

# Technical Perspective

## Anonymity Is Not Privacy

By Vitaly Shmatikov

WE LIVE IN an era of data abundance. Every aspect of our online and offline behavior—every click on a Web site, every relationship in online social networks, every bit of information we disclose about ourselves—is captured and analyzed by multiple entities, ranging from Internet service providers to advertising agencies to credit bureaus. With this dramatic increase in data collection, the companies holding our data face the responsibility for protecting our privacy, especially as they sell and exchange information about us.

Existing privacy protection technologies are overwhelmingly based on anonymization. They remove a few data attributes (such as names) that could be used to identify individuals and consider the resulting anonymized datasets safe from privacy violations. This approach is pervasive in academic literature, as well as industry practices. Whether it's the chief privacy officer of a major online social network testifying to a U.S. Senate committee that there is a critical distinction between the use of information in “personally identifiable form” and the use, sharing, and dissemination of information in “non-personally identifiable form,” or a popular Web site informing its customers that it shares non-personally identifiable information about them with hundreds of advertisers, the safe-keepers of our data act as if anonymity were equivalent to privacy. But is it, really?

To build meaningful protections for sensitive individual data, we must ask the right questions. What does it mean to compromise privacy? How can a potential adversary access and/or influence the data, both before and after anonymization? What are the adversary's capabilities, and what information might she employ to reverse anonymity? Unfortunately, many existing privacy technologies suffer from a certain poverty of imagination. For example, they assume the only way to reidentify anonymized re-

**The following paper is an object lesson in how to do data privacy research.**

records is to link them with an external dataset by matching common demographic attributes. As a consequence, anonymization is easily broken by creative adversaries who use a different attack model.


The following paper by Lars Backstrom, Cynthia Dwork, and Jon Kleinberg is a landmark in privacy research because it asks all of the above questions and gives unexpected answers. The authors demonstrate fragility of data anonymization, invent several new techniques for reidentifying anonymized nodes in social networks, and radically change our understanding of what constitutes personally identifiable information.

Their first contribution is to investigate the meaning of anonymity in graph-structured data, which are very different from relational datasets traditionally considered in privacy research. They focus on online social networks, but their results apply broadly to telephone call graphs, survey data, and, in general, almost any dataset containing information about relationships between people.

Their second contribution is the insight that the basic topological structure of the social graph can act as an identifier. They show that patterns of social links—whether arising naturally or artificially introduced into the social network by the adversary—tend to be unique and efficiently recognizable even in a completely ano-

nymized graph, yet without knowing the pattern, it is difficult to determine whether the graph contains such a structure. This idea is very powerful because graph structure is not a “personally identifiable” attribute by any meaning of the term. Nevertheless, Backstrom et al. show how it can be used to reidentify (sub)graphs that have been anonymized according to the best legal standards and satisfy the strongest anonymity properties.

Their third contribution is a new class of attacks on anonymity; in particular, an active attack in which the adversary deliberately introduces random links into the social network so that the resulting subgraph can be recognized even after all information about identities has been erased from the network. Existing privacy technologies fail to account for the possibility that the adversary may influence the data prior to anonymization and thus do not provide a defense against this threat.

The era of data abundance is bringing new kinds of sensitive data about individuals, new understanding of privacy risks, new attacks, and new defenses. This work provides us with valuable insights in all of these areas. By showing that the basic structure of our social relationships can be as identifying as a name, they debunk the naive belief that simple removal of identifiers renders the data non-personally identifiable. The authors carry out a rigorous theoretical analysis of anonymity in social networks (including interesting connections to graph theory) and accompany it by the empirical evaluation of their reidentification techniques on a large, real-world social network of the LiveJournal blogging service. Their paper is an object lesson in how to do data privacy research. It should be required reading for anyone interested in this area. 

Vitaly Shmatikov (shmat@cs.utexas.edu) is an associate professor at the University of Texas at Austin.

© 2011 ACM 0001-0782/11/12 \$10.00