arXiv:1208.4586v2 [cs.CR] 1 Feb 2013

Differentially Private Data Analysis of Social Networks via Restricted Sensitivity

Jeremiah Blocki Avrim Blum Anupam Datta
Or Sheffet
Carnegie Mellon University
{jblocki@cs, avrim@cs, danupam@andrew, osheffet@cs}.cmu.edu *

February 4, 2013

Abstract

We introduce the notion of *restricted sensitivity* as an alternative to global and smooth sensitivity to improve accuracy in differentially private data analysis. The definition of restricted sensitivity is similar to that of global sensitivity except that instead of quantifying over all possible datasets, we take advantage of any beliefs about the dataset that a querier may have, to quantify over a restricted class of datasets. Specifically, given a query f and a hypothesis $\mathcal{H}$ about the structure of a dataset D, we show generically how to transform f into a new query $f_{\mathcal{H}}$ whose global sensitivity (over all datasets including those that do not satisfy $\mathcal{H}$) matches the restricted sensitivity of the query f. Moreover, if the belief of the querier is correct (i.e., $D \in \mathcal{H}$) then $f_{\mathcal{H}}(D) = f(D)$. If the belief is incorrect, then $f_{\mathcal{H}}(D)$ may be inaccurate.

We demonstrate the usefulness of this notion by considering the task of answering queries regarding social-networks, which we model as a combination of a graph and a labeling of its vertices. In particular, while our generic procedure is computationally inefficient, for the specific definition of $\mathcal{H}$ as graphs of bounded degree, we exhibit efficient ways of constructing $f_{\mathcal{H}}$ using different projection-based techniques. We then analyze two important query classes: *subgraph counting queries* (e.g., number of triangles) and *local profile queries* (e.g., number of people who know a spy and a computer-scientist who know each other). We demonstrate that the restricted sensitivity of such queries can be significantly lower than their smooth sensitivity. Thus, using restricted sensitivity we can maintain privacy whether or not $D \in \mathcal{H}$, while providing more accurate results in the event that $\mathcal{H}$ holds true.

*This research was supported in part by CyLab at Carnegie Mellon University under grants DAAD19-02-1-0389 and W911NF-09-1-0273 from the Army Research Office, the National Science Foundation under grants CCF-1101215 and CCF-1116892, the NSF Science and Technology Center TRUST, and an NSF Graduate Fellowship, as well as by the MSR-CMU Center for Computational Thinking.

1 Introduction

The social networks we inhabit have grown significantly in recent decades with digital technology enabling the rise of networks like Facebook that now connect over 900 million people and house vast repositories of personal information. At the same time, the study of various characteristics of social networks has emerged as an active research area [10]. Yet the fact that the data in a social network might be used to infer sensitive details about an individual, like sexual orientation [15], is a growing concern among social networks' participants. Even in an 'anonymized' unlabeled graph it is possible to identify people based on graph structures [3]. In this paper, we study the feasibility of and design efficient algorithms to release statistics about social networks (modeled as graphs with vertices labeled with attributes) while satisfying the semantic definition of differential privacy [8,9].

A differentially private mechanism guarantees that any two neighboring data sets (i.e., data sets that differ only on the information about a single individual) induce similar distributions over the statistics released. For social networks, we consider two notions of neighboring or adjacent networks: (1) *edge adjacency* stipulating that adjacent graphs differ in just one edge or in the attributes of just one vertex; and (2) *vertex adjacency* stipulating that adjacent networks differ on just one vertex—its attributes or *any number* of edges incident to it.

For any given statistic or query, its global sensitivity measures the maximum difference in the answer to that query over all pairs of neighboring data sets [9]; global sensitivity provides an upper bound on the amount of noise that has to be added to the actual statistic in order to preserve differential privacy. Since the global sensitivity of certain types of queries can be quite high, the notion of smooth sensitivity was introduced to reduce the amount of noise that needs to be added while still preserving differential privacy [18].

However, a key challenge in the differentially private analysis of social networks is that for many natural queries, both global and smooth sensitivity can be very large. In the vertex adjacency model, consider the query "How many people in G_1 are a doctor or are friends with a doctor?" Even if the answer is 0 (e.g., there are no doctors in the social network) there is a neighboring social network G_2 in which the answer is n (e.g., pick an arbitrary person from G_1, relabel him as a doctor, and connect him to everyone). Even in the edge adjacency model, the sensitivity of queries may be high. Consider the query "How many people in G_1 are friends with two doctors who are also friends with each other?" In G_1 the answer may be 0 even if there are two doctors that everyone else is friends with (e.g, the doctors are not friends with each other), but the answer jumps to $n-2$ in a neighboring graph G_2 (e.g, if we simply connect the doctors to each other). In fact, even the first query can have high sensitivity in the edge-adjacency model if we just relabel a high-degree vertex as a doctor.

Yet, while these examples respect the mathematical definitions of neighboring graphs and networks, we note that in a real social network no single individual is likely to be directly connected with everyone else. Suppose that in fact a querier has some such belief $\mathcal{H}$ about the given network ($\mathcal{H}$ is a subset of all possible networks) such that its query f has low sensitivity restricted only to inputs and deviations within $\mathcal{H}$. For example, the querier may believe the following hypothesis ($\mathcal{H}_k$): the maximum degree of any node in the network is at most $k = 5000 \ll n \approx 9 \times 10^8$ (e.g, after reading a study on the anatomy of Facebook [21]). Can one in that case provide accurate answers in the event that indeed $G \in \mathcal{H}$ and yet preserve privacy no matter what (even if $\mathcal{H}$ is not satisfied)?

In this work, we provide a positive answer to this question. We do so by introducing the notion of *restricted sensitivity*, which represents the sensitivity of the query f over only the given subset

	Adjacency	Hypothesis	Query	Sensitivity	Efficient
Theorem 9	Any	Any	Any	$GS_{f_{\mathcal{H}}} = RS_f(\mathcal{H})$	No
Theorem 14	Edge	$\mathcal{H}_k$	Any	$GS_{f_{\mathcal{H}}} = 3RS_f(\mathcal{H})$	Yes
Theorem 18	Vertex	$\mathcal{H}_k$	Any	$S_{f_{\mathcal{H}}} = O(1) \times RS_f(\mathcal{H}_{2k})$	Yes

Table 1: Summary of Results. GS = global sensitivity, RS = restricted sensitivity, and S = smooth bound of local sensitivity.

	Subgraph Counting Query P		Local Profile Query	
Adjacency	Smooth	Restricted	Smooth	Restricted
Edge	$\lvert P\rvert\ k^{\lvert P\rvert-1}$	$\lvert P\rvert\ k^{\lvert P\rvert-1}$	$k+1$	$k+1$
Vertex	$O\left(n^{\lvert P\rvert-1}\right)$	$\lvert P\rvert\ k^{\lvert P\rvert-1}$	$n-1$	$2k+1$

Table 2: Worst Case Smooth Sensitivity over $\mathcal{H}_k$ vs. Restricted Sensitivity $RS_f(\mathcal{H}_k)$.

$\mathcal{H}$, and providing procedures that map a query f to an alternative query $f_{\mathcal{H}}$ s.t. f and $f_{\mathcal{H}}$ identify over the inputs in $\mathcal{H}$, yet the global sensitivity of $f_{\mathcal{H}}$ is comparable to just the restricted sensitivity of f. Therefore, the mechanism that answers according to $f_{\mathcal{H}}$ and adds Laplace random noise preserves privacy for *all* inputs, while giving good estimations of f for inputs in $\mathcal{H}$.

While our general scheme for devising such $f_{\mathcal{H}}$ is inefficient and requires that we construct a separate $f_{\mathcal{H}}$ for each query f, we also design a complementary *projection-based* approach. A projection of $\mathcal{H}$ is a function mapping all possible inputs (e.g., all possible n-node social networks) to inputs in $\mathcal{H}$ with the property that any input in $\mathcal{H}$ is mapped to itself. Therefore, a projection μ allows us to define $f_{\mathcal{H}}$ for any f, simply by composing $f_{\mathcal{H}} = f \circ \mu$. Moreover, if this projection μ satisfies certain smoothness properties, which we define in Section 4, then this function $f_{\mathcal{H}}$ will have its global sensitivity—or at least its smooth sensitivity over inputs in $\mathcal{H}$—comparable to only the restricted sensitivity of f. In particular, for the case $\mathcal{H} = \mathcal{H}_k$ (the assumption that the network has degree at most $k \ll n$), we show we can *efficiently* construct projections μ satisfying these conditions, therefore allowing us to efficiently take advantage of low restricted sensitivity. These results are given in Section 4 and summarized in Table 1.

The next natural question is: how much advantage does restricted sensitivity provide, compared to global or smooth sensitivity, for natural query classes and natural sets $\mathcal{H}$? In Section 5 we consider two natural classes of queries: local profile queries and subgraph counting queries. A local profile query asks how many nodes v in a graph satisfy a property which depends only on the immediate neighborhood of v (e.g, queries relating to clustering coefficients and bridges [10], or queries such as "how many people know two spies who don't know each other?"). A subgraph counting query asks how many copies of a particular subgraph P are contained in the network (e.g., number of triangles involving at least one spy). For the case $\mathcal{H} = \mathcal{H}_k$ for $k \ll n$ we show that the restricted sensitivity of these classes of queries can indeed be much lower than the smooth sensitivity. These results, presented in Section 5, are summarized in Table 2.

1.1 Related Work

Easley and Kleinberg provide an excellent summary of the rich literature on social networks [10]. Previous literature on differentially-private analysis of social networks has primarily focused on

the edge adjacency model in unlabeled graphs where sensitivity is manageable [1]. Triangle counting queries can be answered in the edge adjacency model by efficiently computing the smooth sensitivity [18], and this result can be extended to answer other counting queries [16]. [14] shows how to privately approximate the degree distribution in the edge adjacency model. The Johnson-Lindenstrauss transform can be used to answer all cut queries in the edge adjacency model [5].

The approach taken in the work of Rastogi et al. [19] on answering subgraph counting queries is the most similar to ours. They consider a bayesian adversary whose prior (background knowledge) is drawn from a distribution. Leveraging an assumption about the adversary's prior they compute a high probability upper bound on the local sensitivity of the data and then answer by adding noise proportional to that bound. Loosely, they assume that the presence of an edge does not presence of other edges more likely. In the specific context of a social network this assumption is widely believed to be false (e.g., two people are more likely to become friends if they already have common friends [10]). The privacy guarantees of [19] only hold if these assumptions about the adversaries prior are true. By contrast, we always guarantee privacy even if the assumptions are incorrect.

A relevant approach that deals with preserving differential privacy while providing better utility guarantees for nice instances is detailed in the work of Nissim et al [18] who define the notion of smooth sensitivity. In their framework, the amount of random noise that the mechanism adds to a query's true ansewr is dependent on the extent for which the input database is "nice" – having small *local sensitivity*. As we discuss later, in social networks many natural queries (e.g., local profile queries) even have high local and smooth sensitivity.

2 Preliminaries

2.1 Differential Privacy

We adopt the framework of differential privacy. We use $\mathcal{D}$ to denote the set of all possible datasets. Intuitively, we say two datasets $D, D' \in \mathcal{D}$ are *neighbors* if they differ on the details of a single individual. (See further discussion in Definitions 6 and 7.) We denote the fact that D' is a neighbor of D using $D' \sim D$. We define the *distance* $d(D, D')$ between two databases $D, D' \in \mathcal{D}$ as the minimal non-negative integer k s.t. there exists a path $D_0, D_1, \ldots, D_k$ where $D_0 = D$, $D_k = D'$ and for every $1 \leq i \leq k$ we have that $D_{i-1} \sim D_i$. Given a subset $\mathcal{D}' \subset \mathcal{D}$ we denote the distance of a database D to $\mathcal{D}'$ as $d(D, \mathcal{D}') = \min_{D' \in \mathcal{D}'} d(D, D')$.

Definition 1. *[8] A mechanism A is (ϵ, δ)-differentially private if for every pair of neighboring datasets $D, D' \in \mathcal{D}$ and every subset $S \subseteq Range(A)$ we have that* $\Pr[A(D) \in S] \leq e^{\epsilon} \Pr[A(D') \in S] + \delta$.

Intuitively differential privacy guarantees that an adversary has a very limited ability to distinguish between the output of $A(D)$ and the output of $A(D')$. A query is a function $f : \mathcal{D} \to \mathbb{R}$ mapping the dataset to a real number.

Definition 2. *The* local sensitivity *of a query f at a dataset D is* $LS_f(D) = \max_{D' \sim D} |f(D) - f(D')|$.

Definition 3. *The* global sensitivity *of a query f is* $GS_f = \max_{D \in \mathcal{D}} LS_f(D)$.

[1] Kasiviswanathan, Nissim, Raskhodnikova and Smith have independently been exploring and developed an analysis for node level privacy using an approach similar to ours (personal communication, 2012).

The Laplace mechanism $A(D) = f(D) + Lap(GS_f/\epsilon)$ preseves $(\epsilon, 0)$-differential privacy [9]. This mechanism provides useful answers to queries with low global sensitivity. The primary challenge in the differentially private analysis of social networks is the high global sensitivity of many queries. The local sensitivity $LS_f(D)$ may be significantly lower than the global sensitivity GS_f. However, adding noise proportional to $LS_f(D)$ does not preserve differential privacy because the noise level itself may leak information. A clever way to circumvent this problem is to smooth out the noise level [18].

Definition 4. *[18] A β-smooth upper bound on the local sensitivity of a query f is a function $S_{f,\beta}$ which satisfies (i) $\forall D \in \mathcal{D}, S_{f,\beta}(D) \geq LS_f(D)$, and (ii) $\forall D, D' \in \mathcal{D}$ it holds that $S_{f,\beta}(D) \leq \exp(-\beta d(D, D')) S_{f,\beta}(D')$.*

It is possible to preserve privacy while adding noise proportional to a β-smooth upper bound on the sensitivity of a query. For example, the mechanism $A(D) = f(D) + Lap\left(\frac{2S_{f,\beta}(D)}{\epsilon}\right)$, with $\beta = -\epsilon/2\ln\delta$ preserves (ϵ, δ)-differential privacy [18]. To evaluate A efficiently one must present an algorithm to efficiently compute the β-smooth upperbound $S_{f,\beta}(G)$, a task which is by itself often non-trivial.

2.2 Graphs and Social Networks

Our work is motivated by the challenges posed by differentially private analysis of social networks. As always, a graph is a pair of a set of vertices and a set of edges $G = \langle V, E \rangle$. We often just denote a graph as G, referring to its vertex-set or edge-set as $V(G)$ or $E(G)$ resp. A key aspect of our work is modeling a social network as a *labeled* graph.

Definition 5. *A* social network (G, ℓ) *is a graph with labeling function $\ell : V(G) \to \mathbb{R}^m$. The set of all social networks is denoted $\mathcal{G}$.*

The labeling function ℓ allows us to encode information about the nodes (e.g., age, gender, occupation). For convenience, we assume all social networks are over the same set of vertices, which is denotes as V and has size n, and so we assume $|V| = n$ is public knowledge.[2] Therefore, the graph structures of two social networks are equal if their edge-sets are identical. Similarly, we also fix the dimension m of our labeling.

Defining differential privacy over the labeled graphs $\mathcal{G}$ requires care. What does it mean for two labeled graphs $G_1, G_2 \in \mathcal{G}$ to be neighbors? There are two natural notions: edge-adjacency and vertex adjacency.

Definition 6 (Edge-adjacency)**.** *We say that two social networks (G_1, ℓ_1) and (G_2, ℓ_2) are neighbors if either (i) $E(G_1) = E(G_2)$ and there exists a vertex u such that $\ell_1(u) \neq \ell_2(u)$ whereas for every other $v \neq u$ we have $\ell_1(v) = \ell_2(v)$ or (ii) $\forall v, \ell_1(v) = \ell_2(v)$ and the symmetric difference $E(G_1) \triangle E(G_2)$ contains a single edge.*

In the context of a social network, differential-privacy w.r.t edge-adjacency can, for instance, guarantee that an adversary will not be able to distinguish whether a particular individual has friended some specific pop-singer on Facebook. However, such guarantees do not allow a person to pretend to listen only to high-end indie rock bands, should that person have friended numerous pop-singers on Facebook. This motivates the stronger vertex-adjacency neighborhood model.

[2] Adding or removing vertices could be done by adding one more dimension to the labeling, indicating whether a node is active or inactive.

Definition 7 (Vertex-adjacency)**.** *We say that two social networks (G_1, ℓ_1) and (G_2, ℓ_1) are neighbors if there exists a vertex v_i such that $G_1 - v_i = G_2 - v_i$ and $\ell_1(v_j) = \ell_2(v_j)$ for every $v_j \neq v_i$.*

where for a graph G and a vertex v we denote $G - v$ as the result of removing every edge in $E(G)$ that touches v.

It is evident that any two social networks that are edge-adjacent are also vertex-adjacent. Preserving differential privacy while guaranteeing good utility bounds w.r.t vertex-adjacency is a much harder task than w.r.t edge-adjacency.

Distance Given two social networks (G_1, ℓ_1) and (G_2, ℓ_2), recall that their distance is the minimal k s.t. one can form a path of length k, starting with (G_1, ℓ_1) and ending at (G_2, ℓ_2), with the property that every two consecutive social-networks on this path are adjacent. Given the above two definitions of adjacency, we would like to give an alternative characterization of this distance.

First of all, the set $U = \{v : \ \ell_1(v) \neq \ell_2(v)\}$ dictates $|U|$ steps that we must take in order to transition from (G_1, ℓ_1) to (G_2, ℓ_2). It is left to determine how many adjacent social-networks we need to transition through until we have $E(G_1) = E(G_2)$. To that end, we construct the difference-graph whose edges are the symmetric difference of $E(G_1)$ and $E(G_2)$. Clearly, to transition from (G_1, ℓ_1) to (G_2, ℓ_2), we need to alter every edge in the difference graph. In the edge-adjacency model, a pair of adjacent social networks covers precisely a single edge, and so it is clear that the distance $d\big((G_1, \ell_1), (G_2, \ell_2)\big) = |U| + |E(G_1) \ \triangle \ E(G_2)|$. In the vertex-adjacency model, a single vertex can cover all the edges that touch it, and so the distance between the graphs $G_1 - U$ and $G_2 - U$ is precisely the *vertex cover* of the difference graph. Denoting this vertex cover as $VC(G_1 - U \ \triangle \ G_2 - U)$ we have that $d\big((G_1, \ell_1), (G_2, \ell_2)\big) = |U| + |VC(G_1 - U \ \triangle \ G_2 - U)|$. It is evident that computing the distance of between any two social-networks in the vertex-adjacency model is a NP-hard problem.

To avoid cumbersome notation, from this point on we omit the differentiation between graphs and social networks, and denote networks as graphs $G \in \mathcal{G}$.

3 Restricted Sensitivity

We now introduce the notion of restricted sensitivity, using a hypothesis about the dataset D to restrict the sensitivity of a query. A hypothesis $\mathcal{H}$ is a subset of the set $\mathcal{D}$ of all possible datasets (so in the context of social networks, $\mathcal{H}$ is a set of labeled graphs). We say that $\mathcal{H}$ is true if the true dataset $D \in \mathcal{H}$. Because the hypothesis $\mathcal{H}$ may not be a convex set we must consider all pairs of datasets in $\mathcal{H}$ instead of all pairs of adjacent datasets as in the definition of global sensitivity.

Definition 8. *For a given notion of adjacency among datasets, the* restricted sensitivity *of f over a hypothesis $\mathcal{H} \subset \mathcal{D}$ is*

$$RS_f(\mathcal{H}) = \max_{D_1, D_2 \in \mathcal{H}} \Big(\frac{|f(D_1) - f(D_2)|}{d(D_1, D_2)} \Big) \ .$$

To be clear, $d(D_1, D_2)$ denotes the length of the shortest-path in $\mathcal{D}$ between D_1 and D_2 (not restricting the path to only use $D \in \mathcal{H}$) using the given notion of adjacency (e.g., edge-adjacency or vertex-adjacency). That is, we restrict the set of databases for which we compute the sensitivity, but we do not re-define the distances.

Observe that $RS_f(\mathcal{H})$ may be smaller than $LS_f(D)$ for some $D \in \mathcal{H}$ if D has a neighbor $D' \notin \mathcal{H}$. In fact we often have $LS_f(D) \geq |f(D) - f(D')| \gg RS_f(\mathcal{H})$. As an immediate corollary, in such cases $RS_f(\mathcal{H})$ will be significantly lower than $S_{f,\beta}(D)$, a β-smooth upper bound on $LS_f(D)$.

4 Using Restricted Sensitivity to Reduce Noise

To achieve differential privacy while adding noise proportional to $RS_f(\mathcal{H})$ we must be willing to sacrifice accuracy guarantees for datasets $D \notin \mathcal{H}$. Our goal is to create a new query $f_{\mathcal{H}}$ such that $f_{\mathcal{H}}(D) = f(D)$ for every $D \in \mathcal{H}$ ($f_{\mathcal{H}}$ is accurate when the hypothesis is correct) and $f_{\mathcal{H}}$ either has low global sensitivity or low β-smooth sensitivity over datasets $D \in \mathcal{H}$. In this section, we first give a non-efficient generic construction of such $f_{\mathcal{H}}$, showing that it is always possible to devise $f_{\mathcal{H}}$ whose global sensitivity equals *exactly* the restricted sensitivity of f over $\mathcal{H}$. We then show how for the case of social networks and for the hypothesis $\mathcal{H}_k$ that the network has bounded degree, we can construct functions $f_{\mathcal{H}_k}$ having approximately this property, efficiently.

4.1 A General Construction

We now show how given $\mathcal{H}$ to generically (but not efficiently) construct $f_{\mathcal{H}}$ whose global sensitivity *exactly* equals the restricted sensitivity of f over $\mathcal{H}$.

Theorem 9. *Given any query f and any hypothesis $\mathcal{H} \subset \mathcal{D}$ we can construct a query $f_{\mathcal{H}}$ such that*

1. *$\forall D \in \mathcal{H}$ it holds that $f_{\mathcal{H}}(D) = f(D)$, and*
2. *$GS_{f_{\mathcal{H}}} = RS_f(\mathcal{H})$*

Proof. For each $D \in \mathcal{H}$ set $f_{\mathcal{H}}(D) = f(D)$. Now fix an arbitrary ordering of the set $\{D : D \notin \mathcal{H}\}$, and denote its elements as $D_1, D_2, \ldots, D_m$, where m is the size of the set. For every $D \notin \mathcal{H}$ we define the value of $f_{\mathcal{H}}(D)$ inductively. Denote $\mathcal{T}_i = \mathcal{H} \bigcup \{D_1, ..., D_i\}$. Initially, we are given the values of every $D \in \mathcal{T}_0$. Given $i > 0$, we denote $\Delta_i = RS_{f_{\mathcal{H}}}(\mathcal{T}_i)$. We now prove one can pick the value $f_{\mathcal{H}}(D_i)$ in a way that preserves the invariant that $\Delta_{i+1} = \Delta_i$. By applying the induction m times we conclude that

$$RS_f(\mathcal{H}) = \Delta_0 = \Delta_m = RS_{f_{\mathcal{H}}}(\mathcal{D}) = GS_{f_{\mathcal{H}}} .$$

Fix $i > 0$. Observe that

$$\Delta_{i+1} = \max\left(\Delta_i, \left(\max_{D \in \mathcal{T}_i} \frac{|f_{\mathcal{H}}(D) - f_{\mathcal{H}}(D_{i+1})|}{d(D, D_{i+1})}\right)\right)$$

so to preserve the invariant it suffices to find any value of $f_{\mathcal{H}}(D_{i+1})$ that satisfies that for every $D \in \mathcal{T}_i$ we have $|f_{\mathcal{H}}(D) - f_{\mathcal{H}}(D_{i+1})| \leq \Delta_i \cdot d(D, D_{i+1})$. Suppose for contradiction that no value exists. Then there must be two intervals

$$[f_{\mathcal{H}}(D_1^*) - \Delta_i \cdot d(D_1^*, D_{i+1}),\ f_{\mathcal{H}}(D_1^*) + \Delta_i \cdot d(D_1^*, D_{i+1})]$$
$$[f_{\mathcal{H}}(D_2^*) - \Delta_i \cdot d(D_2^*, D_{i+1}),\ f_{\mathcal{H}}(D_2^*) + \Delta_i \cdot d(D_2^*, D_{i+1})]$$

which don't intersect. This would imply that

$$\frac{|f_{\mathcal{H}}(D_1^*) - f_{\mathcal{H}}(D_2^*)|}{d(D_1^*, D_2^*)} \geq \frac{|f_{\mathcal{H}}(D_1^*) - f_{\mathcal{H}}(D_2^*)|}{d(D_{i+1}, D_1^*) + d(D_{i+1}, D_2^*)} > \Delta_i$$

which contradicts the fact that Δ_i is the restricted sensitivity of $\mathcal{T}_i$. ∎

4.2 Efficient Procedures for $\mathcal{H}_k$ via Projection Schemes

Unfortunately, the construction of Theorem 9 is highly inefficient. Furthermore, this construction deals with one query at a time. We would like to a-priori have a way to efficiently devise $f_{\mathcal{H}}$ for any f. In this section, the way we devise $f_{\mathcal{H}}$ is by constructing a *projection* – a function $\mu : \mathcal{D} \to \mathcal{H}$ with the property that $\mu(D) = D$ for every $D \in \mathcal{H}$. Such μ allows us to canonically convert *any* f into $f_{\mathcal{H}}$ using the naïve definition $f_{\mathcal{H}} = f \circ \mu$. Below we discuss various properties of projections that allow us to derive "good" $f_{\mathcal{H}}$-s. Following each property, we exhibit the existence of such projections μ for the specific case of social networks and $\mathcal{H} = \mathcal{H}_k$, the class of graphs of degree at most k.

Definition 10. *The class $\mathcal{H}_k$ is defined as the set $\{G \in \mathcal{G} : \ \forall v, \ \deg(v) \leq k\}$.*

In many labeled graphs, it is reasonable to believe that $\mathcal{H}_k$ holds for $k \ll n$ because the degree distributions follow a power law. For example, the number of telephone numbers receiving t calls in a day is proportional to $1/t^2$, and the number of web pages with t incoming links is proportional to $1/t^2$ [7, 10, 17]. For these networks it would suffice to set $k = O(\sqrt{n})$. The number of papers that receive t citations is proportional to $1/t^3$ so we could set $k = O(\sqrt[3]{n})$ [10]. While the degrees on Facebook don't seem to follow a power law, the upper bound $k = 5,000$ seems reasonable [21]. By contrast, Facebook had approximately $n = 901,000,000$ users in June, 2012 [1].

4.2.1 Smooth Projection

The first property we discuss is perhaps the simplest and most coveted property such projection can have – smoothness. Smoothness dictates that there exists a global bound on the distance between any two mappings of two neighboring databases.

Definition 11. *A projection $\mu : \mathcal{D} \to \mathcal{H}$ is called* c-smooth *if for any two neighboring databases $D \sim D'$ we have that $d\big(\mu(D), \mu(D')\big) \leq c$.*

Lemma 12. *Let $\mu : \mathcal{D} \to \mathcal{H}$ be a c-smooth projection (i.e., for every $D \in \mathcal{H}$ we have $\mu(D) = D$). Then for every query f, the function $f_{\mathcal{H}} = f \circ \mu$ satisfies that $GS_{f_{\mathcal{H}}} \leq c \cdot RS_f(\mathcal{H})$.*

Proof.

$$\begin{aligned}
GS_{f_{\mathcal{H}}} &= \max_{D_1 \sim D_2} |f_{\mathcal{H}}(D_1) - f_{\mathcal{H}}(D_2)| \\
&= \max_{D_1 \sim D_2} |f(\mu(D_1)) - f(\mu(D_2))| \cdot 1 \\
&\leq \max_{D_1 \sim D_2} |f(\mu(D_1)) - f(\mu(D_2))| \ \frac{c}{d(\mu(D_1), \mu(D_2))} \\
&\leq c \cdot \max_{D_1, D_2 \in \mathcal{H}} \frac{|f(D_1) - f(D_2)|}{d(D_1, D_2)} \\
&= c \cdot RS_f(\mathcal{H})
\end{aligned}$$

□

As we now show, for $\mathcal{H} = \mathcal{H}_k$ and for distances defined via the edge-adjacency model, we can devise an efficient smooth projection.

Claim 13. *In the edge-adjacency model, there exists an efficiently computable 3-smooth projection to $\mathcal{H}_k$.*

The proof of the claim is deferred to the appendix. The high-level idea is to fix a canonical ordering over all edges and then define μ to delete an edge e if and only if there is a vertex v such that (1) e is incident to v and (2) e is not one of the first k edges incident to v. This is then used to achieve the smoothness guarantee. An immediate corollary of Lemma 12 and Claim 13 is the following theorem.

Theorem 14. *(Privacy wrt Edge Changes) Given any query for social networks f, the mechanism that uses the projection μ from Claim 13, and answers the query using $A(f,G) = f(\mu(G)) + Lap(3 \cdot RS_f(\mathcal{H}_k)/\epsilon)$ preserves $(\epsilon, 0)$ privacy for any graph G.*

Now, it is evident that this mechanism has the guarantee that for every $G \in \mathcal{H}_k$ it holds that $\Pr[\ |A(f,G) - f(G)| \leq O(RS_f(\mathcal{H}_k)/\epsilon)] \geq 2/3$. Furthermore, if the querier "lucked out" to ask a query f for which $f(G)$ and $f(\mu(G))$ are close (say, identical), then the same guarantee holds for such G as well. Note however that we *cannot reveal* to the querier whether $f(G)$ and $f(\mu(G))$ are indeed close, as such information might leak privacy.

4.2.2 Projections and Smooth Distances Estimators

Unfortunately, the smooth projections do not always exist, as the following toy-example demonstrates. Fix n graphs, where $d(G_i, G_j) = |i-j|$ for $1 \leq i, j \leq n$, and let $\mathcal{H} = \{G_1, G_n\}$. Because $\mu(G_1) = G_1$ and $\mu(G_n) = G_n$, then there must exist some value i such that $\mu(G_i) \neq \mu(G_{i+1})$, thus every μ cannot be c-smooth for $c < n$.

Note that smooth projections have the property that they also provide a c-approximation of the distance of D to $\mathcal{H}$. Meaning, for every D we have that $d(D,\mathcal{H}) \leq d(D,\mu(D)) \leq c \cdot d(D,\mathcal{H})$. In the vertex adjacency model, however, it is evident that we cannot have a $O(1)$-smooth projection since, as we show in the appendix, it is NP-hard to approximate $d(G,\mathcal{H}_k)$ (see Claim 23), but there does exists an efficient approximation scheme (see Claim 24) of the distance. Yet, we show that it is possible to devise a somewhat relaxed projection s.t. the distance between a database and its mapped image is a smooth function. To that end, we relax a little the definition of projection, allowing it to map instances to some predefined $\bar{\mathcal{H}} \supset \mathcal{H}$.

Definition 15. *Fix $\bar{\mathcal{H}} \supset \mathcal{H}$. Let μ be a projection of $\mathcal{H}$, so μ is a mapping $\mu : \mathcal{D} \to \bar{\mathcal{H}}$ that maps every element of $\mathcal{H}$ to itself ($\forall D \in \mathcal{H}$ we have that $\mu(D) = D$). A* c-smooth distance estimator *is a function $\hat{d}_\mu : \mathcal{D} \to \mathbb{R}$ that satisfies all of the following. (1) For every $D \in \mathcal{H}$ it is defined as $\hat{d}_\mu(D) = 0$. (2) It is lower bounded by the distance of D to its projection: $\forall D \in \mathcal{D},\ \hat{d}_\mu(D) \geq d(D,\mu(D))$. (3) Its value over neighboring databases changes by at most c: $\forall D \sim D',\ \left|\hat{d}_\mu(D) - \hat{d}_\mu(D')\right| \leq c$.*

It is simple to verify that for every $D \in \mathcal{D}$ we have that $\hat{d}_\mu(D) \leq c \cdot d(D,\mathcal{H})$ (using induction on $d(D,\mathcal{H})$). We omit the subscript when μ is specified.

The following lemma suggests that a smooth distance estimator allows us to devise a good smooth-upper bound on the local-sensitivity, thus allowing us to apply the smooth-sensitivity scheme of [18].

Lemma 16. *Fix $\bar{\mathcal{H}} \supset \mathcal{H}$ and let $\mu : \mathcal{D} \to \bar{\mathcal{H}}$ be a projection of $\mathcal{H}$. Let $\hat{d} : \mathcal{D} \to \mathbb{R}$ be an efficiently computable c-smooth distance estimator. Then for every query f, we can define the composition $f_{\mathcal{H}} = f \circ \mu$ and define the function*

$$S_{f_{\mathcal{H}},\beta}(D) = \max_{d \in \mathbb{Z}, d \geq \hat{d}(D)} e^{\left(-\frac{\beta}{c}\left(d - \hat{d}(D)\right)\right)} (2d + c + 1) \cdot RS_f\left(\bar{\mathcal{H}}\right)$$

Then $S_{f_{\mathcal{H}},\beta}$ is an efficiently computable β-smooth upper bound on the local sensitivity of $f_{\mathcal{H}}$. Furthermore, define g as the function $g(x) = \begin{cases} 2\frac{1}{x}e^{-1+\frac{c+1}{2}x}, & 0 \leq x \leq \frac{2}{c+1} \\ c+1, & x > \frac{2}{c+1} \end{cases}$. Then for every D it holds that

$$S_{f,\beta}(D) \leq \exp(\tfrac{\beta}{c}\hat{d}(D)) \cdot g(\beta/c) RS_f(\bar{\mathcal{H}})$$

The proof of Lemma 16 is deferred to the appendix. Like in the edge-adjacency model, we now exhibit a projection and a smooth distance estimator for the vertex-adjacency model.

Claim 17. *In the vertex-adjacency model, there exists a projection $\mu : \mathcal{G} \to \mathcal{H}_{2k}$ and a 4-smooth distance estimator $\hat{d}$, both of which are efficiently computable.*

To construct μ and $\hat{d}$ we start with the linear program that determines a "fractional distance" from a graph to $\mathcal{H}_k$. This LP has $n + \binom{n}{2}$ variables: x_u which intuitively represents whether x_u ought to be removed from the graph or not, and $w_{u,v}$ which represents whether the edge between u and v remains in the projected graph or not. We also use the notation $a_{u,v}$, where $a_{uv} = 1$ if the edge $\{u, v\}$ is in G; otherwise $a_{uv} = 0$.

$$\begin{aligned} \min \textstyle\sum_{v \in V} x_v \quad & s.t. \\ & (1)\ \forall v,\ x_v \geq 0 \\ & (2)\ \forall u, v,\ w_{u,v} \geq 0 \\ & (3)\ \forall u, v,\ a_{uv} \geq w_{uv} \geq a_{uv} - x_u - x_v \\ & (4)\ \forall u, \sum\nolimits_{v \neq u} w_{u,v} \leq k \end{aligned}$$

To convert our fractional solution $(\bar{x}^*, \bar{w}^*)$ to a graph $\mu(G) \in \mathcal{H}_{2k}$ we define $\mu(G)$ to be the graph we get by removing every edge $(u, v) \in E(G)$ whose either endpoint has weight $x^*_u > 1/4$ or $x^*_v \geq 1/4$. We define our distance estimator as $\hat{d}(G) = 4\sum_u x^*_u$. In the appendix we show that μ and $\hat{d}$ satisfy the conditions of claim 17.

As before, combining Lemma 16 with Claim 17 gives the following theorem as an immediate corollary.

Theorem 18. *(Privacy wrt Vertex Adjacency) Given any query for social networks f, the mechanism that uses the projection μ from Claim 13 and the β-smooth upper bound of Lemma 16, and answers the query using $A(f, G) = f(\mu(G)) + Lap(2 \cdot S_{f_{\mathcal{H}},-\epsilon/2\ln\delta}(G)/\epsilon)$ preserves (ϵ, δ) privacy for any graph G.*

Again, it is evident from the definition that the algorithm has the guarantee that for every $G \in \mathcal{H}_k$ it holds that $\Pr[\ |A(f, G) - f(G)| \leq O(g(\frac{\epsilon}{8\ln(1/\delta)})RS_f(\mathcal{H}_{2k})/\epsilon)] \geq 2/3$.

5 Restricted Sensitivity and $\mathcal{H}_k$

Now that we have constructed the machinery of restricted sensitivity, we compare the restricted sensitivity over $\mathcal{H}_k$ with smooth sensitivity for specific types of queries, in order to demonstrate the benefits of our approach. In a nutshell, restricted sensitivity offers a significant advantage over

smooth sensitivity whenever $k \ll n$. I.e., we show that there are queries f s.t. for some $G \in \mathcal{H}_k$ it holds that $RS_f(\mathcal{H}_k) \ll S_{f,\beta}(G)$.

We now define two types of queries. First, let us introduce some notation. A profile is a function that maps a vertex v in a social network (G, ℓ) to $[0, 1]$. Given a set of vertices $\{v_1, v_2, \ldots, v_t\}$, we denote by $G[v_1, v_2, \ldots, v_t]$ the social network derived by restricting G and ℓ to these t vertices. We use $G_v = G[\{v\} \cup \{w \,|(v, w) \in E(G)\}]$ to denote the social network derived by restricting G and ℓ to v and its neighbors. A *local* profile satisfies the constraint $p(v, (G, \ell)) = p(v, G_v)$.

Definition 19. *A* (local) profile query

$$f_p(G, \ell) = \sum\nolimits_{v \in V(G)} p\big(v, (G, \ell)\big)$$

sums the (local) profile p accross all nodes.

Local profile queries are a natural extension of predicates to social networks, which can be used to study many interesting properties of a social network like clustering coefficients [4, 17, 22], local bridges [10, 12] and 2-betweeness [11]. Further dissusion can be found in section C in the appendix. Claim 20 bounds the restricted sensitivity of a local profile query over $\mathcal{H}_k$ (e.g., in the vertex adjacency model a node v can at worst affect the local profiles of itself, its k old neighbors and its k new neighbors). A formal proof of Claim 20 is deferred to the appendix.

Claim 20. *For any local profile query f, we have that $RS_f(\mathcal{H}_k) \leq 2k + 1$ in the vertex adjacency model, and $RS_f(\mathcal{H}_k) \leq k + 1$ in the edge adjacency model.*

By contrast the smooth sensitivity of a local profile query may be as large as $O(n)$ even for graphs in $\mathcal{H}_k$. Consider the local profile query "how many people are friends with a spy?" The $n-1$-star graph G_1 in which a spy v is friends with everyone is adjacent to the empty graph $G_0 \in \mathcal{H}_k$. Therefore, any smooth upper bound $S_{f,\beta}$ must have $S_{f,\beta}(G) \geq n - 1$. It is also worth observing that the assumption $G \in \mathcal{H}_k$ does not necessarily shrink the range of possible answers to a local profile query f (e.g., there are graphs $G \in \mathcal{H}_k$ in which everyone is friends with a spy).

Subgraph queries allows us to ask questions such as "how many triplets of people are all friends when two of them are doctors and the other is a pop-singer?" or "how many paths of length 2 are there connecting a spy and a pop-singer over 40?" The average clustering coefficient of a graph can be computed from the number of triangles and 2-stars in a graph.

Definition 21. *A* subgraph counting *query $f = \langle H, \bar{p} \rangle$ is given by a connected graph H over t vertices and t predicates $p_1, p_2, \ldots, p_t$. Given a social network (G, ℓ), the answer to $f(G, \ell)$ is the size of the set*

$$\big\{v_1, v_2, \ldots, v_t : \; G[v_1, v_2, \ldots, v_t] = H \;\; and \;\; \forall i, \;\; \ell(v_i) \in p_i\big\}$$

The smooth sensitivity of a subgraph counting query may be as high as $O\left(n^{t-1}\right)$ in the vertex adjacency model. Let $f = \langle H, \bar{p} \rangle$ be a subgraph counting query where H is a t-star and each predicate p_i is identically true. Let G_1 be a n-star ($f(G_1) = \binom{n}{c-1}$). Then in the vertex adjacency model there is a neighboring graph G_2 with no edges ($f(G_2) = 0$). We have that $LS_f(G_2) \geq \binom{n}{t-1}$. Observe that $G_2 \in \mathcal{H}_k$. In the appendix we show that the smooth sensitivity of $f = \langle K_3, \bar{p} \rangle$ is *always* greater than n when each predicate p_i is identically true (see claim 25). By contrast Claim 22 bounds the restricted sensitivity of subgraph counting queries. The proof is deferred to the appendix.

Claim 22. *Let $f = \langle H, \bar{p} \rangle$ be subgraph counting query and let $t = |H|$ then $RS_f(\mathcal{H}_k) \leq tk^{t-1}$ in the edge adjacency model and in the vertex adjacency model.*

While the assumption $G \in \mathcal{H}_k$ may shrink the range of a subgraph counting query f, the restricted sensitivity of f will typically be much smaller than this reduced range. For example, if $f(G)$ counts the number of triangles in G then $f(G) \leq nk^2$ for any $G \in \mathcal{H}_k$, while $RS_f(\mathcal{H}_k) \leq 3k^2 \ll nk^2$.

6 Future Questions/Directions

Efficient Mappings: While we can show that there doesn't exist an efficiently computable $O(1)$-smooth projection $\mu : \mathcal{G} \to \mathcal{H}_k$, we don't know whether the construction of Claim 17 can be improved. Meaning, there could be a mapping $\mu : \mathcal{G} \to \bar{\mathcal{H}}$ for some $\bar{\mathcal{H}} \supset \mathcal{H}_k$, whether the solution itself, the set of vertices that dominate the removed edges, is smooth. In other words, Is there an efficiently computable mapping $\mu : \mathcal{G} \to \bar{\mathcal{H}} \subset \mathcal{H}_k$ which satisfies $|d(\mu(G_1), G_1) - d(\mu(G_2), G_2)| \leq c$ for some constant c? **Multiple Queries:** We primarily focus on improving the accuracy of a single query f. Could the notion of restricted sensitivity be used in conjunction with other mechanisms (e.g., BLR [6], Private Multiplicative Weights mechanism [13], etc.) to accurately answer an entire class of queries? **Alternate Hypotheses:** We focused on the specific hypothesis $\mathcal{H}_k$. What other natural hypthothesis could be used to restrict sensitivity in private data analysis? Given such a hypothesis $\mathcal{H}$ can we efficiently construct a query f_H with low global sensitivity or with low smooth sensitivity over datasets $D \in \mathcal{H}$?

References

[1] Facebook newsroom: Statistics. `http://newsroom.fb.com/Key-Facts`, 2012. Retrieved 6/29/2012.

[2] N. Alon, D. Moshkovitz, and S. Safra. Algorithmic construction of sets for k-restrictions. *ACM Transactions on Algorithms (TALG)*, 2(2):153–177, 2006.

[3] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th international conference on World Wide Web*, pages 181–190. ACM, 2007.

[4] P.S. Bearman and J. Moody. Suicide and friendships among american adolescents. *Journal Information*, 94(1), 2004.

[5] J. Blocki, A. Blum, A. Datta, and O. Sheffet. The johnson-lindenstrauss transform itself preserves differential privacy. In *Proceedings of the 53rd annual IEEE Symposium on Foundations of Computer Science*. IEEE, 2012.

[6] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th annual ACM Symposium on Theory of computing*, STOC '08, pages 609–618, New York, NY, USA, 2008. ACM.

[7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks*, 33(1):309–320, 2000.

[8] C. Dwork. Differential privacy. *Automata, languages and programming*, pages 1–12, 2006.

[9] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *In Proceedings of the 3rd Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[10] D. Easley and J. Kleinberg. *Networks, crowds, and markets*. Cambridge Univ Press, 2010.

[11] L.C. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1979.

[12] M.S. Granovetter. *Getting a job: A study of contacts and careers*. University of Chicago Press, 1995.

[13] M. Hardt and G.N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 61–70. IEEE, 2010.

[14] M. Hay, C. Li, G. Miklau, and D. Jensen. Accurate estimation of the degree distribution of private networks. In *ICDM*, pages 169–178, 2009.

[15] C. Jernigan and B.F.T. Mistree. Gaydar: Facebook friendships expose sexual orientation. *First Monday*, 14(10), 2009.

[16] V. Karwa, S. Raskhodnikova, A. Smith, and G. Yaroslavtsev. Private analysis of graph structure. *PVLDB*, 4(11):1146–1157, 2011.

[17] M.E.J. Newman. The structure and function of complex networks. *SIAM review*, pages 167–256, 2003.

[18] 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. ACM, 2007.

[19] V. Rastogi, M. Hay, G. Miklau, and D. Suciu. Relationship privacy: output perturbation for queries with joins. In *Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 107–116. ACM, 2009.

[20] R. Raz and S. Safra. A sub-constant error-probability low-degree test, and a sub-constant error-probability pcp characterization of np. In *Proceedings of the twenty-ninth annual ACM Symposium on Theory of Computing*, pages 475–484. ACM, 1997.

[21] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the facebook social graph. *Arxiv preprint arXiv:1111.4503*, 2011.

[22] D.J. Watts and S.H. Strogatz. Collective dynamics of small-worldnetworks. *Nature*, 393(6684):440–442, 1998.

A Missing Proofs

Reminder of Claim 13. *In the edge-adjacency model, there exists an efficient way to compute a 3-smooth projection μ to $\mathcal{H}_k$.*

Proof of Claim 13. We construct our smooth-projection μ by first fixing a canonical ordering over all possible edges. Let $e_1^v, ..., e_t^v$ denote the edges incident to v in canonical order. For each edge $e = \{u, v\}$ we delete e if and only if (i) $e = e_j^v$ for $j > k$ or (ii) $e = e_j^u$ for $j > k$ (Intuitively for each v with $\deg(v) \geq k$ we keep this first k edges incident to v and flag the other edges for deletion). If $G \in \mathcal{H}_k$ then no edges are deleted, so $\mu(G) = G$. Suppose that G_1, G_2 are neighbors differing on one edge $e = \{x, y\}$ (wlog, say that e is in G_1). Observe that for every $v \neq x, y$, the same set of edges incident to v will be deleted from both G_1 and G_2. In fact, if $\mu(G_1)$ does not contain e then $\mu(G_1) = \mu(G_2)$. Otherwise, if e is not deleted we may assume then there may be at most one edge e_x (incident to x) and at most one edge e_y (incident to y) that were deleted from $\mu(G_1)$ but not from $\mu(G_2)$. Hence, $d(\mu(G_1), \mu(G_2)) \leq 3$. □

Reminder of Lemma 16. *Fix $\bar{\mathcal{H}} \supset \mathcal{H}$ and let $\mu : \mathcal{D} \to \bar{\mathcal{H}}$ be a projection of $\mathcal{H}$. Let $\hat{d} : \mathcal{D} \to \mathbb{R}$ be an efficiently computable c-smooth distance estimator. Then for every query f, we can define the composition $f_{\mathcal{H}} = f \circ \mu$ and define the function*

$$S_{f_{\mathcal{H}},\beta}(D) = \max_{d \in \mathbb{Z}, d \geq \hat{d}(D)} e^{\left(-\frac{\beta}{c}\left(d - \hat{d}(D)\right)\right)} (2d + c + 1) \cdot RS_f\left(\bar{\mathcal{H}}\right)$$

Then $S_{f_{\mathcal{H}},\beta}$ is an efficiently computable β-smooth upper bound on the local sensitivity of $f_{\mathcal{H}}$. Furthermore, define g as the function $g(x) = \begin{cases} 2\frac{1}{x}e^{-1+\frac{c+1}{2}x}, & 0 \leq x \leq \frac{2}{c+1} \\ c+1, & x > \frac{2}{c+1} \end{cases}$. Then for every D it holds that

$$S_{f,\beta}(D) \leq \exp(\tfrac{\beta}{c}\hat{d}(D)) \cdot g(\beta/c) RS_f(\bar{\mathcal{H}})$$

Proof of Lemma 16. First, we show that indeed $S_{f_{\mathcal{H}},\beta}$ is an upper bound on the local sensitivity of $f_{\mathcal{H}}$. Fix any $D \in \mathcal{D}$ and indeed

$$\begin{aligned}
LS_{f_{\mathcal{H}}}(D) &= \max_{D' \sim D} \left| f_{\mathcal{H}}(D) - f_{\mathcal{H}}\left(D'\right) \right| \\
&= \max_{D' \sim D} \left| f(\mu(D)) - f(\mu(D')) \right| \\
&\leq \max_{D' \sim D} RS_f(\bar{\mathcal{H}}) \cdot d(\mu(D), \mu(D')) \\
&\leq \max_{D' \sim D} RS_f(\bar{\mathcal{H}}) \cdot \\
&\qquad \left(d(D, \mu(D)) + d(D, D') + d(D', \mu(D')) \right) \\
&\leq RS_f(\bar{\mathcal{H}}) \cdot \left(\hat{d}(D) + 1 + \max_{D' \sim D} \hat{d}(D') \right) \\
&\leq RS_f(\bar{\mathcal{H}}) \cdot \left(2\hat{d}(D) + c + 1 \right) \\
&\leq \max_{d \geq \hat{d}(D)} e^{-\beta c\left(d - \hat{d}(D)\right)} (2d + c + 1) RS_f\left(\bar{\mathcal{H}}\right) \\
&= S_{f_{\mathcal{H}},\beta}(D) \ .
\end{aligned}$$

Next we prove that $S_{f_{\mathcal{H}},\beta}$ is β-smooth. Let D_1 and D_2 be two neighboring databases, and wlog assume $\hat{d}(D_2) \leq \hat{d}(D_1)$. Then

$$\begin{aligned}
&\frac{S_{f_{\mathcal{H}},\beta}(D_1)}{S_{f_{\mathcal{H}},\beta}(D_2)} \\
&= \frac{\max_{d \geq \hat{d}(D_1)} e^{\left(-\frac{\beta}{c}\left(d-\hat{d}(D_1)\right)\right)} \left(2d+c+1\right) RS_f\left(\bar{\mathcal{H}}\right)}{\max_{d \geq \hat{d}(D_2)} e^{\left(-\frac{\beta}{c}\left(d-\hat{d}(D_2)\right)\right)} \left(2d+c+1\right) RS_f\left(\bar{\mathcal{H}}\right)}
\end{aligned}$$

Let d_0 be the value of d on which the maximum of numerator is obtained. Then

$$\begin{aligned}
&\frac{S_{f_{\mathcal{H}},\beta}(D_1)}{S_{f_{\mathcal{H}},\beta}(D_2)} \\
&= \frac{\exp\left(-\frac{\beta}{c}\left(d_0 - \hat{d}\left(D_1\right)\right)\right)\left(2d_0+c+1\right) RS_f\left(\bar{\mathcal{H}}\right)}{\max_{d \geq \hat{d}(D_2)} \exp\left(-\frac{\beta}{c}\left(d - \hat{d}\left(D_2\right)\right)\right)\left(2d+c+1\right) RS_f\left(\bar{\mathcal{H}}\right)} \\
&\leq \frac{\exp\left(-\frac{\beta}{c}\left(d_0 - \hat{d}\left(D_1\right)\right)\right)\left(2d_0+c+1\right) RS_f\left(\bar{\mathcal{H}}\right)}{\exp\left(-\frac{\beta}{c}\left(d_0 - \hat{d}\left(D_2\right)\right)\right)\left(2d_0+c+1\right) RS_f\left(\bar{\mathcal{H}}\right)} \\
&= \exp\left(-\tfrac{\beta}{c}(\hat{d}(D_2) - \hat{d}(D_1))\right) \leq \exp(\beta)
\end{aligned}$$

where the last inequality uses the smoothness property, i.e. that $\hat{d}(D_2) - \hat{d}(D_1) \geq -c$.

Finally, we wish to prove the global upper bound on $S_{f_{\mathcal{H}},\beta}$, i.e., that for every $D \in \mathcal{D}$

$$S_{f_{\mathcal{H}},\beta}\left(D\right) \leq \exp(\tfrac{\beta}{c}\hat{d}(D)) \cdot g\left(c/\beta\right) RS_f\left(\bar{\mathcal{H}}\right) \ .$$

Fix D and define $h(x) = \exp\left(-\frac{\beta}{c}x\right)(2x+c+1)$, so that $S_{f_{\mathcal{H}},\beta} = \exp\left(\frac{\beta}{c}\hat{d}(D)\right) RS_f(\bar{\mathcal{H}}) \cdot \max_{d \geq d_0} h(d)$. Taking the derivative of h we have

$$h'(x) = e^{-\frac{\beta}{c}x}\left(-2x\tfrac{\beta}{c} - \beta - \tfrac{\beta}{c} + 2\right)$$

which means that $h(x)$ is maximized at $x_0 = \frac{c}{\beta} - \frac{c+1}{2}$. In the case that $x_0 < 0$ (i.e. for $\beta/c > \frac{2}{c+1}$) we can upper bound the function $h(x)$ with $h(0) = c+1$ for every $x \geq 0$. Otherwise, we have that $h(x) \leq h(x_0)$ for every $x \geq 0$, and indeed $h(x_0) = 2\frac{c}{\beta}e^{-1+\frac{\beta}{c}\cdot\frac{c+1}{2}} = g(\beta/c)$.

To conclude the proof, observe that computing $S_{f_{\mathcal{H}},\beta}\left(D\right)$ is just a simple optimization once $\hat{d}\left(D\right)$ is known, much like the derivation done above. So since $\hat{d}$ is efficiently computable, we have that $S_{f_{\mathcal{H}},\beta}$ is efficiently computable. □

Reminder of Claim 17. *In the vertex-adjacency model, there exists a projection $\mu : \mathcal{G} \to \mathcal{H}_{2k}$ and a 4-smooth distance estimator $\hat{d}$, both of which are efficiently computable.*

Proof of Claim 17. We first prove that μ is a projection mapping every graph to a graph in $\mathcal{H}_{2k}$. Suppose that some $v \in G$ has degree $\geq 2k$, then clearly $x^*_v \leq 1/4$, for otherwise we would have removed all of the edges touching v. Observe that every edge we keep has $w^*_{u,v} \geq 1-1/4-1/4 = 1/2$.

Consequently, we can have at most $2k$ edges with $w_{u,v} \geq \frac{1}{2}$ because of the constraint $\sum_u w_{u,v} \leq k$. So there are at most $2k$ edges incident to v in $\mu(G)$.

Now, let us prove that $\hat{d}$ satisfies all of the requirements of a 4-smooth distance estimator. First, if $G \in \mathcal{H}_k$ then the optimal solution of the LP is the all zero vector, so $\hat{d}(G) = 0$ for all graphs of max-degree $\leq k$. Secondly, observe that in the process of computing $\mu(G)$, every edge that is removed from G can be "charged" to a vertex v with $x^*_v \geq 1/4$. If follows that

$$d(G, \mu(G)) \leq \sum_{v: x^*_v \geq 1/4} 1 \quad \leq \sum_{v: x^*_v \geq 1/4} 4x^*_v \quad \leq 4 \sum_v x^*_v \quad = \hat{d}(G) \ .$$

Lastly, fix any neighboring $G_1, G_2 \in \mathcal{G}$, and let v be the vertex whose edges differ in G_1 and G_2. Clearly, if $\bar{x}^*$ is a solution for $LP(G_1)$, then we set $y_v = 1$ for $i = 1...d$ and $y_v = x^*_v$ otherwise. Now $\vec{y}$ is a feasible (not necessarily optimal) solution to $LP(G_2)$. It is simple to infer that

$$\begin{aligned} \hat{d}(G_2) - \hat{d}(G_1) &= \hat{d}(G_2) - 4 \sum_u x^*_u \\ &\leq 4 \sum_u y_u - 4 \sum_u x^*_u \leq 4 \sum_u |(y_u - x^*_u)| \\ &= 4 |y_v - x^*_v| \leq 4 \end{aligned}$$

□

Reminder of Claim 20. *For any local profile query f, we have that $RS_f(\mathcal{H}_k) \leq 2k+1$ in the vertex adjacency model and $RS_f(\mathcal{H}_k) \leq k+1$ in the edge adjacency model.*

Proof of Claim 20. Consider a local profile query f_p.

(Label change) Let $G_1, G_2 \in \mathcal{H}$ be two graphs with the same exact edge set, but with labeling functions ℓ_1, ℓ_2 that are different on a single vertex. Let v be the vertex whose label differs on G_1 and G_1, and let N_v denote the set of its (at most k) neighbors. Then for every $u \notin \{v\} \cup N_v$ we have that $p(u, (G_1, \ell_1)) = p(u, (G_2, \ell_2))$. Hence, $|f_p(G_1) - f_p(G_2)| \leq |\{v\} \cup N_v| \leq k+1$.

(Vertex Adjacency) Let $G_1, G_2 \in \mathcal{H}$ be any two neighboring labeled graphs such that $G_1 - v = G_2 - v$. Let N^1_v (resp. N^2_v) denote the neighborhood of v then for any $y \notin N^1_v \cup N^2_v$ we have that $p(y, (G_1, \ell_1)) = p(y, (G_2, \ell_2))$. Hence, $|f_p(G_1) - f_p(G_2)| \leq \left|N^1_v \cup N^2_v \cup \{v\}\right| \leq 2k+1$.

(Edge Adjacency) Let $G_1, G_2 \in \mathcal{H}$ be any two neighboring labeled graphs. Wlog, there is an edge $e = \{u, v\}$ such that $E(G_1) = E(G_2) \cup \{e\}$. In order to have a vertex y s.t. $p(y, (G_1, \ell_1)) \neq p(y, (G_2, \ell_2))$ we need that the edge e appears in graph we get by restricting the social network to set of y and its neighbors. It follows that the only vertices whose local profile can change are in the union $\{u, v\} \cup \left(N_u \cap N_v\right)$. Hence, $|f(G_1) - f(G_2)| \leq |\{u\} \cup \{v\}| + |N_u \setminus \{v\}| \leq 2 + k - 1 = k+1$.
□

Reminder of Claim 22. *Let $f = \langle H, \bar{p} \rangle$ be subgraph counting query and let $t = |H|$ then*

$$RS_f(\mathcal{H}_k) \leq tk^{t-1}$$

in the edge adjacency model and in the vertex adjacency model.

Proof of Claim 22. (Sketch) Let $G_1, G_2 \in \mathcal{H}_k$ be neighbors and let v be a vertex such that $G_1 - v = G_2 - v$, and let N_i denote the neighbors of v in G_i. Any copy of H which occurs in G_1 but not in G_2 must contain v. Because H is connected we can bound the number of G_1 copies of H. We can start with v, and we pick one of the t vertices of H to be mapped to v. Denote this vertex as

v_0. Now, we proceed inductively. We pick a vertex $v \in H$, connected to the set $\{v_0, v_1, \ldots, v_{i-1}\}$. The vertex v_i must be assigned to a vertex in G which is incident to some specific vertex of the i vertices that we already mapped. Because we have bounded degree, then there are at most k options from which to choose v_i. We obtain the bound: $f(G_1) - f(G_2) \leq t \prod_{i=1}^{t-1} k = tk^{t-1}$. □

B Additional Claims

Claim 23. *(Privacy wrt Vertex Adjacency) Unless $P = NP$ there is no efficiently computable mapping $\mu : \mathcal{G} \to \mathcal{H}_k$ such that*

1. $\forall G \in \mathcal{H}_k, \mu(G) = G$.
2. $\forall G \in \mathcal{G}, d(G, \mu(G)) \leq O(\ln(k)\, d(G, \mathcal{H}_k))$.

Proof. (Sketch) Our reduction is from the minimum set cover problem. It is NP-hard to approximate the minimum set cover problem to a factor better than $O(\log n)$ [2, 20]. Given a set cover instance with sets $S_1, ..., S_m$ and universe $U = \{x_1, ..., x_n\}$ we set $m_i = |\{j : x_i \in S_j|$ and $k = n+1$. We construct our labeled graph G as follows:

1. Add a node for each S_i.
2. Add a node for each x_j.
3. Add the edge $\{x_j, S_i\}$ if and only if $x_j \in S_i$.
4. For each x_i, create $k + 1 - m_i$ fresh nodes $y_1, ..., y_{k-m_i}$ and add each edge $\{y_j, x_i\}$.

Intuitively each node x_j has $k+1$ incident edges. By deleting all of the edges incident to the node S_i we can fix all of the nodes $x \in S_i$. Hence, $d(G, (\mathcal{H}_k))$ corresponds exactly to the size of the minimum set cover. □

Claim 24. *(Privacy wrt Vertex Adjacency) There is an efficiently computable projection $\mu : \mathcal{G} \to \mathcal{H}_k$ such that for every $G \in \mathcal{G}$ it holds that $d(G, \mu(G)) \leq \left(\ln\left(2d^2 + kd\right)\right) d(G, \mathcal{H}_k)$,*

Proof. (Sketch) We use a greedy algorithm to create μ. Define the potential of a graph G as follows

$$\phi(G) = \sum_{v \in G : \deg(v) \geq k} (\deg(v) - k) \ .$$

Our algorithm μ starts by guessing a value d for $d(G, \mathcal{H}_k)$ and deleting any vertex with degree $\geq k + d + 1$ (these vertices *must* be deleted because the degree will be at least $k+1$ after deleting d other vertices). Then μ repeatedly picks the vertex v with the highest potential and eliminates all incident edges, where the potential of a vertex v is $\phi(G) - \phi(G - v)$. Let ϕ_i denote the potential after round i (ϕ_0 is the potential after deleting vertices with degree $\geq k + d + 1$). Observe that $\phi_0 \leq 2d^2 + kd$ if d is correct because (i) there are d vertices we can delete to drop the potential to 0 and (2) deleting a single vertex v decreases the potential by at most $\deg(v) + \deg(v) - k \leq 2d + k$. Also observe that in any round there always exists some vertex whose removal decreases the potential by at least $(1 - 1/d)\, \phi_i$ so we have $\phi_i \leq \phi_{i-1}\left(1 - \frac{1}{d}\right)$. Once $i \geq d \ln\left(2d^2 + dk + d\right)$ we have $\phi_i \leq 1$. □

The reduction in Claim 24 might be used to produce a function $f_{\mathcal{H}}(G) = \mu(f(G))$ with low smooth-sensitivity over the nice graphs $\mathcal{H}_k$. Unfortunately, we don't know of any efficient algorithm to compute the smooth upper bound for such $f_{\mathcal{H}}$.

Claim 25. *Let $f = \langle K_3, \bar{p} \rangle$ be a subgraph counting query with predicates p_i that are identically true. In the vertex adjacency model for any β smooth upper bound on the local sensitivity of f and* any *graph G we have*

$$S^*_{f^P,\beta}(G) \geq \exp(-2\beta)(n-2) \ .$$

Proof. Let G be given. Pick $v_1, v_2 \in V(G)$ and let G_1 be obtained from G by adding all possible edges incident to v_1 and let G_2 be obtained from G_1 by deleting all edges incident to v_2. Finally, let G_3 be obtained from G_2 by adding all possible edges incident to v_2. Now the local sensitivity of f at G_2 is at least $n-2$,

$$\begin{aligned} LS_f(G_2) &= \max_{G':d(G_2,G')=1} \left| f(G_2) - f\left(G'\right) \right| \\ &\geq f(G_3) - f(G_2) \geq n-2 \end{aligned}$$

Plugging this lower bound into the definition of β smooth sensitivity we obtain the required result:

$$S^*_{f^P,\beta}(G) \geq e^{-\beta d(G,G_2)} LS_{f^P}(G_2) \geq e^{-2\beta}(n-2)$$

□

C Local Profile Queries

Local profile queries are a natural extension of predicates to social networks, which can be used to study many interesting properties of a social network like clustering coefficients, local bridges and 2-betweeness). The clustering coefficient $c(v)$ [17, 22] of a node v (e.g., the probability that two randomly selected friends of v are friends with each other)

has been used to identify teenage girls who are more likely to consider suicide [4]. One explanation, is that it becomes an inherent source of stress if a person has many friends who are not friends with each other [10]. Observe that $c(v)$ *is* a local profile query. An edge $\{v, w\}$ is a local bridge if its endpoints have no friends in common. A local profile could score a vertex v based on the number local bridges incident to v. A marketing agency may be interested in identifying nodes that are incident to many local bridges because local bridges "provide their endpoints with access to parts of the network - and hence sources of information - that they would otherwise be far away from [10]." For example, a 1995 study showed that the best job leads often come from aquaintances rather than close friends [12]. 2-betweeness (a variant of betweeness [11]) measures the centrality of a node. We say that the 2-betweeness of a vertex v is the probability that the a randomly chosen shortest path between two randomly chosen neighbors of v $x, y \in G_v$ goes through v.