Community Detection via Random and Adaptive Sampling

Seyoung Yun (MSR-Inria)
with Alexandre Proutiere (KTH/Inria)
Community detection in networks

Objective: Extract K communities in a network of n nodes from random observations

Observations:
1. A graph (see Neeman’s talk -- stochastic block model)
2. This talk: a more general sampling framework
Stochastic Block (SB) model

- The graph is built by considering each pair of nodes once
  - If in the same community: put an edge with probability $p$
  - Else: put an edge with probability $q < p$
Stochastic Block (SB) model

- The graph is built by considering each pair of nodes once
  - If in the same community: put an edge with probability $p$
  - Else: put an edge with probability $q < p$
Stochastic Block (SB) model

- The graph is built by considering each pair of nodes once
  - If in the same community: put an edge with probability $p$
  - Else: put an edge with probability $q < p$
Stochastic Block (SB) model

- The graph is built by considering each pair of nodes once
  - If in the same community: put an edge with probability $p$
  - Else: put an edge with probability $q < p$
The graph is built by considering each pair of nodes once
- If in the same community: put an edge with probability $p$
- Else: put an edge with probability $q < p$
Stochastic Block (SB) model

- The graph is built by considering each pair of nodes once
  - If in the same community: put an edge with probability $p$
  - Else: put an edge with probability $q < p$
Sampling Framework

• The interaction of two nodes can be sampled repeatedly
• Sample for a given node pair: Bernoulli with mean $p$ if nodes are in the same cluster, with mean $q$ otherwise
• Sample budget: $T$
Sampling Strategies

• Non-adaptive Random Strategies
  – The pair of nodes sampled in round $t$ does not depend on past observations, and is chosen uniformly at random
  – S1: sampling with replacement
  – S2: sampling without replacement

• Adaptive Strategies
  – The pair of nodes sampled in round $t$ depends on past observations

• Stochastic Block Model: random sampling without replacement, and $T = n(n-1)/2$
Objectives

Proportion of misclassified nodes under $\pi$: $\varepsilon^\pi(n, T)$

1. Find conditions on $p$, $q$, $n$, and $T$ for accurate solvability
   -- there exists an algorithm detecting clusters accurately, i.e.,
   \[
   \lim_{n \to \infty} \mathbb{E}[\varepsilon^\pi(n, T)] = 0.
   \]

2. How can adaptive sampling improve accurate solvability?

3. Design optimal algorithms
Solvability and the SBM

**Not Solvable**
\[ \varepsilon(n) = \frac{1}{2} \]

**Solvable**
\[ \varepsilon(n) < \frac{1}{2} \]

**Accurately Solvable**
\[ \varepsilon(n) \to 0 \]
Solvability and the SBM

\[ \varepsilon(n) = \frac{1}{2} \]

\[ \varepsilon(n) < \frac{1}{2} \]

\[ \varepsilon(n) \rightarrow 0 \]
Sparse graphs

• SB model: sparse case \( p = \frac{a}{n}, \quad q = \frac{b}{n} \)

**Theorem** (Mossel-Neeman-Sly 2012)

If \( a - b < \sqrt{2(a + b)} \), then the problem is not solvable.

Conjectured by Decelle-Krzakala-Moore-Zdeborova 2012

**Theorem** (Massoulie 2013)

If \( a - b > \sqrt{2(a + b)} \), then there exists an algorithm leading to clusters that are positively correlated with the true clusters.
SBM

\[ \frac{n(p - q)^2}{p + q} = 2 \]

Not Solvable
\[ \varepsilon(n) = 1/2 \]

Solvable
\[ \varepsilon(n) < 1/2 \]

Accurately solvable
\[ \varepsilon(n) \rightarrow 0 \]
Solvability – Generic Random Sampling

\[ \varepsilon^\pi(n, T) = \frac{1}{2}. \]

\[ \varepsilon^\pi(n, T) < \frac{1}{2}. \]

Accurately solvable
\[ \varepsilon^\pi(n, T) \to 0 \]
Solvability – Adaptive Sampling

\[ \epsilon^\pi(n, T) < \frac{1}{2} \]

\[ \epsilon^\pi(n, T) \rightarrow 0 \]
Fundamental limits

- Random sampling:

\[
\kappa_1(n, T) = T \frac{2(n - 2)}{n(n - 1)} \min\{KL(q, p), KL(p, q)\}
+ 2\sqrt{\frac{4T(n - 2)}{n(n - 1)}} \left[\min\{q, 1 - p\} \left(\log \frac{p(1 - q)}{q(1 - p)}\right)^2 + \left(\log(\min\{\frac{p}{q}, \frac{1 - q}{1 - p}\})\right)^2\right]
\]

**Theorem**  Under Random sampling strategy S1 or S2, for any clustering algorithm \(\pi\), we have:

\[
\mathbb{E}[\varepsilon^{\pi}(n, T)] \geq \frac{1}{8} \exp(-\kappa_1(n, T)),
\]
Fundamental limits

- Non-adaptive random sampling -- necessary conditions for asymptotically accurate detection:
  \[ \frac{T}{n} = \omega(1), \quad \frac{T}{n} \min(KL(q, p), KL(p, q)) = \omega(1), \]

- Dense interaction: \( p, q = \Theta(1) \)
  \[ T(p - q)^2 / n = \omega(1) \]

- Sparse interaction: \( p, q = o(1) \)
  \[ T(p - q)^2 / (pn) = \omega(1) \]
Fundamental limits

• Adaptive sampling:

**Theorem**  For asymptotically accurate detection, we need:

\[
\min\{p, 1 - q\} \frac{T}{n} = \Omega(1) \quad \text{and} \quad \frac{T}{n} \max(KL(q, p), KL(p, q)) = \omega(1).
\]

• Example: \( p = \frac{a \log n}{n}, \quad q = \frac{b \log n}{n} \)

  – Non-adaptive sampling: \( \frac{T}{n} = \omega\left(\frac{n}{\log(n)}\right) \)

  – Adaptive sampling: \( \frac{T}{n} = \omega\left(\frac{n}{\log(n)}\right) \)
Fundamental limits

• Adaptive sampling:

**Theorem**  For asymptotically accurate detection, we need:

$$\min\{p, 1 - q\} \frac{T}{n} = \Omega(1) \quad \text{and} \quad \frac{T}{n} \max(KL(q, p), KL(p, q)) = \omega(1).$$

• Example: \( p = \frac{\log n}{n} \quad q = \frac{\sqrt{\log n}}{n} \)

  – Non-adaptive sampling: \( \frac{T}{n} = \omega\left(\frac{n}{\log(n)}\right) \)

  – Adaptive sampling: \( \frac{T}{n} = \Omega\left(\frac{n}{\log(n)}\right) \)
Algorithms for non-adaptive sampling

• Spectral algorithms (extension of Coja-Oghlan’s algorithm)

1. From random samples, build an observation matrix
2. Trimming (remove nodes with too many interactions)
3. Spectral decomposition (find the largest eigenvalues and corresponding eigenvectors)
4. Greedy improvement (for each node compare the number of interactions with the various clusters)
Performance

**Theorem** Assume that:

\[
\frac{(p - q)^2 \alpha T}{p} \frac{\alpha T}{n} = \omega(1), \quad \frac{(p - q)^2 \alpha T}{p} \frac{\alpha T}{n} \geq \log\left(\frac{p}{n} \frac{T}{n}\right).
\]

Then with high probability:

\[
\varepsilon^{SP}(n, T) \leq 8 \exp\left(-\frac{(p - q)^2 \alpha T}{20p} \frac{\alpha T}{n}\right).
\]

- The algorithm is asymptotically accurate under the necessary conditions for accurate detection in the case of random sampling.
- The necessary conditions for accurate detection are tight!
Algorithms for adaptive sampling

• Spatial coupling idea: find reference kernels and build the clusters from these kernels

1. Kernels: select n/log(n) nodes and use T/5 samples to classify these nodes (using the previous spectral algorithm)

2. Select one of remaining nodes. Sample T/3n pairs between the selected nodes to each kernel. Classify the node.

3. Repeat 2. until no remaining node or budget
Performance

**Theorem** Assume that:

\[
\frac{(p - q)^2}{p + q} \frac{T}{n} = \Omega(1), \quad \frac{T}{n} \max(KL(q, p), KL(p, q)) = \omega(1).
\]

Then with high probability:

\[
\varepsilon^{ASP}(n, T) \leq \exp \left( -\frac{T}{6n} (KL(q, p) + KL(p, q)) \right).
\]

- The algorithm is asymptotically accurate under the necessary conditions for accurate detection in the case of adaptive sampling.
- The necessary conditions for accurate detection are tight!
Solvability – Generic Random Sampling

\[ \varepsilon^\pi(n, T) = \frac{1}{2}. \]

\[ \varepsilon^\pi(n, T) < \frac{1}{2}. \]

\[ \frac{T}{n} \min(KL(p, q), KL(q, p)) = \omega(1) \]
Solvability – Adaptive Sampling

\[ T \frac{\max(KL(p, q), KL(q, p))}{n} = \omega(1) \]

\[ \varepsilon^\pi(n, T) < \frac{1}{2} \]

\[ \varepsilon^\pi(n, T) \to 0 \]
SBM: A Numerical Example

- \( n = 4000 \)

\begin{align*}
  p &= 0.01, \quad q = 0.005 \\
  p &= 0.1, \quad q = 0.05
\end{align*}
Summary

• A generic sampling framework extending the SBM
• Necessary conditions for asymptotically accurate detection
• Asymptotically optimal joint sampling and clustering algorithms
  – Spectral method is optimal
• Arbitrary sample budget:
  – Quantify the impact of lack of information (some pair of nodes not observed)
  – Required budget for detection in very sparse regimes (circumventing the phase transition problem)
• Extensions
  – Beyond the SBM: different $p$’s in different clusters
  – Overlapping communities
Thanks!

seyoung.yun@inria.fr

alepro@kth.se