Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

A spectral algorithm with additive clustering for the recovery of overlapping communities in networks

Emilie Kaufmann [a,*], Thomas Bonald [b,1], Marc Lelarge [c,1]

[a] CNRS & CRIStAL, Univ. Lille, France
[b] Telecom ParisTech, France
[c] Inria & Ecole Normale Supérieure, France

ARTICLE INFO

Article history:
Received 19 January 2017
Received in revised form 18 October 2017
Accepted 23 October 2017
Available online xxxx

Keywords:
Community detection
Stochastic blockmodel

ABSTRACT

This paper presents a novel *spectral algorithm with additive clustering*, designed to identify overlapping communities in networks. The algorithm is based on geometric properties of the spectrum of the expected adjacency matrix in a random graph model that we call *stochastic blockmodel with overlap* (SBMO). An adaptive version of the algorithm, that does not require the knowledge of the number of hidden communities, is proved to be consistent under the SBMO when the degrees in the graph are (slightly more than) logarithmic. The algorithm is shown to perform well on simulated data and on real-world graphs with known overlapping communities.

1. Introduction

Many datasets (e.g., social networks, gene regulation networks) take the form of graphs whose structure depends on some underlying *communities*. The commonly accepted definition of a community is that nodes tend to be more densely connected within a community than with the rest of the graph. Communities are often hidden in practice and recovering the community structure directly from the graph is a key step in the analysis of these datasets. Spectral algorithms are popular methods for detecting communities [26], that consist in two phases. First, a *spectral embedding* is built, where the n nodes of the graph are projected onto some low dimensional space generated by well-chosen eigenvectors of some matrix related to the graph (e.g., the adjacency matrix or a Laplacian matrix). Then, a *clustering algorithm* (e.g., k-means or k-median) is applied to the n embedded vectors to obtain a partition of the nodes into communities.

It turns out that the structure of many real datasets is better explained by *overlapping* communities. This is particularly true in social networks, in which the neighborhood of any given node is made of several social circles, that naturally overlap [19]. Similarly, in co-authorship networks, authors often belong to several scientific communities and in protein–protein interaction networks, a given protein may belong to several protein complexes [21]. The communities do not form a partition of the graph and new algorithms need to be designed. This paper presents a novel spectral algorithm, called spectral algorithm with additive clustering (SAAC). The algorithm consists in a spectral embedding based on the adjacency matrix of the graph, coupled with an additive clustering phase designed to find overlapping communities. The proposed algorithm does not require the knowledge of the number of communities present in the network, and can thus be qualified as adaptive.

* Corresponding author.
E-mail address: emilie.kaufmann@univ-lille1.fr (E. Kaufmann).
[1] Thomas Bonald and Marc Lelarge are members of the LINCS, Paris, France. See www.lincs.fr.

https://doi.org/10.1016/j.tcs.2017.12.028
0304-3975/

SAAC belongs to the family of model-based community detection methods, that are motivated by a random graph model depending on some underlying set of communities. In the non-overlapping case, spectral methods have been shown to perform well under the stochastic block model (SBM), introduced by Holland and Leinhardt [12]. Our algorithm is inspired by the simplest possible extension of the SBM to overlapping communities, we refer to as the stochastic blockmodel with overlaps (SBMO). In the SBMO, each node is associated to a binary membership vector, indicating all the communities to which the node belongs. We show that exploiting an additive structure in the SBMO leads to an efficient method for the identification of overlapping communities. To support this claim, we provide consistency guarantees when the graph is drawn under the SBMO, and we show that SAAC exhibit state-of-the-art performance on real datasets for which ground-truth communities are known.

The paper is structured as follows. In Section 2, we cast the problem of detecting overlapping communities into that of estimating a membership matrix in the SBMO model, introduced therein. In Section 3, we compare the SBMO with alternative random graph models proposed in the literature, and review the algorithms inspired by these models. In Section 4, we exhibit some properties of the spectrum of the adjacency matrix under SBMO, that motivate the new SAAC algorithm, introduced in Section 5, where we also formulate theoretical guarantees for an adaptive version of the algorithm. Section 6 illustrates the performance of SAAC on both real and simulated data and we discuss sparse SBMO in Section 7.

Notation We denote by $||x||$ the Euclidean norm of a vector $x \in \mathbb{R}^d$. For any matrix $M \in \mathbb{R}^{n\times d}$, we let M_i denote its i-th row and $M_{\cdot,j}$ its j-th column. For any $\mathcal{S} \subset \{1,\dots,d\}$, $|\mathcal{S}|$ denotes its cardinality and $\mathbb{1}_{\mathcal{S}} \in \{0,1\}^{1\times d}$ is a row vector such that $(\mathbb{1}_{\mathcal{S}})_{1,i} = \mathbb{1}_{\{i\in\mathcal{S}\}}$. The Frobenius norm of a matrix $M \in \mathbb{R}^{n\times d}$ is

$$||M||_F^2 = \sum_{i=1}^{n} ||M_i||^2 = \sum_{j=1}^{d} ||M_{\cdot,j}||^2 = \sum_{1\le i,j\le n} M_{i,j}^2.$$

The spectral norm of a symmetric matrix $M \in \mathbb{R}^{d\times d}$ with eigenvalues $\lambda_1,\dots,\lambda_d$ is $||M|| = \max_{i=1..d} |\lambda_i|$. We denote by $\mathfrak{S}_K$ the set of permutation of $\{1,\dots,K\}$ and for $\sigma \in \mathfrak{S}_K$, by $P_\sigma \in \mathbb{R}^{K\times K}$ the permutation matrix associated to σ, defined by $(P_\sigma)_{k,l} = \delta_{\sigma(k),l}$.

2. The stochastic blockmodel with overlaps (SBMO)

2.1. The model

For any symmetric matrix $A \in [0,1]^{n\times n}$, let $\hat{A}$ be some random symmetric binary matrix whose entries $(\hat{A}_{i,j})_{i\le j}$ are independent Bernoulli random variables with respective parameters $(A_{i,j})_{i\le j}$. Then $\hat{A}$ is the adjacency matrix of an undirected random graph with expected adjacency matrix A. In all the paper, we restrict the hat notation to variables that depend on this random graph. For example, the empirical degree of node i observed on the random graph and the expected degree of node i are respectively denoted by

$$\hat{d}_i = \sum_{j=1}^{n} \hat{A}_{i,j} \quad \text{and} \quad d_i = \sum_{j=1}^{n} A_{i,j}.$$

Similarly, we write $\hat{D} = \mathrm{Diag}(\hat{d}_i)$, $D = \mathrm{Diag}(d_i)$, and

$$\hat{d}_{\max} := \max_i \sum_{j=1}^{n} \hat{A}_{i,j}, \quad d_{\max} = \max_i \sum_{j=1}^{n} A_{i,j}.$$

The stochastic block model (SBM) with n nodes and K communities depends on some mapping $k : \{1,\dots,n\} \to \{1,\dots,K\}$ that associates nodes to communities and on some symmetric community connectivity matrix $B \in [0,1]^{K\times K}$. In this model, two nodes i and j are connected with probability

$$A_{i,j} = B_{k(i),k(j)} = B_{k(j),k(i)}.$$

Introducing a membership matrix $Z \in \{0,1\}^{n\times K}$ such that $Z_{i,k} = \mathbb{1}_{\{k(i)=k\}}$, the expected adjacency matrix can be written

$$A = ZBZ^T.$$

The stochastic blockmodel with overlap (SBMO) is a slight extension of this model, in which Z is only assumed to be in $\{0,1\}^{n\times K}$ and $Z_i \neq 0$ for all i. Compared to the SBM, the rows of the membership matrix Z are no longer constrained to have only one non-zero entry. Since these n rows give the communities of the respective n nodes of the graph, this means that each node can now belong to several communities. Note that the SBMO yields implicit constraints on B and Z, that should satisfy for all $i,j \in \{1,\dots,n\}$ that $Z_i B Z_j^T$ is smaller than 1.

Please cite this article in press as: E. Kaufmann et al., A spectral algorithm with additive clustering for the recovery of overlapping communities in networks, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2017.12.028