2016 American Control Conference (ACC)
Boston Marriott Copley Place
July 6-8, 2016. Boston, MA, USA

Diffusing Private Data over Networks

Fragkiskos Koufogiannis and George J. Pappas

Abstract— **The emergence of social and technological networks has enabled rapid sharing of data and information. This has resulted in significant privacy concerns where private information can be either leaked or inferred from public data. The problem is significantly harder for social networks where we may reveal more information to our friends than to strangers. Nonetheless, our private information can still leak to strangers as our friends to their friends and so on. In order to address this important challenge, in this paper, we present a privacy-preserving mechanism that enables private data to be diffused over a network. In particular, whenever a user wants to access another users' data, other than a neighbor, the proposed mechanism returns a differentially private response that ensures that the amount of private data leaked, depends on the distance between the two users in the network. While allowing global statistics to be inferred by network analysts, our mechanism guarantees that no individual user, or a group of users, can harm the privacy guarantees of any other user. We illustrate our mechanism with an example on a Facebook ego-network where a user shares her infection status.**

I. Introduction

In the era of social networks, individuals' profiles include an increasing amount of private information. Besides users' intention to share this information for social interaction, their private data enables systems such as location-based services and collaborative recommender engines, that is, systems that are not part of their friendship network. Therefore, although users consent to share their private data with their friends, when this is not the case, severe privacy concerns are raised.

Traditionally, these privacy concerns are mitigated by restricting access rights (e.g. on Facebook); more precisely, only users indicated as friends are granted access to each user's personal information. However, such an approach has severe limitations as follows: first, this scheme is inflexible since users cannot be partitioned into exactly two groups, i.e. friends and strangers. Instead, privacy concerns gradually increase from family members and friends, to acquaintances, and finally, strangers. Second, a scheme based on access rights keeps sensitive information local, which limits the ability of inferring statistics of the whole network. For instance, consider network analysts who are interested in statistics over the whole population of the social network such as population density maps and epidemic monitoring. This limits the utility of the network. Hence, an alternative mechanism that allows global statistics on the whole population and respecting individuals' privacy is needed.

Mechanisms for providing privacy guarantees are differential privacy [1] and information theoretic privacy [2]. However, most of the previous approaches do not consider variable privacy levels of a network, where the level of privacy depends on friendship distance. Hereafter, we consider a network where users wish to share their private data under privacy guarantees, where the strength of these guarantees is quantified by the distance on the graph. Within the context of a social network, users wish to communicate accurate information with little privacy guarantees with their close friends, whereas, they desire strong privacy guarantees whenever their private data is communicated to distant areas of the network. From the network analyst's point of view, statistics over the whole network need to be possible while ensuring the privacy guarantees.

Multiple privacy-preserving frameworks that formalize privacy guarantees have appeared in the literature, e.g. [2], [3]. Commonly, privacy-preserving approaches add artificial noise to the accessed private data. This noise is designed such that the resulting response conveys little information about the private data. Specifically, an information-theoretic approach [2] constrains the mutual information between the private data and the released signal. Similarly, differential privacy [3], [1] requires that the statistics of the noisy response should be *almost* independent of perturbations of the private data. In this work, we adopt the framework of differential privacy because of its strong privacy guarantees, yet the underlying problem can be formulated under other privacy frameworks.

Within differential privacy, an extensive family of privacy-preserving mechanisms has emerged. The application range of these mechanisms varies from solving linear problems [4], [5], distributed convex optimization [6], Kalman filtering [7], and consensus that preserves the network topology [8] to smart metering [9], [10] and traffic flow estimation [11]. In particular, the problems introduced in the aforementioned line of research share a common underlying abstract problem that can be stated as follows: given the private data and a predefined privacy requirement, we need to design a differentially private algorithm, called mechanism, which accurately approximates a desired quantity. Then, a single sample from the mechanism is published and is used as a proxy for the exact response, so, a curious user cannot *confidently* infer the original private data. Instead of considering a single privacy level and assuming that the response are publicly released, i.e. everyone receives the same response, in this paper, we consider the novel problem

Authors are with the Department of Electrical and Systems Engineering, University of Pennsylvania, PA, USA.

This work was supported in part by the TerraSwarm Research Center, one of six centers supported by the STARnet phase of the Focus Center Research Program (FCRP) a Semiconductor Research Corporation program sponsored by MARCO and DARPA.

978-1-4673-8681-4/$31.00 ©2016 AACC

of assigning different privacy levels for different users. Moreover, contrary to publishing the proxies, we assume that they are securely communicated to each user. Therefore, the aforementioned works do not address the problem introduced here. In [12], [13], multi-component private data and different privacy levels for different components are considered, i.e. in a user's profile, typically, stronger privacy is required for the component representing salary compared to that of age. Contrary to previous works that focus on variable privacy for different components of one's data, our paper focuses on different privacy levels depending on friendship status. The work closest to ours is [14], where the problem of relaxing the privacy level after e.g. supplementary payments to the owners of the sensitive data. Although some of the tools in [14] are leveraged to provide a solution, here, we consider a different problem which is the problem of releasing sensitive data to multiple parties with different privacy levels has not been studied before.

The paper is organized as follows. Section II informally describes the problem of diffusing private data across a social network, then, provides a model of the system, reviews differential privacy, and derives a formal statement of the problem. Section III introduces a composite mechanism based on a Markov stochastic process and presents low-complexity algorithmic implementations of this mechanism. We demonstrate our approach with an illustrative examples in Section IV which considers a Facebook ego-network where a user shares her infection status.

II. Problem Formulation

Here, the problem of releasing private information over networks (i.e. social networks) is formulated. First, we provide an informal description of the problem whose formal statements are presented in Problem 1 and Problem 2 in the end of this section. Let a network be represented as a graph $G = (\mathcal{V}, \mathcal{E})$, where each node $i \in \mathcal{V}$ is a user and each edge $(i, j) \in \mathcal{E}$ represents a friendship relation between users i and j. Also, we assume that each user i owns a sensitive data $u_i \in \mathcal{U}$, where $\mathcal{U}$ is the set of possible private data, and wishes to share their private data with the rest of the users under privacy guarantees. Specifically, user i generates an approximation y_{ij} of u_i and securely communicates y_{ij} to user j. More specifically, each user i requires her data u_i to be $\epsilon(d_{ij})$-differential privacy against user j (differential privacy is overviewed in Subsection II-B), where d_{ij} is a distance function $d_{ij} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_+$ and $\epsilon : \mathbb{R}_+ \to \mathbb{R}_+$ is a decreasing function that converts distance d to a privacy level $\epsilon(d)$. Therefore, we need to design a mechanism that generates *accurate*[*] responses $\{y_{ij}\}_{j\in\mathcal{V}}$ while satisfying different privacy constraints for different recipients based on the distance on the network..

In order to formalize these statements as in Problems 1 and 2, we need to revisit some concepts and known results. Subsequently, modeling assumptions are presented in Subsection II-A, whereas differential privacy is briefly reviewed in II-B. We present a conventional approach, i.e. a scheme based on access rights in Subsection II-C, and Subsection II-D formally presents the problem of diffusing private data over networks.

[*]Here, accuracy is meant in the expected mean-squared error sense.

A. System Model

Consider a social network represented as a graph G with $|\mathcal{V}| = N$ nodes. For simplicity, we assume that the graph is undirected and unweighted, although this assumption can be removed. Each node $i \in \mathcal{V}$ represents a user and $(i, j) \in \mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents the friendship relation between users i and j. Each user owns a private data $u_i \in \mathcal{U}$. Typical examples include:

1) *Timestamps:* let $u_i \in \mathbb{R}$ be a real-valued representation of a timestamp such as date of birth, e.g. *Unix time* [15] is a popular way of mapping timestamps to integers;
2) *Location:* let $u_i \in \mathbb{R}^2$ be the GPS coordinates of the residence of an individual i;
3) *Binary states:* let $u_i \in \{0, 1\}$ indicate user's i status such as infected or healthy, married or single etc.

Further, we want the severity of the privacy concerns to scale with the distance between two nodes. Typical choices for the distance function $d : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_+$ are as follows:

1) *Shortest Path Distance:* let d_{ij} be the length of the minimum path connecting nodes i and j;
2) *Resistance Distance:* let d_{ij} be the resistance between nodes i and j, where the edges of graph G are associated with unit resistors [16].

A more extended model can incorporate additional information such as family relationships and directed edges (e.g. blocked users) can be also incorporated to better represent social network scenarios, as previously introduced.

B. Differential Privacy

Differential privacy is a formal framework that provides rigorous privacy guarantees. Differentially private algorithms add noise in order to make it hard whether someone's data has been used in the computation. The dependency of this noisy response on the sensitive data is required to be bounded, as formally stated in Definition 1. The strength of this bound is quantified by the non-negative parameter $\epsilon \in [0, \infty)$, called *privacy level*, where smaller values of ϵ imply stronger privacy guarantees. Moreover, an adjacency relation $\mathcal{A}$ is a symmetric binary relation over the set of private data $\mathcal{U}$ which includes the pairs of private data (u, u') that should be rendered *almost* indistinguishable. Further, a mechanism[†] $Q : \mathcal{U} \to \Delta(\mathcal{Y})$ is a randomized map from the space of private data to the space of responses.

Definition 1 (Differential Privacy [1])**.** *Let $\epsilon > 0$, $\mathcal{U}$ be the space of private data, and $\mathcal{A} \subseteq \mathcal{U} \times \mathcal{U}$ be an adjacency*

[†]For a set $\mathcal{T}$ and a rich-enough σ-algebra $\mathcal{T}$ on it, we denote the set of all probability measures on $(T, \mathcal{T})$ with $\Delta(T)$. Specifically, for Euclidean spaces $T = \mathbb{R}^n$, we consider the Borel's σ-algebra.

relation. The mechanism $Q : \mathcal{U} \to \Delta(\mathcal{Y})$ is ϵ-differential privacy if:

$$\mathbb{P}(Qu \in \mathcal{S}) \leq e^{\epsilon}\, \mathbb{P}(Qu' \in \mathcal{S}), \quad \textit{for all } \mathcal{S} \subseteq \mathcal{Y},$$

for all adjacent inputs $(u, u') \in \mathcal{A}$.

In this work, we consider real-valued sensitive data $\mathcal{U} = \mathbb{R}^n$ and the following adjacency relation:

$$(u, u') \in \mathcal{A}_2 \Leftrightarrow \|u - u'\|_2 \leq \alpha, \tag{1}$$

where $\alpha \in \mathbb{R}_+$ is a small constant. Practically, adjacency relation $\mathcal{A}_2$ requires that, given the output of mechanism Q, a curious user should not be able to infer the private input u within a radius of α. One such mechanism is the Laplace mechanism which is near-optimal [17], [10], is used as a building block for many mechanisms, and is described next.

Definition 2 (Laplace Mechanism [1])**.** *Consider the mechanism $Q : \mathbb{R}^T \to \Delta\left(\mathbb{R}^T\right)$ that adds Laplace distributed noise:*

$$Qu = u + V, \text{ where } V \sim Lap\left(\frac{\alpha}{\epsilon}\right),$$

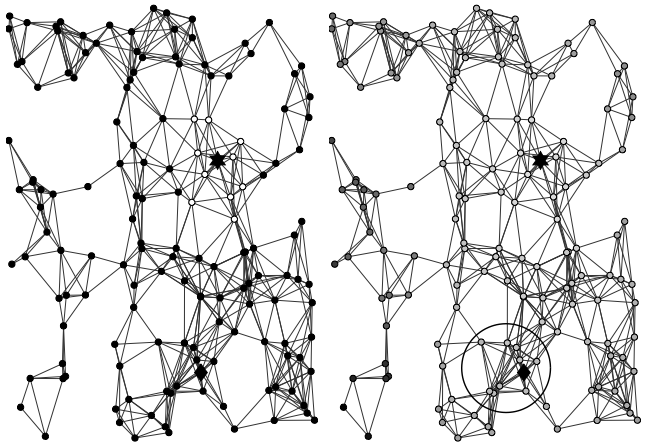
where Lap(b) has density $\mathbb{P}(V = v) = e^{-\frac{\|v\|_2}{b}}$. Then, mechanism Q is ϵ-differential private under adjacency relation $\mathcal{A}_2$.

C. Access Rights Scheme

Now, we describe a typical approach for handling privacy concerns in social network while highlighting its limitations and motivating the need for a more sophisticated privacy-aware approach. Figure 1a shows a synthetic network with 150 nodes, where the starred node wishes to share her sensitive information with the rest of the network. Privacy concerns can be handled by regulating access privileges. For example, friends of a user can access her data, whereas every other user cannot. Such a scheme has limitations. On one hand, users are coarsely partitioned to friends and strangers as depicted in Figure 1a; friends of the star-labeled user are colored white whereas strangers are colored black. On the other hand, the distance between two users can be more finely quantified by a real-valued function, and each user has access only to neighboring information. Although restricting access rights ensures privacy concerns, computing global statistics on the network is impossible, limiting the global utility of the network. Indeed, any estimator of global quantities (mean value, histogram etc.) will to be biased. Therefore, the user may choose to collaborate, merge their local information, and damage any privacy guarantees. Figure 1b overcomes these limitations by defining a distance function $d : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_+$ which quantifies the strength of the privacy concerns. In this case, users share privacy-aware versions of their profile with every member of the network.

D. Diffusing Sensitive Information over a Social Network

Under the modeling introduced in Subsection II-A, we pose the problem of designing a mechanism that diffuses private data over a network that can be formally stated as follows.



(a) An access right scheme. (b) A distance-based scheme.

Fig. 1: A synthetic network with 150 nodes and 1256 edges is shown. Each node represents a user of the network and each edge indicates a friendship. The user indicated with the star wishes to share her sensitive information with the rest of the network. Privacy concerns can be addressed by managing access privileges. Under an access right scheme (Figure 1a), only friends of the starred user (blue nodes) are granted access to the exact information, whereas any other member (red nodes) have no access. Such a scheme partitions users to only two groups; friends and strangers. Moreover, each user has access only to local information and cannot estimate the global state of the network. Therefore, any estimator constructed by the diamond user will be independed of the data of the starred user and, thus, biased. On the other hand, Figure 1b proposes an approach where users' privacy concerns scale with the distance from others. Friends (blue nodes) receive a less noisy versions of the private data, whereas strangers (red nodes) receive only heavily perturbed versions. Despite the increased noise, estimates of aggregate statistics are possible. However, coalitions might be encouraged and initial privacy guarantees can quickly degrade. For example, users within the circle can combine their estimates and infer the private data of the starred user.

Problem 1. *Design a privacy-aware mechanism $Q : \mathcal{U} \to \Delta\left(\mathcal{U}^N\right)$ that privately releases user's i sensitive data $u_i \in \mathcal{U}$ over a social network. Specifically, design mechanism Q that generates N responses $\{y_j\}_{j=1}^N$, where y_{ij} is the securely communicated response to user j. Further, for the adjacency relation* (1) *(where, for simplicity, $\alpha = 1$), the mechanism Q needs to satisfy the following properties:*

- *Variable Privacy: The mechanism must generate the response y_{ij} for private data u_i which $\epsilon(d_{ij})$-differential private.*
- *Optimal Utility: Response y_{ij} must be an accurate approximation of the sensitive data u_i, i.e. for real-valued private data, it should minimize the expected squared-error*

$$\mathbb{E}_Q \|y_{ij} - u_i\|_2^2.$$

Specifically, whenever individual i shares her sensitive information to another individual j, she requires $\epsilon(d_{ij})$-differential privacy, where $\epsilon(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ is a decreasing function that coverts a distance d to a privacy level $\epsilon(d)$. People residing close (w.r.t. a distance) to individual i receive a loose privacy constraint $\epsilon_{ij} \gg 1$, whereas strangers get

noisier versions $\epsilon_{ij} \ll 1$.

Problem 1 admits a straightforward but unsatisfying approach. Let $y_j = u_i + V$, where $V \sim \text{Lap}(d_{ij})$, independently for each user $j \in \mathcal{V}$. Subsequently, a group of users $j \in A \subseteq \mathcal{U}$ have the incentive to collaborate share their estimates $\{y_j\}_{j\in A}$ in order to derive a more accurate estimator y_A of u_i described by

$$y_A = \sum_{j\in A} w_j y_j.$$

Figure 1b depicts a group of users forming such a coalition. The possibly large group A resides far away from the user indicated by the star, $d_{ij} \gg 1, \forall j \in A$. Although each user j in the group A receives a highly noisy estimate of u_i, estimator y_A is more accurate. The composition theorem of differential privacy [1] guarantees only $\left(\sum_{j\in A}\epsilon(d_{ij})\right)$-privacy which can be rather looser than each of the $\epsilon(d_{ij})$-privacy guarantees; larger values of ϵ imply less privacy.

Therefore, Problem 1 is subject to coalition attacks. Thus, we restate Problem 1 by requiring that any group A that exchanges their estimates $\{y_j\}_{j\in A}$ cannot produce a better estimator of u_i than the best estimator among the group y_{j^*}, where $j^* = \arg\min_{j\in A} d_{ij}$ is the user closest to user i. This problem can be stated as follows:

Problem 2. *Design a privacy-aware mechanism $Q : \mathcal{U} \to \Delta\left(\mathcal{U}^N\right)$ that releases a approximation of user's i sensitive data $u_i \in \mathcal{U}$ over a social network. Specifically, mechanism $\mathcal{M}$ generates N responses $\{y_j\}_{j=1}^N$ and securely communicates response y_j to user j. Mechanism Q needs to satisfy:*

- *Privacy: For any group of users $A \subseteq \mathcal{V}$, response $\{y_j\}_{j\in A}$ must be $\max_{j\in A}\epsilon(d_{ij})$-differential private.*
- *Performance: Response y_j must be an accurate approximation of the sensitive data u_i.*

III. Main Results

In this section, we approach the problem of diffusing private data over a network. Subsection III-A derives the needed theoretical results and establishes that the accuracy of each estimate y_{ij} depends *only* on the distance d_{ij}. Moreover, algorithmic implementations of the composite mechanism Q should scale for very large social networks. Subsection III-B provides algorithmic implementations of the mechanism Q with complexity $O\left(\ln\left(\frac{\min_{i,j\in V}\epsilon(d_{ij})}{\max_{i,j\in V}\epsilon(d_{ij})}\right)\right)$.

A. A Private Stochastic Process

For real-valued private data, we derive a composite mechanism in closed form that generates the responses y_{ij} that user j receives as proxies to user's i private data u_i. Additionally, we prove that the accuracy of the response y_{ij} depends solely on the distance d_{ij} between nodes i and j. Specifically, the variance in Equation (2) does not depend on any other parameters of the network (e.g. size) or the rest of the responses $\{y_{ik}\}_{k\in\mathcal{V}\setminus\{j\}}$. Furthermore, assuming the existence of an algorithm for computing the distance d_{ij}, the response y_{ij} can be generated during run-time. This property is crucial, since it circumvents the $O\left(N^2\right)$ memory requirements of a static implementation.

$$\mathbb{E}\left(\|y_{ij} - u_i\|_2\right) = \frac{2}{\epsilon(d_{ij})^2}, \tag{2}$$

where $\epsilon(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ is a decreasing function which converts distance d to a privacy level $\epsilon(d)$. Then, Theorem 3 defines the underlying composite mechanism.

Theorem 3. *Let $d_{ij} \in \mathbb{R}_+$ denote the distance between users i and j, and $u_i \in \mathbb{R}$ be the sensitive data of user i. Then, consider the mechanism Q that generates the responses:*

$$y_{ij} = u_i + V^{(i)}_{\epsilon(d_{ij})},$$

where $\{V^{(i)}_\epsilon\}_{\epsilon>0}$ is a sample of the Markov stochastic process $\{V_\epsilon\}_{\epsilon>0}$ defined as follows:

$$\begin{gathered} V_{\epsilon_3} \perp V_{\epsilon_1} \mid V_{\epsilon_2}, \quad \mathbb{P}(V_\epsilon = v) \propto e^{-\epsilon|v|}, \text{ and} \\ \mathbb{P}(V_{\epsilon_2} = v_2 | V_{\epsilon_1} = v_1) \propto \delta(v_2 - v_1) + e^{-\epsilon_2|v_2 - v_1|}, \end{gathered} \tag{3}$$

where $0 < \epsilon_3 \leq \epsilon_2 \leq \epsilon_1$. Then, mechanism Q provides a solution to Problem 2. In particular, it has the following properties:

- *The variance of response y_{ij} is $2\,\epsilon(d_{ij})^{-2}$ and, thus, depends only on the distance between users i and j.*
- *For any subset of users $A \subseteq \mathcal{V}$, the mechanism that releases the responses $\{y_{ij}\}_{j\in A}$ is $\left(\max_{j\in\mathcal{V}}\epsilon(d_{ij})\right)$-differential private.*

Proof. The main idea is introducing correlation among the responses $\{y_{ij}\}_{j\in V}$ according to Equations (3). First, we prove that the stochastic process $\{V_\epsilon\}_{\epsilon>0}$ is well-defined since Equations (3) are consistent:

$$\begin{aligned} &\mathbb{P}(V_3 = z | V_1 = x) = \\ &\qquad \int_{\mathbb{R}} \mathbb{P}(V_3 = z | V_2 = y)\, \mathbb{P}(V_2 = y | V_1 = x)\, dx. \end{aligned}$$

Next, we prove that process $\{V_\epsilon\}_{\epsilon>0}$ possesses the desired properties:

- The first of Equations (3) implies that the accuracy of the response to user j depends only on the assigned privacy level $\epsilon(d_{ij})$:

$$\mathbb{E}\left[y_{ij} - u_i\right] = \mathbb{E}\left[V_{\epsilon(d_{ij})}\right] = \frac{2}{\epsilon(d_{ij})^2}.$$

- For a subset of users $A = \{a_1, \ldots, a_{|A|}\} \subseteq \mathcal{V}$, we assume, without loss of generality, that user a_1 is residing closest to user i:

$$a_1 \in \arg\max_{j\in A} \epsilon(d_{ij}).$$

Then, the responses $\{y_{ij}\}_{j\in A}$ can be expressed as

follows:

$$\begin{bmatrix} y_{ia_1} \\ y_{ia_2} \\ \vdots \\ y_{i,a_{|A|}} \end{bmatrix} = y_{ia_1} + \begin{bmatrix} 0 \\ y_{ia_2} - y_{ia_1} \\ \vdots \\ y_{ia_{|A|}} - y_{ia_1} \end{bmatrix} \tag{4}$$

$$= u_i + V_{\epsilon(d_{i1})} + \begin{bmatrix} 0 \\ V_{\epsilon(d_{ia_2})} - V_{\epsilon(d_{ia_1})} \\ \vdots \\ V_{\epsilon(d_{ia_{|A|}})} - V_{\epsilon(d_{ia_1})} \end{bmatrix},$$

where the random variables $V_{\epsilon(d_{ia_j})} - V_{\epsilon(d_{ia_1})}$, with $j \in \{2, \dots, |A|\}$ are independent of u_i and $V_{\epsilon(d_{1a_1})}$ and are given by the second of the Equations (3). Therefore, the mechanism that releases the response in (4) can be seen as a post-processing on the Laplace mechanism that releases the response y_{ia_1}. Composition theorem [1] establishes that the mechanism in (4) is $\epsilon(d_{ia_1})$-differential private.

This completes the proof. □

A major consequence of Theorem 3 is that mechanism Q does not incentivize coalitions. Specifically, consider a group of curious users $A \subseteq \mathcal{V}$ who wish to estimate u_i more accurately and, thus, share their knowledge $\{y_{ij}\}_{j \in A}$. In practice, such a group can be fake accounts of a single real but distant (in the sense of d) user. Then, given this shared knowledge, the best estimator is:

$$\hat{u}_i = y_{ij^*}|_{j^* \in \arg\min_{j \in A} d_{ij}}.$$

Therefore, user j^* is not benefited by such a coalition and, thus, she has no incentive to share her information y_{ij^*}.

B. Algorithmic Implementation

Sampling from a continuous-domain stochastic process can often be performed only approximately. For example, consider the Brownian motion $\{B_t, t \in [0,1]\}$ which, for sampling purposes, requires storing an *infimum* of real values. Contrary to Brownian motion, the private process $\{V_\epsilon\}_{\epsilon>0}$ rarely changes value and is, thus, *lazy*. More formally, restricted to a sufficiently small interval, the stochastic process $\{V_\epsilon\}_{\epsilon \in [\epsilon_1, \epsilon_2]}$ is constant with high probability. Proposition 4 characterizes the distribution of the number of jumps n in a bounded interval.

Proposition 4. *The number of jumps that the process $\{V_\epsilon\}_{\epsilon>0}$ performs in the interval $[\epsilon_1, \epsilon_2]$ is Poisson distributed with mean value $2 \ln\left(\frac{\epsilon_2}{\epsilon_1}\right)$.*

$$\mathbb{P}\left(n \textit{ jumps in } [\epsilon_1, \epsilon_2]\right) = \frac{(2x)^n}{n!} e^{-2x},$$

where $x = \ln\left(\frac{\epsilon_2}{\epsilon_1}\right)$.

Proof. Consider the backwards conditional distribution, where $0 < \epsilon_1 < \epsilon_2$:

$$\mathbb{P}(V_{\epsilon_1} = x | V_{\epsilon_2} = y) = \left(\frac{\epsilon_1}{\epsilon_2}\right)^2 \delta(x-y) + \left\{1 - \left(\frac{\epsilon_1}{\epsilon_2}\right)^2\right\} \frac{\epsilon_1}{2} e^{-\epsilon_1 |x-y|}. \tag{5}$$

Let $a_n(x)$ denote the probability that the process performs n jumps in the interval $[\epsilon, e^x \epsilon]$. This probability is invariant of ϵ since, according to distribution (5), the probability of a jump event is governed by the ratio e^x. The probability of landing on exactly value y after performing a jump away of it is zero. Then:

$$a_0(x) = \mathbb{P}(0 \text{ jumps in } [\epsilon, e^x \epsilon]) = e^{-2x}.$$

A limiting argument is used to compute $a_1(x)$. We discretize the interval $[\epsilon_1, \epsilon_2]$ by considering the points $\epsilon^{(k)} = \epsilon_1 + \frac{\epsilon_2 - \epsilon_1}{M} k$, for $k \in \{0, \dots, M\}$. Then:

$$\begin{aligned} a_1(x) = \lim_{M \to \infty} \sum_{k=1}^{M} & \mathbb{P}\left(\text{no jumps in } [\epsilon_1, \epsilon^{(k-1)}]\right) \\ & \mathbb{P}\left(1 \text{ jump in } [\epsilon^{(k-1)}, \epsilon^{(k)}]\right) \mathbb{P}\left(\text{no jumps in } [\epsilon^{(k)}, \epsilon_2]\right) \\ = 2x e^{-2x}. & \end{aligned}$$

A similar argument provides a recurrent equation:

$$\begin{aligned} a_n(x) &= \lim_{M \to \infty} \sum_{i=1}^{M-1} a_{n-1}\left(\frac{ix}{M}\right) a_1\left(\frac{x}{M}\right) \\ &\qquad a_0\left(x\left(1 - \frac{i+1}{M}\right)\right) \\ &= \int_0^1 2x e^{-2x(1-\xi)} a_{n-1}(x\xi) d\xi \\ &= \frac{(2x)^n}{n!} e^{-2x}. \end{aligned} \tag{6}$$

Therefore, for a fixed interval, the number n of jumps is characterized by distribution (6), which is the Poisson distribution with mean value $2 \ln\left(\frac{\epsilon_2}{\epsilon_1}\right)$. □

Corollary 5. *Process $\{V_\epsilon\}_{\epsilon>0}$ performs $\mathbb{E}(n) = 2 \ln 2 \approx 1.39$ jumps (in expectation, with variance Var$(n) = 2 \ln 2$) for every doubling of the privacy level ϵ, i.e. in the interval $[\epsilon, 2\epsilon]$.*

This laziness renders samples from the process highly-compressible. Indeed, given the locations $\{\epsilon^{(i)}\}_{i=1}^N$ of the jumps and the values[‡] $\{V_{\epsilon_-^{(i)}}\}_{i=1}^n$ near those points a sample can be *exactly* reconstructed. The number n of jumps over a bounded interval $[\epsilon_1, \epsilon_2]$ is itself a random variable and captures the memory needs of a system that allows release of sensitive data under multiple privacy levels.

Furthermore, Proposition 4 suggests an efficient algorithm for directly sampling from the process $\{V_\epsilon\}_{\epsilon \in [\epsilon_1, \epsilon_2]}$, which we present in Algorithm 1. Algorithm 1 draws a sample $\{v_\epsilon\}_{\epsilon \in [\epsilon_1, \epsilon_2]}$ from the stochastic process V_ϵ over a

[‡]We use the notation $V_{\epsilon_-} = \lim_{\tau \uparrow \epsilon} V_\tau$ and $V_{\epsilon_+} = \lim_{\tau \downarrow \epsilon} V_\tau$.

bounded interval $\epsilon \in [\epsilon_1, \epsilon_2]$. This sample $\{v_\epsilon\}$ is the main object that performs diffusion of private data; whenever a user j requests user's i private data u_i residing d_{ij} away, the estimator $y_{ij} = u_i + v_{\epsilon(d_{ij})}$.

The algorithm initializes a trace of the process by sampling from the Laplace distribution. Then, the algorithm extends this trace backwards in ϵ by sampling for the location of the next jump. The logarithm of the positions where jumps occur define a Poisson process with rate $\lambda = 2$ and, thus, the length $\delta\epsilon = \ln \epsilon^{(i)} - \ln \epsilon^{(i+1)}$ of the interval until the next jump is exponentially distributed with density $\delta\epsilon \sim \lambda e^{-\lambda\,\delta\epsilon}$. Finally, conditioned on the event of a jump at $\epsilon^{(i)}$, the size $\delta v = V_{\epsilon_-^{(i)}} - V_{\epsilon_+^{(i)}}$ of the jump is Laplace distributed with parameter $\frac{1}{\epsilon^{(i)}}$. The algorithm recycles until the level ϵ_1 is reached. Additionally, responses y_{ij} are generated upon request, and, thus, there is no excessive memory requirement $O(N^2)$ for storing all the responses $\{y_{ij}\}_{i,j\in\mathcal{V}}$. The number of iterations that Algorithm 1 performs is a random variable and is characterized by Proposition 4.

Algorithm 1 Sampling from the stochastic process V_ϵ over a bounded interval $\epsilon \in [\epsilon_1, \epsilon_2]$ can be efficiently and exactly performed.

Require: Privacy levels ϵ_1 and ϵ_2, such that $\epsilon_2 > \epsilon_1 > 0$.
  function SAMPLEPRIVATEPROCESS1D(ϵ_1, ϵ_2)
    $k \leftarrow 1$
    $\epsilon^{(1)} \leftarrow \epsilon_2$
    $v^{(1)} \sim \text{Laplace}\left(\frac{1}{\epsilon_2}\right)$
    while $\epsilon^{(k)} > \epsilon_1$ **do**
      $\delta\epsilon \sim \text{Exponential}(2)$
      $\epsilon^{(k+1)} \leftarrow e^{-\delta\epsilon}\,\epsilon^{(k)}$
      $\delta v \sim \text{Laplace}\left(\frac{1}{\epsilon^{(k+1)}}\right)$
      $v^{(k+1)} \leftarrow v^{(k)} + \delta v$
      $k \leftarrow k + 1$
    end while
    Return $\{(\epsilon^{(i)}, v^{(i)}\}_{i=1}^{k}$
  end function

IV. Example: Facebook Ego-Networks

We present an application that depicts diffusion of private data over a network. This example shows that bits of private information can be spread over the whole network, which allows users to estimate global quantities, such as epidemic spreading, while providing strong privacy guarantees.

In this section, we present an application of diffusing sensitive data on a real network. Specifically, an "ego-network" [18] is a the sub-graph $G = (\mathcal{V} \cup \{\text{Alice}\}, \mathcal{E})$ of Facebook induced by a single user, Alice, and her friends V. Figure 3 plots such an ego-network, where the bottom-left node is the user whose neighborhood is captured. The rest of the nodes represent Alice's friends, edges represent friendships between her friends, whereas, the edges between Alice and her friends are omitted for clarity. We assume that Alice's infection status is captured by a single bit $u \in \{0, 1\}$.

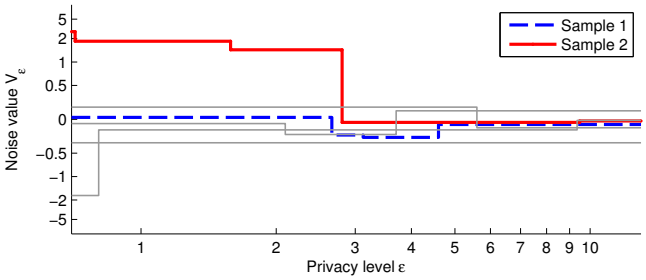


Fig. 2: Two samples of the stochastic process generated by Algorithm 1. The samples are private information; a malicious user i can subtract the noise $w_{\epsilon(d_i)}$ from the received response y_i and exactly infer the private data u.

Then, Alice wishes to share this information with her friends in a privacy-preserving way.

For each friend $i \in \mathcal{V}$, the distance d_i is calculated by a central authority. Values $\{d_i\}_{i\in\mathcal{V}}$ are independent of the private data u, and can be computed without any privacy requirements. In particular, values d_i quantify the strength of the friendship between Alice and friend i and are evaluated according to Equation (7).

$$d_{ij} = \Gamma_{ii} + \Gamma_{jj} - 2\Gamma_{ij}, \tag{7}$$

where $\Gamma \in \mathbb{R}^{n\times n}$ is the pseudo-inverse of the Laplace matrix L of the network. Due to space limitations, we use the fact that our technique allows post-processing of the responses y_{ij} and, thus, is applicable for private bits.

Initially, Alice executes Algorithm 1 in order to generate a single sample $\{w_\epsilon : \epsilon \in [\underline{\epsilon}, \bar{\epsilon}]\}$ of the stochastic process $\{V_\epsilon : \epsilon > 0\}$, where $\underline{\epsilon}$ (resp. $\bar{\epsilon}$) is a lower (resp. upper) bound of the quantity $\min_{i\in V} \epsilon(d_i)$ (resp. $\max_{i\in V} \epsilon(d_i)$). Function $\epsilon(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ is a decreasing function which converts distances d_i to privacy levels $\epsilon_i = \epsilon(d_i)$. In this example, we chose $\epsilon(d) = exp(-3.3d + 4)$ which leads to privacy levels within $[.5, 15]$. Next, individual responses are generated during runtime. Whenever user i requests access to the sensitive data u, the response y_i is securely communicated to user i:

$$y_i = \Pi_{\{0,1\}}(u + w_{\epsilon(d_i)}),$$

where Π_S is the projection operator on the set S.

Figure 2 depicts two executions of Algorithm 1, whereas, Figure 3 plots the ego-network centered around Alice. In particular, Alice is shown on the bottom-left corner and each friend i is plotted at distance d_i from her. The blue and red circles mark the jumps of the stochastic process for the two samples w_ϵ^{blue} and w_ϵ^{red}. Counter-intuitively, friends i lying within two consecutive blue circles receive *exactly* the same response y_i although they are assigned different privacy levels $\epsilon(d_i)$. The paradox is settled by noticing that the boundary circles are random variables themselves. Therefore, users receiving identical responses have different confidence levels.

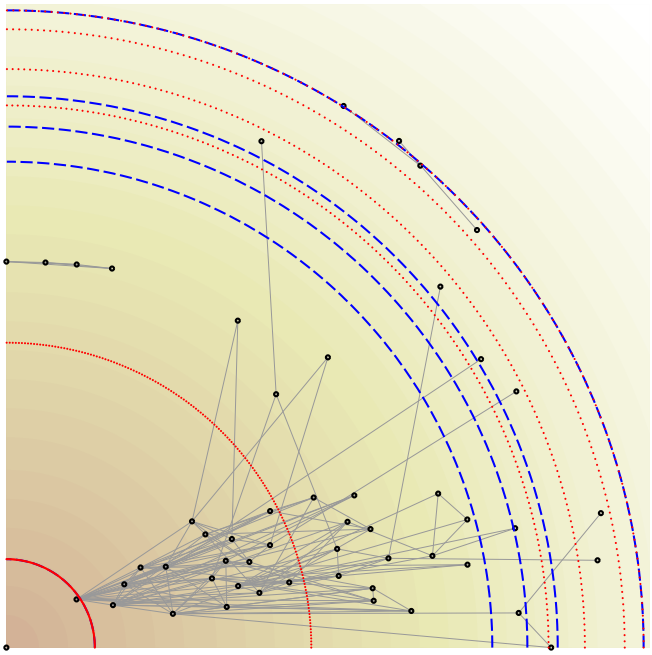


Fig. 3: An ego-network is the part of the Facebook network that is visible from a fixed user A (ego), shown in the bottom-left corner of the plot. Each friend i is plotted at distance d_i. The locations of the jumps of the two samples shown in Figure 2 are depicted by the blue and red circles. Although users residing within consecutive circles receive identical responses y_i, they are assigned different privacy levels $\epsilon(d_i)$ and, thus, have different confidence levels.

V. Conclusions

In this work, we considered the case of a network where each user owns a private data $u \in \mathbb{R}^T$ such as her salary or her infection status and wishes to share approximations of this private data with the rest of the network under differential privacy guarantees. Specifically, we assumed that user i requires $\epsilon(d_{ij})$-differential privacy against against user j, where $\epsilon(\cdot)$ is a decreasing function and d_{ij} is the distance induced by the underlying network between users i and j. In this context, we derived a composite mechanism that generates the response y_{ij} as user's j approximation of user's i private data. The accuracy of the response y_{ij} depends only on the allocated privacy level $\epsilon(d_{ij})$ and not on the size or other parameters of the network. An important property of our proposed mechanism is the resilience to coalitions where we considered a group of users combining their received responses for more accurate approximations. Practically, this means that scenarios where an adversarial user creates multiple fake accounts cannot weaken the privacy guarantees. Algorithms for sampling from this composite mechanism were also provided. In particular, the complexity of these algorithms is independent of the size of the network, which renders them scalable, and is dictated only by the extreme privacy levels $\min_{i\in\mathcal{V}} \epsilon(d_{ij})$ and $\max_{i\in\mathcal{V}} \epsilon(d_{ij})$. Finally, we provided an illustrative examples where a user shares her infection status with her Facebook ego-network. This work focused on the privacy aspect of the problem of diffusing private data over networks. Future work includes the joint problem of accurately estimating formally-defined global quantities while preserving privacy of users' data.

References

[1] C Dwork and A Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 2013.

[2] L Sankar, RS Rajagopalan, S Mohajer, and VH Poor. Smart meter privacy: A theoretical framework. In *IEEE Transactions on Smart Grid*, 2013.

[3] C Dwork. Differential privacy. In *Automata, Languages and Programming*, 2006.

[4] J Hsu, Z Huang, A Roth, and ZS Wu. Jointly private convex programming. *arXiv preprint arXiv:1411.0998*, 2014.

[5] S Han, U Topcu, and GJ Pappas. Differentially private convex optimization with piecewise affine objectives. In *IEEE Conference on Decision and Control*, 2014.

[6] MT Hale and M Egerstedty. Differentially private cloud-based multi-agent optimization with constraints. In *American Control Conference (ACC), 2015*, 2015.

[7] J Le Ny and GJ Pappas. Differentially private filtering. *Automatic Control, IEEE Transactions on*, 2014.

[8] V Katewa, A Chakrabortty, and V Gupta. Protecting privacy of topology in consensus networks. In *American Control Conference (ACC), 2015*, 2015.

[9] G Acs and C Castelluccia. I have a dream!(differentially private smart metering). In *Information Hiding*, 2011.

[10] F Koufogiannis, S Han, and GJ Pappas. Computation of privacy-preserving prices in smart grids. In *IEEE Conference on Decision and Control*, 2014.

[11] J Le Ny, A Touati, and GJ Pappas. Real-time privacy-preserving model-based estimation of traffic flows. In *ICCPS'14: ACM/IEEE 5th International Conference on Cyber-Physical Systems*, 2014.

[12] M Alaggan, S Gambs, and AM Kermarrec. Heterogeneous differential privacy. *arXiv preprint arXiv:1504.06998*, 2015.

[13] H Ebadi, D Sands, and G Schneider. Differential privacy: Now it's getting personal. In *Proceedings of the 42nd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, 2015.

[14] F Koufogiannis, S Han, and GJ Pappas. Gradual release of sensitive data under differential privacy. *arXiv preprint arXiv:1504.00429*, 2015.

[15] Wikipedia. Unix time — wikipedia, the free encyclopedia, 2015. [Online; accessed 13-May-2015].

[16] D Babić, DJ Klein, I Lukovits, S Nikolić, and N Trinajstić. Resistance-distance matrix: A computational algorithm and its application. *International Journal of Quantum Chemistry*, 90(1):166–176, 2002.

[17] Y Wang, Z Huang, S Mitra, and GE Dullerud. Entropy-minimizing mechanism for differential privacy of discrete-time linear feedback systems. In *IEEE Conference on Decision and Control*, 2014.

[18] J Leskovec and JJ Mcauley. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems*, 2012.