Learning Influence-Receptivity Network Structure with Guarantee

Ming Yu
University of Chicago

Varun Gupta
University of Chicago

Mladen Kolar
University of Chicago

Abstract

Traditional works on community detection from observations of information cascade assume that a single adjacency matrix parametrizes all the observed cascades. However, in reality the connection structure usually does not stay the same across cascades. For example, different people have different topics of interest, therefore the connection structure depends on the information/topic content of the cascade. In this paper we consider the case where we observe a sequence of noisy adjacency matrices triggered by information/event with different topic distributions. We propose a novel latent model using the intuition that a connection is more likely to exist between two nodes if they are interested in similar topics, which are common with the information/event. Specifically, we endow each node with two node-topic vectors: an influence vector that measures how influential/authoritative they are on each topic; and a receptivity vector that measures how receptive/susceptible they are to each topic. We show how these two node-topic structures can be estimated from observed adjacency matrices with theoretical guarantee on estimation error, in cases where the topic distributions of the information/event are known, as well as when they are unknown. Experiments on synthetic and real data demonstrate the effectiveness of our model and superior performance compared to state-of-the-art methods.

1 INTRODUCTION

Uncovering latent network structure is an important research area in network model and has a long history

[33, 7]. For a p node network, traditional approaches usually assume a single $p \times p$ adjacency matrix, either binary or real-valued, that quantifies the connection intensity between nodes, and aim to learn the community structure from it. For example, in Stochastic Block Model (SBM) [17] we assume that nodes within a group have an edge with each other with probability p_0 while nodes across groups have an edge with probability q_0 where $p_0 > q_0$. In information diffusion we observe the propagation of information among nodes and aim to recover the underlying connections between nodes [28, 14, 13]. In time-varying networks we allow the connections and parameters to change over time [23, 2]. In this paper, we consider the case where we have a sequence of information/event/collaboration with different topics, and we observe a noisy adjacency matrix for each of them. The connection between nodes varies under each topic distribution and this cannot be captured by only one adjacency matrix. For example, each researcher has her own research interests and would collaborate with others only on the areas they are both interested in. Specifically, suppose researcher 1 is interested in computational biology and information theory; researcher 2 is interested in computational biology and nonparametric statistics; researcher 3 is interested in information theory only. Then if researcher 1 wants to work on computational biology, she would collaborate with researcher 2; while if the topic is on information theory, then she would collaborate with researcher 3. As another example, suppose student 1 is interested in music and sports while student 2 is interested in music and chess. If the topic of a University event is music, then there will be an edge between these two students; however, if the topic of the event is sports or chess, then there would not be an edge between them.

Intuitively, for a specific information/event/collaboration, there will be an edge between two nodes if and only if they are both interested in the topic of this information/event/collaboration. In this paper we model this intuition by giving each node two node-topic vectors: one *influence* vector (how authoritative they are on each topic) and one *receptivity* vector (how susceptible they are on each topic). In addition, each information/event/collaboration is

associated with a distribution on topics. The influence and receptivity vectors are fixed but different topic distributions result in different adjacency matrices among nodes. In this paper we consider both cases where the topic distribution may or may not be known, and provide algorithms to estimate the node-topic structure with theoretical guarantees on estimation error. In particular, we show that our algorithm converges to the true values up to statistical error. Our node-topic structure is easier to interpret than a large adjacency matrix among nodes, and the result can be used to make targeted advertising or recommendation systems.

Notation In this paper we use p to denote the number of nodes in the network; we assume there are K topics in total, and we observe n adjacency matrices under different topic distributions. We use subscript $i \in \{1, \ldots, n\}$ to index samples/observations; subscript $j, \ell \in \{1, \ldots, p\}$ to index nodes; and subscript $k \in \{1, \ldots, K\}$ to index topic. For any matrix A, we use $\|A\|_0 = |(j, k) : A_{jk} \neq 0|$ to denote the number of nonzero elements of A. Also, for any d, I_d is the identity matrix with dimension d.

2 MODEL

Our model to capture the node-topic structure in networks is built on the intuition that, for a specific information/event/collaboration, there would be an edge between two nodes if they are interested in similar topics, which are also common with that of the information/event/collaboration. Furthermore, the connection is *directed* where an edge from node 1 to node 2 is more likely to exist if node 1 is influential/authoritative in the topic, and node 2 is receptive/susceptible to the topic. For example, an eminent professor would have a large influence value (but maybe a small receptivity value) on his/her research area, while a high-producing, young researcher would have a large receptivity value (but maybe a small influence value) on his/her research area. Note that the notion of "topic" can be very general. For example it can be different immune systems: different people have different kinds of immune systems, and a disease is more likely to propagate between people with similar and specific immune system.

Our node-topic structure is parametrized by two matrices $B_1, B_2 \in \mathbb{R}^{p \times K}$. The matrix B_1 measures how much a node can infect others (the *influence* matrix) and the matrix B_2 measures how much a node can be infected by others (the *receptivity* matrix). We use b_{jk}^1 and b_{jk}^2 to denote the elements of B_1 and B_2, respectively. Specifically, b_{jk}^1 measures how influential node j is on topic k, and b_{jk}^2 measures how receptive node j is on topic k. We use b_k^1 and b_k^2 to denote the columns of B_1 and B_2, respectively.

Each observation i is associated with a topic distribution $m_i = (m_{i1}, \ldots, m_{iK})$ on the K topics satisfying $m_{i1}, \ldots, m_{iK} \geq 0$ and $m_{i1} + \ldots + m_{iK} = 1$. The choice of K can be heuristic and pre-specified or alternatively can be decided by methods such as in [18] which learn the distribution over the number of topics. For each observation i, the true adjacency matrix is given by

$$\left(x_i^*\right)_{j\ell} = \sum_{k=1}^{K} b_{jk}^1 \cdot m_{ik} \cdot b_{\ell k}^2, \tag{1}$$

or in matrix form,

$$X_i^* = B_1 \cdot M_i \cdot B_2^\top, \tag{2}$$

where M_i is a diagonal matrix

$$M_i = \text{diag}(m_{i1}, m_{i2}, \ldots, m_{iK}).$$

The interpretation of the model is straightforward from (1). For an observation i on topic k, there will be an edge $j \to \ell$ if and only if node j tends to infect others on topic k (large b_{jk}^1) and node ℓ tends to be infected by others on topic k (large $b_{\ell k}^2$). This intuition applies to each topic k and the final value is the summation over all the K topics.

If we do not consider self connections, we can zero out the diagonal elements and get

$$X_i^* = B_1 M_i B_2^\top - \text{diag}(B_1 M_i B_2^\top).$$

For notational simplicity, we still stick to (2) for the definition of X_i^* in the subsequent sections. The data consists of n observations $\{X_i\}_{i=1}^n$ satisfying

$$X_i = X_i^* + E_i, \tag{3}$$

where the noise term E_i are mean 0 and independent across i. They are not necessarily identically distributed and can follow an unstructured distribution. The observations X_i can be either real-valued or binary. For binary observations we are interested in the existence of a connection only, while for real-valued observation we are also interested in how strong the connection is, i.e. larger values indicate stronger connections.

Related Works There is a vast literature on uncovering latent network structures. The most common and basic model is the Stochastic block model (SBM) [17] where connections are assumed to be dense within group and are sparse across groups. The exact recovery of SBM can be solved using maximum likelihood method but is NP-hard. Many practical algorithms have been proposed for SBM such as Modularity method, EM algorithm, Spectral clustering, etc [6, 21, 29, 26, 30, 32]. Many variants and extensions of SBM have also been developed to better fit real world network structures, including Degree-corrected

block model (DCBM) [22], Mixed membership stochastic block models (MMSB) [4], Degree Corrected Mixed Membership (DCMM) model [20], etc. Other models include information diffusion [28, 14, 13, 38], time-varying networks [23, 2], conjunctive Boolean networks [11, 19, 10], graphical models [1, 5, 36], buyer-seller networks [24, 34, 31], etc. [16] and [15] assume a "logistic" model based on covariates to determine whether an edge exists or not. However, most of the existing work focuses on a single adjacency matrix and ignores the node-topic structure. In [35] the authors propose a node-topic model for information diffusion problem, but it requires the topic distribution to be known and lacks theoretical guarantees.

In [3] the authors study multiple adjacency matrices but it still falls into the SBM framework. The number of blocks need to be predefined (the performance is sensitive to this value) and the output is the block information. As a contrast, our model outputs the (numeric) influence-receptivity information for each node and these nodes do not need to form blocks. Also, their work does not utilize the topic information.

In terms of topic-based network inference, a closely related work is [9] where the authors use K adjacency matrices to describe the network structure. However it ignores the node-topic structure and can only deal with the case where the topic distributions are known, while our method is able to learn the topic distribution and the network structure simultaneously. In Section 6 and 7 we show that our method outperforms this model on both synthetic and real dataset.

Another closely related work is [8] where the authors propose the graph embedding model which also gives each node two K dimensional "embedding" vector. However, our model is different in the following senses: 1. The topic information of our model is easier to interpret than the "embedding" vectors. The whole framework of our model is more interpretable: we know all the topics information and the topics of interest for each node. 2. We provide a generative model and thorough theoretical result (error analysis). 3. The graph embedding model focuses on only one observation while our model focuses on n observations with each observation having a different topic distribution. In our model, the influence and receptivity vectors interact with topic information, while the graph embedding model cannot deal with that.

If we add up all the adjacency matrices X_i to a single matrix X, then it is similar to the mixed membership stochastic block model (MMSB [4]) where $X = B * M * B^\top$ and M can be non-diagonal. Compared to MMSB, our model allows for asymmetry by considering "influence" and "receptivity"; our model considers that

information with a different topic can have different adjacency matrices; also, our model can be used to predict a future adjacency matrix given the topics. Finally, when we have n adjacency matrices, it is usually better to analyze them individually instead of adding them up, which may lead to information loss.

3 OPTIMIZATION

In this paper we consider the loss function

$$f(B_1, B_2) = \frac{1}{2n} \sum_{i=1}^{n} \left\| X_i - B_1 M_i B_2^\top \right\|_F^2. \quad (4)$$

Using the notation $B_1 = [b_1^1, ..., b_K^1]$ and $B_2 = [b_1^2, ..., b_K^2]$, we can rewrite (2) as

$$X_i^* = B_1 M_i B_2^\top = \sum_{k=1}^{K} m_{ik} \cdot b_k^1 b_k^{2\top}.$$

Denote $\Theta_k = b_k^1 b_k^{2\top}$; with some abuse of notation we can rewrite the loss function (4) as

$$f(\Theta) = f(\Theta_1, ..., \Theta_K) = \frac{1}{2n} \sum_{i=1}^{n} \left\| X_i - \sum_{k=1}^{K} m_{ik} \cdot \Theta_k \right\|_F^2. \quad (5)$$

From (5) we can see that solving for B_1, B_2 is equivalent to solving for rank-1 matrix factorization problem on Θ_k. This model is therefore not identifiable on B_1 and B_2, since if we multiply column k of B_1 by some scalar γ and multiply column k of B_2 by $1/\gamma$, the matrix X_i^* remains unchanged for any i, since $b_k^1 b_k^{2\top}$ does not change. Hence the loss function also remains unchanged. Therefore we need an additional regularization term to ensure a unique solution. Here we borrow an idea from matrix factorization literature and introduce the following regularization term

$$g(B_1, B_2) = \frac{\lambda}{2} \cdot \sum_{k=1}^{K} \left(\left\| b_k^1 \right\|_2^2 - \left\| b_k^2 \right\|_2^2 \right)^2.$$

This regularization term forces the 2-norm of each column of B_1 and B_2 to be the same. Intuitively, this means that, for each topic k, the total magnitudes of "influence" and "receptivity" are the same. This acts like a conservation law that the total amount of output should be equal to the total amount of input. At the minimizer, this regularization term is 0, and therefore we can pick any $\lambda > 0$. The final optimization problem is given by

$$\text{minimize} \quad \frac{1}{2n} \sum_{i=1}^{n} \left\| X_i - B_1 M_i B_2^\top \right\|_F^2$$

$$+ \frac{\lambda}{2} \cdot \sum_{k=1}^{K} \left(\left\| b_k^1 \right\|_2^2 - \left\| b_k^2 \right\|_2^2 \right)^2$$

$$\text{subject to} \quad B_1, B_2 \geq 0$$

Algorithm 1 Alternating proximal gradient descent

> **Initialize** $B_1^{(0)}$, $B_2^{(0)}$
> **for** $t = 1, ..., T$ **do**
> $$B_1^{(t+0.5)} = \left[B_1^{(t)} - \eta \nabla_{B_1} f\big(B_1^{(t)}, B_2^{(t)}\big) \right.$$
> $$\left. - \eta \nabla_{B_1} g\big(B_1^{(t)}, B_2^{(t)}\big) \right]_+$$
> $$B_1^{(t+1)} = \text{Hard}\big(B_1^{(t+0.5)}, s\big)$$
> $$B_2^{(t+0.5)} = \left[B_2^{(t)} - \eta \cdot \nabla_{B_2} f\big(B_1^{(t)}, B_2^{(t)}\big) \right.$$
> $$\left. - \eta \cdot \nabla_{B_2} g\big(B_1^{(t)}, B_2^{(t)}\big) \right]_+$$
> $$B_2^{(t+1)} = \text{Hard}\big(B_2^{(t+0.5)}, s\big)$$
> **end for**

Initialization. We initialize by solving the convex relaxation problem (5) without the rank-1 constraint on Θ_k, and apply rank-1 SVD on estimated $\widehat{\Theta}_k$, i.e., we keep only the largest singular value: $[u_k, s_k, v_k] = $ rank-1 SVD of $\widehat{\Theta}_k$. The initialization is given by $B_1^{(0)} = [u_1 s_1^{1/2}, ..., u_K s_K^{1/2}]$ and $B_2^{(0)} = [v_1 s_1^{1/2}, ..., v_K s_K^{1/2}]$. Being a convex relaxation, we can find the global minimum $\widehat{\Theta}_k$ of problem (5) by using gradient descent algorithm.

Algorithm. After the initialization, we alternately apply proximal gradient method [27] on B_1 and B_2 until convergence. In practice, each node would be interested in only a few topics and hence we would expect B_1 and B_2 to be sparse. To encourage sparsity we need an additional hard thresholding step on B_1 and B_2. The overall procedure is given in Algorithm 1. The operation $\text{Hard}(B, s)$ keeps the largest s elements of B and zeros out others; the operation $[B]_+$ keeps all positive values and zeros out others.

4 THEORETICAL RESULT

In this section we derive the theoretical results for our algorithm. We denote B_1^* and B_2^* as the true value and $\Theta_k^* = b_k^{1*} b_k^{2*\top}$ as the corresponding true rank-1 matrices. In this section we assume the topic distribution M_i is known. The case where M_i is unknown is considered in Section 5. All the detailed proofs are relegated to the Appendix. We start by stating the following two mild assumptions on the parameters of the problem.

Topic Condition (TC). Denote the Hessian matrix on Θ as

$$H_\Theta = \frac{1}{n} \begin{bmatrix} \sum_i m_{i1}^2 & \sum_i m_{i1} m_{i2} & \cdots & \sum_i m_{i1} m_{iK} \\ \sum_i m_{i1} m_{i2} & \sum_i m_{i2}^2 & \cdots & \sum_i m_{i2} m_{iK} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_i m_{i1} m_{iK} & \sum_i m_{i2} m_{iK} & \cdots & \sum_i m_{iK}^2 \end{bmatrix}.$$

We require $H_\Theta \succeq \mu_\Theta \cdot I_K$ for some constant $\mu_\Theta > 0$.

Intuitively, this condition requires that, the correlation among topic distributions in the n observations cannot be too large. This makes sense because if several topics are highly correlated with each other among the n observations, then clearly we cannot distinguish them. If we vectorize each Θ_k, the Hessian matrix of $f(\Theta)$ with respect to Θ is a $p^2 K$ by $p^2 K$ matrix and it can be shown that this Hessian matrix is given by $H_\Theta \otimes I_{p^2}$ where $\otimes$ is the Kronecker product. With this condition, the objective function (5) is strongly convex in Θ.

An immediate corollary of this condition is that the diagonal elements of H_Θ must be at least μ_Θ, i.e., for each topic k, we have $\frac{1}{n} \sum_{i=1}^n m_{ik}^2 \geq \mu_\Theta$. This means that at least a constant proportion of the observed data should focus on this topic. The necessity of this condition is also intuitive: if we only get tiny amount of data on some topic, then we cannot expect to recover the structure for that topic accurately.

Sparsity Condition (SC). Both B_1^* and B_2^* are sparse: $\|B_1^*\|_0 = \|B_2^*\|_0 = s^*$. (We use a single s^* for notational simplicity, but is not required).

Subspace distance. For matrix factorization problems, it is common to measure the subspace distance because the factorization $\Theta_k = b_k^1 b_k^{2\top}$ is not unique. Here since we know that Θ_k are exactly rank-1 and we have non-negativity constraints on B_1, B_2, we would not suffer from rotation issue (the only way to rotate scalar is ± 1, but with non-negative constraint, -1 is impossible). Therefore the subspace distance between $B = [B_1, B_2]$ and $B^* = [B_1^*, B_2^*]$ is just defined as

$$d^2(B, B^*) = \min_{o_k \in \{\pm 1\}} \sum_{k=1}^K \|b_k^1 - b_k^{1*} o_k\|_2^2 + \|b_k^2 - b_k^{2*} o_k\|_2^2$$
$$= \|B_1 - B_1^*\|_F^2 + \|B_2 - B_2^*\|_F^2.$$

Statistical error. Denote

$$\Omega = \big\{ \Delta : \Delta = [\Delta_1, ..., \Delta_K] \in \mathbb{R}^{pK \times p}, \text{rank}(\Delta_k) = 2,$$
$$\|\Delta_k\|_0 = s, \|\Delta\|_F = 1 \big\}.$$

The statistical error on Θ is defined as

$$e_{\text{stat}, \Theta} = \sup_{\Delta \in \Omega} \big\langle \nabla f_\Theta(\Theta^*), \Delta \big\rangle$$
$$= \sup_{\Delta \in \Omega} \sum_{k=1}^K \Big\langle \frac{1}{n} \sum_{i=1}^n E_i \cdot m_{ik}, \Delta_k \Big\rangle. \tag{7}$$

where E_i is the error matrix in (3) and $\langle A, B \rangle = \text{tr}(A^\top B)$ is the matrix inner product. Intuitively, this statistical error measures how much accuracy we can expect for the estimator. If we are within $c \cdot e_{\text{stat}}$ distance with the true value, then we are already optimal.

The statistical error depends on the sparsity level s. In practice, s is a hyperparameter and one can choose it as a relatively large value to avoid missing true nonzero values. If s is too large, then we include too many false positive edges. This usually does not affect performance too much, since these false positive edges tend to have small values. However, we lose some sparsity and hence interpretability. If we further assume that each node is interested in at least one but not most of the topics, then we have $s = O(p)$ and we can choose $s = c \cdot p$ where c can be a small constant. In this way, the effect of choosing s is minimal.

In this way we transform the original problem to a standard matrix factorization problem with K rank-1 matrices $\Theta_1, \ldots, \Theta_K$. A function $f(\cdot)$ is termed to be strongly convex and smooth if there exist constant μ and L such that

$$\frac{\mu}{2}\big\|Y - X\big\|_F^2 \le f(Y) - f(X) - \langle \nabla f(X), Y - X \rangle$$
$$\le \frac{L}{2}\big\|Y - X\big\|_F^2.$$

The objective function (5) is strongly convex and smooth in Θ. Since the loss function (5) is quadratic on each Θ_k, it is easy to see that the conditions are equivalent to $\mu \cdot I_K \preceq H_\Theta \preceq L \cdot I_K$. The lower bound is satisfied according to assumption (TC) with $\mu = \mu_\Theta$, and the upper bound is trivially satisfied with $L = L_\Theta = 1$. Therefore we see that the objective function (5) is strongly convex and smooth in Θ. The following lemma quantifies the accuracy of the initialization.

Lemma 1. *Suppose $\widehat{\Theta} = (\widehat{\Theta}_1, \ldots, \widehat{\Theta}_K)$ are the global minimum of the convex relaxation (5), then we have*

$$\sum_{k=1}^{K} \|\Theta_k^* - \widehat{\Theta}_k\|_F^2 \le \frac{2}{\mu_\Theta}\big\|\nabla f(\Theta^*)\big\|_F.$$

The bound we obtain from Lemma 1 scales with $n^{-1/2}$ and therefore can be small as long as we have enough samples. We are then ready for our main theorem. The following Theorem 2 shows that the iterates of Algorithm 1 converge linearly up to statistical error.

Theorem 2. *Suppose conditions (SC) and (TC) hold. We set the sparsity level $s = cs^*$. If the step size η satisfies $\eta \le \frac{1}{16\|B^{(0)}\|_2^2} \cdot \min\left\{\frac{1}{2(\mu_\Theta + L_\Theta)}, 1\right\}$, then for large enough n, after T iterations, we have*

$$d^2\big(B^{(T)}, B^*\big) \le \beta^T d^2\big(B^{(0)}, B^*\big) + C \cdot e_{\text{stat},\Theta}^2, \quad (8)$$

for some constant $\beta < 1$ and constant C.

Remark 3. *Although we focus on the simplest loss function (4), our analysis works for any general loss functions $f(B_1 M B_2^\top)$, as long as the initialization is good and the (restricted) strongly convex and smoothness conditions are satisfied. See [37] for more details.*

Remark 4. *For time complexity of Algorithm 1, calculating the gradient takes $O(p^2 K)$ time and hence taking average over all samples takes $O(np^2 K)$ time. The initialization step involves SVD; but we do not need to obtain the full decomposition since for each Θ_k we only need the singular vector corresponding to the largest singular value. Finally, the number of iteration T is such that β^T has the same order with the statistical error, which gives $T < O(\log n)$.*

5 LEARNING NETWORK AND TOPIC DISTRIBUTIONS JOINTLY

So far we have assumed that the topic distributions m_i for each sample i are given and fixed. However, sometimes we do not have such information. In this case we need to learn the topic distributions and the network structure simultaneously.

We denote m_i^* as the true topic distribution of observation i and $M = [m_1, ..., m_n]$ is the stack of all the topic distributions. The algorithm for joint learning is simply alternating minimization on B_1, B_2 and M. For fixed M, the optimization on B_1, B_2 is the same as before, and can be solved using Algorithm 1. For fixed B_1, B_2, it is straightforward to see that the optimization on M is separable for each i. For each i, we solve the following optimization problem to estimate $M_i = \text{diag}(m_i)$:

$$\begin{aligned}
&\text{minimize} \quad \big\|X_i - B_1 M_i B_2^\top\big\|_F^2 \\
&\text{subject to} \quad m_i \ge 0, 1^\top \cdot m_i = 1
\end{aligned} \quad (9)$$

This problem is convex in M_i and can be easily solved using projected gradient descent. Namely in each iteration we do gradient descent on M_i and then project to the simplex. The overall procedure is summarized in Algorithm 2. With some abuse of notation we write

$$f(\Theta, M) = \frac{1}{2n}\sum_{i=1}^{n}\Big\|X_i - \sum_{k=1}^{K} m_{ik} \cdot \Theta_k\Big\|_F^2. \quad (10)$$

Besides the scaling issue mentioned in Section 3, the problem now is identifiable only up to permutation of the position of the topics. However we can always permute M^* to match the permutation obtained in M. From now on we assume that these two permutations match and ignore the permutation issue. The statistical error on M is defined as

$$e_{\text{stat},M}^2 = \sum_{i,k}\Big[\nabla_{m_{ik}} f(\Theta^*, M^*)\Big]^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\langle E_i, \Theta_k^*\rangle^2.$$

Algorithm 2 Learning network structure and topic distributions jointly

---

Initialize B_1, B_2
while *tolerance* $> \epsilon$ **do**
    Optimize M according to (9) using projected
                       gradient descent.
    Optimize B_1, B_2 according to Algorithm 1
end while

The problem is much harder with unknown topic distribution. Similar to condition (TC), we need the following assumption on the Hessian matrix on M.

Diffusion Condition (DC). Denote the Hessian matrix on M as

$$H_M = \begin{bmatrix} \langle \Theta_1^*, \Theta_1^* \rangle & \langle \Theta_1^*, \Theta_2^* \rangle & \ldots & \langle \Theta_1^*, \Theta_K^* \rangle \\ \langle \Theta_2^*, \Theta_1^* \rangle & \langle \Theta_2^*, \Theta_2^* \rangle & \ldots & \langle \Theta_2^*, \Theta_K^* \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \Theta_K^*, \Theta_1^* \rangle & \langle \Theta_K^*, \Theta_2^* \rangle & \ldots & \langle \Theta_K^*, \Theta_K^* \rangle \end{bmatrix},$$

where $\langle A_1, A_2 \rangle = \mathrm{tr}(A_1^\top A_2)$ is the inner product of matrices A_1, A_2. We require that $H_M \succeq \mu_M \cdot I_K$ for some constant $\mu_M > 0$.

With this condition, the objective function (5) is strongly convex in M. The intuition is similar as in condition (TC). We require that Θ_k can be distinguished from each other.

Initialization. Denote $\overline{X}$ as the sample mean of X_i. We perform rank-K SVD on $\overline{X}$ and obtain $[\widetilde{U}, \widetilde{S}, \widetilde{V}] = \mathrm{rank}\text{-}K$ SVD of $\overline{X}$. Denote $\widetilde{X} = \widetilde{U}\widetilde{S}\widetilde{V}^\top = \sum_{k=1}^K \widetilde{\sigma}_k \widetilde{u}_k \widetilde{v}_k^\top$ and we initialize with $\Theta_k^{(0)} = K \cdot \widetilde{\sigma}_k \widetilde{u}_k \widetilde{v}_k^\top$. For this initialization to be reasonable, we assume the following condition.

Orthogonality Condition (OC). Let $B_1^* = Q_1 R_1$ and $B_2^* = Q_2 R_2$ be the QR decomposition of B_1^* and B_2^*, respectively. Denote A^* as a diagonal matrix with diagonal elements $\frac{1}{n} \sum_{i=1}^n m_{ik}^*$. Denote $R_1 A^* R_2^\top = A_{\mathrm{diag}} + A_{\mathrm{off}}$ where A_{diag} captures the diagonal elements and A_{off} captures the off-diagonal elements. We require that $\|A_{\mathrm{off}}\|_F \leq \rho_0$ for some constant ρ_0. Moreover, we require that $\frac{1}{n} \sum_{i=1}^n m_{ik}^* \leq \eta/K$ for some η.

The detailed rationale of this initialization approach and the Orthogonality Condition is relegated to Appendix B. It is useful to point out that the condition (OC) is for this specific initialization method only. In practice we can also use other initialization methods, for example we can do alternating gradient descent on Θ and M based on the objective function (10). This method also works reasonably well in practice. The following lemma shows that $\Theta_k^{(0)}$ is indeed a good initialization for Θ_k^*.

Lemma 5. *Suppose the condition (OC) is satisfied, then the initialization $\Theta_k^{(0)}$ satisfies*

$$\left\| \Theta_k^{(0)} - \Theta_k^* \right\|_F \leq 2\widetilde{C}K\rho_0 + (\eta - 1)\sigma_{\max},$$

for some constant $\widetilde{C}$ where $\sigma_{\max} = \max_k \|\Theta_k^*\|_2$.

The initialization $\Theta_k^{(0)}$ is no longer $\sqrt{n}$-consistent. Nevertheless it is not required. With this initialization, we then follow Algorithm 2 and estimate B_1, B_2 and M alternatively. Note that when estimating B_1 and B_2, we run Algorithm 1 for large enough T so that the first term in (8) is small compared to the second term. These T iterations for Algorithm 1 are one iteration for Algorithm 2 and we use $B^{[t]} = [B_1^{[t]}, B_2^{[t]}]$ and $M^{[t]}$ to denote the iterates we obtained from Algorithm 2. Denote $d^2(M, M^*) = \frac{1}{n} \sum_{i=1}^n \sum_{k_0=1}^K (m_{ik_0} - m_{ik_0}^*)^2$. We obtain the following theorem on estimation error for jointly learning.

Theorem 6. *Suppose the conditions in Theorem 2 hold and suppose condition (DC) and (OC) hold. For large enough n, after T iterations of Algorithm 2 we have*

$$d^2(B^{[T]}, B^*) + d^2(M^{[T]}, M^*) \leq \frac{C_1 e_{\mathrm{stat},M}^2 + C_2 e_{\mathrm{stat},\Theta}^2}{1 - \beta_0}$$
$$+ \beta_0^T \left[d^2(B^{[0]}, B^*) + d^2(M^{[0]}, M^*) \right],$$

for some constant $\beta_0 < 1$, which shows linear convergence up to statistical error.

6 SIMULATION

In this section we evaluate our model and algorithms on synthetic datasets. We first consider the setting where the topics are known and we consider $p = 200$ nodes with $K = 10$ topics. The true matrices B_1^* and B_2^* are generated row by row where we randomly select 1-3 topics for each row and set a random value generated from $\mathsf{Uniform}(1, 2)$. All the other values are set to be 0. This gives sparsity level $s^* = 2p = 400$ in expectation, and we set $s = 2s^*$ in the algorithm as the hard thresholding parameter. For each observation, we randomly select 1-3 topics and assign each selected topic a random value $\mathsf{Uniform}(0, 1)$, and 0 otherwise. We then normalize this vector to get the topic distribution m_i. The true value X_i^* is generated according to (2). Note that X_i^* is also a sparse matrix. We consider two types of observation: real valued observation and binary valued observation. For real valued observation, we generate X_i (equivalently, set E_i) in the following way: first we randomly select 10% of the nonzero values in X_i^* and set to 0 (miss some edges); second for each of the remaining nonzero values, we generate an independent random number $\mathsf{Uniform}(0.3, 3)$ and multiply with the original value (observe edges with noise); finally we
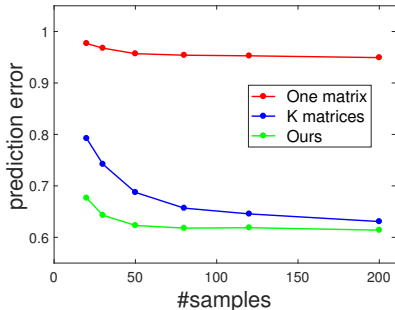
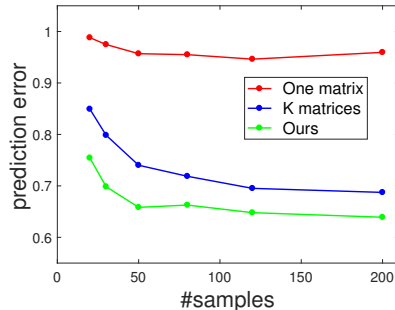Figure 1: Prediction error for real-valued observation, with known topics

Figure 2: Prediction error for real-valued observation, with unknown topics

randomly select 10% of the zero values in X_i^* and set them as $\mathsf{Uniform}(0,1)$ (false positive edges). For binary observations, we treat the true values in X_i^* as probability of observing an edge, and generate X_i as $X_i = \mathsf{Bernoulli}(X_i^*)$. For those true values greater than 1 we just set X_i to be 1. Finally we again pick 10% false positive edges.

We vary the number of observations $n \in \{20, 30, 50, 80, 120, 200\}$ and compare our model with the following two state-of-the-art methods. The first method is inspired by [13] which ignores the topic information and uses one $p \times p$ matrix to capture the entire dataset (termed "One matrix"). This matrix is given by $\overline{X}$. The second method is inspired by [9] which considers the topic information and assigns each topic a $p \times p$ matrix (termed "K matrices"). However it still ignores the node-topic structure. For this model, we ignore the rank constraint and return the matrix Θ_k given by the initialization procedure. Note that "One matrix" method has p^2 parameters, "K matrices" has $p^2 K$ parameters, but our method has only $2pK$ parameters. Since we usually have $K \ll p$, we are able to use much fewer parameters to capture the network structure, and would not suffer too much from overfitting. For fair comparison, we also do hard thresholding on each of these $p \times p$ matrices with parameter $4p$. The comparison is done by evaluating the objective function on independent test dataset (prediction error). This prediction error is given by $\frac{1}{n} \sum_i ||X_i - \widehat{X}_i||_F^2$, where X_i is the observed value and $\widehat{X}_i$ is the predicted value. The predicted values take different forms for each method. For "One matrix" it is just $\overline{X}$; for "K matrices" it is the weighted sum of the K estimated matrices for each topic; for our model, the prediction is obtained by plugging in the estimated B_1 and B_2 into (2). Figure 1 and Figure 3 (in appendix D) show the comparison results for real valued and binary observation, respectively. Each result is based on 20 replicates. We can see that our method has the best prediction error since we are able to utilize the topic

information and the structure among nodes and topics; "One matrix" method completely ignores the topic information and ends up with bad prediction error; K matrices" method ignores the structure among nodes and topics and suffers from overfitting. As sample size goes large, "K matrices" method will behave closer to our model in terms of prediction error, since our model is a special case of the K matrices model. However, it still cannot identify the structure among nodes and topics and is hard to interpret.

We then consider the setting where the topics are unknown. We initialize and estimate B_1, B_2 and M according to the procedure described in Section 5; for "One matrix" method, the estimator is still given by $\overline{X}$; for "K matrices" method, we estimate Θ and M by alternating gradient method on the objective function (10). All the other setups are the same as the previous case. Figure 2 and Figure 4 (in appendix D) show the comparison results for real valued and binary observation, respectively. Again we see that our model behaves the best. These results demonstrate the superior performance of our model and algorithm compared with existing state-of-the-art methods.

Finally we check the running time of our method experimentally. Here we fix $n = 500, T = 50$ and vary K and p. The empirical running time is given in Table 2, where we see a linear dependency on K and quadratic dependency on p, in line with the claim in remark 4.

7 APPLICATION TO ARXIV DATA

In this section we evaluate our model on real dataset. The dataset we use is the ArXiv collaboration and citation network dataset on high energy physics theory [25, 12]. This dataset covers papers uploaded to ArXiv high energy physics theory category in the period from 1993 to 2003, and the citation network for each paper. For our experiment we treat each author as a node and each publication as an observation. For each publication i, we set the observation matrix X_i

Table 1: The influence matrix B_1 for citation dataset

	black hole energy chains	quantum model field theory	gauge theory field effective	algebra space group structure	states space noncommutative boundary	string theory supergravity supersymmetric
Christopher Pope		0.359		0.468		0.318
Arkady Tseytlin	0.223	0.565		0.25		
Emilio Elizalde		0.109				
Cumrun Vafa			0.85	0.623	0.679	0.513
Ashok Das		0.155	0.115	1.07		
Renata Kallosh	0.44	0.512		0.326	0.382	
Ian I. Kogan			0.193			
Barton Zwiebach	0.16			0.222	0.383	0.236
Ashoke Sen	0.319			0.523		0.571

Table 2: Running time (in second) with different K and p

	$K = 10$	$K = 20$	$K = 40$
$p = 100$	1.7	2.2	3.1
$p = 200$	4.1	5.2	7.5
$p = 400$	13.0	16.0	22.0

Table 3: Comparison of the 3 methods on test links for citation dataset

	train	test	# para	# nonzero
One matrix [13]	7.628	8.223	40000	7695
K matrices [9]	5.861	8.415	240000	19431
Our method	8.259	8.217	2400	1200

in the following way: the component $(x_i)_{j\ell} = 1$ if this paper is written by author j and cited by author ℓ, and $(x_i)_{j\ell} = 0$ otherwise. Since each paper has only a few authors, we consider a variant of our original model as $X_i^* = \left[B_1 M_i B_2^\top \right] \odot A_i$ where operator $\odot$ is component-wise product and $A_i \in \mathbb{R}^{p \times p}$ is an indictor matrix with $(a_i)_{j\ell} = 1$ if j is the author of this paper, and $(a_i)_{j\ell} = 0$ otherwise. This means for each paper, we only consider the influence behavior of its authors.

For our experiment we consider the top 200 authors with about top 10000 papers in terms of number of citations, and split the papers into 8000 training set and 2000 test set. We first do Topic modeling on the abstracts of all the papers and extract $K = 6$ topics as well as the topic distribution on each paper. We then treat this topic information as *known* and apply our Algorithm 1 to the training set and learn the two node-topic matrices. Due to space limits we only show matrix B_1 of 10 authors in Table 1. Detailed B_1 and B_2 matrices are relegated to Table 4 and Table 5 in Appendix C. The keywords of the 6 topics are shown at the head of the two tables and the first column of the two tables is the name of the author.

We then compare the node-topic structure to the research interests and publications listed by the authors themselves on their website. The comparison results show that our model is able to capture the research topics accurately. For example, Christopher Pope reports quantum gravity and string theory; Arkady Tseytlin reports quantum field theory; Emilio Elizalde reports quantum physics; Cumrun Vafa reports string theory;

Ashoke Sen reports string theory and black holes as their research areas in their webpages. These are all successfully captured by our method.

Finally we compare the result with "One matrix" and "K matrices" methods on test set. The comparison result is given in Table 3 for training error, testing error, number of total parameters, and number of nonzero parameters. Since our model has much fewer parameters, it has the largest training error. However we can see that our model has the best test error, and both the other two methods do not generalize to test set and suffer from overfitting. These results demonstrates that the topic information and node-topic structure do exist, and our model is able to capture them.

8 CONCLUSION

In this paper we propose an influence-receptivity model and show how this structure can be estimated with theoretical guarantee. Experiments show superior performance of our model on synthetic and real data, compared with existing methods. This influence-receptivity model also provides much better interpretability.

There are several future directions we would like to pursue. Currently the topic information is either learned from topic modeling and fixed, or is (jointly) learned by our model where we ignore the text information. It would be of interest to combine the influence-receptivity structure and topic modeling to provide more accurate results. Another extension would be allowing dynamic influence-receptivity structure over time.

References

[1] *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.

[2] Amr Ahmed and Eric P Xing. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106(29):11878–11883, 2009.

[3] Edo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.

[4] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014, 2008.

[5] Rina Foygel Barber and Mladen Kolar. Rocket: Robust confidence intervals via kendall's tau for transelliptical graphical models. *ArXiv e-prints, arXiv:1502.07641*, February 2015.

[6] Peter J Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[7] Ronald S Burt. The network structure of social capital. *Research in organizational behavior*, 22:345–423, 2000.

[8] Siheng Chen, Sufeng Niu, Leman Akoglu, Jelena Kovačević, and Christos Faloutsos. Fast, warped graph embedding: Unifying framework and one-click algorithm. *arXiv preprint arXiv:1702.05764*, 2017.

[9] Nan Du, Le Song, Hyenkyun Woo, and Hongyuan Zha. Uncover topic-sensitive information diffusion networks. In *Artificial Intelligence and Statistics*, pages 229–237, 2013.

[10] Zuguang Gao, Xudong Chen, and Tamer Başar. Controllability of conjunctive boolean networks with application to gene regulation. *IEEE Transactions on Control of Network Systems*, 5(2):770–781, 2018.

[11] Zuguang Gao, Xudong Chen, and Tamer Başar. Stability structures of conjunctive boolean networks. *Automatica*, 89:8–20, 2018.

[12] Johannes Gehrke, Paul Ginsparg, and Jon Kleinberg. Overview of the 2003 kdd cup. *ACM SIGKDD Explorations Newsletter*, 5(2):149–151, 2003.

[13] Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1019–1028. ACM, 2010.

[14] Manuel Gomez-Rodriguez, Le Song, Hadi Daneshmand, and Bernhard Schölkopf. Estimating diffusion networks: Recovery conditions, sample complexity & soft-thresholding algorithm. *Journal of Machine Learning Research*, 2015.

[15] Peter D Hoff. Multiplicative latent factor models for description and prediction of social networks. *Computational and mathematical organization theory*, 15(4):261, 2009.

[16] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

[17] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[18] Wei-Shou Hsu and Pascal Poupart. Online bayesian moment matching for topic modeling with unknown number of topics. In *NIPS*, 2016.

[19] Abdul Salam Jarrah, Reinhard Laubenbacher, and Alan Veliz-Cuba. The dynamics of conjunctive and disjunctive boolean network models. *Bulletin of Mathematical Biology*, 72(6):1425–1447, 2010.

[20] Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. Estimating network memberships by simplex vertex hunting. *arXiv preprint arXiv:1708.07852*, 2017.

[21] Brian Karrer and Mark EJ Newman. Message passing approach for general epidemic models. *Physical Review E*, 82(1):016101, 2010.

[22] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

[23] Mladen Kolar, Le Song, Amr Ahmed, and Eric P Xing. Estimating time-varying networks. *The Annals of Applied Statistics*, pages 94–123, 2010.

[24] Rachel E Kranton and Deborah F Minehart. A theory of buyer-seller networks. *American economic review*, 91(3):485–508, 2001.

[25] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM, 2005.

[26] Krzysztof Nowicki and Tom A B Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association*, 96(455):1077–1087, 2001.

[27] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[28] Manuel Gomez Rodriguez, David Balduzzi, and Bernhard Schölkopf. Uncovering the temporal dynamics of diffusion networks. *arXiv preprint arXiv:1105.0697*, 2011.

[29] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, pages 1878–1915, 2011.

[30] Tom AB Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification*, 14(1):75–100, 1997.

[31] Achim Walter, Thomas Ritter, and Hans Georg Gemünden. Value creation in buyer–seller relationships: Theoretical considerations and empirical results from a supplier's perspective. *Industrial marketing management*, 30(4):365–377, 2001.

[32] Yuchung J Wang and George Y Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.

[33] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[34] David T Wilson. An integrated model of buyer-seller relationships. *Journal of the academy of marketing science*, 23(4):335–345, 1995.

[35] Ming Yu, Varun Gupta, and Mladen Kolar. An influence-receptivity model for topic based information cascades. *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1141–1146, 2017.

[36] Ming Yu, Mladen Kolar, and Varun Gupta. Statistical inference for pairwise graphical models using score matching. In *Advances in Neural Information Processing Systems*, pages 2829–2837, 2016.

[37] Ming Yu, Zhaoran Wang, Varun Gupta, and Mladen Kolar. Recovery of simultaneous low rank and two-way sparse coefficient matrices, a nonconvex approach. *arXiv preprint arXiv:1802.06967*, 2018.

[38] Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multidimensional hawkes processes. In *Artificial Intelligence and Statistics*, pages 641–649, 2013.