Computing epidemic metrics with edge differential privacy

George Z. Li
University of Maryland

Dung Nguyen
University of Virginia

Anil Vullikanti
University of Virginia

Abstract

Metrics such as the outbreak size in an epidemic process on a network are fundamental quantities used in public health analyses. The datasets used in such models used in practice, e.g., the contact network and disease states, are sensitive in many settings. We study the complexity of computing epidemic outbreak size within a given time horizon, under edge differential privacy. These quantities have high sensitivity, and we show that giving algorithms with good utility guarantees is impossible for general graphs. To address these hardness results, we consider a smaller class of graphs with similar properties as social networks (called expander graphs) and give a polynomial-time algorithm with strong utility guarantees. Our results are the first to give any non-trivial guarantees for differentially private infection size estimation.

1 Introduction

Epidemic models, especially network based, are commonly used in public health analyses, e.g., (Marathe and Vullikanti, 2013; Adiga et al., 2020). Such a model is defined on a contact network $G = (V, E)$, where V denotes the population, and E denotes the set of edges, on which infections could spread. In the Susceptible-Infected-Recovered (SIR) model of disease spread, an infected node $v \in V$ spreads the infection to each susceptible neighbor, independently, with some probability. The simplest metrics of interest in epidemic analyses are expected number of infections, peak size (which are at a population level), and individual risk of infection (which is at an individual level). There has been a lot of previous work on forecasting such metrics, e.g., (Mamidi et al., 2021; Adiga et al., 2022). One limitation of previous approaches is the lack of data privacy guarantees. These methods use diverse kinds of datasets as inputs, including mobility traces, contacts, and infection status. Some individual risk prediction models were based on electronic medical record data, e.g., (Mamidi et al., 2021), and contact tracing apps relied on contacts inferred through phones, e.g., (Akinbi et al., 2021). Many of these are very sensitive datasets. For instance, many individuals might prefer to keep their contacts and disease states private; indeed, privacy is considered to be one of the reasons the deployment of contact tracing apps were not adopted by a large fraction of the population (Akinbi et al., 2021). Our work attempts to address this limitation via the framework of differential privacy (Dwork et al., 2006).

1.1 Our Contributions

We initiate the study of differentially private estimation of the expected number of infections in a graph G resulting from s source infections, denoted by $f(G)$. If the initial infections set is given, we show that any differentially private algorithm must incur an additive $\Theta(n)$ factor in general (see Section 6). Therefore, we focus on the setting where the s source infections are selected randomly. For this setting, we show that we can beat the $\Omega(n)$ lower bound, in contrast to hardness results in the fixed-source setting. We show that the global sensitivity of $f(G)$ in the random initial infection setting is $\Theta(n/s)$ (see Section 3). Combined with the Laplace mechanism (Dwork et al., 2014), this implies an differentially private algorithm with $\Theta(n/s)$ expected additive error. While this is significantly better than a naive bound of $\Omega(n)$, it can lead to poor utility in general.

This motivates alternative approaches to global sensitivity such as various local sensitivity based methods. We first show that the smooth sensitivity technique (Nissim et al., 2007) can be applied to our problem, and can greatly improve utility over global sensitivity-based methods. Since the smooth sensitivity is difficult to compute, we show computing a multiplicative approximation for smooth sensitivity suffices

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

for privacy. We then give a quasi-polynomial time algorithm for computing approximate smooth sensitivity for our problem.

To address the slow running time, we give a generalization of the classical propose-test-release approach, and use it to give a polynomial-time algorithm with strong utility guarantees for a class of sparse but well-connected graphs called expanders. Since many social networks are believed to satisfy approximate expansion properties, this gives the first positive result for estimating infection size with differential privacy for social network graphs. Additionally, we believe that our generalized propose-test-release framework will be of independent interest, since it was the only approach which improved over the quasi-polynomial runtime.

1.2 Related Work

There is no prior work on the problem we study here, and there has been little work on private computation of epidemic metrics. The most closely related work is the work on the Individual Risk Prediction problem (Harrison et al., 2023), which involves privately predicting the probability of infection for a node in the next Δ time steps. Recent work by Liu and Smith (2023) develops a federated learning method for this problem, while guaranteeing node differential privacy. While this method could be used to determine the expected outbreak size by summing the probabilities, the accuracy would be much worse than directly estimating the size. Specifically, the performance of their algorithm degrades very rapidly for large Δ, which would be needed for estimating the full outbreak size. This work is orthogonal to ours.

More generally, there has been a lot of work on private algorithms for computing a variety of graph statistics (e.g., degree distribution and counts of subgraphs) in node, edge and attribute privacy models (Kasiviswanathan et al., 2013; Mülle et al., 2015; Imola et al., 2021; Blocki et al., 2013; Ji et al., 2019; Hay et al., 2009; Zhang et al., 2015; Dhulipala et al., 2022, 2023). Since graph statistics generally have high global sensitivity (e.g., the number of triangles in a graph has a sensitivity of $n-2$ under edge privacy), using the Laplace mechanism can be very inaccurate. Consequently, more advanced techniques based such as smooth sensitivity (Nissim et al., 2007; Karwa et al., 2014), ladder functions (Zhang et al., 2015), propose-test-release (Dwork et al., 2014), and inverse sensitivity (Asi and Duchi, 2020) have been developed to provide much more accurate counts for some problems, but are computationally expensive (see (Li et al., 2023) for a recent survey on private graph algorithms). Our work contributes to and generalizes some approaches in this line of work.

2 Preliminaries

2.1 Independent Cascades Model

We consider the *Independent Cascades* (IC) model, which is the simplest instance of the more general Susceptible-Infected-Recovered (SIR), on a contact graph $G = (V, E)$ (Marathe and Vullikanti, 2013). In this model, each node v is in one of Susceptible (S), Infectious (I) or Recovered (R) state. In the beginning ($t = 0$), we have a subset $I_0 \subseteq V$ of *source* nodes in the infected state, with $s := |I_0|$, and all remaining nodes are in the Susceptible state. Let I_t denote the set of nodes which are infected at time t. At each time-step t, an infected node u can infect each susceptible neighbor v with probability p, independent of other neighbors of v. An infected node v is assumed to recover after one time step, so all nodes in I_t are Recovered in the next timestep. The expected total number of infections $f(G) = \mathbb{E}[|\bigcup_t I_t|]$ is one of the most basic metrics in epidemic analyses, which we will try to estimate given an input graph and transmission probability p.

2.2 Differential Privacy Model

We use the notion of differential privacy (Dwork et al., 2014), which is one of the most widely used standards of privacy, and has been extended to graph data (Blocki et al., 2013; Kasiviswanathan et al., 2013). We focus on the *edge privacy* model (Blocki et al., 2013), which guarantees differential privacy for the contact data between two individuals. Formally, edge-privacy is defined as follows.

Definition 2.1. *We say that two graphs G, G' are edge-neighbors, i.e., $G \sim G'$, if they differ in exactly one edge, i.e., $|E(G)\Delta E(G')| = 1$, where $E(G)$ denotes the set of edges of graph G.*

Definition 2.2. *Let $\mathcal{G}$ denote the set of all undirected graphs. A (randomized) algorithm $M : \mathcal{G} \to R$ is (ϵ, δ)-edge differentially private if for all subsets $S \subset R$ of its output space, and for all $G, G' \in \mathcal{G}$, with $G \sim G'$, we have $Pr[M(G) \in S] \leq e^{\epsilon} Pr[M(G') \in S] + \delta$.*

One of the most common mechanisms for guaranteeing differential privacy is the Laplace mechanism (Dwork et al., 2014), which adds suitably scaled Laplace noise to a statistic to guarantee privacy. We now state the mechanism formally in the context of graph algorithms.

Definition 2.3. *Let $h : \mathcal{G} \to \mathbb{R}$ be any graph statistic. The Laplace mechanism $\mathcal{M}_h$ is defined as $\mathcal{M}_h(G) = h(G) + Lap(GS_h/\epsilon)$, where $GS_h = \max_{G \sim G'} |h(G) - h(G')|$ is the global sensitivity of h.*

Lemma 2.4. *The Laplace mechanism $\mathcal{M}_h$ is ϵ-edge *differentially private and has expected error which scales with the global sensitivity (i.e., $\mathbb{E}[|\mathcal{M}_h(G) - h(G)|] = c \cdot GS_h/\epsilon$ for some absolute constant $c > 0$).*

2.3 Problem Definition

In our problem, we will be given a contact graph $G = (V, E)$ and transmission probability p. We will also be given a parameter s, indicating the number of starting infections I_0. Our goal is to estimate the expected number of infections in the graph G when the s source nodes are chosen uniformly at random. We denote this quantity by $f_s(G)$. Formally, our goal is to obtain an (α, β)-approximation for computing $f_s(G)$.

Definition 2.5. *We say $\hat{f}(G)$ is an (α,β)-approximation if $f_s(G) \leq \hat{f}_s(G) \leq \alpha \cdot f_s(G) + \beta$.*

We make some remarks on our problem definition. First, we note that our model has random sources instead of a fixed set of sources given as input. This is justified in Section 6, where we show strong lower bounds for the fixed-sources model: $\Omega(n)$ additive error is necessary even for expander graphs. Second, we note that our notion of approximation contains multiplicative and additive error. This is the standard one in differential privacy, since additive noise is needed to guarantee privacy while multiplicative approximation is necessary even in the non-private setting.

3 Analysis of Global Sensitivity

In this section, we show that the global sensitivity of $f_s(G)$ is $\Theta(n/s)$ when $p = 1$, where s is the number of random source nodes. We will show that this implies a differentially private algorithm for estimating the expected number of infections with multiplicative error $1 + \eta$ and additive error $\Theta(n/s)$, for any given $\eta > 0$. We first give a proof of the global sensitivity bound.

Lemma 3.1. *For $p = 1$, the global sensitivity of $f_s(G)$ is $\Theta(n/s)$ for each s.*

Proof. Consider the vector $x(G) \in [0,1]^V$ where $x_v(G)$ for each $v \in V$ is the probability that v is infected under the random seeded starting infections. Let $C_1, \ldots, C_r$ denote the connected components of G and let $C(v)$ denote the connected component node $v \in V$ lies in. We can calculate $x_v = 1 - (1 - \frac{|C(v)|}{n})^s$ and note that $f_s(G)$ is exactly the ℓ_1-norm of $x(G)$.

Now, we calculate the global sensitivity of the statistic $f_s(G)$ under the addition or removal of an edge. As mentioned above, this is exactly the ℓ_1-sensitivity of the vector $x(G)$. Since we are considering the global sensitivity, we can without loss of generality consider only edge additions; let's say the edge $(u, v) \in E$ is added to the graph G and suppose $C(u) \neq C(v)$. Now, consider the changes in the vector $x(G)$. For $w \in C(u)$, the probability of infection increases from $1 - (1 - \frac{|C(u)|}{n})^s$ to $1 - (1 - \frac{|C(u)|+|C(v)|}{n})^s$. Similarly, for $w \in C(v)$, the probability of infection increases from $1 - (1 - \frac{|C(v)|}{n})^s$ to $1 - (1 - \frac{|C(u)|+|C(v)|}{n})^s$. Since the remaining probabilities don't change, the ℓ_1-sensitivity of x is $A_1 + A_2$, where

$$A_1 = |C(u)| \cdot \left[(1 - \frac{|C(u)|}{n})^s - (1 - \frac{|C(u)| + |C(v)|}{n})^s\right]$$

$$A_2 = |C(v)| \cdot \left[(1 - \frac{|C(v)|}{n})^s - (1 - \frac{|C(u)| + |C(v)|}{n})^s\right]$$

The global sensitivity is the maximum of $A_1 + A_2$ over all possible $|C(u)|, |C(v)| \in [n]$ such that $|C(u)| + |C(v)| \leq n$. This can clearly be computed in polynomial time, so we now have an efficient and private algorithm for computing the expected infections size.

We will now try to upper bound this expression to obtain accuracy guarantees. Then the expression above can be upper bounded by the following by using $1-x \leq e^{-x}$ for all $x \in \mathbb{R}$:

$$|C(u)| \cdot e^{-|C(u)|s/n} + |C(v)| \cdot e^{-|C(v)|s/n} \qquad (1)$$

Since the two terms in the above expression are independent with respect to $|C(u)|, |C(v)|$, we can optimize them separately. Simple calculus then gives us that the expression is maximized when $|C(u)| = \frac{n}{s}$, so we have an upper bound of $\frac{n}{es}$. A similar procedure shows that the second term is also upper bounded by $\frac{n}{es}$, so the entire expression is upper bounded by $\frac{2n}{es}$, so the global sensitivity is $O(n/s)$. Taking $|C(u)| = |C(v)| = n/s$ in the expressions for A_1 and A_2 also shows that the global sensitivity is $\Omega(n/s)$, proving Lemma 3.1. □

Next, we use our bound on the global sensitivity in the deterministic case to give an algorithm with nontrivial guarantees in the general probabilistic case.

Theorem 3.2. *Given any constant $\eta > 0$, there exists a polynomial-time ϵ-edge differentially private algorithm which computes a $(1 + \eta, O(n/s\epsilon))$-approximation for $f_s(G)$ in expectation.*

Proof. Observe that the running an SIR process on the graph with probability p is equivalent to removing each edge with probability p and then running an SIR process on the resulting graph with probability 1. Given this observation, our algorithm proceeds as follows. We sample N graphs $G_1, \ldots, G_N$, where each G_i is obtained from G by retaining edge with probability p. Then we compute the average $\tilde{f}_s(G) = \frac{1}{N}\sum_{i=1}^{N} f_s(G_i)$. By Lemma 3.1, we know the global sensitivity of each

$f_s(G_i)$ is $O(n/s)$ so the global sensitivity of the average $\tilde{f}_s(G)$ must also be $O(n/s)$. Our algorithm concludes by outputting $\tilde{f}_s(G)$ using the Laplace mechanism, which guarantees edge-differential privacy by Lemma 2.4. By a standard argument via Chernoff-Hoeffding bounds, the average is within a $1 \pm \eta$ multiplicative factor of $f_s(G)$ when $N = \Omega(n^2)$, giving us the multiplicative approximation guarantee. The additive approximation guarantee follows directly from combining the bound on the sensitivity with the utility of the Laplace mechanism (Lemma 2.4). □

4 Smooth Sensitivity and Its Difficulties

In this section, we explore a local sensitivity-based approach to the problem, called *smooth sensitivity.* We show that one can compute a good approximation for smooth sensitivity, which we show suffices for guaranteeing differential privacy. The drawback of this approach is that it is difficult to obtain polynomial run-time, and all algorithms here require quasi-polynomial time to implement. This illustrates the difficulties when applying local sensitivity-based approaches, which we overcome in Section 5.1. We note that a similar problem arises when using an approximate version of *inverse sensitivity* to estimate the infection size, but we omit the details here as it is unrelated to our main results.

First, let us recall the definition of smooth sensitivity from Nissim et al. (2007).

Definition 4.1. *For $\beta > 0$, the β-Smooth Sensitivity of f is*

$$S^*_{f,\beta}(x) = \max_y (LS_f(y)e^{-\beta d(x,y)}),$$

where $LS_f(y)$ is the local sensitivity of f at y.

Nissim et al. (2007) show that for some $\beta = \beta(\epsilon, \delta)$, adding Laplace noise calibrated by the β-Smooth Sensitivity to the statistic $f(x)$ suffices to guarantee (ϵ, δ)-differential privacy. However, this mechanism requires the exact calculation of the β-Smooth Sensitivity, which is intractable in our setting. We demonstrate here that instead of the exact calculation, we may use the approximation of the Smooth Sensitivity to calibrate the noise. This slightly generalizes recent work by Nguyen et al. (2023), which used approximate smooth sensitivity for faster subgraph counting.

To state and prove our results, we first define our notion of approximating the smooth sensitivity and recall the definition of an admissible noise distributions from Nissim et al. (2007).

Definition 4.2. *We say that $\tilde{S}_{f,\beta}$ is a (γ, τ, δ')-approximation of $S^*_{f,\beta}$ if with probability at least $1-\delta'$, we have the following for all datasets x:*

$$S^*_{f,\beta}(x) \le \tilde{S}_{f,\beta}(x) \le \exp(\gamma) \cdot S^*_{f,\beta}(x) + \tau.$$

We may drop the parameter τ in the notation when $\tau = 0$.

Definition 4.3. *A probability distribution h on $\mathbb{R}$ is (α, β)-admissible if for all $\Delta, \lambda \in \mathbb{R}$ with $|\Delta| \le \alpha, |\lambda| \le \beta$, and for all measurable $S \subset \mathbb{R}$, we have the following*

$$\Pr_{Z\sim h}[Z \in S] \le \exp(\epsilon/2) \cdot \Pr_{Z\sim h}[Z \in S + \Delta] + \delta/2 \quad (2)$$

$$\Pr_{Z\sim h}[Z \in S] \le \exp(\epsilon/2) \cdot \Pr_{Z\sim h}[Z \in \exp(\lambda) \cdot S] + \delta/2. \quad (3)$$

In particular, the Lap(1) distribution is $(\frac{\epsilon}{2}, \frac{\epsilon}{2\ln(2/\delta)})$-admissible.

Now, we are ready to state the privacy guarantee of approximate smooth sensitivity mechanisms.

Theorem 4.4. *Let $\tilde{S}_{f,\beta}$ be a (γ, τ, δ')-approximation of $S^*_{f,\beta}$, assume we have a lower bound ρ on the smallest value of the smooth sensitivity $\min_x S^*_{f,\beta}(x)$, and let Z be sampled from some $(\alpha, \gamma+\beta+\tau/(\rho e^{\gamma}))$-admissible distribution. Then mechanism $M_f(x) = f(x) + \frac{\tilde{S}(x)}{\alpha}Z$ is $(\epsilon, \frac{e^{\epsilon/2}+1}{2}\delta + 2\delta')$-differentially private.*

Proof. Fix a pair of neighboring datasets $x \sim y$. Let $\mathcal{E}$ be the event that $S^*_{f,\beta}(x) \le \tilde{S}_{f,\beta}(x) \le e^{\gamma} S^*_{f,\beta} + \tau$ and $\mathcal{E}'$ be the event that $S^*_{f,\beta}(y) \le \tilde{S}_{f,\beta}(y) \le e^{\gamma} S^*_{f,\beta}(y) + \tau$. From the definition of our notion of approximation, we have that $\Pr[\mathcal{E} \text{ and } \mathcal{E}'] \ge 1 - 2\delta'$ by the union bound. We will prove that conditioned on events $\mathcal{E}$ and $\mathcal{E}'$, we have that $M_f(x)$ is $(\epsilon, \frac{e^{\epsilon/2}+1}{2}\delta)$-differentially private. Since $\mathcal{E}$ and $\mathcal{E}'$ both occur with probability at least $1 - 2\delta'$, this implies the desired result directly by a standard characterization of differential privacy.

For the remainder of the proof, condition on the events that $\mathcal{E}$ and $\mathcal{E}'$ occur. Fix some measurable subset $S \subseteq \mathbb{R}$ of the co-domain of f. We have that $\Pr[M_f(x) \in S]$ is upper bounded by

$$\begin{aligned}
&\Pr\left[Z \in \frac{\alpha}{\tilde{S}_{f,\beta}(x)} \cdot [S - f(x)]\right] \\
&\overset{(a)}{\le} \exp(\epsilon/2) \cdot \Pr\left[Z \in \frac{\alpha}{\tilde{S}_{f,\beta}(x)} \cdot [S - f(y)]\right] + \frac{\delta}{2} \\
&\overset{(b)}{\le} \exp(\epsilon) \cdot \Pr\left[Z \in \frac{\alpha}{\tilde{S}_{f,\beta}(y)} \cdot [S - f(y)]\right] + \frac{\delta}{2}(1 + e^{\epsilon/2}) \\
&= \exp(\epsilon) \cdot \Pr[M_f(y) \in S] + \frac{\delta}{2}(1 + e^{\epsilon/2}).
\end{aligned}$$

In inequality (a), we used the Sliding Property with $\Delta = \alpha \cdot |f(x) - f(y)|/\tilde{S}_{f,\beta}(x) \le \alpha$ and in inequality (b), we used the Dilation Property with $\lambda =$

$\log[\tilde{S}_{f,\beta}(x)/\tilde{S}_{f,\beta}(y)]$, which satisfies the constraints of the Dilation property because

$$\log\left[\frac{\tilde{S}_{f,\beta}(x)}{\tilde{S}_{f,\beta}(y)}\right] \leq \log\left[\frac{e^{\gamma}S^*_{f,\beta}(x)+\tau}{S^*_{f,\beta}(y)}\right] \tag{4}$$

$$= \log\left[e^{\gamma+\beta}\right] + \log\left[1+\frac{\tau}{e^{\gamma}\cdot\min_x S^*(x)}\right] \tag{5}$$

$$\leq \gamma+\beta+\tau/(\rho e^{\gamma}). \tag{6}$$

Finally, the two equalities follow by definition of the mechanism $M_f(x)$, completing the proof. □

We now show how to approximately compute the smooth sensitivity for our problem in quasi-polynomial time. Specifically, fix some constant $\eta > 0$ and take $N = \Omega(n^2)$ samples $G_1, \dots, G_N \sim G(p)$ of subgraphs of G so that the average infections $\tilde{f}(G) = \frac{1}{N}\sum_{i=1}^N f(G_i)$ is within an $1 \pm \eta$ factor of $f(G)$. We will show how to approximately compute the smooth sensitivity of $\tilde{f}(G)$. Specifically, assume $\epsilon = O(1)$ and $\delta = 1/\text{poly}(n)$; we will show how to compute the β-smooth sensitivity of $\tilde{f}$ for $\beta = 1/O(\log n)$.

Recall the definition of the β-smooth sensitivity for f:

$$S^*_{\tilde{f},\beta}(G) = \max_{G'}\left(\text{LS}_{\tilde{f}}(G')\cdot e^{-\beta d(G,G')}\right)$$
$$= \max_{k=0,\dots,n^2}\ \max_{G':d(G,G')=k}\left(\text{LS}_{\tilde{f}}(G')\cdot e^{-\beta k}\right).$$

In the above equations, the second equality follows since $d(G, G')$ lies between 1 and n^2. Since $\text{LS}_{\tilde{f}}(G') \leq n$ for all graphs G', observe that whenever $k = \Omega(\log^2 n)$, we have that $\text{LS}_{\tilde{f}}(G')\cdot e^{-\beta k} \leq 1/\text{poly}(n)$. Hence, it suffices to compute the following approximation which only considers $k < \Theta(\log^2 n)$; this gives (γ, τ, δ')-approximation for $\gamma = 0$, $\tau = 1/\text{poly}(n)$, and $\delta' = 1/\text{poly}(n)$ for arbitrarily large polynomials in n. Formally, our approximation is defined as follows:

$$\tilde{S}_{\tilde{f},\beta}(G) = \max_{k=0,\dots,C\log^2 n/\epsilon}\ \max_{G':d(G,G')=k}\left(\text{LS}_{\tilde{f}}(G')\cdot e^{-\beta k}\right).$$

This approximation can easily be computed in $\tilde{O}(n^{\log^2 n})$ time by trying all possible G', since computing the local sensitivity of $\tilde{f}$ can be done in polynomial time.

In order to apply Theorem 4.4, we need a lower bound ρ on the smooth sensitivity of any input. To do this, we simply augment any input graph G with an additional isolated node before inputting the graph into the algorithm for computing approximate smooth sensitivity. This way, the algorithm may assume that all input graphs have at least one isolated node, so that the local sensitivity is always at least $\rho = 1$. Using this algorithm for computing approximate smooth sensitivity with the results of Theorem 4.4 gives a new private algorithm[1] for estimating the expected infection size. Unfortunately. it is unclear how to obtain an approximation for the smooth sensitivity faster than trying all possible G', since the structure of the local sensitivity of $\tilde{f}$ is so complex.

5 Polynomial-time Algorithm

In this section, we will give a polynomial-time algorithm for estimating the expected number of infections with differential privacy. Our approach will be based on the Propose-Test-Release framework of Dwork and Lei (2009), which is another instance of the local sensitivity-based methods explored in the previous section. We show that a suitable generalization of the framework suffices to give a polynomial-time algorithm. We then show that the utility guarantees of the algorithm match the ones given in the previous section up to poly-logarithmic factors. For ease of exposition, we will consider the case where there is only one random source (i.e., $s = 1$). Our claims and method extends to general s, but the expressions are more complicated.

5.1 Propose-Test-Release Framework

We first review the propose-test release framework for an arbitrary statistic $f(X)$ on database $X \in \mathcal{X}^n$.

1. Propose a bound β on the local sensitivity.
2. Compute distance $\gamma = d(X, \{X' : \text{LS}_f(X') \geq \beta\})$ to the closest database X' with $\text{LS}_f(X') \geq \beta$.
3. Compute noisy distance $\hat{\gamma} = \gamma + \text{Lap}(1/\epsilon)$.
4. If $\hat{\gamma} \leq \ln(1/\delta)/\epsilon$, return $\perp$. Otherwise, return $f(X) + \text{Lap}(\beta/\epsilon)$.

Dwork and Lei (2009) showed that the above mechanism preserves $(2\epsilon, \delta)$-differential privacy. Furthermore, the mechanism can provide much stronger utility guarantees if the bound β is much less than the global sensitivity. We generalize the framework slightly for our application. One difficulty in applying the framework is that the bound β on the local sensitivity needs to be derived analytically for a specific class of input databases. To provide more generality, we can instead use a noisy binary search to

[1] It turns out that this algorithm only incurs polylogarithmic additive error on a class of graphs called *expander graphs* (defined in Section 5.1), but we omit the proof since it the algorithm is superseded by the one in Section 5.1.

find a near-optimal bound β given the input database X. The other difficulty is computing the distance γ; this often cannot be computed in an efficient manner. Instead, we propose that it suffices to replace $\gamma(X)$ with any non-negative statistic $\phi(X)$ with sensitivity 1 which is a lower bound on $\gamma(X)$. In some cases such as ours, the algorithm designer can choose a $\phi(X)$ which can be efficiently computed and well approximates the original distance $\gamma(X)$; this freedom enables us to give computationally efficient algorithms using propose-test-release in new settings. Our mechanism is formalized below.

1. Let $\epsilon' = \epsilon / \log_2(\mathrm{GS}_f)$, where GS_f is an upper bound on the global sensitivity
2. Binary search $\beta \in \{1, \ldots, \lceil \mathrm{GS}_f \rceil\}$ for an upper bound on the local sensitivity $\mathrm{LS}_f(X)$.
   a. Compute lower bound $\phi(X)$ on distance $\gamma(X) = d(X, \{X' : \mathrm{LS}_f(X') \geq \beta\})$.
   b. Add Laplacian noise to the lower bound $\hat{\phi}(X) = \phi(X) + \mathrm{Lap}(1/\epsilon')$.
   c. If $\hat{\phi} \leq \ln(1/\delta)/\epsilon'$, increase guess β. Otherwise, decrease guess β.
3. Let $\hat{\beta}$ be the smallest β where $\phi(X) > \ln(1/\delta)/\epsilon'$.
4. Return $f(X) + \mathrm{Lap}(\hat{\beta}/\epsilon)$.

We will first show that the above algorithm is still $(2\epsilon, \delta)$-differentially private. In the next subsection, we will illustrate how the above variant of propose-test-release can be applied to our problem of estimating the expected number of infections in a contact network.

Lemma 5.1. *The above algorithm is* $(2\epsilon, \delta)$*-DP.*

Proof. Let's first analyze each iteration of the binary search. Since we have assumed that the statistic $\phi(X)$ has sensitivity 1, the output $\hat{\phi}(X)$ is ϵ'-differentially private by the privacy guarantees of the Laplace mechanism. Consequently, the decision of whether to increase or decrease β in the binary search is also ϵ'-differentially private by post-processing. By basic composition of adaptive mechanisms, we have $(\epsilon, 0)$-differential privacy for the output of the binary search.

Now consider step 6 of the mechanism and let $\hat{\beta}$ be as defined in the algorithm. Suppose that $\hat{\beta} \geq \mathrm{LS}_f(X)$; then by the privacy guarantees of the Laplace mechanism, we have ϵ-differential privacy. Now suppose that $\hat{\beta} < \mathrm{LS}_f(X)$; then we have $\gamma(X) = 0$ by definition which implies $\phi(X) = 0$ since $0 \leq \phi(X) \leq \gamma(X)$ for all X by assumption on ϕ. But by the PDF of the Laplace distribution, the probability that $\hat{\phi}(X) \geq \log(1/\delta)/\epsilon'$ is at most δ. As a result, we can conclude that step 6 is (ϵ, δ)-differentially private. Again by basic composition of adaptive mechanisms, we can conclude that our generalized propose-test-release framework is $(2\epsilon, \delta)$-differentially private. □

5.2 PTR for Infection Size Estimation

We will apply the generalized propose-test-release framework described in Section 5.1 to our problem of interest. The primary difficulty in applying the generalized propose-test-release framework is finding a non-negative statistic $\phi(X)$ which lower bounds $\gamma(X)$ and is efficiently computable. Recall that $\gamma(X)$ is defined to be the minimum number of edges which need to be added or removed such that the local sensitivity of f exceeds β. Thus, we need to understand the local sensitivity $\mathrm{LS}_f(G)$ and how it is affected by adding or removing edges from the graph before designing a good function $\phi(\cdot)$. As before, we first consider the case where the transmission is deterministic (Lemma 5.3). Then we will use results from the deterministic case to give an algorithm for the general problem with probabilistic transmission (Lemma 5.4).

For our algorithm, we will need the following result from Aissi et al. (2017):

Lemma 5.2. *For an undirected graph $G = (V, E)$, we say that a cut $(S, V - S)$ is a b-balanced cut if there is at least b vertices on the smaller side of the cut. The minimum b-balanced cut of a graph (and its corresponding size) can be computed in polynomial time (Aissi et al., 2017).*

Next, we define our statistic $\phi(G, \beta)$. For each $B \in [n^2]$ and each $k \leq B$, we will compute the following quantities: let $U_1(B, k) = \sum_{i=3}^{k+2} |C_i(G)|$ and let $U_2(B, k)$ denote the maximum b such that there is a b-balanced cut of size $B - k$. Finally, we define $\phi(G, \beta)$ as the minimum B such that there exists $k \leq B$ where $U_1(B, k) + U_2(B, k) \geq \beta$. We will now prove that this is a feasible statistic:

Lemma 5.3. *Assume $p = 1$. There exists a statistic $\phi(G, \beta)$ such that $0 \leq \phi(G, \beta) \leq \gamma(G, \beta)$ for all G, β. Further, $\phi(G, \beta)$ has sensitivity 1 and a polynomial time algorithm.*

Proof. Recall that the statistic f we wish to estimate is the expected number of infections in the giant component of G. Since there is only one random source ($s = 1$), we can write f as a function of $|C_1(G)|$ as follows: $f(G) = |C_1(G)| \cdot |C_1(G)|/n$. In particular, observe that f is a 1-Lipschitz function of $|C_1(G)|$ since $|C_1(G)| \leq n$ always, so it suffices to consider the local sensitivity of $|C_1(G)|$ and how it's affected by addition or removing edges from the graph. More formally, define $\tilde{f}(G) = |C_1(G)|$ to be our auxiliary statistic. Since f is a 1-Lipschitz function of $\tilde{f}$, we have that

$\mathrm{LS}_f(G) \leq \mathrm{LS}_{\tilde{f}}(G)$. This implies that the distance $\gamma_f(G,\beta)$ to the closest dataset which has large local sensitivity is lower bounded by $\gamma_{\tilde{f}}(G,\beta)$. As a result, it suffices to find a statistic $\phi_{\tilde{f}}(G,\beta)$ which lower bounds $\gamma_{\tilde{f}}(G,\beta)$ in order to obtain our desired statistic $\phi_f(G,\beta)$. For the remainder of the proof, we will be working with the auxiliary statistic $\tilde{f}(G) = |C_1(G)|$.

Now let β be given; we wish to find the minimum number of edges to add or remove from G to obtain G' satisfying $\mathrm{LS}_{\tilde{f}}(G') \geq \beta$. We will first characterize its local sensitivity $\mathrm{LS}_{\tilde{f}}(G)$. If one edge is removed from G, the size of the giant component can only change if G is a bridge graph; the local sensitivity will then be the size of the smaller half of the bridge graph. If one edge is added to G, the size of the largest component can change by at most $|C_2(G)|$ by adding an edge connecting the first and second largest components. We will compute a lower bound $\hat{\phi}_{\tilde{f}}(G)$ on the minimum number of edges to add or remove from G to obtain G' so that $|C_2(G')| \geq \beta$. We claim $\phi_{\tilde{f}}(G) \coloneqq \hat{\phi}_{\tilde{f}}(G) - 1$ is a valid lower bound for $\gamma_{\tilde{f}}(G)$. This is because the two quantities differ by at most 1, by adding or removing the single edge which connects the two components.

As a consequence of the above claim, we only need to reason about increasing $|C_2(G')|$ by adding or removing edges to G. Suppose for each B and k, we can show that $U_1(B,k) + U_2(B,k)$ is an upper bound of how much $|C_2(G)|$ can increase. We claim that this implies that the output $\hat{\phi}_{\tilde{f}}(G)$ of the algorithm is a lower bound for the number of edge changes required to obtain $|C_2(G')| \geq \beta$. Indeed, if it was possible to use less than $\hat{\phi}_{\tilde{f}}(G)$ edge changes to obtain $|C_2(G')| \geq \beta$. Then using those exact edge changes, we can obtain $U_1(B,k) + U_2(B,k) \geq \beta$, contradicting the assumption that $\hat{\phi}_{\tilde{f}}(G)$ was the minimal B output by the algorithm. Thus, it suffices to show that $U_1(B,k) + U_2(B,k)$ is a valid upper bound. The remainder of the proof focuses on showing this claim. The outline is as follows. We show that given a graph, the amount which k edge additions can increase $|C_2(G)|$ is upper bounded by $\mathrm{LS}_{add}(G) \leq U_1(B,k)$. In order to make use of the $B-k$ edge removals, our goal is to find $B-k$ edges to remove so that the impact $\mathrm{LS}_{add}(G)$ of the edge additions is maximized. We will then show that the amount which $\mathrm{LS}_{add}(G)$ can be increase by $B-k$ edge removals is at most $U_2(B,k)$. Combining the results gives our desired claim.

When adding edges, the optimal choice for increasing $|C_2(G)|$ is to iteratively add an edge connecting the second largest component $C_2(G')$ with the third largest component $C_3(G')$; this can be proven by a standard exchange argument. In particular, one can observe that $\mathrm{LS}_{add}(G)$ is a monotone and Lipschitz function of $|C_i(G)|$ for all $i > 1$. Furthermore, it is clear that the amount which the k edge additions can increase $|C_2(G')|$ without using the budget of $B-k$ for edge removals is upper bounded by $\mathrm{LS}_{add}(G) \leq \sum_{i=3}^{k+2} |C_i(G)|$. Next, we wish to find $B-k$ edges to remove first, so that the k edge additions afterwards can optimally increase $|C_2(G')|$. Since $\mathrm{LS}_{add}(G)$ is a monotone function of all $|C_i(G)|$ for $i > 1$, removing edges from within $C_i(G)$ for any $i > 1$ can never increase $\mathrm{LS}_{add}(G)$. Thus, we may assume without loss of generality that all edges are removed from $C_1(G)$. Since $\mathrm{LS}_{add}(G)$ is a Lipschitz function of each $|C_i(G)|$ for $i > 1$, the amount which $\mathrm{LS}_{add}(G)$ increases is at most the amount which $|C_1(G)|$ decreases (since $\sum_{i=1}^{n} |C_1(G)|$ is fixed). □

Next, we use the results in Lemma 5.3 to solve the general problem. Let N denote the number of samples which we will specify later. As in the non-private version of the problem, let $G_1, \ldots, G_N$ be N samples of the infection process, taking each edge in G with probability p. Our estimate of the number of infections in G will be $f_p(G) = \frac{1}{N}\sum_{i=1}^{N} f(G_i)$. We will now apply the generalized propose-test-release framework to our statistic $f_p(G)$, which uses the following definition of $\phi(G,\beta)$:

Lemma 5.4. *Let $\phi(G,\beta) = \min_{i\in[n]} \phi(G_i,\beta)$, where $\phi(G_i,\beta)$ is defined as in Lemma 5.3. Then $\phi(G,\beta)$ is a valid lower bound on $\gamma(G,\beta)$ and has sensitivity 1.*

Proof. Observe that we have

$$\phi(G,\beta) = \min_{i\in[n]} \phi(G_i,\beta) \leq \min_{i\in[n]} \gamma(G_i,\beta)$$

since we already have that $\phi(G_i,\beta) \leq \gamma(G_i,\beta)$ for each $i \in [n]$ by Lemma 5.3. Thus, it suffices to show that $\min_{i\in[n]} \gamma(G_i,\beta) \leq \gamma(G,\beta)$. Suppose for the sake of contradiction that $\gamma(G,\beta) < \min_{i\in[n]} \gamma(G_i,\beta)$. Then there exists a sequence of γ edge additions or removals from G to obtain G' where $\mathrm{LS}_{f_p}(G') \geq \beta$. But note that we have $\mathrm{LS}_{f_p}(G') \leq \frac{1}{N}\sum_{i=1}^{N} \mathrm{LS}_f(G_i')$, so there exists some $i \in [n]$ such that $\mathrm{LS}_f(G_i') \geq \beta$. But note that $d(G_i, G_i') \leq d(G, G')$ since G (resp. G') contains an edge if and only if G_i (resp. G_i') contains the edge. Consequently, we can conclude that $\gamma(G,\beta)$ edges also suffices to transform G_i into some G_i' so that $\mathrm{LS}_f(G_i') \geq \beta$. But this implies that $\min_{i\in[n]} \gamma(G_i,\beta) \leq \gamma(G,\beta)$, contradicting our original assumption. Our claim that $\phi(G,\beta)$ has sensitivity 1 follows directly since $\phi(G_i,\beta)$ has sensitivity 1, so we are done. □

Now that we have a valid statistic $\phi(G,\beta)$, we can apply the generalized propose-test-release framework from Section 5.1 to obtain a polynomial-time private mechanism for releasing an estimate of the expected number of infections. In the next subsection, we will analyze its utility in a special class of graphs.

5.3 Discussion of Utility Guarantees

Let us first define the class of graphs we will work with.

Definition 5.5. *A (n, d, λ)-spectral expander is a d-regular graph on n vertices where its eigenvalues (that is, the eigenvalues of its normalized adjacency matrix) $1 = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ satisfy $\lambda = \max\{|\lambda_2|, |\lambda_n|\}$.*

Spectral expanders have many nice properties which will enable us to give theoretical guarantees when computing the expected outbreak size. For example, the size of the giant component of G under percolation is of size $\Theta(n)$ when an outbreak occurs in G. Such a property is necessary for our approach to work since we made a simplifying assumption that the smaller components only contribute to lower order terms.

Lemma 5.6. *Let G be a (n, d, λ)-spectral expander and let $G(p)$ denote the resulting (random) subgraph formed by retaining each edge of G with probability p. Then for $p = \frac{1+\alpha}{d}$, there is (with high probability) a unique giant component of size $\Theta(n)$ and all other components are of size $O(\log n)$.*

Additionally, spectral expanders have the nice property that the resulting giant component after percolation is a vertex expander for sets which are not too small (Diskin and Krivelevich, 2022). Using this property, we can show that our statistic $\phi(G, \beta)$ is not too small when $\beta = \text{polylog}(n)$. As a result, we can obtain high accuracy estimates of the expected outbreak size.

Lemma 5.7. *Let $\alpha > 0$ be a small enough constant and let $\beta > 0$ be such that $\beta \leq \alpha^4$. Let $G = (V, E)$ be a (n, d, λ)-spectral expander with $\lambda \leq \delta$. Fix $p \geq \frac{1+\alpha}{d}$ and let $G(p)$ denote the resulting (random) subgraph formed by retaining each edge of G with probability p. If C_1 is the largest component of $G(p)$, then there exists an absolute constant $c > 0$ such that with high probability for any $S \subseteq L_1$ with $\frac{16\ln(n)}{\alpha^2} \leq |S| \leq \frac{\alpha^2 n}{50}$, we have $|N_{G(p)}(S)| \geq \frac{c\alpha^2|S|}{\ln(1/\alpha)}$.*

Now that we have stated the necessary properties, let us prove our utility guarantees.

Theorem 5.8. *Let G be a (n, d, λ)-spectral expander. Then our algorithm is an $(1+\eta, poly\log(n)/\epsilon)$-approximation algorithm for estimating the expected number of infections, with high probability.*

Proof. Consider $\beta' = C\log^3(n)$ for some sufficiently large constant C; we will show that for all $\beta \geq \beta'$, we have $\hat{\phi} \geq \ln(1/\delta)/\epsilon'$ in Line 5 with high probability for our statistic ϕ described above. As a consequence, we have that $\hat{\beta} \leq \beta'$ in Line 6, so our additive error is at most $O(\beta' \cdot \log(n)/\epsilon) = O(\log^3(n))/\epsilon$ with high probability by the tail bounds of a Laplace distribution. Since we have that $f(X)$ is within a $1 \pm \eta$ multiplicative factor of the true expected number of infections by a standard Chernoff-Hoeffding argument, we will have our desired approximation guarantees.

Fix $\beta \geq \beta'$. We will show that the lower bound ϕ on the distance $d(X, \{X' : \text{LS}_f(X') \geq \beta\}$ satisfies

$$\phi \geq C' \log^2(n)/\epsilon'$$

for some large enough constant C'. This would imply that with high probability, we have $\hat{\phi}(X) > \ln(1/\delta)/\epsilon'$ by the concentration of the Laplace distribution and the assumption that $\delta = 1/\text{poly}(n)$. The claim that $\phi \geq C' \log^2(n)/\epsilon'$ follows by the properties of spectral expander graphs discussed before. If we use edge additions, we know that all non-giant components are of size $O(\log n)$ so it requires at least $\Omega(\log^2(n))$ edge additions in order to increase the local sensitivity to be greater than β. If we use edge deletions, then it requires $\Omega(\log^2(n))$ edge deletions since the sampled graph $G(p)$ is an edge expander with high probability (see Theorem 5.7). Thus, we have proven that $\phi \geq C' \log^2(n)/\epsilon'$ and we are done. □

6 Lower Bounds

In this section, we'll give lower bounds for other natural settings in which one may wish to estimate the expected infection size. For example, our model assumes a random set of sources. We show when the set of source nodes are given, there are strong lower bounds even when the underlying graph is a spectral expander. Similarly, we show in the Appendix that if we instead wish to guarantee attribute privacy instead of edge privacy (in the fixed-source model), there are similar lower bounds showing that incurring $\Theta(n)$ additive error is inevitable. These lower bounds explain and justify our use of the random source model in our paper.

6.1 Lower Bound for Fixed Sources

Here, we justify our use of the random source SIR model by showing if we allow for fixed sources, any differentially private algorithm must incur $\Theta(n)$ additive error. We remark that our lower bound even applies to spectral expanders in the interesting regime of sparse graphs, which most social network graphs satisfy. This is in contrast to our positive results in the previous section for the random source model.

Theorem 6.1. *Let $\epsilon > 0$ be a constant and let $\delta = o(1)$. Then any (ϵ, δ)-differentially private algorithm for estimating the expected number infections* *with a fixed source must incur an additive $\Theta(n)$ expected error, even if the underlying graph is a constant degree spectral expander.*

Proof. Suppose for contradiction that there exists some (ϵ, δ)-differentially private algorithm M which guarantees sublinear additive error for the problem with a fixed source. Take any (n, d, λ)-spectral expander G for some constant d and λ with some fixed source node s and let the transmission probability be $p = 1$. Let G' denote the graph with all edges incident on s removed from G and note that G' is a d-edge neighbor of G. Since G is an expander, we know that it is connected so the expected number of infections in G with source s is n. Because the M has sublinear expected error, we know that

$$\Pr[A(G) \in o(n)] \leq o(1).$$

But by group privacy properties, we have that

$$\Pr[A(G') \in o(n)] \leq \exp(d\epsilon) \cdot o(1) + de^{(d-1)\epsilon}\delta.$$

Since we have assumed that $\epsilon, d = \Theta(1)$ and $\delta = o(1)$, the right hand side is $o(1)$ so $A(G') \in \Theta(n)$ with high probability. But in G', the source node is an isolated vertex so the expected number of infections is 1, so the expected error of the algorithm when run on G' is $\Theta(n)$, a contradiction. □

7 Conclusion

Our work initiates the study of estimating the expected infection size of an epidemic, modeled by an SIR model, under edge differential privacy. Our main result is a polynomial-time edge-differentially private algorithm for the problem with only poly-logarithmic additive error on expander graphs. We believe our algorithms perform well on real-world graphs, because they often have good expansion properties. However, we note that this isn't immediately implied by our theoretical results, as our results only apply for d-regular graphs. It remains interesting to test if our algorithm actually performs well in practice. Another interesting question is whether or not looser graph properties suffice for guaranteeing sublinear additive error for this problem. As a concrete starting case, it would be interesting to see if edge expansion in general (not necessarily regular) graphs suffices.

Acknowledgements

George Z. Li was supported by NSF grants CNS-2317194 and CCF-1918749. Dung Nguyen and Anil Vullikanti was partially supported in part by the following grants: National Science Foundation Grants CCF-1918656 (Expeditions), OAC-1916805 (CINES), IIS-1955797, and CNS-2317193, VDH Grant PV-BII VDH COVID-19 Modeling Program VDH-21-501-0135, DTRA subcontract/ARA S-D00189- 15-TO-01-UVA, NIH 2R01GM109718-07, CDC MIND cooperative agreement U01CK000589.

References

A. Adiga, D. Dubhashi, B. Lewis, M. Marathe, S. Venkatramanan, and A. Vullikanti. Mathematical models for covid-19 pandemic: a comparative analysis. *Journal of the Indian Institute of Science*, pages 1–15, 2020.

A. Adiga, G. Kaur, B. Hurt, L. Wang, P. Porebski, S. Venkatramanan, B. Lewis, and M. Marathe. Enhancing covid-19 ensemble forecasting model performance using auxiliary data sources. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1594–1603. IEEE, 2022.

H. Aissi, A. R. Mahjoub, and R. Ravi. Randomized contractions for multiobjective minimum cuts. In K. Pruhs and C. Sohler, editors, *25th Annual European Symposium on Algorithms, ESA 2017, September 4-6, 2017, Vienna, Austria*, volume 87 of *LIPIcs*, pages 6:1–6:13. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017. doi: 10.4230/LIPIcs.ESA.2017.6. URL https://doi.org/10.4230/LIPIcs.ESA.2017.6.

A. Akinbi, M. Forshaw, and V. Blinkhorn. Contact tracing apps for the covid-19 pandemic: a systematic literature review of challenges and future directions for neo-liberal societies. *Health Information Science and Systems*, 9:1–15, 2021.

H. Asi and J. C. Duchi. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

J. Blocki, A. Blum, A. Datta, and O. Sheffet. Differentially private data analysis of social networks via restricted sensitivity. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, page 87–96, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318594. doi: 10.1145/2422436.2422449. URL https://doi.org/10.1145/2422436.2422449.

L. Dhulipala, Q. C. Liu, S. Raskhodnikova, J. Shi, J. Shun, and S. Yu. Differential privacy from locally adjustable graph algorithms: k-core decomposition, low out-degree ordering, and densest subgraphs. In *2022 IEEE 63rd Annual Symposium on Foundations* *of Computer Science (FOCS)*, pages 754–765. IEEE Computer Society, 2022. doi: 10.1109/FOCS54457.2022.00077.

L. Dhulipala, G. Z. Li, and Q. C. Liu. Near-optimal differentially private k-core decomposition, 2023.

S. Diskin and M. Krivelevich. Expansion in supercritical random subgraphs of expanders and its consequences, 2022. URL https://arxiv.org/abs/2205.04852.

C. Dwork and J. Lei. Differential privacy and robust statistics. In M. Mitzenmacher, editor, *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pages 371–380. ACM, 2009. doi: 10.1145/1536414.1536466. URL https://doi.org/10.1145/1536414.1536466.

C. Dwork, F. McSherry, K. Nissim, and A. D. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors, *Third Theory of Cryptography Conference, TCC 2006*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006. URL https://doi.org/10.1007/11681878_14.

C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

G. Harrison, J. Chen, H. Mortveit, S. Hoops, P. Porebski, D. Xie, M. Wilson, P. Bhattacharya, A. Vullikanti, L. Xiong, and M. Marathe. Synthetic data to support us-uk prize challenge for developing privacy enhancing methods: Predicting individual infection risk during a pandemic, 2023.

M. Hay, C. Li, G. Miklau, and D. Jensen. Accurate estimation of the degree distribution of private networks. In *2009 Ninth IEEE International Conference on Data Mining*, pages 169–178, 2009. doi: 10.1109/ICDM.2009.11.

J. Imola, T. Murakami, and K. Chaudhuri. Locally differentially private analysis of graph statistics. In *30th USENIX Symposium on Security*, 2021.

T. Ji, C. Luo, Y. Guo, J. Ji, W. Liao, and P. Li. Differentially private community detection in attributed social networks. In *Asian Conference on Machine Learning*, pages 16–31. PMLR, 2019.

V. Karwa, S. Raskhodnikova, A. Smith, and G. Yaroslavtsev. Private analysis of graph structure. *ACM Transactions on Database Systems (TODS)*, 39(3):1–33, 2014.

S. P. Kasiviswanathan, K. Nissim, S. Raskhodnikova, and A. Smith. Analyzing graphs with node differential privacy. In *Proceedings of the 10th Theory of Cryptography Conference*, TCC'13, pages 457–476. Springer-Verlag, 2013. URL http://dx.doi.org/10.1007/978-3-642-36594-2_26.

Y. Li, M. Purcell, T. Rakotoarivelo, D. Smith, T. Ranbaduge, and K. S. Ng. Private graph data release: A survey. *ACM Comput. Surv.*, jan 2023. ISSN 0360-0300. doi: 10.1145/3569085. URL https://doi.org/10.1145/3569085. Just Accepted.

K. Liu and V. Smith. Solution documentation of team puffle (cmu) at the us/uk pets prize challenge, 2023. https://hackmd.io/@kenziyuliu/pets-challenge.

T. K. K. Mamidi, T. K. Tran-Nguyen, R. L. Melvin, and E. A. Worthey. Development of an individualized risk prediction model for covid-19 using electronic health record data. *Frontiers in big Data*, 4: 675882, 2021.

M. Marathe and A. Vullikanti. Computational epidemiology. *Communications of the ACM*, 56(7):88–96, 2013.

Y. Mülle, C. Clifton, and K. Böhm. Privacy-integrated graph clustering through differential privacy. In *EDBT/ICDT Workshops*, volume 157, 2015.

D. Nguyen, M. M. Halappanavar, V. Srinivasan, and A. Vullikanti. Faster approximate subgraph counts with privacy. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Fqg9vGWy4k.

K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84, 2007.

J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. Private release of graph statistics using ladder functions. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, page 731–745. Association for Computing Machinery, 2015. doi: 10.1145/2723372.2737785. URL https://doi.org/10.1145/2723372.2737785.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, see Section 2]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]