Side Information in Recovering a Single Community: Information Theoretic Limits

Hussein Saad and Aria Nosratinia
Department of Electrical Engineering,
The University of Texas at Dallas, Richardson, TX 75083-0688, USA,
E-mail: hussein.saad@utdallas.edu; aria@utdallas.edu.

***Abstract*—We consider a generalization of the community detection problem, where for inference we have access to an additional non-graphical side information about the label of each node. Specifically, we study the effect of side information on the information theoretic limits of recovering a hidden community of size K inside a graph consisting of n nodes with $K = o(n)$. Two asymptotic recovery metrics are considered, namely, weak recovery and exact recovery. We consider side information in the form of a vector of length F, where each component in the vector belongs to a set with finite cardinality and the components are generated i.i.d. according to some conditional distribution. We assume F to be a function of n, while the conditional probabilities of the outcomes are independent of n. We show when and by how much side information can improve the information theoretic limits of weak and exact recovery by providing tight necessary and sufficient conditions for both weak and exact recovery. Furthermore, we show that, under certain conditions, any algorithm achieving weak recovery can also achieve exact recovery if followed by a local voting procedure.**

***Index Terms*—Community detection, Stochastic block model, Side information, Information Theoretic Limits.**

I. Introduction

Detecting communities (or clusters) in graphs is a fundamental problem that has been studied in various fields, statistics [1], [2], computer science [3], [4], [5], [6] and theoretical statistical physics [7]. In this paper, we consider the problem of finding a single sub-graph (community) hidden in a large graph, where the community size is much smaller than the graph size. Examples for application of finding a hidden community problem are fraud activity detection in [8], [9] and correlation mining [10].

Several models are now being studied for random graphs that exhibit a community structure; a survey can be found in [11]. A widely used model in the context of community detection is the stochastic block model (SBM) [12]. In this paper, we use the stochastic block model for one community, also known as the planted dense sub-graph model [13], [14], [15], [16]. The stochastic block model for one community is characterized by the following parameters: n is the number of nodes in the graph, K is the size of the community, p is the probability of having an edge between any two nodes inside the community, and q is the probability of having an edge otherwise. The goal is to recover/detect the hidden community upon observing the graph edges.

The problem of finding a hidden community upon observing *only* the graph has been studied in [13], [14], [15]. The information theoretic limits of *weak recovery* (expected number of misclassified nodes is $o(K)$) and *exact recovery* (probability of correctly recovering all the labels converges to one) of a hidden community have been established in [14]. The limit of the belief propagation (BP) algorithm for weak recovery and exact recovery has been also established in [15], [13].

The literature on community detection has, for the most part, concentrated on purely graphical observations. However, in many practical applications, non-graphical relevant information is available that can aid the inference. For example, social networks such as Facebook and Twitter have access to much information other than the graph edges. A citation network has the authors names, keywords, and abstracts of papers. This paper presents new results on the utility of side information in community detection, in particular shedding light on the conditions under which side information can improve the information theoretic limits of weak and exact recovery of a hidden community.

A few results have recently appeared in the literature regarding the community detection problem in the presence of additional (non-graphical) information. In the context of detecting two or more communities: (1) [17] studied the effect of noisy label information on the performance of a belief propagation algorithm. (2) Cai *et. al* [18] demonstrated regimes for BP to achieve weak recovery upon observing a vanishing fraction of labels. Neither of [17], [18] includes a converse, so they do not establish the phase transition. (3) [19], [20] studied the effect of side information in scalar and vector form on the phase transition of exact recovery for the binary symmetric communities. (4) [21] studied the effect of side information only in vector form with a specific length on the phase transition of exact recovery for more than two communities. (5) In the context of detecting a single hidden community: Kadavankandy *et al.* [16] studied the effect of noisy labels with vanishing noise on the performance of belief propagation in detecting a single-community.

The present work is motivated by the following observations in the problem of detecting a hidden community. The effect of side information on the information theoretic limits of weak and exact recovery has not been available to date, Most of the work done on two or more communities assumed only binary side information, but practical scenarios motivate the study of more general side information whose alphabet does not match the number/identity of communities. Moreover, the work done on two or more communities focused only on information limits of exact recovery and not weak recovery.

The contributions of the paper in the problem of detecting a hidden community and can be summarized as follows:

- For the information theoretic limits of weak recovery with side information, we provide necessary and sufficient conditions that are tight. We also show some cases where side information will not improve the information theoretic limits of weak recovery. Moreover, we show that under the same sufficient conditions, weak recovery is achievable even when the size of the community is random and unknown.
- For the information theoretic limits of exact recovery with side information, we provide necessary and sufficient conditions that are tight. We also show that any algorithm that achieves weak recovery can achieve also exact recovery under certain conditions. Furthermore, we provide an example for a popular model of side information, namely, noisy labels and compare the new limits for exact recovery to the ones obtained without side information.

II. System Model and Definitions

We consider the stochastic block model for a hidden community with side information. Let $\mathcal{G}(n, K, p, q)$ denote the ensemble of graphs with n nodes, a hidden community C^* with size $|C^*| = K$ and an edge between a pair of nodes is drawn with probability p if both nodes are in C^* and probability q otherwise. Finally, for each node $i \in \{1, \cdots, n\}$, a vector of length F containing side information is observed. We assume all components of the vector have the same finite alphabet and are i.i.d. conditioned on the labels.

We define $P = Bernoulli(p)$, $Q = Bernoulli(q)$ and $G(V, E)$ to be a graph realization of $\mathcal{G}(n, K, p, q)$. Let x_i denote the label of node $i \in \{1, \cdots, n\}$, where $x_i = 1$ if $i \in C^*$ and $x_i = 0$ if $i \notin C^*$ and define $\boldsymbol{x}^* \in \{0,1\}^n$ denote the vector of the true labels. Let $y_{i,f}, f \in \{1, \cdots, F\}$ be the f^{th} component of the vector of side information of node i and define $\boldsymbol{y}_f$ be a vector of length n denoting the side information of all the nodes for component f. For a given node i and component f, let V and U denote the probability distribution of $y_{i,f}$, conditioned on $i \in C^*$ and $i \notin C^*$, respectively. Define $L_2(i, f) = \log(\frac{V}{U}(y_{i,f}))$ to be the log-likelihood ratio of $y_{i,f}$ with respect to V and U. Finally, let G_{ij} denote a random variable denoting the existence of an edge between nodes i and j in the graph and define $L_1(i, j) = \log(\frac{P}{Q}(G_{ij}))$ denote the log-likelihood ratio of edge G_{ij} with respect to P and Q[1].

In this paper, we focus on the problem of recovery the hidden community upon observing $G(V, E)$ and the vector of nodes side information by $\boldsymbol{y_1}, \cdots, \boldsymbol{y_F}$. Let $\hat{\boldsymbol{x}}(G, \boldsymbol{y_1}, \cdots, \boldsymbol{y_F})$ be an estimator of $\boldsymbol{x}^*$ given $G(V, E)$ and $\boldsymbol{y_1}, \cdots, \boldsymbol{y_F}$. The following assumption and definitions of recovery metrics are used throughout the paper:

Assumption 1. *As $n \to \infty$: $K \to \infty$ such that $K = o(n)$, $p \geq q$, $\frac{p}{q} = \theta(1)$, $\limsup_{n\to\infty} p < 1$ and L_2 is bounded.*

An estimator $\hat{\boldsymbol{x}}(G, \boldsymbol{y_1}, \cdots, \boldsymbol{y_F})$ is said to achieve exact recovery if, as $n \to \infty$, $\mathbb{P}(\hat{\boldsymbol{x}} = \boldsymbol{x}^*) \to 1$. Also, an estimator $\hat{\boldsymbol{x}}(G, \boldsymbol{y_1}, \cdots, \boldsymbol{y_F})$ is said to achieve weak recovery if, as $n \to \infty$, $\frac{d(\hat{\boldsymbol{x}}, \boldsymbol{x}^*)}{K} \to 0$ in probability, where $d(.,.)$ denotes the hamming distance. It was shown in [14] that the latter definition is equivalent to the existence of an estimator $\hat{\boldsymbol{x}}$ such that $\mathbb{E}[d(\hat{\boldsymbol{x}}, \boldsymbol{x}^*)] = o(K)$. We will use this equivalence throughout this paper.

Finally, for the ease of notation we define the following:

$$\psi_{QU}(t, m_1, m_2) \triangleq m_1 \log(\mathbb{E}_Q[e^{tL_1}]) + m_2 \log(\mathbb{E}_U[e^{tL_2}]) \tag{1}$$

$$\psi_{PV}(t, m_1, m_2) \triangleq m_1 \log(\mathbb{E}_P[e^{tL_1}]) + m_2 \log(\mathbb{E}_V[e^{tL_2}]) \tag{2}$$

$$E_{QU}(\theta, m_1, m_2) \triangleq \sup_{t\in[0,1]} t\theta - \psi_{QU}(t, m_1, m_2) \tag{3}$$

$$E_{PV}(\theta, m_1, m_2) \triangleq \sup_{t\in[-1,0]} t\theta - \psi_{PV}(t, m_1, m_2) \tag{4}$$

Remark 1. ***Due to space limitations, most of the proofs are provided online [22] for the reviewers.***

III. Weak Recovery

Theorem 1. *Suppose Assumption 1 holds. If*

$$(K-1)D(P||Q) + FD(V||U) \to \infty \text{ and}$$
$$\liminf_{n\to\infty} (K-1)D(P||Q) + 2FD(V||U) > 2\log(\frac{n}{K}) \tag{5}$$

then weak recovery is possible. If weak recovery is possible, then:

$$(K-1)D(P||Q) + FD(V||U) \to \infty \text{ and}$$
$$\liminf_{n\to\infty} (K-1)D(P||Q) + 2FD(V||U) \geq 2\log(\frac{n}{K}) \tag{6}$$

Remark 2. *Theorem 1 shows that if F grows with n slow enough, e.g., F is fixed and independent of n or $F = o(\log(\frac{n}{K}))$, then the information theoretic limits are the same as the ones characterized in [14] without side information. This holds because by assumption L_2 is bounded which implies that $D(V||U)$ is bounded, and hence,* (5) *and* (6) *can be simplified to $KD(P||Q) \to \infty$ and $\liminf_{n\to\infty}(K - 1)D(P||Q) \geq 2\log(\frac{n}{K})$, which are the same limits as in [14].*

Remark 3. *If the components of the vector of side information are not i.i.d. conditioned on the labels, we conjecture that a necessary and sufficient conditions for weak recovery would be:*

$$(K-1)D(P||Q) + \sum_{f=1}^{F} D(V_f||U_f) \to \infty \text{ and}$$

$$\liminf_{n\to\infty} (K-1)D(P||Q) + 2\sum_{f=1}^{F} D(V_f||U_f) > 2\log(\frac{n}{K})$$

where V_f and U_f are the conditional probability distribution of each component in the vector of side information.

[1] Throughout the paper, we use L_1 and L_2 to denote the random variables of the log-likelihood ratio of the graph edge and the outcome of side information.

Proof.

Necessary Conditions: Provided online [22].

Sufficient Conditions: The sufficient conditions for weak recovery is derived for the maximum likelihood (ML) estimator. Note that although ML is optimal for exact recovery by definition, it is not optimal for weak recovery. Before we go into the proof, we need to derive the rule of ML for recovering a hidden community with side information.

For any subsets $S, T \subset \{1, \cdots, n\}$, define:

$$e_1(S,T) \triangleq \sum_{(i<j):(i,j)\in(S\times T)\cup(T\times S)} L_1(i,j) \tag{7}$$

$$e_2(S) \triangleq \sum_{i\in S}\sum_{f=1}^{F} L_2(i,f) \tag{8}$$

Using these definitions, it is easy to write the rule of the maximum likelihood using the log-likelihood function of the observations $G, \boldsymbol{y_1}, \cdots, \boldsymbol{y_F}$ given the labels $\boldsymbol{x}$ as follows:

$$\hat{C}_{ML} = \arg\max_{C\subset\{1,\cdots,n\}} \{e_1(C,C) + e_2(C) \ : |C| = K\} \tag{9}$$

Let $R = |\hat{C}_{ML} \cap C^*|$ denote the intersection of $\hat{C}_{ML}$ with the true hidden community C^*. Thus, we can write the difference as $|\hat{C}_{ML} \triangle C^*| = 2(K - R)$, and hence, to show that ML achieve weak recovery, it is sufficient to show that there exists $\epsilon = o(1)$, such that $\mathbb{P}\big(R \leq (1-\epsilon)K\big) \leq o(1)$. To show this, we need to bound the error event of ML. The complete proof is provided online [22]. □

A. Sufficient Conditions for Random Community Size

In this section, we show that the conditions of Theorem 1 is sufficient for weak recovery even when $|C^*|$ is random. This is needed for the proof of the sufficient conditions for exact recovery. We will continue using $\hat{C}$ as the estimator defined in (9) although here it is not actually ML estimator because $|C^*|$ need not be K.

Lemma 1. *Suppose that Assumption 1 and the conditions of Theorem 1 hold. Moreover, assume the size of the community is random such that:*

$$\mathbb{P}(||C^*| - K| \leq \frac{K}{\log(K)}) \geq 1 - o(1) \tag{10}$$

then,

$$\mathbb{P}(\frac{|\hat{C}\triangle C^*|}{K} \leq 2\epsilon + \frac{1}{\log(K)}) \geq 1 - o(1) \tag{11}$$

where $\epsilon = \frac{1}{\sqrt{\min(\log(K),(K-1)D(P||Q)+FD(V||U))}} = o(1)$.

Proof. Provided online [22]. □

IV. Exact Recovery

Theorem 2. *Suppose Assumption 1 holds. If* (5) *and the following hold:*

$$\liminf_{n\to\infty} E_{QU}(\log(\frac{n}{K}), K, F) > \log(n) \tag{12}$$

then exact recovery is possible. If exact recovery is possible, then (5) *and the following hold:*

$$\liminf_{n\to\infty} E_{QU}(\log(\frac{n}{K}), K, F) \geq \log(n) \tag{13}$$

TABLE I
Algorithm for exact recovery.

Algorithm 1
1: Input: $n \in \mathbb{N}$, $K > 0$, distributions P, Q, V, U, G, $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_F$, $\delta \in (0,1) : n\delta, \frac{1}{\delta} \in \mathbb{N}$.
2: Partition $\{1, \cdots, n\}$ into $\frac{1}{\delta}$ subsets S_k of size $n\delta$ each, $k = 1, \cdots, \frac{1}{\delta}$.
3: Weak Recovery: For each $k = 1, \cdots, \frac{1}{\delta}$, let G_k and $\boldsymbol{y}_1^k, \cdots, \boldsymbol{y}_F^k$ denote the sub-graph and the part of side information, respectively, restricted to $\{1, \cdots, n\}\backslash S_k$. Run an estimator capable of weak recovery with inputs $(n(1-\delta), \lceil K(1-\delta)\rceil, P, Q, V, U, G_k, \boldsymbol{y}_1^k, \cdots, \boldsymbol{y}_F^k$ and let $\hat{C}_k$ be the output.
4: Voting Procedure: For each $k = 1, \cdots, \frac{1}{\delta}$, compute $r_i = (\sum_{j\in\hat{C}_k} L_1(ij)) + \sum_{f=1}^{F} L_2(i,f)$ for all $i \in S_k$ and return $\tilde{C}$: the set of K indices in $\{1, \cdots, n\}$ with the largest value of r_i.

Remark 4. *Theorem 2 shows if F grows with n slow enough, e.g., F is fixed and independent of n or $F = o(K)$, the information theoretic limits are the same as the ones characterized in [14] without side information. To see this, note that since $t \in [0,1]$ and L_2 are bounded, this implies that $E_{QU}(\log(\frac{n}{K}), K, F) = K(1+o(1))\sup_{t\in[0,1]} \frac{1}{K}\log(\frac{n}{K}) - \log(\mathbb{E}_Q[e^{tL_1}])$ which is the same limit characterized in [14] for exact recovery without side information.*

Proof of Sufficient Conditions of Theorem 2.

The sufficient conditions are derived for Algorithm I that achieves exact recovery in two steps. First, we apply an algorithm that achieves weak recovery for a random community size, like the one presented in Section III-A. Then, a local voting procedure is performed. This shows that exact recovery is achieved for any algorithm that can achieve weak recovery on a random community size followed by the voting procedure.

The following theorem gives sufficient conditions under which Algorithm I achieves exact recovery. The proof of the sufficiency part of Theorem 2 will be given after the proof of the theorem.

Theorem 3. *Let $\tilde{C}$ be the output of Algorithm I using an estimator for weak recovery $\hat{C}_k$ such that as $n \to \infty$:*

$$\mathbb{P}\big(|\hat{C}_k \triangle C_k^*| \leq \delta K \text{ for } 1 \leq k \leq \frac{1}{\delta}\big) \to 1 \tag{14}$$

where $C_k^* = C^* \cap (\{1, \cdots, n\}\backslash S_k)$. Suppose that (12) *and Assumption 1 hold. Then, $\mathbb{P}(\tilde{C} = C^*) \to 1$ as $n \to \infty$.*

Proof. To prove Theorem 3, we need the following Lemma.

Lemma 2. *Suppose that* (12) *and Assumption 1 hold. Let $\{W_l\}$ and $\{\tilde{W}_l\}$ denote sequence of i.i.d. copies of L_1 under P and Q, respectively. Also, for any node i, let Z and $\tilde{Z}$*

denote $\sum_{f=1}^{F} L_2(i,f)$ *under* V *and* U*, respectively. Then, for sufficiently small, but constant,* δ *and* $\gamma = \frac{\log(\frac{n}{K})}{K}$*:*

$$\mathbb{P}\Big(\sum_{l=1}^{K(1-\delta)} \tilde{W}_l + \tilde{Z} \geq K(1-\delta)\gamma \Big) = o(\frac{1}{n}) \quad (15)$$

$$\mathbb{P}\Big(\sum_{l=1}^{K(1-2\delta)} W_l + \sum_{l=1}^{\delta K} \tilde{W}_l + Z \leq K(1-\delta)\gamma \Big) = o(\frac{1}{K}) \quad (16)$$

Proof. Provided online [22] □

Now we prove Theorem 3. From the statement of the theorem, we have the conditions of Lemma 2 satisfied, and hence, (15) and (16) hold. Define the event $F = \{|\hat{C}_k \triangle C_k^*| \leq \delta K\}$. On F, we have:

$$\begin{aligned} |\hat{C}_k \cap C_k^*| &\geq |\hat{C}_k| - |\hat{C}_k \triangle C_k^*| = \lceil K(1-\delta) \rceil - |\hat{C}_k \triangle C_k^*| \\ &\geq K(1-2\delta) \end{aligned}$$

Thus, on the event F, r_i (from Algorithm I) for $i \in C^*$ is stochastically greater than or equal to $(\sum_{l=1}^{K(1-2\delta)} W_l) + (\sum_{l=1}^{K\delta} \tilde{W}_l) + Z$. For $i \notin C^*$, r_i has the same distribution as $(\sum_{l=1}^{K(1-\delta)} \tilde{W}_l) + \tilde{Z}$. Thus, by Lemma 2, with probability converging to 1, $r_i > K(1-\delta)\gamma$ for all $i \in C^*$ and $r_i < K(1-\delta)\gamma$ for all $i \notin C^*$. Hence, $\mathbb{P}(\tilde{C} = C^*) \to 1$ as $n \to \infty$. This concludes the proof of Theorem 3. □

To complete the proof of the sufficient conditions of Theorem 2, it suffices to verify (14) when $\hat{C}_k$ for each k is the ML estimator for C_k^* based on observing G_k and $\boldsymbol{y}_1^k, \cdots, \boldsymbol{y}_F^k$. The distribution of $|C_k^*|$ is obtained by sampling the indicies of the original graph without replacement. Hence, from [14], we have for any convex function ϕ: $\mathbb{E}[\phi(|C_k^*|)] \leq \mathbb{E}[\phi(Binomial(n(1-\delta), \frac{K}{n}))]$. Therefore, Chernoff bounds for Binomial($n(1-\delta), \frac{K}{n}$) also holds for $|C_k^*|$. Thus, we have:

$$\begin{aligned} &\mathbb{P}\big(||C_k^*| - (1-\delta)K| \geq \frac{K}{\log(K)}\big) \\ &\leq \mathbb{P}\big(|Binomial(n(1-\delta), \frac{K}{n}) - (1-\delta)K| \geq \frac{K}{\log(K)}\big) \\ &\overset{(a)}{\leq} o(1) \end{aligned} \quad (17)$$

where (a) holds by Chernoff bounds which states that for $X \sim Binomial(n,p)$: $\mathbb{P}(X \geq (1+\eta)np) \leq e^{-\eta^2 \frac{np}{3}}$ and $\mathbb{P}(X \leq (1-\eta)np) \leq e^{-\eta^2 \frac{np}{2}}$ for all $\eta \in [0,1]$.

Since (5) holds, then we have: $\liminf_{n\to\infty} \lceil (1-\delta)K \rceil D(P||Q) + 2FD(V||U) > 2\log(\frac{n}{K})$ for sufficiently small δ. This result and (17) show that ML achieves weak recovery with K replaced $\lceil (1-\delta)K \rceil$ in Lemma 1. Thus, for any $1 \leq k \leq \frac{1}{\delta}$,

$$\mathbb{P}\big(\frac{|\hat{C}_k \triangle C_k^*|}{K} \leq 2\epsilon + \frac{1}{\log(K)}\big) \geq 1 - o(1) \quad (18)$$

with $\epsilon = o(1)$. Since δ is constant, thus, by the union bound over all $1 \leq k \leq \frac{1}{\delta}$, we have:

$$\mathbb{P}\big(\frac{|\hat{C}_k \triangle C_k^*|}{K} \leq 2\epsilon + \frac{1}{\log(K)} \quad \forall 1 \leq k \leq \frac{1}{\delta}\big) \geq 1 - o(1) \quad (19)$$

Since $\epsilon = o(1)$, the desired (14) holds.

□

Proof of Necessary Conditions of Theorem 2.

The outline is as follows: First, assuming that the maximum likelihood (ML) detector exactly recovers the community, i.e. $\mathbb{P}(\text{ML fails}) = o(1)$, we find conditions for the failure of ML in recovery the community. Then, we express these conditions, generally for bounded and unbounded side information, in terms of large deviations inequalities. Next, we relate these inequalities to the parameters of the graph and the side information. Recall that ML is optimal for exact recovery since C^* is chosen uniformly.

Let $i_o = \arg\min_{i \in C^*} e_1(i, C^*) + \sum_{f=1}^{F} L_2(i,f)$ and define $F_M = \{\min_{i \in C^*} e_1(i,C^*) + \sum_{f=1}^{F} L_2(i,f) \leq \max_{j \notin C^*} e(j, C^* \backslash \{i_o\}) + \sum_{f=1}^{F} L_2(j,f)\}$. Define $\tilde{C} = C^* \backslash \{i_o\} \cup j$ for $j \notin C^*$. Then, using (9), we have:

$$\begin{aligned} &e_1(\tilde{C},\tilde{C}) + e_2(\tilde{C}) - e_1(C^*,C^*) + e_2(C^*) = \\ &(e(j, C^* \backslash \{i_o\}) + \sum_{f=1}^{F} L_2(j,f)) - (e_1(i,C^*) + \sum_{f=1}^{F} L_2(i,f)) \overset{(a)}{\geq} 0 \end{aligned} \quad (20)$$

where (a) holds by assuming F_M happens. Hence, F_M implies the failure of ML. Then, we have:

$$\mathbb{P}(F_M) \leq \mathbb{P}(\text{ML fails}) \overset{(a)}{=} o(1) \quad (21)$$

where (a) holds by assumption that ML achieves exact recovery. The following lemma characterizes general necessary conditions for exact recovery for both bounded and unbounded side information.

Lemma 3. *Suppose Assumption 1 holds. Let* $\{W_l\}$ *and* $\{\tilde{W}_l\}$ *denote sequence of i.i.d. copies of* L_1 *under* P *and* Q*, respectively. Also, for any node i, let* Z *and* $\tilde{Z}$ *denote a copy of* $\sum_{f=1}^{F} L_2(i,f)$ *if node i belong to* C^* *or does not belong to* C^**, respectively. Let* $K_o \to \infty$ *such that* $K_o = o(K)$*. Then, for an estimator* $\hat{C}$ *to achieve exact recovery, i.e.,* $\mathbb{P}(\hat{C} = C^*) \to 1$*, there exists a threshold* θ_n *such that for all sufficiently large* n*:*

$$\mathbb{P}\Big(\sum_{l=1}^{K-K_o} W_l + Z \leq (K-1)\theta_n - \tilde{\theta_n} \Big) \leq \frac{2}{K_o} \quad (22)$$

$$\mathbb{P}\Big(\sum_{l=1}^{K-1} \tilde{W}_l + \tilde{Z} \geq (K-1)\theta_n \Big) \leq \frac{1}{n-K} \quad (23)$$

where $\tilde{\theta_n} = (K_o - 1)D(P||Q) + 6\sigma$ *for* $\sigma^2 = K_o var_P(L_1)$ *and* $var_P(L_1)$ *denote the variance of* L_1 *under* P*.*

Proof. Provided online [22] □

In view of Lemma 3, we now need to show for which parameters of the graph and side information there exists θ_n such that (22) and (23) hold. We will show that if (12) does not hold, then there does not exist θ_n such that (22) and (23)

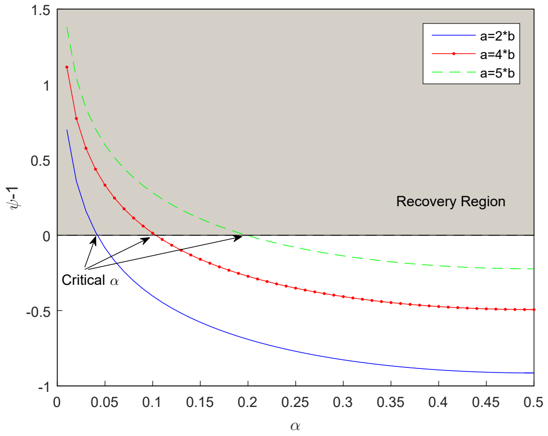

Fig. 1. Exact recovery threshold, $\psi-1$ for different values of α at $c = b = 1$.

hold simultaneously. This is shown in the extended version provided online [22].

□

A. Example

To illustrate our results, we compare our results to the information theoretic limits of exact recovery without side information in the following regime:

$$K = \frac{cn}{\log(n)}, \; q = \frac{b\log^2(n)}{n}, \; p = \frac{a\log^2(n)}{n} \tag{24}$$

for fixed positive $c, a \geq b$ as $n \to \infty$. In the above regime, we have $KD(P||Q) \approx \log(n)$, and hence, weak recovery is always information theoretic possible without side information, and by extension, with side information. Moreover, exact recovery is information theoretic possible if and only if:

$$\sup_{t\in[0,1]} \; tc(a-b) + bc - bc(\frac{a}{b})^t > 1 \tag{25}$$

We focus on side information with two possible outcomes, where each component of the vector of side information is the true label passed though a binary symmetric channel with cross-over probability α. Thus, exact recovery with side information is possible if and only if:

$$\sup_{t\in[0,1]} tc(a-b) + bc - bc(\frac{a}{b})^t - \\ \frac{F}{\log(n)} \log((1-\alpha)^t\alpha^{(1-t)} + (1-\alpha)^{(1-t)}\alpha^t) > 1 \tag{26}$$

The last displayed equation shows that if $F = o(\log(n))$, then exact recovery is possible if and only if (25) holds, and hence, side information does not improve the information theoretic limits of exact recovery. If F is not $o(\log(n))$, note that $\log((1-\alpha)^t\alpha^{(1-t)} + (1-\alpha)^{(1-t)}\alpha^t)$ is always negative since $t \in [0,1]$, and hence, (26) $\geq$ (25).

Let $F = \log(n))$ and $\psi = \sup_{t\in[0,1]} tc(a-b)+bc-bc(\frac{a}{b})^t - \log((1-\alpha)^t\alpha^{(1-t)}+(1-\alpha)^{(1-t)}\alpha^t)$. Figure 1 shows the curve $\psi-1$ for different values of α. From the figure, it can be shown that side information helps BP to achieve recovery in regimes where it was known to fail without side information.

References

[1] A. Zhang and H. Zhou, "Minimax rates of community detection in stochastic block models," *arXiv:1507.05313*, Nov. 2015.

[2] P. J. Bickel and A. Chen, "A nonparametric view of network models and newmangirvan and other modularities," *Proceedings of the National Academy of Sciences*, vol. 106, no. 50, pp. 21 068–21 073, 2009.

[3] J. X. Y. Chen, "Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices," *ICML, In proceedings of*, Feb. 2014.

[4] A. Coja-oghlan, "Graph partitioning via adaptive spectral techniques," *Comb. Probab. Comput.*, vol. 19, no. 2, pp. 227–284, Mar. 2010.

[5] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[6] J. Chen and B. Yuan, "Detecting functional modules in the yeast proteinprotein interaction network," *Bioinformatics*, vol. 22, no. 18, pp. 2283–2290, Sep. 2006.

[7] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications," *Phys. Rev. E*, vol. 84, p. 066106, Dec. 2011.

[8] A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos, "Copycatch: Stopping group attacks by spotting lockstep behavior in social networks," in *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web*, 05 2013, pp. 119–130.

[9] D. H. Chau, S. Pandit, and C. Faloutsos, "Detecting fraudulent personalities in networks of online auctioneers," in *Proceedings of the 10th European Conference on Principle and Practice of Knowledge Discovery in Databases*, ser. PKDD'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 103–114.

[10] H. Firouzi, B. Rajaratnam, and A. H. III, "Predictive correlation screening: Application to two-stage predictor design in high dimension," in *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, C. M. Carvalho and P. Ravikumar, Eds., vol. 31. Scottsdale, Arizona, USA: PMLR, 29 Apr–01 May 2013, pp. 274–288. [Online]. Available: http://proceedings.mlr.press/v31/firouzi13a.html

[11] S. Fortunato, "Community detection in graphs," *arXiv:0906.0612v2*, Jan. 2010.

[12] E. Abbe, A. Bandeira, and G. Hall, "Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms," *arXiv:1503.00609*, Mar. 2015.

[13] A. Montanari, "Finding one community in a sparse graph," *arXiv:1502.05680v2*, Jul. 2015.

[14] B. Hajek, Y. Wu, and J. Xu, "Information limits for recovering a hidden community," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 4729–4745, Aug 2017.

[15] J. X. B. Hajek, Y. Wu, "Recovering a hidden community beyond the spectral limit in $o(|e|\log^*|v|)$ time," *arXiv:1510.02786v2*, Jun. 2017.

[16] A. Kadavankandy, K. Avrachenkov, L. Cottatellucci, and R. Sundaresan, "The power of side-information in subgraph detection," *arXiv:1611.04847v3*, Mar. 2017.

[17] E. Mossel and J. Xu, "Local algorithms for block models with side information," in *ACM Conference on Innovations in Theoretical Computer Science*, ser. ITCS '16. New York, NY, USA: ACM, 2016, pp. 71–80. [Online]. Available: http://doi.acm.org/10.1145/2840728.2840749

[18] T. Tony Cai, T. Liang, and A. Rakhlin, "Inference via message passing on partially labeled stochastic block models," *arXiv:1603.06923v1*, Mar. 2016.

[19] H. Saad, A. Abotabl, and A. Nosratinia, "Exact recovery in the binary stochastic block model with binary side information," in *Allerton Conference on Communications, Control, and Computing*, Oct 2017.

[20] ——, "Side information in the binary stochastic block model: Exact recovery," *arXiv:1708.04972v1*, Aug 2017.

[21] A. R. Asadi, E. Abbe, and S. Verd, "Compressing data on graphs with clusters," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 1583–1587.

[22] H. Saad and A. Nosratinia, "Side information in recovering a single community: Information theoretic limits," 2017. [Online]. Available: http://www.utdallas.edu/%7Ehussein.saad/IT.pdf