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. 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.” 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.


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

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


© 2012 by the authors http://repository.cmu.edu/jpc
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 [3], 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-
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.

<table>
<thead>
<tr>
<th>Persons</th>
<th>Variables</th>
<th>Persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$x_{11}$</td>
<td>$y_{12}$ $y_{13}$ ··· $y_{1n}$</td>
</tr>
<tr>
<td>2</td>
<td>$x_{21}$</td>
<td>$y_{21}$ $y_{23}$ ··· $y_{2n}$</td>
</tr>
<tr>
<td>3</td>
<td>$x_{31}$</td>
<td>$y_{31}$ $y_{12}$ $y_{33}$ ··· $y_{3n}$</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>$n$</td>
<td>$x_{n1}$</td>
<td>$y_{n1}$ $y_{n2}$ $y_{n3}$ ··· $y_{np}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variables</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{12}$</td>
<td>··· $x_{1p}$</td>
</tr>
<tr>
<td>$x_{22}$</td>
<td>··· $x_{2p}$</td>
</tr>
<tr>
<td>$x_{32}$</td>
<td>··· $x_{3p}$</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>$x_{np}$</td>
<td>:</td>
</tr>
</tbody>
</table>

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. 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.

---

References


