Community detection in networks

- Networks with community structures arise in many applications.
- Task: Find underlying communities based on network topology.
- Applications: Friend or movie recommendation in online social networks.

Cluster recovery under stochastic blockmodel

Vast literature on stochastic blockmodel [Holland et al. ’83] and planted partition model [Condon-Karp ’01]:
- [Bickel-Chen ’09] [Rohe et al. ’10] [Jin ’12] [Mossel et al. ’12] (Mossel et al. ’13) [Mossel et al. ’14] [Ca-Li ’14] [Guedon-Vershynin ’14] [Arias-Castro-Verzelen ’14] [Lei-Rinaldo ’14] [Le-Levina-Vershynin ’15]...
- [McSherry ’01] [Coja-Oghlan ’10] [Chaudhuri et al. ’12] [Ames ’12] [Chen-Sanghavi-Xu ’12] [Heimlicher et al. ’12] [Anandkumar et al. ’13] [Lelarge et al. ’13] [Massoulié ’13][Vinayak-Cynam-Hassibi ’14] [Abbe et al. ’14] [Yun-Proustiere ’14] [Abbe-Sandron ’15] [Chin-Rao-Vu ’15]...

This paper focuses on a single community

- One cluster of size \( K \) plus \( n - K \) outliers.
- Connectivity \( p \) within cluster and \( q \) otherwise.
- Also known as Planted Dense Subgraph model.
- \( p = 1, q = \gamma \) corresponds to Planted Clique model.

Planted clique hardness hypothesis

\( H_0 : \text{Bern}(\gamma) \) vs \( H_1 : \text{Bern}(1) \)

[Alon et al. ’98] [Dekel et al. ’10] [Deshpande-Montanari ’13].

- Intermediate regime: \( \log n < K < \sqrt{\pi n} \), \( \gamma = 0(1) \).
- Detection is possible but believed to have high computational complexity: [Alon et al. ’11] [Feldman et al. ’13]...
- Many (worst-case) hardness results assuming Planted Clique hardness with \( \gamma = \frac{2}{3} \).
- Detecting sparse principal component: [Berthet-Rigollet ’13].
- Detecting sparse matrix: [Ma-Wu ’13].
- Cryptography: [Applebaum et al. ’10].

Hardness for detecting a single cluster

Assuming Planted Clique hardness for any constant \( \gamma > 0 \):

\[ K = \Theta(n^\alpha) \]

\( \alpha = \frac{\beta - 1}{2} \approx 0.15 \) \( \alpha = \frac{\beta - 1}{2} \approx 0.25 \)

Main result: Detecting a single cluster in the red regime is at least as hard as detecting a clique of size \( K = o(\sqrt{n}) \).

Corollary of main result: Recovering a single cluster in the red regime is at least as hard as detecting a clique of size \( K = o(\sqrt{n}) \).

About the spectral barrier

[Nadakuditi-Newman ’12]

\[ \lambda = \Theta(n^\alpha) \]

Eigenvalue distribution of \( A \):

\[ \sigma = \sqrt{\lambda^2 + (n - K)\lambda} \]

Conjecture [Chen-Xu ’14]: no polynomial-time algorithm can recover beyond the spectral barrier. (Our corollary partially resolves this conjecture.)

Formal statement of hardness of detecting a clique

\( \gamma \): edge probability in Planted Clique

Theorem

Assume Planted Clique Hypothesis holds for all \( 0 < \gamma \leq 1/2 \). Let \( \alpha > 0 \) and \( \beta < 1 \) be such that

\[ \alpha < \beta < \frac{1}{2} \]

Then there exists a sequence \( \{(N, K, q)\}_{i \in \mathbb{N}} \)

satisfying \( \lim_{i \to \infty} \frac{\log K}{\log n} = \alpha \) and \( \lim_{i \to \infty} \frac{\log K}{\log n} = \beta \) such that for any sequence of randomized polynomial-time tests \( \Omega \) for the PDS\( (N, K, 2q, q) \) problem, the Type-I+II error probability is lower bounded by 1.

Proof requires a polynomial time reduction

\( h : A_{n \times n} \mapsto \bar{A}_{n \times n} \)

\( H_0 : \text{Bern}(\gamma) \) vs \( \text{Bern}(q) \)

\( H_1 : \text{clique} K \)

Need \( h : A \mapsto A \) agnostic to the clique and computable in polynomial time.

Given an integer \( \ell \), two probability distributions \( P, Q \) on \( \{0, 1, \ldots, \ell^2\} \).

How to choose \( P, Q^2 \)? Matching \( H_0 : (1 - \gamma)Q + \gamma P = \text{Binomial}(\ell^2, q) \).

Matching \( H_1 \): approximately \( P \approx \text{Binomial}(\ell^2, p) \) in total variation distance.

Lemma (Bound the total variation distance)

\[ \ell, n \in \mathbb{N}, k \in [n] \text{ and } \gamma \in (0, 1/2) \].

Let \( N = n, K = k \), \( P = 2q \) and \( \omega_0 = \log(1/\gamma) \). Assume that \( 16q^2 \leq 1 \) and \( k \geq 6e^{-1} \).

If \( G \sim G(n, \gamma) \), then \( G \sim G(N, q) \).

If \( G \sim G(n, k, 1/\gamma) \), then \( d_{TV}(P_D, G(N, q)) \leq \omega_0 + \frac{k^2(q^2)^{\omega_0} + \sqrt{\omega_0} - 1}{\sqrt{\omega_0}} \).

Please see paper for more information and references.

Thanks!