Title:	Differentially Private Analysis of Graphs
Name:	Sofya Raskhodnikova, Adam Smith
Affil./Addr.	Pennsylvania State University
Keywords:	Graphs, privacy, subgraph counts, degree distribution
SumOriWork:	2013; Blum, Blocki, Datta, Sheffet 2013; Kasiviswanatan, Nissim, Raskhodnikova, Smith 2013; Chen, Zhou 2015; Raskhodnikova, Smith 2015; Borgs, Chayes, Smith

Differentially Private Analysis of Graphs

Sofya Raskhodnikova, Adam Smith
Pennsylvania State University

Years aud Authors of Summarized Original Work

2013; Blum, Blocki, Datta, Sheffet
2013; Kasiviswanatan, Nissim, Raskhodnikova, Smith
2013; Chen, Zhou
2015; Raskhodnikova, Smith
2015; Borgs, Chayes, Smith

Keywords

Graphs, privacy, subgraph counts, degree distribution

Problem Definition

Many datasets can be represented by graphs, where nodes correspond to individuals and edges capture relationships between them. On one hand, such datasets contain potentially sensitive information about individuals; on the other hand, there are significant public benefits from allowing access to *aggregate* information about the data. Thus, analysts working with such graphs are faced with two conflicting goals: protecting privacy of individuals and publishing accurate aggregate statistics. This article describes algorithms for releasing accurate graph statistics, while preserving a rigorous notion of privacy, called *differential privacy*.

Differential privacy was introduced by Dwork et al. [6]. It puts a restriction on the algorithm that processes sensitive data and publishes the output. Intuitively, differential privacy requires that, for every individual, the output distribution of the algorithm is roughly the same whether or not this individual's data is present in the dataset. Next, we give a formal definition of differential privacy, specialized to datasets represented by graphs.

Two graphs are called *neighbors* if one can be obtained from the other by removing a node and its adjacent edges. Given a parameter $\epsilon > 0$, an algorithm A is *ϵ-node differentially private* if for all neighbor graphs G and G' and for all sets S of possible outputs produced by A:

$$\Pr[A(G) \subseteq S] \leq e^{\epsilon} \cdot \Pr[A(G) \subseteq S].$$

This variant of differential privacy is called *node-differential privacy* because neighbor graphs are defined with respect to node removals. Analogously, we can define *edge differential privacy* by letting graphs be neighbors if they differ in exactly one edge. Intuitively, edge differential privacy protects edges (which represent connections between people), whereas node-differential privacy protects nodes together with their adjacent edges (that is, all information pertaining to individuals). Node-differential privacy is a stronger privacy definition, but it is much harder to attain because it requires the output distribution of the algorithm to hide much larger differences in the input graph.

We would like to design differentially private algorithms (preferably, node-differentially private) that compute accurate graph statistics on a large family of realistic graphs. Typically, graphs that contain sensitive information, such as friendships, sexual relationships, and communication patterns, are sparse. Some examples of graph statistics we would like to compute on these graphs are the number of edges, small subgraph counts, and the degree distribution.

Most work on the topic considers an analyst who wants to evaluate a real-valued function f on the private input graph G (for example, the number of triangles or the number of connected components in G). The goal is to release as good an approximation as possible to the true value $f(G)$. Differentially private algorithms must be randomized, so we try to minimize the expectation of the random variable $error_A(G) = |A(G) - f(G)|$. We will also discuss work on algorithms that release higher-dimensional summaries (that is, output a real vector).

Bibliographical notes Edge privacy was first studied by Nissim et al. [16], and the distinction between node and edge privacy was laid out by Hay et al. [9]. Edge differentially private algorithms for a variety of tasks have been widely investigated. Examples include subgraph counts, degree distributions, and parameters of generative statistical models. Gehrke et al. [7] investigated a notion whose strength lies between edge and node privacy: node privacy for bounded-degree graphs. (The focus of their work is a generalization of differential privacy, called *zero-knowledge privacy*.)

Until recently, no node-differentially private algorithms (where privacy guarantees hold with respect to all graphs) were known that compute accurate graph statistics on realistic (namely, sparse) graphs. The first such algorithms were designed independently by Blocki et al. [3], Kasiviswanathan et al. [11] and Chen and Zhou [5]. Those algorithms look at releasing one real-valued statistic at a time. Two more recent works focus on higher-dimensional node-private releases: Raskhodnikova and Smith [17] and Borgs et al. [4].

This encyclopedia entry focuses on node-differentially private algorithms, since these offer the strongest privacy guarantees. Progress, however, continues on edge-private algorithms; see Lin and Kifer [13], Karwa and Slavkovic [10], Lu and Miklau [14] and Zhang et al. [18] for recent results.

Key Results

The main difficulty in the design of node-private algorithms is that techniques based on *local sensitivity* of a function (which are the basis of the best edge-private algorithms) yield node-private algorithms whose error on "typical" inputs swamps the statistic that one wants to release. The local sensitivity of a function f is a discrete analogue of the derivative of f—it measures how much the value of f can change when the input graph is replaced with its neighbor. On sparse graphs, the local sensitivity can be larger than the value of the function. Any method whose error is proportional to the local sensitivity will have large relative error.

Focus on a "preferred subset" To get around the challenge of high local sensitivity, two works [3; 11] independently designed algorithms that are given a set S of "nice" graphs that hopefully contains G (for example, graphs with an upper bound on the maximum degree). These algorithms are private on *all* graphs and return an accurate answer *on graphs in* S. What makes this approach work is that S is selected so that the sensitivity of f is small when restricted to inputs in S.

Let $\mathbb{G}$ denote the set of all labeled, undirected graphs. We will call $S \subseteq \mathbb{G}$ the "preferred" subset. Define the *Lipschitz constant* (also called the *restricted sensitivity*) of f on S to be

$$\Delta_f(S) = \sup_{G,G' \in S} \frac{\|f(G') - f(G)\|_1}{d_{node}(G, G')},$$

where d_{node} is the node distance between two graphs—the number of vertex insertions and deletions needed to go from G to G'. Blocki et al. [3] and Kasiviswanathan et al. [11] give methods for adding noise proportional to the Lipschitz constant of f on S.

Theorem 1 ([3; 11]). *For every $S \subseteq \mathbb{G}$, function $f : S \to \mathbb{R}$, and $\epsilon > 0$, there exists an algorithm A_S that is ϵ-differentially private (for all inputs) and such that, for all $G \in S$,*

$$\mathbb{E}\,|A_S(G) - f(G)| = O(\Delta_S(f)/\epsilon^2).$$

Moreover, for $S = \mathbb{G}_{\mathbb{D}}$ (the set of D-bounded graphs), the running time of A is the running time for one evaluation of f plus a fixed polynomial in the size of G.

The same works [3; 11] also give generic reductions showing that given any algorithm that is ϵ-differentially private when restricted to graphs in S, one can design an algorithm A that has similar behavior on graphs in S but is ϵ'-differentially private for *all* inputs, for ϵ' not too much larger than ϵ.

"Down" Sensitivity Rather than focusing on a single "nice" subset, some works [5; 17] sought to add noise proportional to a quantity related to, but usually much smaller than, the local sensitivity.

Define the *down sensitivity* (called *empirical global sensitivity* when first defined by Chen and Zhou [5]) of f at a graph G to be the Lipschitz constant of f *when restricted to the set of induced subgraphs of* G. Specifically, we write $G \preceq H$ to denote that G is an induced subgraph of H (that is, G can be obtained by deleting a set of vertices from H) and define the down sensitivity to be

$$DS_f(G) = \max_{H,H' \text{neighbors}, H \preceq H' \preceq G} |f(G') - f(G)|.$$

By carefully (and privately) selecting the "preferred" subset based on the input, one can add noise essentially proportional to the down sensitivity.

Theorem 2 ([17]). *For every monotone function $f : \mathbb{G} \to \mathbb{R}$ and $\epsilon > 0$, there is an algorithm A_f that is ϵ-differentially private and such that, for all $G \in \mathbb{G}$,*

$$\mathbb{E}\,|A_f(G) - f(G)| = \frac{DS_f(G)+1}{\epsilon} \cdot O(\log\log \max_{G'} DS_f(G')).$$

Moreover, A_f can be made efficient when f is a generalized linear query (a class that includes counting occurrences of a fixed subgraph).

The down sensitivity is low for many commonly studied statistics in graphs that satisfy α-*decay*, a condition on the degree distribution that is satisfied by known generative models (including those that generate "scale-free"). (See [11] for a definition of α-decay.)

Lipschitz Extensions and Higher-dimensional Releases The main technical tool in the down-sensitivity-based results [5; 17] is the construction of *efficient* (that is, polynomial-time computable) *Lipschitz extensions* of the function f from subsets S of graphs to the space of all graphs. Kasiviswanathan et al. [11] and Chen and Zhou [5] give efficient Lipschitz extensions of several useful functions (including graph counts) that return a single real value. Raskhodnikova and Smith [17] give efficient Lipschitz extensions of higher-dimensional functions, namely, the degree distribution and adjacency matrix of a graph.

Borgs et al. [4] use the Lipschitz extension technique together with the *exponential mechanism* to provide the first node-differentially private algorithms for fitting high-dimensional statistical models to a given graph (specifically, they consider *stochastic block models* and generalizations thereof).

Applications

The algorithms discussed above address a real problem: datasets containing sensitive information about relationships among a collection of individuals are often valuable sources of information, but publishing useful summaries about such data without leaking individual information is difficult. Even when the graphs are "anonymized" by removing all obviously identifying information, such as names, addresses, birthdays, and zip codes, they present a privacy risk. For example, [1; 15] give de-anonymization attacks based only on unlabeled links. Node-differentially private algorithms offer a principled method for releasing information about a network while providing rigorous privacy guarantees (though some authors argue that even stronger notions may be needed [12; 7]).

Open Problems

Gupta et al. [8]; Blocki et al. [2] give edge differentially private algorithms for releasing a data structure that approximates the sizes of all cuts in the input graph in the following sense: for any cut, with high probability, the estimated cut size is accurate (the first reference gives weaker approximation guarantees with a stronger quantifier order: with high probability, all cut sizes are accurate). It is open whether a node-differentially private algorithm can obtain similar results.

For datasets that do not contain information about relationships, but only contain personal attributes that come from a relatively small set, differentially private algorithms can output a large number of statistics at once (see "Query Release via

Online Learning" and "Geometric Approaches to Answering Queries" cross-referenced below). It is open how to do achieve similar results for graph statistics, even with edge differential privacy.

Finally, all algorithms we discussed release numerical graph statistics. The subject of differentially private synthetic graphs is largely unexplored. See [10; 13] for initial results.

Cross-References

Beyond Worst Case Sensitivity in Private Data Analysis
Geometric Approaches to Answering Queries
Privacy and Game Theory
Private Spectral Analysis
Query Release via Online Learning

Acknowledgements

The authors were supported in part by NSF award IIS-1447700, Boston University's Hariri Institute for Computing and Center for Reliable Information Systems and Cyber Security and, while visiting the Harvard Center for Research on Computation & Society, by a Simons Investigator grant to Salil Vadhan.

Recommended Reading

1. Backstrom L, Dwork C, Kleinberg J (2007) Wherefore art thou r3579x? anonymized social networks, hidden patterns, and structural steganography. In: Proc. 16th Intl. World Wide Web Conference, pp 181–190
2. Blocki J, Blum A, Datta A, Sheffet O (2012) The Johnson-Lindenstrauss transform itself preserves differential privacy. In: 53rd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2012, New Brunswick, NJ, USA, October 20-23, 2012, IEEE Computer Society, pp 410–419, DOI 10.1109/FOCS.2012.67, URL http://dx.doi.org/10.1109/FOCS.2012.67
3. Blocki J, Blum A, Datta A, Sheffet O (2013) Differentially private data analysis of social networks via restricted sensitivity. In: Innovations in Theoretical Computer Science (ITCS), pp 87–96
4. Borgs C, Chayes JT, Smith A (2015) Private graphon estimation for sparse graphs. arXiv:150606162 [mathST]
5. Chen S, Zhou S (2013) Recursive mechanism: towards node differential privacy and unrestricted joins. In: ACM SIGMOD International Conference on Management of Data, pp 653–664
6. Dwork C, McSherry F, Nissim K, Smith A (2006) Calibrating noise to sensitivity in private data analysis. In: Halevi S, Rabin T (eds) TCC, vol 3876, pp 265–284
7. Gehrke J, Lui E, Pass R (2011) Towards privacy for social networks: A zero-knowledge based definition of privacy. In: Ishai Y (ed) TCC, Springer, Lecture Notes in Computer Science, vol 6597, pp 432–449
8. Gupta A, Roth A, Ullman J (2012) Iterative constructions and private data release. In: TCC
9. Hay M, Li C, Miklau G, Jensen D (2009) Accurate estimation of the degree distribution of private networks. In: Int. Conf. Data Mining (ICDM), pp 169–178
10. Karwa V, Slavkovic A (2014) Inference using noisy degrees: Differentially private -model and synthetic graphs. statME arXiv:1205.4697v3 [stat.ME]
11. Kasiviswanathan SP, Nissim K, Raskhodnikova S, Smith A (2013) Analyzing graphs with node-differential privacy. In: Theory of Cryptography Conference (TCC), pp 457–476
12. Kifer D, Machanavajjhala A (2011) No free lunch in data privacy. In: Sellis TK, Miller RJ, Kementsietsidis A, Velegrakis Y (eds) SIGMOD Conference, ACM, pp 193–204
13. Lin BR, Kifer D (2013) Information preservation in statistical privacy and Bayesian estimation of unattributed histograms. In: ACM SIGMOD International Conference on Management of Data, pp 677–688

14. Lu W, Miklau G (2014) Exponential random graph estimation under differential privacy. In: 20th ACM SIGKDD International Conference on Knowledge discovery and data mining, pp 921–930
15. Narayanan A, Shmatikov V (2009) De-anonymizing social networks. In: IEEE Symp. Security and Privacy, pp 173–187
16. Nissim K, Raskhodnikova S, Smith A (2007) Smooth sensitivity and sampling in private data analysis. In: Symp. Theory of Computing (STOC), pp 75–84, full paper on authors' web sites.
17. Raskhodnikova S, Smith A (2015) High-dimensional Lipschitz extensions and node-private analysis of network data. arXiv:150407912
18. Zhang J, Cormode G, Procopiuc CM, Srivastava D, Xiao X (2015) Private release of graph statistics using ladder functions. In: ACM SIGMOD International Conference on Management of Data, pp 731–745