# 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)

CS Theory Seminar at Northeastern, 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..

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

# The privacy issue

**Facts**

**Network data with human users** is sensitive.
Analyzing network data can **leak sensitive information.**

# The privacy issue

## Facts

**Network data with human users** is sensitive.
Analyzing network data can **leak sensitive information.**

...this leads to **high societal concerns!**
(e.g. Cambridge Analytica Scandal)

# The privacy issue

### Facts

**Network data with human users** is sensitive.
Analyzing network data can **leak sensitive information.**

...this leads to **high societal concerns!**
(e.g. Cambridge Analytica Scandal)

Cambridge
Analytica

### Natural Question

How can we **analyze** network data with human users,
but **respect individual privacy**?

# Privacy and the Loss in Accuracy

**An interesting solution** *(Differential privacy! [DMNS '06])*:
Restrict yourself on statistical algorithms whose output imply provably **almost nothing** for any specific user.

# Privacy and the Loss in Accuracy

**An interesting solution** *(Differential privacy! [DMNS '06])*:
Restrict yourself on statistical algorithms whose output imply provably **almost nothing** for any specific user.

Restricting the class of algorithms may cause **significant accuracy loss!**

# Privacy and the Loss in Accuracy

**An interesting solution** *(Differential privacy! [DMNS '06])*:
Restrict yourself on statistical algorithms whose output imply provably **almost nothing** for any specific user.

Restricting the class of algorithms may cause **significant accuracy loss!**

**Main Motivation: Quantify the trade-off**

How much accuracy is necessarily **sacrificed**
if we **restrict** ourselves to differentially private algorithms?

# 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**
- general lower bounds

# 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**
- general lower bounds

(2) **Erdos-Renyi-Estimation of probability** p:

- Compute (almost) tightly **the optimal estimation rate**
- Uses a **novel extension lemma**, potentially of broad use

# Outline

(1) Node Differential Privacy and Stochastic Block Model

(2) The Statistical Task

(3) Main Results

- Upper Bound for k-SBM (optimal in many regimes)
- Lower Bound for k-SBM, $k \geq 2$
- The case $k = 1$ (Erdos-Renyi case)-an almost tight optimal rate

(4) The Extension Lemma

# Outline

(1) **Node Differential Privacy and Stochastic Block Model**

(2) The Statistical Task

(3) Main Results

- Upper Bound for k-SBM (optimal in many regimes)
- Lower Bound for k-SBM, $k \geq 2$
- The case $k = 1$ (Erdos-Renyi case)-an almost tight optimal rate

(4) The Extension Lemma

# Differential Privacy for Networks

**D.P. algorithms: General Idea**

If two input datasets " *differ only on the data of one user*",
then outputs are *"close"* (in distribution).

# Differential Privacy for Networks

**D.P. network algorithms: General Idea**

If two input **networks** " *differ only on the data of one user*",
then outputs are *"close"* (in distribution).

# Differential Privacy for Networks

**D.P. network algorithms: General Idea**

If two input **networks** " *differ only on the data of one user*",
then outputs are *"close"* (in distribution).

*What does "differ on one user" means for two networks (graphs)?*

# Differential Privacy for Networks

**D.P. network algorithms: General Idea**

If two input **networks** " *differ only on the data of one user*",
then outputs are *"close"* (in distribution).

*What does "differ on one user" means for two networks (graphs)?*

- **(Edge-DP)** Two graphs are close if they differ in an **edge**.
  (a *simple "local"* notion, protects *"relationships of individuals"*.)
  *Big literature* [NRS '07], [GRU '12], [XCT '14] and many others

# Differential Privacy for Networks

**D.P. network algorithms: General Idea**

If two input **networks** " *differ only on the data of one user*", then outputs are *"close"* (in distribution).

*What does "differ on one user" means for two networks (graphs)?*

- **(Edge-DP)** Two graphs are close if they differ in an **edge**.
  (a *simple "local"* notion, protects *"relationships of individuals"*.)
  *Big literature* [NRS '07], [GRU '12], [XCT '14] and many others
- **(Node-DP)** Two graphs are close if they differ in a **node**.
  (a *stronger "global"* notion, protects *"individuals"*.)
  *Limited literature* [KNRS'13], [BBDS '13], [CZ '13] and few others

# Differential Privacy for Networks

**D.P. network algorithms: General Idea**

If two input **networks** " *differ only on the data of one user*",
then outputs are *"close"* (in distribution).

*What does "differ on one user" means for two networks (graphs)?*

- **(Edge-DP)** Two graphs are close if they differ in an **edge**.
  (a *simple "local"* notion, protects *"relationships of individuals"*.)
  *Big literature* [NRS '07], [GRU '12], [XCT '14] and many others
- **(Node-DP)** Two graphs are close if they differ in a **node**.
  (a *stronger "global"* notion, protects *"individuals"*.)
  *Limited literature* [KNRS'13], [BBDS '13], [CZ '13] and few others

**This work is for node-DP!**

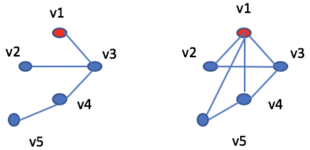
# $\epsilon$-Node Differential Private Algorithms

*Intuition:* If two n-vertex $G, G'$ differ in **one node**, then outputs are **"close"** (in distribution).

# $\epsilon$-Node Differential Private Algorithms

*Intuition:* If two n-vertex $G, G'$ differ in **one node**, then outputs are **"close"** (in distribution).

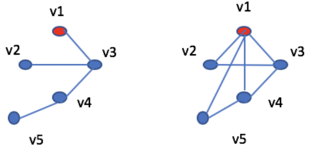
**Node-neighbors**: We call $G, G'$ node-neighbors if they differ only on the neighborhood of one node.



# $\epsilon$-Node Differential Private Algorithms

*Intuition:* If two n-vertex $G, G'$ differ in **one node**, then outputs are **"close"** (in distribution).

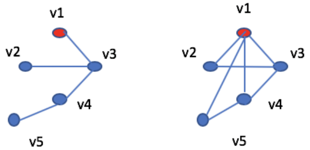
**Node-neighbors**: We call $G, G'$ node-neighbors if they differ only on the neighborhood of one node.




*Node-Distance*: The path metric induced by node-neighbors relation

# $\epsilon$-Node Differential Private Algorithms

*Intuition:* If two n-vertex $G, G'$ differ in **one node**, then outputs are **"close"** (in distribution).
**Node-neighbors**: We call $G, G'$ node-neighbors if they differ only on the neighborhood of one node.




*Node-Distance*: The path metric induced by node-neighbors relation

**Definition**

A randomized $\mathcal{A}$ on n-vertex graphs is $\epsilon$-**node-DP** if for **all** node-neighbors $G, G'$and v in the output space,

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

# k-Stochastic Block Model for Large Networks

Parameters

- n **nodes (users)**
- k **types** (based on characteristics such as *social status*, *cultural background*, *political identity*.)
- $B \in [0, 1]^{k \times k}$ symmetric **probability (frequency) matrix** between the k types.

# k-Stochastic Block Model for Large Networks

Parameters

- n **nodes (users)**
- k **types** (based on characteristics such as *social status, cultural background, political identity*.)
- $B \in [0,1]^{k \times k}$ symmetric **probability (frequency) matrix** between the k types.

The k-Stochastic Block Model

(1) Each node v chooses type(v) from [k] iid **u.a.r.**.

(2) Nodes v, w connect with an edge **w.p.** $B_{\text{type}(v),\text{type}(w)}$ independently.

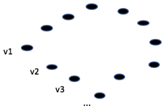



Figure: $n = 12$

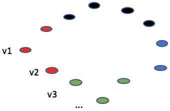



Figure: $k = 4$

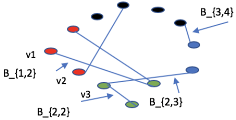



Figure: Assign Edges

# k-Stochastic Block Model for Large Networks

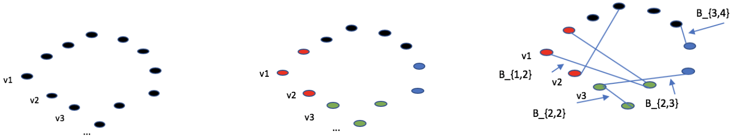
Parameters

- n **nodes (users)**
- k **types** (based on characteristics such as *social status, cultural background, political identity.*)
- $B \in [0,1]^{k \times k}$ symmetric **probability (frequency) matrix** between the k types.

The k-Stochastic Block Model

(1) Each node v chooses type(v) from [k] iid **u.a.r.**.
(2) Nodes $v, w$ connect with an edge **w.p.** $B_{\text{type}(v),\text{type}(w)}$ independently.

If $k = 1$, simple **Erdos-Renyi** model $G(n,p)$!



# Modeling Large Networks: k-Stochastic Block Model

**k-SBM**, $G(n, B)$, for sym. $B \in [0,1]^{k \times k}$:
n **nodes**, k **types** (node's choice u.a.r.),
each edge between $v, w$ **with probability** $B_{type(v),type(w)}$ **independently.**

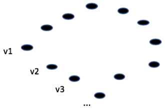



Figure: $n = 12$

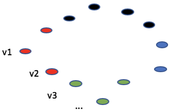



Figure: $k = 4$

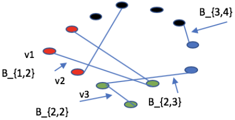



Figure: Assign Edges

# Modeling Large Networks: k-Stochastic Block Model

**k-SBM**, $G(n, B)$, for sym. $B \in [0,1]^{k \times k}$:
n **nodes**, k **types** (node's choice u.a.r.),
each edge between $v, w$ **with probability** $B_{type(v),type(w)}$ **independently.**

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

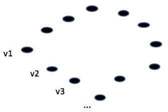



Figure: $n = 12$

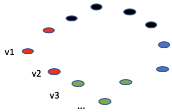



Figure: $k = 4$

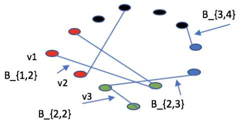



Figure: Assign Edges

# Modeling Large Networks: k-Stochastic Block Model

**k-SBM**, $G(n, B)$, for sym. $B \in [0,1]^{k \times k}$:
n **nodes**, k **types** (node's choice u.a.r.),
each edge between $v, w$ **with probability** $B_{\mathrm{type}(v),\mathrm{type}(w)}$ **independently.**

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

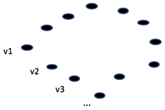



Figure: $n = 12$

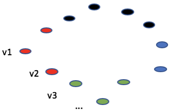



Figure: $k = 4$

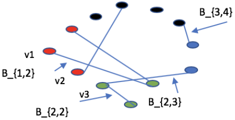



Figure: Assign Edges

*Vast literature* (without privacy): connections with
**community detection (gene expressions, webpage sorting), planted bisection problem, statistical physics models**.

# Outline

(1) Node Differential Privacy and Stochastic Block Model

(2) **The Statistical Task**

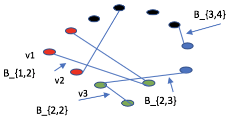
(3) Main Results

- Upper Bound for k-SBM (optimal in many regimes)
- Lower Bound for k-SBM, $k \geq 2$
- The case $k = 1$ (Erdos-Renyi case)-an almost tight optimal rate

(4) The Extension Lemma

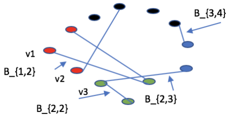
# The Statistical Setting

Let (unknown) $B \in [0, \rho]^{k \times k}$ for (known) k.



# The Statistical Setting

Let (unknown) $B \in [0, \rho]^{k \times k}$ for (known) k.

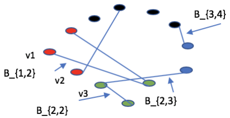



**Task:**

We observe **one** n-vertex sample G from $G(n, B)$.
The goal is to estimate B using an $\epsilon$-**node-DP** $\mathcal{A}(G)$.

# The Statistical Setting

Let (unknown) $B \in [0, \rho]^{k \times k}$ for (known) k.




## Task:

We observe **one** n-vertex sample G from $G(n, B)$.
The goal is to estimate B using an $\epsilon$-**node-DP** $\mathcal{A}(G)$.

## Metric (types-order invariant) for fixed B

$$\delta_2(\mathcal{A}(G), B) = \min_{\pi:[k]\to[k]} \frac{1}{k} \|\mathcal{A}(G)_\pi - B\|_2,$$

where $\mathcal{A}(G)_\pi = \left(\mathcal{A}(G)_{\pi(i),\pi(j)}\right)_{i,j}$.
For $G \sim G(n, B)$, focus on **MSE** $\mathbb{E}_{G\sim G(n,B)}\left[\delta_2\left(\mathcal{A}(G), B\right)^2\right]$.

# The Statistical Question

## Performance of Algorithm for general B

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

$$\mathrm{err}(\mathcal{A}) = \max_{B \in [0,\rho]^{k \times k}} \mathbb{E}_{G \sim G(n,B)} \left[ \delta_2 \left( \mathcal{A}(G), B \right)^2 \right]$$

# The Statistical Question

## Performance of Algorithm for general B

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

$$\mathrm{err}(\mathcal{A}) = \max_{B \in [0,\rho]^{k \times k}} \mathbb{E}_{G \sim G(n,B)} \left[ \delta_2 \left( \mathcal{A}(G), B \right)^2 \right]$$

## The Estimation Rate

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

# The Statistical Question

## Performance of Algorithm for general B

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

$$\mathrm{err}(\mathcal{A}) = \max_{B \in [0,\rho]^{k \times k}} \mathbb{E}_{G \sim G(n,B)} \left[ \delta_2 \left( \mathcal{A}(G), B \right)^2 \right]$$

## The Estimation Rate

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

- Note that we assume G is generated from k-SBM.
- In paper, we generalize to the **agnostic setting** to fitting k-SBM to a $k'$-SBM for unknown $k' > k$.

# Outline

(1) Node Differential Privacy and Stochastic Block Model

(2) The Statistical Task

(3) **Main Results**

- **Upper Bound for** k-**SBM (optimal in many regimes)**
- Lower Bound for k-SBM, $k \geq 2$
- The case $k = 1$ (Erdos-Renyi case)-an almost tight optimal rate

(4) The Extension Lemma

# Upper Bound for k-SBM

**Theorem (informal, (BCSZ FOCS '18))**

*For any* $\epsilon > 0$,

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

.

# Upper Bound for k-SBM

**Theorem (informal, (BCSZ FOCS '18))**

*For any* $\epsilon > 0$,

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

.

- *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 $\epsilon$-node-DP algorithm proposed in (BCS '15).

# Upper Bound for k-SBM

**Theorem (informal, (BCSZ FOCS '18))**

*For any* $\epsilon > 0$,

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

# Upper Bound for k-SBM

**Theorem (informal, (BCSZ FOCS '18))**

*For any $\epsilon > 0$,*

$$\mathcal{R}_k(\epsilon) = \underbrace{O\left(\rho\left(\frac{k^2}{n^2} + \frac{\log k}{n}\right)\right)}_{\textit{Optimal non-private rate!}} + O\left(\frac{\rho^2(k-1)^2 \log n}{n\epsilon} + \frac{1}{n^2\epsilon^2}\right)$$

# Upper Bound for k-SBM

**Theorem (informal, (BCSZ FOCS '18))**

*For any* $\epsilon > 0$,

$$\mathcal{R}_k(\epsilon) = \underbrace{O\left(\rho\left(\frac{k^2}{n^2} + \frac{\log k}{n}\right)\right)}_{\textit{Optimal non-private rate!}} + O\left(\frac{\rho^2(k-1)^2 \log n}{n\epsilon} + \frac{1}{n^2\epsilon^2}\right)$$

Comments:

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

# Upper Bound: Proof Idea

(1) (BCS'15) algorithm is a (quite non-trivial) combination of **exponential, laplace mechanism and Lipschitz extensions** ideas applied to the optimal non-private algorithm (KTV'17).

# Upper Bound: Proof Idea

(1) (BCS'15) algorithm is a (quite non-trivial) combination of **exponential, laplace mechanism and Lipschitz extensions** ideas applied to the optimal non-private algorithm (KTV'17).

(2) All these mechanisms provide **additive error** guarantees.

# Upper Bound: Proof Idea

(1) (BCS'15) algorithm is a (quite non-trivial) combination of **exponential, laplace mechanism and Lipschitz extensions** ideas applied to the optimal non-private algorithm (KTV'17).

(2) All these mechanisms provide **additive error** guarantees.

(3) Adjust the analysis from (KTV'17)- a delicate net argument- to show that it **is not much affected** by additive errors.

# Upper Bound: Tightness?

**Theorem (informal (BCSZ FOCS '18) )**

*For any* $\epsilon > 0$,

$$\mathcal{R}_k(\epsilon) = \underbrace{O\left(\rho\left(\frac{k^2}{n^2} + \frac{\log k}{n}\right)\right)}_{\textit{Optimal non-private rate!}} + O\left(\frac{\rho^2 (k-1)^2 \log n}{n\epsilon} + \frac{1}{n^2\epsilon^2}\right)$$

# Upper Bound: Tightness?

**Theorem (informal (BCSZ FOCS '18) )**

*For any* $\epsilon > 0$,

$$\mathcal{R}_k(\epsilon) = \underbrace{O\left(\rho\left(\frac{k^2}{n^2} + \frac{\log k}{n}\right)\right)}_{\textit{Optimal non-private rate!}} + O\left(\frac{\rho^2 (k-1)^2 \log n}{n\epsilon} + \frac{1}{n^2\epsilon^2}\right)$$

**What about the $\epsilon$-dependent parts?**

# Upper Bound: Tightness?

**Theorem (informal (BCSZ FOCS '18) )**

*For any* $\epsilon > 0$,

$$\mathcal{R}_k(\epsilon) = \underbrace{O\left(\rho\left(\frac{k^2}{n^2} + \frac{\log k}{n}\right)\right)}_{\textit{Optimal non-private rate!}} + O\left(\frac{\rho^2(k-1)^2 \log n}{n\epsilon} + \underbrace{\frac{1}{n^2\epsilon^2}}_{\textit{necessary}, k\geq 2}\right)$$

**What about the $\epsilon$-dependent parts?**

We prove $\frac{1}{n^2\epsilon^2}$ is (almost) necessary if $k \geq 2$.

# Outline

(1) Node Differential Privacy and Stochastic Block Model

(2) The Statistical Task

(3) **Main Results**

- Upper Bound for k-SBM (optimal in many regimes)
- **Lower Bound for** k-**SBM,** $k \geq 2$
- The case $k = 1$ (Erdos-Renyi case)-an almost tight optimal rate

(4) The Extension Lemma

# A lower bound for $k \geq 2$: A variant of SBM

We prove that the term $\frac{1}{n^2\epsilon^2}$ is **necessary** under a small model change.

# A lower bound for k $\geq$ 2: A variant of SBM

We prove that the term $\frac{1}{n^2\epsilon^2}$ is **necessary** under a small model change.

Recall for some underlying probability matrix B:

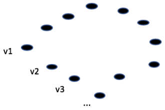



Figure: $n = 12$

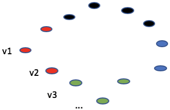



Figure: $k = 4$ u.a.r.

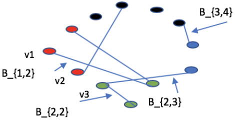



Figure: Assign Edges

.

# A lower bound for $k \geq 2$: A variant of SBM

We prove that the term $\frac{1}{n^2\epsilon^2}$ is **necessary** under a small model change.

Recall for some underlying probability matrix B:

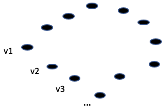



Figure: $n = 12$

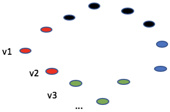



Figure: $k = 4$
**close to** u.a.r.

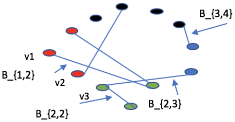



Figure: Assign Edges

.

## New k-SBM

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

# A lower bound for $k \geq 2$: Result

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

# A lower bound for k $\geq$ 2: Result

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

**Proposition (informal, (BCSZ FOCS '18) )**

*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 $k \geq 2$: Result

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

**Proposition (informal, (BCSZ FOCS '18) )**

*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 $\mathrm{Bern}(q)$.

# A lower bound for $k \geq 2$: Result

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

**Proposition (informal, (BCSZ FOCS '18) )**

*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).

**How crucial is that $k \geq 2$?**

# A lower bound for $k \geq 2$: Result

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

**Proposition (informal, (BCSZ FOCS '18) )**

*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).

**How crucial is that $k \geq 2$?**

If $k = 1$, the rate can be improved to $\frac{1}{n^3\epsilon^2}$.

# Outline

(1) Node Differential Privacy and Stochastic Block Model

(2) The Statistical Task

(3) **Main Results**

- Upper Bound for k-SBM (optimal in many regimes)
- Lower Bound for k-SBM, $k \geq 2$
- **The case** $k = 1$ **(Erdos-Renyi case)-an almost tight optimal rate**

(4) The Extension Lemma

# The case $k = 1$: Learning privately Erdos Renyi graphs

**A fundamental open problem**

Observe simply a $G(n, p)$: estimate **privately** $p \in [0, 1]$

# The case $k = 1$: Learning privately Erdos Renyi graphs

## A fundamental open problem

Observe simply a $G(n, p)$: estimate **privately** $p \in [0, 1]$

## Task for $k = 1$

Compute

$$\mathcal{R}_1(\epsilon) = \min_{\mathcal{A}\ \epsilon\text{–node-DP}} \max_{p \in [0,1]} \mathbb{E}_{G \sim G_{n,p}} \left[ |\mathcal{A}(G) - p|^2 \right].$$

# The case $k = 1$: Learning privately Erdos Renyi graphs

$$\mathcal{R}_1(\epsilon) = \min_{\mathcal{A}\ \epsilon\text{-node-DP}} \max_{p \in [0,1]} \mathbb{E}_{G \sim G_{n,p}} \left[ |\mathcal{A}(G) - p|^2 \right] = ?$$

## The case $k = 1$: Learning privately Erdos Renyi graphs

$\mathcal{R}_1(\epsilon) = \min_{\mathcal{A}\ \epsilon\text{–node-DP}} \max_{p \in [0,1]} \mathbb{E}_{G \sim G_{n,p}} \left[ |\mathcal{A}(G) - p|^2 \right] =?$

### Standard Techniques

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

$\mathcal{R}_1(\epsilon) = \min_{\mathcal{A}\ \epsilon\text{-node-DP}} \max_{p \in [0,1]} \mathbb{E}_{G \sim G_{n,p}} \left[ |\mathcal{A}(G) - p|^2 \right] = ?$

**Standard Techniques**

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

**Upper bound** by Laplace mechanism (and our main result for $k = 1$).

# The case $k = 1$: Learning privately Erdos Renyi graphs

$$\mathcal{R}_1(\epsilon) = \min_{\mathcal{A}\ \epsilon\text{–node-DP}} \max_{p \in [0,1]} \mathbb{E}_{G \sim G_{n,p}} \left[ |\mathcal{A}(G) - p|^2 \right] = ?$$

## Standard Techniques

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

**Upper bound** by Laplace mechanism (and our main result for $k = 1$).
**Lower bounds** by standard methods.

# The case $k = 1$: Learning privately Erdos Renyi graphs

$$\mathcal{R}_1(\epsilon) = \min_{\mathcal{A}\ \epsilon\text{–node-DP}} \max_{p\in[0,1]} \mathbb{E}_{G\sim G_{n,p}} \left[ |\mathcal{A}(G) - p|^2 \right] = ?$$

## Standard Techniques

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

**Upper bound** by Laplace mechanism (and our main result for $k = 1$).
**Lower bounds** by standard methods.

What is the true $\epsilon$-dependent rate?!

## The case $k = 1$: Learning privately Erdos Renyi graphs

$$\mathcal{R}_1(\epsilon) = \min_{\mathcal{A}\ \epsilon\text{–node-DP}} \max_{p \in [0,1]} \mathbb{E}_{G \sim G_{n,p}} \left[ |\mathcal{A}(G) - p|^2 \right] = ?$$

### Standard Techniques

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

**Upper bound** by Laplace mechanism (and our main result for $k = 1$).
**Lower bounds** by standard methods.

What is the true $\epsilon$-dependent rate?! **(Almost tight) answer:** $\frac{1}{n^3 \epsilon^2}$

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

**Theorem (BCSZ FOCS '18)**

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

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

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

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

# Upper Bound: The Laplace estimator (suboptimal)

$\mathcal{R}_1(\epsilon) = \min_{\mathcal{A}\ \epsilon\text{–node-DP}} \max_{p \in [0,1]} \mathbb{E}_{G \sim G_{n,p}} \left[ |\mathcal{A}(G) - p|^2 \right] = ?$

## Laplace Estimator

For any f, $f(G) + \mathrm{Lap}(\frac{\Delta}{\epsilon})$ is $\epsilon$-node-DP for

$$\Delta = \max_{G,G' : d_v(G,G')=1} |f(G) - f(G')|.$$

# Upper Bound: The Laplace estimator (suboptimal)

$\mathcal{R}_1(\epsilon) = \min_{\mathcal{A}\ \epsilon\text{–node-DP}} \max_{p \in [0,1]} \mathbb{E}_{G \sim G_{n,p}} \left[ |\mathcal{A}(G) - p|^2 \right] = ?$

## Laplace Estimator

For any f, $f(G) + \mathrm{Lap}(\frac{\Delta}{\epsilon})$ is $\epsilon$-node-DP for

$$\Delta = \max_{G,G' : d_v(G,G')=1} |f(G) - f(G')|.$$

Best non-private estimator is $e(G) = \frac{|E[G]|}{\binom{n}{2}}$.

# Upper Bound: The Laplace estimator (suboptimal)

$\mathcal{R}_1(\epsilon) = \min_{\mathcal{A}\ \epsilon\text{–node-DP}} \max_{p\in[0,1]} \mathbb{E}_{G\sim G_{n,p}} \left[|\mathcal{A}(G) - p|^2\right] = ?$

**Laplace Estimator**

For any f, $f(G) + \mathrm{Lap}(\frac{\Delta}{\epsilon})$ is $\epsilon$-node-DP for

$$\Delta = \max_{G,G':d_v(G,G')=1} |f(G) - f(G')|.$$

Best non-private estimator is $e(G) = \frac{|E[G]|}{\binom{n}{2}}$.

We consider the $\epsilon$-node-DP, $e(G) + \mathrm{Lap}(\frac{\Delta}{\epsilon})$ for

$$\Delta = \max_{G,G':d_v(G,G')=1} |e(G) - e(G')|$$

# Upper Bound: The Laplace estimator (suboptimal)

$\mathcal{R}_1(\epsilon) = \min_{\mathcal{A}\ \epsilon\text{–node-DP}} \max_{p\in[0,1]} \mathbb{E}_{G\sim G_{n,p}}\left[|\mathcal{A}(G) - p|^2\right] = ?$

## Laplace Estimator

For any f, $f(G) + \mathrm{Lap}(\frac{\Delta}{\epsilon})$ is $\epsilon$-node-DP for

$$\Delta = \max_{G,G':d_v(G,G')=1} |f(G) - f(G')|.$$

Best non-private estimator is $e(G) = \frac{|E[G]|}{\binom{n}{2}}$.

We consider the $\epsilon$-node-DP, $e(G) + \mathrm{Lap}(\frac{\Delta}{\epsilon})$ for

$$\Delta = \max_{G,G':d_v(G,G')=1} |e(G) - e(G')| = \Theta\left(n/\binom{n}{2}\right) = \Theta(\frac{1}{n}).$$

# Upper Bound: The Laplace estimator (suboptimal)

$\mathcal{R}_1(\epsilon) = \min_{\mathcal{A}\ \epsilon\text{–node-DP}} \max_{p\in[0,1]} \mathbb{E}_{G\sim G_{n,p}} \left[|\mathcal{A}(G) - p|^2\right] = ?$

## Laplace Estimator

For any f, $f(G) + \mathrm{Lap}(\frac{\Delta}{\epsilon})$ is $\epsilon$-node-DP for

$$\Delta = \max_{G,G':d_v(G,G')=1} |f(G) - f(G')|.$$

Best non-private estimator is $e(G) = \frac{|E[G]|}{\binom{n}{2}}$.

We consider the $\epsilon$-node-DP, $e(G) + \mathrm{Lap}(\frac{\Delta}{\epsilon})$ for

$$\Delta = \max_{G,G':d_v(G,G')=1} |e(G) - e(G')| = \Theta\left(n/\binom{n}{2}\right) = \Theta(\frac{1}{n}).$$

**Same upper bound**

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

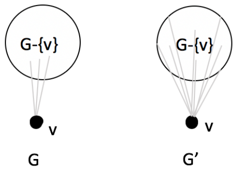
# Upper Bound: Learning from the Laplace estimator

- Recall: $e(G) + \mathrm{Lap}(\frac{\Delta}{\epsilon})$ for $\Delta = \max_{G,G':d_v(G,G')=1} |e(G) - e(G')| = \Theta(\frac{1}{n})$.

# Upper Bound: Learning from the Laplace estimator

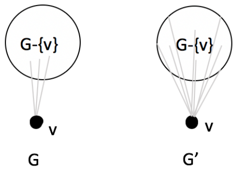
- Recall: $e(G) + \mathrm{Lap}(\frac{\Delta}{\epsilon})$ for $\Delta = \max_{G,G':d_v(G,G')=1} |e(G) - e(G')| = \Theta(\frac{1}{n})$.

- $\Delta = \Theta(\frac{1}{n})$ requires $\Theta(n)$ change in one degree and no change in rest. **Deviation $\Theta(n)$ from average degree in $G$ or $G'$.**



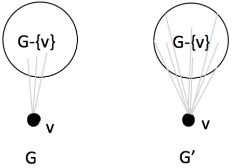
# Upper Bound: Learning from the Laplace estimator

- Recall: $e(G) + \mathrm{Lap}(\frac{\Delta}{\epsilon})$ for $\Delta = \max_{G,G':d_v(G,G')=1} |e(G) - e(G')| = \Theta(\frac{1}{n})$.
- $\Delta = \Theta(\frac{1}{n})$ requires $\Theta(n)$ change in one degree and no change in rest. **Deviation $\Theta(n)$ from average degree in $G$ or $G'$.**
- All degrees from $G(n,p)$ are $(n-1)p + O(\sqrt{n \log n})$: $\sqrt{n \log n}$-window. $\Rightarrow$ $G$ or $G'$ very **atypical** for any Erdos-Renyi graph.



# Upper Bound: Learning from the Laplace estimator

- Recall: $e(G) + \mathrm{Lap}(\frac{\Delta}{\epsilon})$ for
  $\Delta = \max_{G,G':d_v(G,G')=1} |e(G) - e(G')| = \Theta(\frac{1}{n})$.
- $\Delta = \Theta(\frac{1}{n})$ requires $\Theta(n)$ change in one degree and no change in rest. **Deviation $\Theta(n)$ from average degree in $G$ or $G'$.**
- All degrees from $G(n,p)$ are $(n-1)p + O(\sqrt{n \log n})$: $\sqrt{n \log n}$-window.
  $\Rightarrow$ $G$ or $G'$ very **atypical** for any Erdos-Renyi graph.
- How to exclude **atypical graphs**?
  (Challenge: need to be **private for all pairs** of graphs: )



# Upper Bound: Improving the Laplace estimator

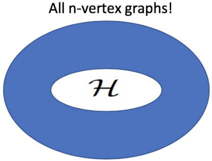
We construct a **subset** $\mathcal{H}$ of all n-vertex graphs

- **typical** for ER graphs

$$\max_{p\in[0,1]} \mathbb{P}_{G\sim G(n,p)}(G \notin \mathcal{H}) = O(\frac{1}{n^2}),$$

- with **lower sensitivity**

  $\max_{G,G'\in\mathcal{H}:d(G,G')=1} |e(G)-e(G')| = O(\sqrt{n\log n}/\binom{n}{2}) = O(\frac{\sqrt{\log n}}{n^{\frac{3}{2}}}).$

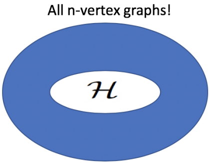


# Upper Bound: Improving the Laplace estimator

- (Privacy in $\mathcal{H}$): Let $\hat{\mathcal{A}}(G) = e(G) + \mathrm{Lap}(\frac{2\sqrt{\log n}}{n^{\frac{3}{2}}\epsilon})$ ,$G \in \mathcal{H}$.

  (1) $\frac{\epsilon}{2}$-node-DP estimator on $\mathcal{H}$ and

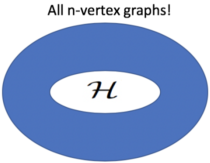
  (2) For $G \in \mathcal{H}$, $\mathbb{E}\left[(\hat{\mathcal{A}}(G) - p)^2\right] = O(\frac{1}{n^2} + \frac{\log n}{n^3\epsilon^2})$.

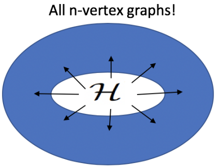


# Upper Bound: Improving the Laplace estimator

- (Privacy in $\mathcal{H}$): Let $\hat{\mathcal{A}}(G) = e(G) + \mathrm{Lap}(\frac{2\sqrt{\log n}}{n^{\frac{3}{2}}\epsilon})$ ,$G \in \mathcal{H}$.
  (1) $\frac{\epsilon}{2}$-node-DP estimator on $\mathcal{H}$ and
  (2) For $G \in \mathcal{H}$, $\mathbb{E}\left[(\hat{\mathcal{A}}(G) - p)^2\right] = O(\frac{1}{n^2} + \frac{\log n}{n^3\epsilon^2})$.

- (Privacy in the whole space+ same accuracy:)
  **Extension lemma** which **extends** $\hat{\mathcal{A}}$ to $\mathcal{A}$
  (1) $\epsilon$-node-DP estimator on evert n-vertex graph and
  (2) $\mathcal{A}(G) = \hat{\mathcal{A}}(G)$ in distribution, when $G \in \mathcal{H}$.



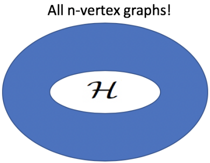



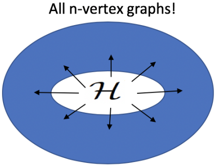
# Upper Bound: Improving the Laplace estimator

- (Privacy in $\mathcal{H}$): Let $\hat{\mathcal{A}}(G) = e(G) + \mathrm{Lap}(\frac{2\sqrt{\log n}}{n^{\frac{3}{2}}\epsilon})$ ,$G \in \mathcal{H}$.
  (1) $\frac{\epsilon}{2}$-node-DP estimator on $\mathcal{H}$ and
  (2) For $G \in \mathcal{H}$, $\mathbb{E}\left[(\hat{\mathcal{A}}(G) - p)^2\right] = O(\frac{1}{n^2} + \frac{\log n}{n^3\epsilon^2})$.

- (Privacy in the whole space+ same accuracy:)
  **Extension lemma** which **extends** $\hat{\mathcal{A}}$ to $\mathcal{A}$
  (1) $\epsilon$-node-DP estimator on evert n-vertex graph and
  (2) $\mathcal{A}(G) = \hat{\mathcal{A}}(G)$ in distribution, when $G \in \mathcal{H}$.

  $\mathcal{H}$ typical for $G_{n,p}$ implies $\mathbb{E}_{G \sim G_{n,p}}\left[|\mathcal{A}(G) - p|^2\right] = O\left(\frac{1}{n^2} + \frac{\log n}{n^3\epsilon^2}\right)$.



## The case $k = 1$: Main Result and Extension Lemma

### Theorem (BCSZ FOCS '18 )

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

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

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

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

# The case $k = 1$: Main Result and Extension Lemma

**Theorem (BCSZ FOCS '18 )**

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

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

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

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

Special Importance for Upper Bound

**Extension Lemma**

*Extended private algorithm from typical instances to private algorithm on the whole space.*

# Outline

(1) Node Differential Privacy and Stochastic Block Model

(2) The Statistical Task

(3) Main Results

- Upper Bound for k-SBM (optimal in many regimes)
- Lower Bound for k-SBM, $k \geq 2$
- The case $k = 1$ (Erdos-Renyi case)-an almost tight optimal rate

(4) **The Extension Lemma**

# 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

# The extension lemma: beyond networks

Technical challenge with *designing* differential private algorithms:

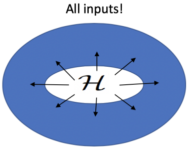
- **Privacy** constraint should hold for **any** pair of datasets
- **Accuracy** guarantee suffice to hold for **typical** datasets of our input distribution.

*Key contribution:* Suffices to be **private** only for **typical** datasets

**Proposition ("Extending any DP Algorithm, (BCSZ FOCS'18) )**

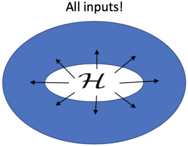
*Let $\hat{\mathcal{A}}$ $\epsilon$-DP on a subset of the input space $\mathcal{H} \subseteq \mathcal{M}$. Then there exists $\mathcal{A}$ defined on $\mathcal{M}$ which is 1) $2\epsilon$-DP on $\mathcal{M}$ and 2) $\forall \mathrm{D} \in \mathcal{H}$, $\mathcal{A}(\mathrm{D}) \overset{d}{=} \hat{\mathcal{A}}(\mathrm{D})$.*

Generalizes "extensions": (KNRS'13), (BBDS'13), (CZ'13), (RS'15).
**Note** on arXiv: *"Private Algorithms Can Always Be Extended"*



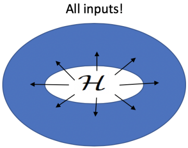
# Proof Ideas of Extension Lemma

- Differential-privacy can be translated into an $\epsilon$-**Lipschitz condition**. (small input changes leads to small output changes)



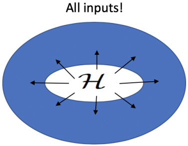
# Proof Ideas of Extension Lemma

- Differential-privacy can be translated into an $\epsilon$-**Lipschitz condition**. (small input changes leads to small output changes)
  $\max_v \frac{\mathbb{P}(\mathcal{A}(G)=v)}{\mathbb{P}(\mathcal{A}(G')=v)} \leq e^{\epsilon d_v(G,G')} \Rightarrow$
  $\max_v |\log \mathbb{P}(\mathcal{A}(G) = v) - \log \mathbb{P}\left(\mathcal{A}(G') = v\right)| \leq \epsilon d_v(G, G')$



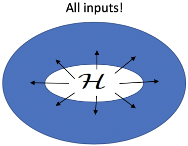
# Proof Ideas of Extension Lemma

- Differential-privacy can be translated into an $\epsilon$-**Lipschitz condition**. (small input changes leads to small output changes)
  $\max_v \frac{\mathbb{P}(\mathcal{A}(G)=v)}{\mathbb{P}(\mathcal{A}(G')=v)} \leq e^{\epsilon d_v(G,G')} \Rightarrow$
  $\max_v |\log \mathbb{P}(\mathcal{A}(G)=v) - \log \mathbb{P}\left(\mathcal{A}(G')=v\right)| \leq \epsilon d_v(G,G')$
- Lipschitz extensions are well-studied in **functional analysis.**



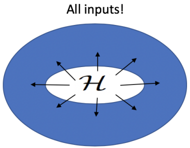
# Proof Ideas of Extension Lemma

- Differential-privacy can be translated into an $\epsilon$-**Lipschitz condition**. (small input changes leads to small output changes)
  $\max_v \frac{\mathbb{P}(\mathcal{A}(G)=v)}{\mathbb{P}(\mathcal{A}(G')=v)} \leq e^{\epsilon d_v(G,G')} \Rightarrow$
  $\max_v |\log \mathbb{P}(\mathcal{A}(G) = v) - \log \mathbb{P}\left(\mathcal{A}(G') = v\right)| \leq \epsilon d_v(G, G')$
- Lipschitz extensions are well-studied in **functional analysis.**
- Standard result: functions with $\ell_\infty(\Gamma)$-**output space** *can always be Lipschitz-extented* with the same Lip constant.



# Proof Ideas of Extension Lemma

- Differential-privacy can be translated into an $\epsilon$-**Lipschitz condition**. (small input changes leads to small output changes)
  $\max_v \frac{\mathbb{P}(\mathcal{A}(G)=v)}{\mathbb{P}(\mathcal{A}(G')=v)} \leq e^{\epsilon d_v(G,G')} \Rightarrow$
  $\max_v |\log \mathbb{P}(\mathcal{A}(G)=v) - \log \mathbb{P}(\mathcal{A}(G')=v)| \leq \epsilon d_v(G,G')$
- Lipschitz extensions are well-studied in **functional analysis.**
- Standard result: functions with $\ell_\infty(\Gamma)$-**output space** *can always be Lipschitz-extented* with the same Lip constant.
- $\epsilon$-DP has **almost this property** but not exactly.
  Yet similar proof (alongside with measure-theory techniques) works by doubling the Lip constant.



# 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 $\rightarrow +\infty$) 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 $\rightarrow +\infty$) and agnostic learning in the paper!*

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

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

# Open Questions

(1) **SBM**: Study of the term $\rho^2 \frac{(k-1)^2 \log n}{n\epsilon}$ (exponential mechanism term).

# Open Questions

(1) **SBM**: Study of the term $\rho^2 \frac{(k-1)^2 \log n}{n\epsilon}$ (exponential mechanism term).

(2) Complete the lower bound proof of $\frac{1}{n^3\epsilon^2}$-rate for **Erdos-Renyi** model.

# Open Questions

(1) **SBM**: Study of the term $\rho^2 \frac{(k-1)^2 \log n}{n\epsilon}$ (exponential mechanism term).

(2) Complete the lower bound proof of $\frac{1}{n^3\epsilon^2}$-rate for **Erdos-Renyi** model.

(3) **Time-efficiency!**
No non-trivial efficient private algorithm even for 2-SBM!

# Open Questions

(1) **SBM**: Study of the term $\rho^2 \frac{(k-1)^2 \log n}{n\epsilon}$ (exponential mechanism term).

(2) Complete the lower bound proof of $\frac{1}{n^3\epsilon^2}$-rate for **Erdos-Renyi** model.

(3) **Time-efficiency!**
No non-trivial efficient private algorithm even for 2-SBM!

(4) Statistically optimal private results using **extension lemma!**
-Recent work on **private estimation of** $\mathcal{N}(\mu, \Sigma)$ [KV'18],[KLSU'18]
What is the **tight statistical rate**? Ongoing work with C. Tzamos.

# Open Questions

(1) **SBM**: Study of the term $\rho^2 \frac{(k-1)^2 \log n}{n\epsilon}$ (exponential mechanism term).

(2) Complete the lower bound proof of $\frac{1}{n^3\epsilon^2}$-rate for **Erdos-Renyi** model.

(3) **Time-efficiency!**
No non-trivial efficient private algorithm even for 2-SBM!

(4) Statistically optimal private results using **extension lemma!**
-Recent work on **private estimation of** $\mathcal{N}(\mu, \Sigma)$ [KV'18],[KLSU'18]
What is the **tight statistical rate**? Ongoing work with C. Tzamos.

# Thank you!!

# The case $k = 1$: Lower Bound Sketch

**Theorem (BCSZ FOCS '18 )**

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

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

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

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

# Lower Bound: Estimation to Testing

## Goal

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

Proof for u.a.r n vertices, m edges (conditional ER), call it $G(n, m)$.

# Lower Bound: Estimation to Testing

**Goal**

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

Proof for u.a.r n vertices, m edges (conditional ER), call it $G(n, m)$.

- **From estimation to testing:** If

$$\mathcal{R}_1(\epsilon) = o\left(\alpha_n^2\right)$$

then we can distinguish between $G_{n,p}$ and $G_{n,p+\alpha_n}$.

# Lower Bound: Estimation to Testing

## Goal

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

Proof for u.a.r n vertices, m edges (conditional ER), call it $\mathsf{G(n,m)}$.

- **From estimation to testing:** If

$$\mathcal{R}_1(\epsilon) = \mathsf{o}\left(\alpha_{\mathsf{n}}^2\right)$$

  then we can distinguish between $\mathsf{G_{n,p}}$ and $\mathsf{G_{n,p+\alpha_n}}$.
- **General privacy limitation:** By using $\epsilon$ (-node)-DP algorithms, inputs of (node-)distance at most $\mathsf{O}\left(\frac{1}{\epsilon}\right)$ are indistinguishable!

# Lower Bound: Estimation to Testing

## Goal

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

Proof for u.a.r n vertices, m edges (conditional ER), call it $G(n, m)$.

- **From estimation to testing:** If

$$\mathcal{R}_1(\epsilon) = o\left(\alpha_n^2\right)$$

then we can distinguish between $G_{n,p}$ and $G_{n,p+\alpha_n}$.

- **General privacy limitation:** By using $\epsilon$ (-node)-DP algorithms, inputs of (node-)distance at most $O\left(\frac{1}{\epsilon}\right)$ are indistinguishable!

## Random Graphs Question

For which $\alpha_n$, $G_{n,p}$ and $G_{n,p+\alpha_n}$ have node-distance $O\left(\frac{1}{\epsilon}\right)$?

# Lower Bound: Coupling Random Graphs

## Goal and an Easy Coupling

Need couple $G_{n,p}$ and $G_{n,p+\alpha_n}$ with *node-distance* $O\left(\frac{1}{\epsilon}\right)$.

Each edge $\alpha_n$-probability slack, easy to couple with $O\left(\alpha_n\binom{n}{2}\right)$ new edges.

# Lower Bound: Coupling Random Graphs

## Goal and an Easy Coupling

Need couple $G_{n,p}$ and $G_{n,p+\alpha_n}$ with *node-distance* $O\left(\frac{1}{\epsilon}\right)$.

Each edge $\alpha_n$-probability slack, easy to couple with $O\left(\alpha_n \binom{n}{2}\right)$ new edges.

- **Easy** $\alpha_n = \frac{1}{n^2 \epsilon} \left(\Rightarrow \mathcal{R}_1(\epsilon) = \Omega\left(\frac{1}{n^4 \epsilon^2}\right)\right)$
  $O\left(\frac{1}{\epsilon}\right)$ new edges, hence *node-distance* $O\left(\frac{1}{\epsilon}\right)$.

# Lower Bound: Coupling Random Graphs

## Goal and an Easy Coupling

Need couple $G_{n,p}$ and $G_{n,p+\alpha_n}$ with *node-distance* $O\left(\frac{1}{\epsilon}\right)$.

Each edge $\alpha_n$-probability slack, easy to couple with $O\left(\alpha_n \binom{n}{2}\right)$ new edges.

- **Easy** $\alpha_n = \frac{1}{n^2 \epsilon}$ $\left(\Rightarrow \mathcal{R}_1(\epsilon) = \Omega\left(\frac{1}{n^4 \epsilon^2}\right)\right)$
  $O\left(\frac{1}{\epsilon}\right)$ new edges, hence *node-distance* $O\left(\frac{1}{\epsilon}\right)$.
- **Harder** $\alpha_n = \frac{1}{n^{\frac{3}{2}} \epsilon}$ $\left(\Rightarrow \mathcal{R}_1(\epsilon) = \Omega\left(\frac{1}{n^3 \epsilon^2}\right)\right)$
  $O\left(\frac{\sqrt{n}}{\epsilon}\right)$ new edges, can we assign $\sqrt{n}$-edges per vertex?

## Proposition (Key Step)

*For appropriate choice of* $m = \Theta(n^2)$, *there is a coupling between* $G(n, m)$ *and* $G(n, m + o(\sqrt{n}))$ *where instances are always node-neighbors.*