H. Langweg, M. Meier, B.C. Witt, D. Reinhardt (Hrsg.): Sicherheit 2018,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2018 273

Towards a Differential Privacy Theory for Edge-Labeled Directed Graphs

Jenni Reuben[1]

Abstract:

Increasingly, more and more information is represented as graphs such as social network data, financial transactions and semantic assertions in Semantic Web context. Mining such data about people for useful insights has enormous social and commercial benefits. However, the privacy of the individuals in datasets is a major concern. Hence, the challenge is to enable analyses over a dataset while preserving the privacy of the individuals in the dataset. Differential privacy is a privacy model that offers a rigorous definition of privacy, which says that from the released results of an analysis it is '*difficult*' to determine if an individual contributes to the results or not. The differential privacy model is extensively studied in the context of relational databases. Nevertheless, there has been growing interest in the adaptation of differential privacy to graph data. Previous research in applying differential privacy model to graphs focuses on unlabeled graphs. However, in many applications graphs consist of labeled edges, and the analyses can be more expressive, which now takes into account the labels. Thus, it would be of interest to study the adaptation of differential privacy to edge-labeled directed graphs. In this paper, we present our foundational work towards that aim. First we present three variant notions of an individual's information being/not being in the analyzed graph, which is the basis for formalizing the differential privacy guarantee. Next, we present our plan to study particular graph statistics using the differential privacy model, given the choice of the notion that represent the individual's information being/not being in the analyzed graph.

Keywords: Differential privacy, graphs, labels, analyze, utility.

1 Introduction

Analyzing the information collected in statistical databases provides valuable insights in the medical and social science research. However, the challenge is how to ensure that the public release of the results from the analysis does not compromise the privacy of the individual contributors of the dataset. This challenge is referred to as the data privacy problem and being extensively studied both in the statistics and computer science community [AW89, AS00, Be80, Ch05, Sw02].

Increasingly, more and more information is represented using graph structures, for example social network data, financial transactions and semantic assertions in the Semantic Web context. Many recent studies have investigated the data privacy problem in graph data [ZPL08].

[1] Department of Mathematics and Computer Science, Karlstad University, Sweden, jenni.reuben@kau.se

doi:10.18420/sicherheit2018_24

Among the studied privacy models, the Differential Privacy model provides a mathematical definition of data privacy that is guaranteed to the participants of a database, independent of the auxiliary information available to an adversary[3]. Intuitively, a trusted curator of the database uses a privacy preserving mechanism that satisfies the definition of differential privacy for releasing the results of the analysis performed on the database. The definition of differential privacy states that the result is 'essentially' the same when an individual participates or refrains from participating in the database. Thus, the publicly released results provide meaningful insights about the underlying population of the database yet obscure any one individual's contribution.

Previous research in applying differential privacy theory in the context of releasing graph statistics focuses on graphs with unlabeled edges. One typical example of the queries over the graph data is, '$COUNT$ all edges'. However, in many applications, the graphs consist of edges that are labeled. Accordingly, there are analyses that take into account these labels, for example to get '$COUNT$ of the edges that have certain label(s)' . It would thus be of interest to learn how to apply differential privacy to graphs with labeled edges, where different edges may have different labels. The main goal of this paper is to define the foundations for applying differential privacy to edge-labeled directed graphs.

2 Preliminaries

As a basis for presenting our work in this paper, in this section we recall the main notions defined for differential privacy both in the relational database settings [Dw08, Dw06] and in graphs without edge labels [Ka13, NRS07, TC12].

A database D is a set of rows. Consider a database D_I that contains information about a set of individuals I, where the information about each individual is captured as a separate row. Now, consider another database $D_J = D_{I \pm x}$ that includes/excludes information of one random individual x. So, D_J and D_I differ by one row and they are called neighboring databases [Dw08]. A trust worthy curator uses a privatized mechanism K that takes a database D as input and produces a result KD, which gives nearly zero evidence about whether the input database is D_I or D_J, thus obfuscating any one individual's contribution to the result.

Definition 1 (Differential Privacy [Dw08]) *A privatized mechanism K is said to give* ϵ-differential privacy*, if for any pair of databases D_I and D_J that differ by one row, and for all $S \subseteq RangeK$, it holds that: $PrKD_I \in S \leq e^{\epsilon} \times PrKD_J \in S$*

where the probabilities represent the random choices made by the privatized mechanism K and $RangeK$ denotes the set of all possible outputs of K.

[3] In this context, an adversary is an entity that intends to compromise the privacy of the participants of a database.

Dwork et al. [Dw06] presented a privatized mechanism K for continuous-valued queries, that is the classes of queries that map the database to vectors of real numbers. If the true response of a query function f is fD, for achieving ϵ-differential privacy the privatized mechanism then distorts this true response by adding appropriately chosen noise before disclosing it to the public. The noise that needs to be added to the true response is given by the sensitivity of the query function f and the chosen value of ϵ. The function sensitivity specifies what is the maximum difference that the privatized mechanism needs to bridge in the form of additive noise such that from the noisy response it is difficult to attribute that the input database is D_I or D_J. If the value of ϵ is set to a very small value, then the noise that need to be added increases. Similarly, the amount of the additive noise will be large if the sensitivity of the function is greater.

The original differential privacy definition remains essentially unchanged for graph data. However the notion of neighboring databases on which Definition 1 is based on need to be adapted to graph data. In the literature, there appear two variants of differential privacy definitions that formalize the privacy guarantee for two different notions of what it means for a pair of graphs to differ by one unit. One definition is *edge privacy*, which formalizes the differential privacy guarantee for any two graphs that differ by at most one edge [NRS07]. The second definition is *node privacy*, which deals with any pair of graphs that differ by a single node including all its adjacent edges [Ka13, TC12]. Inspired by these definitions, in the next section, we present three variants of differential privacy definitions that guarantee various levels of privacy protections for edge-labeled directed graphs.

3 Differential Privacy for Edge-labeled Directed Graphs

Let L be an infinite set of possible edge labels. An edge-labeled directed graph, hereafter simply a graph, is a tuple $G = V, E$, where V is a set of vertices, and E is a set of edges such that $E \subseteq V \times L \times V$. The privacy guarantee formalized in Definition 1 builds on the notion that the response for a query over a database is 'essentially' the same for any two databases that are neighbors (i.e they differ by a row). For graph data, as presented in Section 2 there are different possibilities to represent what it means for two graphs to differ by one unit. The following definitions specify what it means for a pair of graphs being neighbors for formalizing the differential privacy guarantee for edge-labeled directed graphs. Each possible definition of neighboring graphs provides a different semantic interpretations of the differential privacy guarantee. Hence it is important to study the privacy/utility trade-off of the chosen graph neighbor definition. First, we adapt the 'edge privacy' definition of unlabeled graphs. Accordingly, two graphs are edge-neighbors, if in one of them one edge is included/excluded independent of its label.

Definition 2 (Edge-neighboring Graphs) *Graphs* $G = V, E$ *and* $G^{'} = V^{'}, E^{'}$ *are* edge-neighbors *if* $V = V^{'}$ *and* $E^{'} = E - \{e\}$ *for some edge* $e \in E^{'}$.

Second, in accordance with the 'node privacy' definition of unlabeled graphs, two graphs are node-neighbors if one of them is obtained from the other by adding/removing one arbitrary node and all of its labeled edges.

Definition 3 (Node-neighboring Graphs) *Graphs* $G = V, E$ *and* $G^{'} = V^{'}, E^{'}$ *are* node-neighbors *if* $V^{'} = V - x$ *and* $E^{'} = E - \{v_1, l, v_2 \mid v_1 = x \vee v_2 = x\}$ *for some* $x \in V$.

In the third adaptation, the differential privacy guarantee is built on the notion of graphs that differ by a set of labeled outedges of a node. The intuition is, in some of the applications of edge-labeled directed graphs such as RDF, an entity is represented by its associations and particular associations, indicated by certain labels, 'uniquely' identify that entity.

Definition 4 (QL-Outedge Neighboring Graphs) *Let* QL *be a subset of* L. *Graphs* $G = V, E$ *and* $G^{'} = V^{'}, E^{'}$ *are* QL-outedge-neighbors *if* $V = V^{'}$ *and* $E^{'} = E - \{v_1, l, v_2 \mid v_1 = x \text{ and } l \in QL\}$ *for some* $x \in V$.

Example 1. Let $QL = \{b\}$, given the QL, in Fig 1, 2 graphs $G = V, E$ and $G^{'} = V^{'}, E^{'}$ are QL-outedge neighbors, because in $G^{'}$ for the vertex 'y', there does not exists any of the outedges with the labels in QL. Similarly in Fig 3, 4, let $QL^{'} = \{a, b\}$, in $G^{'''} = V^{'}, E^{'}$ for the vertex 'y' all its outedges with all the labels in $QL^{'}$ are excluded. So, graphs $G^{''}$ and $G^{'''}$ are $QL^{'}$-outedge neighbors.

Given the different definitions of what it means for two edge-labeled directed graphs to be neighbors, the privacy guarantee of a privatized mechanism is formalized as:

Definition 5 (Differential privacy for edge-labeled directed graphs) *A privatized mechanism* K *for edge-labeled directed graphs satisfies* ϵ-edge differential privacy *(respectively,* ϵ-node differential privacy*, or* ϵ-QL-edge-labeled differential privacy *for some* $QL \subseteq L$*), if for every pair of graphs* G *and* $G^{'}$ *that are edge-neighbors (respectively, node-neighbors, or* QL*-outedge neighbors), and for all* $S \subseteq RangeK$*, it holds that:* $PrKG \in S \leq e^{\epsilon} \times PrKG^{'} \in S$.

where the probabilities represent the random choices made by K *and* $RangeK$ *denotes the set of all possible outputs of* K.

4 Discussion and Future Outlook

Differential privacy for graph data offers a strong mathematical privacy guarantee for releasing graph statistics, but the semantics of the privacy protection rests on the definition of neighboring graphs. Thus, the choice of the definition of neighboring graphs impacts the privacy/utility trade-off offered by the privatized mechanism.

In the case of edge-neighbors (i.e Definition 2), edge disclosure is protected by the privatized mechanism. For some applications such as analysis on email communication graphs where

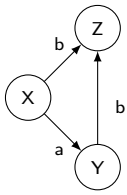

Fig. 1: G

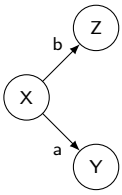

Fig. 2: $G^{'}$

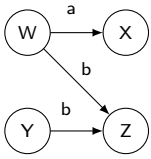

Fig. 3: $G^{''}$

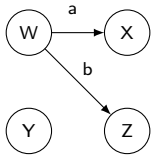

Fig. 4: $G^{'''}$

the relationships are sensitive, the edge-neighbors definition is sufficient. The node-neighbors definition (i.e Definition 3), mirrors the notion of neighboring databases formalized in Definition 1. However, the structural properties of graphs may introduce a huge gap between two neighboring graphs, consequently the function sensitivity is large, which need to be obfuscated by the privatized mechanism in order to protect the input graph that produces the result. The type of graph data analyses that a privatized mechanism can support under this definition yet producing accurate results may thus be limited. In many applications of edge-labeled directed graphs, the labels play a significant part in defining the relationships among the nodes. We assume that edges with labels from a particular domain-specific subset of all labels, 'uniquely' identify a node in an edge-labeled directed graph. Further, we focus on outedges of a node because it represents the contributions that this node makes to the graph dataset. Hence, this semantically captures the notion of an individual being in one graph but not in another graph similar to the private data represented as tuple in the relational databases. We propose that this definition of neighboring graphs offers another level of privacy protection than the edge-neighbors definition. Further, we hypothesize that under this definition the noise required to bridge the gap between the two neighboring graphs will be less than the node-neighbor definition, thus increasing the utility of the results returned by the privatized mechanism. Nevertheless, it would be interesting to study the type of graph statistics that are accurate and how accurate the results are under this scheme of things.

To test the hypothesis, as a next step we begin to focus on degree distribution as a particular graph statistics over edge-labeled directed graphs. Degree distribution of a graph gives a simplistic understanding of the structure of a graph. Degree distribution is a vector of real numbers that represent the degrees of the nodes in a graph. We plan to employ the privatized mechanism introduced by Dwork et al. [Dw06] to answer the degree distribution queries over edge-labeled directed graphs. Most importantly, we plan to study the privacy/utility trade-off of this privatized mechanism when the different neighbor definitions are chosen (i.e., Definition 2 versus Definition 3 versus Definition 4 with different QL). To this end, we plan to generate different edge-labeled graphs by systematically varying the structural characteristics, which constitute the datasets that are protected by the envisioned privatized mechanism. Based on these graphs, we aim to analyze the accuracy of the degree distribution query under the edge-neighbors versus the QL-outedge-neighbors (for different QL). Further, we aim to analyze the privacy/utility trade-off of our mechanism versus the Hay et al.'s

mechanism [Ha09], which supports degree distribution queries but requires a post processing step for improving the utility of the results. We also plan to evaluate the privacy/utility trade-off of our mechanism over a set of real-world graphs.

As a long-term goal, we move on to study other types of graph analyses, in particular, analyses that take into account the edge labels (e.g., in the case of degree distribution, to estimate the degree distribution that represent the edges with certain labels). From the results of these experiments that analyze the privacy/utility trade-off when different neighboring edge-labeled graph definitions are chosen, we aim to investigate different ways to optimize the privacy/utility trade-off in particular for QL-outedge neighbor definition.

Acknowledgments. I thank my advisor Olaf Hartig for the discussions and feedback that enable this work. I also thank Simone Fischer-Hübner for her feedback.

References

[AS00] Agrawal, Rakesh; Srikant, Ramakrishnan: Privacy-preserving Data Mining. In: Proc. of the 2000 ACM SIGMOD Int. Conf. on Management of Data. 2000.

[AW89] Adam, Nabil R.; Worthmann, John C.: Security-control Methods for Statistical Databases: A Comparative Study. ACM Comput. Surv., 21(4):515–556, 1989.

[Be80] Beck, Leland L.: A Security Machanism for Statistical Database. ACM ToDS, 5(3), 1980.

[Ch05] Chawla, Shuchi; Dwork, Cynthia; McSherry, Frank; Smith, Adam D; Wee, Hoeteck: Toward Privacy in Public Databases. In: TCC. volume 3378, 2005.

[Dw06] Dwork, Cynthia; Mcsherry, Frank; Nissim, Kobbi; Smith, Adam: Calibrating noise to sensitivity in private data analysis. In: Proc. of 3rd TCC. 2006.

[Dw08] Dwork, Cynthia: Differential Privacy: A Survey of Results. In: TAMC: Proc. 5th Int. Conf.˙ 2008.

[Ha09] Hay, M.; Li, C.; Miklau, G.; Jensen, D.: Accurate Estimation of the Degree Distribution of Private Networks. In: 9th IEEE Int. Conf. on DM. 2009.

[Ka13] Kasiviswanathan, S. P.; Nissim, K.; Raskhodnikova, S.; Smith, A.: Analyzing Graphs with Node Differential Privacy. In: Proc. of 10th TCC (2013). 2013.

[NRS07] Nissim, K.; Raskhodnikova, S.; Smith, A.: Smooth Sensitivity and Sampling in Private Data Analysis. In: 39th ACM Symp. on Theory of Computing. 2007.

[Sw02] Sweeney, L.: k-anonymity: A model for protecting privacy. Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10:557–570, 2002.

[TC12] Task, C.; Clifton, C.: A Guide to Differential Privacy Theory in Social Network Analysis. In: Int. Conf. on Advances in SN Analysis and Mining. 2012.

[ZPL08] Zhou, B.; Pei, J.; Luk, W.: A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data. SIGKDD Ex. Nl., 2008.