Private Graphon Estimation for Sparse Graphs

Christian Borgs   Jennifer T. Chayes   Adam Smith

Microsoft Research, New England   Pennsylvania State University

Privacy for Network Data

Many types of data can be represented as graphs where
- nodes correspond to individuals
- edges capture relationships
  - "Friendships" in online social network
  - Financial transactions
  - Email communication
  - Health networks (of doctors and patients)
  - Romantic relationships

image source: http://community.expressor-software.com/blogs/mtarallo/36-extracting-data-facebook-social-graph-expressor-tutorial.html

Image source: Bearman, Moody, Stovel, American J. Sociology (2004)

Such graphs contain potentially sensitive information.

This paper: Algorithms for learning complex, nonparametric generative graph models subject to strong, node-level privacy guarantees

Goal: Estimation of Graphons

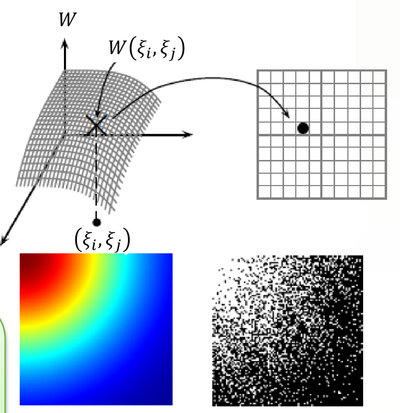
Graphons provide a complex generative model for graphs
- Extremely general
- Includes stochastic block models as special cases
- Deep connections to limits of graph sequences (e.g., [BCLSV'08, '12])

A graphon is a function $W: [0,1]^2 \to \mathbb{R}^+$. Typical examples
- k-block graphons (constant on the cells of a $k \times k$ grid)
- Smooth graphons (e.g., Hölder continuous)

"W-random", (a.k.a. "latent position") graphs

A graphon W defines a family of distributions on graphs:
- Given a size n and target density $\rho \in \mathbb{R}^+$, define $G_n(\rho W)$:
  - Select $\xi_1, \dots, \xi_n \in [0,1]$ uniformly, i.i.d.
  - Form matrix $H \in [0,1]^{n\times n}$, where $H_{ij} = \min(1, \rho\, W(\xi_i, \xi_j))$
  - For each $i, j \in [n]$ add edge (i,j) to G with prob H_{ij} (independently)

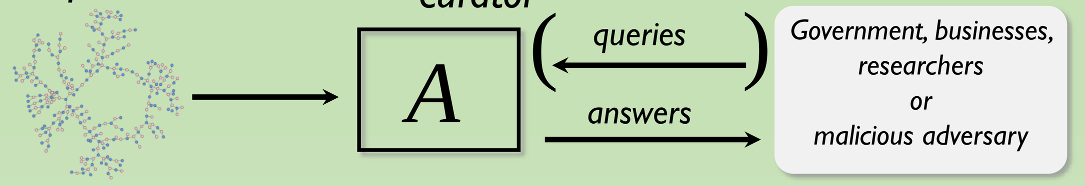
W-random graphs provide a rich, nonparametric model for graphs
- The set [0,1] can model any set of "vertex types"
- Captures all exchangeable graph distributions in the limit
  - Similar the role of IID distributions in de Finetti's theorem [DJ'09]



Goal: given $G \sim G_n(\rho W)$, where $\rho \in (0,1)$ and $\|W\|_1 = 1$, estimate ρ and W

image source http://scholar.harvard.edu/stanleychan/research

Differential Privacy for Graphs



- G and G' are neighbors if they differ in one person's data
- Neighboring datasets induce close distributions on outputs

Definition [DMNS'06]: Randomized algorithm A is ϵ-differentially private if, for all data sets G and G' that "differ in one element" and for all events S,

$$\Pr[A(G) \in S] \le e^{\epsilon} \cdot \Pr[A(G') \in S].$$

Edge- vs node-level privacy [HLMJ'09]

- **Edge** differential privacy

Two graphs are neighbors if they differ in *one edge*.

- **Node** differential privacy

Two graphs are neighbors if one can be obtained from the other by deleting *one node and its adjacent edges*.

Node privacy is stronger, but few node-private algorithms are known.

Why strong guarantees?

Anonymized data isn't.

Some attacks in the literature (citations in paper)
- Reidentifying individuals based on external sources
  - Social networks
  - Computer networks
- Composition attacks
- Reconstruction attacks
- Membership attacks

Node-level differential private alg's provably resist all of these.

Previous work on private graph analysis

- Many works on edge privacy [NRS'07]
  - Wide variety of functionalities: cut estimation, subgraph counts, …
  - Estimators for high-dimensional models, starting with [MW'13]
- Few works on node privacy
  - Initial results assume known degree bound for privacy [GHLP'12]
  - Existing works focus on subgraph counts [BBDS'13, KNRS'13, CZ'13]
  - Estimation of degree distribution [RS'15]
- Main challenge for node private algorithms: sensitivity
  - In sparse graphs, most natural analyses can be completely disrupted by adding a vertex with arbitrary set of edges
  - Private algorithms must be insensitive even in worst case

There were no previously known node-private algorithms for fitting high-dimensional network models

Main Result

Node-differentially private estimator A_ϵ that is **consistent for every bounded graphon**:

$$A_\epsilon(G_n(\rho_n W)) \xrightarrow{P} W \text{ as } n \to \infty$$

as long as average degree $n\rho_n = \omega(\log n)$.

- First estimation result of this generality even without privacy.
  - Previous results made additional assumptions on W

Measuring Convergence: δ_2 metric

- Many graphons generate the same distribution on graphs
  - Relabeling the points in [0,1] doesn't change $G_n(W)$
- Distance on graphons is defined up to "permutations" of [0,1]

$$\delta_2(W, W') = \inf_{\substack{\phi:[0,1]\to[0,1] \\ \text{measure-preserving},1-1}} \|W^\phi - W'\|_2$$

- Here W^ϕ denotes map $(x,y) \mapsto W(\phi(x), \phi(y))$

Oracle and sampling errors

- Our algorithm approximates W using a block graphon
  - Goal: compete with best k-block approximation to W

$$\epsilon_k^{(O)}(W) = \inf_{\substack{k-\text{block} \\ \text{graphons } B}} \delta_2(B, W)$$

"Oracle" error: no better block approximation

- Our algorithm approximates H, then W
  - Goal: compete with approximation to W provided by matrix H

$$\epsilon_n(W) \approx \delta_2(H, W) \; {}^*$$

* Real definition more complicated; see paper.

"Sampling" error. This random variable depends on $\xi_1, \dots, \xi_n$

- For every graphon W, these errors go to 0
  - $\epsilon_k^{(O)}(W) \to 0$ as $k \to \infty$ and $\epsilon_n(W) \xrightarrow{a.s.} 0$ as $n \to \infty$.

Precise bounds

- Our algorithm takes inputs
  - ϵ: privacy parameter
  - Λ: upper bound on W
  - k: number of blocks in estimated graphon
  - G: input graph, assumed to be drawn from $G_n(\rho W)$

Theorem 1: Let $W: [0,1]^2 \to [0, \Lambda]$ be a graphon, let $\rho \in (0,1)$ such that $\rho\Lambda < 1$ and $\rho n > 6\log n$, and assume $k \le \min\left(n\sqrt{\rho/2}, e^{\rho n/2}\right)$. Then

$$\delta_2(\widehat{W}, W) \le \epsilon_k^{(O)}(W) + 2\epsilon_n(W) + O_P\left(\sqrt[4]{\frac{\Lambda^2 \log k}{\rho n}} + \Lambda\sqrt{\frac{k^2 \log n}{n\epsilon}} + \frac{\Lambda}{n\rho\epsilon}\right).$$

- In paper: better bound for a nonprivate version of our algorithm
  - Improves previously known nonprivate bounds [OW '14]
  - Recently improved by [KTV '15]
- For specific families of graphons, bounds on $\epsilon_k^{(O)}(W)$ and $\epsilon_n(W)$:

Upper bounds for	$\epsilon_k^{(O)}(W)$	$\epsilon_n(W)$
k-block graphons	0	$O_P(\sqrt[4]{k/n})$
α-Hölder-continuous	$O(k^{-\alpha})$	$O_P(n^{-\alpha/2})$

Techniques and Proof

Least squares estimator

- We introduce and study a restricted least squares estimator
  - This nonprivate algorithm forms basis of our private algorithm
- On input ϵ, Λ, k, G:
  - $\hat\rho \leftarrow \|G\|_1$ (average density of input)
  - $\hat{B} \leftarrow \underset{\substack{k-\text{block graphons } B \\ \text{with entries} \lesssim \hat\rho\Lambda}}{\operatorname{argmin}} \hat\delta_2(B, G)$
  - Return $\widehat{W} \leftarrow \frac{1}{\hat\rho}\hat{B}$.

$\hat\delta_2$ is a "finite" version of δ_2, where we minimize over assignments of vertices in G to the blocks of B

$$\hat\delta_2(B, G) = \min_{\pi:[n]\to[k]} \|B_\pi - G\|_2$$

- Previous work [OW '14] studied maximum likelihood estimator, which is unstable when W takes small, nonzero values (not suitable for our setting)

Applying the "exponential mechanism"

Exponential mechanism [MT'07] : generic method for private optimization
- Replace "argmin" with sampling from Gibbs-like measure
- Naïve application in our case: $\Pr(\tilde{B} = B) \propto \exp\left(-\frac{\epsilon}{\Delta} \cdot \hat\delta_2(B, G)\right)$

Challenge: For privacy, parameter Δ needs to upper bound changes in $\hat\delta_2$
- Node privacy requires limiting the influence of any single node
- We need $\Delta \ge \max_{G,G' \text{ different by 1 node}} |\hat\delta_2(B,G) - \hat\delta_2(B,G')|$
- Minimal value of Δ is huge, so mechanism returns useless results
  - We make several changes to achieve small Δ

Lipschitz extensions for node stability

- Main technical tool: Lipschitz extensions of graph statistics
  - Let $\mathcal{G}$ be the set of all labeled, finite, undirected graphs
  - Let $\mathcal{G}_d \subseteq \mathcal{G}$ be set of graphs of maximum degree d
    - Partially ordered set under vertex-induced inclusion
    - Metric structure: $d(G, G')$ =number of vertices that must be deleted from G and/or G' to get identical graphs ("vertex distance")
- Lemma [KNRS '13]: If $f: \mathcal{G}_d \to \mathbb{R}$ is monotone and c-Lipschitz, then there exists $f': \mathcal{G} \to \mathbb{R}$ such that
  - f' agrees with f on $\mathcal{G}_d$
  - f' is monotone and c-Lipschitz.

Proof outline of main result

- Run exponential mechanism using Lipschitz extension of $\hat\delta_2(B, G)$ as score
  - Also restrict to matrices with entries bounded by $\rho\Lambda$
- Main steps
  - Show uniform concentration of scores around expectation
  - Show bound on effect of Lipschitz extension
  - Show expectation of $\hat\delta_2(B, G)$ "close" to $\delta_2(B, W)$ with high probability
- Novel aspects
  - Explicit relation to $\epsilon_n(W)$ and $\epsilon_k^{(O)}(W)$
  - Convergence for all bounded graphs
  - Use of Lipschitz extension in exponential mechanism

Conclusions & Consequences

- Private consistent graphon estimation is possible
  - Graphon estimators can be robust to changes in individual nodes
- Design of **robust private** estimator led to better **nonprivate** estimation in sparse graphs
  - Improved [OW] for small ρ and removed requirement that densities be high

Cuts and other estimation tasks

- δ_2 bounds other metrics on graphons and graphs
- Estimation in δ_2 metric allows for estimation of
  - Subgraph frequencies (number of triangles, clustering coefficient, …)
  - Density of every multi-way cut [BCCZ'14]

Open Questions

- Can our bounds be achieved **efficiently**?
  - Best current algorithms take exponential time
    - Private algorithms [this paper]
    - Non private algorithms [OW '14, this paper, KTV '15]
  - Known efficient algorithms are not private and have higher error, e.g., [C'15, AS'15]
- Can **private** algorithms achieve **nonprivate** rates?
  - Independent work [KTV '15] gave optimal nonprivate algorithms for several parameter ranges
  - Private algorithms are currently worse by polynomials in k, n

References (partial list)

See paper for full discussion of related work and citations.

[AS'15] E. Abbe and C. Sandon. Recovering communities in the general stochastic block model without knowing the parameters. arXiv:1503.00609, 2015

[BDK'07] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x? anonymized social networks, hidden patterns, and structural steganography. In *WWW* 2007.

[BBDS'13] J. Blocki, A. Blum, A. Datta, and O. Sheffet. Differentially private data analysis of social networks via restricted sensitivity. In *Innovations in Theoretical Computer Science (ITCS)* 2013

[BCCZ'14] C. Borgs, J. T. Chayes, H. Cohn, and Y. Zhao. An L_p theory of sparse graph convergence II: LD convergence, quotients, and right convergence. arXiv:1408.0744, 2014.

[BCLSV'08] C. Borgs, J. T. Chayes, L. Lovasz, V. Sos, and K. Vesztergombi. Convergent graph sequences I: Subgraph frequencies, metric properties, and testing. *Advances in Math.*, 2008.

[BCL'10] C. Borgs, J. T. Chayes, and L. Lovasz. Moments of two-variable functions and the uniqueness of graph limits. *Geometric And Functional Analysis*, 19(6), 2010.

[BCLSV'12] C. Borgs, J. T. Chayes, L. Lovasz, V. Sos, and K. Vesztergombi. Convergent graph sequences II: Multiway cuts and statistical physics. *Ann. of Math.* 176, 2012.

[C'15] S. Chatterjee. Matrix estimation by universal singular value thresholding. *Annals of Statistics*, 43(1), 2015

[CZ'12] S. Chen and S. Zhou. Recursive mechanism: towards node differential privacy and unrestricted joins. In *SIGMOD* 2013

[DJ'08] P. Diaconis and S. Janson. Graph limits and exchangeable random graphs. *Rendiconti di Matematica*, 28, 2008

[DN'03] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *PODS* 2003.

[DMNS'06] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC* 2006.

[GHLP'12] J. Gehrke, M. Hay, E. Liu, , R. Pass. Zero-knowledge privacy. In *TCC* 2012.

[HLMJ'09] M. Hay, C. Li, G. Miklau, and D. Jensen. Accurate estimation of the degree distribution of private networks. In *ICDM* 2009.

[KNRS'13] S. P. Kasiviswanathan, K. Nissim, S. Raskhodnikova, and A. Smith. Analyzing graphs with node- differential privacy. In *TCC* 2013.

[KTV'15] O. Klopp, A. Tsybakov, and N. Verzelen. Oracle inequalities for network models and sparse graphon estimation. arXiv:1507.04118, 2015.

[M'34] E. J. McShane. Extension of range of functions. Bull. Amer. Math. Soc., 40(12):837–842, 1934.

[MT'07] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *FOCS* 2007.

[MW'12] D. Mir and R. Wright. A differentially private estimator for the stochastic kronecker graph model. In EDBT/ICDT Workshops 2012.

[NS'07] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *IEEE Symp. Security and Privacy* 2009

[NRS'07] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC* 2007.

[RS'15] S. Raskhodnikova and A. Smith. High-dimensional Lipschitz extensions and node-private analysis of network data. arXiv:1504.07912, 2015

[WO'13] P. Wolfe and S. C. Olhede. Nonparametric graphon estimation. arXiv:1309.5936, 2013.

Acknowledgments

A.S. was supported by NSF award IIS-1447700 and a Google Faculty Award. Part of this work was done while visiting Boston University's Hariri Institute for Computation and Harvard University's Center for Research on Computation and Society.