Privacy-Aware Data Management in Information Networks

Michael Hay
Cornell University
Ithaca, NY
mhay@cs.cornell.edu

Kun Liu
Yahoo! Labs
Santa Clara, CA
kun@yahoo-inc.com

Gerome Miklau
U. of Massachusetts Amherst
Amherst, MA
miklau@cs.umass.edu

Jian Pei
Simon Fraser University
Burnaby, BC Canada
jpei@cs.sfu.ca

Evimaria Terzi
Boston University
Boston, MA
evimaria@cs.bu.edu

ABSTRACT

The proliferation of information networks, as a means of sharing information, has raised privacy concerns for enterprises who manage such networks and for individual users that participate in such networks. For *enterprises*, the main challenge is to satisfy two competing goals: releasing network data for useful data analysis and also preserving the identities or sensitive relationships of the individuals participating in the network. Individual users, on the other hand, require *personalized* methods that increase their awareness of the visibility of their private information.

This tutorial provides a systematic survey of the problems and state-of-the-art methods related to both enterprise and personalized privacy in information networks. The tutorial discusses privacy threats, privacy attacks, and privacy-preserving mechanisms tailored specifically to network data.

Categories and Subject Descriptors

H.2.0 [**DATABASE MANAGEMENT**]: Security, integrity, and protection

General Terms

Algorithms, Security

Keywords

Networks, Privacy, Anonymization, Differential Privacy

1. INTRODUCTION

A network dataset is a graph representing a set of entities and the connections between them. Network data can describe a variety of domains: a *social network* might describe individuals connected by friendships; an *information network* might describe a set of articles connected by citations; a *communication network* might describe Internet hosts related by traffic flows.

The ability for enterprises to collect network data has increased rapidly, creating the possibility for a wide range of compelling data analysis tasks: studying disease transmission, measuring a publication's influence, evaluating a network's resiliency to faults and attacks, etc. While members of the enterprise may be able to perform these analyses, they often wish to enlist outside experts. In addition, the ability to disseminate network data, and/or release the results of analyses, supports experimental repeatability and advances scientific investigation. Unfortunately, access to network data is extremely constrained because many networks contain sensitive information about their participants.

The first topic in this tutorial focuses on privately managing *enterprise* network data. We investigate the problem of limiting the disclosure of sensitive information while publishing network data sets for useful analysis, or alternatively, releasing the results of network analyses.

At the same time, individuals increasingly participate in online information networks: complex systems in which their personal information and interactions with others are recorded and displayed publicly. The second part of the tutorial identifies the privacy risks to individuals due to these public networked interactions. We focus our discussion on measures for quantifying users' privacy risks as well as mechanisms for helping them identify appropriate privacy settings.

2. TUTORIAL OUTLINE

Our tutorial is organized as follows.

2.1 Information-network data

The tutorial begins with a brief introduction to network data that includes social-network data (e.g., Facebook, LinkedIn), instant-messenger networks, collaboration networks, etc. We review examples of large-scale networks currently being collected and analyzed. We provide examples of key analysis tasks as well as the nature of the sensitive information contained in network data sets.

2.2 Threats and attacks for network data

Because network analysis can be performed in the absence of entity identifiers (e.g., name, social security number), a natural strategy for protecting sensitive information is to replace identifying attributes with synthetic identifiers. We refer to this procedure as *naive anonymization*. This com-

SIGMOD'11, June 12–16, 2011, Athens, Greece.

mon practice attempts to protect sensitive information by breaking the association between the real-world identity and the sensitive data.

We review a number of attacks on naively anonymized network data which can re-identify nodes, disclose edges between nodes, or expose properties of nodes (e.g., node features). These attacks include: *matching attacks*, which use external knowledge of node features [20, 14, 39, 27]; *injection attacks*, which alter the network prior to publication [1]; and *auxiliary network attacks*, which use publicly available networks as an external information source [25].

2.3 Publishing networks privately

For enterprises that wish to publish their network data without revealing sensitive information, the above attacks demonstrate that simply removing identifiers prior to publication fails to protect privacy. Thus, enterprises must consider more complex transformations of the data. An active area of research has focused on designing algorithms that transform network data so that it is safe for publication.

These algorithms have two primary objectives. First, the transformations should protect privacy, which is typically demonstrated by proving that the transformed network resists certain attacks. Second, the utility of the data should be preserved by the transformation—i.e., salient features of the network should be minimally distorted. While the privacy objective makes some distortion inevitable, most algorithms are designed to minimize distortion (measured in various ways, depending on the algorithm). In most cases, utility is not provably guaranteed but rather assessed empirically, for instance through a comparison of the transformed network to the original in terms of a measure of graph distance, or in terms of the difference in various network statistics, such as average shortest path lengths, clustering coefficient and degree distribution. In addition to protecting privacy and preserving utility, algorithm runtime and scalability are also important considerations.

One of the first algorithms proposed was that of Liu and Terzi [20], which transforms the network through edge insertions. Edges are added until nodes cannot be distinguished by their degree: specifically, for each node, there are at least $k-1$ other nodes that share its degree. This prevents an adversary with knowledge of node degree from re-identifying a target node beyond a set of k candidates. (The transformed network may also resist attacks from adversaries with richer auxiliary information, but the algorithm provides no formal guarantee.) To preserve utility, the algorithm attempts to find the minimal set of edge insertions necessary to achieve the privacy objective.

In recent years, many other algorithms have been proposed (cf. surveys [13, 32, 38]). They can be organized based on two key design decisions: the kind of data transformation and the privacy objective. Algorithms transform the network using one of several kinds of alteration: *directed alterations* transform the network through addition and deletion of edges [6, 20, 27, 39]; *network generalization* summarizes the network data in terms of node groups [5, 7, 8, 14]; *random alteration* transforms the network stochastically via random edge additions, deletions, or rewirings [15, 22, 31, 34]. There is also work comparing different alteration strategies [33].

While a common privacy objective is preventing re-identification [5, 6, 14, 20, 27, 39], other work seeks to prevent the disclosure of sensitive information, including edges [7, 8, 22, 34, 35, 36]. The techniques make different assumptions about adversary capabilities, from auxiliary information limited to node attributes [7, 8], node degree [20], or immediate (labelled) neighborhood [27], to arbitrary structural information [5, 6, 14, 39]. In addition, recent work looks at the privacy risks of releasing multiple views of a dynamic network [3, 4].

2.4 Answering network queries privately

Instead of publishing a transformed network, an alternative strategy allows users to query the data through a controlled interface. Prior work on querying private data has investigated two techniques: auditing and perturbation. Query auditing, in which queries are denied if the answers may lead to privacy breaches, has proven to be computationally hard and surprisingly subtle (because even a denial can lead to a privacy breach) [24]. Perturbation injects random noise into the query answers, creating uncertainty about the state of the underlying database. If appropriately calibrated, an individual's sensitive information can be hidden by the noise. Recent work in *differential privacy* has characterized a relationship between the magnitude of the noise and quantifiable privacy protection [10]. The noise depends both on the number of queries and on query *sensitivity*, which measures how much an individual's data can affect the answer.

There has been much research (cf. a recent review [9]) on differentially private mechanisms for querying tabular data, where a person's private information is encapsulated in a single database record. Network data poses new challenges because private information may span multiple records. The choice of differential object (a single edge vs. a node and incident edges) has profound implications on the semantics of the differential privacy guarantee and the resulting accuracy of query answers [12, 17].

The research on private network analysis is in its early stages. Our tutorial will describe the sensitivity of some common network statistics and discuss the limitations of answering high sensitivity queries under differential privacy. We also review recent results on privately computing network statistics like the frequency of degrees [12, 16] and subgraph motifs [26, 28], as well as the challenges of designing a social recommendation system over private network data [23]. Finally, we discuss interesting directions for future work, including the possibility of using query answering mechanisms as a basis for generating synthetic network data.

2.5 Privacy management for individual users

As the number of online social-networking users explodes, securing individuals' privacy to avoid threats such as *identity theft* and *digital stalking* becomes an increasingly important issue. Unfortunately, even sophisticated users who value privacy often compromise it to improve their digital presence. They know that loss of control over their personal information poses a long-term threat, but they cannot assess the risk accurately enough to compare it with the short-term gain. Even worse, tracking privacy controls in online services is often a complicated and time-consuming task that confuses many users. We dedicate this section of the tutorial to privacy concerns from the individual users' viewpoint. Specifically, we introduce models and algorithms to measure the potential privacy risks for online users due to the information they share explicitly or implicitly [2, 19,

21, 37]. We discuss mechanisms that help users better manage their privacy settings [11, 30]. We also review some work for users to trade their privacy for better services [29]. Finally, we discuss the privacy implications and design issues of microtargeted advertising, which offers the benefits of fine-grained audience targeting [18]. We conclude with open problems and future directions in personalized privacy for online information networks.

3. REFERENCES

[1] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *WWW*, pages 181–190, Banff, Alberta, Canada, May 2007.

[2] J. Becker and H. Chen. Measuring privacy risk in online social networks. In *Proceedings of Web 2.0 Security and Privacy Workshop (W2SP'09)*, Oakland, CA, 2009.

[3] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava. Prediction promotes privacy in dynamic social networks. In *Workshop on Online Social Networks (WOSN)*, 2010.

[4] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava. Privacy in dynamic social networks. In *WWW*, 2010.

[5] A. Campan and T. M. Truta. A clustering approach for data and structural anonymity in social networks. In *PinKDD*, 2008.

[6] J. Cheng, A. W.-C. Fu, and J. Liu. k-isomorphism: Privacy preserving network publication against structural attacks. In *SIGMOD*, 2010.

[7] G. Cormode, D. Srivastava, S. Bhagat, and B. Krishnamurthy. Class-based graph anonymization for social network data. In *VLDB*, 2009.

[8] G. Cormode, D. Srivastava, T. Yu, and Q. Zhang. Anonymizing bipartite graph data using safe groupings. *Proceedings of the VLDB Endowment*, 1(1):833–844, 2008.

[9] C. Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86–95, 2011.

[10] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Third Theory of Cryptography Conference*, 2006.

[11] L. Fang and K. LeFevre. Privacy wizards for social networking sites. In *WWW*, Raleigh, NC, April 2010.

[12] M. Hay, C. Li, G. Miklau, and D. Jensen. Accurate estimation of the degree distribution of private networks. In *Proceedings of the 2009 IEEE International Conference on Data Mining (ICDM'09)*, pages 169–178, Miami, FL, December 2009.

[13] M. Hay, G. Miklau, and D. Jensen. *Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques*, chapter Enabling Accurate Analysis of Private Network Data. Chapman & Hall/CRC Press, 2010.

[14] M. Hay, G. Miklau, D. Jensen, D. Towsley, and C. Li. Resisting structural re-identification in anonymized social networks. *VLDB Journal*, 2010.

[15] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava. Anonymizing social networks. Technical report, University of Massachusetts Amherst, 2007.

[16] M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially-private queries through consistency. *Proceedings of the VLDB Endowment*, 2010.

[17] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *SIGMOD*, 2011.

[18] A. Korolova. Privacy violations using microtargeted ads: A case study. In *IEEE International Workshop on Privacy Aspects of Data Mining (PADM'2010)*, pages 474–482, 2010.

[19] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. Inferring private information using social network data. In *WWW*, pages 1145–1146, Madrid, Spain, 2009.

[20] K. Liu and E. Terzi. Towards identity anonymization on graphs. In *SIGMOD*, pages 93–106, Vancouver, Canada, June 2008.

[21] K. Liu and E. Terzi. A framework for computing the privacy score of users in online social networks. In *Proceedings of the 2009 IEEE International Conference on Data Mining (ICDM'09)*, pages 288–297, Miami, FL, December 2009.

[22] L. Liu, J. Wang, J. Liu, and J. Zhang. Privacy preservation in social networks with sensitive edge weights. In *Proceedings of 2009 SIAM International Conference on Data Mining (SDM'09)*, pages 954–965, Sparks, NV, April 2009.

[23] A. Machanavajjhala, A. Korolova, and A. D. Sarma. Personalized social recommendations - accurate or private? In *VLDB*, 2011.

[24] S. Nabar, K. Kenthapadi, N. Mishra, and R. Motwani. A survey of query auditing techniques for data privacy. *Privacy-Preserving Data Mining*, pages 415–431, 2008.

[25] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *Proceedings of the 30th IEEE Symposium on Security and Privacy*, pages 173–187, Oakland, CA, May 2009.

[26] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, pages 75–84, New York, NY, USA, 2007. ACM.

[27] J. Pei and B. Zhou. Preserving privacy in social networks against neighborhood attacks. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering (ICDE'08)*, pages 506–515, Cancun, Mexico, April 2008.

[28] V. Rastogi, M. Hay, G. Miklau, and D. Suciu. Relationship privacy: output perturbation for queries with joins. In *Symposium on Principles of Database Systems (PODS)*, pages 107–116, Providence, RI, June 2009.

[29] J.-M. Seigneur and C. D. Jensen. Trading privacy for trust. In *Trust Management*, volume 2995/2004 of *Lecture Notes in Computer Science*, pages 93–107. Springer, 2004.

[30] A. C. Squicciarini, M. Shehab, and F. Paci. Collective privacy management in social networks. In *WWW*, pages 521–530, Madrid, Spain, 2009.

[31] L. Wu, X. Ying, and X. Wu. Reconstruction from randomized graph via low rank approximation. In *Proceedings of 2010 SIAM International Conference on Data Mining (SDM'10)*, Columbus, OH, April 2010.

[32] X. Wu, X. Ying, K. Liu, and L. Chen. A survey of algorithms for privacy- preservation of graphs and social networks. In *Managing and Mining Graph Data*. Kluwer Academic Publishers, 2010.

[33] X. Ying, K. Pan, X. Wu, and L. Guo. Comparisons of randomization and k-degree anonymization schemes for privacy preserving social network publishing. In *Proceedings of the 3rd SIGKDD Workshop on Social Network Mining and Analysis (SNA-KDD'09)*, 2009.

[34] X. Ying and X. Wu. Randomizing social networks: a spectrum preserving approach. In *Proceedings of 2008 SIAM International Conference on Data Mining (SDM'08)*, pages 739–750, Atlanta, GA, April 2008.

[35] X. Ying and X. Wu. On link privacy in randomizing social networks. In *PAKDD*, 2009.

[36] E. Zheleva and L. Getoor. Preserving the privacy of sensitive relationships in graph data. In *Workshop on Privacy, Security, and Trust in KDD (PinKDD)*, pages 153–171, 2007.

[37] E. Zheleva and L. Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *WWW*, pages 531–540, Madrid, Spain, April 2009.

[38] B. Zhou, J. Pei, and W.-S. Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *SIGKDD Explorations*, 2008.

[39] L. Zou, L. Chen, and M. T. Özsu. K-automorphism: A general framework for privacy preserving network publication. *Proceedings of the VLDB Endowment*, 2(1):946–957, 2009.