Journal of Machine Learning **ISSN**: 2790-2048(e), 2790-203X(p)

Semi-Supervised Clustering of Sparse Graphs: Crossing the Information-Theoretic Threshold

Junda Sheng * [1] and Thomas Strohmer † [2]

[1]Department of Mathematics, University of California, Davis, CA 95616-5270, USA.
[2]Department of Mathematics and Center of Data Science and Artificial Intelligence Research, University of California, Davis, CA 95616-5270, USA.

Abstract. The stochastic block model is a canonical random graph model for clustering and community detection on network-structured data. Decades of extensive study on the problem have established many profound results, among which the phase transition at the Kesten-Stigum threshold is particularly interesting both from a mathematical and an applied standpoint. It states that no estimator based on the network topology can perform substantially better than chance on sparse graphs if the model parameter is below a certain threshold. Nevertheless, if we slightly extend the horizon to the ubiquitous semi-supervised setting, such a fundamental limitation will disappear completely. We prove that with an arbitrary fraction of the labels revealed, the detection problem is feasible throughout the parameter domain. Moreover, we introduce two efficient algorithms, one combinatorial and one based on optimization, to integrate label information with graph structures. Our work brings a new perspective to the stochastic model of networks and semidefinite program research.

Keywords:
Clustering,
Semi-supervised learning,
Stochastic block model,
Kesten-Stigum threshold,
Semidefinite programming.

Article Info.:
Volume: 3
Number: 1
Pages: 64 - 106
Date: March/2024
doi.org/10.4208/jml.230624

Article History:
Received: 24/06/2023
Accepted: 29/02/2024

Communicated by:
Qianxiao Li

1 Introduction

Clustering has long been an essential subject of many research fields, such as machine learning, pattern recognition, data science, and artificial intelligence. In this section, we include some background information on its general setting and the semi-supervised approach.

1.1 Clustering on graphs

The basic task of clustering or community detection in its general form is, given a (possibly weighted) graph, to partition its vertices into several densely connected groups with relatively weak external connectivity. This property is sometimes also called assortativity. Clustering and community detection are central problems in machine learning and data science with various applications in scientific research and industrial development. A considerable amount of data sets can be represented in the form of a network that consists of

*Corresponding author. sheng@math.ucdavis.edu
†strohmer@math.ucdavis.edu

interacting nodes, and one of the first features of interest in such a situation is to understand which nodes are similar, as an end or as a preliminary step towards other learning tasks. Clustering is used to find genetically similar sub-populations [60], to segment images [65], to study sociological behavior [59], to improve recommendation systems [47], to help with natural language processing [30], etc. Since the 1970s, in different communities like social science, statistical physics, and machine learning, a large diversity of algorithms have been developed such as:

- Hierarchical clustering algorithms [38] build a hierarchy of progressive communities, by either recursive aggregation or division.
- Model-based statistical methods, including the celebrated expectation-maximization (EM) clustering algorithm proposed in [23], fit the data with cluster-exhibiting statistical models.
- Optimization approaches identify the best cluster structures regarding carefully designed cost functions, for instance, minimizing the cut [34] and maximizing the Girvan-Newman modularity [58].

Multiple lines of research intersect at a simple random graph model, which appears under many different names. In the machine learning and statistics literature around social networks, it is called stochastic block model (SBM) [36], while it is known as the planted partition model [17] in theoretical computer science and referred to as inhomogeneous random graph model [14] in the mathematics literature. Moreover, it can also be interpreted as a spin-glass model [22], a sparse-graph code [5], a low-rank random matrix model [50], and more.

The essence of SBM can be summarized as follows: Conditioned on the vertex labels, edges are generated independently and the probability only depends on which clusters the pairs of vertices belong to. We consider its simplest form, namely the symmetric SBM consisting of two blocks, also known as the planted bisection model.

Definition 1.1 (Planted Bisection Model)**.** *For $n \in \mathbb{N}$ and $p, q \in (0,1)$, let $\mathcal{G}(n,p,q)$ denote the distribution over graphs with n vertices defined as follows. The vertex set is partitioned uniformly at random into two subsets S_1, S_2 with $|S_i| = n/2$. Let E denote the edge set. Conditional on this partition, edges are included independently with probability*

$$\mathrm{P}\left((i,j) \in E \,|\, S_1, S_2\right) = \begin{cases} p, & \text{if} \quad \{i,j\} \subseteq S_1 \quad \text{or} \quad \{i,j\} \subseteq S_2, \\ q, & \text{if} \quad i \in S_1, \quad j \in S_2 \quad \text{or} \quad i \in S_2, \quad j \in S_1. \end{cases} \tag{1.1}$$

Note that if $p = q$, the planted bisection model is reduced to the so-called Erdős-Rényi random graph where all edges are generated independently with the same probability. Hence, there exists no cluster structure. But if $p \gg q$, a typical graph will have two well-defined clusters. The scale of p and q also plays a significant role in the resulting graph, which will be discussed in detail later. They govern the amount of signal and noise in the graph's generating process. As the key parameters that researchers work with, they depict various regimes and thresholds.