Group Differential Privacy-preserving Disclosure of Multi-level Association Graphs

Balaji Palanisamy
School of Information Sciences
University of Pittsburgh
Pittsburgh, USA
Email: bpalan@pitt.edu

Chao Li
School of Information Sciences
University of Pittsburgh
Pittsburgh, USA
Email: chl205@pitt.edu

Prashant Krishnamurthy
School of Information Sciences
University of Pittsburgh
Pittsburgh, USA
Email: prashk@pitt.edu

***Abstract*—Traditional privacy-preserving data disclosure solutions have focused on protecting the privacy of individual's information with the assumption that all aggregate (statistical) information about individuals is safe for disclosure. Such schemes fail to support group privacy where aggregate information about a group of individuals may also be sensitive and users of the published data may have different levels of access privileges entitled to them. We propose the notion of ϵ_g-*Group Differential Privacy* that protects sensitive information of groups of individuals at various defined privacy levels, enabling data users to obtain the level of access entitled to them. We present a preliminary evaluation of the proposed notion of group privacy through experiments on real association graph data that demonstrate the guarantees on group privacy on the disclosed data.**

I. Introduction

In the age of Big Data, organizations and governments can obtain rich information and insights by mining large volumes of data that get generated at an unprecedented velocity, volume and scale. Data privacy becomes a critical barrier in effectively leveraging large-scale data analytics due to serious privacy risks. Publishing and maintaining data that contains sensitive information about individuals is a challenging problem. Such sensitive datasets may include private information such as medical information, patient records, census information or sales transactions made by customers. Private data often arise in the form of associations between entities in real world such as the drugs purchased by patients in a pharmacy store or the movies rated by viewers in a movie rating database or the publications authored by authors in a double-blind review conference [1]. Such associations are best captured as bipartite association graphs with nodes representing the entities (e.g., drugs and patients) and the edges correspond to the associations between them (e.g., Patient Bob purchased the Insulin drug).

Differential privacy [2], [3] provides a model to quantify the disclosure risks by ensuring that the published statistical data does not depend on the presence or absence of an individual record in the dataset. In the past, data privacy schemes [2], [3], [4] have largely focused on applying differential privacy to protect the privacy of individuals' information while supporting aggregate (statistical) queries on groups of individuals. Such schemes were developed with an intrinsic assumption that all aggregate(statistical) information about individuals are safe for disclosure and therefore, become inapplicable in scenarios when the aggregate information itself can be sensitive and needs protection. In general, we consider that sensitive information may arise as: (i) an individual sensitive value indicating an individual's private information (e.g., did buyer 'Bob' purchase the drug 'insulin'?) or (ii) a statistical value representing some sensitive statistics about a group/subgroup of individuals (e.g., the total number of 'Psychiatric' drugs made by buyers in a given neighborhood represented by a zipcode). While traditional privacy preserving mechanisms have solely focused on protecting individual's sensitive values, our work takes a new perspective on privacy-preserving data publishing focusing the problem of privacy protection when aggregate (statistical) information about a group of individuals is private and needs protection.

In this paper, we propose the notion of ϵ_g-group differential privacy that provides guaranteed protection of aggregate information of a group of individuals in a given dataset. We present a preliminary evaluation of the proposed notion of group privacy on real association graph data that demonstrate the guarantees on group privacy on the disclosed data.

II. Group differential privacy

In this section, we review the definition of conventional individual differential privacy and present the proposed notion of ϵ_g-group differential privacy.

A. Individual Differential Privacy

Differential privacy is a classical privacy definition [2], [3] that makes very conservative assumptions about the adversary's background knowledge and bounds the allowable error in a quantified manner. In general, differential privacy is designed to protect a single individual's privacy by considering adjacent data sets which differ only in one record. A data set D can be considered as a subset of records from the universe U, represented by $D \in \mathbb{N}^{|U|}$, where $\mathbb{N}$ stands for the non-negative set and D_i is the number of element i in $\mathbb{N}$. For example, if $U = \{a, b, c\}$, $D_1 = \{a, b, c\}$ and $D_2 = \{a, c\}$ can be represented as $\{1, 1, 1\}$ and $\{1, 0, 1\}$ respectively. Based on this representation, it is appropriate to use l_1 distance (Manhattan distance) to measure the distance between data sets, which leads us the notion of adjacent data sets as follows.

Definition *1 (*Adjacent Data set*): Two data sets D_1, D_2 are adjacent if $||D_1 - D_2||_1 = 1$.*

ϵ - differential privacy is designed to protect the privacy between adjacent data sets which differ only in one record. In other words, it protects the individual-level privacy and we define ϵ_i - individual differential privacy as:

DEFINITION *2 (*INDIVIDUAL DIFFERENTIAL PRIVACY*): A randomized algorithm $\mathcal{A}$ guarantees ϵ_i-differential privacy if for all adjacent data sets D_1 and D_2 differing by at most one record, and for all possible results $\mathcal{S} \subseteq Range(\mathcal{A})$,*

$$Pr[\mathcal{A}(D_1) = \mathcal{S}] \leq e^{\epsilon_i} \times Pr[\mathcal{A}(D_2) = \mathcal{S}]$$

where the probability space is over the randomness of $\mathcal{A}$.

B. Group Differential Privacy

In this work, we extend the conventional notion of individual differential privacy to protect group privacy at various group granularity levels. We focus on the scenarios where one needs to protect group-level privacy in addition to individual privacy, where a group consists of a set of individuals. We define the proposed notion of ϵ_g - group differential privacy by considering adjacent data sets from a group privacy perspective. Specifically, we consider the universe, $U = \cup_{i=1}^{n} G_i$ is partitioned into n non-overlapping subgroups $G = \{G_1, ..., G_n\}$ with each record of U joining only one subgroup $G_i \in G$. Therefore, the overall data set space can be represented as $D = \{D_i | D_i = \cup_{i \in I} G_i, G_i \in G, I \subseteq \{1, ..., n\}\}$. This leads to a number of group-level adjacent data sets. Formally, group-level adjacent data sets are defined as:

DEFINITION *3 (*GROUP-LEVEL ADJACENT DATA SETS*): Two data sets D_1 and D_2 are group-level adjacent data sets of each other if $\exists G_i \in G$ such that $D_1 = D_2 \cup G_i$.*

Thus the notion of ϵ_g- group differential privacy based on group level adjacent datasets is defined as:

DEFINITION *4 (*GROUP DIFFERENTIAL PRIVACY*): A randomized algorithm $\mathcal{A}$ guarantees ϵ_g-group differential privacy if for all adjacent data sets D_1 and D_2 differing by at most one group, $G_i \in G$, and for all possible results $\mathcal{S} \subseteq Range(\mathcal{A})$,*

$$Pr[\mathcal{A}(D_1) = \mathcal{S}] \leq e^{\epsilon_g} \times Pr[\mathcal{A}(D_2) = \mathcal{S}]$$

where the probability space is over the randomness of $\mathcal{A}$.

III. GROUP PRIVACY-AWARE DISCLOSURE

In this section, we demonstrate the effectiveness of the proposed notion of group differential privacy in the context of bipartite association graphs that capture real world associations between entities. We apply the notion of group differential privacy to the disclosure of bipartite graphs with the objective of releasing different levels of group-level aggregate information conforming to group differential privacy guarantees. The data transformation process consists of two phases. We first partition the given bipartite graph through several rounds of specialization to form multiple levels of groups of subgraphs through Exponential Mechanism [4]. Here, the top level grouping consists of all nodes on one side of the graph (left or right side of the bipartite graph). In the subsequent levels, the nodes are continuously partitioned into two parts through an Exponential Mechanism [4]. In the second phase, the transformation process injects noise to the subgraphs induced by each group level through a Gaussian Mechanism [3] so that group differential privacy can be guaranteed.

We present a preliminary experimental evaluation of the effectiveness of the group differential privacy aware disclosure scheme using a real association dataset, DBLP (*http://dblp.uni-trier.de/xml/*), which contains 1295100 authors, 2281341 papers and 6384117 associations. We partition the entire data set, represented by a bipartite graph, for nine times to form nine group levels. Specifically, the group level 9 is the entire dataset and the group level 1 is the most fine-grained group level. Each group in level i is split to 4 subgroups in level $i-1$. Here two sub groups correspond to the left side nodes of the bipartite graph and the other two sub groups refer to the right side nodes of the bipartite graph. The level 0 is the individual user level where each group contains only one node. Based on a Gaussian Mechanism [3], different amounts of noises are injected to the count query result (what is the number of associations in the dataset?) to protect differential privacy of the different level of groups, denoted by $l_{9,i}$, where $i \in [0, 7]$. The performance is measured by the relative error rate $RER = \frac{|P-T|}{T}$, where P and T denotes the perturbed and actual answers respectively.

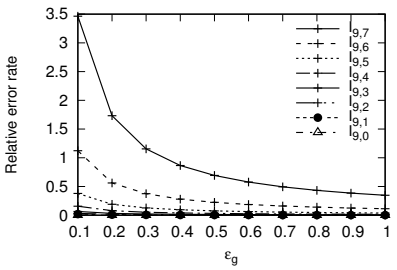

Fig. 1: Impact of ϵ_g

In Figure 1, we change the privacy budget ϵ_g to evaluate the impact of RER of different group levels. When $\epsilon_g = 0.999$, all the eight levels show small relative error, RER, and level $I_{9,1}$ only generates 0.2% RER. The RER increases to 0.33% at level $I_{9,2}$ and finally reaches 35% at level $I_{9,7}$. As can be seen, the users with lowest privilege, who can only get information of $I_{9,7}$, requiring protection of differential privacy of group level L_7, are given highly perturbed information. However, the RER for $I_{9,6}$ and $I_{9,5}$ reduce to 11% and 4% respectively. The higher the privilege a user has, the more accurate and sensitive information she can obtain. When ϵ_g is decreased, RER for all the information levels gradually increases. When ϵ_g goes down to 0.1, since the budget is highly restricted, more noise has to be injected, which makes RER for all the information levels increase significantly, especially for $I_{9,7}$ and $I_{9,6}$. However, even in this extreme case, the information levels from $I_{9,5}$ to $I_{9,0}$ still show acceptable utility with low RER, which shows that the group differential privacy can be protected for group levels generated through the specialization, even if the privacy budget is considerably small. Overall, we infer that the proposed techniques are effective, scalable and provide the required guarantees on group privacy.

REFERENCES

[1] G. Cormode, D. Srivastava, T. Yu, *et al.* Anonymizing Bipartite Graph Data using Safe Groupings In *VLDB*, 2008.

[2] C. Dwork , F. McSherry , K Nissim , *et al.* Calibrating noise to sensitivity in private data analysis. *Theory of cryptography*, Springer Berlin Heidelberg, 265-284, 2006.

[3] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4), 211-407, 2013.

[4] F. McSherry and K. Talwar. Mechanism design via differential privacy. *Foundations of Computer Science (FOCS'07)*, 48th Annual IEEE Symposium on. IEEE, 94-103, 2007.