
Diffusion Network Inference for Cross-layer Cascades

Siyu Huang*

Department of Statistics
The Pennsylvania State University
sph6006@psu.edu

Yubai Yuan*†

Department of Statistics
The Pennsylvania State University
yvy5509@psu.edu

Abdul Basit Adeel

Department of Sociology and Criminology
The Pennsylvania State University
axa6372@psu.edu

Abstract

A cascade over a network refers to the diffusion process where behavior changes occurring in one part of an interconnected population lead to a series of sequential changes throughout the entire population. In recent years, there has been a surge in interest and efforts to understand and model cascade mechanisms since they motivate many significant research topics across different disciplines. The propagation structure of cascades is governed by underlying diffusion networks that are often hidden. Inferring diffusion networks thus enables interventions in cascading process to maximize information propagation and provides insights into the Granger causality of interaction mechanisms among individuals. In this project, we propose a novel double network mixture model for inferring latent diffusion network in presence of strong cascade heterogeneity. The new model represents cascade pathways as a distributional mixture over diffusion networks that capture different cascading patterns at the population level. We develop a data-driven optimization method to infer diffusion networks using only visible temporal cascade records, avoiding the need to model complex and heterogeneous individual states. Both statistical and computational guarantees are established for the proposed method. We apply the proposed model to analyze research topic cascades in social sciences across U.S. universities and uncover the latent research topic diffusion network among top U.S. social science programs.

1 Introduction

Cascades over network refer to the diffusion processes where behavior changes in a part of an interconnected population lead to a series of sequential changes throughout the entire population. In recent years, there are surging interests and efforts to understand and model the cascade mechanism since it motivates many significant research topics in different areas, including social influence [12, 13, 20], information propagation via social media [32, 1], diffusion of policy and social norms [4, 33], viral marketing [25, 9], and contagion of infectious diseases [21, 35].

One fundamental problem is to understand the diffusion networks that govern cascade propagation patterns. However, diffusion networks are often hidden and need to be inferred from observed cascading behaviors. For example, in the case of infectious diseases, we can observe when an individual is infected but need to impute the missing information on who infects this individual.

*Co-first authors.

†Corresponding author.

More importantly, real-world cascading behaviors often exhibit strong heterogeneity and are jointly governed by different diffusion patterns. For example, in information cascade and epidemiology, cascades can diffuse among population via different transmission channels building on various social relations, and thus lead to heterogeneous propagation speeds and scales [24, 30]. Furthermore, cascading heterogeneity originates from the variability of individuals' statuses in engaging the cascade [39, 11]. In social media like X and Instagram, the spreading pattern and the speed of messages heavily depend on users' activity status. An active user will respond more instantly to interesting messages and accelerate the information cascade compared to inactive users [10]. As for volatility cascades in financial markets, structure heterogeneity in volatility diffusion depends on different time horizons of the agents in the market [41]. In these scenarios, individual statuses determine the transmission channels engaging in the cascades, and changes of individual statuses can also change the downstream cascade diffusion patterns. To conclude from these examples, diffusion patterns can exhibit combinatorial complexity as population grows. We present an example of cascade diffusion when transmission channels of individuals vary in Figure (1).

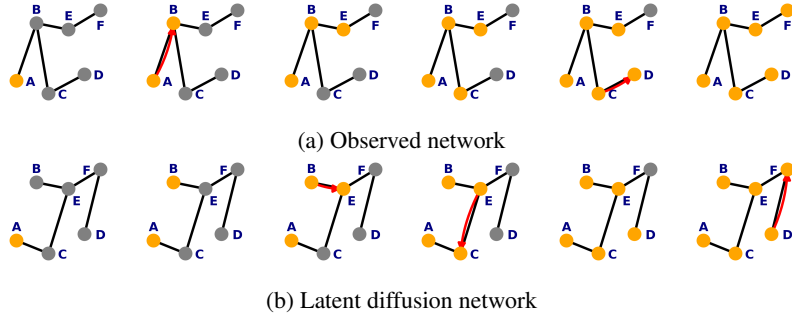


Figure 1: A cascade diffuses through nodes $\{A, B, C, D, E, F\}$ on a two-layer network. Yellow nodes: activated nodes; gray nodes: inactivated nodes. $\{B, D\}$ are activated via the observed network and $\{E, C, F\}$ are activated via the latent network.

In this paper, we propose a novel double network mixture model to infer multiple diffusion networks simultaneously from heterogeneous cascade data. The proposed model introduces a distributional mixture of diffusion networks to capture the heterogeneous cascading patterns, where diffusion networks provide complementary connection information. The main advantage of the distributional mixture is to avoid modeling the complex individual status changes. Specifically, the proposed model can describe the diffusion process over multi-layer networks where cascades propagate across different layers alternatively. Compared to existing methods, the proposed method can uncover latent diffusion networks even when the number of diffusion patterns is exponential to the number of nodes. Furthermore, the parameter estimation in our model can be solved by a sequence of convex optimization problems, which leads to both statistical and computational guarantees for our diffusion network estimation.

2 Related works

Various directed probabilistic graphical models have been developed to infer diffusion networks from observed cascade samples [17, 19, 31, 8, 22, 18]. Generally, these models treat infection time as a continuous random variable and construct the likelihood of cascade samples based on the target diffusion network under local Markov assumption. To capture heterogeneous diffusion patterns, several multi-pattern cascade models have been proposed [37, 40], where cascade samples are adaptively clustered into groups and each group corresponds to a distinct diffusion network. Among these methods, ConNle [29], NetRate [31], and MMRate [37] are popular representatives. Specifically, ConNle employs a maximum likelihood formulation via convex programming, incorporating an l_1 -type penalty to promote sparsity in the inferred network. Building upon ConNle, NetRate explicitly represents diffusion as a continuous-time probabilistic process, characterized by edge-specific transmission rates governing edge-wise diffusion probabilities. MMRate further extends this framework by accommodating multiple distinct diffusion patterns, assuming multiple heterogeneous latent networks, with each cascade diffusing via one network according to a certain probability.

3 Methodology

3.1 Continuous-time cascade on single network

Consider a network with N nodes where each node has two infection conditions in a single cascade: infected (activated) and uninfected (inactivated), and an infected node will always remain infected. A cascade is a N -dimensional temporal record $\mathbf{t} = (t_1, t_2, \dots, t_N)$, where t_i is the infection time of the i -th node. Instead of infinite time horizon, we observe a cascade within a finite time window of length T , i.e., $\forall i, t_i \in [t_0, t_0 + T]$, where $t_0 := \min_{1 \leq i \leq N} \{t_i\}$ is the infection time of source node. We denote the infection time of the nodes not infected in the observation window as $t_0 + T$. Without loss of generality, we assume $t_0 = 0$ in this paper.

A cascade diffuses nodewise over edges of a diffusion network. The continuous-time model formulates the cascade transmission from an infected node j to another infected node i with survival analysis models. Specifically, given a node j being infected at time t_j , let $f(t_i | t_j, \lambda_{ji})$ denote the likelihood of node i being infected by node j at time t_i , where $t_i \geq t_j$. The transmission rate λ_{ji} represents how fast cascades diffuse from node j to node i . Accordingly, the cumulative probability function is $F(t_i | t_j, \lambda_{ji}) = \int_{t_j}^{t_i} f(s | t_j, \lambda_{ji}) ds$. We consider the hazard rate function defined as

$$H(t_i | t_j, \lambda_{ji}) = \frac{f(t_i | t_j, \lambda_{ji})}{1 - F(t_i | t_j, \lambda_{ji})} = \frac{f(t_i | t_j, \lambda_{ji})}{S(t_i | t_j, \lambda_{ji})}, \quad S(t_i | t_j, \lambda_{ji}) = \exp\left(-\int_{t_j}^{t_i} H(t | t_j) dt\right),$$

where $S(t_i | t_j, \lambda_{ji}) = 1 - F(t_i | t_j, \lambda_{ji})$ is the survival function that indicates the probability of node i not being infected by node j until time t_i . Typical parametric forms of hazard rate functions are Exp model: $H(t_i | t_j, \lambda_{ji}) = \lambda_{ji}$; Pow model: $H(t_i | t_j, \lambda_{ji}) = \lambda_{ji} \frac{1}{t_i - t_j}$; Ray model: $H(t_i | t_j, \lambda_{ji}) = \lambda_{ji}(t_i - t_j)$. Given that λ_{ji} controls the likelihood and speed of transmission between node j to node i , the global cascading pattern over network can be captured by matrix $\mathbf{\Lambda} = (\lambda_{ij}) \in \mathbb{R}_+^{N \times N}$, where rows represent sender nodes and columns represent receiver nodes. Note that $\mathbf{\Lambda}$ can be asymmetric, i.e., $\lambda_{ij} \neq \lambda_{ji}$, when node i and j are different in the capacity of launching transmission.

The diffusion process is typically modeled by independent cascade models [22], where any infected nodes can infect a node independently and a node stays infected once another node infects it. Therefore, one can formulate the likelihood of node i being infected by potential parent nodes $\{j : t_j < t_i\}$ at time t_i as

$$\begin{aligned} P_I(t_i; \mathbf{\Lambda}_{\cdot i}) &:= P(t_i | \{t_j : t_j < t_i\}) = \sum_{j: t_j < t_i} f(t_i | t_j, \lambda_{ji}) \prod_{k: t_k < t_i, k \neq j} S(t_i | t_k, \lambda_{ki}) \quad (1) \\ &= \sum_{j: t_j < t_i} H(t_i | t_j, \lambda_{ji}) \prod_{k: t_k < t_i} S(t_i | t_k, \lambda_{ki}) \end{aligned}$$

On the other hand, if node i is not activated by any parent nodes in the observation window, then the corresponding likelihood for $\{i : t_i = T\}$ is

$$P_U(T; \mathbf{\Lambda}_{\cdot i}) := P(t_i | \{t_j : t_j < t_i\}) = \prod_{j: t_j < T} S(T | t_j, \lambda_{ji}), \quad (2)$$

Since infections of each node are conditionally independent given the corresponding parent nodes, we can decompose the likelihood of a cascade \mathbf{t} as the multiplication of a series of conditional probabilities:

$$P(\mathbf{t}; \mathbf{\Lambda}) = \prod_{i=1}^N P(t_i | \{t_j : t_j < t_i\}; \mathbf{\Lambda}_{\cdot i}) = \prod_{i=1}^N \{ \mathbb{1}(t_i < T) P_I(t_i; \mathbf{\Lambda}_{\cdot i}) + \mathbb{1}(t_i \geq T) P_U(T; \mathbf{\Lambda}_{\cdot i}) \}$$

Therefore, the full likelihood of independent cascade samples $\{\mathbf{t}^{(c)}\}_{c=1}^C$ is $\prod_{c=1}^C P(\mathbf{t}^{(c)}; \mathbf{\Theta})$. The advantage of the above model is that each cascade sample induces a directed acyclic graph, whose local Markov property allows the decomposition of likelihood and thus a reduction in computational complexity of inferring the diffusion network $\mathbf{\Theta}$.

3.2 Double network mixture model

In many applications such as social media, there exist multiple networks among an interconnected population that reflect different types of relations. Cascade diffusion pattern can change over different networks, and cascade can proceed on different networks alternatively and simultaneously due to the inter-layer interactions. The diffusion process on multi-layer networks can greatly increase the degree of freedom in possible diffusion pathways, which leads to strong heterogeneity in cascade observations. To model the multi-network cascade behavior, we propose the **double network mixture model**. Consider two diffusion networks Θ and Ψ among the same population with N units, where Θ_{ij} and Ψ_{ij} are the transmission rate from node i to node j corresponding to two relations. We introduce double diffusion indicators $Z_i^c \in \{0, 1\}$ such that

$$Z_i^c \stackrel{\text{i.i.d}}{\sim} \text{Bern}(\pi_i), \quad i = 1, \dots, N, \quad c = 1, \dots, C.$$

For cascade c , node i is activated via diffusion pathway on network Θ if $Z_i^c = 1$, and $Z_i^c = 0$ if via network Ψ . Denote $\pi = \{\pi_i\}_{i=1}^N \in [0, 1]^N$ where π_i is the probability of node i engaging cascade via network Θ . Then the diffusion pathway of the cascade c can be represented as

$$\Theta^c \in \mathbb{R}_+^{N \times N} = (Z_1^c \Theta_{\cdot 1} + (1 - Z_1^c) \Psi_{\cdot 1}, \dots, Z_N^c \Theta_{\cdot N} + (1 - Z_N^c) \Psi_{\cdot N}), \quad c = 1, \dots, C. \quad (3)$$

Therefore, Θ^c is column-wise mixture of Θ and Ψ , which can vary for different cascades. Different to conventional mixture model, the proposed method allows Z_i^c to vary across different cascade samples and nodes. Therefore, the proposed model (3) can generate up to 2^N different types of diffusion patterns, which grows exponentially as network size increases. Therefore, the proposed method has the principled heterogeneity modeling for the diffusion patterns. The likelihood of the proposed double mixture model can be also explicitly formulated. Following P_I and P_U in (1), we have the probability for node i being infected in the c -th cascade given specific network as:

$$P(t_i^c \mid Z_i^c; \{t_j^c : t_j^c < t_i^c\}, \Theta^c) = [P_I(t_i^c; \Theta_{\cdot i})]^{Z_i^c} \times [P_I(t_i^c; \Psi_{\cdot i})]^{1-Z_i^c}, \quad (4)$$

and following (2) the probability for node i not being infected is

$$P(T \mid Z_i^c; \{t_j^c : t_j^c < T\}, \Theta^c) = [P_U(T; \Theta_{\cdot i})]^{Z_i^c} \times [P_U(T; \Psi_{\cdot i})]^{1-Z_i^c}. \quad (5)$$

Denote $\mathbf{Z} = \{Z_i^c\}_{i=1, c=1}^{N, C}$, the joint distribution of diffusion indicators is

$$P(\mathbf{Z}) = \prod_{c=1}^C \prod_{i=1}^N P(Z_i^c) = \prod_{c=1}^C \prod_{i=1}^N \pi_i^{Z_i^c} (1 - \pi_i)^{1-Z_i^c}. \quad (6)$$

Then we have the joint distribution of cascade samples and diffusion indicators \mathbf{Z} as

$$\prod_{c=1}^C P(\mathbf{t}^{(c)}, \mathbf{Z}; \Omega) = \prod_{c=1}^C \left\{ \prod_{i: t_i^c \leq T} [\pi_i P_I(t_i^c; \Theta_{\cdot i})]^{Z_i^c} [(1 - \pi_i) P_I(t_i^c; \Psi_{\cdot i})]^{1-Z_i^c} \right. \quad (7)$$

$$\left. \times \prod_{j: t_j^c \geq T} [\pi_j P_U(T; \Theta_{\cdot j})]^{Z_j^c} [(1 - \pi_j) P_U(T; \Psi_{\cdot j})]^{1-Z_j^c} \right\}, \quad (8)$$

where $\Omega = (\pi, \Theta, \Psi)$ denote model parameters. And the marginal distribution of cascade samples is

$$\prod_{c=1}^C P(\mathbf{t}^{(c)}; \Omega) = \prod_{c=1}^C \left\{ \prod_{i: t_i^c \leq T} [\pi_i P_I(t_i^c; \Theta_{\cdot i}) + (1 - \pi_i) P_I(t_i^c; \Psi_{\cdot i})] \times \prod_{j: t_j^c \geq T} [\pi_j P_U(T; \Theta_{\cdot j}) + (1 - \pi_j) P_U(T; \Psi_{\cdot j})] \right\}. \quad (9)$$

Another advantage of our model is that the posterior distribution of diffusion indicator $\{Z_i^c\}$ can be calculated in an explicit form

$$\hat{\pi}_i^c := P(Z_i^c = 1 \mid \mathbf{t}^{(c)}) = \begin{cases} \frac{\pi_i P_I(t_i^c; \Theta_{\cdot i})}{\pi_i P_I(t_i^c; \Theta_{\cdot i}) + (1 - \pi_i) P_I(t_i^c; \Psi_{\cdot i})}, & \text{if } t_i^c < T \\ \frac{\pi_i P_U(T; \Theta_{\cdot i})}{\pi_i P_U(T; \Theta_{\cdot i}) + (1 - \pi_i) P_U(T; \Psi_{\cdot i})}, & \text{if } t_i^c \geq T \end{cases} \quad (10)$$

3.3 Model identification

In this subsection, we establish the identifiability of the double network mixture model. Denote $\mathcal{R}_i^\Theta := \{j \in \{1, \dots, N\} \mid \Theta_{ji} > 0\}$ and $\mathcal{R}_i^\Psi := \{j \in \{1, \dots, N\} \mid \Psi_{ji} > 0\}$ as the sets of nodes that can directly reach node i on Θ and Ψ , and $t_{\mathcal{R}_i} \in [0, T]^{\lvert \mathcal{R}_i^\Theta \cup \mathcal{R}_i^\Psi \rvert} := \{t_j \mid j \in \mathcal{R}_i^\Theta \cup \mathcal{R}_i^\Psi\}$ as the infectious times of these nodes. We have the following identifiability result for model in (9).

Proposition 1. *For each node i , $i = 1, \dots, N$, assume that 1) $\|\Theta_{\cdot i}\|_1 + \|\Psi_{\cdot i}\|_1 > 0$ and there exists $j \neq i$ such that $\Theta_{ji} \neq \Psi_{ji}$, and 2) survival function satisfies $S(t_i \mid t_j, \lambda_{ji}) = \exp\{\lambda_{ji}h(t_i - t_j)\}$ for some differentiable function $h(\cdot)$. Then, the parameters $(\pi_i, \Theta_{\cdot i}, \Psi_{\cdot i})$ associated with node i in (9) are identifiable, i.e., if*

$$\begin{aligned} \pi_i P_I(t; \Theta_{\cdot i}) + (1 - \pi_i) P_I(t; \Psi_{\cdot i}) &= \tilde{\pi}_i P_I(t; \tilde{\Theta}_{\cdot i}) + (1 - \tilde{\pi}_i) P_I(t; \tilde{\Psi}_{\cdot i}), \text{ or} \\ \pi_i P_U(t; \Theta_{\cdot i}) + (1 - \pi_i) P_U(t; \Psi_{\cdot i}) &= \tilde{\pi}_i P_U(t; \tilde{\Theta}_{\cdot i}) + (1 - \tilde{\pi}_i) P_U(t; \tilde{\Psi}_{\cdot i}) \end{aligned}$$

for any t_i and $t_{\mathcal{R}_i}$, then $\pi_i = \tilde{\pi}_i$, $\Theta_{\cdot i} = \tilde{\Theta}_{\cdot i}$, and $\Psi_{\cdot i} = \tilde{\Psi}_{\cdot i}$.

Assumption 2) can be satisfied by popular risk models including Exp model, Pow model, Ray model, and other additive risk models of information propagation, such as kernel hazard functions [10] and feature-enhanced hazard functions [36]. Proposition 1 shows that the network preferences $\{\pi_i\}_{i=1}^N$ are identifiable and diffusion networks Θ and Ψ are column-wise identifiable. However, similar to the labeling non-identifiability issue in finite mixture model [23], Θ and Ψ may still not be globally identifiable without structural constraints due to column permutation. Specifically, the data distribution does not change if we swap $\Theta_{\cdot i}$ and $\Psi_{\cdot i}$ in Θ and Ψ with other columns fixed.

Layer-specific network structure constraint Motivated by real-world applications, one can interpret Θ as the diffusion pathways over an observed social network $\mathbf{A} \in \{0, 1\}^{N \times N}$. Therefore, we can impose support constraints on Θ as

$$\Theta_{ij} \geq 0 \text{ if } \mathbf{A}_{ij} = 1; \Theta_{ij} = 0 \text{ if } \mathbf{A}_{ij} = 0.$$

Due to the sparse nature of real-world social networks, the support constraint also implicitly imposes sparsity constraint on Θ . On the other hand, one can interpret Ψ as the diffusion pathways via latent social relations of individuals that are not captured by the social network \mathbf{A} . The magnitude of Ψ_{ij} reflects the social distance between individual i and j in terms of their latent factors. It has been found that social distance typically has or can be approximated by low-rank structure [34, 26, 27], since high-dimensional social factors can always be embedded into a low dimensional latent space that preserves social distances [34]. Therefore, we impose low-rank structure on Ψ as

$$\text{rank}(\Psi) \leq r, 1 \leq r \ll N.$$

Imposing the above support constraint and low-rank structure allows Θ and Ψ to capture complementary diffusion patterns driven by different types of relations. In addition, the structure constraints solve the above column-wise permutation issue [6]. Specifically, we introduce the matrix subspace $\Lambda_1(\Theta) = \{N \in \mathbb{R}^{N \times N} \mid \text{support}(N) \subseteq \mathbf{A}\}$. We also perform SVD on $\Psi = U \Sigma V^\top$ where $U, V \in \mathbb{R}^{N \times r}$ and r is the rank of Ψ . Then, we define another matrix subspace as $\Lambda_2(\Psi) = \{U X^\top + Y V^\top \mid X, Y \in \mathbb{R}^{N \times k}\}$. We have the following result:

Proposition 2. *Given that the assumptions in Proposition 1 hold, then Θ and Ψ are identifiable if:*

$$\max_{N \in \Lambda_1(\Theta), \|N\|_\infty \leq 1} \|N\|_2 \times \max_{N \in \Lambda_2(\Psi), \|N\|_2 \leq 1} \|N\|_\infty < 1, \quad (11)$$

where $\|\cdot\|_2$ and $\|\cdot\|_\infty$ denote matrix operation norm and largest element in magnitude.

The first term in (23) controls the rank of Θ given a fixed sparsity level, where a larger value indicates a lower rank. The second term controls the sparsity level of Ψ given a fixed rank, where a larger value indicates a lower sparsity level. Intuitively, networks Θ and Ψ can be globally identified given that they are well-separated in terms of either rank or sparsity.

3.4 Model estimation

Combining the distribution of cascade samples in Section 3.2 and the network structure constraints in Section 3.3, we estimate the model parameters Ω via constrained likelihood maximization as

$$\arg \max_{\Omega = \{\Theta, \Psi, \pi\}} \frac{1}{C} \sum_{c=1}^C \log P(\mathbf{t}^{(c)}; \Omega) \quad \text{s.t. } \Theta \odot (\mathbf{I} - \mathbf{A}) = \mathbf{0}, \text{ rank}(\Psi) \leq r,$$

where \mathbf{I} is a N -by- N matrix with all elements being 1. However, both the likelihood function and the rank regularization are difficult to directly optimize. Therefore, we maximize the evidence lower bound of the log-likelihood function and replace the low-rank penalty with its convex relaxation as nuclear norm $\|\cdot\|_*$. The optimization problem can thus be reformulated as follow

$$\arg \max_{\mathbf{q}(\mathbf{Z}), \mathbf{\Omega}} \mathbf{E}_{\mathbf{q}(\mathbf{Z})} \left[\frac{1}{C} \sum_{c=1}^C \log P(\mathbf{t}^{(c)}, \mathbf{Z}; \mathbf{\Omega}) \right] - [\mathbf{E}_{\mathbf{q}(\mathbf{Z})} \log \mathbf{q}(\mathbf{Z})] \text{ s.t. } \mathbf{\Theta} \odot (\mathbf{I} - \mathbf{A}) = \mathbf{0}, \|\mathbf{\Psi}\|_* \leq \rho,$$

where $\mathbf{E}_{\mathbf{q}(\mathbf{Z})}$ is the expectation of \mathbf{Z} over distribution $\mathbf{q}(\mathbf{Z})$. The above objective function can be optimized via EM algorithm by iteratively updating $\mathbf{q}(\mathbf{Z})$ and $\mathbf{\Omega}$. Specifically, with the estimated parameters $\mathbf{\Omega}^{(s)}$ from the s -th step:

$$\textbf{E-step: } \mathbf{q}(\mathbf{Z}; \mathbf{\Omega}^{(s)}) = \prod_{c=1}^C \prod_{i=1}^N P(Z_i^c | \mathbf{t}^{(c)}; \mathbf{\Omega}^{(s)})$$

$$\textbf{M-step: } \mathbf{\Omega}^{(s+1)} = \arg \max_{\mathbf{\Omega}} \frac{1}{C} \sum_{c=1}^C \mathbf{E}_{\mathbf{q}(\mathbf{Z}; \mathbf{\Omega}^{(s)})} [\log P(\mathbf{t}^{(c)}, \mathbf{Z}; \mathbf{\Omega})] \text{ s.t. } \mathbf{\Theta} \odot (\mathbf{I} - \mathbf{A}) = \mathbf{0}, \|\mathbf{\Psi}\|_* \leq \rho.$$

In E-step, the posterior distribution of network indicators $\hat{\pi}_i^c = P(Z_i^c | \mathbf{t}^{(c)}; \mathbf{\Omega}^{(s)})$ can be explicitly updated via (10). In M-step, the objective function $Q(\mathbf{\Omega} | \mathbf{\Omega}^{(s)}) := \frac{1}{C} \sum_{c=1}^C \mathbf{E}_{\mathbf{q}(\mathbf{Z}; \mathbf{\Omega}^{(s)})} [\log P(\mathbf{t}^{(c)}, \mathbf{Z}; \mathbf{\Omega})]$ can be decomposed as $Q(\mathbf{\Omega} | \mathbf{\Omega}^{(s)}) = Q_1(\mathbf{\Theta} | \mathbf{\Omega}^{(s)}) + Q_2(\mathbf{\Psi} | \mathbf{\Omega}^{(s)}) + Q_3(\mathbf{\pi} | \mathbf{\Omega}^{(s)})$. The arguments $\mathbf{\Theta}$, $\mathbf{\Psi}$, and $\mathbf{\pi}$ can thus be updated parallelly in M-step as

$$\text{M.1 : } \mathbf{\Theta}^{(s+1)} = \arg \max_{\mathbf{\Theta}} Q_1(\mathbf{\Theta} | \mathbf{\Omega}^{(s)}) \text{ s.t. } \mathbf{\Theta} \odot (\mathbf{I} - \mathbf{A}) = \mathbf{0} \quad (12)$$

$$\text{M.2 : } \mathbf{\Psi}^{(s+1)} = \arg \max_{\mathbf{\Psi}} Q_2(\mathbf{\Psi} | \mathbf{\Omega}^{(s)}) \text{ s.t. } \|\mathbf{\Psi}\|_* \leq \rho \quad (13)$$

$$\text{M.3 : } \mathbf{\pi}^{(s+1)} = \arg \max_{\mathbf{\pi}} Q_3(\mathbf{\pi} | \mathbf{\Omega}^{(s)}) \quad (14)$$

When the structural network \mathbf{A} is not observed, one can impose l_1 -norm penalty to pursue sparsity structure in $\mathbf{\Theta}$. Accordingly, the constraint in optimization (12) is replaced by $\|\mathbf{\Theta}\|_1 \leq s'$, where $s' > 0$ is the sparsity tuning parameter. The main advantage of the proposed relaxation is that the M-step becomes a series of convex optimization problems. Specifically, denote parameter spaces $\Xi_{\mathbf{\Theta}}(s) := \{\mathbf{\Theta} \in [0, \beta_1]^{N \times N} | \mathbf{\Theta} \odot (\mathbf{I} - \mathbf{A}) = \mathbf{0}\}$ with $s = \|\mathbf{A}\|_0$, $\Xi_{\mathbf{\Psi}}(\rho) := \{\mathbf{\Psi} \in [0, \beta_2]^{N \times N} | \|\mathbf{\Psi}\|_* \leq \rho\}$, and $\Xi_{\mathbf{\pi}} := \{\mathbf{\pi} \in [\epsilon, 1 - \epsilon]^N\}$, where β_1, β_2 are nonnegative constants and $0 < \epsilon < 0.5$. We have the following result:

Theorem 3.1. *The parameter spaces $\Xi_{\mathbf{\Theta}}(s)$, $\Xi_{\mathbf{\Psi}}(\rho)$, and $\Xi_{\mathbf{\pi}}$ are convex sets for any $s > 0$, $\rho > 0$, and $Q_3(\mathbf{\pi} | \mathbf{\Omega}^{(s)})$ is concave on $\mathbf{\pi}$. If the hazard function $H(t | t', \lambda)$ satisfies $\frac{\partial H^2(t | t', \lambda)}{\partial \lambda^2} = 0$ for $t \geq t'$, then $Q_1(\mathbf{\Theta} | \mathbf{\Omega}^{(s)})$ and $Q_2(\mathbf{\Psi} | \mathbf{\Omega}^{(s)})$ are concave on $\mathbf{\Theta}$ and $\mathbf{\Psi}$, respectively. Furthermore, if for any node i , the probabilities of being source node $P(v) > 0$ for $v \in \mathcal{R}$ where \mathcal{R} denotes the set of nodes from which i is reachable via a directed path, then $\mathbb{E}_t[Q_1(\mathbf{\Theta} | \mathbf{\Omega}^{(s)})]$ and $\mathbb{E}_t[Q_2(\mathbf{\Psi} | \mathbf{\Omega}^{(s)})]$ are also strictly concave in terms of $\mathbf{\Theta}$ and $\mathbf{\Psi}$, respectively.*

Theorem 3.1 guarantees that the optimization in M-step has a unique and optimal solution. Combining the theorem with the convergence guarantee of EM algorithm for convex ancillary function $Q(\mathbf{\Omega} | \mathbf{\Omega}^{(s)})$ [2, 38], the above likelihood maximizer $\hat{\mathbf{\Omega}}$ is guaranteed to converge to the true $\mathbf{\Omega}$. In addition, the convexity assumption on hazard function can be satisfied by popular risk models including Exp model, Pow model, Ray model, and other additive risk models. We summarize and provide details for the above optimization process in the Appendix.

4 Numerical experiments on synthetic cascading data

We investigate the performance of the proposed method in recovering the diffusion networks based on synthetic cascading data. The performance of global transmission rates estimation on a network $\mathbf{\Theta}$ is measured by normalized mean absolute error (MAE) as $\text{MAE}(\mathbf{\Theta}) = \sum_{i,j} |\hat{\mathbf{\Theta}}_{ij} - \mathbf{\Theta}_{ij}| / \mathbf{\Theta}_{ij}$. The

performance of network probability π estimation is measured as $\text{MAE}(\pi) = \sum_i |\hat{\pi}_i - \pi_i|/\pi_i$. In addition, we investigate the performance of recovering the structure of diffusion network. Specifically, given a diffusion network Θ we consider three topology estimation metrics accuracy, precision, and recall as $\text{Acc}(\Theta) = \sum_{i,j} |\mathbf{I}(\hat{\Theta}_{ij}) - \mathbf{I}(\Theta_{ij})| / (\sum_{i,j} \mathbf{I}(\hat{\Theta}_{ij}) + \sum_{i,j} \mathbf{I}(\Theta_{ij}))$, $\text{Pre}(\Theta) = \sum_{i,j} (\mathbf{I}(\hat{\Theta}_{ij}) \cdot \mathbf{I}(\Theta_{ij})) / \sum_{i,j} \mathbf{I}(\hat{\Theta}_{ij})$, and $\text{Recall}(\Theta) = \sum_{i,j} (\mathbf{I}(\hat{\Theta}_{ij}) \cdot \mathbf{I}(\Theta_{ij})) / \sum_{i,j} \mathbf{I}(\Theta_{ij})$, where $\mathbf{I}(\alpha) = 1$ if $\alpha > 0$ and $\mathbf{I}(\alpha) = 0$ otherwise. In the following numerical experiments, we fix the size of diffusion networks at $N = 200$.

4.1 Benchmark comparison under different network topologies

We compare the proposed method with baseline methods including NetRate [31], MMRate [37], and ConNle [29] on diffusion network recovery. We generate diffusion networks Θ and Ψ to mimic different structure of real-world networks. Specifically, we consider the diffusion network Θ and its support \mathbf{A} as random network, network with community structure, and scale-free network. On the other hand, we fix the latent diffusion network to be both low-rank (rank 5) and sparse (edge density 0.05). Given Θ and Ψ , we generate $C = 2,000$ independent cascade samples based on double mixture model with Exp transmission model and observation window length being $T = 10$.

Table 1 shows the proposed method outperforms both NetRate and MMRate by achieving lower MAE of transmission rate estimation on diffusion network Θ under three network topologies and lower MAE of corresponding latent diffusion network Ψ . MAE comparison does not include ConNle since it only estimates network topology. In addition, we compare the proposed method with baseline methods in terms of network topology recovery on both Ψ and joint network $\hat{\Theta} \cup \hat{\Psi}$ for a fair comparison since benchmark methods NetRate and ConNle do not distinguish the different diffusion networks by design. Table 1 shows the proposed method achieves higher accuracy on topology recovery via both latent network $\hat{\Psi}$ and joint network $\hat{\Theta} \cup \hat{\Psi}$ than NetRate, MMRate, and ConNle under different underlying structures in Θ . It also shows that under different settings, our method significantly outperforms all other methods in precision while remains higher recall than MMRate and ConNle, and achieves similar recall to NetRate. Since only MMRate differentiates diffusion networks and estimates probabilities of cascading diffusing over a specific network, we compare MAE(π) from the proposed method with that from MMRate. Table 1 shows that our method also outperforms MMRate in estimation of network selection probability under different network topologies.

Table 1: Diffusion network estimations from different methods under three Θ topologies (standard deviation in parenthesis, best performance highlighted in blue).

		MAE_{Θ}	MAE_{Ψ}	MAE_{π}	Acc_{Ψ}	Pre_{Ψ}	Rec_{Ψ}	$Acc_{\Theta \cup \Psi}$	$Pre_{\Theta \cup \Psi}$	$Rec_{\Theta \cup \Psi}$
OurAlg	Rand	0.315(0.011)	0.544(0.007)	0.047(0.002)	0.842(0.006)	0.757(0.008)	0.948(0.006)	0.889(0.006)	0.829(0.006)	0.958(0.004)
	Com	0.264(0.005)	0.507(0.004)	0.049(0.002)	0.846(0.005)	0.767(0.007)	0.942(0.004)	0.888(0.004)	0.832(0.006)	0.954(0.003)
	Scale	0.312(0.006)	0.545(0.006)	0.050(0.002)	0.833(0.005)	0.750(0.007)	0.937(0.005)	0.883(0.004)	0.826(0.007)	0.949(0.003)
NetRate	Rand	0.865(0.001)	2.998(0.026)	—	0.172(0.002)	0.095(0.001)	0.964(0.005)	0.207(0.003)	0.116(0.002)	0.942(0.004)
	Com	0.851(0.000)	3.103(0.014)	—	0.180(0.002)	0.099(0.001)	0.975(0.004)	0.216(0.002)	0.122(0.001)	0.955(0.003)
	Scale	0.855(0.000)	2.897(0.018)	—	0.179(0.002)	0.099(0.001)	0.929(0.003)	0.220(0.002)	0.125(0.001)	0.922(0.003)
MMRate	Rand	0.637(0.007)	1.490(0.030)	0.888(0.133)	0.201(0.023)	0.120(0.015)	0.632(0.023)	0.265(0.028)	0.164(0.019)	0.695(0.019)
	Com	0.630(0.008)	1.665(0.036)	0.881(0.137)	0.219(0.017)	0.130(0.012)	0.705(0.008)	0.281(0.020)	0.173(0.015)	0.755(0.006)
	Scale	0.668(0.006)	1.477(0.027)	0.928(0.028)	0.211(0.019)	0.126(0.013)	0.650(0.009)	0.278(0.023)	0.173(0.018)	0.710(0.007)
ConNle	Rand	—	—	—	0.519(0.006)	0.397(0.005)	0.752(0.010)	0.638(0.006)	0.530(0.006)	0.800(0.008)
	Com	—	—	—	0.519(0.005)	0.398(0.005)	0.748(0.007)	0.638(0.005)	0.532(0.007)	0.796(0.006)
	Scale	—	—	—	0.523(0.006)	0.402(0.005)	0.748(0.009)	0.646(0.006)	0.543(0.006)	0.797(0.008)

4.2 Network recovery under different transmission models

In this subsection, we investigate the performance of our diffusion network estimation when cascade samples are generated from popular transmission models including Exp, Pow, and Ray models, respectively. Table 2 illustrates the transmission rates recovery and network topology recovery on Ψ under different transmission models and cascade sample sizes C . As the sample size C increases, both the parameter estimations (MAE) and topology recovery metrics (Acc, Pre, Rec) improve under different transmission models. In addition, the degree of improvement decreases as more cascade samples become available. This pattern indicates the consistency of the proposed diffusion network estimators, and the convergence of the proposed EM-type optimization. Notice that the diffusion network recovery based on cascade samples generated from Pow transmission model is better than

Exp and Ray model. This is because the Pow model introduces an additional parameter as time lag lower bound, which lowers the variation of activate time and overestimation of transmission rates.

Table 2: Diffusion network estimations from the proposed method under different cascade models and sample sizes.

		MAE_{Θ}	MAE_{Ψ}	MAE_{π}	Acc_{Ψ}	Prc_{Ψ}	Rec_{Ψ}
Exp	C = 500	0.320(0.019)	0.871(0.019)	0.095(0.006)	0.564(0.010)	0.431(0.011)	0.817(0.010)
	C = 1000	0.235(0.013)	0.667(0.007)	0.091(0.005)	0.743(0.009)	0.653(0.014)	0.862(0.007)
	C = 1500	0.216(0.012)	0.601(0.008)	0.085(0.004)	0.817(0.006)	0.746(0.008)	0.903(0.006)
	C = 2000	0.187(0.011)	0.562(0.007)	0.082(0.004)	0.857(0.004)	0.794(0.006)	0.930(0.005)
Ray	C = 500	0.287(0.018)	1.053(0.056)	0.152(0.008)	0.605(0.018)	0.525(0.024)	0.714(0.013)
	C = 1000	0.211(0.015)	0.990(0.100)	0.130(0.008)	0.712(0.019)	0.605(0.026)	0.866(0.012)
	C = 1500	0.179(0.011)	0.629(0.091)	0.134(0.004)	0.795(0.017)	0.702(0.026)	0.916(0.007)
	C = 2000	0.188(0.012)	0.605(0.081)	0.137(0.004)	0.807(0.011)	0.709(0.019)	0.937(0.011)
Pow	C = 500	0.171(0.008)	0.433(0.012)	0.144(0.010)	0.882(0.015)	0.832(0.024)	0.939(0.009)
	C = 1000	0.129(0.006)	0.328(0.007)	0.131(0.007)	0.956(0.004)	0.932(0.007)	0.982(0.004)
	C = 1500	0.112(0.005)	0.290(0.004)	0.125(0.008)	0.969(0.002)	0.946(0.004)	0.993(0.002)
	C = 2000	0.106(0.006)	0.271(0.004)	0.129(0.011)	0.973(0.003)	0.950(0.005)	0.997(0.001)

5 Real cascading data analysis

In this section, we study the cascading patterns of research topics in sociology by discovering the diffusion networks among US universities. Geographic proximity is known to facilitate idea exchanges through collaborations and citations among colleagues within the same institution or nearby locations [3], which can be due to dependence on shared research resources and the need for coordination [28]. It is also known that research diffusion can happen via the latent network of scholars, which is also sometimes called the "invisible colleges" [7] to highlight the role of informal networks and latent ties. Our goal is to infer the latent research topic diffusion network and compare its difference with the geographic network in terms of topic cascading.

Data preparation and preprocess We select universities in both the list of 1965 ASA Guide to Graduate Departments of Sociology and the list of 2022 US News Best Sociology Programs in America. In addition, we exclude the universities whose sociology programs were established after 1965. Based on the above conditions, we finalize $N = 104$ universities, which are considered as the target population of our study. We then create a geographical network $\mathbf{A} \in \{0, 1\}^{N \times N}$ among the selected universities, where $A_{ij} = 1$ if university i and j are located in the same state.

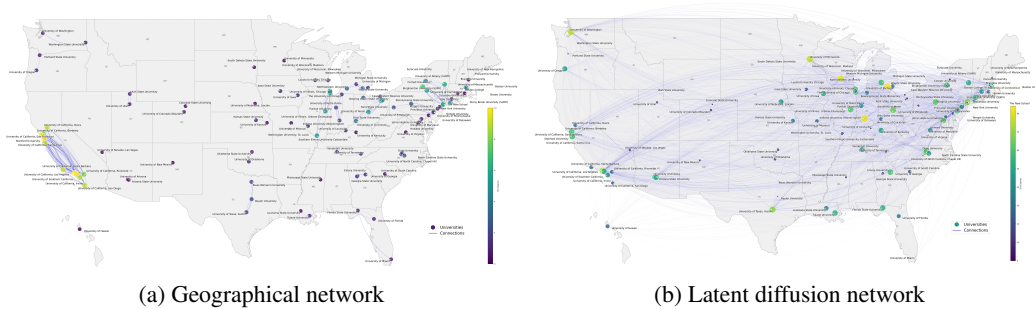


Figure 2: Cascade transmission rates over geographical network Θ and latent diffusion network Ψ .

To construct research topics, we use Elsevier’s Scopus API to compile a dataset of 29,725 unique articles from 23 top generalist sociology journals, including information on authors, their affiliations, article titles, and abstracts. Multiple keywords are extracted from the titles and abstracts of articles. After these preprocesses, we obtain 3,033 unique research topics that cover major research fields in sociology. For each research topic c , we construct the corresponding cascade samples as $(t_1^c, t_2^c, \dots, t_N^c)$ where $t_i^c := \tilde{t}_i^c - \tilde{t}_0^c$, with \tilde{t}_0^c being the publication date of the first article that involves topic c , and \tilde{t}_i^c being the publication date of the first article that involves topic c and is published by any scholar affiliated with university i . If university i never publishes any article that involves topic c , we set t_i^c to be year 2022.

Diffusion networks inference We simultaneously estimate the geographic diffusion network Θ and the latent diffusion network Ψ based on the proposed model. Figure (2) illustrates the inferred diffu-

sion networks $\hat{\Theta}$ and $\hat{\Psi}$, respectively. The size of nodes represents the out-degree of corresponding universities on networks. We see that the geographic diffusion network has a strong local community structure, especially among universities in California, the Great Lakes, and the northeast coast. Figure (2) also shows that the latent diffusion network demonstrates a decentralized structure and contains many connections between the east and west coasts. This captures the national collaboration and academic mobility. We also compare estimated pairwise transmission rates on diffusion networks $\hat{\Theta}$ and $\hat{\Psi}$ in Figure (3). In summary, the latent diffusion network is denser than the geographic network, and the magnitude of transmission rates on the geographic network is larger and has more variation compared to the latent diffusion network.

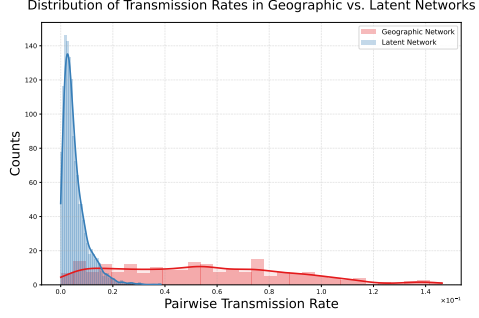


Figure 3: Comparison of distributions of pairwise transmission rates from geographic network Θ_{ij} and latent network Ψ_{ij} .

To provide an interpretation of the inferred geographic diffusion network Θ and latent diffusion network Ψ , we connect universities' positions on the network with universities' U.S. News University rankings on the sociology program. We choose these rankings since they are a systematic and popular summary of academic factors. Figure (4a) illustrates the association between program ranking and the node-wise betweenness centrality on latent diffusion network. The Betweenness centrality of a node i is defined as $\sum_{j \neq i \neq k} \sigma_{jk}(i) / \sigma_{jk}$, where σ_{jk} is the total number of shortest paths from node j to k , and $\sigma_{jk}(i)$ is the number of those paths that pass through i . Based on the figure, the universities'

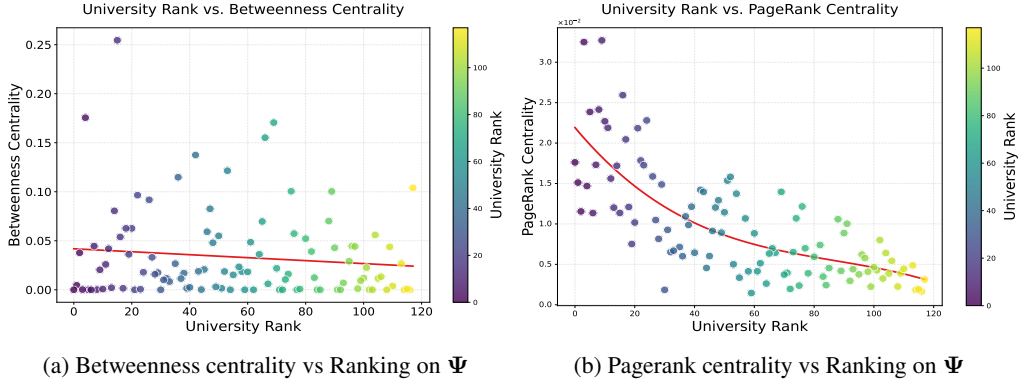


Figure 4: Node-wise centrality vs Ranking on latent network Ψ .

betweenness centralities are relatively uniform and do not significantly decrease as ranking increases. This suggests that there does not exist any strong community or centralized topology in the latent network. When universities are grouped according to their ranking (top 20, middle 60, low 20 etc.), each group has different high-betweenness universities, serving as the bridges in idea diffusion. We also investigate the relation between university rankings and Pagerank centrality. Pagerank centrality measures the influence of a node on a network based on how many influential nodes it connects to. Figure (4b) shows that the universities' Pagerank centralities decrease as their rankings increase, which suggests that the idea exchanges among high-ranked universities are more frequent and faster than those of other universities. This pattern appears because high-ranked universities have a higher level of research activities. University prestige is also considered as an important proxy for the quality of ideas, which increases the likelihood of research ideas from higher-ranked universities to diffuse.

6 Scalability analysis

The implicit convex nature of our objective function allows the optimization process to be as efficient as gradient descent, even though the optimization of our method is based on EM-type update. The E-step in our algorithm has an explicit form as in (10). Furthermore, we have proved in Theorem

3.1 that the objective function $Q(\Omega)$ in the M-step is strictly concave and in each EM-iteration, we only perform gradient ascent once instead of maximizing $Q(\Omega)$ until convergence. Under these conditions, established proofs in [2] show that the above EM-update is equivalent to gradient decent, and our optimization thus enjoys gradient descent’s geometric convergence rate.

Computational complexity of the proposed method is $K \cdot O(N^2 \cdot C) + K \cdot O(N^3)$, where N is network size, C is sample size, and K is the number of iterations executed. In practice, we require $C \gg N$ for reasonable estimation. The first term thus dominates and the computational complexity is approximately the same as that of gradient descent. We can replace the standard SVD operation in the M -step by randomized SVD or Lanczos algorithm. Then, we can further reduce the computational complexity of the second term from $O(N^3)$ to $O(rN^2)$ or $O((r+l)N^2)$ where $r \ll N$ is the rank of network Ψ , N is the size of network, and l is a constant usually between 10 and 20.

We also numerically investigate both time and estimation performance of the proposed method in recovering diffusion networks based on synthetic cascading data in large network settings. We adopt the same criteria used in Section 4 to evaluate estimation performance and use average time per iteration to evaluate time performance. We generate diffusion networks Θ and Ψ of different sizes in a similar way to Section 4.2, fixing Ψ to be both low-rank (rank 5) and sparse (edge density 0.01). Given Θ and Ψ , we generate $C = 50,000$ independent cascade samples based on double mixture model with Exp transmission model and observation window length being $T = 10$.

Table 3: Execution time of different methods under different network sizes (unit: second per iteration).

	$N = 500$	$N = 1000$	$N = 2000$	$N = 4000$
OurAlg	2.395(0.001)	8.201(0.019)	33.442(0.031)	132.955(0.046)
NetRate	1.390(0.001)	4.929(0.011)	20.039(0.030)	79.992(0.040)

Table 3 shows that the computation time of the proposed method per iteration is proportional to the squared network size N^2 . This aligns with the earlier theoretical analysis of computational complexity. In addition, for all network sizes $N = 500, 1000, 2000, 4000$, the ratio between computational time per iteration of the proposed method and NetRate [31] is approximately a constant 1.65. Since NetRate [31] is well-known for its scalability, this result demonstrates the computational efficiency and scalability of the proposed method.

Table 4 illustrates the estimation performance of the proposed method for large networks. For networks of all sizes $N = 500, 1000, 2000$, the proposed method outperforms NetRate [31] on both transmission rate estimation and network topology recovery. For networks of sizes $N = 500, 1000$, the proposed method achieves estimation performances comparable to those of smaller networks in Section 4. Additionally, when sample size is fixed, the performance of the proposed method decreases, indicating the need for more samples on larger networks to achieve accurate estimations.

Table 4: Diffusion network estimations from the proposed method under different network sizes.

		MAE_{Θ}	MAE_{Ψ}	Acc_{Ψ}	Pre_{Ψ}	Rec_{Ψ}
OurAlg	$N = 500$	0.398(0.001)	0.711(0.019)	0.894(0.003)	0.820(0.005)	0.983(0.001)
	$N = 1000$	0.411(0.005)	0.797(0.007)	0.779(0.001)	0.666(0.001)	0.939(0.003)
	$N = 2000$	0.437(0.003)	0.814(0.008)	0.571(0.002)	0.523(0.002)	0.630(0.002)
NetRate	$N = 500$	0.743(0.014)	1.053(0.038)	0.078(0.008)	0.041(0.004)	0.818(0.007)
	$N = 1000$	0.755(0.010)	1.094(0.019)	0.086(0.001)	0.045(0.001)	0.914(0.007)
	$N = 2000$	0.759(0.003)	1.096(0.002)	0.083(0.000)	0.044(0.000)	0.913(0.000)

7 Conclusion

In this paper, we propose a novel double network mixture model for heterogeneous cascading process on multi-layer networks. Our method can identify the latent diffusion network complementary to the observed network. Due to its convex formulation, our method has both statistical and computational guarantee in terms of estimating diffusion networks. A major future work is to generalize the mixture graph model to a system with more than two-layer networks and derive the model identification conditions. Extending the proposed method to inference on time-varying networks with time-dependent parameters $\{\Theta(t), \Psi(t), \pi(t)\}$ or nonparametric transmission models are also interesting directions for future works.

References

- [1] Eytan Adar and Lada A Adamic. Tracking information epidemics in blogspace. In *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pages 207–214. IEEE, 2005.
- [2] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77 – 120, 2017.
- [3] Pierre-Alexandre Balland and David Rigby. The geography of complex knowledge. *Economic geography*, 93(1):1–23, 2017.
- [4] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. Learning from the behavior of others: Conformity, fads, and informational cascades. *Journal of economic perspectives*, 12(3):151–170, 1998.
- [5] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.
- [6] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.
- [7] Diana Crane. *Invisible Colleges; Diffusion of Knowledge in Scientific Communities*. University of Chicago Press, Chicago,, 1972.
- [8] Hadi Daneshmand, Manuel Gomez-Rodriguez, Le Song, and Bernhard Schoelkopf. Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm. In *International conference on machine learning*, pages 793–801. PMLR, 2014.
- [9] Nan Du, Le Song, Manuel Gomez Rodriguez, and Hongyuan Zha. Scalable influence estimation in continuous-time diffusion networks. *Advances in neural information processing systems*, 26, 2013.
- [10] Nan Du, Le Song, Ming Yuan, and Alex Smola. Learning networks of heterogeneous influence. *Advances in neural information processing systems*, 25, 2012.
- [11] Wenjing Duan, Bin Gu, and Andrew B Whinston. Informational cascades and software adoption on the internet: an empirical investigation. *MIS quarterly*, pages 23–48, 2009.
- [12] John RP French. The bases of social power. *Studies in social power/University of Michigan Press*, 1959.
- [13] Noah E Friedkin and Eugene C Johnsen. *Social influence network theory: A sociological examination of small group dynamics*, volume 33. Cambridge University Press, 2011.
- [14] Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The annals of applied statistics*, 1(2):302–332, 2007.
- [15] P. E. Gill, W. Murray, and M. A. Saunders. Snopt: An sqp algorithm for large-scale constrained optimization. *SIAM Review*, 47:99–131, 2005.
- [16] Philip E. Gill, Walter Murray, Michael A. Saunders, and Elizabeth Wong. *User’s Guide for SNOPT Version 7.7: Software for Large-Scale Nonlinear Programming*. Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112, USA, March 2018. CCoM Technical Report 18-1.
- [17] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):1–37, 2012.
- [18] Manuel Gomez-Rodriguez, Jure Leskovec, and Bernhard Schölkopf. Modeling information propagation with survival theory. In *International conference on machine learning*, pages 666–674. PMLR, 2013.

- [19] Manuel Gomez Rodriguez, Jure Leskovec, and Bernhard Schölkopf. Structure and dynamics of information pathways in online media. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 23–32, 2013.
- [20] Márton Karsai, Gerardo Iñiguez, Riivo Kikas, Kimmo Kaski, and János Kertész. Local cascades induced global contagion: How heterogeneous thresholds, exogenous effects, and unconcerned behaviour govern online adoption spreading. *Scientific reports*, 6(1):27178, 2016.
- [21] Matt J Keeling and Ken TD Eames. Networks and epidemic models. *Journal of the royal society interface*, 2(4):295–307, 2005.
- [22] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003.
- [23] Daeyoung Kim and Bruce G Lindsay. Empirical identifiability in finite mixture models. *Annals of the Institute of Statistical Mathematics*, 67:745–772, 2015.
- [24] Alden S Klov Dahl, John J Potterat, Donald E Woodhouse, John B Muth, Stephen Q Muth, and William W Darrow. Social networks and infectious disease: The colorado springs study. *Social science & medicine*, 38(1):79–88, 1994.
- [25] Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5–es, 2007.
- [26] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *journal of the association for information science and technology* (2007). *Google Scholar Google Scholar Digital Library Digital Library*, 2007.
- [27] Aditya Krishna Menon and Charles Elkan. Link prediction via matrix factorization. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part II* 22, pages 437–452. Springer, 2011.
- [28] Sarah Morrison-Smith and Jaime Ruiz. Challenges and barriers in virtual teams: a literature review. *SN Applied Sciences*, 2(6):1096, 2020.
- [29] Seth A Myers and Jure Leskovec. On the convexity of latent social network inference. *Advances in Neural Information Processing Systems*, 2, 2010.
- [30] Jeongha Oh, Anjana Susarla, and Yong Tan. Examining the diffusion of user-generated content in online social networks. *Available at SSRN 1182631*, 2008.
- [31] Manuel Gomez Rodriguez, Jure Leskovec, David Balduzzi, and Bernhard Schölkopf. Uncovering the structure and temporal dynamics of information propagation. *Network Science*, 2(1):26–65, 2014.
- [32] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704, 2011.
- [33] Charles R Shipan and Craig Volden. The mechanisms of policy diffusion. *American journal of political science*, 52(4):840–857, 2008.
- [34] Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.
- [35] Jacco Wallinga and Peter Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of epidemiology*, 160(6):509–516, 2004.
- [36] Liaoruo Wang, Stefano Ermon, and John E Hopcroft. Feature-enhanced probabilistic models for diffusion network inference. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II* 23, pages 499–514. Springer, 2012.

- [37] Senzhang Wang, Xia Hu, Philip S Yu, and Zhoujun Li. Mmrate: Inferring multi-aspect diffusion networks with multi-pattern cascades. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1246–1255, 2014.
- [38] Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint arXiv:1412.8729*, 2014.
- [39] Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.
- [40] Ming Yu, Varun Gupta, and Mladen Kolar. Estimation of a low-rank topic-based model for information cascades. *Journal of Machine Learning Research*, 21(71):1–47, 2020.
- [41] Gilles Zumbach and Paul Lynch. Heterogeneous volatility cascade in financial markets. *Physica A: Statistical Mechanics and its Applications*, 298(3-4):521–529, 2001.

A Additional details on methodology

A.1 Detailed expression of objectives in M-step

$Q_1(\Theta | \Omega^{(s)})$, $Q_2(\Psi | \Omega^{(s)})$, and $Q_3(\pi | \Omega^{(s)})$ can be formulated in detail as:

$$\begin{aligned} Q_1(\Theta | \Omega^{(s)}) &= \frac{1}{C} \sum_{c=1}^C \left\{ \sum_{i:t_i^c \leq T} \hat{\pi}_i^c \log P_I(t_i; \Theta_{\cdot i}) + \sum_{j:t_j^c > T} \hat{\pi}_j^c \log P_U(T; \Theta_{\cdot j}) \right\} \\ Q_2(\Psi | \Omega^{(s)}) &= \frac{1}{C} \sum_{c=1}^C \left\{ \sum_{i:t_i^c \leq T} (1 - \hat{\pi}_i^c) \log P_I(t_i; \Psi_{\cdot i}) + \sum_{j:t_j^c > T} (1 - \hat{\pi}_j^c) \log P_U(T; \Psi_{\cdot j}) \right\} \\ Q_3(\pi | \Omega^{(s)}) &= \frac{1}{C} \sum_{c=1}^C \left\{ \sum_{i=1}^N \left[\hat{\pi}_i^c \log \pi_i + (1 - \hat{\pi}_i^c) \log(1 - \pi_i) \right] \right\}, \end{aligned}$$

where P_I, P_U follow the definitions in Section 3.1.

A.2 Optimization algorithm and discussions

We summarize the optimization of the reformulated problem in Section 3.4 via EM algorithm as Algorithm 1.

Algorithm 1 First-order projected EM algorithm

Require: initialization $\Omega^{(0)} = \{\Theta^{(0)}, \Psi^{(0)}, \pi^{(0)}\}$, observed network \mathbf{A} , low-rank penalty μ , learning rate λ , and stopping criterion ϵ .

while $Q(\Omega^{(s+1)} | \Omega^{(s)}) - \mu \|\Psi^{(s+1)}\|_* - Q(\Omega^{(s)} | \Omega^{(s)}) + \mu \|\Psi^{(s)}\|_* > \epsilon$ **do**.

E-step: update $\hat{\pi}_i^c = P(Z_i | t^c; \Omega^{(s)})$ via its posterior distribution based on $\Omega^{(s)}$ for $i = 1, \dots, N$, $c = 1, \dots, C$.

M-step: decompose $Q(\Omega | \Omega^{(s)}) = Q_1(\Theta | \Omega^{(s)}) + Q_2(\Psi | \Omega^{(s)}) + Q_3(\pi | \Omega^{(s)})$

 M.1: Update Θ via $Q_1(\Theta | \Omega^{(s)})$:

$$\Theta^{(s+1)} \leftarrow \max\{\Theta^{(s)} + \lambda \frac{\partial Q_1}{\partial \Theta} \Big|_{\Theta=\Theta^{(s)}} \odot \mathbf{A}, \mathbf{0}\}$$

 M.2: Update Ψ via $Q_2(\Psi | \Omega^{(s)})$:

$$\text{M.2.1: } \Psi \leftarrow \Psi + \lambda \frac{\partial Q_2}{\partial \Psi} \Big|_{\Psi=\Psi^{(s)}}$$

 M.2.2: perform SVD decomposition on $\Psi = U\Lambda V^\top$

$$\text{M.2.3: } \Psi^{(s+1)} \leftarrow \max\{U \text{diag}(\Lambda - \lambda\mu)_+ V, \mathbf{0}\}$$

 M.3: Update π via $Q_3(\pi | \Omega^{(s)})$:

$$\pi_i^{(s+1)} = \frac{\sum_{c=1}^C \hat{\pi}_i^c}{C}, \quad i = 1, \dots, N$$

end while

In Algorithm 1, we utilize projected gradient ascent and proximal gradient ascent to update diffusion networks Θ and Ψ in M.1 and M.2 of M-step, where the latter can be implemented via singular value soft-thresholding operation [14, 5]. For computational efficiency and stability, we utilize the first-order EM algorithm [2] such that the ELBO is increased via one-step gradient ascent instead of maximized in M-step. In addition, both the gradients $\frac{\partial Q_1}{\partial \Theta}$ and $\frac{\partial Q_2}{\partial \Psi}$ have closed forms and can be efficiently calculated. The closed form gradients are provided in A.3.

Likelihood-based parameter tuning The low-rank penalty μ in the above Algorithm 1 can be selected in a data-adapted manner. Specifically, we can first randomly separate the total cascade samples into a training subset C_{train} and a validation subset C_{val} , and estimate model parameters $\hat{\Omega}_\mu$ on C_{train} given a specific μ . Then we calculate the log-likelihood of validation samples as $\frac{1}{|C_{val}|} \sum_{c \in C_{val}} \log P(t^{(c)}, \hat{\Omega}_\mu)$, and select μ such that

$$\mu = \arg \max_{\mu \in \mathbf{G}} \frac{1}{|C_{val}|} \sum_{c \in C_{val}} \log P(t^{(c)}, \hat{\Omega}_\mu)$$

where \mathbf{G} is a grid of candidates for penalty values.

Sparsity pursuit on Θ When the observed network \mathbf{A} is not sparse enough or the sample size of cascade is relatively small, one can further add l_1 regularizer on Θ in the objective function $Q_1(\Theta \mid \Omega^{(s)})$. Accordingly, the matrix Θ is updated via an additional soft-thresholding operator in the M.1 step of Algorithm 1. Previous results shows that adopting the l_1 regularization can improve the performance and efficiency to recover the diffusion network structure [8, 31].

A.3 Closed form gradients in Algorithm 1

Denote $(S_c^\Theta)_{ji} = S(t_i^c \mid t_j^c, \Theta_{ji})$, $(S_c^\Psi)_{ji} = S(t_i^c \mid t_j^c, \Psi_{ji})$, $(H_c^\Theta)_{ji} = H(t_i^c \mid t_j^c, \Theta_{ji})$, and $(H_c^\Psi)_{ji} = H(t_i^c \mid t_j^c, \Psi_{ji})$, where $S(\cdot \mid \cdot, \cdot)$ and $H(\cdot \mid \cdot, \cdot)$ are survival and hazard functions, respectively.

In addition, we introduce two cascade sample indicator variables $\mathbb{I}^{(1)}(c)_{ji}$ and $\mathbb{I}^{(2)}(c)_{ji}$ for each node pair (j, i) such that

$$\mathbb{I}^{(1)}(c)_{ji} = \begin{cases} 1, & \text{if } t_j^c < t_i^c, \text{ and, } t_j^c \leq T, \\ 0, & \text{otherwise} \end{cases}$$

and

$$\mathbb{I}^{(2)}(c)_{ji} = \begin{cases} 1, & \text{if } t_j^c < t_i^c, t_j^c \leq T, \text{ and, } t_i^c \leq T, \\ 0, & \text{otherwise} \end{cases}$$

Then, $\frac{\partial Q_1}{\partial \Theta}$ and $\frac{\partial Q_2}{\partial \Psi}$ over each element in Θ and Ψ , respectively, can be explicitly formulated as:

$$\begin{aligned} \frac{\partial Q_1}{\partial (\Theta)_{ji}} &= \sum_{c=1}^C [\mathbb{I}^{(1)}(c)_{ji} \times \mathbb{I}(A_{ji} = 1) \frac{\hat{\pi}_i^c}{(S_c^\Theta)_{ji}} \frac{\partial (S_c^\Theta)_{ji}}{\partial \Theta_{ji}} + \\ &\quad \mathbb{I}^{(2)}(c)_{ji} \times \mathbb{I}(A_{ji} = 1) \frac{\hat{\pi}_i^c}{\sum_{k: t_k^c < t_i^c} (H_c^\Theta)_{ki}} \frac{\partial (H_c^\Theta)_{ji}}{\partial \Theta_{ji}}], \\ \frac{\partial Q_2}{\partial (\Psi)_{ij}} &= \sum_{c=1}^C [\mathbb{I}^{(1)}(c)_{ji} \frac{1 - \hat{\pi}_i^c}{(S_c^\Psi)_{ji}} \frac{\partial (S_c^\Psi)_{ji}}{\partial \Psi_{ji}} + \mathbb{I}^{(2)}(c)_{ji} \frac{1 - \hat{\pi}_i^c}{\sum_{k: t_k^c < t_i^c} (H_c^\Psi)_{ki}} \frac{\partial (H_c^\Psi)_{ji}}{\partial \Psi_{ji}}] \end{aligned}$$

The above gradients can be calculated by matrix operations after translating two indicator functions into matrices.

B Additional details on experimental setup

B.1 Benchmark comparison under different network topologies

In this experiment, we generate diffusion networks with different topologies. Specifically, we consider the diffusion network Θ and its support \mathbf{A} as random network, network with community structure, and scale-free network. For random network, we generate \mathbf{A} via Erdos-Renyi model with edge generation probability being 0.01. For the community structure, we generate \mathbf{A} with a stochastic block model that contains four equal-sized communities. The generation probability of within-community edge is 0.05 and that of between-community edge is 0.01. For scale-free network, we generate \mathbf{A} via Barabási–Albert model where we set the number of edges to attach from a new node to existing nodes to be 1. After generating support \mathbf{A} , we set its diagonal to be 0 and assign to each non-zero edge a weight that follows $Unif(1, 5)$ to construct Θ .

We generate the latent diffusion network as $\Psi = \Psi_1 \Psi_2^\top$ where $\Psi_1, \Psi_2 \in \mathbb{R}_+^{N \times 5}$. We fix the proportion of non-zero edges in Ψ_1, Ψ_2 at 0.1 and sample weights from $Unif(1, 2)$ for these non-zero edges. The generated Ψ has an edge density of 0.05 and transmission rates between 1 and 8 on non-zero edges.

In addition, we impute edges on \mathbf{A} to ensure that each row and column of $\Theta \cup \Psi$ has at least one non-zero element, i.e., the combined network is connected. Given Θ and Ψ , we generate $C = 2,000$ independent cascade samples based on the double mixture model with Exp transmission model and observation window length of $T = 10$.

B.2 Network recovery under different transmission models

In this experiment, we generate the latent diffusion network Ψ with low-rank structure using the same procedure as in B.1.

Given Ψ and its support B , we then generate Θ 's support A by forcing one overlap with B per column, i.e. $\{(i, j) \mid A_{ij}B_{ij} = 1\}$, and one non-overlap with B per column, i.e. $\{(i, j) \mid A_{ij} = 1, B_{ij} = 0\}$. For Exp and Pow transmission models, we create Θ by sampling transmission rates from $Unif(2, 5)$ for non-zero edges of A . We slightly modify the generation process for Ray model by decreasing the transmission rates on non-zero edges of Θ and Ψ . We instead sample weights from $Unif(0.1, 0.8)$ for non-zero edges in Ψ_1, Ψ_2 and weights from $Unif(0.02, 2)$ for those in Θ . This modification aims to force significantly large differences in infection time that contribute to large differences in probability of infection among potential parents.

We also add edges in Θ to avoid all-zero row or column in $\Theta \cup \Psi$ to ensure diffusion network connectivity. The resulting Θ and Ψ have edge density at 0.01 and 0.05, respectively. Based on Θ and Ψ , we generate different numbers of cascade samples at $C = 500, 1000, 1500, 2000$ with a fix observation window length at $T = 10$. In the Pow transmission model, we select delay parameter $\delta = 1$.

B.3 Network recovery under large network settings

In this experiment, for each network size $N = 500, 1000, 2000, 4000$, we generate the latent diffusion network Ψ following the same procedure as in B.2 but fixing its rank at 5 and edge density at 0.01. Given Ψ , we also follow the same procedure in B.2 to generate Θ . The resulting Θ has an edge density of 0.001. Based on Θ and Ψ , we generate $C = 50000$ independent cascade samples with a fixed observation window length of $T = 10$.

To cope with memory limits when the network size N and the sample size C are large, in each outer iteration, we stream the data in batches of size B : we process one batch at a time to compute the contributions (likelihood terms, gradients) required by both the E-step and the M-step, accumulate these quantities across all $\lceil M/B \rceil$ batches, and then carry out the parameter update using the aggregated totals—thus matching the effect of a full-batch EM update while keeping memory usage bounded. When more memory is available, one can increase the batch size B accordingly for better parallelization on GPU and thus shorter execution time per iteration.

C Additional numerical results

C.1 Network recovery under different support overlap

In this subsection, we investigate the performance of our proposed method under different levels of overlap between Θ and Ψ , defined as $\text{overlap}(\Theta, \Psi) = \sum_{i,j} \mathbf{I}(A_{ij} \times B_{ij}) / \sum_{i,j} \mathbf{I}(A_{ij} + B_{ij})$, where A, B are the support of Θ, Ψ , respectively.

We generate $\Psi = \Psi_1 \Psi_2^T$ following similar procedure in B.1 except that we sample weights of non-zero edges in Ψ_1 and Ψ_2 from $Unif(0.2, 1.5)$ and increase its rank to 30 to better control its overlap degree with Θ . The edge density of the generated Ψ is 0.025 and the transmission rates on its non-zero edges range from 0.05 to 2.

Given Ψ and its support B , we generate different supports A following the same procedure in B.2 while changing their overlaps and non-overlaps per column with B so that $\text{overlap}(\Theta, \Psi)$ takes three levels at about 0.1, 0.3, 0.5. Then we create Θ by sampling transmission rates from $Unif(0.1, 0.2) \cup (1.9, 2.0)$ for the overlap $\{(i, j) \mid A_{ij}B_{ij} = 1\}$, and sampling from $Unif(1, 2)$ for non-overlap support $\{(i, j) \mid A_{ij} = 1, B_{ij} = 0\}$.

We also add edges in Θ with transmission rates from $Unif(1, 2)$ to avoid all-zero row or column in $\Theta \cup \Psi$ to ensure diffusion network connectivity. Based on Θ and Ψ , we generate different numbers of cascade samples from Exp transmission model, and fix the observation window length at $T = 10$.

Table 5 shows that the diffusion network recovery performance of the proposed method improves as sample sizes increases. Additionally, as overlap level between Θ and Ψ increases, the transmission rates estimations MAE on both Θ, Ψ and π decreases while estimation of Ψ 's support improves.

This pattern uncovers a trade-off between network differentiation and topology recovery in network mixture model. As the overlap level increases, the joint diffusion complexity on Θ and Ψ decreases due to the shared diffusion pathways and support information. On the other hand, it becomes more difficult and requires more samples to differentiate Θ and Ψ on their overlap component.

Table 5: Diffusion network estimations from the proposed method under different degrees of overlapping between diffusion networks.

overlap(Θ, Ψ)		MAE_{Θ}	MAE_{Ψ}	MAE_{π}	Acc_{Ψ}	Pre_{Ψ}	Rec_{Ψ}
0.1	C = 500	0.325(0.014)	0.510(0.018)	0.314(0.013)	0.690(0.007)	0.553(0.009)	0.919(0.005)
	C = 1000	0.309(0.015)	0.414(0.016)	0.300(0.016)	0.793(0.008)	0.669(0.010)	0.974(0.003)
	C = 1500	0.307(0.016)	0.391(0.018)	0.297(0.016)	0.827(0.013)	0.711(0.018)	0.989(0.002)
	C = 2000	0.277(0.012)	0.360(0.012)	0.288(0.015)	0.853(0.004)	0.748(0.007)	0.994(0.002)
0.3	C = 500	0.393(0.012)	0.577(0.017)	0.435(0.016)	0.764(0.008)	0.647(0.010)	0.933(0.008)
	C = 1000	0.388(0.018)	0.519(0.016)	0.454(0.015)	0.852(0.006)	0.756(0.010)	0.977(0.003)
	C = 1500	0.387(0.013)	0.507(0.018)	0.456(0.012)	0.887(0.008)	0.804(0.013)	0.989(0.002)
	C = 2000	0.366(0.017)	0.489(0.022)	0.452(0.018)	0.908(0.003)	0.835(0.004)	0.994(0.002)
0.5	C = 500	0.458(0.014)	0.618(0.014)	0.533(0.017)	0.810(0.017)	0.712(0.023)	0.940(0.006)
	C = 1000	0.438(0.018)	0.555(0.013)	0.555(0.011)	0.884(0.034)	0.809(0.053)	0.975(0.006)
	C = 1500	0.442(0.018)	0.556(0.026)	0.560(0.011)	0.919(0.027)	0.863(0.046)	0.986(0.004)
	C = 2000	0.421(0.019)	0.526(0.019)	0.557(0.011)	0.935(0.030)	0.886(0.052)	0.992(0.004)

C.2 Performance under different time window length

In this subsection, we investigate the network recovery performance under different observation window lengths, which ranges in $T \in \{1, 2, 3, 5, 10\}$.

We generate Θ and Ψ following the same procedure in C.1 and fix their overlap level at about 0.30. We generate $C = 1500$ cascade samples from transmission models Exp, Pow, Ray.

We illustrate the performance in Figure 5 and 6. In general, both transmission rate estimations and network topology recovery improve as the observation window becomes T longer since effective diffusion information increases within each cascade sample. Specifically, by Figure 5, the proposed method achieves better MAE of Θ and Ψ under Ray (green line with shaded one standard deviation error range) and Pow (yellow line) cascade model. MAE criterion under Exp model is not strictly monotone, but its overall trend matches that of Pow and Ray. Figure 6 shows that accuracy and precision of Ψ also increases as T increases and then remains stable. The results in this subsection suggest to set a large T in practice, which is also consistent with our theoretical analysis.

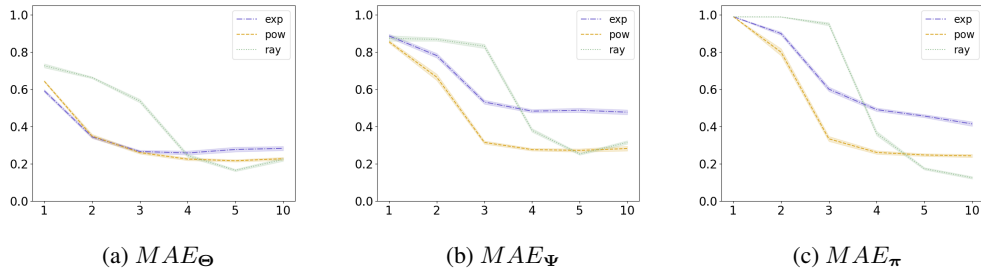


Figure 5: Parameter estimation under different time window lengths.

C.3 Network recovery without support information

In this subsection, we investigate the performance of the proposed method when support \mathbf{A} of diffusion network Θ is unobserved.

We consider diffusion networks Θ and Ψ of size $N = 100$, generated by similar scheme as in C.1. Since the size of network is reduced, rank of Ψ_1, Ψ_2 , where $\Psi = \Psi_1 \Psi_2^T$, is accordingly reduced to 15. The generated Θ and Ψ have edge densities of 0.02 and 0.05, respectively. Non-zero edges of Θ and Ψ have ranges 0.1 to 2 and 0.05 to 2.2, respectively. The level of overlap between the two networks is

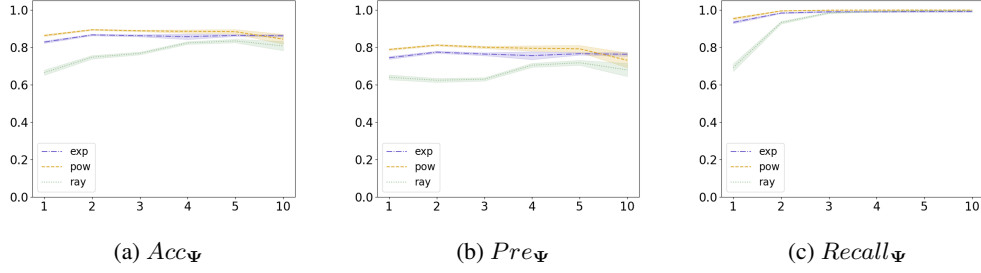


Figure 6: Topology recovery of latent network estimation under different time window lengths.

fixed at 0.15. We generate cascade samples of different sizes $C = 2000, 3000, 4000, 5000$ and from all three different transmission models Exp, Pow, and Ray.

Table 6 shows that when support of Θ is not observed, performance of the proposed method on transmission rate estimation still increases as number of samples increases for all three cascade generation models. Benchmark values in Table 6 are obtained from the proposed method with information on the support of Θ . Figure 7 and 8 show the relationship between precision and recall of the estimated Θ and Ψ when their entries below some varying thresholds are truncated to 0. As sample size increases, the performance of the proposed method on network topology recovery increases (shown by the larger area under P-R curves). Although the proposed method without support information of Θ cannot achieve as good performance as when the information is available, its performance on both transmission rate estimation and network topology recovery still exhibits converging pattern to the benchmark performance as sample size increases.

Table 6: Parameter estimation from the proposed method when support of Θ is unknown.

		MAE_{Θ}	Acc_{Θ}	MAE_{Ψ}	Acc_{Ψ}
Exp	C = 2000	0.607(0.028)	0.628(0.013)	0.672(0.027)	0.803(0.015)
	C = 3000	0.533(0.038)	0.702(0.016)	0.604(0.030)	0.883(0.007)
	C = 4000	0.447(0.107)	0.742(0.053)	0.520(0.089)	0.902(0.009)
	C = 5000	0.327(0.113)	0.787(0.047)	0.414(0.095)	0.904(0.024)
	Benchmark	0.088(0.005)	0.976(0.003)	0.210(0.008)	0.926(0.006)
Ray	C = 2000	0.665(0.044)	0.573(0.021)	0.699(0.039)	0.601(0.020)
	C = 3000	0.553(0.112)	0.680(0.053)	0.592(0.093)	0.699(0.040)
	C = 4000	0.292(0.054)	0.824(0.027)	0.376(0.042)	0.811(0.016)
	C = 5000	0.232(0.022)	0.858(0.012)	0.326(0.017)	0.837(0.010)
	Benchmark	0.067(0.002)	1.000(0.000)	0.177(0.002)	0.941(0.023)
Pow	C = 2000	0.604(0.064)	0.732(0.026)	0.537(0.051)	0.855(0.010)
	C = 3000	0.447(0.051)	0.786(0.043)	0.423(0.037)	0.871(0.008)
	C = 4000	0.331(0.062)	0.822(0.025)	0.344(0.048)	0.879(0.010)
	C = 5000	0.346(0.047)	0.791(0.022)	0.341(0.034)	0.869(0.010)
	Benchmark	0.061(0.002)	0.976(0.003)	0.175(0.008)	0.916(0.002)

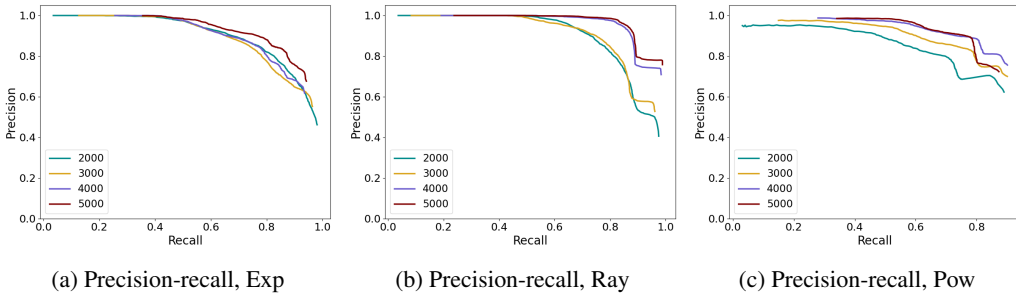


Figure 7: Precision-recall plot of Θ

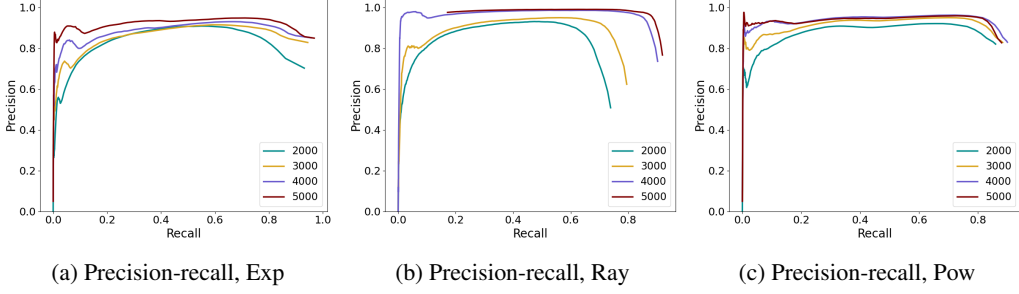


Figure 8: Precision-recall plot of Ψ

D Additional details on method implementation

To give proper initializations to our proposed method, we use estimation result Λ from NetRate [31]. For experiments in Section 4.1, 4.2, C.1, and C.2, we initialize $\Theta_0 = \Lambda \odot A$ and $\Psi_0 = \Lambda \odot (1 - A)$, where A is the support of Θ .

We employ a new initialization method in C.3 since information on A , the support of Θ , is missing. We initialize Ψ_0 as the truncated SVD of Λ at rank r . We then initialize Θ_0 as the top $\gamma\%$ entries of the residual $\Lambda - \Psi_0$. Both r and γ are tunable hyperparameters, where $1 \leq r \leq N$ and $0 < \gamma < 100$. In the optimization process of C.3, instead of forcing gradient updates of Θ taking non-zero values only on A , we introduce an additional l_1 regularizer on Θ in the loss function to promote its sparsity and update Θ via an additional l_1 soft-thresholding operator.

We also provide a short discussion in this section on the general strategy to choose the best hyperparameters $[\lambda_\Theta, \lambda_\Psi, \mu, \rho_\Theta, \rho_\Psi]$, where λ denotes learning rate, μ denotes low-rank penalty, ρ denotes l_1 penalty, and the subscript denotes the network associated with.

In general, as network size, sample size, or edge density increases, smaller $\lambda_\Theta, \lambda_\Psi$ are required for stable and good performance of the algorithm. In terms of different cascade transmission models, the optimal learning rates for Pow model have the largest magnitude, while those for Ray model have the smallest. Level of overlap, length of observation window, or the rank of Ψ does not significantly affect learning rate choices.

For fixed $\{\Theta, \Psi, \pi\}$, there exist best threshold $\lambda_\Psi \cdot \mu$ for singular value soft-thresholding operation and best thresholds $\lambda_\Theta \cdot \rho_\Theta, \lambda_\Psi \cdot \rho_\Psi$ for l_1 soft-thresholding operation. Thus, for fixed $\{\Theta, \Psi, \pi\}$, we need to adjust penalties inversely proportional to learning rates for the best performance of the proposed algorithm. l_1 penalties should be increased as transmission rates on Θ, Ψ increase, while μ should be increased as rank of Ψ increases.

E Proofs for theoretical results

Proposition 1

Proof. To prove the column-wise identification between $\Theta_{\cdot i}$ and $\Psi_{\cdot i}$, we only need to consider node i 's direct parents in two networks as $\mathcal{R}_i^\Theta := \{j \in \{1, \dots, N\} \mid \Theta_{ji} > 0\}$ and $\mathcal{R}_i^\Psi := \{j \in \{1, \dots, N\} \mid \Psi_{ji} > 0\}$ as the sets of nodes that can directly reach node i on Θ and Ψ . Denote the activation times of the direct parents of node i as $t_{\mathcal{R}_i} \in [0, T]^{|\mathcal{R}_i^\Theta \cup \mathcal{R}_i^\Psi|} := \{t_j \mid j \in \mathcal{R}_i^\Theta \cup \mathcal{R}_i^\Psi\}$.

If $\pi_i P_I(t; \Theta_{\cdot i}) + (1 - \pi_i) P_I(t; \Psi_{\cdot i}) = \tilde{\pi}_i P_I(t; \tilde{\Theta}_{\cdot i}) + (1 - \tilde{\pi}_i) P_I(t; \tilde{\Psi}_{\cdot i})$, then we have

$$\pi_i \sum_j H(t_j; \Theta_{ji}) \prod_j S(t_j; \Theta_{ji}) + (1 - \pi_i) \sum_j H(t_j; \Psi_{ji}) \prod_j S(t_j; \Psi_{ji}) \quad (15)$$

$$= (1 - \tilde{\pi}_i) \sum_j H(t_j; \tilde{\Theta}_{ji}) \prod_j S(t_j; \tilde{\Theta}_{ji}) + (1 - \tilde{\pi}_i) \sum_j H(t_j; \tilde{\Psi}_{ji}) \prod_j S(t_j; \tilde{\Psi}_{ji}), \quad (16)$$

where the summation and multiplication are over $j \in \mathcal{R}_i^\Theta \cup \mathcal{R}_i^\Psi$ with $p := |\mathcal{R}_i^\Theta \cup \mathcal{R}_i^\Psi|$. Given that survival function satisfies $S(t_j; \lambda_{ji}) = \exp\{\lambda_{ji} h(t_i - t_j)\}$ for some non-negative function $h(\cdot)$, then

the infection likelihood and hazard function are

$$f(t_j; \lambda_{ji}) = \lambda_{ji} \exp\{-\lambda_{ji} h(t_i - t_j)\} h'(t_i - t_j); \quad H(t_j; \lambda_{ji}) = \lambda_{ji} h'(t_i - t_j).$$

Denote $w_1(t_j) = \pi_i h'(t_i - t_j) \prod_k S(t_k; \Theta_{ki})$, $w_2(t_j) = (1 - \pi_i) h'(t_i - t_j) \prod_k S(t_k; \Psi_{ki})$, $\tilde{w}_1(t_j) = \tilde{\pi}_i h'(t_i - t_j) \prod_k S(t_k; \tilde{\Theta}_{ki})$, and $\tilde{w}_2(t_j) = (1 - \tilde{\pi}_i) h'(t_i - t_j) \prod_k S(t_k; \tilde{\Psi}_{ki})$. Then (15) can be re-written as

$$\sum_j w_1(t_j) \Theta_{ji} + \sum_j w_2(t_j) \Psi_{ji} - \sum_j \tilde{w}_1(t_j) \tilde{\Theta}_{ji} - \sum_j \tilde{w}_2(t_j) \tilde{\Psi}_{ji} = 0, \quad (17)$$

which should hold for any $t_{\mathcal{R}_i}$. Consider a $4p \times 4p$ matrix Γ with each row being $[w_1(t_1), \dots, w_1(t_p), w_2(t_1), \dots, w_2(t_p), \tilde{w}_1(t_1), \dots, \tilde{w}_1(t_p), \tilde{w}_2(t_1), \dots, \tilde{w}_2(t_p)]$ where each row takes different values of $t_{\mathcal{R}_i}$. Based on assumption that there exist $j^* \neq i$ such that $\Theta_{j^*i} \neq \Psi_{j^*i}$, then up to permutation between $\tilde{\Theta}_{\cdot i}$ and $\tilde{\Psi}_{\cdot i}$, the columns corresponding to different t_j will be linear independent. Without loss of generality, we only need to investigate the $4p \times 4$ submatrix Γ_j with each row being $[w_1(t_j), w_2(t_j), \tilde{w}_1(t_j), \tilde{w}_2(t_j)]$. We prove that $w_1(t_j) = \tilde{w}_1(t_j)$, $w_2(t_j) = \tilde{w}_2(t_j)$ and $\Theta_{ji} = \tilde{\Theta}_{ji}$, $\Psi_{ji} = \tilde{\Psi}_{ji}$ by contradiction argument. For two different values t_j and t'_j , we have

$$\frac{w_1(t_j)}{w_1(t'_j)} = \frac{w_2(t_j)}{w_2(t'_j)} \Rightarrow \frac{\prod_j S(t_j, \Theta_{ji})}{\prod_j S(t'_j, \Theta_{ji})} = \frac{\prod_j S(t_j, \Psi_{ji})}{\prod_j S(t'_j, \Psi_{ji})} \quad (18)$$

$$\Rightarrow \sum_j (\Theta_{ji} - \Psi_{ji})(h(t_i - t_j) - h(t_i - t'_j)) = 0, \quad (19)$$

and we can choose $t_j \neq t'_j$ for those j^* such that $\Theta_{j^*i} \neq \Psi_{j^*i}$. Then the above equation implies $\Theta_{j^*i} = \Psi_{j^*i}$, which cause contradiction. Therefore, the rank of Γ_j is larger than 1.

If there exists a, b such that $a \times w_1(t_j) + b \times w_2(t_j) = \tilde{w}_1(t_j)$ hold for any $t_{\mathcal{R}_i}$. Then we let $t_j = t_i$ for all $j \in \mathcal{R}_i^\Theta \cup \mathcal{R}_i^\Psi$ except j^* . Then we have

$$a \times \pi_i h'(t_i - t_{j^*}) \exp\{h(t_i - t_{j^*}) \Theta_{j^*i}\} + b \times (1 - \pi_i) h'(t_i - t_{j^*}) \exp\{h(t_i - t_{j^*}) \Psi_{j^*i}\} \quad (20)$$

$$= \tilde{\pi}_i h'(t_i - t_{j^*}) \exp\{h(t_i - t_{j^*}) \tilde{\Theta}_{j^*i}\} \quad (21)$$

$$\Rightarrow a' \exp\{h(t_i - t_{j^*}) (\Theta_{j^*i} - \tilde{\Theta}_{j^*i})\} + b' \exp\{h(t_i - t_{j^*}) (\Psi_{j^*i} - \tilde{\Psi}_{j^*i})\} = 1, \quad (22)$$

where $a' = a \frac{\pi_i}{\tilde{\pi}_i}$, $b' = b \frac{1-\pi_i}{1-\tilde{\pi}_i}$. Notice that at least one of $\Theta_{j^*i} - \tilde{\Theta}_{j^*i}$ and $\Psi_{j^*i} - \tilde{\Psi}_{j^*i}$ is not zero and the monotonicity of $\exp(\lambda t)$ in terms of t , the above exponential function equation cannot hold as long as $h(t_i - t)$ is not constant over t . Therefore, the rank of Γ_j can not be 2. Using the same argument above, we can show the rank of Γ_j can not be 3 as well.

If the rank of Γ_j is 4, i.e., Γ_j is full rank, then Γ is also full rank by applying above argument on each $j \in \mathcal{R}_i^\Theta \cup \mathcal{R}_i^\Psi$ except j^* . Based on (17), we have $\Gamma[\Theta_{\cdot i}^\top, \Psi_{\cdot i}^\top, \tilde{\Theta}_{\cdot i}^\top, \tilde{\Psi}_{\cdot i}^\top]^\top = \mathbf{0}$, which leads to $\Theta_{\cdot i}^\top = \Psi_{\cdot i}^\top = \tilde{\Theta}_{\cdot i}^\top = \tilde{\Psi}_{\cdot i}^\top = \mathbf{0}$ and contradicts to assumption $\|\Theta_{\cdot i}\|_1 + \|\Psi_{\cdot i}\|_1 > 0$. Therefore, $w_1(t_j) = \tilde{w}_1(t_j)$, $w_2(t_j) = \tilde{w}_2(t_j)$ and $\Theta_{ji} = \tilde{\Theta}_{ji}$, $\Psi_{ji} = \tilde{\Psi}_{ji}$ for all j . Finally, by the definition of $w_1(t_j)$ and $\tilde{w}_1(t_j)$, we can derive $\pi_i = \tilde{\pi}_i$. The statement based on P_U can be similarly derived. \square

Proposition 2

Proof. Given Proposition 1, we have the column-wise identification of $\Theta_{\cdot i}$ and $\Psi_{\cdot i}$, and only the identification of Θ and Ψ up to the column exchange between Θ and Ψ . Notice that Proposition 1 guarantee the identification of $C = \Theta + \Psi$. Then if

$$\max_{N \in \Lambda_1(\Theta), \|N\|_\infty \leq 1} \|N\|_2 \times \max_{N \in \Lambda_2(\Psi), \|N\|_2 \leq 1} \|N\|_\infty < 1, \quad (23)$$

we have $\Lambda_1(\Theta) \cap \Lambda_2(\Psi) = \emptyset$ based on Proposition 1 in [6]. Notice that $\Theta \in \Lambda_1(\Theta)$ and $\Psi \in \Lambda_2(\Psi)$, then Θ and Ψ are identifiable conditioning on C is identifiable. \square

Theorem 3.1

Proof. We first show $\Xi_{\Theta}(s)$, $\Xi_{\Psi}(\rho)$, and Ξ_{π} are convex sets for any $s > 0$, $\rho > 0$ by checking that the convex combination of any two elements in $\Xi_{\Theta}(s)$, $\Xi_{\Psi}(\rho)$, or Ξ_{π} still lies in the set.

For any $\Theta_1, \Theta_2 \in \Xi_{\Theta}(s)$ and $0 \leq \lambda \leq 1$,

$$0 \leq \lambda(\Theta_1)_{ij} + (1 - \lambda)(\Theta_2)_{ij} \leq \lambda\beta_1 + (1 - \lambda)\beta_1 = \beta_1 \quad (24)$$

$$(\lambda\Theta_1 + (1 - \lambda)\Theta_2) \odot (I - A) = \lambda[\Theta_1 \odot (I - A)] + (1 - \lambda)[\Theta_2 \odot (I - A)] = 0$$

$\Xi_{\Theta}(s)$ is thus a convex set. Note that $s > 0$ avoids the trivial case where A is a zero matrix.

It is easy to argue that Ξ_{π} is a convex set using similar argument to (24).

For any $\Psi_1, \Psi_2 \in \Xi_{\Psi}(\rho)$ and $0 \leq \lambda \leq 1$,

$$\|\lambda\Psi_1 + (1 - \lambda)\Psi_2\|_* \leq \lambda\|\Psi_1\|_* + (1 - \lambda)\|\Psi_2\|_* = \rho \quad (25)$$

By (25) and similar argument to (24), $\Xi_{\Psi}(\rho)$ is a convex set.

Next, we show $Q_3(\pi \mid \Omega^{(s)})$ is concave over π . Recall the formulation of $Q_3(\pi \mid \Omega^{(s)})$ in A.1

$$Q_3(\pi \mid \Omega^{(s)}) = \frac{1}{C} \sum_{c=1}^C [(\hat{\pi}^c)^T \log(\pi) + (\mathbf{1} - \hat{\pi}^c)^T \log(\mathbf{1} - \pi)]$$

When $\pi \in \Xi_{\pi}$, both $\log(\pi)$ and $\log(\mathbf{1} - \pi)$ are concave over π . From the above formulation, $Q_3(\pi \mid \Omega^{(s)})$ is concave over π by linearity of concavity.

Without loss of generality, we show $Q_1(\Theta \mid \Omega^{(s)})$ is concave on Θ under the assumption that hazard function $H(t \mid t', \lambda)$ satisfies $\partial H^2(t \mid t', \lambda) / \partial \lambda^2 = 0$ for $t \geq t'$.

Note that this assumption immediately implies the concavity of $H(t \mid t', \lambda)$ and the log concavity of $S(t \mid t', \lambda)$ over λ .

Recall the definition of P_I and P_U in Section 3.1 we have

$$\log P_I(t_i; \Theta_{\cdot i}) = \log \left(\sum_{j: t_j < t_i} H(t_i \mid t_j, \Theta_{ji}) \prod_{k: t_k < t_i} S(t_i \mid t_k, \Theta_{ki}) \right) \quad (26)$$

$$\log P_U(T; \Theta_{\cdot i}) = \log \left(\prod_{j: t_j < T} S(T \mid t_j, \Theta_{ji}) \right) \quad (27)$$

Rewrite (26) and (27) in the following form,

$$\log P_I(t_i; \Theta_{\cdot i}) = \log \left(\sum_j H(t_i \mid t_j, \Theta_{ji}) \right) + \sum_k \log S(t_i \mid t_k, \Theta_{ki})$$

$$\log P_U(T; \Theta_{\cdot i}) = \sum_j \log S(T \mid t_j, \Theta_{ji})$$

Then by concavity and monotonicity of log function, composition rule for concavity, and linearity of concavity, both $\log P_I(t_i; \Theta_{\cdot i})$ and $\log P_U(T; \Theta_{\cdot i})$ are concave over $\Theta_{\cdot i}$.

Concavity of $Q_1(\Theta \mid \Omega^{(s)})$ over Θ immediately follows from composition rule for concavity and the proof for $Q_3(\pi \mid \Omega^{(s)})$. Concavity of $Q_2(\Theta \mid \Psi^{(s)})$ directly follows this proof by changing Θ to Ψ .

By the above proof, the linear combination $\sum_{j: t_j^c > T} \hat{\pi}_j^c \log P_U(T; \Theta_{\cdot j})$ where $\hat{\pi}_j^c$ depends on $\Omega^{(s)}$ instead of Θ is concave over Θ . Therefore, $\mathbb{E}_t(\sum_{j: t_j^c > T} \hat{\pi}_j^c \log P_U(T; \Theta_{\cdot j}))$ is concave in terms of Θ . In addition, notice that $\frac{\partial^2 \mathbb{E}_t[Q_1(\Theta \mid \Omega^{(s)})]}{\partial^2 \Theta}$ is a $(N^2 - N) \times (N^2 - N)$ diagonal block matrix as $\frac{\partial^2 \mathbb{E}_t[Q_1(\Theta \mid \Omega^{(s)})]}{\partial \Theta_{ij} \partial \Theta_{kl}} = 0$ if $j \neq l$, which leads to N blocks $\left(\frac{\partial^2 \mathbb{E}_t[Q_1(\Theta \mid \Omega^{(s)})]}{\partial \Theta_{ji} \partial \Theta_{ki}} \right)_{(N-1) \times (N-1)}$, $i = 1, \dots, N$. And the i th block corresponds to the second derivative of

$$\frac{1}{C} \sum_{c=1}^C \hat{\pi}_i^c \left(\mathbf{1}_{\{t_i^c \leq T\}} \log P_I(t_i; \Theta_{\cdot i}) + \mathbf{1}_{\{t_i^c > T\}} \log P_U(T; \Theta_{\cdot i}) \right)$$

Based on the above arguments, we only need to show that the strictly concavity for each block in $\frac{\partial^2 \mathbb{E}_t[\mathcal{Q}_1(\boldsymbol{\Theta} | \boldsymbol{\Omega}^{(s)})]}{\partial^2 \boldsymbol{\Theta}}$, i.e., $\mathbb{E}_t(\frac{1}{C} \sum_{c=1}^C \mathbf{1}_{\{t_i^c \leq T\}} \hat{\pi}_i^c \log \mathbf{P}_I(t_i; \boldsymbol{\Theta}_{\cdot i}))$. We then set C to be larger enough such that $C_0 := \sum_{c=1}^C \mathbf{1}_{\{t_i^c \leq T\}} > N - 1$ and set the index of cascade where i is infected as $1, \dots, C_0$, and

$$\mathbb{E}_t\left(\frac{1}{C} \sum_{c=1}^C \mathbf{1}_{\{t_i^c \leq T\}} \hat{\pi}_i^c \log \mathbf{P}_I(t_i; \boldsymbol{\Theta}_{\cdot i})\right) = \mathbb{E}_t\left(\frac{1}{C} \sum_{c=1}^{C_0} \hat{\pi}_i^c \log \mathbf{P}_I(t_i; \boldsymbol{\Theta}_{\cdot i})\right)$$

Then we denote $\mathcal{Q} = \frac{1}{C} \sum_{c=1}^{C_0} \hat{\pi}_i^c \log \mathbf{P}_I(t_i; \boldsymbol{\Theta}_{\cdot i})$ and follow the argument of Lemma 10 in [8] to prove \mathcal{Q} is positive definite. Following their notations and $\alpha := \boldsymbol{\Theta}_{\cdot i}$, we can write

$$\mathcal{Q} = \mathbf{D}(\alpha) + \frac{1}{C} \mathbf{X}(\alpha) [\mathbf{X}(\alpha)]^\top$$

where $\mathbf{D}(\alpha) = \frac{1}{C} \sum_{c=1}^{C_0} \mathbf{D}(\mathbf{t}^c; \alpha)$ and $\mathbf{D}(\mathbf{t}^c; \alpha)$ is a diagonal matrix with $[\mathbf{D}(\mathbf{t}^c; \alpha)]_{jj} = -\hat{\pi}_i^c S''(t_i^c | t_j^c; \alpha_k) - \hat{\pi}_i^c h^{-1}(\mathbf{t}^c, \alpha) H''(t_i^c | t_j^c; \alpha_k)$, and $h(\mathbf{t}^c, \alpha) = \sum_{j: t_j^c < t_i^c} H(t_i^c | t_j^c; \alpha_k)$. In addition, $\mathbf{X}(\alpha)$ is a $(N - 1)$ -by- C_0 matrix as

$$\mathbf{X}(\alpha) = [\mathbf{X}(\mathbf{t}^1; \alpha) | \mathbf{X}(\mathbf{t}^2; \alpha) | \dots | \mathbf{X}(\mathbf{t}^{C_0}; \alpha)],$$

where each column $\mathbf{X}(\mathbf{t}^c; \alpha) = \sqrt{\hat{\pi}_i^c} h^{-1}(\mathbf{t}^c, \alpha) \nabla_\alpha h(\mathbf{t}^c, \alpha)$. Given the assumption that the probabilities of being source node $\mathbf{P}(v) > 0$ for $v \in \mathcal{R}$ where \mathcal{R} denotes the set of nodes from which i is reachable via a directed path, we can follow the same argument in Lemma 10 of [8] to show \mathcal{Q} is strictly concave in terms of $\boldsymbol{\Theta}$. Then $\mathbb{E}_t(\mathcal{Q})$ is also strictly concave. The strong concavity of $\mathbb{E}_t[\mathcal{Q}_2(\boldsymbol{\Psi} | \boldsymbol{\Omega}^{(s)})]$ in terms of $\boldsymbol{\Psi}$ can be similarly proved. \square

F Information on computer resources

Numerical experiments in this paper were carried out in Google Colab on a single Nvidia A100 GPU with 80GB of memory available. Regardless of failed experiments, the entire numerical experiment section took approximately 50 hours of computation time.

In addition, to evaluate the benchmark algorithm ConNle, we used SNOPT (version 7.7) as the underlying constrained optimization solver [15, 16].

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: [The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.](#)

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: [The paper discusses the limitations of this work in the Conclusion section.](#)

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: [All assumptions are stated and all proofs are provided in either Methodology section or Appendix.](#)

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: [The paper fully discloses all the information needed to reproduce the main experimental results. Code for experiments are provided as supplementary material.](#)

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: [The paper provides open access to the data and code with sufficient instructions to faithfully reproduce the main experimental results.](#)

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: [The paper specifies all the training and test details necessary to understand the results.](#)

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: [The paper reports standard deviations where appropriate with suitable explanations.](#)

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: [The paper provides information on computer resources in the Appendix.](#)

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: [The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics](#)

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: [In Real cascading data analysis section, we apply our method to analyze the research topic cascades in social science across U.S. universities and uncover the latent research topic diffusion networks among the top U.S. social science programs.](#)

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: [This paper uses open-domain data and code, properly crediting the license and creators.](#)

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: [New code accompanying this paper is well documented and submitted as supplementary material.](#)

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]