Differentially Private Community Detection for Stochastic Block Models

Mohamed Seif*, Dung Nguyen*, Anil Vullikanti, Ravi Tandon

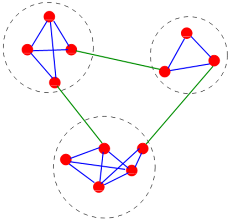
University of Arizona, University of Virginia

*Equal contributions

Background: Community Detection

- In general, a community is a subgraph that is
  - well-connected inside
  - sparsely connected to other communities
- Community Detection: the problem of finding similarity classes of vertices in a network by having access to measurements of local interactions [1],[2]

- Communities are often measured by metrics such as cut and modularity
- [Barnes 1982] uses minimum bisection cut
- [MacQueen 1967] minimizes total intra-cluster distance (k-means)
- [Newman 2014] maximizes modularity



Simple communities in a graph. Image from [2]

[1] Abbe 2017

[2] Fortunato 2010

Applications

Community Detection has been used in many domains:[1]

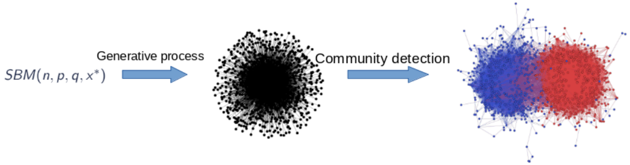
- Bio-informatics
- Recommender systems
- NLP
- Social network analysis
- Caveat: In many applications, communities may overlap and be not well-separated

[1] Fortunato 2010

Stochastic Block Model (SBM)

- Community Detection using metrics is ad-hoc
  - Community Detection in SBM is a systematic approach
- SBM is a random graph model
  - Nodes are divided into multiple (r) blocks (communities)
  - Connections between all pairs of nodes are generated with probability
    - p if endpoints are in the same block (intra-community)
    - q if endpoints are in different blocks (inter-community)
- Other related models: Degree-Corrected SBM, Symmetric Binary SBM, Graphon (infinite communities),...
- Binary Symmetric SBM is the simplest variant (two equal-sized blocks)
  - Is a canonical model for graph algorithms
  - Is extensively studied for theoretical boundaries, efficient algorithms, etc [1]
  - even that, Community Detection with Differential Privacy in BSSBM is still challenging and open

[1] Abbe 2017



Community Detection from graph generated by an SBM. Graph images from [1]

- **Community Detection:** To output the underlying communities, given an input graph $G = (V, E)$, assumed to be generated by an SBM.
- **Edge-Differential Privacy.** To guarantee (ϵ, δ)-Differential Privacy in the Edge-Privacy model, i.e., for any two graphs $G \sim G'$ that differ by exact one edge
  $\forall O \subseteq Range(f) : \Pr[f(G) \in O] \leq e^{\epsilon} \Pr[f(G') \in O] + \delta$.
- **Exact Recovery** [1] [2]
  - When the ground-truth labeling σ^* is recovered correctly.

[1] Abbe 2017

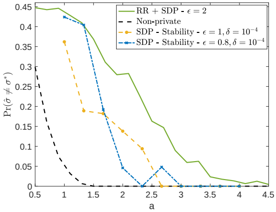
[2] Abbe 2015

Exact recovery

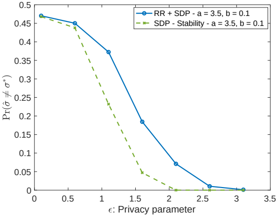
- Threshold in the dense regime: $p = \frac{a \log n}{n}, q = \frac{b \log n}{n}$
- Exact Recovery is possible when:
  - Binary Symmetric SBM: $\sqrt{a} - \sqrt{b} \geq \sqrt{2}$
  - r communities: $\sqrt{a} - \sqrt{b} \geq \sqrt{r}$
- Multiple Exact Recovery approaches in the non-private settings (not inclusive list):
  - Minimum bi-section cut (Maximum Likelihood Estimator - MLE)
  - Semi-definite programming (SDP)
  - Spectral methods

Contributions

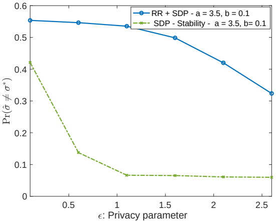
- First edge-DP algorithms for community detection in SBMs with rigorous bounds on recoverability in the dense regime
- Design and Analysis of DP mechanisms based on the Stability of Estimators
- Other approaches: Sampling based methods (Exponential Mechanism, Bayesian Estimator), Randomized Response mechanism
- Empirical results on synthetic and real-world networks



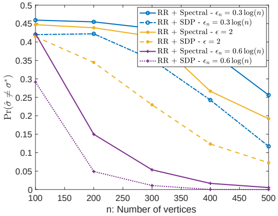
(a) Impact of changing a where $r = 2$ and $n = 100$.



(b) Impact of ϵ where $r = 2$ and $n = 200$.



(c) Impact of ϵ where $r = 3$ and $n = 200$.



(d) SDP vs. Spectral method for $r = 2$ communities.

Stability and Privacy

- **Stability.** (Informally) A graph G is k-stable under some function f if flipping up to k connections of G does not affect the value of f
    - f can be MLE or SDP Relaxation
- When $k = 0$, G is unstable
- **Main idea**: If G is generated by an SBM with appropriate parameters, G is $\Omega(\log n)$-stable under MLE and SDP Relaxation, hence the Stability mechanism performs Exact Recovery w.h.p..

Proposed Mechanisms Summary

	MLE-Stability	SDP-Stability
ϵ	$\mathcal{O}(1)$	$\mathcal{O}(1)$
δ	$1/n^2$	$1/n^2$
$\sqrt{a}-\sqrt{b}\geq$	$\sqrt{2}\cdot\sqrt{1+3/2\epsilon}$	$\sqrt{2}\cdot\sqrt{2+3/2\epsilon}$
Time	$\mathcal{O}(\exp(n))$	$n^{(\mathcal{O}(\log(n)))}$

	Bayesian	Exponential	RR + SDP
ϵ	$\Omega(\log(a/b))$	$\mathcal{O}(1)$	$\Omega(\log(n))$
δ	0	0	0
$\sqrt{a}-\sqrt{b}\geq$	$\frac{2}{(\sqrt{2}-1)(1-e^{-\epsilon_0})}$	$\frac{2}{(\sqrt{2}-1)\epsilon}$	$\sqrt{2}\times\frac{\sqrt{e^{\epsilon}+1}}{\sqrt{e^{\epsilon}-1}}+\frac{1}{\sqrt{e^{\epsilon}-1}}$
Time	$\mathcal{O}(\exp(n))$	$\mathcal{O}(\exp(n))$	$\mathcal{O}(\text{poly}(n))$

Conclusion

- We study the community detection problem for Stochastic Block Models with edge-differential privacy
- First Exact Recovery algorithms with edge-DP for community detection in SBMs: Stability-based, Sampling-based, Randomized Response based
- Analyze the rigorous bounds on recoverability in the dense regime
    - Proving the Stability properties of Maximum Likelihood Estimator and SDP Relaxation under edge-perturbation
- Conduct experiments to confirm the utility of the proposed algorithms in both synthetic and real-world networks