The Stochastic Topic Block Model for the Clustering of Vertices in Networks with Textual Edges

Rawya ZREIK[1,2] , Pierre LATOUCHE[1] & Charles BOUVEYRON [2]

[1] *Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne*
[2]*Laboratoire MAP5, UMR CNRS 8145, Université Paris Descartes*

Abstract. Due to the significant increase of communications between individuals via social media (Facebook, Twitter, Linkedin) or electronic formats (email, web, e-publication) in the past two decades, network analysis has become a unavoidable discipline. Many random graph models have been proposed to extract information from networks based on person-to-person links only, without taking into account information on the contents. This paper describes the stochastic topic block model (STBM) as introduced in [2], a probabilistic model for networks with textual edges. We address here the problem of discovering meaningful clusters of vertices that are coherent from both the network interactions and the text contents. A classification variational expectation-maximization (C-VEM) algorithm is proposed to perform inference. Finally, we demonstrate the effectiveness of our methodology on a real-word data set.

Keywords. Textual edges, stochastic block model, latent Dirichlet allocation, classification variational expectation-maximization. . .

1 Introduction

Ranging from communication to co-authorship networks, it is nowadays particularly frequent to observe networks with textual edges. It is obviously of strong interest to be able to model and cluster the vertices of those networks using information on both the network structure and the text contents. Techniques able to provide such a clustering would allow a deeper understanding of the studied networks. As a motivating example, Figure 1 shows a network made of 3 ”communities” of vertices where one of the communities can in fact be split into two separate groups based on the topics of communication between nodes of these groups (see legend of Figure 1 for details). Despite the important efforts in both network and text analysis, only a few methods have focused on the joint modeling of network vertices and textual edges.

Here we propose a new generative model for the clustering of vertices in networks with textual edges, such as communication or co-authorship networks. Contrary to existing approaches which are based on either too simple or highly-parametrized models, our model relies on the stochastic block model (SBM) [3] which offers a sufficient flexibility, with a reasonable complexity. This model is one of the few able to recover different topological structures such as communities, stars or disassortative clusters. Regarding the modeling of texts, our approach is based on the latent Dirichlet allocation (LDA) model proposed

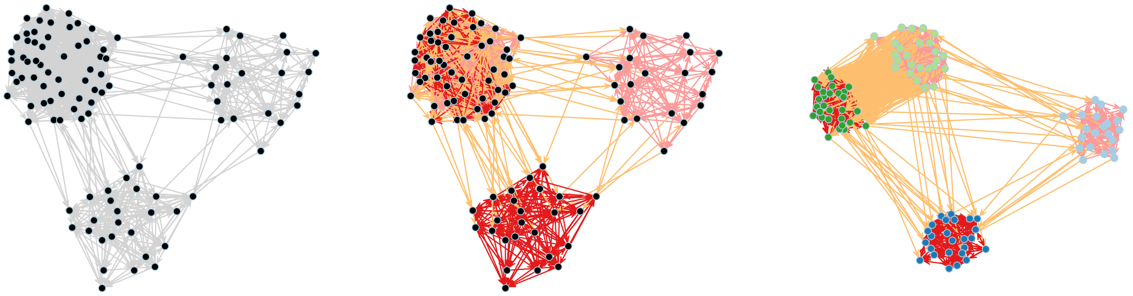

Figure 1: A sample network made of 3 "communities". The left panel only shows the observed (binary) edges in the network, the center panel shows the network with only the partition of edges into 3 topics (edge colors indicate the majority topics of texts). The right panel shows the network with the clustering of its nodes (vertex colors indicate the groups) and the majority topic of the edges.

by [1], in which the words are conditioned on the latent topics. Thus, the proposed modeling is able to exhibit node partitions that are meaningful both regarding the network structure and the topics, with a model of limited complexity, highly interpretable, for both directed and undirected networks. In addition, the proposed inference procedure allows the use of our model on large-scale networks.

2 The STBM model

This section presents the notations used in the paper and describes the STBM model. The joint distributions of the model to create edges and the corresponding documents are also given.

2.1 Context and notations

A directed network with M vertices, described by its $M \times M$ adjacency matrix A, is considered. Thus, $A_{ij} = 1$ if there is an edge from vertex i to vertex j, 0 otherwise. The network is assumed not to have any self-loop and therefore $A_{ii} = 0$ for all i. If an edge from i to j is present, then it is characterized by a set of D_{ij} documents, denoted $W_{ij} = (W_{ij}^d)_d$. Each document W_{ij}^d is made of a collection of N_{ij}^d words $W_{ij}^d = (W_{ij}^{dn})_n$. In the directed scenario considered, W_{ij} for instance can model a set of emails or text messages sent from actor i to actor j. Note that all the methodology proposed in this paper easily extends to undirected networks. In such a case, $A_{ij} = A_{ji}$ and $W_{ij}^d = W_{ji}^d$ for all i and j. The set W_{ij}^d of documents can then for example model books or scientific papers written by both i and j. In the following, we denote $W = (W_{ij})_{ij}$ the set of all exchanged documents, for all the edges present in the network.

Our goal is to cluster the vertices into Q latent groups sharing homogeneous connection profiles, *i.e.* find an estimate of the set $Y = (Y_1, \ldots, Y_M)$ of latent variables Y_i. The clustering task then consists in building groups of vertices having similar trends to connect to others. Therefore, Two nodes in the same cluster should have the same trend to connect to others, and when connected, the documents they are involved in should be made of words related to similar topics.

2.2 Modeling the presence of edges

In order to model the presence of edges between pairs of vertices, a stochastic block model [3] is considered. Thus, the vertices are assumed to be spread into Q latent clusters such that $Y_{iq} = 1$ if vertex i belongs to cluster q, and 0 otherwise. In practice, the binary vector Y_i is assumed to be drawn from a multinomial distribution

$$Y_i \sim \mathcal{M}\left(1, \rho = (\rho_1, \ldots, \rho_Q)\right),$$

where ρ denotes the vector of class proportions. By construction, $\sum_{q=1}^{Q} \rho_q = 1$ and $\sum_{q=1}^{Q} Z_{iq} = 1, \forall i$.

An edge from i to j is then sampled from a Bernoulli distribution, depending on their respective clusters

$$A_{ij} | Y_{iq} Y_{jr} = 1 \sim \mathcal{B}(\pi_{qr}). \tag{1}$$

In words, if i is in cluster q and j in r, then A_{ij} is 1 with probability π_{qr}.

All vectors Y_i are sampled independently, and given $Y = (Y_1, \ldots, Y_M)$, all edges in A are assumed to be independent. This leads to the following joint distribution

$$p(A, Y | \rho, \pi) = p(A | Y, \pi) p(Y | \rho).$$

2.3 Modeling the construction of documents

As mentioned previously, if an edge is present from vertex i to vertex j, then a set of documents $W_{ij} = (W_{ij}^d)_d$, characterizing the oriented pair (i, j), is assumed to be given. The STBM model relies on two concepts at the core of the SBM and LDA models respectively. On the one hand, a generalization of the SBM model would assume that any kind of relationships between two vertices can be explained by their latent clusters only. On the other hand, in the LDA model, the main assumption is that words in documents are drawn from a mixture distribution over topics, each document d having its own vector of topic proportions θ_d. The STBM model combines these two concepts to introduce a new generative procedure for documents in networks.

Each pair of clusters (q, r) of vertices is first associated with a vector of topic proportions $\theta_{qr} = (\theta_{qrk})_k$ sampled independently from a Dirichlet distribution

$$\theta_{qr} \sim \operatorname{Dir}\left(\alpha = (\alpha_1, \ldots, \alpha_K)\right),$$

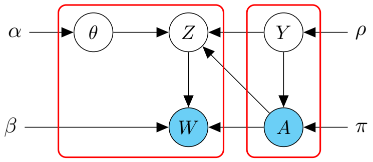

Figure 2: Graphical representation of the stochastic topic block model.

such that $\sum_{k=1}^{K} \theta_{qrk} = 1, \forall (q, r)$. We denote $\theta = (\theta_{qr})_{qr}$. The nth word W_{ij}^{dn} of documents d in W_{ij} is then associated with a latent topic vector Z_{ij}^{dn} assumed to be drawn from a multinomial distribution, depending on the latent vectors Y_i and Y_j

$$Z_{ij}^{dn} | \left\{ Y_{iq} Y_{jr} A_{ij} = 1, \theta \right\} \sim \mathcal{M}\left(1, \theta_{qr} = (\theta_{qr1}, \ldots, \theta_{qrK})\right). \tag{2}$$

Note that $\sum_{k=1}^{K} Z_{ij}^{dnk} = 1, \forall (i, j, d), A_{ij} = 1$. Equations (1) and (2) are related: they both involve the construction of random variables depending on the cluster assignment of vertices i and j. Thus, if an edge is present ($A_{ij} = 1$) and if i is in cluster q and j in r, then the word W_{ij}^{dn} is in topic k ($Z_{ij}^{dnk} = 1$) with probability θ_{qrk}.

Then, given Z_{ij}^{dn}, the word W_{ij}^{dn} is assumed to be drawn from a multinomial distribution

$$W_{ij}^{dn} | Z_{ij}^{dnk} = 1 \sim \mathcal{M}\left(1, \beta_k = (\beta_{k1}, \ldots, \beta_{kV})\right), \tag{3}$$

where V is the number of (different) words in the vocabulary considered and $\sum_{v=1}^{V} \beta_{kv} = 1, \forall k$ as well as $\sum_{v=1}^{V} W_{ij}^{dnv} = 1, \forall (i, j, d, n)$. Therefore, if W_{ij}^{dn} is from topic k, then it is associated with word v of the vocabulary ($W_{ij}^{dnv} = 1$) with probability β_{kv}. Equations (2) and (3) lead to the following mixture model for words over topics

$$W_{ij}^{dn} | \left\{ Y_{iq} Y_{jr} A_{ij} = 1, \theta \right\} \sim \sum_{k=1}^{K} \theta_{qrk} \mathcal{M}\left(1, \beta_k\right),$$

where the $K \times V$ matrix $\beta = (\beta_{kv})_{kv}$ of probabilities does not depend on the cluster assignments. Note that words of different documents d and $d^{'}$ in W_{ij} have the same mixture distribution which only depends on the respective clusters of i and j. We also point out that words of the vocabulary appear in any document d of W_{ij} with probabilities.

$$\mathbb{P}(W_{ij}^{dnv} = 1 | Y_{iq} Y_{jr} A_{ij} = 1, \theta) = \sum_{k=1}^{K} \theta_{qrk} \beta_{kv}.$$

Figure 2 presents the graphical model for STBM.

3 Inference

We aim at maximizing the log-likelihood

$$\log p(A, W, Y | \rho, \pi, \beta) = \log \sum_{Z} \int_{\theta} p(A, W, Y, Z, \theta | \rho, \pi, \beta) d\theta, \tag{4}$$

with respect to the model parameters (ρ, π, β) and the set $Y = (Y_1, \ldots, Y_M)$ of cluster membership vectors. Note that Y is not seen here as a set of latent variables over which the log-likelihood should be integrated out, as in standard expectation maximization (EM) or variational EM algorithms. Moreover, the goal is not to provide any approximate posterior distribution of Y given the data and model parameters. Conversely, Y is seen here as a set of (binary) vectors for which we aim at providing estimates. This choice is motivated by a key property of the STBM model: for a given Y, the full joint distribution factorizes into a LDA like term and SBM like term. In particular, given Y, words in W can be seen as being drawn from a LDA model with $D = Q^2$ documents.

4 Analysis of the Enron email network

We consider here a classical communication network, the Enron data set, which contains all email communications between 149 employees of the famous company from 1999 to 2002. Here, we focus on the period September, 1st to December, 31th, 2001.

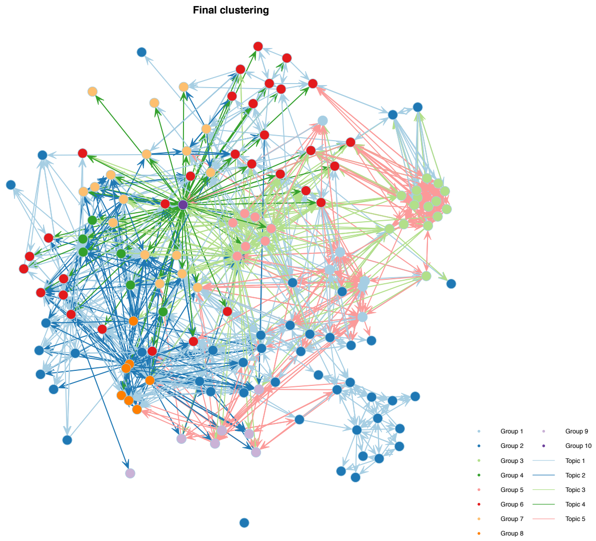

Figure 3: Clustering result with STBM on the Enron data set (Sept.-Dec. 2001).

The data set considered here contains 20 940 emails sent between the $M = 149$ employees. All messages sent between two individuals were coerced in a single meta-message. Thus, we end up with a data set of 1 234 directed edges between employees, each edge carrying the text of all messages between two persons.

The C-VEM algorithm we developed for STBM was run on these data for a number Q of groups from 1 to 14 and a number K of topics from 2 to 20. Figure 3 shows the clustering obtained with STBM for 10 groups of nodes and 5 topics. As previously, edge colors refer to the majority topics for the communications between the individuals. The

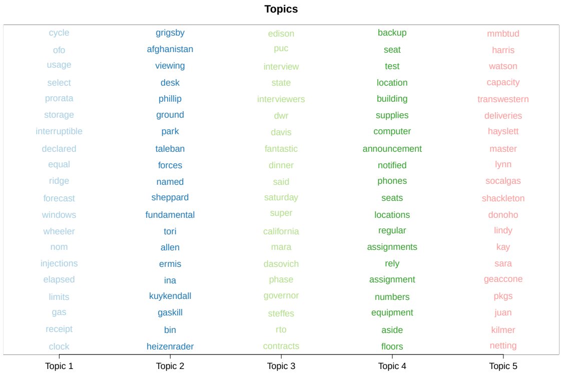

Figure 4: Most specific words for the 5 found topics with STBM on the Enron data set.

found topics can be easily interpreted by looking at the most specific words of each topic, displayed in Figure 4. In a few words, we can summarize the found topics as follows:
- Topic 1 seems to refer to the financial and trading activities of Enron,
- Topic 2 is concerned with Enron activities in Afghanistan (Enron and the Bush administration were suspected to work secretly with Talibans up to a few weeks before the 9/11 attacks),
- Topic 3 contains elements related to the California electricity crisis, in which Enron was involved, and which almost caused the bankruptcy of SCE-corp (Southern California Edison Corporation) early 2001,
- Topic 4 is about usual logistic issues (building equipment, computers, ...),
- Topic 5 refers to technical discussions on gas deliveries.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] C. Bouveyron, L. Latouche, and R. Zreik. The stochastic topic block model for the clustering of vertices in networks with textual edges. *Statistics and Computing*, page pages DOI, 2016.

[3] K. Nowicki and T. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.