Understanding the Success of Graph-based Semi-Supervised Learning using Partially Labelled Stochastic Block Model

Avirup Saha[1*], **Shreyas Sheshadri**[2*], **Samik Datta**[2], **Niloy Ganguly**[1], **Disha Makhija**[2] and **Priyank Patel**[2]
[1]Indian Institute of Technology, Kharagpur
[2]Flipkart Internet Private Limited
avirupsaha@iitkgp.ac.in, {s.shreyas, samik.datta}@flipkart.com, niloy@cse.iitkgp.ac.in, {disha.makhiji, priyank.patel}@flipkart.com

Abstract

With the proliferation of learning scenarios with an abundance of instances, but limited amount of high-quality labels, semi-supervised learning algorithms came to prominence. Graph-based Semi-Supervised Learning (G-SSL) algorithms, of which Label Propagation (LP) is a prominent example, are particularly well-suited for these problems. The premise of LP is the existence of homophily in the graph, but beyond that nothing is known about the efficacy of LP. In particular, there is no characterisation that connects the structural constraints, volume and quality of the labels to the accuracy of LP. In this work, we draw upon the notion of recovery from the literature on community detection, and provide guarantees on accuracy for partially-labelled graphs generated from the Partially-Labelled Stochastic Block Model (PLSBM). Extensive experiments performed on synthetic data verify the theoretical findings.

1 Introduction

In several practical learning scenarios – e.g., e-commerce, genomics, image search, speech recognition and natural language processing – out of the large number of available instances, only a handful are typically labelled. The cost associated with labelling necessitates employing semi-supervised learning (SSL) in such cases. Graph-based SSL (G-SSL) is an influential approach to SSL that seeks to exploit *homophily* [McPherson *et al.*, 2001]), in order to estimate the labels of the unlabelled instances. First introduced in [Zhu *et al.*, 2003], and subseauently popularised in [Zhu, 2005], Label Propagation (LP), enjoyed widespread empirical success in G-SSL and sparked a series of extensions: [Talukdar and Crammer, 2009], [Chakrabarti *et al.*, 2014], [Stretcu *et al.*, 2019].

Sadly, theoretical understanding of the efficacy of these algorithms has been lacking. Till date, to the best of our knowledge, only [Yamaguchi and Hayashi, 2017] attempts to characterise the performance of LP, by proving equivalence with the problem of community recovery in PLSBM, an extension of the Stochastic Block Model (SBM) [Abbe *et al.*, 2015] to the case of partially-labelled graphs. However, beyond equivalence, [Yamaguchi and Hayashi, 2017] does not quantify the efficacy in terms of well-understood notions of predictive performance, such as accuracy. We address this gap by noting an equivalence between the notion of accuracy and that of community recovery [Abbe and Sandon, 2015; Ke and Honorio, 2018]. Moreover, we characterise the conditions on graph structure, volume and quality of available labels that lead to upper- and lower-bounds on accuracy for LP.

Our contributions are listed below (see Table 1 for a quick reference):

- We extend the notion of success and failure beyond equivalence ([Yamaguchi and Hayashi, 2017]) by connecting it to 100% accuracy, $\geq 100(1-\vartheta)\%$ accuracy, and $\leq 100(1-\delta)\%$ accuracy, respectively, for arbitrary $\vartheta, \delta \in (0,1)$. See §3.2 for details.
- We characterise the condition on graph structure, label volume and quality, that leads to $\leq 100(1-\delta)\%$ accuracy by connecting it to the notion of failure of recovery ([Ke and Honorio, 2018]) (§4).
- We state similar necessary conditions that lead to 100% accuracy by connecting it to the notion of exact recovery ([Saad and Nosratinia, 2018]) (§5).
- We state a well-positioned conjecture on similar necessary conditions that lead to $100(1-\vartheta)\%$ accuracy by connecting it to the notion of partial recovery ([Yun and Proutiere, 2014]), for the case of $\vartheta = s/n$, where $s = o(n)$ (§6).
- We extend the failure of recovery result obtained in [Ke and Honorio, 2018] to the case of partially-labelled SBM with $K \geq 2$. See Theorem 2.
- We provide extensive experimental evidence support the theory (including the conjecture) on synthetic partially-labelled graphs. See §7 for details.

Proofs and additional experimental evidences are available in the supplementary material [1].

[*]Equal contribution

[1]See following URL: http://bit.ly/PLSBM_IJCAI20

Domain	Source	Equivalence Guarantee	Failure Guarantee	Success Guarantee	
		$\widehat{X_{DLP}} = \widehat{X_{MAP}}$	Acc. $\leq 100(1-\delta)\%$	Acc. $\geq 100(1-\frac{s}{n})\%$	Acc. $= 100\%$
G-SSL	Present Work	☑ (§2.3: $PLSBM(n, \epsilon_l, \cdots)$)	☑ (§4: PLSBM, $K \geq 2$)	☑ (§6: PLSBM)	☑ (§5: PLSBM)
	[Yamaguchi and Hayashi, 2017]	☑ ($PLSBM(n, n_l, \cdots)$)	☒	☒	☒
Community Detection	[Ke and Honorio, 2018]	☒	☑ (SBM, $K = 2$)	☒	☒
	[Yun and Proutiere, 2014]	☒	☒	☑ (SBM)	☒
	[Saad and Nosratinia, 2018]	☒	☒	☒	☑ (m-ary SBM)

Table 1: A compendium of novel theoretical results presented in this work.

2 Background

We begin by recounting the equivalence theorem derived in [Yamaguchi and Hayashi, 2017].

Let (A, X) denote a partially-labelled graph on n nodes, $[n]$, where the first n_l nodes, $[n_l]$, carry a label in $[K]$. $X \in \{0,1\}^{n\times K}$ denotes the labels on nodes, when available. For nodes $i \in [n_l]$, X_i is the one-hot representation of the corresponding label $x_i \in [K]$, whereas for the unlabelled nodes, $j \in [n] \setminus [n_l]$, $X_j = \mathbf{0}$ ($\mathbf{0}$ denotes the K-dimensional zero vector). $A \in \{0,1\}^{n\times n}$ denotes the adjacency matrix, where $A_{i,j} = 1$ indicates the presence of edge between nodes i and j (we assume that there are no self-loops: $A_{i,i} = 0, \forall i \in [n]$). Graph-based semi-supervised learning (G-SSL) attempts at recovering the label, $\widehat{X}$, of the hitherto unlabelled nodes, $[n] \setminus [n_l]$.

2.1 Discrete Label Propagation

Definition 1 (DLP [Yamaguchi and Hayashi, 2017])**.** *DLP estimates the true labels of the nodes, $\widehat{X_{DLP}} \in \{0,1\}^{n\times K}$, as a solution to the following (combinatorial) optimisation problem:*

$$\arg\min_{Y\in\{0,1\}^{n\times K}} \frac{1}{2}\sum_{i=1}^{n} \|X_i - Y_i\|_2^2 + \frac{\lambda}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{i,j}\,\|Y_i - Y_j\|_2^2 \tag{1}$$

A relaxation, $Y \in [0,1]^{n\times K}$, and a subsequent rounding leads to the original Label Propagation algorithm [Zhu, 2005]. It is worth noting that the solution to LP, $\widehat{X_{LP}} \in \{0,1\}^{n\times K}$, is a worse optimiser of Equation 1 than $\widehat{X_{DLP}}$ because of the intermediate relaxation and rounding.

2.2 Partially-Labelled Stochastic Block Model

Definition 2 (**PLSBM** [Yamaguchi and Hayashi, 2017])**.** *$PLSBM(n, n_l, \alpha, K, Q)$ posits a generative model over the family of partially-labelled graphs, (A, X), defined as follows:*

1. *Each node $i \in [n]$ is assigned a label, $x_i \in [K]$, (corresponding to its community membership) with uniform probability $1/K$, yielding a set of balanced communities in expectation* [2].
2. *The entries of the adjacency matrix, A, are generated i.i.d., with $\Pr(A_{i,j} = 1 \mid x_i, x_j) = Q_{x_i,x_j}$, where $Q \in (0,1)^{K\times K}$ is a symmetric matrix. In other words, $A_{i,j} \mid X \sim \mathrm{Bern}(\cdot|Q_{x_i,x_j})$, i.i.d..*
3. *For the nodes in $[n_l]$, (potentially noisy) labels, $x' \in [K]$, are revealed i.i.d. as follows: $\Pr(x'_i = x_i) = 1-\alpha$ and $\Pr(x'_i = j) = \frac{\alpha}{K-1}, \forall j \neq x_i$. The rest of the nodes, indexed $\{n_l + 1, \cdots, n\}$ remain unlabelled.*

In §4, §5, §6, we use another equivalent formulation of PLSBM ([Saad and Nosratinia, 2018]), $PLSBM(n, \epsilon_l, \alpha, K, Q)$, with the following modification: instead of revealing the label of the nodes in $[n_l]$, we reveal labels (with or without noise, as before) for each node in $[n]$ with probability ϵ_l. With $\epsilon_l = \frac{n_l}{n}$, in expectation, labels for n_l nodes are revealed under $PLSBM(n, \epsilon_l, \alpha, K, Q)$.

Given $(A, X') \sim PLSBM(n, n_l, \alpha, K, Q)$, we let $\widehat{X_{MAP}} = \arg\max_X \Pr(X|A, X'; \epsilon_l, \alpha, Q)$ denote the MAP estimate of the (latent) membership vector X (obtained through an inference procedure such as variational EM [Yamaguchi and Hayashi, 2017]). Similarly, we let $\widehat{X_{MLE}} = \arg\max_X \Pr(X, A, X'; \epsilon_l, \alpha, Q)$ denote the corresponding MLE estimate. Following [Abbe *et al.*, 2015], we note that $\widehat{X_{MAP}} = \widehat{X_{MLE}}$, when $\gamma = 1/K\mathbf{1}$.

2.3 Equivalence of DLP and PLSBM

The following result builds a bridge between the two seemingly disparate disciplines of graph-based semi-supervised learning and graph generative models:

Theorem 1 ([Yamaguchi and Hayashi, 2017])**.** *Given a realisation $(A, X') \sim PLSBM(n, n_l, \alpha, K, \mu\mathbf{I} + \nu(\mathbf{1}\mathbf{1}^T - \mathbf{I}))$, and number of nodes in each cluster, $n_k, \forall k \in [K]$, $\widehat{X_{DLP}}(\lambda) = \widehat{X_{MAP}}(\alpha, \mu, \nu)$, if the following conditions hold:*

1. *If $\lambda > 0$, then $\mu > \nu$; else if $\lambda < 0$, then $\mu < \nu$; else if $\lambda = 0$, then $\mu = \nu$; and*
2. $\lambda \ln \frac{\alpha(K-1)}{1-\alpha} = \ln \frac{\mu(1-\nu)}{\nu(1-\mu)}$

Discussion. It is not hard to see that Theorem 1 holds for $PLSBM(n, \epsilon_l, \alpha, K, Q)$ as well, since the log-likelihood is identical to that of $PLSBM(n, n_l, \alpha, K, Q)$, barring a constant. This extension of Theorem 1, with $\widehat{X_{MAP}}$ substituted with $\widehat{X_{MLE}}$, will be used in the subsequent sections (§4, §5, §6).

[2] PLSBM is typically defined with $x_i \sim \mathrm{Mult}(\cdot \mid \gamma)$, i.i.d. However, the restricted notion with $\gamma = 1/K\mathbf{1}$ suffices for the present work.

In what follows, we strengthen Theorem 1 by further connecting the notion of *recovery* in the context of community detection (see §3.1) to that of *accuracy* in the graph-based semi-supervised learning (see §3.2). §3.3 provides an outline of the theoretical results we have obtained.

3 Preparation

We begin with a brief refresher of the relevant definitions from the literature ([Abbe and Sandon, 2015] and [Saad and Nosratinia, 2018]) on the recovery of communities.

3.1 Recovery of Communities in PLSBM

Exact Recovery

The correct community memberships for all the nodes must be inferred with probability tending to 1 as $n \to \infty$ in order to attain *exact recovery*. In other words, $\lim_{n\to\infty} \Pr(X = \widehat{X}) = 1$, where the probability is defined by the generative model of $PLSBM(n, \epsilon_l, \alpha, K, Q)$.

[Saad and Nosratinia, 2018] furnishes exact recovery results under a more general generative model of partially-labelled graphs (*m-ary side information*) for $K = 2$, which we adapt in §5 to the case of PLSBM.

Partial Recovery

The correct community membership for at least s nodes must be inferred with probability tending to 1 as $n \to \infty$ in order to attain *partial recovery*. In other words, $\lim_{n\to\infty} \Pr(\ell^{0\text{-}1}(X, \widehat{X}) \leq s/n) = 1$, where $\ell^{0\text{-}1}(X, \widehat{X}) = \frac{1}{n}\sum_{i\in[n]} \mathbf{1}_{\{x_i \neq \hat{x}_i\}}$ is the usual 0-1 loss counting the fraction of errors. To the best of our knowledge, no result exists for the partially-labelled graphs that we study. [Yun and Proutiere, 2014] furnishes a partial recovery result for the SBM with $K = 2$, when $s = o(n)$. We dwell upon this topic in §6, where we put forth a well-reasoned conjecture based on this result.

Failure of Recovery

The correct community membership of at least $100(1-\delta)\%$ of the nodes (in expectation) must be incorrectly inferred in order to attain *failure of recovery*, for any given $\delta \in [0, 1]$. In other words, for every node $\Pr(\widehat{x}_i \neq x_i) \geq \delta$. [Ke and Honorio, 2018] studies failure of recovery from an information-theoretic perspective and derives algorithm-agnostic results for SBM with $K = 2$. In §4, we refine the result for partially-labelled graphs (PLSBM) and extend it to the case of $K \geq 2$. We refer the interested reader to [Abbe, 2017] for a comprehensive survey on this topic, in addition to the above works.

3.2 Success and Failure of G-SSL

We now connect the notion of *accuracy* in the G-SSL parlance to that of recovery in community detection, and lend precise definitions to the notions of its success and failure:

Total success. It is easy to see that exact recovery leads to 100% accuracy – this is the strongest possible notion of success for G-SSL. Note that this guarantee on accuracy only holds with probability tending to 1 as $n \to \infty$.

Partial success. Since partial recovery places an upper-bound on the 0-1 loss, $\ell^{0\text{-}1}(X, \widehat{X}) \leq s/n$, it leads to a lower-bound, $100(1 - \frac{s}{n})\%$, accuracy. Note that this guarantee on accuracy only holds with probability tending to 1 as $n \to \infty$.

Failure. Failure to recover, guaranteeing $\Pr(\widehat{x}_i \neq x_i) \geq \delta, \forall i \in [n]$, puts a lower-bound on the 0-1 loss, $\ell^{0\text{-}1}(X, \widehat{X}) \geq \delta$ in expectation. Therefore, it translates to an upper-bound of $100(1-\delta)\%$ on accuracy. Note that this guarantee holds in expectation, irrespective of the value of n.

3.3 Outline of the Theoretical Results

In what follows, we connect the three notions of recovery (§3.1) to the corresponding notions of accuracy (§3.2). §4 employs information-theoretic tools deployed in [Ke and Honorio, 2018] to prove an algorithm-agnostic non-asymptotic upper-bound on accuracy; however, it extends the result presented therein to the case of PLSBM, and to the case of $K \geq 2$ communities. §5 adapts the result presented in [Saad and Nosratinia, 2018] to the case of PLSBM. §6 puts forth a well-reasoned conjecture that proposes to extend the result presented in [Yun and Proutiere, 2014] to the case of PLSBM. At a high level, all the three results state one structural constraint (involving Q) for each regime of the volume and quality of labels (governed by ϵ_l and α, respectively) that is either necessary for recovery (or, in some cases, sufficient as well). The corresponding guarantee on accuracy of LP is then made by invoking Theorem 1.

We refer the reader to Table 1 for a summary of the results presented in the remainder of this work, and a highlight of our contributions both w.r.t. the literature on the success/failure of G-SSL ([Yamaguchi and Hayashi, 2017]) in the top two rows (shaded in gray), as well as the literature on community detection ([Ke and Honorio, 2018], [Saad and Nosratinia, 2018], [Yun and Proutiere, 2014]) that we adapt in the bottom two rows (below the rule). Inside the bracket, we provide a pointer to the corresponding section, as well as highlight the difference with the corresponding paper that we adapt (marked with ☑ in the same column).

4 Failure of Recovery & Up to $100(1-\delta)\%$ Accuracy

In this section, we aim to provide a sufficient condition for the failure of recovery and, thus, furnish an upper-bound on the accuracy. We choose to prove algorithm-agnostic and non-asymptotic bounds and resort to information-theoretic techniques employed in [Ke and Honorio, 2018]. Note that their proof applies to SBM *with no side-information* and is limited to the $K = 2$ setting — two limitations that we alleviate in this work. Note that we do not consider bounds specific to any inference technique such as MLE ([Saad and Nosratinia, 2018] and [Abbe *et al.*, 2015]) or those that are asymptotic ([Abbe and Sandon, 2015]). We state the theorem below:

Theorem 2. *Given* $(A, X') \sim PLSBM(n, \epsilon_l, \alpha, K, Q)$, *where* Q *is a weakly assortative matrix as defined by* $\min_k (Q_{k,k} - \max_{k'} Q_{k,k'}) \geq 0$, *every algorithm fails to estimate the true membership,* X, *with* $\Pr(\widehat{X} \neq X) \geq \delta$, *for*

any given $\delta \in [0,1]$ *if:*

$$D_{\max}^{\text{Bern}}(Q) \leq \frac{2(1-\delta)\log K + 4\epsilon_l}{n-1} - \frac{2\epsilon_l}{n-1}\left(\alpha(K-1) + \frac{1}{\alpha(K-1)}\right) - \frac{\log 2}{\binom{n}{2}} \quad (2)$$

Where $\text{Bern}(\cdot|p)$ *is the Bernoulli distribution with parameter* $p \in [0,1]$ *(and support in* $\{0,1\}$*) and* $D_{\max}^{\text{Bern}}(Q) = \max_{i,j\in[K]} \text{KL}(\text{Bern}(Q_{i,i})||\text{Bern}(Q_{i,j}))$

Sketch of Proof. We make use of Fano's inequality, $\Pr(\widehat{X} \neq X) \geq 1 - \frac{I(A,X';X)+\log 2}{K \log n}$, to lower-bound $\Pr(\widehat{X} \neq X)$. Next, we upper-bound the mutual information, $I(A, X'; X)$, relying on a combination of independence/additivity, uniform bounding. □

Discussion. Theorem 2 characterises the parameter space of $PLSBM(n, \epsilon_l, \alpha, K, Q)$ that leads to a failure of inference by furnishing a structural constraint (involving Q) for a regime of labelling defined by the volume (ϵ_l) and quality (α) of labels. Note that $\alpha(K-1) + \frac{1}{\alpha(K-1)}$ increases with decrease in α (signifying gradual increase in the quality of the labels). Hence with increasing $\epsilon_l > 0$ (signifying a gradual transition in the volume of labels from rare to abundant), the product term $\epsilon_l(\alpha(K-1) + \frac{1}{\alpha(K-1)})$ increases, thus reducing the RHS and destroying the inequality unless $D_{\max}^{\text{Bern}}(Q)$ is very low. In other words, abundance of high-quality (low noise) labels helps community recovery, as expected.

Discussion. $(A, X') \sim PLSBM(n, \epsilon_l, \alpha, 2, a\frac{\log n}{n}\mathbf{I} + b\frac{\log n}{n}(\mathbf{1}\mathbf{1}^T - \mathbf{I}))$, and the weak assortativity constraint, $a \geq b$, allows us to simplify and bound the term $D_{max}^{\text{Bern}}(Q)$ in Equation 2 using the inequality, $\log x \leq x - 1$, to yield:

$$\frac{n-1}{n}\frac{(a-b)^2}{b\left(1 - b\frac{\log n}{n}\right)} \leq \frac{1}{\log n}\left(2(1-\delta)\log 2 + 2\epsilon_l(2 - (\alpha + \frac{1}{\alpha})) - \frac{2\log 2}{n}\right) \quad (3)$$

Note that in the limit $n \to \infty$, the RHS tends to 0, whereas the LHS tends to $\frac{(a-b)^2}{b}$. Therefore, the only way to guarantee failure is $a = b$, which turns the intra- and inter-community connectivity identical making the communities completely indistinguishable.

Discussion. Demanding a pronounced failure will require us to set $1 - \delta$ low, so that the accuracy upper-bound is tightened. This will lower the RHS and demand an increasing degree of indistinguishability (by lowering $D_{\max}^{\text{Bern}}(Q)$) for failure of recovery.

Discussion. This information theoretic result holds regardless of the algorithm employed for exact recovery, one of which is the MLE algorithm which has been shown by [Yamaguchi and Hayashi, 2017] to be the same as LP [Bengio *et al.*, 2006] since the MAP objective functions are equivalent. This means that, under the condition presented in Theorem 2, LP will misclassify any given node on the graphs generated by PLSBM with high probability, and therefore should not be applied at all.

Corollary 1 (Up to $100(1-\delta)\%$ accuracy of LP)**.** *By Theorem 2, if the volume and the quality of labels (governed by* ϵ_l *and* α*, respectively) fail to balance the structural separation (governed by* $D_{\max}^{\text{Bern}}(Q)$*), every algorithm would fail to infer* $\widehat{X}$ *effectively (since* $\Pr(\widehat{X} \neq X) \geq 1 - \delta$*). Therefore, MLE will not succeed, as well: i.e.* $\Pr(\widehat{X_{MLE}} \neq X) \geq 1 - \delta$*. By Theorem 1 and its extension,* $\widehat{X_{DLP}} = \widehat{X_{MLE}}$ *for* $PLSBM(n, n_l, \alpha, K, \mu\mathbf{I} + \nu(\mathbf{1}\mathbf{1}^T - \mathbf{I}))$ *with given community sizes,* $\{n_k \mid k \in [K]\}$*, for an appropriate choice of* λ*. Together, it guarantees that in expectation, the labels inferred by LP will have atmost* $100(1-\delta)\%$ *accuracy.*

Corollary 1 strengthens the precision of Theorem 1 considerably by lending a precise notion to the term failure, rather than capturing it by a far weaker notion of $\widehat{X_{DLP}} \neq \widehat{X_{MLE}}$ ([Yamaguchi and Hayashi, 2017]).

5 Exact Recovery & 100% Accuracy

We now characterise the regime in which LP attains 100% accuracy on partially-labelled networks generated from *symmetric* PLSBM with $K = 2$. To this end, we adapt the result on exact recovery ([Saad and Nosratinia, 2018]) for a more general model of partial labelling, *m-ary side-information*, to the case of symmetric PLSBM.

We commence by formally defining the restricted PLSBM setting i.e., PLSBM with two symmetric clusters.

Definition 3 (Symmetric PLSBM with $K = 2$)**.** *Symmetric PLSBM is* $PLSBM(n, \epsilon_l, \alpha, 2, a\frac{\log n}{n}\mathbf{I} + b\frac{\log n}{n}(\mathbf{1}\mathbf{1}^T - \mathbf{I}))$*, where each of the two clusters are of size* $n/2$*.*

Definition 4 (m-ary SBM)**.** *Instead of just* 2 *symbols, as in the case with symmetric PLSBM with* $K = 2$*, m-ary side-information encodes* x_i' *with an alphabet of size* m*:* $[m]$*. The crossover,* $\Pr(x_i' \mid x_i) = M_{x_i, x_i'}$*, is governed by a* stochastic matrix *(each row sums to* 1*),* $M \in \mathbb{R}^{2\times m}$*.*

It is easy to see that symmetric PLSBM with $K = 2$ can be equivalently viewed as an m-ary side-information with $m = 3$, where the alphabet is $\{-1, +1, 0\}$. This enables us to formulate the following theorem:

Theorem 3. *Given* (A, X') *sampled from a symmetric PLSBM, if MLE exactly recovers the true membership* X*, then:*

1. $\left(\sqrt{a} - \sqrt{b}\right)^2 > 2$*, if* $(\epsilon_l, \alpha) \in \mathcal{R}_1(\epsilon_l, \alpha)$*,*
2. $\left(\sqrt{a} - \sqrt{b}\right)^2 + 2\beta > 2$*, with* $\beta > 0$*, if* $(\epsilon_l, \alpha) \in \mathcal{R}_2(\epsilon_l, \alpha; \beta)$*,*
3. $\eta(a, b, |\beta_1|) + 2\beta_2 > 2$*, with* $0 < |\beta_1| < T\frac{(a-b)}{2}$ *and* $\beta_2 \geq 0$*, if* $(\epsilon_l, \alpha) \in \mathcal{R}_3(\epsilon_l, \alpha; \beta_1, \beta_2)$*,*

Where $T = \log\frac{a}{b}$*,* $\eta(a, b, \beta) = a + b + \beta - \frac{2\gamma}{T} + \frac{\beta}{T}\log\left(\frac{\gamma+\beta}{\gamma-\beta}\right)$ *and* $\gamma = \sqrt{\beta^2 + abT^2}$*.*

Before we provide a sketch of the proof, we need to define the regimes, $\mathcal{R}_1(\epsilon_l, \alpha)$, $\mathcal{R}_2(\epsilon_l, \alpha; \beta)$ and $\mathcal{R}_3(\epsilon_l, \alpha; \beta_1, \beta_2)$, of label volume and quality precisely. Intuitively, a low value of ϵ_l (equivalently, a low value of $\log \frac{1}{1-\epsilon_l}$, since $\epsilon_l \in [0, 1]$) signifies rarity in labels. Similarly, a too high ($\alpha \approx 1$) or a too low ($\alpha \approx 0$) value of α would be informative - since it would either favour the incorrect label always, or would favour the correct label. Since $\log \frac{\alpha}{1-\alpha}$ attains its maximum at $\alpha = 1/2$, this term captures the lack of information. Since Theorem 3 is inherently asymptotic, we need to bound these quantities as n grows: e.g. between $\log \frac{1}{1-\epsilon_l} = o(\log n)$ and $\log \frac{1}{1-\epsilon_l} = \theta(\log n)$, the former indicates a rare regime, whereas the latter signifies (relative) abundance. We operationalise these intuitions as follows:

1. $\mathcal{R}_1(\epsilon_l, \alpha)$: Either $\log(1-\epsilon_l) \in o(\log n)$, or, $\{\log \epsilon_l(1-\alpha), \log \epsilon_l \alpha, \log \frac{\alpha}{(1-\alpha)}\} \subset o(\log n)$.
2. $\mathcal{R}_2(\epsilon_l, \alpha; \beta)$: Either $\log(1-\epsilon_l) \in -\beta \log n + o(\log n)$, or, $\log \frac{\alpha}{(1-\alpha)} \in o(\log n)$ and $\{\log \epsilon_l(1-\alpha), \log \epsilon_l \alpha\} \subset -\beta \log n + o(\log n)$.
3. $\mathcal{R}_3(\epsilon_l, \alpha; \beta_1, \beta_2)$: $\log \frac{(1-\alpha)}{\alpha} \in \beta_1 \log n + o(\log n)$, and, if $\beta_1 > 0$ then $\log \epsilon_l(1-\alpha) \in -\beta_2 \log n + o(\log n)$ or if $\beta_1 < 0$ then $\log \epsilon_l \alpha \in -\beta_2 \log n + o(\log n)$

Sketch of Proof. In the case of symmetric PLSBM with $K = 2$, the symbol $y \in [m]$ is informative [3] if $f_1^y = \log \frac{M_{+1,y}}{M_{-1,y}}$ is $O(\log n)$, and is non-informative if $f_1^y = o(\log n)$. By Theorem 5 of [Saad and Nosratinia, 2018], this translates to the corresponding condition on $\log \frac{(1-\alpha)}{\alpha}$. Similarly, the symbol y is rare if $f_2^y = \log M_{+1,y}$ or $f_3^y = \log M_{-1,y}$ is $O(\log n)$ ($o(\log n)$ leads to the abundance). By Theorem 5 of [Saad and Nosratinia, 2018], this translates to the corresponding conditions on $\log(1-\epsilon_l)$, $\log \epsilon_l(1-\alpha)$, $\log \epsilon_l \alpha$. □

Corollary 2 (100% accuracy of LP)**.** *By Theorem 3, if $\widehat{X_{MLE}}$ exactly recovers the true labels X, the necessary conditions on the symmetric PLSBM parameters must hold and align with the three regimes, $\mathcal{R}_1(\epsilon_l, \alpha)$, $\mathcal{R}_2(\epsilon_l, \alpha; \beta)$ and $\mathcal{R}_3(\epsilon_l, \alpha; \beta_1, \beta_2)$. By Theorem 1 and its extension, $\widehat{X_{DLP}} = \widehat{X_{MLE}}$ for symmetric PLSBM, since $n_1 = n_2 = \frac{n}{2}$, for an appropriate choice of λ for each of the three regimes (e.g., $\lambda = \omega(\log n)$ when $(\epsilon_l, \alpha) \in \mathcal{R}_1(\epsilon_l, \alpha)$). Therefore, if $\widehat{X_{DLP}}$ recovers the true labels X with 100% accuracy, the symmetric PLSBM parameters have to lie in one of the three regimes* [4].

Corollary 2 strengthens Theorem 1 by precisely defining the notion of success beyond $\widehat{X_{DLP}} = \widehat{X_{MLE}}$ as found in [Yamaguchi and Hayashi, 2017]. §6 further generalises the notion of success.

[3] Our usage of the terms *informative* and *rare* follows [Saad and Nosratinia, 2018]

[4] Note that this is not sufficient. However, the experimental evidence presented in Figure 1 supports the characterisation.

6 Partial Recovery & at Least $100(1 - s/n)\%$ Accuracy

In this section, we present a well-reasoned *conjecture* for the partial recovery of PLSBM, based on the results on partial recovery of SBM, *without partial labelling*, presented in [Yun and Proutiere, 2014]. Figure 1 furnishes experimental evidences in its support. We formulate the following corollary based on Theorem 1 of [Yun and Proutiere, 2014]:

Corollary 3. *For a symmetric $SBM(n, 2, p\mathbf{I} + q(\mathbf{1}\mathbf{1}^T - \mathbf{I}))$ with $p = a\frac{\log n}{n}$ and $q = b\frac{\log n}{n}$, where $a > b$, the number of nodes with wrongly inferred community memberships under a certain spectral algorithm does not exceed s ($s = o(n)$) with high probability in the asymptotic sense, if:*

$$(\sqrt{a} - \sqrt{b})^2 \geq \frac{2a}{\log(a \log n)} + \frac{2 \log n/s}{\log n}$$

After scaling both sides with $\frac{\log n}{\log n/s}$, taking $\liminf_{n\to\infty}$ we recover the necessary condition, $(\sqrt{a} - \sqrt{b})^2 > 2$, of Theorem 3, barring the scaling factor on the LHS (and the strict inequality):

$$\liminf_{n\to\infty} \frac{\log n}{\log n/s} (\sqrt{a} - \sqrt{b})^2 \geq 2$$

This, further, tempts us to extend this observation to the case of a symmetric PLSBM with $K = 2$ in *all the three regimes of Theorem 3*, as well as to the MLE estimator (instead of the spectral algorithm that Corollary 3 relies upon), and put forth the following conjecture (see Figure 1 for the experimental evidence in its favour):

Conjecture 1. *Given (A, X') sampled from a symmetric PLSBM, if $\widehat{X_{MLE}}$ makes at most s ($s = o(n)$) mistakes, asymptotically, and with high probability, then:*

1. $\liminf_{n\to\infty} \frac{\log n}{\log n/s} (\sqrt{a} - \sqrt{b})^2 \geq 2$, *if* $(\epsilon_l, \alpha) \in \mathcal{R}_1(\epsilon_l, \alpha)$,
2. $\liminf_{n\to\infty} \frac{\log n}{\log n/s} \left((\sqrt{a} - \sqrt{b})^2 + 2\beta \right) \geq 2$, *with* $\beta > 0$, *if* $(\epsilon_l, \alpha) \in \mathcal{R}_2(\epsilon_l, \alpha; \beta)$,
3. $\liminf_{n\to\infty} \frac{\log n}{\log n/s} \left(\eta(a, b, |\beta_1|) + 2\beta_2 \right) \geq 2$, *with* $0 < |\beta_1| \leq T\frac{(a-b)}{2}$ *and* $\beta_2 \geq 0$, *if* $(\epsilon_l, \alpha) \in \mathcal{R}_3(\epsilon_l, \alpha; \beta_1, \beta_2)$,

Where $T = \log \frac{a}{b}$, $\eta(a, b, \beta) = a + b + \beta - \frac{2\gamma}{T} + \frac{\beta}{T} \log \left(\frac{\gamma+\beta}{\gamma-\beta} \right)$ and $\gamma = \sqrt{\beta^2 + abT^2}$.

Corollary 4 (At least $100(1 - s/n)\%$ accuracy of LP)**.** *Modulo Conjecture 1, if $\widehat{X_{MLE}}$ recovers X barring at most s ($s = o(n)$; e.g. $s = \sqrt{n}$) mistakes with high probability as $n \to \infty$, then the necessary conditions on the symmetric PLSBM parameters must hold and align with the three regimes, $\mathcal{R}_1(\epsilon_l, \alpha)$, $\mathcal{R}_2(\epsilon_l, \alpha; \beta)$ and $\mathcal{R}_3(\epsilon_l, \alpha; \beta_1, \beta_2)$. By Theorem 1 and its extension, $\widehat{X_{DLP}} = \widehat{X_{MLE}}$ for symmetric PLSBM, since $n_1 = n_2 = \frac{n}{2}$, for an appropriate choice of λ for each of the three regimes. Therefore, if $\widehat{X_{DLP}}$ recovers*

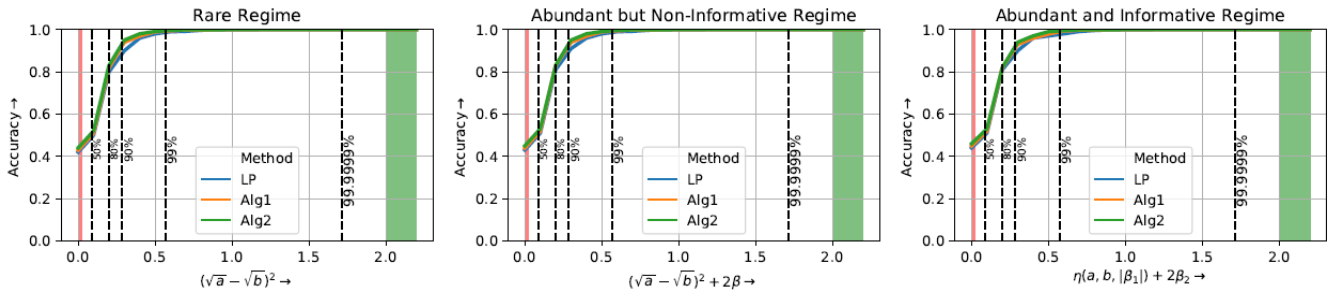

Figure 1: Variation of accuracy of LP, `Alg1` and `Alg2` with the MAC (see §7) for the three regimes of Theorem 3.

the true labels X with $100(1-{}^{s}/_{n})\%$ accuracy, the symmetric PLSBM parameters have to lie in one of the three regimes [5].

Conjecture 1 helps us relax the definition of success presented in §5, and helps us bolster Theorem 1 with a precise and widely applicable definition of success beyond $\widehat{X_{DLP}} = \widehat{X_{MLE}}$ as presented in [Yamaguchi and Hayashi, 2017].

7 Empirical Results

We consider synthetic graphs generated by a symmetric 2-cluster PLSBM for the three regimes of Theorem 3 with an assortativity matrix $Q = \left[\begin{smallmatrix} p & q \\ q & p \end{smallmatrix}\right]$, where $p = a\frac{\log n}{n}$ and $q = b\frac{\log n}{n}$, with $a > b$. We define a *modified assortativity coefficient* (MAC) as $(\sqrt{a}-\sqrt{b})^2, (\sqrt{a}-\sqrt{b})^2 + 2\beta, \eta(a, b, |\beta_1|) + 2\beta_2$ for the regimes, $\mathcal{R}_1(\epsilon_l, \alpha)$, $\mathcal{R}_2(\epsilon_l, \alpha; \beta)$ and $\mathcal{R}_3(\epsilon_l, \alpha; \beta_1, \beta_2)$, respectively. From Theorem 3 we see that as we vary the MAC in each regime from 0 to 2, we expect the accuracy of LP to rise gradually from 50% (or less) to 100%, with the intermediate values governed by Conjecture 1. The results shown in Figure 1 demonstrate that this is indeed the case, where in addition to LP [Zhu, 2005], we consider two other algorithms adapted for exact and partial recovery of PLSBM, for comparison: a) `Alg1` – exact recovery algorithm proposed in [Saad and Nosratinia, 2018] which employs a weak recovery algorithm ([Massoulié, 2014]) with a label flipping procedure to achieve exact recovery; b) `Alg2` – it is similar to `Alg1`, but with the Spectral Partition algorithm ([Yun and Proutiere, 2014]) swapped with [Massoulié, 2014] instead.

For generating these graphs, we set $n = 10^7$, $b = 1$, and varied a accordingly to get the desired values of the MAC. For regime $\mathcal{R}_1$, we set the number of partially labeled nodes, $n_l = \log n$ and for regimes $\mathcal{R}_2$ and $\mathcal{R}_3$, we set $n_l = \sqrt{n}$. For the regime $\mathcal{R}_2$, the label corruption probability, α was set to 0.1, whereas for the regime $\mathcal{R}_3$ it was set to 10^{-5}.

7.1 Observations

We mark the regimes of failure (acc. $\leq 100(1-\delta)\%, \delta = 0.5$) and success (acc. $= 100\%$) as predicted by the theory in Figure 1 in light red and light green, respectively. As per Theorem 3, the success regime is the entire region where the MAC is > 2. From Corollary 1, in the asymptotic case as $n \to \infty$ the failure regime becomes infinitesimally small. Since the value of n is large ($= 10^7$), the failure regime becomes quite narrow and Figure 1 shows that accuracy of our algorithms drops below 50% even outside it. This illustrates the fact that Corollary 1 only gives us a regime of guaranteed failure and does not imply that failure will not occur outside the regime. However this is probably due to the imperfection of the existing algorithms. Hence we can reasonably expect that as G-SSL algorithms get more sophisticated, the empirical point of failure will get pushed further and further towards the theoretical failure regime.

Following Conjecture 1, the values of the MAC for different lower bounds of guaranteed accuracy are obtained by simply scaling by a factor of $\frac{\log n}{\log(n/s)}$. That is, for an accuracy guarantee of $100(1 - {}^{s}/_{n})\%$, the MAC must be above $2\frac{\log(n/s)}{\log n}$. We delineate a few such values of the MAC as black (dotted) vertical lines on the graphs and find them to be quite accurate lower bounds of accuracy of LP, thus furnishing empirical evidence in its favour. We also observe that 99% accuracy is achieved long before the theoretical success regime of exact recovery (acc. $= 100\%$), which highlights the importance of the partial recovery result. Since we get *almost identical* graphs for the three cases of Theorem 3, these results corroborate the theory presented in the paper in the realm of a symmetric 2-cluster PLSBM.

8 Conclusion

In this paper we have imparted new meaning and formalisation to the notions of success and failure of LP by relating its accuracy to the notions of recovery in community detection. The novel theoretical results presented characterise the efficacy of LP based on graph structure, label volume and quality. Experiments on synthetic datasets generated from PLSBM confirmed the theory, as well as furnished insights on the regimes that the theory does not apply to. However, analysis of a broader family of G-SSL algorithms on real-world graphs remains an interesting open problem.

Acknowledgements

This work has been supported by the following projects:

1. Project "AI/ML Techniques for Online Shopping" sponsored by Flipkart Internet Private Limited.
2. Project "Learning Representations from Network Data" sponsored by Intel Corporation.

[5] Note that this is not sufficient. However, the experimental evidence presented in Figure 1 supports the characterisation.

References

[Abbe and Sandon, 2015] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 670–688. IEEE, 2015.

[Abbe *et al.*, 2015] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. volume 62, pages 471–487. IEEE, 2015.

[Abbe, 2017] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[Bengio *et al.*, 2006] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. 11 label propagation and quadratic criterion. *researchgate.net*, 2006.

[Chakrabarti *et al.*, 2014] Deepayan Chakrabarti, Stanislav Funiak, Jonathan Chang, and Sofus A Macskassy. Joint inference of multiple label types in large networks. *arXiv preprint arXiv:1401.7709*, 2014.

[Ke and Honorio, 2018] Chuyang Ke and Jean Honorio. Information-theoretic limits for community detection in network models. In *Advances in Neural Information Processing Systems*, pages 8324–8333, 2018.

[Massoulié, 2014] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703. ACM, 2014.

[McPherson *et al.*, 2001] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[Saad and Nosratinia, 2018] Hussein Saad and Aria Nosratinia. Community detection with side information: Exact recovery under the stochastic block model. volume 12, pages 944–958. IEEE, 2018.

[Stretcu *et al.*, 2019] Otilia Stretcu, Krishnamurthy Viswanathan, Dana Movshovitz-Attias, Emmanouil Platanios, Sujith Ravi, and Andrew Tomkins. Graph agreement models for semi-supervised learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8710–8720. Curran Associates, Inc., 2019.

[Talukdar and Crammer, 2009] Partha Pratim Talukdar and Koby Crammer. New regularized algorithms for transductive learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 442–457. Springer, 2009.

[Yamaguchi and Hayashi, 2017] Yuto Yamaguchi and Kohei Hayashi. When does label propagation fail? a view from a network generative model. In *IJCAI*, pages 3224–3230, 2017.

[Yun and Proutiere, 2014] Se-Young Yun and Alexandre Proutiere. Accurate community detection in the stochastic block model via spectral algorithms. *arXiv preprint arXiv:1412.7335*, 2014.

[Zhu *et al.*, 2003] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.

[Zhu, 2005] Xiaojin Zhu. *Semi-supervised learning with graphs*. PhD thesis, Carnegie Mellon University, language technologies institute, school of, 2005.