Local Differential Privacy for Physical Sensor Data and Sparse Recovery

Anna C. Gilbert
Department of Mathematics
University of Michigan
Ann Arbor, MI USA
annacg@umich.edu

Audra McMillan
Department of Mathematics
University of Michigan
Ann Arbor, MI USA
amcm@umich.edu

Abstract—In this work, we exploit the ill-posedness of linear inverse problems to design algorithms to release differentially private data or measurements of the physical system. We discuss the spectral requirements on a matrix such that only a small amount of noise is needed to achieve privacy and contrast this with the ill-conditionedness. We then instantiate our framework with several diffusion operators and explore recovery via \( \ell_1 \) constrained minimisation. Our work indicates that it is possible to produce locally private sensor measurements that both keep the exact locations of the heat sources private and permit recovery of the “general geographic vicinity” of the sources.

Index Terms—sparse signal recovery, graph diffusion, differential privacy, graph tomography

I. INTRODUCTION

Imagine dripping a few drops of ink into a glass of water. The ink drops spread out, forming complicated tendrils that coil back on each other, expanding quickly, until all of the ink has diffused and the liquid is a slightly darker shade than its original colour. There is no physical process by which you can make the diffusing ink coalesce back into its original droplets. This intuition is at the heart of what we call computational cloaking. Because it is physically impossible to reconstruct the ink droplet exactly, we should be able to hide or keep private in a precise sense its original location. When mathematicians and physicists refer to cloaking, they mean transformation optics [16], the design of optical devices with special customised effects on wave propagation. In this work, we exploit the ill-posedness of inverse problems to design algorithms to release differentially private data or measurements of the physical system.

We are motivated by the explosion in the power and ubiquity of lightweight (thermal, light, motion, etc.) sensors. These data offer important benefits to society. For example, thermal sensor data now plays an important role in controlling HVAC systems and minimising energy consumption in smart buildings [3], [23]. Simultaneously, we have begun to understand the extent to which our privacy is compromised by allowing this increased level of data collection. The field of privacy-preserving data analytics has developed to help alleviate these privacy concerns [10]. A particular notion of privacy called differential privacy has emerged as a gold standard for privacy.

To continue with the example of thermal sensor data, one might consider sources of heat to be people, whose locations we aim to keep private. Our work indicates that it is possible to produce locally private sensor measurements that both keep the exact locations of the heat sources private and permit recovery of the “general vicinity” of the sources. That is, the locally private data can be used to recover an estimate, \( \hat{f} \), that is close to the true source locations, \( f_0 \), in the Earth Mover Distance (EMD). This is the second aspect to our work: algorithms that reconstruct sparse signals with error guarantees with respect to Earth Mover Distance (rather than the more traditional \( \ell_1 \) or \( \ell_2 \) error in which accurate recovery is insurmountable).

II. PRELIMINARIES

A. Source localization

Suppose that we have a vector \( f_0 \) of length \( n \) that represents the strengths and positions of our “sources.” The \( i \)th entry represents the strength of the source at position \( i \). Further, suppose that we take \( m \) linear measurements of our source vector; we observe

\[
y = Mf_0
\]

where \( M \) represents some generic linear physical transformation of our original data. Let us also assume that the source vector \( f_0 \) consists of at most \( k \) sources (or \( k \) non-zero entries). The straightforward linear inverse problem is to determine \( f_0 \), given \( M \) and a noisy version of \( y \). More precisely, given noisy measurements \( \tilde{y} = Mf_0 + N(0, \sigma^2 I_m) \), can we produce an estimate \( \hat{f} \) that is still “useful”?

For physical processes such as diffusion, intuitively, we can recover the approximate “geographic vicinity” of the source. This is exactly the concept of closeness captured by the Earth Mover Distance (EMD). Thus, in this paper, we aim to recover \( \hat{f} \) that is close to \( f_0 \) in the EMD. The EMD can be defined between any two probability distributions on a finite discrete metric space \( (\Omega, d(\cdot, \cdot)) \). It computes the amount of work required to transform one distribution into the other.

Definition 1: [27] Let \( P = \{(x_1, p_1), \cdots, (x_n, p_n)\} \) and \( Q = \{(x_1, q_1), \cdots, (x_n, q_n)\} \) be two probability distributions on the discrete space \( \{x_1, \cdots, x_n\} \). Now, let

\[
f^* = \arg \min_{f \in [0,1]^{n \times n}} \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij} d(x_i, x_j)
\]
s.t. $f_{ij} \geq 0 \forall i, j \in [m]$, $\sum_{j=1}^{n} f_{ij} \leq p_i \forall i \in [m]$,

$$\sum_{i=1}^{n} f_{ij} \leq q_i \forall i \in [n], \text{ and } \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij} = 1.$$ then $\text{EMD}(P, Q) = \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij}^* d(x_i, x_j)$.

**B. Differential Privacy**

To understand our definition of cloaking, we give a very brief introduction to differential privacy in this section. A more in-depth introduction can be found in [10]. Differential privacy has emerged over the past decade as the leading definition of privacy for privacy-preserving data analysis. A database is a vector $D$ in $D^n$ for some data universe $D$. We call two databases $D, D'$ adjacent or “neighbouring” if $||D - D'||_0 = 1$.

**Definition 2 ((e, δ)-Differential Privacy):** [8] A randomised algorithm $A$ is $(e, \delta)$-differentially private if for all adjacent databases $D, D'$ and events $E$,

$$\mathbb{P}(A(D) \in E) \leq e^e \mathbb{P}(A(D') \in E) + \delta.$$ To understand this definition suppose the database $D$ contains some sensitive information about Charlie and the data analyst, Lucy, produces some statistic $A(D)$ about the database $D$ via a differentially private algorithm. Then Lucy can give Charlie the following guarantee: an adversary given access to the output $A(D)$ can not determine whether the database was $D$ or $D'$, where $D$ has Charlie’s true data and $D'$ has Charlie’s data replaced with an arbitrary element of $D$.

**C. Computational cloaking precisely**

First, we clarify exactly what “data” we would like to keep private. We consider the coordinates of $f_0$ to be our data, that is the locations of the sources are what we would like to keep private. We assume that there exists a metric $d(\cdot, \cdot)$ on the set of possible source locations, which induces the EMD on the set of source vectors. For the remainder of this paper, we will assume that metric $d$ is such that every pair of source locations is connected by a path that travels via neighbours.

When the matrix $M$ represents a physical process, we usually cannot hope to keep the existence of a source private and also recover an estimation to $f_0$ that is close in the EMD. However, it may be possible to keep the exact location private while allowing recovery of the “general vicinity” of the source. In fact, we will show in Section IV-A that this is possible for diffusion on the discrete 1-dimensional line and in Section IV-B that we can generalise these results to diffusion on a general graph $G$. We are going narrow our definition of “neighbouring” databases to capture this idea.

**Definition 3:** For $\alpha > 0$, two source vectors $f_0$ and $f'_0$ are $\alpha$-neighbours if $\text{EMD}(f_0, f'_0) \leq \alpha$.

The larger $\alpha$ is, the less stringent the neighbouring condition is, so the more privacy we are providing. This definition has two important instances. We can move a source of weight 1 by $\alpha$ units, hiding the location of a large heat source (like a fire) within a small area. Also, we can move a source with weight $\alpha$ by 1 unit, hiding the location that small heat source (like a person) over a much larger area. We will usually drop the $\alpha$ when referring to neighbouring vectors.

A locally differentially private algorithm is a private algorithm in which the individual data points are made private before they are collated by the data analyst. In many of our motivating examples the measurements $y_i$ are at distinct locations prior to being transmitted by a data analyst (for example, at the sensors). Thus, the “local” part of the title refers to the fact that we consider algorithms were each measurement, $y_i$, is made private individually. This is desirable since the data analyst (e.g. landlord, government) is often the entity the consumer would like to be protected against. Also, it is often the case that the data must be communicated via some untrusted channel [13], [30]. Usually this step would involve encrypting the data, incurring significant computational and communication overhead. However, if the data is made private prior to being sent, then there is less need for encryption. We then wish to use this locally differentially private data to recover an estimate to the source vector that is close in the EMD. The structure of the problem is outlined in the following diagram:

\[ f_0 \xrightarrow{M} \{ y_1 \xrightarrow{A} \tilde{y}_1, \ldots, y_m \xrightarrow{A} \tilde{y}_m \} \xrightarrow{R} \hat{f} \]

Design algorithms $A$ and $R$ such that:

1) (Privacy) For all neighbouring source vectors $f_0$ and $f'_0$, indices $i$, and Borel measurable sets $E$ we have

$$\mathbb{P}(A((M f_0)_i) \in E) \leq e^e \mathbb{P}(A((M f'_0)_i) \in E) + \delta.$$ 2) (Utility) $\text{EMD}(f_0, \hat{f})$ is small.

**D. Related work**

An in-depth survey on differential privacy and its links to machine learning and signal processing can be found in [28]. The body of literature on general and local differential privacy is vast and so we restrict our discussion to work that is directly related. There is growing body of literature of differentially private sensor data, for example [11], [18], [20], [24], [31]. Much of this work is concerned with differentially private release of aggregate statistics derived from sensor data and the difficulty in maintaining privacy over a period of time (called the continual monitoring problem).

Connections between privacy and signal recovery have been explored previously in the literature. In [9], Dwork, et al. considered the recovery problem with noisy measurements where the matrix $M$ has i.i.d. standard Gaussian entries. Newer results of Bun et al. [4] can be interpreted in a similar light where $M$ is a binary matrix. Compressed sensing has also been used in the privacy literature as a way to reduce the amount of noise needed to maintain privacy [21], [26]. There are also several connections between sparse signal recovery and inverse problems [5], [12], [17], [19]. The heat...
source identification problem is severely ill-conditioned and, hence, it is known that noisy recovery is impossible in the common norms like ℓ₁ and ℓ₂. This has resulted in a lack of interest in developing theoretical bounds [22], thus the mathematical analysis and numerical algorithms for inverse heat source problems are still very limited.

To the best of the authors knowledge, the two papers that are most closely related to this paper are Li, et al. [22] and Beddiaf, et al. [1]. Both of these papers attempt to circumvent the condition number lower bounds by changing the error metric to capture “the recovered solution is geographically close to the true solution”, as in this paper. Our algorithm is the same as Li, et al., who also consider the Earth Mover Distance (EMD). Our upper bound is a generalisation of theirs to source vectors with more than one source. Beddiaf, et al. follow a line of work that attempts to find the sources using ℓ₂-minimisation and regularisation.

III. PRIVACY OF MEASUREMENTS

Because we assume that our sensors are lightweight computationally, the algorithm A is simply for each sensor to add Gaussian noise locally to its measurements before sending to the central node¹. The question then is; how much noise should we add to maintain privacy? The following lemma says, essentially, that the standard deviation of the noise added to a statistic should be proportional to how much the statistic can vary between neighbouring data sets. Let \( g : D^n \to \mathbb{R}^n \) be a function and let \( \Delta_2 g = \max_{D,D'} \| g(D) - g(D') \|_2 \) (called the ℓ₂ sensitivity of \( g \)).

**Lemma 1 (The Gaussian Mechanism):** [10] Let \( \epsilon > 0 \), \( \delta > 0 \) and \( \sigma = \frac{2\ln(1.25/\delta)}{\epsilon} \Delta_2 g \) then

\[
A(D) \sim g(D) + N(0,\sigma^2 I_n)
\]

is an \((\epsilon,\delta)\)-differentially private algorithm.

Let us apply the Gaussian mechanism to the general linear inverse problem. As we discussed previously, ill-conditioned source localization problems behave poorly under addition of noise. Intuitively, this should mean we need only add a small amount of noise to mask the original data. We show that this statement is partially true. There is, however, a fundamental difference between the notion of a problem being ill-conditioned (as defined by the condition number) and being easily kept private. Let \( M_i \) be the \( i \)th column of \( M \).

**Proposition 1:** With \( \alpha > 0 \) and the definition of \( \alpha \)-neighbours presented in Definition 3, we have

\[
A(M f_0) \sim M f_0 + \frac{2\ln(1.25/\delta)}{\epsilon} \Delta_2(M) N(0,I_m)
\]

is a \((\epsilon,\delta)\)-differentially private algorithm where

\[
\Delta_2(M) = \max_{e_i,e_j \text{ neighbours}} \| M_i - M_j \|_2
\]

**Proof:** Suppose \( f_0 \) and \( f'_0 \) are \( \alpha \)-neighbours and let \( f_{kl} \) be the optimal flow from \( f_0 \) to \( f'_0 \) (as defined in Definition 1) so \( f_0 = \sum_{k,l} f_{kl} e_k \) and \( f'_0 = \sum_{k,l} f_{kl} e_l \), where \( e_k \) are the standard basis vectors. Then

\[
\| M f_0 - M f'_0 \|_2 \leq \sum_{k,l} f_{kl} \| M e_k - M e_l \|_2
\]

To the best of the authors knowledge, the two papers that are most closely related to this paper are Li, et al. [22] and Beddiaf, et al. [1]. Both of these papers attempt to circumvent the condition number lower bounds by changing the error metric to capture “the recovered solution is geographically close to the true solution”, as in this paper. Our algorithm is the same as Li, et al., who also consider the Earth Mover Distance (EMD). Our upper bound is a generalisation of theirs to source vectors with more than one source. Beddiaf, et al. follow a line of work that attempts to find the sources using ℓ₂-minimisation and regularisation.

The following lemma gives a characterization of \( \Delta_2(M) \) in terms of the spectrum of \( M \). It verifies that the matrix \( M \)

¹ Gaussian noise is not the only option to achieve privacy. There has been some work on the optimal type of noise to add to achieve privacy [14].
must be almost rank 1, in the sense that the spectrum should be dominated by the largest singular value.

**Lemma 3:** If $\Delta_2(M) \leq \nu$, then $\|M_i\|_2 - \|M_j\|_2 \leq \frac{\nu}{\alpha}$ for any pair of neighbouring locations $e_i$ and $e_j$, and $|\sum_{i\notin \min\{m,n\}} s_i| \leq \frac{(n+1)^{3/2}\nu}{\alpha}$, where $\rho$ is the diameter of the space of source locations.

Conversely, if $|\sum_{i\notin \min\{m,n\}} s_i| \leq \frac{\nu}{\alpha}$ and $\|M_i\|_2 - \|M_j\|_2 \leq \frac{\nu}{\alpha}$ then $\Delta_2(M) \leq 4\nu$.

**Proof:** Let $e_i$ and $e_j$ be neighbouring sources. Now, assume $\Delta_2(M) \leq \nu$ then $\|M_i\|_2 - \|M_j\|_2 \leq \|M_i - M_j\|_2 \leq \frac{\nu}{\alpha}$. Suppose wlog that $\max_m \|M_i\|_2 = \|M_1\|_2$ and let $M' = [M_1 \cdots M_k]$ be the matrix whose columns are all duplicates of the first column of $M$. Recall that the trace norm of a matrix is the sum of its singular values and for any matrix, $\|M\|_\text{tr} \leq \sqrt{\min\{m,n\}}\|M\|_F$ and $\|M\|_2 \leq \|M\|_F$. Since $M'$ is rank 1, $\|M'\|_\text{tr} = \|M'\|_2 = s_{\min\{m,n\}}$, thus,

$$
\min\{m,n\} - 1 \leq \sum_{i=1}^{\min\{m,n\} - 1} s_i \leq \|M'\|_\text{tr} - \|M'\|_\text{tr} + \|M'\|_\text{tr} - s_{\min\{m,n\}} \\
\leq \|M' - M\|_\text{tr} + \|M'\|_2 - \|M\|_2 \\
\leq (\sqrt{\min\{m,n\}} + 1)\|M' - M\|_F \\
\leq (\sqrt{\min\{m,n\}} + 1)\rho(n-1)\frac{\nu}{\alpha}
$$

Conversely, suppose $|\sum_{i\notin \min\{m,n\}} s_i| \leq \frac{\nu}{\alpha}$ and $\|M_i\|_2 - \|M_j\|_2 \leq \frac{\nu}{\alpha}$. Using the SVD we know, $M = \sum s_k U_k \otimes V_k$, where $U_k$ and $V_k$ are the left and right singular values, respectively.

$$
\|M_i - M_j\|_2 = \|\sum_{k} s_k(U_k)_i V_k - \sum_{k} s_k(U_k)_j V_k\|_2 \\
\leq \|\sum_{k} s_{\min\{m,n\}} [(U_1)_i - (U_1)_j]\|_\infty + \nu
$$

Also, $\|M_i\|_2 - \|M_j\|_2 \geq \|s_{\min\{m,n\}} [(U_1)_i - (U_1)_j]\|_\infty - \frac{\nu}{\alpha}$, so $|\sum_{i\notin \min\{m,n\}} s_i| \leq 3\frac{\nu}{\alpha\min\{m,n\}}$.

**IV. Recovery Algorithm**

We claimed that the private data is both useful and privacy-preserving. As has been studied extensively in the compressed sensing literature, one can recover a sparse vector by solving a convex optimization problem (typically, $\ell_1$ minimization) when the measurement matrix $M$ satisfies the restricted isometry property (RIP) [6]. Even if the matrix $M$ does not satisfy the RIP, using $\ell_1$ minimization promotes sparsity and can be executed efficiently. This use of constrained $\ell_1$-minimization to recover source vectors with the heat kernel was introduced in Li, et al. [22], who studied the case of a 1-sparse source vector.

**Input:** $M$, $\sigma > 0$, $\hat{y}$

**Output:** $\hat{f} \in [0,1]^n$

1. $\hat{f} = \text{arg min}_{f \in [0,1]^n} \|f\|_1 \text{ s.t. } \|Mf - \hat{y}\|_2 \leq \sigma \sqrt{n}$

**Algorithm 1:** $R$: Constrained $\ell_1$ minimization recovery algorithm

**Theorem 1:** Suppose that $f_0$ is a source vector, $\hat{y} = R(\hat{y})$ and assume the following:

<table>
<thead>
<tr>
<th>Variable</th>
<th>EMD$(\frac{f_0}{|f_0|_1}, \frac{f}{|f|_1})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>$O(1 + \frac{\nu}{\sqrt{n}})$</td>
</tr>
<tr>
<td>$m$</td>
<td>$O(1)$</td>
</tr>
<tr>
<td>$T$</td>
<td>$\min {1, O(1 + \frac{\nu}{\sqrt{\nu}} + T^{2.5}e^{-\sigma^2/4\sigma^2})}$</td>
</tr>
</tbody>
</table>

The bound $\sigma \sqrt{m}$ in Algorithm 1 is chosen to ensure $f_0$ is a feasible point with high probability.

**A. Diffusion on the unit interval**

Let us define the linear physical transformation explicitly for heat source localization. To distinguish this special case from the general, let us denote the measurement matrix by $A$ (instead of $M$). For heat diffusion, we have a diffusion constant $\mu$ and a time $t$ at which we take our measurements.

Let $T = \mu t$ in what follows. Let $g(x,t) = \frac{1}{\sqrt{4\pi t}} e^{-x^2/4t}$. Let $n > 0$ and suppose the support of $f$ is contained in the discrete set $\{\frac{1}{n}, \cdots, 1\}$. Let $m > 0$ and suppose we take $m$ measurements at locations $\frac{1}{m}, \cdots, 1$ so $y$ is the measurement of the sensor at location $\frac{1}{m}$ at time $t$ and we have

$$y = Af_0 \text{ where } A_{ij} = g\left(\frac{i}{n} - \frac{j}{m}, t\right).$$

The heat kernel, $A$, is severely ill-posed due to the fact that as heat dissipates, the measurement vectors for different source vectors become increasingly close [32]. Figure 1 shows the typical behaviour of Algorithm 1 with the matrix $A$. As can be seen in the figure, this algorithm returns an estimate $\hat{f}$ that is indeed close to $f_0$ in the EMD but not close in more traditional norms like the $\ell_1$ and $\ell_2$ norms. This phenomenon was noticed by Li, et al., who proved that if $f_0$ consists of a single source then $\text{EMD}(f_0, \hat{f})$ is small where $\hat{f} = R(\hat{y})$ [22].

**Proposition 2:** With the definition of neighbours presented in Definition 3 and restricting to $f_0 \in [0,1]^n$ we have

$$\Delta_2(A) = O\left(\frac{\alpha \sqrt{n}}{T^{1.5}}\right)$$

**Theorem 1:** Suppose that $f_0$ is a source vector, $\hat{y} = R(\hat{y})$ and assume the following:
We first upper bound the KL-divergence between the distribution of measurements generated by two source vectors. For every pair of source vectors in $f, f'$, we have

$$\inf_{\hat{f}} \mathbb{E}[\text{KL}(f, \hat{f})] \leq \Omega \left( \frac{1}{2} \sqrt{\frac{1}{m} \text{KL}(f, f')^2} \right).$$

where $\text{KL}(f, f')$ is the KL-divergence between the distributions generated by $f$ and $f'$, and $\hat{f}$ is the estimator we use. The details can be found in the extended version of this paper [15].

In order to obtain a recovery bound for the private data, we set $\sigma = \frac{\sqrt{m} \text{KL}(f, f')}{\sqrt{T} + 1}$. The asymptotics of this bound are contained in Table I. It is interesting to note that, unlike in the constant $\sigma$ case, the error increases as $T \to 0$ (as well as when $T \to \infty$). This is because as $T \to 0$ the inverse problem becomes less ill-conditioned so we need to add more noise.

The following theorem gives a lower bound on the estimation error of the noisy recovery problem.

**Theorem 2:** We have

$$\inf_{\hat{f}} \mathbb{E}[\text{KL}(f, \hat{f})] \leq \Omega \left( \frac{1}{2} \sqrt{\frac{1}{m} \text{KL}(f, f')^2} \right).$$

where $\inf_{\hat{f}}$ is the infimum over all estimators $\hat{f} : \mathbb{R}^m \to [0, 1]^n$, $\sup_{f_0}$ is the supremum over all source vectors in $[0, 1]^n$ and $\hat{y}$ is sampled from $y + N(0, \sigma^2 I_m)$.

We first upper bound the KL-divergence between the distributions on measurements generated by two source vectors $f_0$ and $f'_0$ in terms of $\text{KL}(f_0, f'_0)$. We then find a small class of source vectors such that for every pair of source vectors in the class, we can upper bound the KL-divergence and lower bound the EMD. The details can be found in the extended version.

Note that this lower bound matches our upper bound asymptotically in $\sigma$ and is slightly loose in $T$. It varies by a factor of $\sqrt{m}$ from our theoretical upper bound. Experimental results (contained in the extended version) suggest that the error decays like $O(1 + \frac{1}{\sqrt{m}})$.

### TABLE II: Examples of $\Delta_2(A_G)$ for some standard graphs.

<table>
<thead>
<tr>
<th>Graph Type</th>
<th>$\Delta_2(A_G)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete Graph</td>
<td>$e^{-2rn} + \left(\frac{e^{-r}-e^{-r}}{n-1}\right)^2 + \left(\frac{e^{-r}-e^{-r}}{n-1}\right)^2$</td>
</tr>
</tbody>
</table>

### B. Diffusion on graphs

In this section, we generalise to diffusion on an arbitrary graph. As usual, our aim is to protect the exact location of a source, while allowing the neighbourhood to be revealed. Diffusion on graphs models not only heat spread in a graph, but also the path of a random walker in a graph and the spread of rumours, viruses, or information in a social network. If a graph displays a community structure, then we would like to determine which community the source is in, without being able to isolate an individual person within that community. A motivating example, is that we would like people to be able to spread information without fear of retribution, but also be able to approximately locate the source of misinformation.

Let $G$ be a connected, undirected graph with $n$ nodes. The $n \times n$ matrix $W$ contains the edge weights so $W_{ij}$ is the weight of the edge between node $i$ and node $j$ and the diagonal matrix $D$ has $D_{ii}$ equal to the sum of the $i$-th row of $W$. The graph Laplacian is $L = D - W$. As above, we also have a parameter controlling the rate of diffusion $\tau$. Then if the initial distribution is given by $f_0$ then the distribution after diffusion is given by the linear equation $y = e^{-\tau L} f_0$ [29]. We will use $A_G$ to denote the matrix $e^{-\tau L}$. Note that, unlike in the previous section, we have no heat leaving the domain (i.e., the boundary conditions are different).

The graph $G$ has a metric on the nodes given by the shortest path between any two nodes. Recall that in Lemma 3 we can express the amount of noise needed for privacy, $\Delta_2(A_G)$, in terms of the spectrum of $A_G$. Let $s_1 \leq s_2 \leq \cdots \leq s_{\min(n,m)}$ be the eigenvalues of $L$ then $e^{-s_1 \tau} \leq e^{-s_2 \tau} \leq \cdots \leq e^{-s_{\min(n,m)} \tau}$ are the eigenvalues of $A_G$. For any connected graph $G$, the Laplacian $L$ is positive semidefinite and 0 is an eigenvalue with multiplicity 1 and eigenvector the all-ones vector.

**Lemma 4:** For any graph $G$, $\Delta_2(A_G) \leq \sum_{k=1}^{\min(n,m)} e^{-s_k \tau} |(u_k)_i - (u_j)_k|$, where $u_i$ is the $i$-th row of the matrix whose columns are the left singular vectors of $L$.

**Proof:** With set-up as in Lemma 3 we have

$$\|(A_G)_{ij} - (A_G)_{ij}\|_2 = \|\sum_k e^{-s_k \tau} ((U_k)_i - (U_k)_j) V_k\|_2 \leq \sum_k e^{-s_k \tau} |(U_k)_i - (U_k)_j|.$$
The eigenvalue of $L_s$, $s_2$, (called the algebraic connectivity) is related to the connectivity of the graph, in particular the graphs expanding properties, maximum cut, diameter and mean distance [25]. As the graph becomes more connected, the rate of diffusion increases so the amount of noise needed for privacy decreases. The dependence on the rows of the matrix is intriguing as these rows arise in several other areas of numerical analysis. Their $\ell_2$ norms are called leverage scores [7]. This is a connection we will explore in future work.

Figure 2 shows the average behaviour of Algorithm 1 on a graph with community structure. Preliminary experiments suggest that provided $\tau$ is not too large or too small and $\epsilon$ is not too small, the correct community is recovered.

An extended version of this paper can be found at arXiv:1706.05916v3.

REFERENCES


