Revealing Network Structure, Confidentially (Improved Rates for Node-Private Graphon Estimation)

Ilias Zadik[1], joint work with Christian Borgs[2], Jennifer Chayes[2] and Adam Smith[3]

[1]Massachusetts Institute of Technology (MIT),
[2]Microsoft Research (MSR) and
[3]Boston University (BU)

59th Symposium of Foundations of Computer Science (FOCS), 2018

Introduction

Large and complicated networks arise everywhere in society! For example,

- the Facebook graph,
- the disease transmission graph
- the collaboration graph
- and many others..

Introduction

Large and complicated networks arise everywhere in society! For example,

- the Facebook graph,
- the disease transmission graph
- the collaboration graph
- and many others..

Analysis of Networks: Important across fields (sociology, medicine etc), rich in theory (random graphs, graph algorithms etc)

Introduction

Large and complicated networks arise everywhere in society! For example,

- the Facebook graph,
- the disease transmission graph
- the collaboration graph
- and many others..

Analysis of Networks: Important across fields (sociology, medicine etc), rich in theory (random graphs, graph algorithms etc)

Privacy on Networks: Huge concern (e.g. Cambridge Analytica Scandal) and also rich in theory. Many open questions for networks!!

This work: Limits of Network Estimation under Privacy

New algorithms and **impossibility results**
for estimating complex network models,
subject to rigorous **privacy constraints** (**node differentially privacy**.)

This work: Limits of Network Estimation under Privacy

New algorithms and **impossibility results**
for estimating complex network models,
subject to rigorous **privacy constraints** (**node differentially privacy**.)

(1) **Stochastic Block Model-Estimation of probability matrix:**
-new analysis of recent private algorithm (BCS'15)
-matches in many regimes the **optimal non-private estimation rate**

This work: Limits of Network Estimation under Privacy

New algorithms and **impossibility results**
for estimating complex network models,
subject to rigorous **privacy constraints** (**node differentially privacy**.)

(1) **Stochastic Block Model-Estimation of probability matrix:**
-new analysis of recent private algorithm (BCS'15)
-matches in many regimes the **optimal non-private estimation rate**

(2) **Erdos-Renyi-Estimation of probability** p:
-Compute tightly **the optimal estimation rate**
-Uses a **novel extension lemma**, potentially of broad use

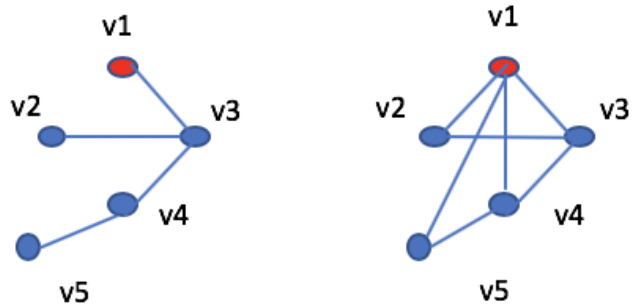
Node Differential Private Algorithms

Intuition: If n-vertex G, G' differ on one user's (node's) data then the outputs of the algorithm are close (in distribution).

Node Differential Private Algorithms

Intuition: If n-vertex G, G' differ on one user's (node's) data then the outputs of the algorithm are close (in distribution).

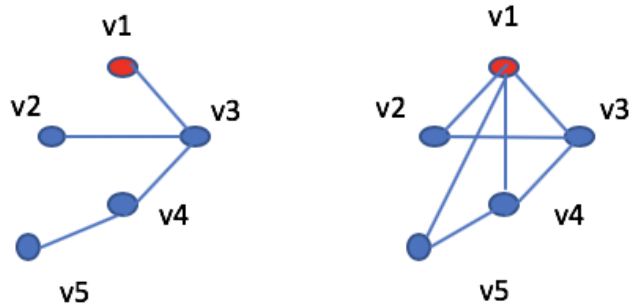
Node-neighbors: We call G, G' node-neighbors if they differ only on the neighborhood of one node.



Node Differential Private Algorithms

Intuition: If n-vertex G, G' differ on one user's (node's) data then the outputs of the algorithm are close (in distribution).

Node-neighbors: We call G, G' node-neighbors if they differ only on the neighborhood of one node.



Definition

A randomized $\mathcal{A}$ on n-vertex graphs is ϵ-node-DP if for node-neighbors $G, G'$and S,

$$\exp(-\epsilon)\,\mathbb{P}\left(\mathcal{A}(G') \in S\right) \leq \mathbb{P}\left(\mathcal{A}(G) \in S\right) \leq \exp(\epsilon)\,\mathbb{P}\left(\mathcal{A}(G') \in S\right).$$

Modeling Large Networks: k-Stochastic Block Model

1-SBM (Erdos Renyi) $G(n, p)$: n nodes every edge appears independently with probability p.

Modeling Large Networks: k-Stochastic Block Model

1-SBM (Erdos Renyi) $G(n, p)$: n nodes every edge appears independently with probability p.

2-SBM $G(n, \begin{bmatrix} p_{1,1} & p_{1,2} \\ p_{2,1} & p_{2,2} \end{bmatrix})$ with $p_{1,2} = p_{2,1}$: n nodes, **2 groups** (each node chooses u.a.r.), and each edge between v_i, v_j **with probability** $p_{group(v_i),group(v_j)}$.

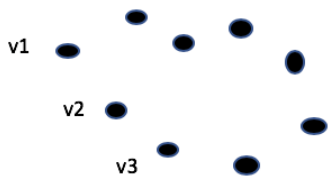


Figure: 9 vertices

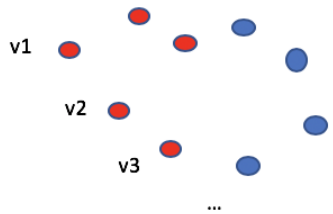


Figure: 2 groups

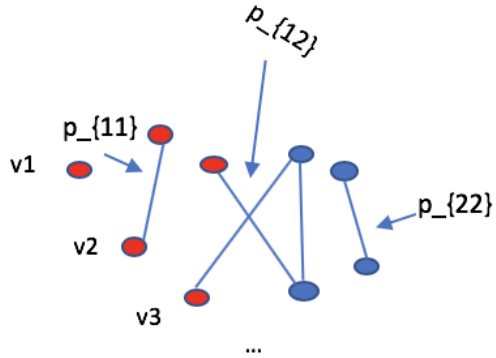


Figure: Assign Edges

Modeling Large Networks: k-Stochastic Block Model

k-SBM, $G(n, B)$, for sym. $B \in [0,1]^{k \times k}$:
n nodes, k **groups** (node choice u.a.r.),
each edge between v_i, v_j **with probability** $B_{group(v_i),group(v_j)}$.

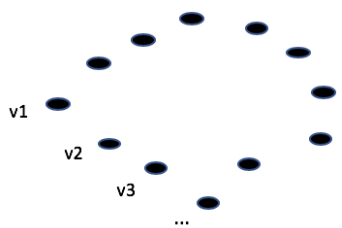


Figure: n = 12

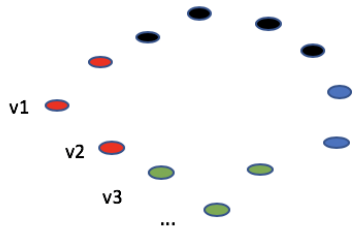


Figure: k = 4

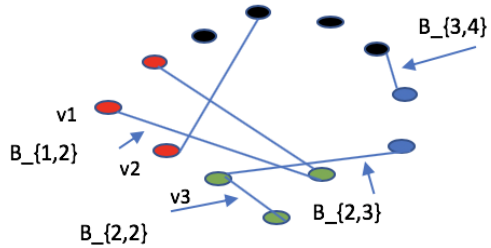


Figure: Assign Edges

.

Modeling Large Networks: k-Stochastic Block Model

k-SBM, $G(n, B)$, for sym. $B \in [0,1]^{k \times k}$:
n nodes, k **groups** (node choice u.a.r.),
each edge between v_i, v_j **with probability** $B_{group(v_i),group(v_j)}$.

Constraint (!) : (ρ-**sparse**) **k-SBM**, $G(n, B)$, where $B \in [0, \rho]^{k \times k}$.

(Vast literature - planted bisection, planted clique, graph limits etc)

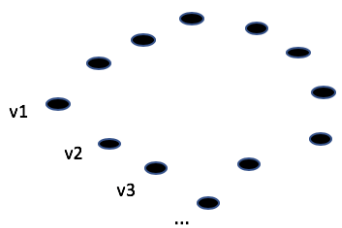


Figure: $n = 12$

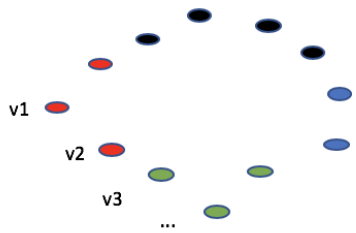


Figure: $k = 4$

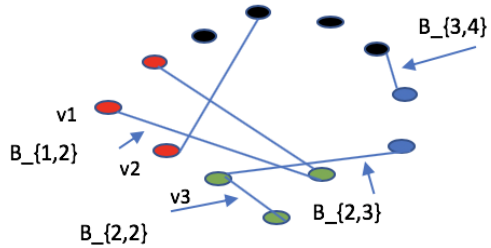


Figure: Assign Edges

.

The Statistical Question

Task: From one sample G from $G(n, B)$ **estimate** B using an ϵ-**node-DP estimator** $\mathcal{A}$.

The Statistical Question

Task: From one sample G from $G(n, B)$ **estimate** B using an ϵ-**node-DP estimator** $\mathcal{A}$.

Each $\mathcal{A}$ has (worst-case over B) error

$$\mathrm{err}(\mathcal{A}) = \max_{B \in [0,\rho]^{k \times k}} \mathbb{E}_{G \sim G(n,B)} \left[\frac{1}{k^2} \|\mathcal{A}(G) - B\|_2^2 \right].$$

The Statistical Question

Task: From one sample G from $G(n, B)$ **estimate** B using an ϵ-**node-DP estimator** $\mathcal{A}$.

Each $\mathcal{A}$ has (worst-case over B) error

$$\mathrm{err}(\mathcal{A}) = \max_{B \in [0,\rho]^{k \times k}} \mathbb{E}_{G \sim G(n,B)} \left[\frac{1}{k^2} \|\mathcal{A}(G) - B\|_2^2 \right].$$

The Estimation Rate

$$R_k(\epsilon) = \min_{\mathcal{A}\ \epsilon\text{–node-DP}} \mathrm{err}(\mathcal{A}).$$

The Statistical Question

Task: From one sample G from $G(n, B)$ **estimate** B using an ϵ-**node-DP estimator** $\mathcal{A}$.

Each $\mathcal{A}$ has (worst-case over B) error

$$\mathrm{err}(\mathcal{A}) = \max_{B \in [0,\rho]^{k \times k}} \mathbb{E}_{G \sim G(n,B)} \left[\frac{1}{k^2} \|\mathcal{A}(G) - B\|_2^2 \right].$$

The Estimation Rate

$$R_k(\epsilon) = \min_{\mathcal{A}\ \epsilon\text{–node-DP}} \mathrm{err}(\mathcal{A}).$$

(For agnostic learning see paper!)

k-SBM Upper Bound

Theorem (informal)

For any $\epsilon > 0$,

$$\mathcal{R}_k(\epsilon) = O\left(\rho\left(\frac{k^2}{n^2} + \frac{\log k}{n}\right)\right) + O\left(\rho^2 \frac{(k-1)^2 \log n}{n\epsilon} + \frac{1}{n^2\epsilon^2}\right).$$

.

k-SBM Upper Bound

Theorem (informal)

For any $\epsilon > 0$,

$$\mathcal{R}_k(\epsilon) = O\left(\rho\left(\frac{k^2}{n^2} + \frac{\log k}{n}\right)\right) + O\left(\rho^2 \frac{(k-1)^2 \log n}{n\epsilon} + \frac{1}{n^2\epsilon^2}\right).$$

.

- *Intuition:* $\frac{k^2}{n^2}$ parametric rate for B, $\frac{\log k}{n} = \frac{\log k^n}{n^2}$ combinatorial rate
- Via a new detailed analysis of an ϵ-node-DP algorithm proposed in (BCS '15).

k-SBM Upper Bound: Optimality in many regimes

Theorem (informal)

For any $\epsilon > 0$,

$$\mathcal{R}_k(\epsilon) = \underbrace{O\left(\rho\left(\frac{k^2}{n^2} + \frac{\log k}{n}\right)\right)}_{\text{optimal non-private rate (KTV'17)}} + O\left(\rho^2 \frac{(k-1)^2 \log n}{n\epsilon} + \frac{1}{n^2\epsilon^2}\right).$$

k-SBM Upper Bound: Optimality in many regimes

Theorem (informal)

For any $\epsilon > 0$,

$$\mathcal{R}_k(\epsilon) = \underbrace{O\left(\rho\left(\frac{k^2}{n^2} + \frac{\log k}{n}\right)\right)}_{\text{optimal non-private rate (KTV'17)}} + O\left(\rho^2 \frac{(k-1)^2 \log n}{n\epsilon} + \frac{1}{n^2\epsilon^2}\right).$$

Comments:

- (GLZ'14), (MS'17), (KTV'17): Optimal ϵ-independent part.
- Many regimes (e.g. ϵ, k constant and $\frac{1}{n} < \rho < \frac{1}{\log n}$):
  -(BCS'15) algorithm, optimal over **all** algorithms!
  -**No additional error** with privacy!

A lower bound for $k \geq 2$

Suppose each node $i \in [n]$ chooses the group in a close to uniform way. (Say each group has probability in $[\frac{1}{4k}, \frac{4}{k}]$.)

Proposition (informal)

For $k \geq 2$ *and any* $\epsilon > 0$,

$$\mathcal{R}_k^*(\epsilon) = \Omega\left(\frac{1}{n^2\epsilon^2}\right),$$

where $\mathcal{R}_k^*$ *stands for the rate for the new variant of the SBM.*

A lower bound for $k \geq 2$

Suppose each node $i \in [n]$ chooses the group in a close to uniform way. (Say each group has probability in $[\frac{1}{4k}, \frac{4}{k}]$.)

Proposition (informal)

For $k \geq 2$ *and any* $\epsilon > 0$,

$$\mathcal{R}_k^*(\epsilon) = \Omega\left(\frac{1}{n^2\epsilon^2}\right),$$

where $\mathcal{R}_k^*$ *stands for the rate for the new variant of the SBM.*

Proof: reduction to privately estimating $q \in [0, 1]$ out of n samples from Bern(q).

The case k = 1: Learning privately Erdos Renyi graphs

Observe simply a $G(n, p)$: estimate **privately** p!

The case k = 1: Learning privately Erdos Renyi graphs

Observe simply a G(n, p): estimate **privately** p!

$$\mathcal{R}_1(\epsilon) = ?$$

The case k = 1: Learning privately Erdos Renyi graphs

Observe simply a G(n, p): estimate **privately** p!

$$\Omega\left(\frac{1}{n^2}+\frac{1}{n^4\epsilon^2}\right)=\mathcal{R}_1(\epsilon)=O\left(\frac{1}{n^2}+\frac{1}{n^2\epsilon^2}\right).$$

Upper bound by main result, Laplace noise to edge density, median of degrees etc.
Lower bounds, by vanilla methods such as packing arguments.

The case k = 1: Learning privately Erdos Renyi graphs

Observe simply a G(n, p): estimate **privately** p!

$$\Omega\left(\frac{1}{n^2}+\frac{1}{n^4\epsilon^2}\right)=\mathcal{R}_1(\epsilon)=O\left(\frac{1}{n^2}+\frac{1}{n^2\epsilon^2}\right).$$

Upper bound by main result, Laplace noise to edge density, median of degrees etc.
Lower bounds, by vanilla methods such as packing arguments.

What is the true ϵ-dependent rate?!

The case $k = 1$: $\frac{1}{n^4\epsilon^2} \leq \epsilon - \text{dep.} \leq \frac{1}{n^2\epsilon^2}$

Theorem

For $\epsilon > \frac{\log n}{n}$,

$$\mathcal{R}_1(\epsilon) = O(\frac{1}{n^2} + \frac{\log n}{n^3\epsilon^2}).$$

Many novel techniques including a general extension lemma (next slide)!

The case k = 1: $\frac{1}{n^4\epsilon^2} \leq \epsilon - \text{dep.} \leq \frac{1}{n^2\epsilon^2}$

Theorem

For $\epsilon > \frac{\log n}{n}$,

$$\mathcal{R}_1(\epsilon) = O(\frac{1}{n^2} + \frac{\log n}{n^3\epsilon^2}).$$

Many novel techniques including a general extension lemma (next slide)!

Proposition (n^3 is tight!)

Furthermore, if G *is sampled u.a.r. from graphs with a fixed number of edges (conditional Erdos Renyi) for* ϵ *constant,*

$$\mathcal{R}_1'(\epsilon) = \Omega(\frac{1}{n^3\epsilon^2}).$$

The extension lemma: beyond networks

Technical challenge with *designing* differential private algorithms:

- **Privacy** constraint should hold for **any** pair of datasets
- **Accuracy** guarantee suffice to hold for **typical** datasets of our input distribution.

The extension lemma: beyond networks

Technical challenge with *designing* differential private algorithms:

- **Privacy** constraint should hold for **any** pair of datasets
- **Accuracy** guarantee suffice to hold for **typical** datasets of our input distribution.

Key contribution: Suffices to be **private** only for **typical** datasets of our input distribution!

The extension lemma: beyond networks

Technical challenge with *designing* differential private algorithms:

- **Privacy** constraint should hold for **any** pair of datasets
- **Accuracy** guarantee suffice to hold for **typical** datasets of our input distribution.

Key contribution: Suffices to be **private** only for **typical** datasets of our input distribution!

Proposition ("Extending Private Algorithms at ϵ-cost")

Let $\hat{\mathcal{A}}$ ϵ-DP on a subset of the input space $\mathcal{H} \subseteq \mathcal{M}$. Then there exists $\mathcal{A}$ defined on $\mathcal{M}$ which is 1) 2ϵ-DP on $\mathcal{M}$ and 2) for every $\mathsf{D} \in \mathcal{H}$, $\mathcal{A}(\mathsf{D}) \stackrel{\mathsf{d}}{=} \hat{\mathcal{A}}(\mathsf{D})$.

Generalizes "extensions" from (KNRS'13), (BBDS'13), (CZ'13), (BCS'15), (RS'15).

Summary of Contributions

(1) We focus on optimal private estimation of **Stochastic Block Model** and **Erdos Renyi** models.

Summary of Contributions

(1) We focus on optimal private estimation of **Stochastic Block Model** and **Erdos Renyi** models.

(2) **Stochastic Block Model:** new analysis of existing algorithm (BCS'15) matches **optimal non-private rate** in many regimes. *Graphons (*k*-SBM for* $k \to +\infty$*) and agnostic learning in the paper!*

Summary of Contributions

(1) We focus on optimal private estimation of **Stochastic Block Model** and **Erdos Renyi** models.

(2) **Stochastic Block Model:** new analysis of existing algorithm (BCS'15) matches **optimal non-private rate** in many regimes. *Graphons (*k*-SBM for* $k \to +\infty$*) and agnostic learning in the paper!*

(3) **Erdos-Renyi**: "almost" tight optimal rate.

Summary of Contributions

(1) We focus on optimal private estimation of **Stochastic Block Model** and **Erdos Renyi** models.

(2) **Stochastic Block Model:** new analysis of existing algorithm (BCS'15) matches **optimal non-private rate** in many regimes. *Graphons (*k*-SBM for* $k \to +\infty$*) and agnostic learning in the paper!*

(3) **Erdos-Renyi**: "almost" tight optimal rate.

(4) Proved an **extension lemma** - potentially of broad use.

Summary of Contributions

(1) We focus on optimal private estimation of **Stochastic Block Model** and **Erdos Renyi** models.

(2) **Stochastic Block Model:** new analysis of existing algorithm (BCS'15) matches **optimal non-private rate** in many regimes. *Graphons (*k*-SBM for* $k \to +\infty$*) and agnostic learning in the paper!*

(3) **Erdos-Renyi**: "almost" tight optimal rate.

(4) Proved an **extension lemma** - potentially of broad use.

Thank you!!