Clustering in Partially Labeled Stochastic Block Models via Total Variation Minimization

Alexander Jung
Department of Computer Science, Aalto University
Espoo, Finland; firstname.lastname(at)aalto.fi

***Abstract*—A main task in data analysis is to organize data points into coherent groups or clusters. The stochastic block model is a probabilistic model for the cluster structure. This model prescribes different probabilities for the presence of edges within a cluster and between different clusters. We assume that the cluster assignments are known for at least one data point in each cluster. In such a partially labeled stochastic block model, clustering amounts to estimating the cluster assignments of the remaining data points. We study total variation minimization as a method for this clustering task. We implement the resulting clustering algorithm as a highly scalable message passing protocol. We also provide a condition on the model parameters such that total variation minimization allows for accurate clustering.**

I. Introduction

Many application domains generate data with an intrinsic network structure [1], [2]. One of the main workhorses for processing such networked data is the stochastic block model (SBM) [3]. The SBM is a generative (probabilistic) model for the network structure of data and offers a principled approach to community detection or clustering methods [4], [5].

The SBM extends the Erdős-Rényi (ER) random graph model by prescribing an intrinsic cluster structure. The cluster assignments of nodes (data points) are considered as labels associated with nodes. Clustering algorithms are obtained from inference methods for the SBM which estimate the labels from the observed links between data points [6], [7].

Most existing clustering methods using SBM only take network structure into account. However, in some applications we might have a good idea of the (difference in the) cluster assignments for a few data points. The partially labeled SBM (PLSBM) assumes that cluster assignments of a certain fraction of the nodes are known. Clustering methods for the labeled SBM have been studied previously [6], [8].

We consider partially labeled SBM for the extreme case of having access to the cluster assignment of exactly one data point for each cluster. The recovery of the cluster assignments for all remaining data points is then based on interpreting the cluster assignments as piece-wise constant graph signals. Piece-wise constant graph signals can be recovered efficiently using total variation minimization.

Our main contributions are:

- a message passing method to learn cluster assignments of partially labeled networked data.
- a precise condition on the SBM parameters such that out method accurately recovers cluster assignments.

Notation: For a real number $x \in \mathbb{R}$ we define its quantized version $\mathrm{round}\{x\} \in \mathbb{Z}$ as the integer such that $x = \mathrm{round}\{x\} + r$ with $-1/2 \leq r < 1/2$. The maximum and minimum of two numbers x, y is denoted as $x \vee y$ and $x \wedge y$, respectively.

II. Problem Formulation

We represent networked data by an undirected *empirical graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The nodes $i \in \mathcal{V} = \{1, \ldots, N\}$ represent data points such as text documents or social network users.

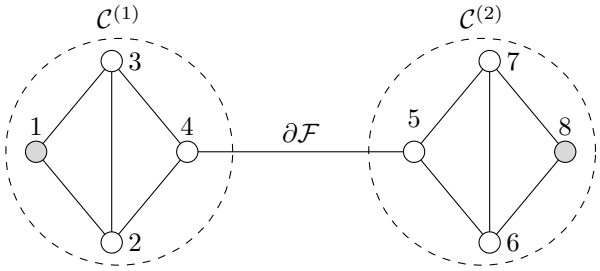

Fig. 1: Empirical graph $\mathcal{G}$ whose nodes $\mathcal{V}$ are grouped into two clusters $\mathcal{C}^{(1)}$ and $\mathcal{C}^{(2)}$ forming the partition $\mathcal{F} = \{\mathcal{C}^{(1)}, \mathcal{C}^{(2)}\}$. Nodes for which the cluster assignments is known are shaded.

Two data points $i, j \in \mathcal{V}$ are connected by an undirected edge $\{i, j\} \in \mathcal{E}$ if they are considered similar, such as documents authored by the same person or social network profiles of befriended users. For ease of notation, we denote the edge set $\mathcal{E}$ by $\{1, \ldots, E := |\mathcal{E}|\}$.

It will be convenient to define a directed version of the empirical graph by orienting each undirected edge $e = \{i, j\}$ to obtain the directed edge (e_+, e_-) with $e_+ := i \wedge j$ and $e_- := i \vee j$.

The neighborhood and degree of a node $i \in \mathcal{V}$ are denoted $\mathcal{N}(i) := \{j : (i, j) \in \mathcal{E}$ and $d_i := |\mathcal{N}(i)|$, respectively. It will be convenient to also define the directed neighbourhoods of a node $i \in \mathcal{V}$ as

$$\mathcal{N}^+(i) := \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}, i < j\}, \text{ and}$$
$$\mathcal{N}^-(i) := \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}, i > j\}. \quad (1)$$

We consider data having an intrinsic cluster structure [2] with a known number K of clusters

$$\mathcal{V} = \mathcal{C}^{(1)} \cup \ldots \cup \mathcal{C}^{(K)}. \quad (2)$$

The ith data point is assigned to the cluster $c^{(i)} \in \{1, \ldots, K\}$.

The SBM interprets the (presence of) edges between two nodes $i,j \in \mathcal{V}$ in the empirical graph $\mathcal{G}$ as realizations of independent random variables $t_{i,j} \in \{0,1\}$. An edge is present ($\{i,j\} \in \mathcal{E}$) between two nodes $i,j \in \mathcal{V}$ if and only if $t_{i,j}=1$.

One version of the SBM prescribes a constant probability $p_{\rm in}$ for placing an edge between nodes in the same cluster,

$$\mathrm{P}\{\underbrace{t_{i,j}=1}_{\{i,j\}\in\mathcal{E}}\} = p_{\rm in} \text{ for } c^{(i)}=c^{(j)}, \tag{3}$$

and another constant probability $p_{\rm out}$ for placing an edge between nodes from different clusters,

$$\mathrm{P}\{\underbrace{t_{i,j}=1}_{\{i,j\}\in\mathcal{E}}\} = p_{\rm out} \text{ for } c^{(i)}\neq c^{(j)}. \tag{4}$$

We propose and study a method for recovering the cluster assignments c_i of all data points $i \in \mathcal{V}$. The recovery is based on the edge set $\mathcal{E}$ and knowledge of the cluster assignments c_i for few data points in a small training set $\mathcal{M} \subseteq \mathcal{V}$.

Our focus is on the extreme case of observing the cluster assignment of exactly one data point in each cluster,

$$\left|\mathcal{M} \cap \mathcal{C}^{(k)}\right| = 1 \text{ for each } k=1,\ldots,K. \tag{5}$$

Thus, we consider a training set $\mathcal{M} = \{i^{(1)},\ldots,i^{(K)}\}$ consisting of exactly K nodes $i^{(k)} \in \mathcal{C}^{(k)}$ for $k=1,\ldots,K$.

We interpret the cluster assignments c_i as the signal values of a piece-wise constant graph signal $\mathbf{c} = \big(c_1,\ldots,c_N\big)^T \in \mathbb{R}^N$. Recovering the cluster assignments for all data points then amounts to the problem of recovering a piece-wise constant graph signal from the knowledge of its values on the training set $\mathcal{M}$.

III. TV Minimization

A recently studied method for recovering piece-wise constant graph signals is based on minimizing total variation (TV)

$$\|\mathbf{c}\|_{\rm TV} := \sum_{\{i,j\}\in\mathcal{E}} |c_j - c_i|. \tag{6}$$

We also define the TV for subsets $\mathcal{S} \subseteq \mathcal{E}$ of edges via

$$\|\mathbf{c}\|_{\mathcal{S}} := \sum_{\{i,j\}\in\mathcal{S}} |c_j - c_i|. \tag{7}$$

We focus on the SBM regime $p_{\rm in} \gg p_{\rm out}$ such that nodes within the same cluster are well connected whereas few links between nodes from different clusters exist. In this case, the cluster assignments form a graph signal $\mathbf{c}$ having a small TV.

It seems natural to recover a graph signal with small TV and known signal values on the set $\mathcal{M}$ via

$$\hat{\mathbf{c}} \in \operatorname*{argmin}_{\tilde{\mathbf{c}}\in\mathbb{R}^N} \underbrace{\sum_{\{i,j\}\in\mathcal{E}} |\tilde{c}_j - \tilde{c}_i|}_{=\|\tilde{\mathbf{c}}\|_{\rm TV}} \text{ s.t. } \tilde{c}_i = c_i \text{ for all } i\in\mathcal{M}. \tag{8}$$

Since the objective function and the constraints in (8) are convex, the optimization problem (8) is a convex optimization problem [9]. In fact, (8) can be reformulated as a linear program [9, Sec. 1.2.2].

The solution to (8) might not be unique but any such solution $\hat{\mathbf{c}}$ provides an estimated cluster assignment $\hat{c}_i$ that is characterized by: (i) it is consistent with known cluster assignments: $\hat{c}_i = c_i$ for all nodes $i \in \mathcal{M}$; and (ii) it has minimum TV (6) among all such cluster assignments.

As shown in [10], TV minimization (8) can be solved iteratively by a scalable message passing method which we have summarized in Algorithm 1. The stopping criterion in Algorithm 1 can be a fixed number of iterations. The number of iterations can be chosen based on the convergence analysis in [11]. Another option is to monitor the decrease of the objective function in (8) and stop is the relative decrease falls below a specified (small) threshold.

Algorithm 1 Clustering in PLSBM

Input: empirical graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$, node set $\mathcal{M}$ with known cluster assignments $\{c_i\}_{i\in\mathcal{M}}$.

Initialize: $r:=0$, $\bar{\mathbf{c}} = \hat{\mathbf{c}}^{(0)} = \hat{\mathbf{c}}^{(-1)} = \hat{\mathbf{c}}^{(0)} := \mathbf{0}$, $\gamma_i := 1/d_i$.

1: **repeat**

2: for all nodes $i\in\mathcal{V}$: $\tilde{c}_i := 2\hat{c}_i^{(r)} - \hat{c}_i^{(r-1)}$

3: for all edges $e=(i,j)\in\mathcal{E}$:

$$\hat{y}_e^{(r+1)} := \hat{y}_e^{(r)} + (1/2)(\tilde{c}_{e^+} - \tilde{c}_{e^-})$$

4: for all edges $e\in\mathcal{E}$:

$$\hat{y}_e^{(r+1)} := \hat{y}_e^{(r+1)} / \max\{1, |\hat{y}_e^{(r+1)}|\}$$

5: for all nodes $i\in\mathcal{V}$:

$$\hat{c}_i^{(r+1)} := \hat{c}_i^{(r)} - \gamma_i \Big[\sum_{j\in\mathcal{N}^+(i)} \hat{y}_{(i,j)}^{(r+1)} - \sum_{j\in\mathcal{N}^-(i)} \hat{y}_{(j,i)}^{(r+1)} \Big]$$

6: for all labeled nodes $i\in\mathcal{M}$: $\hat{c}_i^{(r+1)} := c_i$

7: $r := r+1$

8: for all nodes $i\in\mathcal{V}$: $\bar{c}_i := (1-1/r)\bar{c}_i + (1/r)\hat{c}_i^{(r)}$

9: **until** stopping criterion is satisfied

Output: cluster assignments $\hat{c}_i := \mathrm{round}\{\hat{c}_i^{(r)}\}$ for $i\in\mathcal{V}$

IV. When Does it Work?

We now discuss conditions such that solutions of TV minimization (8) to coincide with the true underlying cluster assignments c_i with high probability. These conditions involve the SBM parameters $p_{\rm in}$, $p_{\rm out}$ and cluster sizes $|\mathcal{C}^{(k)}|$. For deriving these conditions, we need the concept of network flows [12], [13]

Definition 1. *A network is a graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$ with capacities b_e for each edge $e\in\mathcal{E}$. We define one node $s\in\mathcal{V}$ as source and another node $t\in\mathcal{V}$ as the sink. A network flow $f:\mathcal{E}\to\mathbb{R}$ assigns each directed edge $e=(i,j)\in\mathcal{E}$ a number $f_e\in\mathbb{R}$ such that*

- *the flow is conserved at each node which is not a sink or source,*

$$\sum_{j\in\mathcal{N}^+(i)} f_{(i,j)} - \sum_{j\in\mathcal{N}^-(i)} f_{(j,i)} = 0 \text{ for each } i\in\mathcal{V}\setminus\{s,t\}. \tag{9}$$

- *the flow does not exceed the edge capacities,*

$$|f_{(i,j)}| \leq b_{(i,j)} \text{ for each directed edge } (i,j) \in \mathcal{E}. \quad (10)$$

The value of a flow is

$$\mathrm{val} f := \sum_{j:(s,j)\in\mathcal{E}} f_{(s,j)} - \sum_{j:(i,s)\in\mathcal{E}} f_{(i,s)}. \quad (11)$$

The concept of network flows allows to quantify the connectivity of the clusters $\mathcal{C}^{(k)}$ in the empirical graph $\mathcal{G}$. Let us define, for each cluster $\mathcal{C}^{(k)}$ in the original empirical graph $\mathcal{G}$, an associated graph $\mathcal{G}^{(k)}$. The nodes $\mathcal{V}^{(k)}$ of the graph $\mathcal{G}^{(k)}$ contains all nodes in the cluster $\mathcal{C}^{(k)}$ and one additional node denoted $t^{(k)}$,

$$\mathcal{V}^{(k)} := \mathcal{C}^{(k)} \cup \{t^{(k)}\}. \quad (12)$$

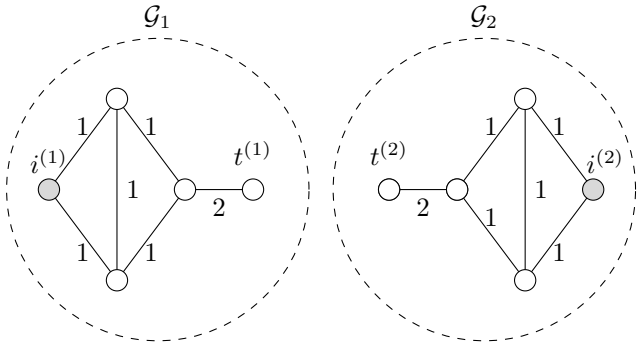

Fig. 2: Graphs $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ associated with the clusters $\mathcal{C}^{(1)}$ and $\mathcal{C}^{(2)}$ of the empirical graph depicted in Figure 1.

The edges $\mathcal{E}^{(k)}$ of the graph $\mathcal{G}^{(k)}$ are obtained by retaining all intra-cluster edges $\{i,j\} \in \mathcal{E}$ with $i,j \in \mathcal{C}^{(k)}$ and then adding an edge from the augmented node $t^{(k)}$ to each of the boundary nodes

$$\partial\mathcal{C}^{(k)} := \{i \in \mathcal{C}^{(k)} : \{i,j\} \in \mathcal{E} \text{ with some } j \notin \mathcal{C}^{(k)}\}. \quad (13)$$

We assign the capacity $b_{(i,j)} = 2$ to all edges of $\mathcal{G}^{(k)}$ which are incident to the node $t^{(k)}$. The remaining edges of $\mathcal{G}^{(k)}$ are assigned capacity $b_{(i,j)} = 1$.

As shown in [14], the solution of (8) coincides with the true underlying cluster assignments if, for each graph $\mathcal{C}^{(k)}$, there exists a network flow of value $2|\partial\mathcal{C}^{(k)}|$ between source node $s = i^{(k)} \in \mathcal{M} \cap \mathcal{C}^{(k)}$ and the augmented node $t^{(k)}$.

Claim 1 (Informal)**.** *The solution of TV minimization* (8) *coincides with the true cluster assignments c_i for all nodes if for each graph $\mathcal{G}^{(k)}$ there is a flow $f^{(k)}$ of value*

$$\mathrm{val}\{f^{(k)}\} = 2|\partial\mathcal{C}^{(k)}|. \quad (14)$$

Using large deviation analysis (see [15, Theorem 2.1]), we can ensure the conditions in Claim 1 with high probability whenever

$$p_{\mathrm{in}} \gg 2p_{\mathrm{out}}(|\mathcal{V}| - |\mathcal{C}^{(k)}|) \text{ for each } k = 1,\ldots,K. \quad (15)$$

Condition (15) characterizes parameter regimes for the SBM such that TV minimization (8), implemented by Algorithm 1, recovers the cluster assignments c_i of all data points $i \in \mathcal{V}$.

V. Numerical Experiments

We verify the (non-rigorous) condition (15) on the SBM parameters for TV minimization (8) succeeding to recover the cluster assignments, by a numerical experiment.

In this experiment we generate an empirical graph $\mathcal{G}$ using a SBM with two clusters $\mathcal{C}^{(1)} = \{1,\ldots,50\}$ and $\mathcal{C}^{(2)} = \{51,\ldots,100\}$. An edge is placed between nodes i,j with probability p_{in} if they are in the same cluster and with probability p_{out} if they are from different clusters.

The cluster assignments c_i form a piece-wise constant graph signal,

$$c_i = \begin{cases} 1 & \text{for } i \in \mathcal{C}^{(1)}, \\ 2 & \text{for } i \in \mathcal{C}^{(2)}. \end{cases} \quad (16)$$

We assume that cluster assignments are known only for nodes $\mathcal{M} = \{1, 51\}$. The cluster assignments of the remaining nodes are then estimated using Algorithm 1.

The (non-rigorous) condition (15) suggests that Algorithm 1 delivers correct cluster assignments with high probability whenever $p_{\mathrm{in}}/p_{\mathrm{out}} \gg 2(|\mathcal{V}| - |\mathcal{C}^{(k)}|)$ for $k = 1, 2$. Inserting the particular SBM parameters used in this experiment yields the condition $p_{\mathrm{in}}/p_{\mathrm{out}} \gg 10$.

Figure 3 depicts the accuracy of Algorithm 1 for varying ratio $p_{\mathrm{in}}/p_{\mathrm{out}}$. We measure the accuracy of Algorithm 1 using the fraction of nodes for which $\hat{c}_i = c_i$ (averaged over 100 i.i.d. simulation runs). The results in Figure 3 agree with the (non-rigorous) condition $p_{\mathrm{in}}/p_{\mathrm{out}} \gg 10$ such that TV minimization correctly recovers the cluster assignments of all nodes.

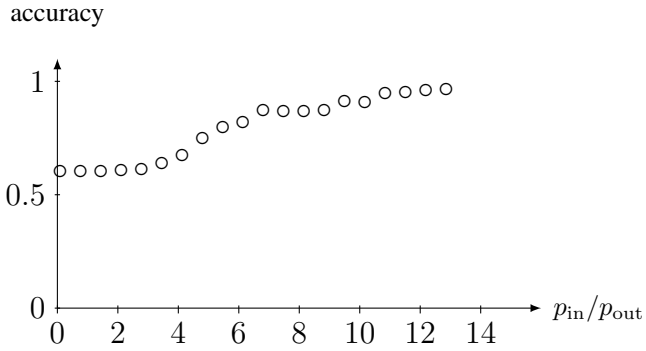

Fig. 3: Clustering accuracy achieved by Algorithm 1 for varying $p_{\mathrm{in}}/p_{\mathrm{out}}$.

The source code underlying this experiment is available at https://github.com/alexjungaalto/ResearchPublic/blob/master/TVMinPLSBM/tvmin_PLSBM.m.

References

[1] S. Cui, A. Hero, Z.-Q. Luo, and J.M.F. Moura, Eds., *Big Data over Networks*, Cambridge Univ. Press, 2016.

[2] M. E. J. Newman, *Networks: An Introduction*, Oxford Univ. Press, 2010.

[3] Emmanuel Abbe, "Community detection and stochastic block models: Recent developments," *Journal of Machine Learning Research*, vol. 18, no. 177, pp. 1–86, 2018.

[4] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, Feb. 2010.

[5] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou, "Achieving optimal misclassification proportion in stochastic block models," *J. Mach. Learn. Res.*, vol. 18, pp. 1–45, 2017.

[6] E. Mossel, J. Neeman, and A. Sly, "Belief propagation, robust reconstruction and optimal recovery of block models," *Ann. App. Prob.*, vol. 26, no. 4, pp. 2211–2256, 2016.
[7] A. A. Amini, A. Chen, P. J. Bickel, and E. Levina, "Pseudo-likelihood methods for community detection in large sparse networks," *Ann. Statist.*, vol. 41, no. 4, pp. 2097–2122, 2013.
[8] T.T. Cai, T. Liang, and A. Rakhlin, "Inference via message passing on partially labeled stochastic block models," *arXiv*, 2016.
[9] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge Univ. Press, Cambridge, UK, 2004.
[10] A. Jung, A O. Hero, A. Mara, S. Jahromi, A. Heimowitz, and Y.C. Eldar, "Semi-supervised learning in network-structured data via total variation minimization," *ArXiv e-prints*, 2019.
[11] A. Jung, "On the complexity of sparse label propagation," *Front. Appl. Math. Stat.*, vol. 4, pp. 22, July 2018.
[12] J. Kleinberg and E. Tardos, *Algorithm Design*, Addison Wesley, 2006.
[13] Dieter Jungnickel, *Graphs, Networks and Algorithms*, Springer Berlin Heidelberg, 4 edition, 2013.
[14] A. Jung, A.O. Hero III, A. Mara, and S. Jahromi, "Semi-supervised learning via sparse label propagation," *arxiv*, 2017.
[15] David R. Karger, "Random sampling in cut, flow, and network design problems," *Mathematics of Operations Research*, vol. 24, no. 2, 1999.