Journal of Privacy and Confidentiality (2012) **4**, Number 2, 1–5

Is the Privacy of Network Data an Oxymoron?

Stephen E. Fienberg*

1 Introduction

While social networks are now a part of everyday life for the vast majority of people using computers, smartphones, and tablets, privacy is but an afterthought. Google+ has in excess of 100 million users a month while Facebook has topped 1 billion. Other more specialized networks such as Linked-in add to the fray. But from a privacy perspective the biggest concern for users should be the efforts to integrate the networking apps into all other forms of online activity as well as the constant effort to link additional data to network information, in addition to the network owners' efforts to market that information to third party vendors. Further, Facebook and other networking sites have already begun to build search capabilities and other facilities into their systems that extend the information they collect from users even further. What protections do users really have?

While most social network users when queried express concern for the privacy of their posted information, they nonetheless post large quantities of potentially embarrassing or at least individually identifiable information and appear to be unconcerned with the control or lack thereof they exert over their posted information. Further, the controls offered form a constantly shifting landscape. Matt McKeon offers a succinct summary of the changes to Facebook's default privacy settings from 2005 through 2010 in striking graphical form, based on his interpretation of the Facebook "Terms of Service" over the years, along with his personal memories of the default privacy settings for different classes of personal data.[1] The trend is striking, and it has continued to the present day. Facebook's approach remains an "opt-out" rather than "opt-in" one, and thus as its scope has expanded so have users' vulnerabilities. While privacy settings may be easier to use today than in the past, many users believe that "Keeping your Facebook info private is getting harder and harder all the time—mostly because Facebook keeps trying to make it public."[2] The problems are many. For example, most users don't realize that when they hide a post or photograph from their profile page, that those posts are not truly hidden and can be visible elsewhere, including on another person's page, and are ultimately easily accessible to external third parties.

In this issue of the *Journal of Privacy and Confidentiality*, Stutsman, Acquisti, and Gross [11] summarize attitudes and practices of the Carnegie Mellon Facebook community over the period 2005–2010. And in a separate study [1], they illustrate

*Department of Statistics, Machine Learning Department, Heinz College, and CYLAB, Carnegie Mellon University, Pittsburgh, PA, `mailto:fienberg@stat.cmu.edu`.

[1] See: http://mattmckeon.com/facebook-privacy/.

[2] Whitson Gordon. "The Always Up-to-Date Guide to Managing Your Facebook Privacy," http://lifehacker.com/5813990/the-always-up+to+date-guide-to-managing-your-facebook-privacy, (accessed 1/25/13).

how information posted by Carnegie Mellon users on their webpages, especially photos, can be easily linked to external online data, facilitating the identification of individuals in other settings such as "anonymous" online dating networks. In many ways these authors describe the tip of the privacy iceberg. In this paper, we turn to a much narrower and technical aspect of privacy protection for network data. The basic message is quite simple: privacy protection is difficult at best and impossible at worst, even when we restrict attention to simple caricatures of the nature of network data. The reason for this is the fundamental dependence of network structures, thus making what might otherwise be protectable personal information vulnerable to attacks. And the identifiable information may be associated with the individuals or with their links, or with the individuals' associations with inclusion in various affinity groups. Photos, which are now tagged and labeled in many networks such as Facebook, are especially vulnerable.

2 Protecting the Privacy of Network Data

There is a growing literature of successful attacks on the privacy of social network data, and we refer the reader to Backstrom et al. [2] and Narayanan and Shmatikov [9] for a discussion of some of these. Similarly, by now numerous authors have proposed methods for protecting either nodes or edges in a graph using a variety of privacy-preserving criteria. Zheleva and Getoor [13] and Zheleva et al. [14] provide excellent descriptions of many of these. Yet, to date there is no real technical fix for the release of real social network data for many reasons—in part because of the complexity and clear identifiability of much of the data posted on social networks, and in part from the fundamental nature of data, i.e., the dependence it induces among the nodes of the network corresponding to actual individuals and their profiles.

We illustrate using the data formatted in Figure 1. The $n \times p$ array in the left-hand side of the figure, with entries $\{x_{ij}\}$, corresponds to the usual multivariate persons by a variable array of values. This matrix contains nodal characteristics or covariates in the network setting. These can be continuous, discrete, or a mixture of the two, and can even be objects such as pictures. We could apply any standard approach to protecting the release of information from such an array, such as the many variants on matrix masking [4], e.g., data perturbation, or more elaborate methods such as Random Orthogonal matrix manipulation [12], as well as invoking any of the usual criteria such as k-anonymity, l-diversity, differential privacy, etc. All of these methods typically take the rows of the array corresponding to individuals as realizations of independent random variables. Unfortunately, this is not the case for network data.

The $n \times n$ adjacency matrix in the right-hand side of Figure 1, with 0-1 entries y_{ij}, describes the dependencies among the nodes. The same standard (by now) privacy protection mechanisms can be applied to this part of the data. The problem is that the two parts of the data are inextricably intertwined since the same persons are represented in both parts.

We can illustrate the difficulties that now arise by considering the approach associ-

Persons	Variables				
1	x_{11}	x_{12}	x_{13}	$\cdots$	x_{1p}
2	x_{21}	x_{22}	x_{23}	$\cdots$	x_{2p}
3	x_{31}	x_{32}	x_{33}	$\cdots$	x_{3p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	x_{n3}	$\cdots$	x_{np}

Persons				
-	y_{12}	y_{13}	$\cdots$	y_{1n}
y_{21}	-	y_{23}	$\cdots$	y_{2n}
y_{31}	y_{32}	-	$\cdots$	y_{3n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_{n1}	y_{n2}	y_{n3}	$\cdots$	-

Figure 1: A Generic Form of Network Data: Standard $n \times p$, persons by variables data array on the left; Adjacency Matrix linking persons (network nodes) on the right.

ated with differential privacy (Dwork et al. [5]). Differential privacy in a network context would appear to require the revealed properties of a graph to look roughly the same if any single node is removed. There are of course different ways we could define node removal especially in light of the dependencies, but none seem to be able to get around the fact that the nodes in the network are linked and aggregate properties of a network graph can change substantially simply by the removal of a single node. One way to address this issue is to attempt to release standard data summaries for networks associated with many statistical models of interest such as the in- and out-degree distributions or the corresponding degree sequences, e.g., see Goldenberg et al. [6]. Nodes with high in-degrees and out-degrees associated in a network graph are especially vulnerable to attack and their removal can also alter the structure of the network graph.

Some of the more recent work on network privacy protection emanating from the differential privacy literature attempts to address these difficulties, but it does so by focussing solely on the right-hand side of Figure 1, the adjacency matrix. Karwa et al. [7] focus on counting queries, such as the number of edges, two-stars, or triangles in the network graph. These often show up as minimal sufficient statistics in exponential random graph models, e.g., see Goldenberg et al. [6]. Karwa and Slavković [8] focus directly on the degree sequence in an undirected network setting, with the goal of releasing synthetic graphs under the β model., for which the degree sequence is a minimal sufficient statistic, e.g., see [10]. Finally, Blocki et al. [3] take a somewhat different tack and replace the usual local or global sensitivity component of differential privacy with the notion of restricted sensitivity using the concepts of edge-adjacency and vertex adjacency.

3 The Facebook Privacy Challenge

What remains a challenge, even in the simple network data scenario of Figure 1, is how to combine the two sides of the figure from the perspective of privacy protection. Only then can we move towards addressing the kinds of data posted on standard social network settings and membership in affinity groups and other forms of individualization. Unfortunately, no formal privacy tools can protect people from themselves. Good privacy controls on social network sites are important but far from sufficient. And once individuals release information about themselves, their friends, and their families, that very information can then be the basis for totally undercutting any promises of privacy protection that a vendor or data owner might make. For some excellent advice on privacy setting for Facebook I recommend a recent article in the New York Times.[1] Facebook and other social networking sites remain a moving target when it comes to privacy protection.

Thus, when it comes to posting on social network sites, the best advice I can offer is *caveat emptor.* The instant gratification that many social network users get from their favorite site is not easily balanced against the long term harm that the release of truly private information might produce.

This is why I do not use Facebook, Google+, or any other networking sites and why I do not tweet. I know that I can be found on Facebook and other social networks, but only through postings created by others. Unfortunately, pictures of me and my biographical information are there and I am powerless to remove them.

Acknowledgments

This research was supported in part by NSF grants DMS-0631589 and BCS0941518 to the Department of Statistics, Carnegie Mellon University, and by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office to the Carnegie Mellon University - Singapore Management University Living Analytics Research Centre.

[1] Somini Sengupta (2013). "Staying Private on the New Facebook," *New York Times*, February 6, 2013, as well as a followup online column by the same author entitled "New Stuff I've Learned Since My Facebook Privacy Tool Kit." *New York Times*, February 7, 2013. `http://www.nytimes.com/2013/02/07/technology/personaltech/protecting-your-privacy-on-the-new-facebook.html?_r=0`.
`http://bits.blogs.nytimes.com/2013/02/07/new-stuff-ive-learned-since-my-facebook-privacy-tool-kit/`.

References

[1] Acquisti, A., Gross, R., and Stutzman, F. (2011). Faces of Facebook: Privacy in the age of augmented reality. In *BlackHat USA, 2011*.

[2] Backstrom, L., Dwork, C., and Kleinberg, J. (2007). 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: ACM. 181–190.

[3] Blocki, J., Blum, A., Datta, A., and Sheffet, O. (2012). Differentially private data analysis of social networks via restricted sensitivity. `http://arxiv.org/abs/1208.4586`

[4] Duncan, G. T., Elliot, M., and Salazar-Gonzalez, J. J. (2011). *Statistical Confidentiality: Principles and Practices*. New York: Springer.

[5] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*. Berlin: Springer.

[6] Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2): 129–233.

[7] Karwa, V., Raskhodnikova, S., Smith, A., and Yaroslavtsev, G. (2011). Private analysis of graph structure. In *Private Analysis of Graph Structure*, vol. 11, 1146–1147.

[8] Karwa, V. and Slavković, A. B. (2012). Differentially private graphical degree sequences and synthetic graphs. In J. Domingo-Ferrer and I. Tinnirello, (eds.), *Privacy in Statistical Databases 2012*, vol. 7556 of *LNCS*. Berlin: Springer. 273–285.

[9] Narayanan, A. and Shmatikov, V. (2009). De-anonymizing social networks. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*.

[10] Rinaldo, A., Petrović, S., , and Fienberg, S. E. (2013). Maximum likelihood estimation in the beta model. *Annals of Statistics*, 41(1): in press.

[11] Stutzman, F., Gross, R., and Acquisti, A. (2012). Silent listeners: The evolution of privacy and disclosure on Facebook. *Journal of Privacy and Confidentiality*, 2(2).

[12] Ting, D., Fienberg, S. E., and Trottini, M. (2008). Random orthogonal matrix masking methodology for microdata release. *International Journal of Information and Computer Security*, 2(1): 86–105.

[13] Zheleva, E. and Getoor, L. (2011). Privacy in social networks: A survey. In C. C. Aggarwal, (ed.), *Social Network Data Analytics*. New York: Springer. 277–306.

[14] Zheleva, E., Terzi, E., and Getoor, L. (2012). *Privacy in Social Networks*. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers.