacks in the future. In particular, it enables "cross-site phishing," where an attacker can trick a legitimate website into displaying a "Connect with Facebook" button which directs to a malicious pop-up window.

Finally, the lack of TLS authentication on Facebook's log-in page enables a middleperson attack redirecting the password submission to an attacker's computer. While the actual HTTP `POST` action which submits credentials is normally protected with TLS, very few users are apt to notice if this has been removed.

D. Malicious Applications

Facebook offers a rich development platform for third-party applications. While users can set specific privacy settings for each application, in practice, most users give all applications full access to their account [4], the default setting. For data collection purposes, a malicious application has effectively phished the profile of every user who adds the application with such permissive settings. The most popular applications on Facebook have over 10 million users, and there exist several companies such as Slide and RockYou which collectively have over 30 million users of their applications [22]—enough to view the vast majority of the network, according to our evaluation in Section V.

Even if users select more restrictive settings, many holes have previously been found in Facebook's platform which allow applications to extract unauthorised data [4]. While

[1] In fact, if two users both have a non-searchable profile, there is no method for them to become friends on the network! The lack of a suitable means of out-of-band friend linking suggests that non-searchable profiles are tacitly discouraged by Facebook.

[2] Facebook is in position to easily deploy photo-based anti-phishing techniques. For example, if a user's profile photo is already publicly searchable, Facebook could display the user's profile photo next to the input box for their password.

Facebook only allows applications to display data in a site-specific "Facebook Markup Language" which is then translated into safe HTML and JavaScript on Facebook servers, hacks have been found in this interface which allow for the execution of arbitrary scripts [23]. There have even been Facebook-specific self-propagating worms [24].

E. Facebook Query Language

Facebook allows application developers to make queries using 'FQL,' a subset of the SQL database query language. While FQL is designed to provide access only to profile information for users registered with an application, we have found several flaws which allow applications with no registered users to gather social graph data from Facebook by repeatedly issuing FQL queries.

Problems arise from FQL's use of user IDs as capabilities which allow for querying some bits of user data. UIDs cannot be considered secret as they are present in URLs and are easily located in public listings. They can also be extracted directly from FQL, by submitting queries such as:

```
SELECT uid, name, affiliations FROM user
    WHERE uid IN (X,Y, ... Z);
```

which will return a list of valid user names and affiliations from a set of UIDs. We found it possible to crawl the UID-space in blocks of 1,000 UIDs, with each query taking $\sim$ 10 seconds to be processed. This means that it would take approximately 2,000 machine-days to exhaust Facebook's UID space of $2 \cdot 10^9$.

It is also possible to retrieve large blocks of UIDs by querying the members of a group:

```
SELECT uid FROM group_member
    WHERE gid = G;
```

This technique returns a maximum of 500 group members, making it less useful for gathering large sets of UIDs.

Once a large list of UIDs has been collected, their friendship connections can be retrieved using large friendship queries manually joining two sets of UIDs:

```
SELECT uid1, uid2 FROM friend
    WHERE uid1 IN (X,Y, ... Z)
    AND uid2 IN (U,V, ... W);
```

We found it possible to query all friendship links between two sets of 1,000 users at a time. Assuming an aggregator has the complete list of valid UIDs, it would take

$$\binom{\frac{N}{1{,}000}}{2} \approx \binom{200,000}{2} \approx 2 \cdot 10^{10}$$

queries to extract the complete graph, which is an infeasibly large number. However, this is still a useful technique because it can be used to find information that is invisible to other methods due to privacy settings. Users must manually opt out of the Facebook platform in order to be hidden from FQL queries, but we have found experimentally that less than 1% of users do so, as few are unaware of this is an avenue for data collection.

V. Evaluating Collection Techniques

To examine the implications of the attacks above, we tested their effectiveness against a crawled sub-network consisting of the original ID-space for students at Stanford university. We crawled this network using the FQL-query method described in Section IV-E. This network consists of 15,043 users and was crawled by a single desktop PC in less than 12 hours. Our crawling script collected only UIDs and links, which were deleted following our experiments.

We conjecture that, in many cases, compromising one regional network is the goal of an aggregator (for example, local marketers or a local police department). In practice, most networks are easily compromised given a single account within the network, due to the default setting of profiles being exposed to members of their own network. We will consider, however, networks for which we start out with no data, and evaluate three possible discovery techniques.

A. Data Collection Techniques

- **False Friendship Requests**—In this scenario, an aggregator has created a false profile and is able to send friendship requests to users of her choice. She begins with random requests, then begins to target high-degree nodes with the network view she builds up.
- **Random Compromise**—In this scenario, the aggregator can compromises random accounts via malicious applications, phishing, malware attacks, or a middleperson attack on a wireless access point.
- **Targeted Compromise**—In this most powerful scenario, the aggregator is able to target specific accounts, for example using "spear-phishing" attacks.

B. Evaluation of Techniques

We tested these methods on our crawled sub-graph, measuring their effectiveness against networks of users with friend-only privacy settings and friend-of-friend privacy settings. For each scenario, we evaluated the number of profiles and friendship links viewable given a limited budget of compromised nodes. The results of the different scenarios may not be directly comparable, since, for example, an aggregator may have a higher budget for random compromises than targeted compromises.

The results of our simulations on the Stanford network are shown in Figures 1–4. Consistent with previous analysis, we observe that a tiny fraction of network compromise gives away most of the network [10] [9]. We found that it in most cases, it is very efficient to extract 50% of user profiles and 90% of friendship links before returns begin to slowly diminish. The level of compromise required to get this 50%/90% benchmark is shown in Table 1, and marked with

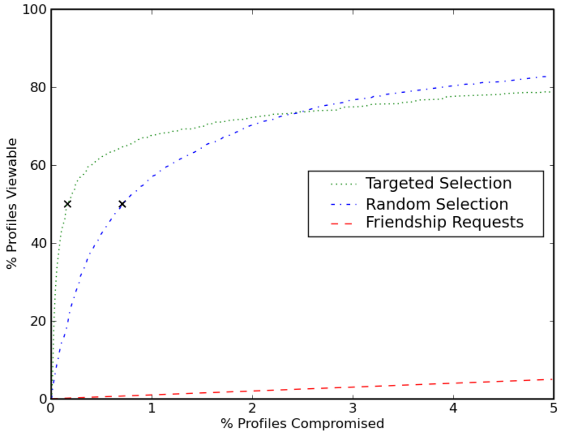

Figure 1. Profiles found, friends-only

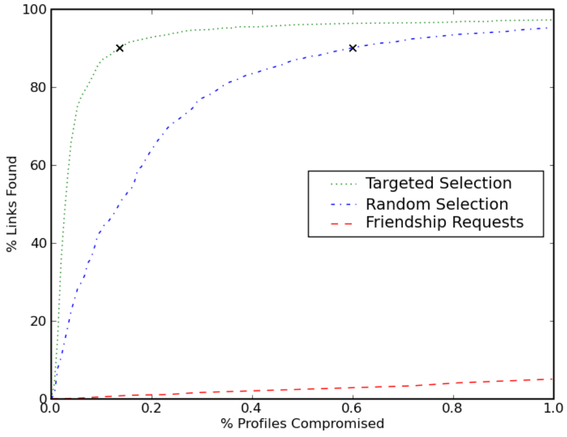

Figure 3. Links found, friends-only

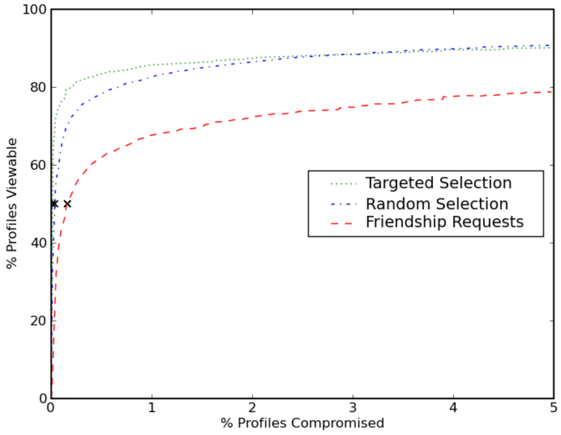

Figure 2. Profiles found, friend-of-friend

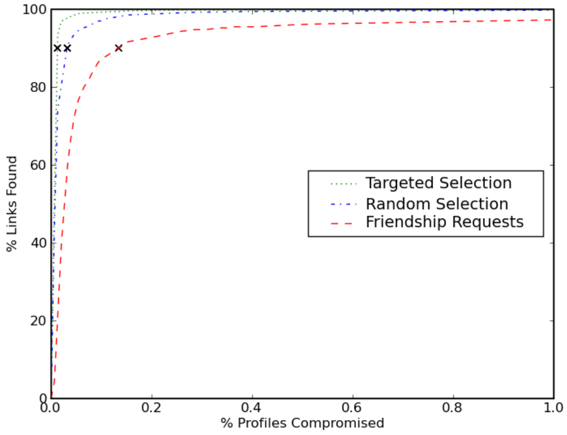

Figure 4. Links found, friend-of-friend

a '×' in Figures 1–4. In the worst case of universal friend-of-friend privacy settings, this level of discovery requires only 6 randomly compromised accounts out of over 15,000. Even with a friends-only privacy policy, we need just 112 random profiles, less than 1% of the total network.

In addition to the low number of accounts needed to compromise the network, we gained several important insights. First, random compromise yields almost as much information as targeted compromise, consistent with the analysis in [9]. This indicates that phishing, which yields an essentially random set of user profiles, can be very effective, given the poor phishing security described in Section IV-C.

Friend-of-friend privacy makes discovery of the network ∼100 times faster than friends-only privacy, consistent with the analysis of 'lookahead' in [9]. Critically, discovery of the network using bogus friendship requests becomes effective only with friend-of-friend privacy settings. This is important because the difficulty of creating a false profile and sending friend requests is far lower than compromising accounts.

Finally, we note that the discovery of a large number of links is significantly easier than discovering a large number of profiles. Given that many of the aggregators described in Section III-B are more interested in links than profiles, this is a compelling finding.

VI. Conclusions

Our analysis indicates that the current state of protection against data crawling of social networks has not kept pace as Facebook and other sites evolve into global networks interfacing with search engines, third-party applications, and

	50% profiles	90% links
Targeted comp., f.o.	0.16%	0.14%
Random comp., f.o.	0.71%	0.60%
Friend req., f.o.	50.0%	19.6%
Targeted comp., f.o.f.	0.01%	0.01%
Random comp., f.o.f.	0.04%	0.03%
Friend req., f.o.f.	0.16%	0.14%

Table I
EFFICIENCY OF DISCOVERY

external sites. We have enumerated the ways in which aggregators can extract both personal data and social graph data from social networks, contrary to most users' expectations and in some cases regardless of a user's privacy settings. We have also demonstrated that only a small percentage of accounts are needed to view the majority of the network, indicating that industrial-scale data collection is possible.

Our results suggest that social networks should limit the number of mechanisms to access user data, combat phishing aggressively, reduce reliance on sub-network membership for access control, and eliminate friend-of-friend data sharing. The greatest problem, however, may be the lack of user understanding of the limited amount of privacy which can actually be enforced by today's sites.

ACKNOWLEDGMENTS

The authors thank Claudia Diaz, Ross Anderson, Frank Stajano, and Robert Watson for helpful comments.

REFERENCES

[1] D. Boyd and N. Ellison, "Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication*, 2007.

[2] G. Hogben ed., "Security Issues and Recommendations for Online Social Networks," *European Network and Information Security Agency Position Paper*, Oct 2007.

[3] "Facebook Privacy Policy," *www.facebook.com/policy.php*, 2009.

[4] A. Felt and D. Evans, "Privacy Protection for Social Networking Platforms," *Workshop on Web 2.0 Security and Privacy*, 2008.

[5] A. Acquisti and R. Gross, "Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook," in *Privacy Enhancing Technologies – PET 2006*, 2006, pp. 36–58.

[6] B. Krishnamurthy and C. E. Wills, "Characterizing Privacy in Online Social Networks," in *WOSN: Workshop on Online Social Networks*, 2008, pp. 37 – 42.

[7] D. H. Chau, S. Pandit, S. Wang, and C. Faloutsos, "Parallel Crawling for Online Social Networks," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 1283–1284.

[8] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks," in *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, 2007, pp. 29–42.

[9] A. Korolova, R. Motwani, S. U. Nabar, and Y. Xu, "Link Privacy in Social Networks," in *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, 2008, pp. 289–298.

[10] G. Danezis and B. Wittneben, "The Economics of Mass Surveillance and the Questionable Value of Anonymous Communications," *WEIS: Workshop on the Economics of Information Security*, 2006.

[11] S. Wasserman and K. Faust, *Social Network Analysis*. Cambridge University Press, 1994.

[12] M. E. J. Newman and M. Girvan, "Finding and Evaluating Community Structure in Networks," *Physical Review E*, vol. 69, p. 026113, 2004.

[13] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore Art Thou R3579x?: Anonymized Social networks, Hidden Patterns, and Structural Steganography," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*. New York, NY, USA: ACM, 2007, pp. 181–190.

[14] A. Narayanan and V. Shmatikov, "How To Break Anonymity of the Netflix Prize Dataset," *CoRR*, vol. abs/cs/0610105, 2006.

[15] W. Xu, X. Zhou, and L. Li, "Inferring Privacy Information via Social Relations," *International Conference on Data Engineering*, 2008.

[16] J. Lindamood and M. Kantarcioglu, "Inferring Private Information Using Social Network Data," *WOSN: Workshop on Online Social Networks*, 2008.

[17] P. Klerks, "The Network Paradigm Applied to Criminal Organisations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands," *ISNA Connections*, pp. 53–65, 2001.

[18] T. Jagatic, N. Johnson, M. Jakobsoon, and F. Menczer, "Social Phishing," *Communications of the ACM*, vol. 50, no. 10, p. 94, 2007.

[19] J. Bonneau, J. Anderson, F. Stajano, and R. Anderson, "Eight Friends Are Enough: Social Graph Approximation via Public Listings," in *SNS '09: Proceeding of the 2nd ACM Workshop on Social Network Systems*, 2009.

[20] "Tor Project," *www.torproject.org*, 2009.

[21] M. Arrington, "Phishing For Facebook," *TechCrunch*, Jan 2008.

[22] "Facebook Application Statistics," *www.allfacebook.com/*, Jan 2009.

[23] A. Felt, "Defacing Facebook: A Security Case Study," *www.cs.virginia.edu/felt/fbook/facebook-xss.pdf*, 2007.

[24] M. Arrington, "Elaborate Facebook Worm Spreading," *TechCrunch*, Aug 2008.