Exact Recovery by Semidefinite Programming in the Binary Stochastic Block Model with Partially Revealed Side Information

Mohammad Esmaeili, Hussein Saad, and Aria Nosratinia

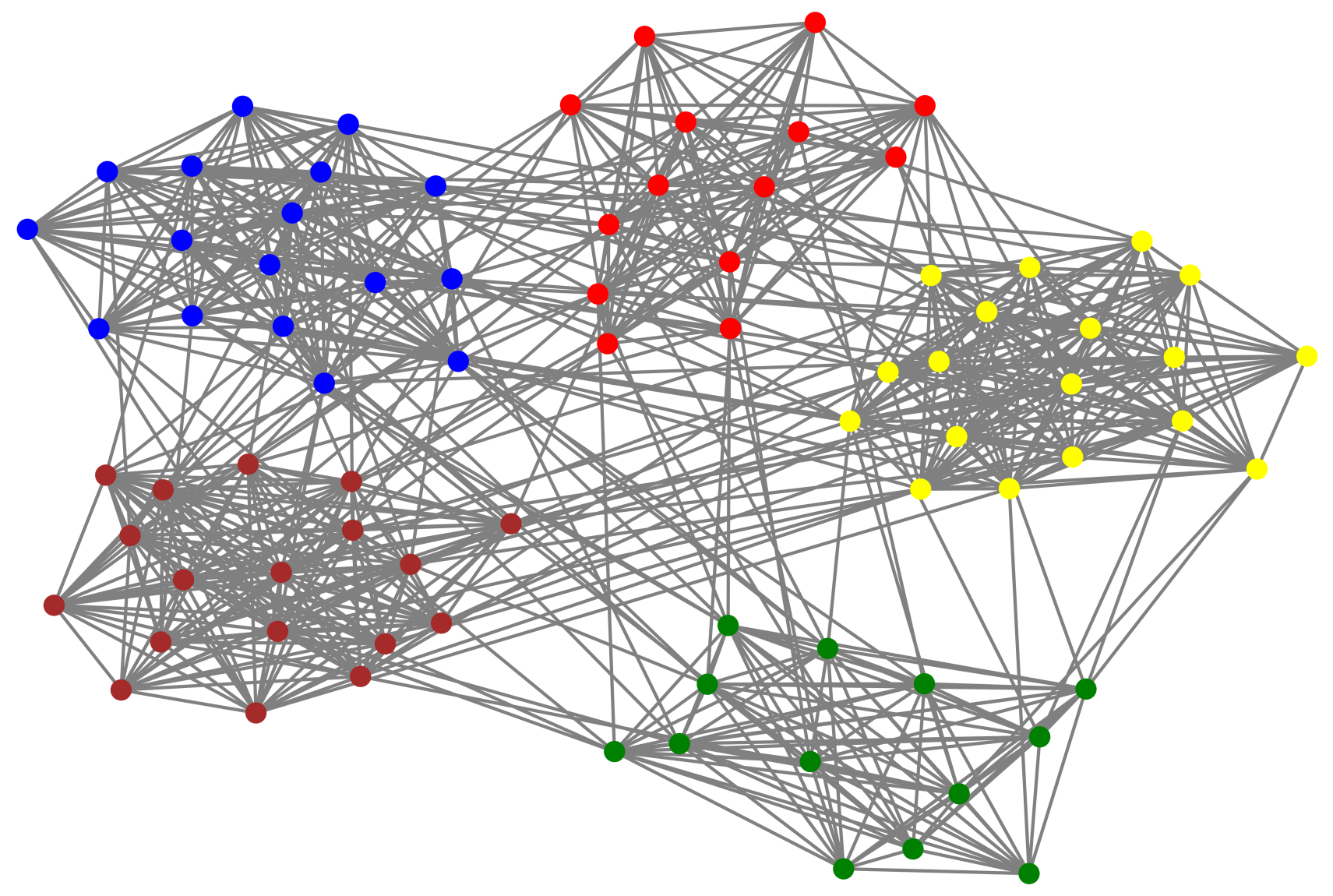
The University of Texas at Dallas

Abstract

Semidefinite programming has been shown to be both efficient and asymptotically optimal in solving community detection problems, as long as observations are purely graphical in nature. In this paper, we extend this result to observations that have both a graphical and a non-graphical component. We consider the binary censored block model with n nodes and study the effect of partially revealed labels on the performance of semidefinite programming. Essentially, we ask the question: do partially revealed labels help the semidefinite programming solution as much as they help the maximum likelihood solution? Our results are two fold. First, we show that partially revealed labels change the phase transition of exact recovery if and only if they grow no slower than $\Omega(\log(n))$. Second, we show that the semidefinite programming relaxation of maximum likelihood can achieve exact recovery down to the optimal threshold under partially revealed labels.

Community Detection with Side Information

- **Networks with community structure:**



- **Community detection problem:** Given a network (e.g., Facebook), exploit its structure to recover hidden communities (clusters).
- **Side Information:** In many practical inference problems, relevant information other than the graph is available (e.g., gender, location, income, etc.)
- **Our Contribution:**
  - Applying semidefinite programming to community detection with side information.
  - We calculate the perfect recovery threshold under the partially-revealed-label side information

Problem Formulation

- n nodes, Graph adjacency G
- Conditional edge probabilities $p = a\frac{\log n}{n}$ $q = b\frac{\log n}{n}$
- Node label $x_i \in \{\pm 1\}$, side information $y_i = x_i$ with prob. $1-\epsilon \in (0,1)$ for $i \in [n]$

- **Maximum likelihood detector:**

$$\begin{aligned} \hat{X} =& \arg\max_X \; X^T G X \\ & \text{subject to} \quad X \in \{\pm 1\}^n \\ & \qquad\qquad\quad X^T \mathbf{1} = 0 \\ & \qquad\qquad\quad X^T Y = Y^T Y. \end{aligned} \tag{1}$$

- Semidefinite relaxation arises from $X^T G X = \mathbf{Tr}(XX^T G)$, substituting $X^T X \to Z$, and relaxing the rank-1 constraint on Z to a positivity constraint.
- **Semidefinite Program:**

$$\begin{aligned} \widehat{Z}_{SDP} =& \arg\max_Z \; \langle Z, G \rangle \\ & \text{subject to} \quad Z \succeq 0 \\ & \qquad\qquad\quad Z_{ii} = 1, \quad i \in [n] \\ & \qquad\qquad\quad \langle Z, W \rangle = (Y^T Y)^2 \\ & \qquad\qquad\quad \langle Z, \mathbf{J} \rangle = 0. \end{aligned} \tag{2}$$

Exact Recovery Conditions

Exact recovery metric:

$$\lim_{n\to\infty} \mathbb{P}(e = 0) = 1$$

Lemma: If there exists $D^* = \text{diag}(d_i^*) \geq 0$, $\lambda^* \in \mathbb{R}$, and $\mu^* \in \mathbb{R}$ such that $S^* = D^* - G + \lambda^* \mathbf{J} + \mu^* W$ satisfies $S^* \succeq 0$, $\lambda_2(S^*) \geq 0$, and $S^* X^* = 0$, then $\widehat{Z}_{SDP} = Z^*$ is the unique solution to (2).

❶ Showing $\widehat{Z}_{SDP} = Z^*$ is optimal

❷ Showing $\widehat{Z}_{SDP} = Z^*$ is unique

❸ Showing $S^* \succeq 0$ and $\lambda_2(S^*) > 0$ with probability at least $1 - o(1)$. In other words, it suffices to show

$$\mathbb{P}\left\{ \inf_{V \perp X^*, \|V\|=1} V^T S^* V > 0 \right\} \geq 1 - o(1).$$

Theorem

Under the binary symmetric stochastic block model and partial revealed labels side information, if

$$\begin{cases} (\sqrt{a} - \sqrt{b})^2 > 2 & \text{when } \log \epsilon = o(\log n) \\ (\sqrt{a} - \sqrt{b})^2 + 2\beta > 2 & \text{when } \log \epsilon = -(\beta + o(1)) \log n \end{cases},$$

then as $n \to \infty$, $\min_{Z^* \in \mathcal{Z}_n} \mathbb{P}(\widehat{Z}_{SDP} = Z^*) \geq 1 - o(1)$.

- **Converse:** Obtained by Saad and Nosratinia (JSTSP'2018)

Simulation Results

Results are asymptotic, but give insight even on relatively small graphs.

With Side Information					Without Side Information				
a	b	β	n	Error Probability	a	b	β	n	Error Probability
3	1	0.2	100	2.1×10^{-2}	3	1	0.0	100	1.4×10^{-1}
3	1	0.2	200	1.6×10^{-2}	3	1	0.0	200	1.2×10^{-1}
3	1	0.2	300	1.3×10^{-2}	3	1	0.0	300	1.0×10^{-1}
3	1	0.2	400	1.1×10^{-2}	3	1	0.0	400	9.5×10^{-2}
3	1	0.2	500	1.0×10^{-2}	3	1	0.0	500	9.1×10^{-2}
3	1	0.8	100	2.7×10^{-4}	6	1	0.0	100	6.7×10^{-5}
3	1	0.8	200	1.5×10^{-4}	6	1	0.0	200	4.0×10^{-5}
3	1	0.8	300	8.6×10^{-5}	6	1	0.0	300	2.3×10^{-5}
3	1	0.8	400	6.4×10^{-5}	6	1	0.0	400	1.1×10^{-5}
3	1	0.8	500	5.0×10^{-5}	6	1	0.0	500	5.0×10^{-6}

Discussion

- Semidefinite programming relaxation of the maximum likelihood estimator can achieve exact recovery down to the optimal threshold
- Partially revealed labels change the phase transition of exact recovery if and only if the information they provide grows no slower than $\Omega(log(n))$.
- The asymptotic results of this paper can also shed light on the performance at finite n.
- The scope of the results were significantly expanded subsequent to the submission of the paper to include noisy-label and general side information, as well as other graph models.