Robust Estimation for Random Graphs

Jayadev Acharya, Cornell University

work with

Ayush Jain, UC San Diego

Gautam Kamath, U Waterloo

Ananda Theertha Suresh, Google Research

Huanyu Zhang, Facebook

Outline

- Problem formulation
- Related work
- Results
- Proof sketch
- Conclusion

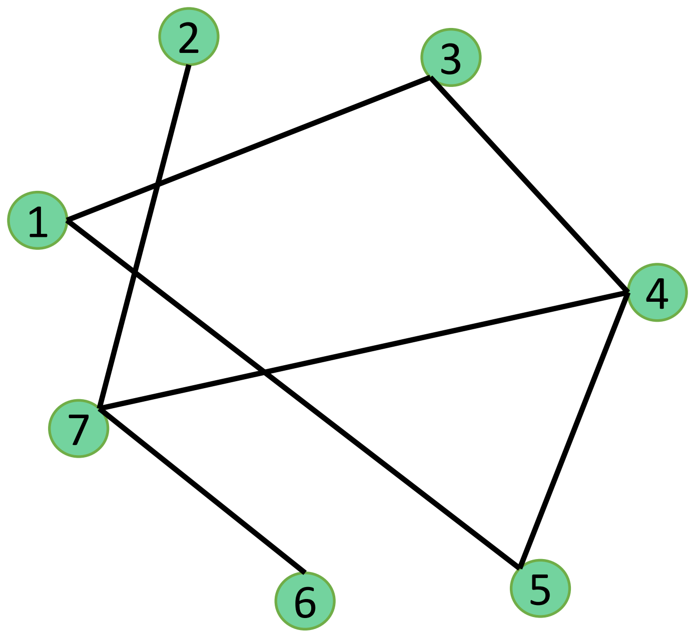
Setup

Problem formulation

$G(n,p)$: Erdős Rényi graphs over n nodes

$\Pr\big((i,j)\text{ exists}\big) = p$ independently

Given $G \sim G(n,p)$, estimate p



Simple estimators

d_j: degree of node j

Mean estimator:

$$\hat{p} = \frac{d_1 + \cdots + d_n}{n(n-1)}$$

Median estimator:

$$\hat{p} = \frac{\mathrm{Median}\{d_1, \ldots, d_n\}}{n-1}$$

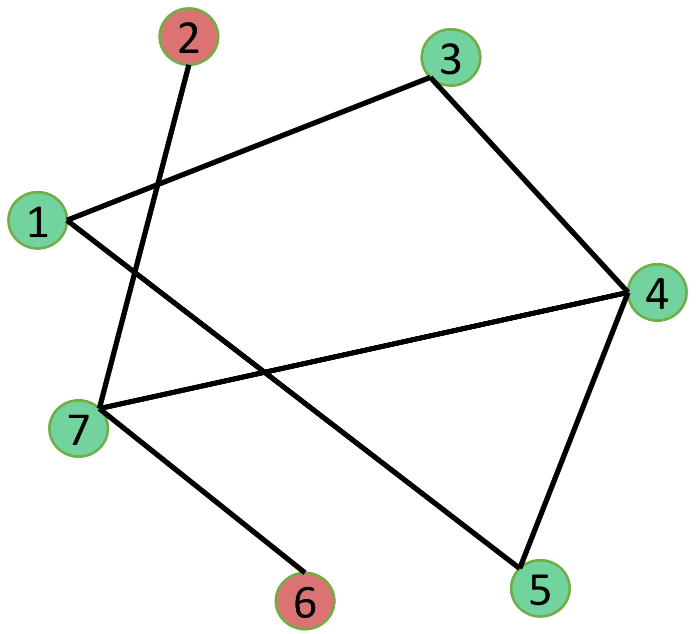
Lemma. For the mean estimator

$$|\hat{p} - p| = \Theta\left(\frac{\sqrt{p(1-p)}}{n}\right)$$

Robust estimation under corruptions

An adversary $\mathcal{A}$:

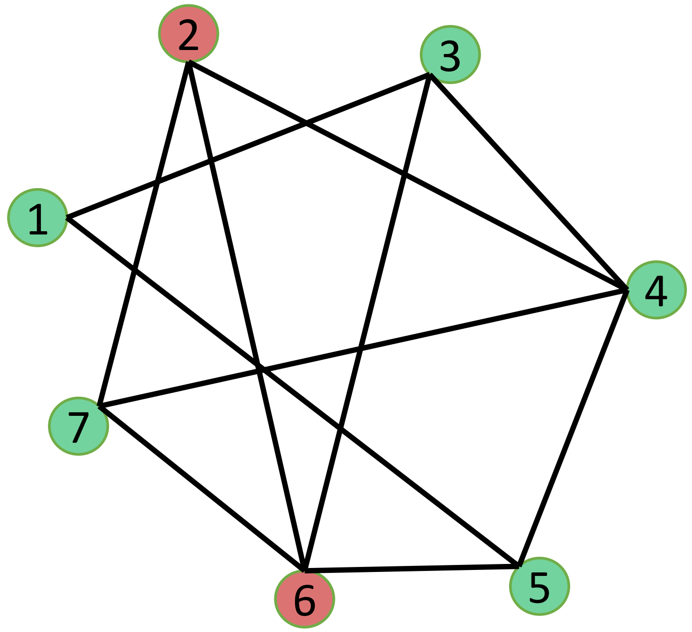
- Looks at G
- Picks a set B nodes with $|B| = \gamma n$
- Changes neighborhood of B as it likes
- We observe resulting graph $\mathcal{A}(G)$



Robust estimation under corruptions

An adversary $\mathcal{A}$:

- Looks at G
- Picks a set B nodes with $|B| = \gamma n$
- Changes neighborhood of B as it likes
- We observe resulting graph $\mathcal{A}(G)$



Given $\mathcal{A}(G)$, estimate p.

Robust estimation

Robust statistics:

Donoho, Hampel, Huber, Rousseeuw, Tukey, …

More recently, computationally efficient multivariate estimation

LRV’16, DKKLMS’16, …

Robust estimation of discrete distributions

QV’17, CLM’19, JO’20

Robust community detection

CL’14

Graph estimation under-differential privacy

BCSZ’18, SU’19

Our Results

Simple estimators with corruptions

For both mean and median estimators:

$$|\hat{p} - p| = \Theta\left(\gamma + \frac{\sqrt{p(1-p)}}{n}\right)$$

Prune-then simple estimators

Prune-then-estimation:

- Remove $c \cdot \gamma$ fraction of nodes with largest and smallest degrees
- Output the mean/median of the remaining subgraph

Lemma. For prune-then-median:

$$|\hat{p} - p| = \Omega\left(\gamma + \frac{\sqrt{p(1-p)}}{n}\right)$$

Lemma. For prune-then-mean:

$$|\hat{p} - p| = \Omega\left(\gamma^2 + \frac{\sqrt{p(1-p)}}{n}\right)$$

Main result: upper bound

Theorem. There exists an algorithm such that

$$|\hat{p} - p| = \tilde{O}\left(\frac{\sqrt{p(1-p)}}{n} + \frac{\gamma\sqrt{p(1-p)}}{\sqrt{n}} + \frac{\gamma}{n}\right).$$

It runs in time $\tilde{O}(\gamma n^3 + n^{2.5})$.

If $p \in \left(\frac{1}{n}, 1 - \frac{1}{n}\right)$ and $\gamma > 1/\sqrt{n}$,

$$|\hat{p} - p| = \tilde{O}\left(\frac{\gamma\sqrt{p(1-p)}}{\sqrt{n}}\right)$$

$$= \tilde{O}\left(\gamma\sqrt{n} \cdot \frac{\sqrt{p(1-p)}}{n}\right)$$

Main result: lower bound

Let $p \in \left(\frac{1}{n}, 1 - \frac{1}{n}\right)$ and $\gamma > 1/\sqrt{n}$

$$\delta(p, \gamma, n) \coloneqq 0.05 \cdot \frac{\gamma\sqrt{p(1-p)}}{\sqrt{n}}$$

There exists $\mathcal{A}$ such that if $G \sim G(n, p)$, and $G' \sim G\big(n, p + \delta(p, \gamma, n)\big)$

$$d_{\mathrm{TV}}\big(\mathcal{A}(G), \mathcal{A}(G')\big) < 0.1.$$

Furthermore, $\mathcal{A}$ corrupts a randomly chosen B.

Upper and lower bounds of up to log factors

(tight in all terms).

Upper Bounds

Upper bound outline

A two-step algorithm:

- A spectral algorithm to output a coarse estimate $\hat{p}$ such that

$$|\hat{p} - p| = \tilde{O}\left(\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$$

- A post-processing step to improve the estimate

Large subsets of uncorrupted nodes are good

A: adjacency matrix of $\mathcal{A}(G)$

For $S \subseteq [n]$,

$A_{S\times S}$: submatrix of A restricted to $S\times S$

p_S: average $A_{S\times S}$ (density of subgraph of $\mathcal{A}(G)$ induced by S)

$(A - p_S)_{S\times S}$: subtracting p_S from each entry in $A_{S\times S}$

$F = [n] \setminus B$: set of uncorrupted nodes

Lemma. W.h.p. simultaneously for all $F' \subset F$: $|F'| > n(1 - 18\gamma)$:

1. $\| (A - p_{F'})_{F'\times F'} \|$ is small
2. $p_{F'}$ is a good estimate of p

Small norm implies good estimate

Let $S \subseteq [n]$ be such that $|S| > n(1 - 9\gamma)$

Lemma. If $\| (A - p_S)_{S \times S} \|$ is small, then p_S is a coarse estimate of p.

Proof sketch:

- $S \cap F$ is a large uncorrupted set => $p_{S \cap F}$ is close to p
- If p_S is far from p, then $p_{S \setminus S \cap F}$ is far from p
- Implies a lower bound on spectral norm

An inefficient coarse estimation:

- Iterate over all large subsets to minimize the spectral norm above

Making it efficient

Suppose $|S| > n(1-9\gamma)$ and v a normalized top eigenvector of $(A-p_S)_{S\times S}$

Main lemma. If $\| \ (A-p_S)_{S\times S} \ \|$ is large, then $\| \ v_{S\cap B} \ \|^2$ is at least a constant.

Algorithm:

- $S = [n]$
- While $|S| > n(1-9\gamma)$:
  - Compute top eigenvector v of $(A-p_S)_{S\times S}$
  - Sample i with probability v_i^2
  - $S \leftarrow S \setminus \{i\}$

Step 2: pruning

S^*: set returned by coarse algorithm such that

$$|p_{S^*} - p| = \tilde{O}\left(\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$$

Pruning:

- Remove $3\gamma n$ nodes with highest and lowest degrees
- Output the mean $\hat{p}$ of the remaining nodes

Theorem.

$$|\hat{p} - p| = \tilde{O}\left(\gamma\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$$

Lower Bounds

Lower bound

Let $p \in \left(\frac{1}{n}, 1 - \frac{1}{n}\right)$ and $\gamma > 1/\sqrt{n}$

$$\delta(p, \gamma, n) := 0.05 \cdot \frac{\gamma\sqrt{p(1-p)}}{\sqrt{n}}$$

There exists an adversary such that if $G \sim G(n, p)$, and $G' \sim G\big(n, p + \delta(p, \gamma, n)\big)$

$$d_{\mathrm{TV}}\big(\mathcal{A}(G), \mathcal{A}(G')\big) < 0.1.$$

A coupling for lower bound

Done if we can convert $G \sim G(n,p)$ to $G' \sim G\big(n, p + \delta(p,\gamma,n)\big)$ by changing γn nodes.

- Node degrees of G are $Bin(n-1, p)$
- Node degrees of G' are $Bin\big(n-1, p + \delta(p,\gamma,n)\big)$

$$d_{\mathrm{TV}}\Big(Bin(n-1,p), Bin\big(n-1, p+\delta(p,\gamma,n)\big)\Big) < \gamma/10$$

If node degrees of $G(n,p)$ are independent:

- There is a coupling between G and G' with $n \cdot \gamma/10$ node changes in expectation

Unfortunately, node degrees are not independent

Directed graphs to the rescue

$DG(n,p)$: directed ER graphs

- Outgoing node degrees are $Bin(n-1,p)$
- Degrees are independent
- $\delta(p,\gamma,n)$ lower bound holds for estimating p for $DG(n,p)$ under γ corruptions

Lemma. For any γ, parameter estimation for $G(n,p)$ is harder than $DG(n,p)$

Thank You

Conclusion:

- Robust estimation task for graph problems
- Almost optimal results for ER parameter estimation

Ongoing work:

- Stochastic block models

Other directions:

- Other random graph models