Clustering Networks with Node and Edge Attributes using Bregman Divergence

Felipe Schreiber Fernandes [1]
Advisors: Daniel Ratton Figueiredo [1]**, Maximilien Dreveton**[2]

[1]Programa de Engenharia de Sistemas e Computação (PESC)
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ

[2]School of Computer and Communication Sciences (IC)
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland

{felipesc,daniel}@cos.ufrj.br, maximilien.dreveton@epfl.ch

Abstract. *Network clustering tackles the problem of identifying sets of nodes (clusters or communities) that have similar connection patterns. However, in many modern scenarios, nodes also have attributes that are correlated with the network structure. Thus, network information (edges) and node information (attributes) can be jointly leveraged to design high-performance clustering algorithms. Under a general model for the network and node attributes, this thesis establishes an information-theoretic criterion for the exact recovery of community labels and characterizes a phase transition determined by the Chernoff-Hellinger divergence of the model. The criterion shows how network and attribute information can be exchanged in order to yield exact recovery (e.g., more reliable network information requires less reliable attribute information). This thesis also presents two iterative clustering algorithms that greedily maximizes the joint likelihood of the model under the assumption that the probability distribution of network edges and node attributes belong to exponential families. Extensive analysis of the two algorithms on both synthetic datasets and real benchmarks highlights their accuracy and performance with respect to other state-of-the-art approaches.*

1. Introduction

Community detection or network clustering–the task of identifying sets of similar nodes in a network–is a fundamental problem in network analysis [Abbe 2017, Fortunato and Hric 2016], with applications in diverse fields such as digital humanities, data science and biology. In the classic formulation, a set of communities must be determined from the connection patterns among the nodes of a single network. A simple random graph model with community structure, the Stochastic Block Model (SBM), has been the canonical model to characterise theoretical limitations for detecting communities and evaluate different community detection algorithms [Abbe 2017].

However, nodes of many real-world networks have attributes or features that can reveal their identity as an individual or within a group. For example, the age, gender and ethnicity of individuals in a social network [Newman and Clauset 2016], the title, keywords and co-authors of papers in a citation network [Sen et al. 2008], or the longitude and latitude of meteorological stations in weather forecast networks [Braun et al. 2022]. In some scenarios, such attributes can be leveraged alone to identify node communities (clusters) without even using the network.

Thus, a modern formulation for community detection must consider both network information (edges) and node information (attributes). Indeed, recent works have addressed this problem by designing community detection algorithms that can effectively leverage both sources of information to improve performance, including techniques based on modularity optimisation [Combe et al. 2015], belief propagation [Deshpande et al. 2018], spectral clustering [Abbe et al. 2022] and iterative likelihood based methods [Braun et al. 2022].

A fundamental problem in this new formulation is fusing both sources of information: how important is network information in comparison to node information given a problem instance? Intuitively, this depends on the noise associated with network edges and node attributes. For example, if edges are reliable then the clustering algorithm should prioritize edges when determining the communities. However, most prior approaches adopt some form of heuristic when merging the two sources of information [Combe et al. 2015, Deshpande et al. 2018]. A rigorous approach to this problem requires a mathematical model, and one has been recently proposed.

The Contextual Stochastic Block Model (CSBM) is a generalization of the classic SBM where nodes and edges can have random attributes that depend on their communities. While the model formulation is general, CSBM has only been rigorously studied in the restrictive setting where edges are binary (present or not) and node attributes follow a Gaussian distribution [Abbe et al. 2022, Braun et al. 2022, Deshpande et al. 2018].

However, real networks often depart from binary edges and Gaussian attributes. Indeed, in many scenarios network edges have weights that reveal information about their interactions and nodes have discrete or non-Gaussian attributes. This work tackles this scenario by considering a CSBM where edges have weights and nodes have attributes that follow arbitrary distributions. Under this general model, this thesis is the first to characterise the phase transition for the exact recovery of community labels. In particular, the Chernoff-Hellinger (CH) divergence, initially defined just for binary networks [Abbe and Sandon 2015], is extended to this more general model. This divergence effectively captures the difficulty of distinguishing different communities and thus plays a crucial role in determining the limits of exact recovery. The analysis reveals an additional term in the divergence that quantifies the information provided by the attributes of the nodes. Moreover, it quantifies the trade-off between network and node information in meeting the threshold for exact recovery.

Characterizing the threshold for exact recovery is important because it provides a theoretical lower bound for recovering the communities with a fraction of errors that goes to zero as the network size increases. This means that no algorithm (independent of running time) can exactly recover the communities if the problem at hand has a CH divergence value below the threshold. However, the knowledge threshold does not lead to an algorithm that can exactly recover communities when the problem at hand has a CH divergence value above the threshold. Thus, designing efficient algorithms to recover algorithms in this scenario is paramount.

As it was first proposed, the CSBM generates weighted complete networks (all possible edges are present) when edge weights follow a continuous distribution. However, most real weighted networks are sparse (only a very small fraction of all possible edges are present). To model sparse weighted networks and to provide a practical community detection algorithm, we consider a CSBM whose weights belong to *zero-inflated distributions*. This means that edges can be absent with some probability (that depends on the communities), and in case an edge is present, its weight follows a distribution from the exponential family. Similarly, node attribute distributions (that also depend on the community) are also assumed to belong to an exponential family.

Assuming edge weights and node attributes follow distributions from the exponential family is motivated by two factors. Firstly, the exponential family encompass a broad range of parametric distributions, including the commonly used Bernoulli, Poisson, Gaussian, and Gamma distributions. Secondly, there exists an intricate connection between distributions in the exponential family and Bregman divergences, which has proven to be a powerful tool in designing algorithms across a variety of problems such as clustering, classification, and dimensionality reduction [Banerjee et al. 2005].

This connection between Bregman divergences and exponential family has been previously explored in the context of clustering dense networks (all possible edges are present) [Long et al. 2007]. In contrast, this thesis proposes two iterative algorithms also based on Bregman divergences but that can be directly applied to both dense and sparse networks. This is a key difference with most prior works which either study dense weighted networks [Mariadassou et al. 2010] or sparse binary networks with Gaussian attributes [Abbe et al. 2022, Deshpande et al. 2018, Stanley et al. 2019]. The lgorithms here proposed attempt to maximize the likelihood function through greedy iterations that lead to changes in the community assignment of nodes. The hard and soft versions of the algorithm reflect how community assignment are treated within the iterations of the algorithm. In hard clustering, at each iteration nodes are assigned to a single community, while in soft clustering, nodes are assigned a distribution over the set of communities. Hard and soft clustering are common approaches in the clustering literature [Long et al. 2007, Abbe et al. 2022], and thus our contribution is to apply them to the model under consideration using Bregman divergence.

Iterative algorithms for community detection such as the ones proposed in this thesis must start with some initial community assignment (either hard or soft). However, this initial assignment often has immense influence on the communties identified by the algorithm. In fact, starting from random community assignments often leads to very poor results. Thus, another important consideration is determining the initial assignment. Another contribution of this thesis is a methodology to generate an initial community assignment from the data.

Last, the performance of the proposed algorithms along with the proposed initialization are assessed using synthetic datasets generated by the CSBM model under consideration. Interestingly, results indicate that the hard clustering algorithm can recover communities effectively and close to the theoretical threshold for exact recovery. Moreover, the performance of the proposed algorithms are compared with state-of-the-art alternatives showing superiority in different scenarios. The proposed algorithms are also applied to real benchmark datasets for sparse networks with node attributes.

2. Related Work

2.1. Exact recovery in SBM with edge weights and node attributes

Community detection in classic SBM (binary edges) is a well-understood problem with strong theoretical results concerning exact recovery and efficient algorithms with guaranteed accuracy [Abbe 2017]. However, extending the classic SBM to weighted networks (non-binary edges) with arbitrary distributions is an ongoing research area. Most existing work in this scenario has been restricted to the *homogeneous model* (aka. *planted partition model*), where edge weights within and across communities are determined by two respective distributions.

In non-homogeneous models, a more complex divergence called the Chernoff-Hellinger (CH) divergence is the appropriate information-theoretic quantity for exact community recovery [Abbe and Sandon 2015]. However, the expression of the Chernoff-Hellinger divergence as originally defined in [Abbe and Sandon 2015] for binary networks does not have an intuitive interpretation, and its extension to non-binary (weighted) networks is challenging.

Another generalization of the SBM allows for nodes to have attributes that provide information about their community, such as the Contextual SBM (CSBM) [Deshpande et al. 2018]. CSBM has only been rigorously studied in the setting where edges are binary and node attributes follow a Gaussian distribution. In this scenario, the phase transition for exact recovery for the community labels has been established [Abbe et al. 2022, Braun et al. 2022, Deshpande et al. 2018]. A natural generalization is to investigate the model where network edges have weights and nodes have attributes that follow arbitrary distributions. Indeed, this is one of the main contributions of this thesis: Expression (3.4) gives a straightforward yet crucial formula for the phase transition for exact recovery, also providing a natural interpretation for the influence of both the network and node attributes. Moreover, Expression (3.4) also applies when no node attribute is available, thus providing the exact recovery threshold for a non-homogeneous model and arbitrary edge weight distribution, a significant advancement in the state of the art.

2.2. Algorithms for clustering weighted networks with node attributes

Algorithms leveraging different approaches have been proposed to tackle community detection in networks with edge weights and node attributes. A common principled approach is to determine the community assignment that maximizes the likelihood function of a model for the data. However, optimizing the likelihood function is computationally intractable even for binary networks. Thus, approximation schemes such as variational inference and pseudo-likelihood methods are often adopted. For instance, [Mariadassou et al. 2010] introduced a variational-EM algorithm for clustering non-homogeneous weighted SBM with arbitrary distributions. Another approach for clustering node-attributed SBM whose edge weights and attribute distributions belong to exponential families is [Long et al. 2007]. These two approaches assume that the network is dense (all edges are present and have non-zero edge weight). However, most real networks are very sparse (most node pairs do not have an edge) and this work focuses on this scenario. Another very recent approach tackling sparse networks is the *IR_sLs* algorithm [Braun et al. 2022], although its theoretical guarantees assume binary networks with Gaussian attributes. The performance of the algorithms here proposed are directly compared to *IR_sLs*, illustrating their superiority in different scenarios.

3. Model and Exact Recovery in CSBM

3.1. Model definition

Consider n nodes partitioned into $K \geq 2$ disjoint sets, called *blocks* or communities. A node-labelling vector $z = (z_1, \cdots, z_n) \in [K]^n$ represents this partitioning so that $z_i \in [K]$ indicates the block (label) of node i. The block of nodes are random variables assumed to be independent and identically distributed such that $\mathbb{P}(z_i = k) = \pi_k$ for some vector $\pi \in (0,1)^K$ verifying that $\sum_k \pi_k = 1$. The nodes interact in unordered pairs giving rise to undirected edges, and $\mathcal{X}$ is the measurable space of all possible edge weights. Additionally, each node has an attribute that is an element of a measurable space $\mathcal{Y}$. Let $X \in \mathcal{X}^{N \times N}$ denote the symmetric matrix such that X_{ij} represents the edge weight between node pair (ij), and by $Y = (Y_1, \cdots, Y_n) \in \mathcal{Y}^n$ the node attribute vector.

Assume that edge weights and attributes are independent conditionally on the community labels of the nodes. Let $f_{k\ell}(x)$ denote the probability that two nodes in blocks k and ℓ have an edge $x \in \mathcal{X}$, and $h_k(y)$ denote the probability that a node in block $k \in [K]$ has an attribute $y \in \mathcal{Y}$. Thus,

$$\mathbb{P}\left(X, Y \,|\, z\right) \;=\; \prod_{1 \leq i < j \leq n} f_{z_i z_j}(X_{ij}) \prod_{i=1}^{n} h_{z_i}(Y_i). \tag{3.1}$$

In the following, the spaces $\mathcal{X}, \mathcal{Y}$ might depend on n, as well as the respective probabilities f, h. The number of nodes n will increase to infinity while K and π are constant. For an estimator $\hat{z} \in [K]^n$ of z, we define the *classification error* as

$$(z, \hat{z}) \;=\; \min_{\tau \in \mathcal{S}_K} (z, \tau \circ \hat{z}),$$

where $\mathcal{S}_K$ is the set of permutations of $[K]$ and $(\cdot, \cdot)$ is the hamming distance between two vectors. An estimator $\hat{z} = \hat{z}(X, Y)$ achieves *exact recovery* if $\mathbb{P}\left((z, \hat{z}) \geq 1\right) = o(1)$.

3.2. Exact recovery threshold

The difficulty of classifying empirical data into one of K possible classes is traditionally measured by the *Chernoff information*. More precisely, in the context of network clustering, let $\mathrm{CH}(a, b) = \mathrm{CH}(a, b, \pi, f, h)$ denote the hardness of distinguishing nodes that belong to block a from block b. This quantity is defined by

$$\mathrm{CH}(a, b) \;=\; \sup_{t \in (0,1)} \mathrm{CH}_t(a, b), \tag{3.2}$$

where

$$\mathrm{CH}_t(a, b) \;=\; (1 - t) \left[\sum_{c=1}^{K} \pi_c \, \mathrm{D}_t\left(f_{bc} \| f_{ac}\right) + \frac{1}{n} \mathrm{D}_t\left(h_b \| h_a\right) \right] \tag{3.3}$$

is the *Chernoff coefficient* of order t across blocks a and b, and $\mathrm{D}_t(f \| g) = \frac{1}{t-1} \log \int f^t(x) g^{1-t}(x) dx$ is the *Rényi divergence* of order t between two probability densities f, g [Van Erven and Harremos 2014]. The key quantity assessing the possibility or impossibility of exact recovery in SBM is then the minimal Chernoff information across all pairs of clusters. We denote it by $I = I(\pi, f, h)$, and it is defined by

$$I = \min_{\substack{a,b \in [K] \\ a \neq b}} \mathrm{CH}(a,b). \tag{3.4}$$

The following Theorem provides the information-theoretic threshold for exact recovery in node-attributed SBM.

Theorem 3.1. *Consider model* (3.1) *with* $\pi_a > 0$ *for all* $a \in [K]$. *Denote by* a^*, b^* *the two hardest blocks to estimate, that is* $\mathrm{CH}(a^*, b^*) = I$. *Suppose that* $t \in (0,1) \mapsto \lim_{n\to\infty} \frac{n}{\log n} \mathrm{CH}_t(a^*, b^*)$ *exists and is strictly concave. Then the following holds:*

(i) exact recovery is information-theoretically impossible if $\lim_{n\to\infty} \frac{n}{\log n} I < 1$*;*

(ii) exact recovery is information-theoretically possible if $\lim_{n\to\infty} \frac{n}{\log n} I > 1$.

The proof for Theorem 3.1 is provided in our paper [Dreveton et al. 2023].

4. Bregman hard clustering of sparse weighted node-attributed networks

In this section, we will propose an algorithm for clustering sparse, weighted networks with node attributes. When present, the weights are sampled from an exponential family, and the node attributes also belong to an exponential family. In Section 4.1, we provide some reminder of exponential families. We derive the likelihood of the model in Section 4.2, and present the algorithm in Section 4.3.

4.1. Exponential family

An exponential family $\mathcal{E}_\psi$ is a parametric class of probability distributions whose densities can be canonically written as $p_{\theta,\psi}(x) = e^{<\theta,x>-\psi(\theta)}$, where the density is taken with respect to an appropriate measure, $\theta \in \Theta$ is a function of the parameters of the distribution that must belong to an open convex space Θ, and ψ is a convex function.

We consider the model defined in (3.1), such that f_{ab} are *zero-inflated distributions* and are given by

$$f_{ab}(x) = (1 - p_{ab})\delta_0(x) + p_{ab} f^*_{ab}(x), \tag{4.1}$$

where $p_{ab} \in [0,1]$ is the edge probability between blocks a and b, $\delta_0(x)$ is the Dirac delta at zero, and f^*_{ab} is a probability density with no mass at zero. Note that this model can represent sparse weighted networks, as edges between nodes in blocks a and b are absent with probability $1 - p_{ab}$.

Finally, suppose that the distributions $\{f^*_{ab}\}$ and $\{h_a\}$ belong to exponential families. More precisely,

$$f^*_{ab}(x) = e^{<\theta_{ab},x>-\psi(\theta_{ab})} \quad \text{and} \quad h_a(y) = e^{<\eta_a,y>-\phi(\eta_a)}, \tag{4.2}$$

for some parameters θ_{ab}, η_a and functions ψ, ϕ.

4.2. Log-likelihood

Given a convex function ψ, the *Bregman divergence* $d_\psi \colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_+$ is defined by

$$d_\psi(x, y) \;=\; \psi(x) - \psi(y) - < x - y, \nabla\psi(y) > .$$

The log-likelihood of the density $p_{\psi,\theta}$ of an exponential family distribution is linked to the Bregman divergence by the following relationship

$$\log p_{\psi,\theta}(x) \;=\; -d_{\psi^*}(x, \mu) + \psi^*(x), \tag{4.3}$$

where $\mu = \mathbb{E}_{p_{\psi,\theta}}(X)$ is the mean of the distribution, and ψ^* denotes the *Legendre transform* of ψ, defined by $\psi^*(t) \;=\; \sup_\theta\{< \theta, t > -\psi(\theta)\}$.

Suppose that X, Y follow the model (3.1) with probability distributions given by (4.1)-(4.2). Let A be a binary matrix such that $A_{ij} = 1(X_{ij} \neq 0)$. We have

$$-\log \mathbb{P}(X, Y \mid z) \;=\; \sum_i \left\{ \frac{1}{2} \sum_{j \neq i} \left[\mathrm{d_{KL}}(A_{ij}, p_{z_i z_j}) + A_{ij} d_{\psi^*}\left(X_{ij}, \mu_{z_i z_j}\right)\right] + d_{\phi^*}(Y_i, \nu_{z_i}) \right\} + c,$$

where the additional term c is a function of X, Y but does not depend on z. Denoting $Z \in \{0,1\}^{n \times K}$ the one-hot membership matrix such that $Z_{ik} = 1(z_i = k)$, observe that $p_{z_i z_j} = \left(ZpZ^T\right)_{ij}$ where p is a symmetric matrix with the edge probabilities between different blocks, $\mu_{z_i z_j} = \left(Z\mu Z^T\right)_{ij}$ where μ is a symmetric matrix with the expected value of the edge weights between different blocks, and $\nu_{z_i} = \left(Z^T \nu\right)_i$ where ν is a vector with the expected value of the attribute for different blocks. Thus, up to some additional constants, the negative log-likelihood $-\log \mathbb{P}(X, Y \mid Z)$ is equal to

$$\sum_i \left\{ \frac{1}{2} \mathrm{d_{KL}}\left(A_{i\cdot}, \left(ZpZ^T\right)_{i\cdot}\right) + \frac{1}{2} d'_{\psi^*}\left(X_{i\cdot}, \left(Z\mu Z^T\right)_{i\cdot}\right) + d_{\phi^*}\left(Y_i, \left(Z^T\nu\right)_i\right) \right\} + c, \tag{4.4}$$

where $d'_{\psi^*}(B, C) = \sum_{j=1}^n 1(B_j \neq 0) d_{\psi^*}(B_j, C_j)$ for two vectors $B, C \in \mathbb{R}^n$.

4.3. Clustering by iterative likelihood maximisation

Following the log-likelihood expression derived in (4.4), we propose an iterative clustering algorithm that places each node in the block maximising $\mathbb{P}\left(X, Y \mid z_{-i}, z_i = a\right)$ for $1 \leq a \leq K$, the likelihood that node i is in community a given the community labels of the other nodes, z_{-i}. Let $Z^{(ia)}$ denote the membership matrix obtained from Z by placing node i in block a, and let $L_{ia}(Z^{(ia)})$ denote the contribution of node i to the negative log-likelihood when node i is placed in block a. Equation (4.4) shows that

$$L_{ia}(Z) \;=\; \frac{1}{2} \mathrm{d_{KL}}\left(A_{i\cdot}, \left(ZpZ^T\right)_{i\cdot}\right) + \frac{1}{2} d'_{\psi^*}\left(X_{i\cdot}, \left(Z\mu Z^T\right)_{i\cdot}\right) + d_{\phi^*}\left(Y_i, \left(Z^T\nu\right)_i\right), \tag{4.5}$$

where the p, μ and ν in the equation above must be estimated from X, Y, and the community membership matrix Z. Let $\hat{p} = \hat{p}(A, Z)$, $\hat{\mu} = \hat{\mu}(X, Z)$, and $\hat{\nu} = \hat{\nu}(Y, Z)$ denote the estimators for p, μ and ν, respectively. Their values can be computed as follows:

$$\begin{aligned} \hat{p}(A, Z) \;&=\; \left(Z^T Z\right)^{-1} Z^T A Z \left(Z^T Z\right)^{-1}, \\ \hat{\mu}(X, Z) \;&=\; \left(Z^T A Z\right)^{-1} Z^T X Z, \\ \hat{\nu}(Y, Z) \;&=\; \left(Z^T Z\right)^{-1} Z^T Y. \end{aligned} \tag{4.6}$$

Algorithm 1 Bregman hard clustering for CSBM.

Input: Edge weights $X \in \mathcal{X}^{n \times n}$, node attributes $Y \in \mathcal{Y}^n$, convex functions ψ^*, ϕ^* (distributions), initial clustering Z_0

Let $Z = Z_0$
repeat
  Compute $\hat{p}, \hat{\mu}, \hat{\nu}$ according to (4.6)
  Let $Z = 0_{n \times K}$
  for $i = 1, \ldots, n$ **do**
    Let $Z^{(ia)}$ be the membership matrix obtained from Z by placing node i in community a
    Find $k^* = \arg\max_{a \in [K]} L_{ia}\left(Z^{(ia)}\right)$, where $L_{ia}\left(Z^{(ia)}\right)$ is defined in (4.5);
    Let $Z_{ik} = 1(k = k^*)$ for all $k = 1, \ldots, K$
  Let $Z = Z$
until *convergence*;
Return: Node-membership matrix Z

Note that the matrix inverse $\left(Z^T Z\right)^{-1}$ can be easily computed since $Z^T Z$ is a K-by-K diagonal matrix. This approach is described in Algorithm 1.

Due to space limitations, the reader is referred to the thesis [Fernandes 2023] for the description of the soft clustering algorithm.

4.4. Initial membership assignment

A fundamental aspect of many likelihood maximization iterative algorithms such as Algorithm 1 is the initial membership assignment, Z_0. This thesis proposes the following novel procedure to determine the initial assignment. Consider matrix $W \in \mathbb{R}^{n \times 2K}$ where the first K columns of W are the first K eigenvectors of the network normalised Laplacian (with edge weights), while the last K columns of W are the first K eigenvectors of the Gram matrix YY^T. This matrix is clustered using k-means algorithm to provide the initial membership assignment. More details can be found in the thesis [Fernandes 2023].

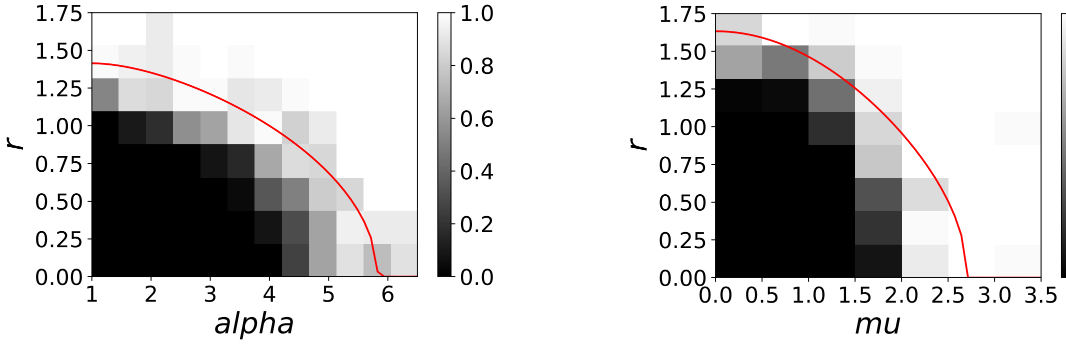
5. Numerical experiments

5.1. Exact recovery of Algorithm 1

Figure 1 shows the performance of Algorithm 1 in terms of exact recovery (fraction of times the algorithm correctly recovers the community of *all* nodes) with the theoretical threshold for exact recovery proved in the paper (red curve in the plots) in two settings: Figure 1a shows binary weight with Gaussian attributes, and Figure 1b shows zero-inflated Gaussian weights with Gaussian attributes. A solid black (resp., white) square means that over 50 trials, the algorithms failed 50 times (resp., succeeded 50 times) at exactly recovering the block structure.

5.2. Comparison with other algorithms

The Adjusted Rand Index (ARI) between the predicted clusters and the ground truth is used to evaluate the performance of the algorithms.



(a) Binary weights with Gaussian attributes

(b) zero-inflated Gaussian weights with Gaussian attributes.

Figure 1. Phase transition of exact recovery. Each pixel represents the empirical probability that Algorithm 1 succeeds at exactly recovering the clusters (over 50 runs), and the red curve shows the theoretical threshold. (a) $n = 500$**,** $K = 2$**,** $= \mathrm{Ber}(\alpha n^{-1} \log n)$**,** $= \mathrm{Ber}(n^{-1} \log n)$**. The attributes are 2d-spherical Gaussian with radius** $(\pm r\sqrt{\log n}, 0)$ **and identity covariance matrix. (b)** $n = 600$**,** $K = 3$**,** $= (1-\rho)\delta_0 + \rho\,\mathrm{Nor}(\mu, 1)$**,** $= (1-\rho)\delta_0 + \rho\,\mathrm{Nor}(0, 1)$ **with** $\rho = 5n^{-1} \log n$**. The attributes are 2d-spherical Gaussian whose means are the vertices of a regular polygon on the circle of radius** $r\sqrt{\log n}$**.**

The following algorithms are considered for comparison: *IR_sLs* algorithm [Braun et al. 2022] is one of the most recent algorithms for node-attributed SBM and it comes with theoretical guarantees (for binary networks with Gaussian attributes); *attSBM* [Stanley et al. 2019] is an EM algorithm designed for binary networks with Gaussian attributes; *EM-GMM* refers to fitting a Gaussian Mixture Model via EM on attribute data Y alone (no network); and *sc* refers to *spectral clustering* on network data X alone (no node information).

Figure 2 shows the results for binary networks with Gaussian attributes. Algorithm 1 successfully learns from both the signal coming from the network and the attributes, even in scenarios where one of them is non-informative. Moreover, Algorithm 1 has better performance than the two other node-attributed clustering algorithms, and those algorithms also show a large variance. Note that *IR_sLs* and *attSBM* are both tailor-made for binary edges and Gaussian attributes. Even in such a setting, Algorithm 1 outperforms them. When the network is weighted and the attributes non-Gaussian, *IR_sLs* and *attSBM* perform poorly (see paper [Dreveton et al. 2023] or thesis [Fernandes 2023]).

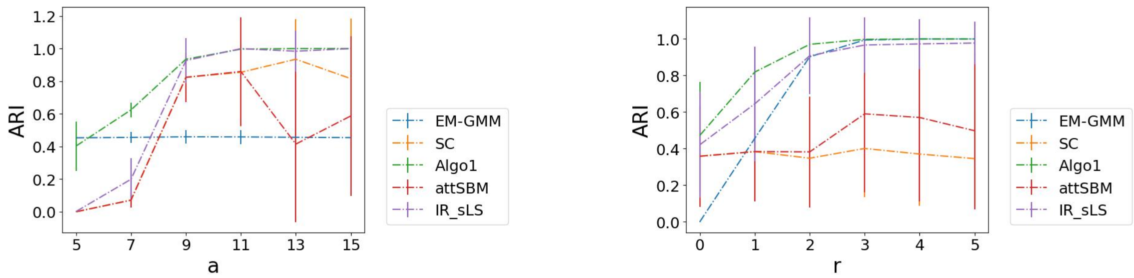
Due to space limitations, evaluation on real datasets have been suppressed and can be found in the related paper [Dreveton et al. 2023] or thesis [Fernandes 2023].

References

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

Abbe, E., Fan, J., and Wang, K. (2022). An ℓ_p theory of pca and spectral clustering. *The Annals of Statistics*, 50(4):2359–2385.

Abbe, E. and Sandon, C. (2015). Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *IEEE FOCS*.



(a) Varying edge distribution.

(b) Varying attribute distributions.

Figure 2. Performance on binary networks with Gaussian attributes. We take $n = 600$, $K = 2$, $= \mathrm{Ber}(an^{-1}\log n)$, $= \mathrm{Ber}(5n^{-1}\log n)$, and 2-dimensional Gaussian attributes with unit variance and mean $(\pm r, 0)$. Results are averaged over 60 runs.

Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J., and Lafferty, J. (2005). Clustering with bregman divergences. *Journal of machine learning research*, 6(10).

Braun, G., Tyagi, H., and Biernacki, C. (2022). An iterative clustering algorithm for the contextual stochastic block model with optimality guarantees. In *ICML*.

Combe, D., Largeron, C., Géry, M., and Egyed-Zsigmond, E. (2015). I-louvain: An attributed graph clustering method. In *Intl. Symp. on Intelligent Data Analysis*.

Deshpande, Y., Sen, S., Montanari, A., and Mossel, E. (2018). Contextual stochastic block models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Dreveton, M., Fernandes, F. S., and Figueiredo, D. R. (2023). Exact recovery and bregman hard clustering of node-attributed stochastic block model. In *NeurIPS*.

Fernandes, F. S. (2023). Clustering networks with node attributes using bregman divergences. Master's thesis, Federal University of Rio de Janeiro (UFRJ).

Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics reports*, 659:1–44.

Long, B., Zhang, Z. M., and Yu, P. S. (2007). A probabilistic framework for relational clustering. In *ACM SIGKDD*, pages 470–479.

Mariadassou, M., Robin, S., and Vacher, C. (2010). Uncovering latent structure in valued graphs: A variational approach. *The Annals of Applied Statistics*, 4(2):715 – 742.

Newman, M. E. and Clauset, A. (2016). Structure and inference in annotated networks. *Nature communications*, 7(1):11863.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., and Eliassi-Rad, T. (2008). Collective classification in network data. *AI magazine*, 29(3):93–93.

Stanley, N., Bonacci, T., Kwitt, R., Niethammer, M., and Mucha, P. J. (2019). Stochastic block models with multiple continuous attributes. *Applied Network Science*, 4(1):1–22.

Van Erven, T. and Harremos, P. (2014). Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.