Winter 12-15-2012

A Framework for Enabling Privacy Preserving Analysis of Graph Properties in Distributed Graphs

Xiaoyun He
Kean University, xhe@kean.edu

A Framework for Enabling Privacy Preserving Analysis of Graph Properties in Distributed Graphs

Xiaoyun He[1]
Nathan Weiss Graduate College, Kean University,
Union, NJ, USA

ABSTRACT

In the real world, many phenomena can be naturally modeled as a graph whose nodes represent entities and whose edges represent interactions or relationships between the entities. Past and ongoing research on graphs has developed concepts and theories that may deepen the understanding of the graph data and facilitate solving many problems of practical interest represented by graphs. However, little of this work takes privacy concerns into account. This paper contributes to privacy preserving graph analysis research by proposing a framework for enabling privacy preserving analysis of graph properties in distributed graphs. The framework is composed of three modules. We discuss the functionality of each module and describe how the modules together ensure the privacy protection while retaining graph properties and answer users' queries pertaining to graph properties.

Keywords: Privacy preserving, graphs, distributed, framework, graph properties.

INTRODUCTION

In the real world, many phenomena can be naturally modeled as a graph whose nodes represent entities and whose edges represent interactions or relationships between the entities. Online social networks, transaction networks in financial services, and peer-to-peer file sharing

[1] Corresponding author. xhe@kean.edu +1 908 737 5998

systems are just a few examples. Moreover, advances in information technology and the ubiquity of networked computers have facilitated the creation of large amount of graph data in many application domains. Past and ongoing research on graphs has developed concepts and theories that may deepen the understanding of the graph data and facilitate solving many problems of practical interest represented by graphs (Cormen et al. 2001). Examples include analyzing graph properties such as computing the shortest path, finding out the maximum-flow between two given nodes in a graph (Cormen et al. 2001), detecting community structures (Newman 2004), etc. With the exception of some recent work, not much has been done in the area of analyzing graph data from a privacy preserving perspective. However, due to the sensitive and/or personal nature of the data, the public release of (or access to) such data, often poses considerable privacy risk to the individuals/entities involved (Backstrom et al. 2007). We note that there has been extensive research focusing on privacy protection of tabular data (Samarati and Sweeney 1998). Due to the interconnectivity of nodes in a graph, privacy preserving graph analysis and anonymization has been recognized as a challenging area (Kleinberg 2007).

Often, the graph data is distributed across different autonomous enterprises. Consider the following scenario. In financial sectors, funds may be transferred from one bank account to another opened in a different bank. Given large volume of transactions, such transfers can eventually form a complex graph structure among bank accounts. For discovering certain illegal financial activities (e.g., money laundering and financial fraud), it requires all the data from all involved banks. However, due to privacy concerns and legal issues, banks would not be willing to disclose their customers' data while without privacy guarantee. Similar concerns among other domains, including telecommunications, law enforcement, homeland security, etc. further

provide motivations for us to address privacy constraints when dealing with distributed graph data (Cauley 2006; Gross and Acquisti 2005; Kleinberg 2007).

The previous studies in the literature have primarily focused on centralized model where a single graph ownership is considered (e.g., Cormode et al. 2008; Hay et al. 2008; Liu and Terzi 2008; Zhou and Pei 2008). Under such centralized graph model, certain perturbation operations are typically performed on the single original graph meanwhile aiming to satisfy specified privacy protection and utility requirements. The perturbed (anonymized) graph is then released so that it can be used for graph analysis and any other purposes.

In particular, the perturbation techniques include adding/deleting edges and/or nodes as well as clustering nodes (Hay et al. 2008; Liu and Terzi 2008; Ying and Wu 2008). Three types of private information disclosure have been identified in the literature (Liu and Terzi 2008): identity disclosure, link disclosure, and attribute disclosure. Identity disclosure occurs when a real-world entity or individual is mapped to a particular node in the released graph. Link disclosure occurs when any new sensitive relationships (e.g., edge existence) between two entities are revealed. In graph data, attributes may refer to the contents that are associated with nodes and/or edges (e.g., edge capacity in the maximum-flow problem). Attribute disclosure occurs when an adversary is able to determine the value of an entity's or a relationship's attribute that is intended to keep private. Identity disclosure often leads to link and attribute disclosure.

As with any privacy preserving data releasing and analysis, utility of perturbed graphs is an important measure of anonymizing quality. Indeed, it should never be ignored when designing and evaluating techniques for privacy protection purposes. For example, adding random noise does not compromise privacy but very likely rendering the transformed data useless. Towards retaining utility, the existing works have mainly focused on preserving one property or the other

of the original graph. In other words, it is quite ad hoc. For instance, Gao et al. only focus on preserving shortest distances in anonymized graph (Gao et al. 2011); Hay et al. mainly aim to retain the degree distribution of the original graph (Hay et al. 2009); and Ying and Wu propose to preserve the eigenvalue of a graph (Ying and Wu 2008). In some cases (Cheng et al. 2010; Liu and Terzi 2008; Zhou and Pei 2008), only empirical analysis of graph properties is conducted on the perturbed graphs. There is no any guarantee on whether or not the perturbed graph will retain the original graph properties.

The above review of the current literature reveals that the research area of privacy preserving graph publishing and analysis is fragmented with ad hoc perturbation techniques when achieving narrowly specified privacy and utility goals. None of the existing works has taken a holistic and systematic approach. Furthermore, the practical distributed model has not been considered in the literature yet, even though it is an area that has recently attracted more attention from academics and practitioners (Cauley 2006; He et al. 2008). To fill the gap, the paper proposes an integrative framework for enabling privacy preserving analysis of graph properties in distributed graphs.

The paper is organized as follows. Following a literature review on privacy preserving graph publishing and analysis, we propose a framework for enabling privacy preserving analysis of graph properties in distributed graphs. Finally, we conclude the paper and provide suggestions for the future work.

LITERATURE REVIEW AND RELATED STUDIES

In this section, we briefly review relevant literature on privacy preserving graph analysis and publishing. In the area of privacy preserving in graphs, specific ad hoc techniques have been proposed to protect pre-defined privacy requirements. Some of the existing work primarily

focuses on preventing node re-identification through anonymization (e.g., Hay et al. 2008; Liu and Terzi 2008; Zhou and Pei 2008). On the other hand, some aim to protect link/association disclosure between nodes through edge adding/deleting and/or node clustering (e.g., Bhagat et al. 2009; Cormode et al. 2008; Ying and Wu 2008).

Specifically, Hay et al. formalize the ad hoc queries of adversarial knowledge and showing the risks of node re-identification in real datasets (Hay et al. 2008). They also propose a technique for generalizing a graph by grouping nodes into partitions and then only publish the number of nodes in each partition, along with the density of edges that exist within and across partitions. Zhou and Pei propose an anonymity approach against 1-neighborhood attack by generalizing labels and adding edges so as to make at least k nodes having the same isomorphic neighborhood subgraphs (Zhou and Pei 2008). Liu and Terzi present a degree anonymity algorithm which guarantees at least k nodes having the same degree in the anonymized graph (Liu and Terzi 2008). We shall note that all the above work is built on the principle of k-anonymity (Samarati and Sweeney 1998). Previously, k-anonymity has been extensively used for tabular data.

Rather than focusing on preventing node reidentification, Ying and Wu study how edge-based graph perturbations would affect the spectrum properties (eigenvalues) of the graph and how the perturbation would protect link privacy (Ying and Wu 2008). Cormode et al. focus on the problem of anonymizing bipartite graph and present so called (k, l)-groupings algorithm to hide the actual mapping between two types of entities. The groupings lead to strong tradeoffs between privacy and data utility (Cormode et al. 2008). This work is further extended by Bhagat et al. (Bhagat et al. 2009). An interaction bipartite graph is considered and each node in the graph is associated with a set of properties. The proposed approach is based on grouping the entities

into classes, and thus masking the mapping between entities and the nodes that represent them in the anonymized graph.

We note that our work extends and complements the above existing work and propose a unified framework in the context of distributed graphs for enabling privacy preserving analysis of graph properties.

THE PROPOSED FRAMEWORK

In this section, we first describe the problem of privacy preserving computation of graph properties in distributed graphs. Then, we provide an overview of the proposed framework followed by the more detailed description of each module in the framework.

Problem Description

In this study, we consider a distributed graph data model. The private graphs are distributed among several data owners. Specifically, we denote these graphs as G_1, G_2, …, G_k owned by data owner 1, data owner 2, …, data owner k, respectively. The interconnections among these graphs result in one single overall graph G. In other words, each of the graphs G_1, G_2, …, G_k is a private subgraph of graph G. The ability of computing the graph properties of graph G would be beneficial to not only each proprietary data owner but also to other users including various relevant stakeholders. Such ability typically requires the availability of each subgraph of graph G. However, due to the privacy and legal concerns as mentioned earlier, each data owner may not be willing to disclose his/her own private subgraph while without being assured of private data protection.

In the above distributed setting, we are interested in developing a framework which facilitates the computation of graph properties of the overall graph G with privacy preservation. In particular, the system based on the framework shall be able to answer a user's utility queries

regarding various graph properties. Those properties may include shortest path length, degree distribution, clustering coefficient, maximum-flow between two given nodes, etc. Specific privacy protection requirements shall be satisfied depending on the needs of data owners. In the following sections, we describe each functional module of the framework. The functionalities are built on both utility and privacy requirements.

Overview of the Framework

The proposed framework consists of three modules as depicted in Figure 1. The first module is Graph Transformation Module. This module is to transform the original graph into a perturbed one so that the required privacy protection is assured while maintaining certain graph properties. The second module is to securely integrate the input graphs so that one single overall graph can be made available without revealing which edge (node) originally belongs to which data owner. This provides further privacy protection in respect to the graph data. Finally, a graph properties query evaluation module takes user's queries and return the value of the querying graph properties.

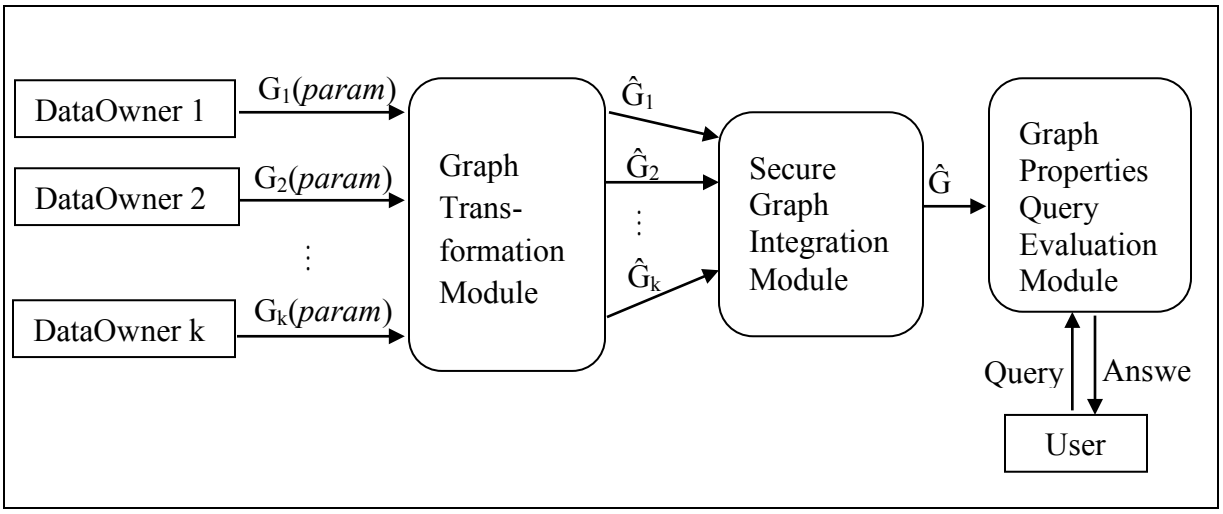

Figure 1. The Proposed Framework

Graph transformation module

In this module, the graph perturbation operations are performed. The inputs of this module are the original private graphs, each of which is owned by the respective data owner. The date owner may specify the particular privacy requirement and graph properties retaining goal that should be satisfied. Such requirements can be passed on to the module through a parameter called *param*.

The transformation functionality of this module is similar to the graph perturbation that is provided in the existing studies (Cheng et al. 2010; Hay et al. 2008; Liu and Terzi 2008; Ying and Wu 2008). For instance, in order to prevent node re-identification, random perturbation technique through a sequence of random edge-deletion followed by edge-insertions is proposed (Hay et al. 2008). However, rather than just focusing on one specific perturbation technique, the transformation function in the module may include the various perturbation techniques that are targeted to meet the corresponding privacy and graph properties retaining requirements.

In essence, the graph transformation module acts like a blackbox. The module can be viewed solely in terms of its input, output and transformation characteristics without any knowledge of its internal workings, that is, its implementation is "opaque". As mentioned earlier, the input includes the original graph with the parameter of specified privacy and graph properties retaining requirements. The module outputs the perturbed version of the original graph.

Secure graph integration module

Since the graphs are distributed among the data owners, the secure graph integration module provides a mechanism that the single integrated overall graph $\hat{G}$ can be made available without allowing any parties to learn any other information beyond what is revealed by the

integrated graph. The implementation of such module typically involves the use of encryption cryptosystem to ensure privacy and security. Such cryptosystems have been studied in the field of computer science (Cramer et al. 2001; Goldreich 2004). Thus, we can directly utilize some of the developed tools to implement this secure integration module (Cramer et al. 2001; Pohlig et al. 1978; Rivest et al. 1976).

Notably, with the privacy and security guarantees resulting from graph transformation module and secure graph integration module, the integrated graph can be released after the integration. That is, users may use their own graph analysis tools to analyze the integrated graph $\hat{G}$. This provides the flexibility as well as achieves the ultimate goal of privacy preserving graph analysis. Nevertheless, we include a graph query evaluation module in the proposed framework for the completeness and the intention of meeting various users' needs.

Graph query evaluation module

In this module, the graph analysis tools are included to analyze graph properties of the final integrated graph. It takes the user's query as an input; such query is evaluated by the graph analysis tools and the module returns the result to the user accordingly. Indeed, analyzing graph properties is a well-studied topic in graph theory and various graph algorithms have been proposed and designed (Chartrand 1985; Cormen et al. 2001). Thus, we can use the tools that already developed for analyzing graph properties or develop the corresponding tools according to the system requirements for implementing the framework.

CONCLUSION AND FUTURE WORK

The research reported here seeks to address the privacy issues involved in the graph data. We particularly focus on the distributed graph model. To fill the gap in the existing literature of privacy preserving graph analysis and publishing, we propose a unified framework for enabling

privacy preserving analysis of graph properties in distributed graphs. The three modules in the framework are intended to ensure privacy protection while retaining graph properties of the graph as well as answer user queries on graph properties.

For future work, we would like to initiate new cases and use the real world graph data to evaluate the framework. Our ultimate goal is to further refine the framework and implement it for practical use.

REFERENCES

Backstrom, L., Dwork, C., and Kleinberg, J. 2007. "Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography," in *Proceedings of the 16th international conference on World Wide Web*, New York, NY, pp. 181–190.

Bhagat, S., Cormode, G., Krishnamurthy, B., and Srivastava, D. 2009. "Class-based graph anonymization for social network data," in *Proceedings of the VLDB Endowment* (2:1) pp. 766–777.

Bhagat, S., Cormode, G., Krishnamurthy, B., and Srivastava, D.2010. "Privacy in dynamic social networks," In M. Rappa, P. Jones, J. Freire, and S. Chakrabarti (eds.), *WWW*, pp. 1059–1060.

Cauley, L., 2006. "NSA has massive database of Americans' phone calls," USA TODAY, http://www.usatoday.com/news/washington/2006-05-10-nsa_x.htm, accessed on 12/2/2011.

Chartrand, G. 1985. *Introductory Graph Theory*. NY: Dover Publication.

Cheng, J., Fu, A. W., and Liu, J. 2010. "K-isomorphism: privacy preserving network publication against structural attacks," in *Proceedings of the 2010 international conference on Management of data*, pp. 459–470.

Cormen, T. H., Leiserson, C. E., and Rivest, R. L. 2001. *Introduction to Algorithms*. New York: McGraw-Hill.

Cormode, G., Srivastava, D. , Yu, T., and Zhang, Q. 2008. "Anonymizing bipartite graph data using safe groupings," In *Proceedings of the VLDB Endowment* (1:1), pp. 833-844.

Cramer, R., Damgard, I., and Nielsen, J. B. 2001. "Multiparty Computation from Threshold Homomorphic Encryption," in *Proceedings of the International Conference on the Theory and Application of Cryptographic Techniques*, pp. 280-299.

Gao, J., Yu, J. X., Jin, R., Zhou, J. ,Wang, T., and Yang, D. 2011. "Neighborhood-privacy protected shortest distance computing in cloud," ," in *Proceedings of the 2011 international conference on Management of data*, pp. 409–42.

Gross, R. and Acquisti, A. 2005."Information revelation and privacy in online social networks," in Proceedings of the 2005 ACM workshop on Privacy in the electronic society, pp.71-80.

Goldreich, O. 2004. *The Foundations of Cryptography*, Cambridge University Press, Cambridge, UK.

Hay, M., Li, C., Miklau, G., and Jensen, D. 2009 . "Accurate estimation of the degree distribution of private networks," in *proceedings of IEEE International Conference on Data Mining*, W.Wang, H. Kargupta, S. Ranka, P. S. Yu, and X.Wu (eds.), pp. 169–178.

Hay, M., Li, C., Miklau, G., Jensen, D., Towsley, D. F., and Weis, P. 2008. "Resisting structural re-identification in anonymized social networks," In *Proceedings of the VLDB Endowment* (1:1), pp. 102–114.

He, X., Shafiq, B. , Vaidya, J. , and Adam, N. R. 2008. "Privacy-preserving link discovery." In *Proceedings of the 2008 ACM symposium on Applied computing*, R. L. Wainwright and H. Haddad(eds.), pp. 909–915.

He, X., Vaidya, J. , Shafiq, B., Adam, N. R., and Atluri V. 2009. "Preserving privacy in social networks: A structure-aware approach," In *Web Intelligence*, pp. 647–654.

Kleinberg, J. M. 2007. "Challenges in mining social network data: processes, privacy, and paradoxes," in *Proceedings of ACM Knowledge Discovery and Data mining*, P. Berkhin, R. Caruana, and X. Wu (eds.), pp. 4–5.

Korolova, A., Motwani, R., Nabar, S. U., and Xu, Y. 2008. "Link privacy in social networks," in *proceedings of IEEE International Conference on Data Engineering*, pp. 1355–1357.

Lindamood, J., Heatherly, R., Kantarcioglu, M., and Thuraisingham, B. 2009. "Inferring private information using social network data," in *proceedings of the 18th International World Wide Web Conference*, Madrid, Spain, pp. 1145–1146.

Liu, K. and Terzi, E. 2008. "Towards identity anonymization on graphs," In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, New York, NY, pp. 93–106.

Liu, L., Wang, J., Liu, J.,and Zhang, J. 2009. "Privacy preservation in social networks with sensitive edge weights," in *proceedings of the 2009 SIAM International Conference on Data Mining*, Sparks, NV, pp. 954–965.

Newman, M. E. J. 2004. "Detecting community structure in networks," *The European Physical Journal B - Condensed Matter* (38:2), pp. 321–330.

Pohlig, S. C. and Hellman, M. E. 1978. "An improved algorithm for computing logarithms over GF(p) and its cryptographic significance", *IEEE Transactions on Information Theory* (IT-24), pp. 106-110.

Rivest, R. L., Shamir, A. and Adleman, L. 1978. "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems", *Communications of the ACM* (21:2), pp. 120-126.

Samarati, P. and Sweeney, L. 1998. "Generalizing data to provide anonymity when disclosing information (abstract)," *inProceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pp. 188.

Ying, X. and Wu, X. 2008. "Randomizing social networks: a spectrum preserving approach," in *proceedings of the 2008 SIAM International Conference on Data Mining*, Atlanta, GA, pp. 739–750.

Zheleva, E. and Getoor, L. 2009. "To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles," in *proceedings of the 18th International World Wide Web Conference*, Madrid, Spain, pp. 531–540.

Zhou, B. and Pei, J. 2008. "Preserving privacy in social networks against neighborhood attacks," in *Proceedings of the 24th International Conference on Data Engineering*, Cancun, Mexico, pp. 506-515.