# Information-Theoretic Lower Bounds for Recovery of Diffusion Network Structures

Keehwan Park
Department of Computer Science
Purdue University
Email: park451@purdue.edu

Jean Honorio
Department of Computer Science
Purdue University
Email: jhonorio@purdue.edu

***Abstract*—We study the information-theoretic lower bound of the sample complexity of the correct recovery of diffusion network structures. We introduce a discrete-time diffusion model based on the Independent Cascade model for which we obtain a lower bound of order $\Omega(k \log p)$, for directed graphs of $p$ nodes, and at most $k$ parents per node. Next, we introduce a continuous-time diffusion model, for which a similar lower bound of order $\Omega(k \log p)$ is obtained. Our results show that the algorithm of [1] is statistically optimal for the discrete-time regime. Our work also opens the question of whether it is possible to devise an optimal algorithm for the continuous-time regime.**

## I. Introduction

In recent years, the increasing popularity of online social network services, such as Facebook, Twitter, and Instagram, allows researchers to access large influence propagation traces. Since then, the influence diffusion on social networks has been widely studied in the data mining and machine learning communities. Several studies showed how influence propagates in such social networks as well as how to exploit this effect efficiently. Domingos et al. [2] first explored the use of social networks in viral marketing. Kempe et al. [3] proposed the influence maximization problem on the Independent Cascade (IC) and Linear Threshold (LT) models, assuming all influence probabilities are known. [4], [5] studied the learning of influence probabilities for a known (fixed) network structure.

The network inference problem consists in discovering the underlying functional network from cascade data. The problem is particularly important since regardless of having some structural side information, e.g., friendships in online social networks, the functional network structure, which reflects the actual influence propagation paths, may look greatly different. Adar et al. [6] first explored the problem of inferring the underlying diffusion network structure. The subsequent researches [7], [8] have been done in recent years and the continuous-time extensions [9]–[11] have also been explored in depth.

**Basic diffusion model.** Consider a directed graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, \ldots, p\}$ is the set of nodes and $\mathcal{E}$ is the set of edges. Next, we provide a short description for the discrete-time IC model [3]. Initially we draw an initial set of active nodes from a source distribution. The process unfolds in discrete steps. When node $j$ first becomes active at time $t$, it independently makes a single attempt to activate each of its outgoing, inactive neighbors $i$, with probability $\theta_{j,i}$. If $j$ succeeds, then $i$ will become active at time $t+1$. If $j$ fails, then it makes no further attempts to activate $i$. And this process runs until no more activations are possible.

**Related works.** Research on the sample complexity of the network inference problem is very recent [1], [12]–[15]. Netrapalli et al. [12] studied the network inference problem based on the discrete-time IC model and showed that for graphs of $p$ nodes and at most $k$ parents per node, $\mathcal{O}(k^2 \log p)$ samples are sufficient, and $\Omega(k \log p)$ samples are necessary. However, as Daneshmand et al. [14] have pointed out, their model only considers the discrete-time diffusion model and the correlation decay condition is rather restrictive since it limits the number of new activations at every step. Abrahao et al. [13] proposed the First-Edge algorithm to solve the network inference problem and also suggested lower bounds but their results are specific to their algorithm, i.e., the lower bounds are not information-theoretic.

In [14], Daneshmand et al. worked on the continuous-time network inference problem with $\ell$-1 regularized maximum likelihood estimation and showed that $\mathcal{O}(k^3 \log p)$ samples are sufficient, using the *primal-dual witness* method. Narasimhan et al. [15] explored various influence models including IC, LT, and Voter models under the Probably Approximately Correct learning framework. Pouget-Abadie et al. [1] studied various discrete-time models under the restricted eigenvalue conditions. They also proposed the first algorithm which recovers the network structure with high probability in $\mathcal{O}(k \log p)$ samples.

It is important to note that, as we will see later in the paper, we show information-theoretic lower bounds of order $\Omega(k \log p)$, confirming that the algorithm in [1] is statistically optimal. However, since their algorithm only considered discrete-time models, developing a new algorithm for continuous-time models with the sufficient condition on the sample complexity of order $\mathcal{O}(k \log p)$ can be an interesting future work.

## II. Ensemble of Discrete-time Diffusion Networks

Lower bounds of the sample complexity for general graphs under the IC and LT models [3] seem to be particularly difficult to analyze. In this paper, we introduce a simple network under IC model, which fortunately allow us to show sample complexity lower bounds that match the upper bounds found in [1] for discrete-time models.

### A. A simple two-layer network

Here we considered the two-layer IC model shown in Figure 1. Although not realistic, the considered model allows to show that even in this simple two-layer case, we require $\Omega(k \log p)$ samples in order to avoid network recovery failure.

In Figure 1, each circle indicates a node and each edge $(j, i)$ with its influence probability $\theta$ indicates that a cascade can be propagated from node $j$ to $i$ or equivalently node $j$ activates $i$ with probability $\theta$. The model assumes that there exists a super source node $s_1$, which is already activated at time zero and at time 1, it independently tries to activate $p$ parent nodes with probability $\theta_0$ and $s_2$ with probability 1. There exist a child node $p+1$, which has exactly $k+1$ parents including $s_2$. Then at time 2, $s_2$ and all direct parents of $p+1$, which have been activated at time 1, independently try to activate the child node $p+1$ with probability $\theta_0$ and $\theta$, respectively. We use $t_i = \infty$ to indicate that a node $i$ has not been activated during the cascading process. Note that these influence probabilities can be generalized without too much effort.

Given the model with unknown edges between parent nodes and the child node $p+1$, and a set of $n$ samples $\boldsymbol{t}^{(\mathbf{1})}, \boldsymbol{t}^{(\mathbf{2})}, \ldots, \boldsymbol{t}^{(\boldsymbol{n})} \in \{1, \infty\}^p \times \{2, \infty\}$, the goal of the learner is to recover the $k$ edges or equivalently to identify the $k \ll p$ direct parents of the child node $p+1$. Each sample is a $(p+1)$-dimensional vector, $\boldsymbol{t} = (t_1, \ldots, t_p, t_{p+1})$, and includes all the activation times of the parent and child nodes. A parent node $i \in \{1, \ldots, p\}$ is either activated at time 1 (i.e., $t_i = 1$) or not (i.e.,$t_i = \infty$). The child node $p+1$ is either activated at time 2 (i.e., $t_{p+1} = 2$) or not (i.e., $t_{p+1} = \infty$).

Now, we define the hypothesis class $\mathcal{F}$ as the set of all combinations of $k$ nodes from $p$ possible parent nodes, that is $|\mathcal{F}| := \binom{p}{k}$. Thus, a hypothesis $\pi$ is the set of $k$ parent nodes such that $\forall i \in \pi$, there exist an edge from $i$ to $p+1$ with influence probability $\theta$. We also let $\pi^c := \{1, \ldots, p\} \backslash \pi$ to be the complement set of $\pi$. Given a hypothesis $\pi$ and a sample $\boldsymbol{t}$, we can write a data likelihood using independence assumptions.

$$\mathbb{P}(\boldsymbol{t}; \pi) = \mathbb{P}(\boldsymbol{t}_\pi)\mathbb{P}(\boldsymbol{t}_{\pi^c})\mathbb{P}(t_{p+1}|\boldsymbol{t}_\pi t_{s_2}) \tag{1}$$

The conditional probability can be expressed as follows.

$$\mathbb{P}(t_{p+1} = 2|\boldsymbol{t}_\pi t_{s_2}) = 1 - (1-\theta)^{\sum_{i\in\pi} \mathbb{1}[t_i=1]}(1-\theta_0)$$
$$\mathbb{P}(t_{p+1} = \infty|\boldsymbol{t}_\pi t_{s_2}) = (1-\theta)^{\sum_{i\in\pi} \mathbb{1}[t_i=1]}(1-\theta_0)$$

where $\mathbb{1}[\cdot]$ is an indicator function. Lastly, for simplicity, we define,

$$\theta := 1 - \theta_0^{\frac{1}{k}} \tag{2}$$

which decreases as the child node $p+1$ has more parents. The latter agrees with the intuition that as we have more parents, the chance of a single parent activating the child node gets smaller.

We will study the information-theoretic lower bounds on the sample complexity of the network inference problem. We will use Fano's inequality in order to analyze the necessary number of samples for any conceivable algorithm in order to avoid failure.

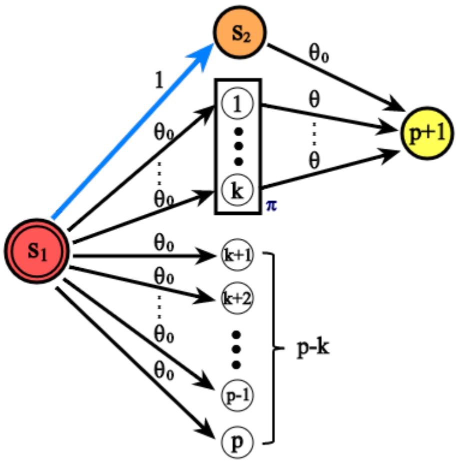

Fig. 1. Diffusion Model with Two Layers.

### B. Lower Bounds with Fano's inequality

First, we will bound the mutual information by using a pairwise Kullback-Leibler (KL) divergence-based bound [16], and show the following lemma.

**Lemma 1.** *Under the settings of the discrete-time diffusion model, for any pair of hypotheses, $\pi, \pi' \in \mathcal{F}$,*

$$\mathbb{KL}(\mathcal{P}_{\boldsymbol{t}|\pi}||\mathcal{P}_{\boldsymbol{t}|\pi'}) \leq \log \frac{1}{\theta_0}$$

*Proof.* First, we notice that the maximum KL divergence between two distributions, $\mathcal{P}_{\boldsymbol{t}|\pi}$ and $\mathcal{P}_{\boldsymbol{t}|\pi'}$ can be achieved when the two sets, $\pi$ and $\pi'$, do not share any node, or equivalently, when there is not any overlapping edge between parent and child nodes. That is, $\pi \cap \pi' = \emptyset$.

Then we compute the KL divergence with the two disjoint parent sets, as follows

$$\mathbb{KL}(\mathcal{P}_{\boldsymbol{t}|\pi}||\mathcal{P}_{\boldsymbol{t}|\pi'}) = \sum_{\boldsymbol{t}\in\{1,\infty\}^p\times\{2,\infty\}} \mathbb{P}(\boldsymbol{t}|\pi) \log \frac{\mathbb{P}(\boldsymbol{t}|\pi)}{\mathbb{P}(\boldsymbol{t}|\pi')}$$

Using Jensen's inequality and Eq (1), we have

$$\begin{aligned}
\mathbb{KL}(\mathcal{P}_{\boldsymbol{t}|\pi}||\mathcal{P}_{\boldsymbol{t}|\pi'}) &\leq \log \left( \sum_{\boldsymbol{t}\in\{1,\infty\}^p\times\{2,\infty\}} \mathbb{P}(\boldsymbol{t}|\pi) \frac{\mathbb{P}(\boldsymbol{t}|\pi)}{\mathbb{P}(\boldsymbol{t}|\pi')} \right) \\
&\leq \log \left( \max_{\boldsymbol{t}\in\{1,\infty\}^p\times\{2,\infty\}} \frac{\mathbb{P}(\boldsymbol{t}|\pi)}{\mathbb{P}(\boldsymbol{t}|\pi')} \right) \\
&= \log \left( \max_{\boldsymbol{t}\in\{1,\infty\}^p\times\{2,\infty\}} \frac{\mathbb{P}(\boldsymbol{t}_\pi)\mathbb{P}(\boldsymbol{t}_{\pi^c})\mathbb{P}(t_{p+1}|\boldsymbol{t}_\pi t_{s_2})}{\mathbb{P}(\boldsymbol{t}_{\pi'})\mathbb{P}(\boldsymbol{t}_{\pi'^c})\mathbb{P}(t_{p+1}|\boldsymbol{t}_{\pi'} t_{s_2})} \right) \\
&= \log \left( \max_{\boldsymbol{t}\in\{1,\infty\}^p\times\{2,\infty\}} \frac{\mathbb{P}(t_{p+1}|\boldsymbol{t}_\pi t_{s_2})}{\mathbb{P}(t_{p+1}|\boldsymbol{t}_{\pi'} t_{s_2})} \right) \quad (3)
\end{aligned}$$

Now as we have argued earlier, the maximum value can be attained when $\pi \cap \pi' = \emptyset$. Without loss of generality, we assume that $\pi$ connects the first $k$ nodes to $p+1$ and $\pi'$ connects the subsequent $k$ nodes to $p+1$. Thus we have

$$\frac{\mathbb{P}(t_{p+1}=2|\boldsymbol{t}_{\pi}t_{s_2})}{\mathbb{P}(t_{p+1}=2|\boldsymbol{t}_{\pi'}t_{s_2})} \le \frac{1-(1-\theta)^{\sum_{i=1}^{k}\mathbb{1}[t_i=1]}(1-\theta_0)}{1-(1-\theta)^{\sum_{i=k+1}^{2k}\mathbb{1}[t_i=1]}(1-\theta_0)}$$

Similarly, we have

$$\frac{\mathbb{P}(t_{p+1}=\infty|\boldsymbol{t}_{\pi}t_{s_2})}{\mathbb{P}(t_{p+1}=\infty|\boldsymbol{t}_{\pi'}t_{s_2})} \le \frac{(1-\theta)^{\sum_{i=1}^{k}\mathbb{1}[t_i=1]}(1-\theta_0)}{(1-\theta)^{\sum_{i=k+1}^{2k}\mathbb{1}[t_i=1]}(1-\theta_0)}$$

We can use the above expressions in order to obtain an upper bound for Eq (3). Thus, by Eq (2) we have

$$\begin{aligned}&\mathbb{KL}(\mathcal{P}_{\boldsymbol{t}|\pi}||\mathcal{P}_{\boldsymbol{t}|\pi'})\\ &\le \log\left(\max\left\{\frac{1-(1-\theta)^k(1-\theta_0)}{\theta_0}, \frac{1-\theta_0}{(1-\theta)^k(1-\theta_0)}\right\}\right)\\ &\le \log\left(\frac{1}{\theta_0}\right)\end{aligned}$$

□

By using the above results, we show that the necessary number of samples for the network inference problem is $\Omega(k\log p)$.

**Theorem 2.** *Suppose that nature picks a "true" hypothesis $\bar{\pi}$ uniformly at random from some distribution of hypotheses with support $\mathcal{F}$. Then a dataset $S$ of $n$ independent samples $\boldsymbol{t}^{(1)}, \boldsymbol{t}^{(2)}, \ldots, \boldsymbol{t}^{(n)} \in \{1,\infty\}^p \times \{2,\infty\}$ is produced, conditioned on the choice of $\bar{\pi}$. The learner then infers $\hat{\pi}$ from the dataset $S$. Under the settings of the two-layered discrete-time diffusion model, there exists a network inference problem of $k$ direct parent nodes such that if $n \le \frac{k\log p - k\log k - 2\log 2}{2\log\frac{1}{\theta_0}}$, then learning fails with probability at least $1/2$, i.e.,*

$$\mathbb{P}[\hat{\pi} \neq \bar{\pi}] \ge \frac{1}{2}$$

*for any algorithm that a learner could use for picking $\hat{\pi}$.*

*Proof.* We first bound the mutual information by the pairwise KL-based bound [16].

$$\begin{aligned}\mathbb{I}(\bar{\pi}, S) &< \frac{1}{|\mathcal{F}|^2}\sum_{\pi\in\mathcal{F}}\sum_{\pi'\in\mathcal{F}}\mathbb{KL}(\mathcal{P}_{S|\pi}||\mathcal{P}_{S|\pi'})\\ &= \frac{n}{|\mathcal{F}|^2}\sum_{\pi\in\mathcal{F}}\sum_{\pi'\in\mathcal{F}}\mathbb{KL}(\mathcal{P}_{\boldsymbol{t}|\pi}||\mathcal{P}_{\boldsymbol{t}|\pi'})\end{aligned}$$

Now from Lemma 1, we can bound the mutual information as follows.

$$\mathbb{I}(\bar{\pi}, S) < n\log\frac{1}{\theta_0} \qquad (4)$$

Finally, by Fano's inequality [17], Eq (4), and the well-known bound, $\log\binom{p}{k} \ge k(\log p - \log k)$, we have

$$\begin{aligned}\mathbb{P}[\hat{f} \neq \bar{f}] &\ge 1 - \frac{n\log\frac{1}{\theta_0} + \log 2}{\log\binom{p}{k}}\\ &\ge 1 - \frac{n\log\frac{1}{\theta_0} + \log 2}{k(\log p - \log k)}\\ &= \frac{1}{2}\end{aligned}$$

By solving the last equality we conclude that, if $n \le \frac{k\log p - k\log k - 2\log 2}{2\log\frac{1}{\theta_0}}$, then any conceivable algorithm will fail with a large probability, $\mathbb{P}[\hat{\pi} \neq \bar{\pi}] \ge 1/2$. □

## III. Ensemble of Continuous-time Diffusion Networks

In this section, we will study the continuous-time extension to the two-layer diffusion model. For this purpose, we introduce a transmission function between parent and child nodes. For the interested readers, Gomez-Rodriguez et al. [10] discuss transmission functions in full detail.

### A. A simple two-layer network

Here we used the same two-layer network structure shown in Figure 1. However, for a general continuous model, the activation time for a child node is dependent on the activation times of its parents. For our analysis, we relax this assumption by considering a fixed time range for each layer. In other words, we first consider a fixed time span, $T$. Then the $p$ parent nodes are only activated between $[0, T]$, and the child node $p+1$ is only activated between $[T, 2T]$. Our analysis for the continuous-time model largely borrows from our understanding of the discrete-time model.

The continuous-time model works as follows. The super source node $s_1$, tries to activate each of the $p$ parent nodes with probability $\theta_0$, and $s_2$ with probability 1. If a parent node gets activated, it picks an activation time from $[0, T]$ based on the transmission function, $f(t;\pi)$. Then, $s_2$ and all the direct parents, which have been activated in $t \in [0, T]$, independently try to activate the child node $p+1$ with probability $\theta_0$ and $\theta$, respectively. If the child node $p+1$ gets activated, it picks an activation time from $[T, 2T]$ based on the transmission function, $f(t;\pi)$.

For the continuous-time model, the conditional probabilities can be expressed as follows.

$$\begin{aligned}&\mathbb{P}(t_{p+1}\in[T,2T]|\boldsymbol{t}_{\pi}t_{s_2}) =\\ &\qquad\left(1-(1-\theta)^{\sum_{i\in\pi}\mathbb{1}[t_i\in[0,T]]}(1-\theta_0)\right)\cdot f(t_{p+1}-T;\pi)\\ &\mathbb{P}(t_{p+1}=\infty|\boldsymbol{t}_{\pi}t_{s_2}) = (1-\theta)^{\sum_{i\in\pi}\mathbb{1}[t_i\in[0,T]]}(1-\theta_0)\end{aligned}$$

Lastly, we define the domain of a sample $\boldsymbol{t}$ to be $\mathcal{T} := ([0,T]\cup\{\infty\})^p \times ([T,2T]\cup\{\infty\})$.

### B. Boundedness of Transmission Functions

We will start with the general boundedness of the transmission functions. The constants in the boundedness condition will be later directly related to the lower bound of the sample complexity. In the later part of the paper, we will provide an example for the exponentially distributed transmission function. Often, transmission functions used in the literature fulfill this assumption, e.g., the Rayleigh distribution [14] and the Weibull distribution for $\mu \ge 1$ [18].

**Condition 1** (Boundedness of transmission functions)**.** *Suppose $t \in [0, T]$ is a transmission time random variable, dependent on its parents $\pi$. The probability density function*

*$f(t;\pi)$ fulfills the following condition for a pair of positive constants $\kappa_1$ and $\kappa_2$.*

$$\min_{t\in[0,T]} f(t;\pi) \geq \kappa_1 > 0$$

$$\max_{t\in[0,T]} f(t;\pi) \leq \kappa_2 < \infty$$

### *C. Lower Bounds with Fano's inequality*

First, we provide a bound on the KL divergence that will be later used in analyzing the necessary number of samples for the network inference problem.

**Lemma 3.** *Under the settings of the continuous-time diffusion model, for any pair of hypotheses, $\pi, \pi' \in \mathcal{F}$,*

$$\mathbb{KL}(\mathcal{P}_{\boldsymbol{t}|\pi}||\mathcal{P}_{\boldsymbol{t}|\pi'}) \leq \log\left(\max\left\{\frac{\kappa_2}{\kappa_1}\left(\frac{1}{\theta_0}-(1-\theta_0)\right), \frac{1}{\theta_0}\right\}\right)$$

*Proof.* We note that the proof is very similar to that of Lemma 1.

$$\begin{aligned}\mathbb{KL}(\mathcal{P}_{\boldsymbol{t}|\pi}||\mathcal{P}_{\boldsymbol{t}|\pi'}) &= \sum_{\boldsymbol{t}\in\mathcal{T}} \mathbb{P}(\boldsymbol{t}|\pi)\log\frac{\mathbb{P}(\boldsymbol{t}|\pi)}{\mathbb{P}(\boldsymbol{t}|\pi')}\\ &\leq \log\left(\max_{\boldsymbol{t}\in\mathcal{T}}\frac{\mathbb{P}(\boldsymbol{t}|\pi)}{\mathbb{P}(\boldsymbol{t}|\pi')}\right)\\ &= \log\left(\max_{\boldsymbol{t}\in\mathcal{T}}\frac{\mathbb{P}(t_{p+1}|\boldsymbol{t}_{\pi}t_{s_2})}{\mathbb{P}(t_{p+1}|\boldsymbol{t}_{\pi'}t_{s_2})}\right) \qquad (5)\end{aligned}$$

Now with the same argument we made in Lemma 1, consider that $\pi$ connects the first $k$ nodes to $p+1$ and $\pi'$ connects the subsequent $k$ nodes to $p+1$. Thus, we have

$$\frac{\mathbb{P}(t_{p+1}\in[T,2T]|\boldsymbol{t}_{\pi}t_{s_2})}{\mathbb{P}(t_{p+1}\in[T,2T]|\boldsymbol{t}_{\pi'}t_{s_2})} \leq \frac{\left(1-(1-\theta)^{\sum_{i=1}^{k}\mathbb{1}[t_i\in[0,T]]}(1-\theta_0)\right)f(t_{p+1}-T;\pi)}{\left(1-(1-\theta)^{\sum_{i=k+1}^{2k}\mathbb{1}[t_i\in[0,T]]}(1-\theta_0)\right)f(t_{p+1}-T;\pi')}$$

Similarly, we have

$$\frac{\mathbb{P}(t_{p+1}=\infty|\boldsymbol{t}_{\pi}t_{s_2})}{\mathbb{P}(t_{p+1}=\infty|\boldsymbol{t}_{\pi'}t_{s_2})} \leq \frac{(1-\theta)^{\sum_{i=1}^{k}\mathbb{1}[t_i\in[0,T]]}(1-\theta_0)}{(1-\theta)^{\sum_{i=k+1}^{2k}\mathbb{1}[t_i\in[0,T]]}(1-\theta_0)}$$

We can use the above expressions in order to obtain an upper bound for Eq (5). Thus, by Eq (2) we have

$$\begin{aligned}&\mathbb{KL}(\mathcal{P}_{\boldsymbol{t}|\pi}||\mathcal{P}_{\boldsymbol{t}|\pi'})\\ &\leq \log\left(\max\left\{\frac{1-(1-\theta)^k(1-\theta_0)}{\theta_0}\frac{\kappa_2}{\kappa_1}, \frac{1-\theta_0}{(1-\theta)^k(1-\theta_0)}\right\}\right)\\ &= \log\left(\max\left\{\frac{\kappa_2}{\kappa_1}\left(\frac{1}{\theta_0}-(1-\theta_0)\right), \frac{1}{\theta_0}\right\}\right)\end{aligned}$$

□

By using the above results, we show that the necessary number of samples for the network inference problem is also $\Omega(k\log p)$ in the continuous-time model.

**Theorem 4.** *Suppose that nature picks a "true" hypothesis $\bar{\pi}$ uniformly at random from some distribution of hypotheses with support $\mathcal{F}$. Then a dataset $S$ of $n$ independent samples $\boldsymbol{t}^{(\boldsymbol{1})}, \boldsymbol{t}^{(\boldsymbol{2})}, \ldots, \boldsymbol{t}^{(\boldsymbol{n})} \in ([0,T]\cup\{\infty\})^p \times ([T,2T]\cup\{\infty\})$ is produced, conditioned on the choice of $\bar{\pi}$. The learner then infers $\hat{\pi}$ from the dataset $S$. Assume that the transmission function $f(t;\pi)$, satisfies Condition 1 with constants $\kappa_1$ and $\kappa_2$. Under the settings of the two-layered continuous-time diffusion model, there exists a network inference problem of $k$ direct parent nodes such that if*

$$n \leq \frac{k\log p - k\log k - 2\log 2}{2\log\left(\max\left\{\frac{\kappa_2}{\kappa_1}\left(\frac{1}{\theta_0}-(1-\theta_0)\right), \frac{1}{\theta_0}\right\}\right)}$$

*then learning fails with probability at least $1/2$, i.e.,*

$$\mathbb{P}[\hat{\pi}\neq\bar{\pi}] \geq \frac{1}{2}$$

*for any algorithm that a learner could use for picking $\hat{\pi}$.*

*Proof.* The proof is very similar to that of Theorem 2. First, by the pairwise KL-based bound [16] and Lemma 3, we have

$$\mathbb{I}(\bar{\pi}, S) < n\log\left(\max\left\{\frac{\kappa_2}{\kappa_1}\left(\frac{1}{\theta_0}-(1-\theta_0)\right), \frac{1}{\theta_0}\right\}\right) \qquad (6)$$

By Fano's inequality [17], Eq (6), and the well-known bound, $\log\binom{p}{k} \geq k(\log p - \log k)$, we have

$$\begin{aligned}&\mathbb{P}[\hat{f}\neq\bar{f}]\\ &\geq 1 - \frac{n\log\left(\max\left\{\frac{\kappa_2}{\kappa_1}\left(\frac{1}{\theta_0}-(1-\theta_0)\right), \frac{1}{\theta_0}\right\}\right)+\log 2}{\log\binom{p}{k}}\\ &\geq 1 - \frac{n\log\left(\max\left\{\frac{\kappa_2}{\kappa_1}\left(\frac{1}{\theta_0}-(1-\theta_0)\right), \frac{1}{\theta_0}\right\}\right)+\log 2}{k(\log p - \log k)}\\ &= \frac{1}{2}\end{aligned}$$

By solving the last equality we conclude that, if $n \leq \frac{k\log p - k\log k - 2\log 2}{2\log\left(\max\left\{\frac{\kappa_2}{\kappa_1}\left(\frac{1}{\theta_0}-(1-\theta_0)\right), \frac{1}{\theta_0}\right\}\right)}$, then any conceivable algorithm will fail with a large probability, $\mathbb{P}[\hat{\pi}\neq\bar{\pi}] \geq 1/2$. □

Lastly, we will present an example for the exponentially distributed transmission function.

**Corollary 5** (Exponential Distribution)**.** *Suppose that nature picks a "true" hypothesis $\bar{\pi}$ uniformly at random from some distribution of hypotheses with support $\mathcal{F}$. Then a dataset $S$ of $n$ independent samples $\boldsymbol{t}^{(\boldsymbol{1})}, \boldsymbol{t}^{(\boldsymbol{2})}, \ldots, \boldsymbol{t}^{(\boldsymbol{n})} \in ([0,T]\cup\{\infty\})^p \times ([T,2T]\cup\{\infty\})$ is produced, conditioned on the choice of $\bar{\pi}$. The learner then infers $\hat{\pi}$ from the dataset $S$. Assume that the transmission function $f(t;\pi) = \frac{\lambda e^{-\lambda t}}{1-e^{-\lambda T}}$ is of the censored (rescaled) exponential distribution form, defined over [0,T].*

*Under the settings of the two-layered continuous-time diffusion model, there exists a network inference problem of $k$ direct parent nodes such that if*

$$n \leq \frac{k \log p - k \log k - 2 \log 2}{2 \log \left( \max \left\{ e^{\lambda T} \left( \frac{1}{\theta_0} - (1 - \theta_0) \right), \frac{1}{\theta_0} \right\} \right)}$$

*then learning fails with probability at least $1/2$, i.e.,*

$$\mathbb{P}[\hat{\pi} \neq \bar{\pi}] \geq \frac{1}{2}$$

*for any algorithm that a learner could use for picking $\hat{\pi}$.*

*Proof.* Since the probability density function should only be defined between $[0, T]$, we need to rescale the probability density function of the standard exponential distribution, $g(t) \sim Exp(\lambda)$, whose cumulative density function is $G(t)$. Given this, we have the censored (rescaled) transmission function,

$$f(t; \pi) = \frac{g(t)}{G(T) - G(0)} = \frac{g(t)}{G(T)} = \frac{\lambda e^{-\lambda t}}{1 - e^{-\lambda T}}$$

From the above, we can obtain the minimum and maximum values of the density function, $\kappa_1$ and $\kappa_2$, in Condition 1 as follows.

$$\kappa_1 = \frac{\lambda e^{-\lambda T}}{1 - e^{-\lambda T}} \quad , \quad \kappa_2 = \frac{\lambda}{1 - e^{-\lambda T}} \quad \Rightarrow \quad \frac{\kappa_2}{\kappa_1} = e^{\lambda T} \tag{7}$$

Finally using Theorem 4 and Eq (7), we show that if

$$n \leq \frac{k \log p - k \log k - 2 \log 2}{2 \log \left( \max \left\{ e^{\lambda T} \left( \frac{1}{\theta_0} - (1 - \theta_0) \right), \frac{1}{\theta_0} \right\} \right)}$$

then any conceivable algorithm will fail with a large probability, $\mathbb{P}[\hat{\pi} \neq \bar{\pi}] \geq 1/2$. □

## IV. Conclusion

We have formulated the two-layered discrete-time and continuous-time diffusion models and derived the information-theoretic lower bounds of the sample complexity of order $\Omega(k \log p)$. Our bound is particularly important since we can infer that the algorithm in [1], which only works under discrete-time settings, is statistically optimal based on our bound.

Our work opens the question of whether it is possible to devise an algorithm for which the sufficient number of samples is $\mathcal{O}(k \log p)$ in continuous-time settings. We also have observed some potential future work to analyze sharp phase transitions for the sample complexity of the network inference problem.

## References

[1] J. Pouget-Abadie and T. Horel, "Inferring graphs from cascades: A sparse recovery framework," in *Proceedings of the 24th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2015, pp. 625–626.

[2] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 57–66.

[3] D. Kempe, J. Kleinberg, and É. 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*. ACM, 2003, pp. 137–146.

[4] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *Proceedings of the 12th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, Part III*, ser. KES '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 67–75. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-85567-5_9

[5] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 241–250.

[6] E. Adar and L. Adamic, "Tracking information epidemics in blogspace," in *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, Sept 2005, pp. 207–214.

[7] S. Myers and J. Leskovec, "On the convexity of latent social network inference," in *Advances in Neural Information Processing Systems*, 2010, pp. 1741–1749.

[8] M. Gomez Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 1019–1028.

[9] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Learning continuous-time information diffusion model for social behavioral data analysis," in *Advances in Machine Learning*. Springer, 2009, pp. 322–337.

[10] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," in *ICML '11: Proceedings of the 28th International Conference on Machine Learning*, 2011.

[11] N. Du, L. Song, M. Gomez-Rodriguez, and H. Zha, "Scalable influence estimation in continuous-time diffusion networks," in *NIPS '13: Advances in Neural Information Processing Systems*, 2013.

[12] P. Netrapalli and S. Sanghavi, "Learning the graph of epidemic cascades," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, no. 1. ACM, 2012, pp. 211–222.

[13] B. Abrahao, F. Chierichetti, R. Kleinberg, and A. Panconesi, "Trace complexity of network inference," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 491–499.

[14] H. Daneshmand, M. Gomez-Rodriguez, L. Song, and B. Schoelkopf, "Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm," in *Proceedings of the... International Conference on Machine Learning. International Conference on Machine Learning*, vol. 2014. NIH Public Access, 2014, p. 793.

[15] H. Narasimhan, D. C. Parkes, and Y. Singer, "Learnability of influence in networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 3168–3176.

[16] Y. B., "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, T. E. Pollard D. and Y. G., Eds. Springer New York, 1997, pp. 423–435.

[17] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.

[18] T. Kurashima, T. Iwata, N. Takaya, and H. Sawada, "Probabilistic latent network visualization: inferring and embedding diffusion networks," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1236–1245.