Journal of Machine Learning Research 17 (2016) 1-29 Submitted 10/14; Revised 4/15; Published 6/16

# Estimating Diffusion Networks: Recovery Conditions, Sample Complexity & Soft-thresholding Algorithm

**Manuel Gomez-Rodriguez**[◇†] MANUELGR@MPI-SWS.ORG
*MPI for Software Systems*
*Paul-Ehrlich-Strasse, 67663 Kaiserslautern*
*Germany*

**Le Song**[◇] LSONG@CC.GATECH.EDU
*College of Computing*
*Georgia Institute of Technology*
*Atlanta, GA 30332, USA*

**Hadi Daneshmand** SEYED.DANESHMAND@INF.ETHZ.CH
*Computer Science Department*
*Universitätstrasse 6, 8092 Zürich*
*Switzerland*

**Bernhard Schölkopf** BS@TUE.MPG.DE
*MPI for Intelligent Systems*
*Spemannstrasse 38, 72076 Tübingen*
*Germany*

**Editor:** Edo Airoldi

## Abstract

Information spreads across social and technological networks, but often the network structures are hidden from us and we only observe the traces left by the diffusion processes, called *cascades*. Can we recover the hidden network structures from these observed cascades? What kind of cascades and how many cascades do we need? Are there some network structures which are more difficult than others to recover? Can we design efficient inference algorithms with provable guarantees?

Despite the increasing availability of cascade data and methods for inferring networks from these data, a thorough theoretical understanding of the above questions remains largely unexplored in the literature. In this paper, we investigate the network structure inference problem for a general family of continuous-time diffusion models using an $\ell_1$-regularized likelihood maximization framework. We show that, as long as the cascade sampling process satisfies a natural incoherence condition, our framework can recover the correct network structure with high probability if we observe $O(d^3 \log N)$ cascades, where $d$ is the maximum number of parents of a node and $N$ is the total number of nodes. Moreover, we develop a simple and efficient soft-thresholding network inference algorithm

. [◇] These authors contributed equally to this work.
. [†] This work was done while Manuel Gomez-Rodriguez was still affiliated with the MPI for Intelligent Systems, Spemannstr. 38, 72076 Tuebingen, Germany
. Preliminary version of this work appeared in proceedings of the 31st International Conference on Machine Learning (ICML '14), 2014.

which demonstrate the match between our theoretical prediction and empirical results. In practice, this new algorithm also outperforms other alternatives in terms of the accuracy of recovering hidden diffusion networks.

## 1. Introduction

Diffusion of information, behaviors, diseases, or irrepresentabilitymore generally, *contagions* can be naturally modeled as a stochastic process that occur over the edges of an underlying network (Rogers, 1995). In this scenario, we often observe the temporal traces that the diffusion generates, called *cascades*, but the edges of the network that gave rise to the diffusion remain unobservable (Adar and Adamic, 2005). For example, blogs or media sites often publish a new piece of information without explicitly citing their sources. Marketers may note when a social media user decides to adopt a new behavior but cannot tell which neighbor in the social network influenced them to do so. Epidemiologist observe when a person gets sick but usually cannot tell who infected her. In all these cases, given a set of cascades and a diffusion model, the network inference problem consists of inferring the edges (and model parameters) of the unobserved underlying network (Gomez-Rodriguez, 2013).

The network inference problem has attracted significant attention in recent years (Saito et al., 2009; Gomez-Rodriguez et al., 2010, 2011, 2013b, 2014; Snowsill et al., 2011; Du et al., 2012a, 2013; Zhou et al., 2013), since it is essential to reconstruct and predict the paths over which information can spread, and to maximize sales of a product or stop infections. Most previous work has focused on developing network inference algorithms and evaluating their performance experimentally on different synthetic and real networks, and a rigorous theoretical analysis of the problem has been missing. However, such analysis is of outstanding interest since it would enable us to answer many fundamental open questions. For example, which conditions are sufficient to guarantee that we can recover a network given a large number of cascades? If these conditions are satisfied, how many cascades are sufficient to infer the network with high probability? Until recently, there has been only two pieces of work along this direction (Netrapalli and Sanghavi, 2012; Abrahao et al., 2013), which leverage less realistic diffusion models. Moreover, none of them is able to identify a recovery condition relating the interaction between the network structure and the cascade sampling process, which we make precise in our paper.

### 1.1 Overview of results

We consider the network inference problem under the continuous-time diffusion model recently introduced by Gomez-Rodriguez et al. (2011), which has been extensively validated in real diffusion data, and, due to its flexibility, has been extended to support textual information (Wang et al., 2012), nonparametric pairwise likelihoods (Du et al., 2012a), topic modeling (Du et al., 2012b), dynamic networks (Gomez-Rodriguez et al., 2013a). We identify a natural irrepresentability condition for such a model which depends on both the network structure, the diffusion parameters and the sampling process of the cascades. This condition captures the intuition that we can recover the network structure if the co-occurrence of a node and its non-parent nodes is small in the cascades. Furthermore, we show that, if this condition holds for the population case, we can recover the network structure using

an $\ell_1$-regularized maximum likelihood estimator and $O(d^3 \log N)$ cascades, where $N$ is the number of nodes in the network and $d$ is the maximum number of parents of a node, with the probability of success approaching 1 in a rate exponential in the number of cascades. Importantly, if this condition also holds for the finite sample case, then the guarantee can be improved to $O(d^2 \log N)$ cascades. Beyond theoretical results, we also propose a new, efficient and simple proximal gradient algorithm to solve the $\ell_1$-regularized maximum likelihood estimation. The algorithm is especially well-suited for our problem since it is highly scalable and naturally finds sparse estimators, as desired, by using soft-thresholding. Using this algorithm, we perform various experiments illustrating the consequences of our theoretical results and demonstrating that it typically outperforms other state-of-the-art algorithms.

### 1.2 Related work

Netrapalli and Sanghavi (2012) propose a maximum likelihood network inference method for a variation of the discrete-time independent cascade model (Kempe et al., 2003) and show that, for general networks satisfying a *correlation decay*, the estimator recovers the network structure given $O(d^2 \log N)$ cascades, and the probability of success is approaching 1 in a rate exponential in the number of cascades. The rate they obtained is on a par with our results. However, their discrete diffusion model is less realistic in practice, and the correlation decay condition implies that, on average, each node can only infect one single node per cascade. Instead, we use a general continuous-time diffusion model (Gomez-Rodriguez et al., 2011), which has been extensively validated in real diffusion data and extended in various ways by different authors (Wang et al., 2012; Du et al., 2012a,b).

Abrahao et al. (2013) propose a simple network inference method, First-Edge, for a slightly different continuous-time independent cascade model (Gomez-Rodriguez et al., 2010), and show that, for general networks, if the cascade sources are chosen uniformly at random, the algorithm needs $O(Nd \log N)$ cascades to recover the network structure and the probability of success is approaching 1 in a rate polynomial in the number of cascades. Additionally, they study trees and bounded-degree networks and show that, if the cascade sources are chosen uniformly at random, the error decreases polynomially as long as $O(\log N)$ and $\Omega(d^9 \log^2 d \log N)$ cascades are recorded respectively. In our work, we show that, for general networks satisfying a natural irrepresentability condition, our method outperforms the First-Edge algorithm and the algorithm for bounded-degree networks in terms of rate and sample complexity.

Gripon and Rabbat (2013) propose a network inference method for unordered cascades, in which nodes that are infected together in the same cascade are connected by a path containing exactly the nodes in the trace, and give necessary and sufficient conditions for network inference. However, they consider a restrictive scenario in which cascades are all three nodes long.

## 2. Continuous-Time Diffusion Model

In this section, we revisit the continuous-time generative model for cascade data introduced by Gomez-Rodriguez et al. (2011). The model associates each edge $j \to i$ with a transmission function, $f(t_i|t_j; \alpha_{ji}) = f(t_i - t_j; \alpha_{ji})$, a density over time parameterized by $\alpha_{ji}$. This is

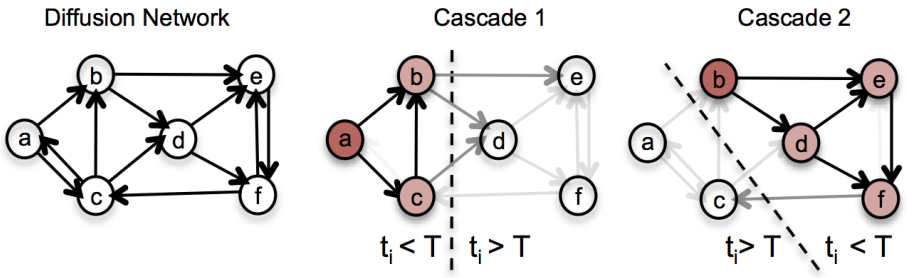



Figure 1: The diffusion network structure (left) is unknown and we only observe cascades, which are $N$-dimensional vectors recording the times when nodes get infected by contagions that spread (right). Cascade 1 is $(t_a, t_b, t_c, \infty, \infty, \infty)$, where $t_a < t_c < t_b$, and cascade 2 is $(\infty, t_b, \infty, t_d, t_e, t_f)$, where $t_b < t_d < t_e < t_f$. Each cascade contains a source node (dark red), drawn from a source distribution $\mathbb{P}(s)$, as well as infected (light red) and uninfected (white) nodes, and it provides information on black and dark gray edges but does not on light gray edges.

in contrast to previous discrete-time models which associate each edge with a fixed infection probability (Kempe et al., 2003). Moreover, it also differs from discrete-time models in the sense that events in a cascade are not generated iteratively in rounds, but event timings are sampled directly from the transmission functions in the continuous-time model.

### 2.1 Cascade generative process

Given a *directed* contact network, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $N$ nodes, the process begins with an infected source node, $s$, initially adopting certain *contagion* (idea, meme or product) at time zero, which we draw from a source distribution $\mathbb{P}(s)$. The contagion is transmitted from the source along her out-going edges to her direct neighbors. Each transmission through an edge entails a *random* transmission time, $\tau$, drawn from an associated pairwise transmission likelihood $f(\tau; \alpha_{ji})$. We assume transmission times are independent, possibly distributed differently across edges, and, in some cases, can be arbitrarily large, $\tau \to \infty$. Then, the infected neighbors transmit the contagion to their respective neighbors, and the process continues. We assume that an infected node remains infected for the entire diffusion process. Thus, if a node $i$ is infected by multiple neighbors, only the neighbor that first infects node $i$ will be the *true parent*. As a result, although the contact network can be an arbitrary directed network, each contagion induces a Directed Acyclic Graph (DAG). Figure 1 illustrates the process and Table 1 gives several examples of well-known parametric transmission likelihoods (Gomez-Rodriguez et al., 2011, 2013a, 2014).

### 2.2 Cascade data

Observations from the model are recorded as a set $C^n$ of cascades $\{\mathbf{t}^1, \ldots, \mathbf{t}^n\}$. Each cascade $\mathbf{t}^c$ is an $N$-dimensional vector $\mathbf{t}^c := (t_1^c, \ldots, t_N^c)$ recording when nodes are infected, $t_k^c \in [0, T^c] \cup \{\infty\}$. Symbol $\infty$ labels nodes that are not infected during observation window $[0, T^c]$ – it does not imply they are never infected. The 'clock' is reset to 0 at the start of each

| Model | Transmission functions $f(\tau;\alpha_{ji})$ | Log survival $y(t+\tau\|t;\alpha_{ji}) = \log S(t+\tau\|t;\alpha_{ji})$ | Hazard $H(t+\tau\|t;\alpha_{ji})$ |
|---|---|---|---|
| Exp | $\begin{cases} \alpha_{j,i} \cdot e^{-\alpha_{j,i}\tau} & \text{if } \tau \geq 0 \\ 0 & \text{otherwise} \end{cases}$ | $-\alpha_{j,i}\tau$ | $\alpha_{j,i}$ |
| Pow | $\begin{cases} \frac{\alpha_{j,i}}{\delta}\left(\frac{\tau}{\delta}\right)^{-1-\alpha_{j,i}} & \text{if } \tau \geq \delta \\ 0 & \text{otherwise} \end{cases}$ | $-\alpha_{j,i}\log\left(\frac{\tau}{\delta}\right)$ | $\frac{\alpha_{j,i}}{\tau}$ |
| Ray | $\begin{cases} \alpha_{j,i}\tau e^{-\frac{1}{2}\alpha_{j,i}\tau^2} & \text{if } \tau \geq 0 \\ 0 & \text{otherwise} \end{cases}$ | $-\alpha_{j,i}\frac{\tau^2}{2}$ | $\alpha_{j,i}\tau$ |

Table 1: Pairwise transmission models

cascade. We assume $T^c = T$ for all cascades; the results generalize trivially. Contagions often propagate simultaneously (Myers and Leskovec, 2012; Prakash et al., 2012) over the same network but we assume each contagion to propagate independently of each other. Finally, we also assume that all activated nodes except the first one are activated by network diffusion, *i.e.*, by previously activated nodes, ignoring external influences (Myers et al., 2012). Refer to Figure 1 for an example.

### 2.3 Likelihood of a cascade

Gomez-Rodriguez et al. (2011) showed that the likelihood of a cascade $\mathbf{t}$ under the continuous-time independent cascade model is

$$f(\mathbf{t};\mathbf{A}) = \prod_{t_i \leq T} \prod_{t_m > T} S(T|t_i;\alpha_{im}) \times \prod_{k:t_k<t_i} S(t_i|t_k;\alpha_{ki}) \sum_{j:t_j<t_i} H(t_i|t_j;\alpha_{ji}), \tag{1}$$

where $\mathbf{A} = \{\alpha_{ji}\}$ denotes the collection of parameters, $S(t_i|t_j;\alpha_{ji}) = 1 - \int_{t_j}^{t_i} f(t-t_j;\alpha_{ji})\,dt$ is the survival function and $H(t_i|t_j;\alpha_{ji}) = f(t_i - t_j;\alpha_{ji})/S(t_i|t_j;\alpha_{ji})$ is the hazard function. The survival terms in the first line account for the probability that uninfected nodes survive to all infected nodes in the cascade up to $T$ and the survival and hazard terms in the second line account for the likelihood of the infected nodes. The survival and hazard functions are simple for several well-known parametric transmission likelihoods, as shown in Table 1. Then, assuming cascades are sampled independently, the likelihood of a set of cascades is the product of the likelihoods of individual cascades given by Eq. 1. For notational simplicity, we define $y(t_i|t_k;\alpha_{ki}) := \log S(t_i|t_k;\alpha_{ki})$, and $h(\boldsymbol{t};\boldsymbol{\alpha}_i) := \sum_{k:t_k\leq t_i} H(t_i|t_k;\alpha_{ki})$ if $t_i \leq T$ and 0 otherwise.

## 3. Network Inference Problem

Consider an instance of the continuous-time diffusion model defined above with a contact network $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$ and associated parameters $\left\{\alpha^*_{ji}\right\}$. We denote the set of parents of node $i$ as $\mathcal{N}^-(i) = \{j \in \mathcal{V}^* : \alpha^*_{ji} > 0\}$ with cardinality $d_i = |\mathcal{N}^-(i)|$ and the minimum positive transmission rate as $\alpha^*_{\min,i} = \min_{j\,:\,\alpha^*_{ji}>0} \alpha^*_{ji}$. Let $C^n$ be a set of $n$ cascades sampled from the model, where the source $s \in \mathcal{V}^*$ of each cascade is drawn from a source distribution

| **Function** | **Infected node** ($t_i < T$) | **Uninfected node** ($t_i > T$) |
|---|---|---|
| $g_i(\mathbf{t};\boldsymbol{\alpha})$ | $\log h(\mathbf{t};\boldsymbol{\alpha}) + \sum_{j:t_j<t_i} y(t_i\|t_j;\alpha_j)$ | $\sum_{j:t_j<T} y(T\|t_j;\alpha_j)$ |
| $[\nabla y_i(\mathbf{t};\boldsymbol{\alpha})]_k$ | $-y'(t_i\|t_k;\alpha_k)$ | $-y'(T\|t_k;\alpha_k)$ |
| $[D(\mathbf{t};\boldsymbol{\alpha})]_{kk}$ | $-y''(t_i\|t_k;\alpha_k) - h(\mathbf{t};\boldsymbol{\alpha})^{-1}H''(t_i\|t_k;\alpha_k)$ | $-y''(T\|t_k;\alpha_k)$ |

Table 2: Functions. $g_i(\mathbf{t};\boldsymbol{\alpha})$ is node $i$'s log-likelihood in a cascade $\mathbf{t}$, $y_i(\mathbf{t};\boldsymbol{\alpha})$ is the logarithm of node $i$'s survivals in a cascade $\mathbf{t}$, $D(\mathbf{t};\boldsymbol{\alpha})$ is a diagonal matrix defined in Eq. 5, $H(t_i|t_k;\alpha_k)$ is the hazard function, and $h(\mathbf{t};\boldsymbol{\alpha})$ denotes the sum of node $i$'s hazard functions in a cascade $\mathbf{t}$.

$\mathbb{P}(s)$. Then, the network inference problem consists of finding the directed edges and the associated parameters using only the temporal information from the set of cascades $C^n$.

This problem has been cast as a maximum likelihood estimation problem (Gomez-Rodriguez et al., 2011)

$$\begin{array}{ll} \text{minimize}_{\mathbf{A}} & -\frac{1}{n}\sum_{c\in C^n} \log f(\mathbf{t}^c;\mathbf{A}) \\ \text{subject to} & \alpha_{ji} \geq 0,\, i,j = 1,\ldots,N, i \neq j, \end{array} \tag{2}$$

where the inferred edges in the network correspond to those pairs of nodes with non-zero parameters, *i.e.* $\hat{\alpha}_{ji} > 0$.

In fact, the problem in Eq. 2 decouples into a set of independent smaller subproblems, one per node, where we infer the parents of each node and the parameters associated with these incoming edges. Without loss of generality, for a particular node $i$, we solve the problem

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{\alpha}_i} & \ell^n(\boldsymbol{\alpha}_i) \\ \text{subject to} & \alpha_{ji} \geq 0,\, j = 1,\ldots,N, i \neq j, \end{array} \tag{3}$$

where the parameters $\boldsymbol{\alpha}_i := \{\alpha_{ji} \,|\, j = 1,\ldots,N, i \neq j\}$ are the relevant variables, and $\ell^n(\boldsymbol{\alpha}_i) = -\frac{1}{n}\sum_{c\in C^n} g_i(\mathbf{t}^c;\boldsymbol{\alpha}_i)$ corresponds to the terms in Eq. 2 involving $\boldsymbol{\alpha}_i$. The function $g(\,\cdot\,;\boldsymbol{\alpha}_i)$ is simple for several well-known parametric transmission likelihoods, including those described in Table 1. For example, for an exponential transmission likelihood,

$$g_i(\mathbf{t};\boldsymbol{\alpha}_i) = \log\left(\sum_{j:t_j<t_i} \alpha_{ji}\right) - \sum_{j:t_j<t_i} \alpha_{ji}(t_i - t_j)$$

for an infected node and $g_i(\mathbf{t};\boldsymbol{\alpha}_i) = -\sum_{j:t_j<T} \alpha_{ji}(T - t_j)$ for an uninfected node. Refer to Table 2 for a general definition of $g(\,\cdot\,;\boldsymbol{\alpha}_i)$. Moreover, in this subproblem, we only need to consider a super-neighborhood $\mathcal{V}_i = \mathcal{R}_i \cup \mathcal{U}_i$ of $i$, with cardinality $p_i = |\mathcal{V}_i| \leq N$, where $\mathcal{R}_i$ is the set of upstream nodes from which $i$ is reachable, $\mathcal{U}_i$ is the set of nodes which are reachable from at least one node $j \in \mathcal{R}_i$. Here, we consider a node $i$ to be reachable from a node $j$ if and only if there is a directed path from $j$ to $i$. We can skip all nodes in $\mathcal{V}\backslash\mathcal{V}_i$ from our analysis because they will never be infected in a cascade before $i$, and thus, the maximum likelihood estimation of the associated transmission rates will always be zero (and correct).

Below, we show that, as $n \to \infty$, the solution, $\hat{\boldsymbol{\alpha}}_i$, of the problem in Eq. 3 is a consistent estimator of the true parameter $\boldsymbol{\alpha}_i^*$. However, it is not clear whether it is possible to

recover the true network structure with this approach given a finite amount of cascades and, if so, how many cascades are needed. We will show that by adding an $\ell_1$-regularizer to the objective function and solving instead the following optimization problem

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{\alpha}_i} & \ell^n(\boldsymbol{\alpha}_i) + \lambda_n ||\boldsymbol{\alpha}_i||_1 \\ \text{subject to} & \alpha_{ji} \geq 0,\, j = 1, \ldots, N, i \neq j, \end{array} \tag{4}$$

we can provide finite sample guarantees for recovering the network structure (and parameters). Our analysis also shows that by selecting an appropriate value for the regularization parameter $\lambda_n$, the solution of Eq. 4 successfully recovers the network structure with probability approaching 1 exponentially fast in $n$.

In the remainder of the paper, we will focus on estimating the parent nodes of a particular node $i$. For simplicity, we will use $\boldsymbol{\alpha} = \boldsymbol{\alpha}_i$, $\alpha_j = \alpha_{ji}$, $\mathcal{N}^- = \mathcal{N}^-(i)$, $\mathcal{R} = \mathcal{R}_i$, $\mathcal{U} = \mathcal{U}_i$, $d = d_i$, $p_i = p$ and $\alpha^*_{\min} = \alpha^*_{\min,i}$.

## 4. Consistency

*Can we recover the hidden network structures from the observed cascades?* The answer is yes. We will show this by proving that the estimator provided by Eq. 3 is consistent, meaning that as the number of cascades goes to **infinity**, we can always recover the true network structure.

More specifically, Gomez-Rodriguez et al. (2011) showed that the network inference problem defined in Eq. 3 is convex in $\boldsymbol{\alpha}$ if the survival functions are log-concave and the hazard functions are concave in $\boldsymbol{\alpha}$. Under these conditions, the Hessian matrix, $\mathcal{Q}^n = \nabla^2 \ell^n(\boldsymbol{\alpha})$, can be expressed as the sum of a nonnegative diagonal matrix $\boldsymbol{D}^n$ and the outer product of a matrix $\boldsymbol{X}^n(\boldsymbol{\alpha})$ with itself, *i.e.*,

$$\mathcal{Q}^n = \boldsymbol{D}^n(\boldsymbol{\alpha}) + \tfrac{1}{n} \boldsymbol{X}^n(\boldsymbol{\alpha}) [\boldsymbol{X}^n(\boldsymbol{\alpha})]^\top. \tag{5}$$

Here the diagonal matrix $\boldsymbol{D}^n(\boldsymbol{\alpha}) = \frac{1}{n} \sum_c \boldsymbol{D}(\mathbf{t}^c; \boldsymbol{\alpha})$ is a sum over a set of diagonal matrices $\boldsymbol{D}(\mathbf{t}^c; \boldsymbol{\alpha})$, one for each cascade $c$ (see Table 2 for the definition of its entries); and $\boldsymbol{X}^n(\boldsymbol{\alpha})$ is the Hazard matrix

$$\boldsymbol{X}^n(\boldsymbol{\alpha}) = \left[ \boldsymbol{X}(\mathbf{t}^1; \boldsymbol{\alpha}) \,|\, \boldsymbol{X}(\mathbf{t}^2; \boldsymbol{\alpha}) \,|\, \ldots \,|\, \boldsymbol{X}(\mathbf{t}^n; \boldsymbol{\alpha}) \right], \tag{6}$$

with each column $\boldsymbol{X}(\mathbf{t}^c; \boldsymbol{\alpha}) := h(\boldsymbol{t}^c; \boldsymbol{\alpha})^{-1} \nabla_{\boldsymbol{\alpha}} h(\boldsymbol{t}^c; \boldsymbol{\alpha})$. Intuitively, the Hessian matrix captures the co-occurrence information of nodes in cascades. Both $\boldsymbol{D}(\mathbf{t}^c; \boldsymbol{\alpha})$ and $\boldsymbol{X}^n(\boldsymbol{\alpha})$ are simple for several well-known transmission likelihoods, including those described in Table 1. For example, for an exponential transmission likelihood, $[\boldsymbol{D}(\mathbf{t}^c; \boldsymbol{\alpha})]_{kk} = 0$ and $[\boldsymbol{X}^n(\boldsymbol{\alpha})]_j = \left( \sum_{k: t_k < t_i} \alpha_{ki} \right)^{-1}$ if $t_j < t_i$ and 0 otherwise. Then, we can prove the following consistency result:

**Theorem 1** *If the source probability $\mathbb{P}(s)$ is strictly positive for all $s \in \mathcal{R}$, then, the maximum likelihood estimator $\hat{\boldsymbol{\alpha}}$ given by the solution of Eq. 3 is consistent.*

**Proof** We check the three criteria for consistency: continuity, compactness and identification of the objective function (Newey and McFadden, 1994). Continuity is obvious. For compactness, since $L \to -\infty$ for both $\alpha_{ij} \to 0$ and $\alpha_{ij} \to \infty$ for all $i, j$ so we lose nothing imposing upper and lower bounds thus restricting to a compact subset. For the identification condition, $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^* \Rightarrow \ell^n(\boldsymbol{\alpha}) \neq \ell^n(\boldsymbol{\alpha}^*)$, we use Lemma 9 and 10 (refer to

Appendices 12.1 and 12.2), which establish that $\boldsymbol{X}^n(\boldsymbol{\alpha})$ has full row rank as $n \to \infty$, and hence $\mathcal{Q}^n$ is positive definite. ■

## 5. Recovery Conditions

In this section, we will find a set of sufficient conditions on the diffusion model and the cascade sampling process under which we can recover the network structure from **finite samples**. These results allow us to address two questions:

- *Are there some network structures which are more difficult than others to recover?*
- *What kind of cascades are needed for the network structure recovery?*

The answers to these questions are intertwined. The difficulty of finite-sample recovery depends crucially on an irrepresentability condition which is a function of both network structure, parameters of the diffusion model and the cascade sampling process. Intuitively, the sources of the cascades in a diffusion network have to be chosen in such a way that nodes without parent-child relation should co-occur less often compared to nodes with such relation. Many commonly used diffusion models and network structures can be naturally made to satisfy this condition.

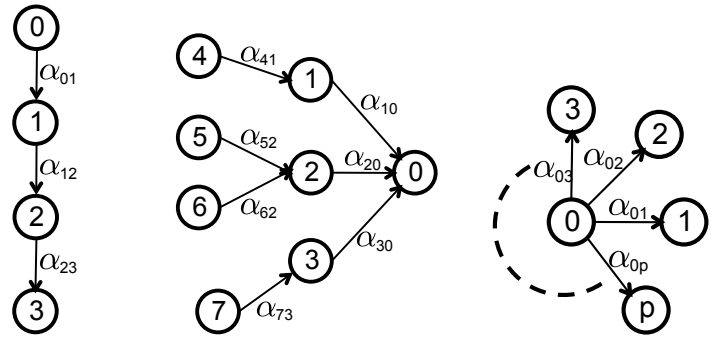
More specifically, we first place two conditions on the Hessian of the population log-likelihood, $\mathbb{E}_c\left[\ell^n(\boldsymbol{\alpha})\right] = \mathbb{E}_c\left[\log g(\mathbf{t}^c; \boldsymbol{\alpha})\right]$, where the expectation here is taken over the distribution $\mathbb{P}(s)$ of the source nodes, and the density $f(\mathbf{t}^c|s)$ of the cascades $\mathbf{t}^c$ given a source node $s$. In this case, we will further denote the Hessian of $\mathbb{E}_c\left[\log g(\mathbf{t}^c; \boldsymbol{\alpha})\right]$ evaluated at the true model parameter $\boldsymbol{\alpha}^*$ as $\mathcal{Q}^*$. Then, we place two conditions on the Lipschitz continuity of $\boldsymbol{X}(\mathbf{t}^c; \boldsymbol{\alpha})$, and the boundedness of $\boldsymbol{X}(\mathbf{t}^c; \boldsymbol{\alpha}^*)$ and $\nabla g(\mathbf{t}^c; \boldsymbol{\alpha}^*)$ at the true model parameter $\boldsymbol{\alpha}^*$. For simplicity, we will denote the subset of indexes associated to node $i$'s true parents as $S$, and its complement as $S^c$. Then, we use $\mathcal{Q}^*_{SS}$ to denote the sub-matrix of $\mathcal{Q}^*$ indexed by $S$ and $\boldsymbol{\alpha}^*_S$ the set of parameters indexed by $S$. Note that $\boldsymbol{\alpha}^*_{S^c} = 0$.

**Condition 1 (Dependency condition):** There exists constants $C_{min} > 0$ and $C_{max} > 0$ such that $\Lambda_{min}\left(\mathcal{Q}^*_{SS}\right) \geq C_{min}$ and $\Lambda_{max}\left(\mathcal{Q}^*_{SS}\right) \leq C_{max}$ where $\Lambda_{min}(\cdot)$ and $\Lambda_{max}(\cdot)$ return the leading and the bottom eigenvalue of its argument respectively. This assumption ensures that two connected nodes co-occur reasonably frequently in the cascades but are not deterministically related.

**Condition 2 (Irrepresentability condition):** There exists a constant $\varepsilon \in (0, 1]$ such that $|||\mathcal{Q}^*_{S^cS}\left(\mathcal{Q}^*_{SS}\right)^{-1}|||_\infty \leq 1 - \varepsilon$, where $|||A|||_\infty = \max_j \sum_k |A_{jk}|$. This assumption captures the intuition that, node $i$ and any of its neighbors should get infected together in a cascade more often than node $i$ and any of its non-neighbors. A similar irrepresentability condition has been proposed on model selection consistency of Lasso (Zhao and Yu, 2006).

**Condition 3 (Lipschitz Continuity):** For any feasible cascade $\mathbf{t}^c$, the Hazard vector $\boldsymbol{X}(\mathbf{t}^c; \boldsymbol{\alpha})$ is Lipschitz continuous in the domain $\{\boldsymbol{\alpha} : \boldsymbol{\alpha}_S \geq \alpha^*_{\min}/2\}$,

$$\|\boldsymbol{X}(\mathbf{t}^c; \boldsymbol{\beta}) - \boldsymbol{X}(\mathbf{t}^c; \boldsymbol{\alpha})\|_2 \leq k_1 \|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2,$$




(a) Chain (b) Tree (c) Star

Figure 2: Example networks.

where $k_1$ is some positive constant. As a consequence, the spectral norm of the difference, $n^{-1/2}(\boldsymbol{X}^n(\boldsymbol{\beta}) - \boldsymbol{X}^n(\boldsymbol{\alpha}))$, is also bounded (refer to appendix 12.3), *i.e.*,

$$|||n^{-1/2}\big(\boldsymbol{X}^n(\boldsymbol{\beta}) - \boldsymbol{X}^n(\boldsymbol{\alpha})\big)|||_2 \leq k_1\|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2. \tag{7}$$

Furthermore, for any feasible cascade $\mathbf{t}^c$, $\boldsymbol{D}(\boldsymbol{\alpha})_{jj}$ is Lipschitz continuous for all $j \in \mathcal{V}$,

$$|\boldsymbol{D}(\mathbf{t}^c;\boldsymbol{\beta})_{jj} - \boldsymbol{D}(\mathbf{t}^c;\boldsymbol{\alpha})_{jj}| \leq k_2\|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2,$$

where $k_2$ is some positive constant.

**Condition 4 (Boundedness):** For any feasible cascade $\mathbf{t}^c$, the absolute value of each entry in the gradient of its log-likelihood and in the Hazard vector, as evaluated at the true model parameter $\boldsymbol{\alpha}^*$, is bounded,

$$\|\nabla g(\mathbf{t}^c;\boldsymbol{\alpha}^*)\|_\infty \leq k_3, \quad \|\boldsymbol{X}(\mathbf{t}^c;\boldsymbol{\alpha}^*)\|_\infty \leq k_4,$$

where $k_3$ and $k_4$ are positive constants. Then the absolute value of each entry in the Hessian matrix $\mathcal{Q}^*$, is also bounded $|||\mathcal{Q}^*|||_\infty \leq k_5$.

**Remarks for condition 1** As stated in Theorem 1, as long as the source probability $\mathbb{P}(s)$ is strictly positive for all $s \in \mathcal{R}$, the maximum likelihood formulation is strictly convex and thus there exists $C_{min} > 0$ such that $\Lambda_{min}(\mathcal{Q}^*) \geq C_{min}$. Moreover, condition 4 implies that there exists $C_{max} > 0$ such that $\Lambda_{max}(\mathcal{Q}^*) \leq C_{max}$.

**Remarks for condition 2** The irrepresentability condition depends, in a non-trivial way, on the network structure, diffusion parameters, observation window and source node distribution. Here, we give some intuition by studying three small canonical examples.

First, consider the chain graph in Fig. 2(a) and assume that we would like to find the incoming edges to node 3 when $T \to \infty$. Then, it is easy to show that the irrepresentability condition is satisfied if $(P_0+P_1)/(P_0+P_1+P_2) < 1-\varepsilon$ and $P_0/(P_0+P_1+P_2) < 1-\varepsilon$, where $P_i$ denotes the probability of a node $i$ to be the source of a cascade. Thus, for example, if the source of each cascade is chosen uniformly at random, the inequality is satisfied. Here, the irrepresentability condition depends on the source node distribution.

Second, consider the directed tree in Fig. 2(b) and assume that we would like to find the incoming edges to node 0 when $T \to \infty$. Then, it can be shown that the irrepresentability condition is satisfied as long as (1) $P_1 > 0$, (2) ($P_2 > 0$) or ($P_5 > 0$ and $P_6 > 0$), and (3) $P_3 > 0$. As in the chain, the condition depends on the source node distribution.

Finally, consider the star graph in Fig. 2(c), with exponential edge transmission functions, and assume that we would like to find the incoming edges to a leave node $i$ when $T < \infty$. Then, as long as the root node has a nonzero probability $P_0 > 0$ of being the source of a cascade, it can be shown that the irrepresentability condition reduces to the inequalities $\left(1 - \frac{\alpha_{0j}}{\alpha_{0i}+\alpha_{0j}}\right) e^{-(\alpha_{0i}+\alpha_{0j})T} + \frac{\alpha_{0j}}{\alpha_{0i}+\alpha_{0j}} < 1 - \varepsilon(1 + e^{-\alpha_{0i}T})$, $j = 1, \ldots, p : j \neq i$, which always holds for some $\varepsilon > 0$. If $T \to \infty$, then the condition holds whenever $\varepsilon < \alpha_{0i}/(\alpha_{0i} + \max_{j:j\neq i} \alpha_{0j})$. Here, the larger the ratio $\max_{j:j\neq i} \alpha_{0j}/\alpha_{0i}$ is, the smaller the maximum value of $\varepsilon$ for which the irrepresentability condition holds. To summarize, as long as $P_0 > 0$, there is always some $\varepsilon > 0$ for which the condition holds, and such $\varepsilon$ value depends on the time window and the parameters $\alpha_{0j}$.

**Remarks for conditions 3 and 4** Well-known pairwise transmission likelihoods such as exponential, Rayleigh or Power-law, used in previous work Gomez-Rodriguez et al. (2011), satisfy conditions 3 and 4.

## 6. Sample Complexity

*How many cascades do we need to recover the network structure?* We will answer this question by providing a sample complexity analysis of the optimization in Eq. 4. Given the conditions spelled out in Section 5, we can show that the number of cascades needs to grow polynomially in the number of true parents of a node, and depends only logarithmically on the size of the network. This is a positive result, since the network size can be very large (millions or billions), but the number of parents of a node is usually small compared the network size. More specifically, for each individual node, we have the following result:

**Theorem 2** *Consider an instance of the continuous-time diffusion model with parameters $\alpha^*_{ji}$ and associated edges $\mathcal{E}^*$ such that the model satisfies condition 1-4, and let $C^n$ be a set of $n$ cascades drawn from the model. Suppose that the regularization parameter $\lambda_n$ is selected to satisfy*

$$\lambda_n \geq 8k_3 \frac{2-\varepsilon}{\varepsilon} \sqrt{\frac{\log p}{n}}. \tag{8}$$

*Then, there exist positive constants $L$ and $K$, independent of $(n, p, d)$, such that if*

$$n > Ld^3 \log p, \tag{9}$$

*then the following properties hold with probability at least $1 - 2\exp(-K\lambda_n^2 n)$:*

1. *For each node $i \in \mathcal{V}$, the $\ell_1$-regularized network inference problem defined in Eq. 4 has a unique solution, and so uniquely specifies a set of incoming edges of node $i$.*
2. *For each node $i \in \mathcal{V}$, the estimated set of incoming edges does not include any false edges and include all true edges.*

*Furthermore, suppose that the finite sample Hessian matrix $\mathcal{Q}^n$ satisfies conditions 1 and 2. Then there exist positive constants $L$ and $K$, independent of $(n, p, d)$, such that the sample complexity can be improved to $n > Ld^2 \log p$ with other statements remain the same.*

**Remarks.** The above sample complexity is proved for each node separately for recovering its parents. Using a union bound, we can provide the sample complexity for recovering the entire network structure by joining these parent-child relations together. The resulting sample complexity and the choice of regularization parameters will remain largely the same,

except that the dependency on $d$ will change from $d$ to $d_{max}$ (the largest number of parents of a node), and the dependency on $p$ will change from $\log p$ to $2 \log N$ ($N$ the number of nodes in the network).

### 6.1 Outline of Analysis

The proof of Theorem 2 uses a technique called primal-dual witness method, previously used in the proof of sparsistency of Lasso (Wainwright, 2009) and high-dimensional Ising model selection (Ravikumar et al., 2010). To the best of our knowledge, the present work is the first that uses this technique in the context of diffusion network inference. First, we show that the optimal solutions to Eq. 4 have shared sparsity pattern, and under a further condition, the solution is unique (proven in Appendix 12.4):

**Lemma 3** *Suppose that there exists an optimal primal-dual solution $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\mu}})$ to Eq. 4 with an associated subgradient vector $\hat{\mathbf{z}}$ such that $||\hat{\mathbf{z}}_{S^c}||_\infty < 1$. Then, any optimal primal solution $\tilde{\boldsymbol{\alpha}}$ must have $\tilde{\boldsymbol{\alpha}}_{S^c} = 0$. Moreover, if the Hessian sub-matrix $\mathcal{Q}^n_{SS}$ is strictly positive definite, then $\hat{\boldsymbol{\alpha}}$ is the unique optimal solution.*

Next, we will construct a primal-dual vector $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\mu}})$ along with an associated subgradient vector $\hat{\mathbf{z}}$. Furthermore, we will show that, under the assumptions on $(n, p, d)$ stated in Theorem 2, our constructed solution satisfies the KKT optimality conditions to Eq. 4, and the primal vector has the same sparsity pattern as the true parameter $\boldsymbol{\alpha}^*$, *i.e.*,

$$\hat{\alpha}_j > 0, \ \forall j : \alpha^*_j > 0, \tag{10}$$

$$\hat{\alpha}_j = 0, \ \forall j : \alpha^*_j = 0. \tag{11}$$

Then, based on Lemma 3, we can deduce that the optimal solution to Eq. 4 correctly recovers the sparsisty pattern of $\boldsymbol{\alpha}^*$, and thus the incoming edges to node $i$.

More specifically, we start by realizing that a primal-dual optimal solution $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\mu}})$ to Eq. 4 must satisfy the generalized Karush-Kuhn-Tucker (KKT) conditions Boyd and Vandenberghe (2004):

$$0 \in \nabla \ell^n(\tilde{\boldsymbol{\alpha}}) + \lambda_n \tilde{\mathbf{z}} - \tilde{\boldsymbol{\mu}}, \tag{12}$$

$$\tilde{\mu}_j \tilde{\alpha}_j = 0, \tag{13}$$

$$\tilde{\mu}_j \geq 0, \tag{14}$$

$$\tilde{z}_j = 1, \ \forall \tilde{\alpha}_j > 0, \tag{15}$$

$$|\tilde{z}_j| \leq 1, \ \forall \tilde{\alpha}_j = 0, \tag{16}$$

where $\ell^n(\tilde{\boldsymbol{\alpha}}) = -\frac{1}{n} \sum_{c \in C^n} \log g(\mathbf{t}^c; \tilde{\boldsymbol{\alpha}})$ and $\tilde{\mathbf{z}}$ denotes the subgradient of the $\ell_1$-norm.

Suppose the true set of parent of node $i$ is $S$. We construct the primal-dual vector $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\mu}})$ and the associated subgradient vector $\hat{\mathbf{z}}$ in the following way

1. We set $\hat{\boldsymbol{\alpha}}_S$ as the solution to the partial regularized maximum likelihood problem

$$\hat{\boldsymbol{\alpha}}_S = \operatorname*{argmin}_{(\boldsymbol{\alpha}_S, 0), \boldsymbol{\alpha}_S \geq 0} \{\ell^n(\boldsymbol{\alpha}) + \lambda_n ||\boldsymbol{\alpha}_S||_1\}. \tag{17}$$

Then, we set $\hat{\boldsymbol{\mu}}_S \geq 0$ as the dual solution associated to the primal solution $\hat{\boldsymbol{\alpha}}_S$.

2. We set $\hat{\boldsymbol{\alpha}}_{S^c} = 0$, so that condition (11) holds, and $\hat{\boldsymbol{\mu}}_{S^c} = \boldsymbol{\mu}^*_{S^c} \geq 0$, where $\boldsymbol{\mu}^*$ is the optimal dual solution to the following problem:

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{\alpha}} & \mathbb{E}_c\left[\ell^n(\boldsymbol{\alpha})\right] \\ \text{subject to} & \alpha_j \geq 0,\, j = 1, \ldots, N, i \neq j. \end{array} \tag{18}$$

Thus, our construction satisfies condition (14).

3. We obtain $\hat{\mathbf{z}}_{S^c}$ from (12) by substituting in the constructed $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{z}}_S$.

Then, we only need to prove that, under the stated scalings of $(n, p, d)$, with high-probability, the remaining KKT conditions (10), (13), (15) and (16) hold.

For simplicity of exposition, we first assume that the dependency and irrepresentability conditions hold for the finite sample Hessian matrix $\mathcal{Q}^n$. Later we will lift this restriction and only place these conditions on the population Hessian matrix $\mathcal{Q}^*$. The following lemma (proven in Appendix 12.5) show that our constructed solution satisfies condition (10):

**Lemma 4** *Under condition 3, if the regularization parameter is selected to satisfy*

$$\sqrt{d}\lambda_n \leq \frac{C^2_{\min}}{6(k_2 + 2k_1\sqrt{C_{\max}})},$$

*and* $\|\nabla_s \ell^n(\boldsymbol{\alpha}^*)\|_\infty \leq \frac{\lambda_n}{4}$*, then,*

$$\|\hat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}^*_S\|_2 \leq \frac{3\sqrt{d}\lambda_n}{C_{\min}} \leq \frac{\alpha^*_{\min}}{2},$$

*as long as* $\alpha^*_{\min} \geq 6\sqrt{d}\lambda_n / C_{\min}$.

Based on this lemma, we can then further show that the KKT conditions (13) and (15) also hold for the constructed solution. This can be trivially deduced from condition (10) and (11), and our construction steps (a) and (b). Note that it also implies that $\hat{\boldsymbol{\mu}}_S = \boldsymbol{\mu}^*_S = 0$, and hence $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}^*$.

Proving condition (16) is more challenging. We first provide more details on how to construct $\hat{\mathbf{z}}_{S^c}$ mentioned in step (c). We start by using a Taylor expansion of Eq. 12,

$$\mathcal{Q}^n(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) = -\nabla \ell^n(\boldsymbol{\alpha}^*) - \lambda_n \hat{\boldsymbol{z}} + \hat{\boldsymbol{\mu}} - \mathbf{R}^n, \tag{19}$$

where $\mathbf{R}^n$ is a remainder term with its $j$-th entry

$$R^n_j = \left[\nabla^2 \ell^n(\bar{\boldsymbol{\alpha}}_j) - \nabla^2 \ell^n(\boldsymbol{\alpha}^*)\right]^T_j (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*),$$

and $\bar{\boldsymbol{\alpha}}_j = \theta_j \hat{\boldsymbol{\alpha}} + (1 - \theta_j)\boldsymbol{\alpha}^*$ with $\theta_j \in [0, 1]$ according to the mean value theorem. Rewriting Eq. 19 using block matrices

$$\begin{pmatrix} \mathcal{Q}^n_{SS} & \mathcal{Q}^n_{SS^c} \\ \mathcal{Q}^n_{S^cS} & \mathcal{Q}^n_{S^cS^c} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}^*_S \\ \hat{\boldsymbol{\alpha}}_{S_c} - \boldsymbol{\alpha}^*_{S^c} \end{pmatrix} = - \begin{pmatrix} \nabla_S \ell^n(\boldsymbol{\alpha}^*) \\ \nabla_{S^c} \ell^n(\boldsymbol{\alpha}^*) \end{pmatrix} - \lambda_n \begin{pmatrix} \hat{\boldsymbol{z}}_S \\ \hat{\boldsymbol{z}}_{S^c} \end{pmatrix} + \begin{pmatrix} \hat{\mu}_S \\ \hat{\mu}_{S^c} \end{pmatrix} - \begin{pmatrix} \mathbf{R}^n_S \\ \mathbf{R}^n_{S^c} \end{pmatrix} \tag{20}$$

and, after some algebraic manipulation, we have

$$\lambda \hat{\mathbf{z}}_{S^c} = -\nabla_{S^c} \ell^n(\boldsymbol{\alpha}^*) + \hat{\boldsymbol{\mu}}_{S^c} - \mathbf{R}^n_{S^c} - \mathcal{Q}^n_{S^cS}(\mathcal{Q}^n_{SS})^{-1}\big( - \nabla_s \ell^n(\boldsymbol{\alpha}^*) - \lambda \hat{\boldsymbol{z}}_S + \hat{\boldsymbol{\mu}}_S - \mathbf{R}^n_S\big). \tag{21}$$

Next, we upper bound $\|\hat{\mathbf{z}}_{S^c}\|_\infty$ using the triangle inequality

$$\begin{aligned} \|\hat{\mathbf{z}}_{S^c}\|_\infty \quad & \leq \lambda_n^{-1}\|\boldsymbol{\mu}^*_{S^c} - \nabla_{S^c}\ell^n(\boldsymbol{\alpha}^*)\|_\infty + \lambda_n^{-1}\|\mathbf{R}^n_{S^c}\|_\infty + \|\mathcal{Q}^n_{S^cS}(\mathcal{Q}^n_{SS})^{-1}\|_\infty \times \big[1 + \lambda_n^{-1}\|\mathbf{R}^n_S\|_\infty \\ & \quad + \lambda_n^{-1}\|\mu^*_S - \nabla_S \ell^n(\boldsymbol{\alpha}^*)\|_\infty\big], \end{aligned}$$

and we want to prove that this upper bound is smaller than 1. This can be done with the help of the following two lemmas (proven in Appendices 12.6 and 12.7):

**Lemma 5** *Given $\varepsilon \in (0,1]$ from the irrepresentability condition, we have,*

$$P\left(\frac{2-\varepsilon}{\lambda_n}\|\nabla \ell^n(\boldsymbol{\alpha}^*) - \boldsymbol{\mu}^*\|_\infty \geq 4^{-1}\varepsilon\right) \leq 2p\exp\left(-\frac{n\lambda_n^2\varepsilon^2}{32k_3^2\left(2-\varepsilon\right)^2}\right), \tag{22}$$

*which converges to zero at rate $\exp(-c\lambda_n^2 n)$ as long as $\lambda_n \geq 8k_3\frac{2-\varepsilon}{\varepsilon}\sqrt{\frac{\log p}{n}}$.*

**Lemma 6** *Given $\varepsilon \in (0,1]$ from the irrepresentability condition, if conditions 3 and 4 holds, $\lambda_n$ is selected to satisfy*

$$\lambda_n d \leq C_{\min}^2 \frac{\varepsilon}{36K(2-\varepsilon)},$$

*where $K = k_1 + k_4k_1 + k_1^2 + k_1\sqrt{C_{\max}}$, and $\|\nabla_s \ell^n(\boldsymbol{\alpha}^*)\|_\infty \leq \frac{\lambda_n}{4}$, then, $\frac{\|\mathbf{R}^n\|_\infty}{\lambda_n} \leq \frac{\varepsilon}{4(2-\varepsilon)}$, as long as $\alpha_{\min}^* \geq 6\sqrt{d}\lambda_n/C_{\min}$.*

Now, applying both lemmas and the irrepresentability condition on the finite sample Hessian matrix $\mathcal{Q}^n$, we have

$$\begin{aligned}\|\hat{\mathbf{z}}_{S^c}\|_\infty &\leq (1-\varepsilon) + \lambda_n^{-1}(2-\varepsilon)\|\mathbf{R}^n\|_\infty + \lambda_n^{-1}(2-\varepsilon)\|\boldsymbol{\mu}^* - \nabla\ell^n(\boldsymbol{\alpha}^*)\|_\infty \\ &\leq (1-\varepsilon) + 0.25\varepsilon + 0.25\varepsilon = 1 - 0.5\varepsilon,\end{aligned}$$

and thus condition (16) holds.

A possible choice of the regularization parameter $\lambda_n$ and cascade set size $n$ such that the conditions of the Lemmas 4-6 are satisfied is $\lambda_n = 8k_3(2-\varepsilon)\varepsilon^{-1}\sqrt{n^{-1}\log p}$ and $n > 288^2k_3^2(2-\varepsilon)^4C_{min}^{-4}\varepsilon^{-4}d^2\log p + \left(48k_3(2-\varepsilon)C_{min}^{-1}(\alpha_{min}^*)^{-1}\varepsilon^{-1}\right)^2 d\log p$.

Last, we lift the dependency and irrepresentability conditions imposed on the finite sample Hessian matrix $\mathcal{Q}^n$. We show that if we only impose these conditions in the corresponding population matrix $\mathcal{Q}^*$, then they will also hold for $\mathcal{Q}^n$ with high probability (proven in Appendices 12.8 and 12.9).

**Lemma 7** *If condition 1 holds for $\mathcal{Q}^*$, then, for any $\delta > 0$,*

$$P\left(\Lambda_{min}\left(\mathcal{Q}_{SS}^n\right) \leq C_{min} - \delta\right) \leq 2d^{B_1}\exp\left(-A_1\frac{\delta^2 n}{d^2}\right),$$

$$P\left(\Lambda_{max}\left(\mathcal{Q}_{SS}^n\right) \geq C_{max} + \delta\right) \leq 2d^{B_2}\exp\left(-A_2\frac{\delta^2 n}{d^2}\right),$$

*where $A_1$, $A_2$, $B_1$ and $B_2$ are constants independent of $(n,p,d)$.*

**Lemma 8** *If $|||\mathcal{Q}_{S^cS}^*\left(\mathcal{Q}_{SS}^*\right)^{-1}|||_\infty \leq 1-\varepsilon$, then,*

$$P\left(\|\mathcal{Q}_{S^cS}^n(\mathcal{Q}_{SS}^n)^{-1}\|_\infty \geq 1-\varepsilon/2\right) \leq p\exp\left(-K\frac{n}{d^3}\right),$$

*where $K$ is a constant independent of $(n,p,d)$.*

Note in this case the cascade set size need to increase to $n > Ld^3\log p$, where $L$ is a sufficiently large positive constant independent of $(n,p,d)$, for the error probabilities on these last two lemmas to converge to zero.

**Algorithm 1** $\ell_1$-regularized network inference

**Require:** $C^n, \lambda_n, K, L$
  **for all** $i \in \mathcal{V}$ **do**
    $k = 0$
    **while** $k < K$ **do**
      $\boldsymbol{\alpha}_i^{k+1} = \left(\boldsymbol{\alpha}_i^k - L\nabla_{\boldsymbol{\alpha}_i}\ell^n(\boldsymbol{\alpha}_i^k) - \lambda_n L\right)_+$
      $k = k + 1$
    **end while**
    $\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i^{K-1}$
  **end for**
  **return** $\{\hat{\boldsymbol{\alpha}}_i\}_{i\in\mathcal{V}}$

## 7. Efficient soft-thresholding algorithm

*Can we design efficient algorithms to solve Eq.* (4) *for network recovery?* Here, we will design a proximal gradient algorithm which is well suited for solving non-smooth, constrained, large-scale or high-dimensional convex optimization problems Parikh and Boyd (2013). Moreover, they are easy to understand, derive, and implement. We first rewrite Eq. 4 as an unconstrained optimization problem:

$$\text{minimize}_{\boldsymbol{\alpha}} \quad \ell^n(\boldsymbol{\alpha}) + g(\boldsymbol{\alpha}),$$

where the non-smooth convex function $g(\boldsymbol{\alpha}) = \lambda_n||\boldsymbol{\alpha}||_1$ if $\boldsymbol{\alpha} \geq 0$ and $+\infty$ otherwise. By rewriting both problems as a sum of a smooth convex function $\ell^n(\boldsymbol{\alpha})$ and a non-smooth convex function $g(\boldsymbol{\alpha})$, the general recipe from Parikh and Boyd (2013) for designing proximal gradient algorithm can be applied directly.

Algorithm 1 summarizes the resulting algorithm. In each iteration of the algorithm, we need to compute $\nabla\ell^n$ (Table 2) and the proximal operator $\mathbf{prox}_{L^k g}(\mathbf{v})$, where $L^k$ is a step size that we can set to a constant value $L$ or find using a simple line search Beck and Teboulle (2009). Using Moreau's decomposition, we have

$$\mathbf{prox}_{L^k g}(\mathbf{v}) = \mathbf{v} - L^k\ \mathbf{prox}_{g^*/L^k}(\mathbf{v}/L^k), \tag{23}$$

where

$$g^*(\mathbf{y}) = \sup_{\mathbf{x}} \left((\mathbf{y} - \lambda_n\mathbf{1})^T x - \mathbf{1}(\mathbf{x} \geq 0)\right) = \begin{cases} \infty & \text{if } \exists i : y_i > \lambda_n \\ 0 & \text{otherwise} \end{cases}$$

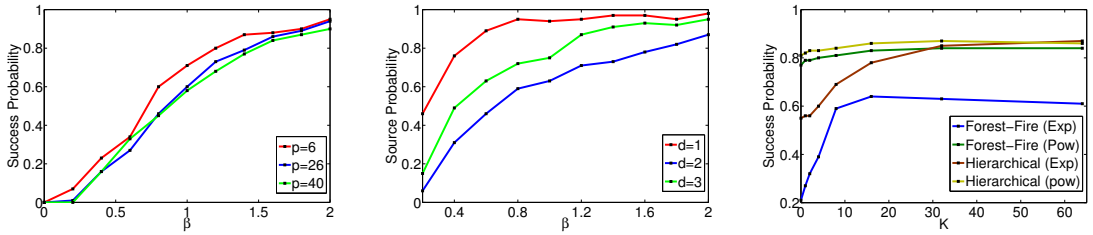
is the conjugate function of $g$. Then,

$$\mathbf{prox}_{g^*/L^k}(\mathbf{v}/L^k) = \operatorname*{argmin}_{\mathbf{y}}\{g^*(\mathbf{y}) + \frac{L^k}{2}\|y - \mathbf{v}/L^k\|_2^2\} = (\mathbf{v} - \lambda_n L^k)_+$$

In summary, the proximal operator for our particular function $g(\cdot)$ is a soft-thresholding operator, $(\mathbf{v} - \lambda_n L^k)_+$, which leads to a sparse optimal solution $\hat{\boldsymbol{\alpha}}$, as desired.

## 8. Experiments

In this section, we first illustrate some consequences of Th. 2 by applying our algorithm to several types of networks, parameters $(n, p, d)$, and regularization parameter $\lambda_n$. Then, we compare our algorithm to two different state-of-the-art algorithms: NetRate Gomez-Rodriguez et al. (2011) and First-Edge Abrahao et al. (2013).



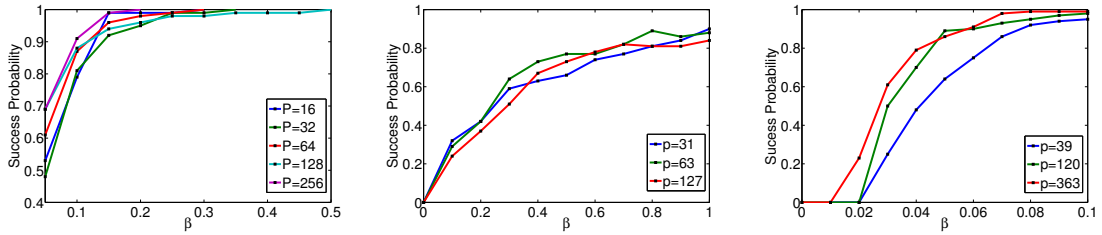

(a) Super-neighborhood $p_i$ (b) In-degrees $d_i$ (c) $\lambda_n$

Figure 3: Success probability vs. # of cascades.

**Experimental Setup** We focus on synthetic networks that mimic the structure of real-world diffusion networks – in particular, social networks. We consider two models of directed real-world social networks: the Forest Fire model (Barabási and Albert, 1999) and the Kronecker Graph model (Leskovec et al., 2010), and use simple pairwise transmission models such as exponential, power-law or Rayleigh. We use networks with 128 nodes and, for each edge, we draw its associated transmission rate from a uniform distribution $U(0.5, 1.5)$. In general, we proceed as follows: we generate a network $\mathcal{G}^*$ and transmission rates $\mathbf{A}^*$, simulate a set of cascades and, for each cascade, record the node infection times. Then, given the infection times, we infer a network $\hat{\mathcal{G}}$. Finally, when we illustrate the consequences of Th. 2, we evaluate the accuracy of the inferred neighborhood of a node $\hat{\mathcal{N}}^-(i)$ using probability of success $P(\hat{\mathcal{E}} = \mathcal{E}^*)$, estimated by running our method of 100 independent cascade sets. When we compare our algorithm to NetRate and First-Edge, we use the $F_1$ score, which is defined as $2PR/(P+R)$, where precision (P) is the fraction of edges in the inferred network $\hat{\mathcal{G}}$ present in the true network $\mathcal{G}^*$, and recall (R) is the fraction of edges of the true network $\mathcal{G}^*$ present in the inferred network $\hat{\mathcal{G}}$.

**Parameters** $(n, p, d)$ According to Th. 2, the number of cascades that are necessary to successfully infer the incoming edges of a node will increase polynomially to the node's neighborhood size $d_i$ and logarithmically to the super-neighborhood size $p_i$. Here, we first infer the incoming links of nodes on the same type of canonical networks as depicted in Fig. 2. We choose nodes the same in-degree but different super-neighboorhod set sizes $p_i$ and experiment with different scalings $\beta$ of the number of cascades $n = 10\beta d \log p$. We set the regularization parameter $\lambda_n$ as a constant factor of $\sqrt{\log(p)/n}$ as suggested by Theorem 2 and, for each node, we used cascades which contained at least one node in the super-neighborhood of the node under study. We used an exponential transmission model and time window $T = 10$. As predicted by Theorem 2, very different $p$ values lead to curves that line up with each other quite well.

Next, we infer the incoming links of nodes of a larger hierarchical Kronecker network. Again, we choose nodes with the same in-degree ($d_i = 3$) but different super-neighboorhod set sizes $p_i$ under different scalings $\beta$ of the number of cascades $n = 10\beta d \log p$. We used an exponential transmission model and $T = 5$. Fig. 3(a) summarizes the results, where, for each node, we used cascades which contained at least one node in the super-neighborhood of the node under study. Similarly as in the case of the canonical networks, very different $p$ values lead to curves that line up with each other quite well.




(a) Chain ($d_i = 1$) (b) Stars ($d_i = 1$) (c) Tree ($d_i = 3$)

Figure 4: Success probability vs. # of cascades. Different super-neighborhood sizes $p_i$.

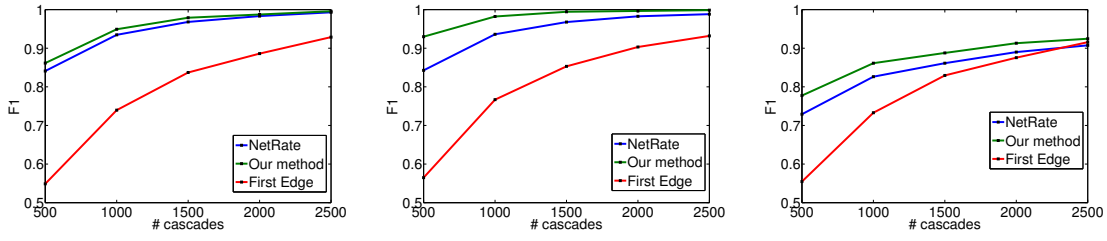
Finally, we infer the incoming links of nodes of a hierarchical Kronecker network with equal super neighborhood size ($p_i = 70$) but different in-degree ($d_i$) under different scalings $\beta$ of the number of cascades $n = 10\beta d \log p$ and choose the regularization parameter $\lambda_n$ as a constant factor of $\sqrt{\log(p)/n}$ as suggested by Theorem 2. We used an exponential transmission model and time window $T = 5$. Figure 3(b) summarizes the results, where we observe that, as predicted by Theorem 2, different $d$ values lead to noticeably different curves.

**Regularization parameter** $\lambda_n$ Our main result indicates that the regularization parameter $\lambda_n$ should be a constant factor of $\sqrt{\log(p)/n}$. Fig. 3(c) shows the success probability of our algorithm against different scalings $K$ of the regularization parameter $\lambda_n = K\sqrt{\log(p)/n}$ for different types of networks using 150 cascades and $T = 5$. We find that for sufficiently large $\lambda_n$, the success probability flattens, as expected from Th. 2. It flattens at values smaller than one because we used a fixed number of cascades $n$, which may not satisfy the conditions of Th. 2.

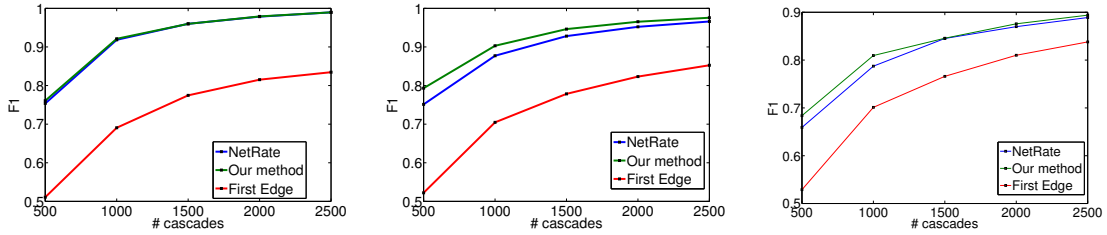
**Comparison with NetRate and First-Edge** Fig. 5 compares the accuracy of our algorithm, NetRate and First-Edge against number of cascades for three hierarchical Kronecker network and three Forest Fire networks, with power-law (Pow), exponential (Exp) and rayleigh (Ray) transmission models, and an observation window $T = 10$. Our method outperforms both competitive methods, finding especially striking the competitive advantage with respect to First-Edge, however, this may be explained by comparing the sample complexity results for both methods: First-Edge needs $O(Nd \log N)$ cascades to achieve a probability of success approaching 1 in a rate polynomial in the number of cascades while our method needs $O(d^3 \log N)$ to achieve a probability of success approaching 1 in a rate exponential in the number of cascades.

## 9. Discussion

Our results can be extended in multiple directions. First, our novel formulation of the diffusion network recovery problem as a $\ell_1$-regularized convex optimization problem establishes a connection between the literature on information diffusion and a vast literature on high dimension sparse recovery problem from machine learning and statistics literature. This connection allows us to borrow analysis frameworks for graphical model structure estimation to analyze information diffusion. In terms of diffusion models, we can extend the current independent cascade model to deal with nonparametric transmission functions (Du




(a) Kronecker Hier., Pow (b) Kronecker Hier., Exp (c) Kronecker Hier., Ray




(d) Forest Fire, Pow (e) Forest Fire, Exp (f) Forest Fire, Ray

Figure 5: $F_1$-score vs. # of cascades.

et al., 2012a), transmission function conditioned on additional features (Du et al., 2013), diffusion models which allows for multiple events (Zhou et al., 2013). All three models will result in convex loss function; and for the former two models, we can employ grouped lasso regularization (Yuan and Lin, 2006), while for the latter model, we can employ a nuclear norm regularization (Recht et al., 2010). These models are more complicated than the independent cascade model we studied in the paper, but analysis for these models can be carried out using a general M-estimation analysis framework (Negahban et al., 2009), since these regularizers are decomposable and one needs to check the restricted strong convexity of the loss function. In terms of the estimation algorithms, we can employ proximal algorithms (Parikh and Boyd, 2013) or the conditional gradient algorithm (Jaggi, 2013) to deal with different type of diffusion models and regularizers. When the data is large, one can consider distributed (Boyd et al., 2011) and online estimation (Nemirovski et al., 2009) procedures.

Our results also bring out interesting further open problems on diffusion network estimations. For instance, the success of the network inference algorithm in Equation (2) relies on the fulfillment of the above mentioned irrepresentability condition on the Hessian, $\mathcal{Q}^*$, of the population log-likelihood $\mathbb{E}[\ell^n]$, where the expectation here is taken over the distribution $\mathbb{P}(s)$ of the source nodes and the random generative process of the diffusion model given a source node $s$. This condition captures the intuition that, node $i$ and any of its neighbors should get infected together in a cascade more often than node $i$ and any of its non-neighbors. Unfortunately, the irrepresentability condition depends, in a non-trivial way, on the network structure, diffusion parameters, and the source distribution $\mathbb{P}(s)$, which are all *unknown* during the network inference stage. Previous work has typically assumed the network structure, diffusion parameters, observation window and source distribution to be fixed, and source locations are sampled *passively* from the latter. However, in practice,

the source locations to sample from may be determined *actively* in a sequential manner, potentially based on the information gathered from previous source locations. Thus an interesting open question is:

> Suppose there exists an unknown $\mathbb{P}(s)$ where the irrepresentability conditions hold for the diffusion model. Under what conditions, can we design an "active" algorithm which samples the source location intelligently and achieves the sample complexity in Theorem 2, or even better sample complexity, *e.g.*, $o(d_i^3 \log N)$?

## 10. Conclusions

Our work contributes towards establishing a theoretical foundation of the network inference problem. Specifically, we proposed a $\ell_1$-regularized maximum likelihood inference method for a well-known continuous-time diffusion model and an efficient proximal gradient implementation, and then show that, for general networks satisfying a natural irrepresentability condition, our method achieves an exponentially decreasing error with respect to the number of cascades as long as $O(d^3 \log N)$ cascades are recorded.

Our work also opens many interesting venues for future work. For example, given a fixed number of cascades, it would be useful to provide confidence intervals on the inferred edges. Further, a detailed theoretical analysis of the irrepresentability condition on large synthetic networks that mimic the structure of real-world diffusion networks, such as Kronecker or Forest-Fire networks, is still missing. Given a network with arbitrary pairwise likelihoods, it is an open question whether there always exists at least one source distribution and time window value such that the irrepresentability condition is satisfied, and, and if so, whether there is an efficient way of finding this distribution. Finally, our work assumes all activations occur due to network diffusion and are recorded. It would be interesting to allow for missing observations, as well as activations due to exogenous factors.

## 11. Acknowledgement

This research was supported in part by NSF/NIH BIGDATA 1R01GM108341-01, NSF IIS1116886, NSF CAREER IIS-1350983 and a Raytheon faculty fellowship to L. Song.

## 12. Appendix

### 12.1 Proof of Lemma 9

**Lemma 9** *Given log-concave survival functions and concave hazard functions in the parameter(s) of the pairwise transmission likelihoods, then, a sufficient condition for the Hessian matrix $\mathcal{Q}^n$ to be positive definite is that the hazard matrix $X^n(\boldsymbol{\alpha})$ is non-singular.*

**Proof** Using Eq. 5, the Hessian matrix can be expressed as a sum of two matrices, $\mathbf{D}^n(\boldsymbol{\alpha})$ and $\mathbf{X}^n(\boldsymbol{\alpha})\mathbf{X}^n(\boldsymbol{\alpha})^\top$. The matrix $\mathbf{D}^n(\boldsymbol{\alpha})$ is trivially positive semidefinite by log-concavity of the survival functions and concavity of the hazard functions. The matrix $\mathbf{X}^n(\boldsymbol{\alpha})\mathbf{X}^n(\boldsymbol{\alpha})^\top$ is positive definite matrix since $\mathbf{X}^n(\boldsymbol{\alpha})$ is full rank by assumption. Then, the Hessian matrix is positive definite since it is a sum a positive semidefinite matrix and a positive definite matrix. ■

### 12.2 Proof of Lemma 10

**Lemma 10** *If the source probability $\mathbb{P}(s)$ is strictly positive for all $s \in \mathcal{R}$, then, for an arbitrarily large number of cascades $n \to \infty$, there exists an ordering of the nodes and cascades within the cascade set such that the hazard matrix $\boldsymbol{X}^n(\boldsymbol{\alpha})$ is non-singular.*

**Proof** In this proof, we find a labeling of the nodes (row indices in $\mathbf{X}^n(\boldsymbol{\alpha})$) and ordering of the cascades (column indices in $\mathbf{X}^n(\boldsymbol{\alpha})$), such that, for an arbitrary large number of cascades, we can express the matrix $\mathbf{X}^n(\boldsymbol{\alpha})$ as $[T\, B]$, where $T \in \mathbb{R}^{p \times p}$ is an upper triangular with nonzero diagonal elements and $B \in \mathbb{R}^{p \times n-p}$. And, therefore, $\mathbf{X}^n(\boldsymbol{\alpha})$ has full rank (rank $p$). We proceed first by sorting nodes in $\mathcal{R}$ and then continue by sorting nodes in $\mathcal{U}$:

- **Nodes in $\mathcal{R}$**: For each node $u \in \mathcal{R}$, consider the set of cascades $C_u$ in which $u$ was a source and $i$ got infected. Then, rank each node $u$ according to the earliest position in which node $i$ got infected across all cascades in $C_u$ in decreasing order, breaking ties at random. For example, if a node $u$ was, at least once, the source of a cascade in which node $i$ got infected just after the source, but in contrast, node $v$ was never the source of a cascade in which node $i$ got infected the second, then node $u$ will have a lower index than node $v$. Then, assign row $k$ in the matrix $\mathbf{X}^n(\boldsymbol{\alpha})$ to node in position $k$ and assign the first $d$ columns to the corresponding cascades in which node $i$ got infected earlier. In such ordering, $\mathbf{X}^n(\boldsymbol{\alpha})_{mk} = 0$ for all $m < k$ and $\mathbf{X}^n(\boldsymbol{\alpha})_{kk} \neq 0$.

- **Nodes in $\mathcal{U}$**: Similarly as in the first step, and assign them the rows $d+1$ to $p$. Moreover, we assign the columns $d+1$ to $p$ to the corresponding cascades in which node $i$ got infected earlier. Again, this ordering satisfies that $\mathbf{X}^n(\boldsymbol{\alpha})_{mk} = 0$ for all $m < k$ and $\mathbf{X}^n(\boldsymbol{\alpha})_{kk} \neq 0$. Finally, the remaining columns $n - p$ can be assigned to the remaining cascades at random.

This ordering leads to the desired structure $[T\, B]$, and thus it is non-singular. ■

### 12.3 Proof of Eq 7.

If the Hazard vector $\boldsymbol{X}(\mathbf{t}^c; \boldsymbol{\alpha})$ is Lipschitz continuous in the domain $\{\boldsymbol{\alpha} : \boldsymbol{\alpha}_S \geq \frac{\alpha^*_{\min}}{2}\}$,

$$\|\boldsymbol{X}(\mathbf{t}^c; \boldsymbol{\beta}) - \boldsymbol{X}(\mathbf{t}^c; \boldsymbol{\alpha})\|_2 \leq k_1 \|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2,$$

where $k_1$ is some positive constant. Then, we can bound the spectral norm of the difference, $\frac{1}{\sqrt{n}}(\boldsymbol{X}^n(\boldsymbol{\beta}) - \boldsymbol{X}^n(\boldsymbol{\alpha}))$, in the domain $\{\boldsymbol{\alpha} : \boldsymbol{\alpha}_S \geq \frac{\alpha^*_{\min}}{2}\}$ as follows:

$$\|\|\frac{1}{\sqrt{n}}\big(\mathbf{X}^n(\boldsymbol{\beta}) - \mathbf{X}^n(\boldsymbol{\alpha})\big)\|\|_2 = \max_{\|\boldsymbol{u}\|_2=1} \frac{1}{\sqrt{n}} \|\boldsymbol{u}\big(\mathbf{X}^n(\boldsymbol{\beta}) - \mathbf{X}^n(\boldsymbol{\alpha})\big)\|_2$$

$$= \max_{\|\boldsymbol{u}\|_2=1} \frac{1}{\sqrt{n}} \sqrt{\sum_{c=1}^{n} \langle \boldsymbol{u}, \mathbf{X}(\mathbf{t}^c; \boldsymbol{\beta}) - \mathbf{X}(\mathbf{t}^c; \boldsymbol{\alpha})\rangle^2} \leq \frac{1}{\sqrt{n}} \sqrt{k_1^2 n \|\boldsymbol{u}\|_2^2 \|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2^2} \leq k_1 \|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2.$$

### 12.4 Proof of Lemma 3

By Lagrangian duality, the regularized network inference problem defined in Eq. 4 is equivalent to the following constrained optimization problem:

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{\alpha}_i} & \ell^n(\boldsymbol{\alpha}_i) \\ \text{subject to} & \alpha_{ji} \geq 0,\, j = 1, \ldots, N, i \neq j, \\ & ||\boldsymbol{\alpha}_i||_1 \leq C(\lambda_n) \end{array} \tag{24}$$

where $C(\lambda_n) < \infty$ is a positive constant. In this alternative formulation, $\lambda_n$ is the Lagrange multiplier for the second constraint. Since $\lambda_n$ is strictly positive, the constraint is active at any optimal solution, and thus $||\boldsymbol{\alpha}_i||_1$ is constant across all optimal solutions.

Using that $\ell^n(\boldsymbol{\alpha}_i)$ is a differentiable convex function by assumption and $\{\boldsymbol{\alpha} : \alpha_{ji} \geq 0, ||\boldsymbol{\alpha}_i||_1 \leq C(\lambda_n)\}$ is a convex set, we have that $\nabla \ell^n(\boldsymbol{\alpha}_i)$ is constant across optimal primal solutions Mangasarian (1988). Moreover, any optimal primal-dual solution in the original problem must satisfy the KKT conditions in the alternative formulation defined by Eq. 24, in particular,

$$\nabla \ell^n(\boldsymbol{\alpha}_i) = -\lambda_n \mathbf{z} + \boldsymbol{\mu},$$

where $\boldsymbol{\mu} \geq 0$ are the Lagrange multipliers associated to the non negativity constraints and $\mathbf{z}$ denotes the subgradient of the $\ell 1$-norm.

Consider the solution $\hat{\boldsymbol{\alpha}}$ such that $||\hat{\mathbf{z}}_{\mathbf{S^c}}||_\infty < 1$ and thus $\nabla_{\alpha_{S^c}} \ell^n(\hat{\boldsymbol{\alpha}}_i) = -\lambda_n \hat{\mathbf{z}}_{\mathbf{S^c}} + \hat{\boldsymbol{\mu}}_{S^c}$. Now, assume there is an optimal primal solution $\tilde{\boldsymbol{\alpha}}$ such that $\tilde{\alpha}_{ji} > 0$ for some $j \in S^c$, then, using that the gradient must be constant across optimal solutions, it should hold that $-\lambda_n \hat{z}_j + \hat{\mu}_j = -\lambda_n$, where $\tilde{\mu}_{ji} = 0$ by complementary slackness, which implies $\hat{\mu}_j = -\lambda_n(1 - \hat{z}_j) < 0$. Since $\hat{\mu}_j \geq 0$ by assumption, this leads to a contradiction. Then, any primal solution $\tilde{\boldsymbol{\alpha}}$ must satisfy $\tilde{\boldsymbol{\alpha}}_{S^c} = 0$ for the gradient to be constant across optimal solutions.

Finally, since $\boldsymbol{\alpha}_{S^c} = 0$ for all optimal solutions, we can consider the restricted optimization problem defined in Eq. 17. If the Hessian sub-matrix $[\nabla^2 L(\hat{\boldsymbol{\alpha}})]_{SS}$ is strictly positive definite, then this restricted optimization problem is strictly convex and the optimal solution must be unique.

### 12.5 Proof of Lemma 4

To prove this lemma, we will first construct a function

$$G(\mathbf{u}_S) := \ell^n(\boldsymbol{\alpha}^*_S + \mathbf{u}_S) - \ell^n(\boldsymbol{\alpha}^*_S) + \lambda_n(\|\boldsymbol{\alpha}^*_S + \mathbf{u}_S\|_1 - \|\boldsymbol{\alpha}^*_S\|_1).$$

whose domain is restricted to the convex set $\mathcal{U} = \{\mathbf{u}_S : \boldsymbol{\alpha}^*_S + \mathbf{u}_S \geq \mathbf{0}\}$. By construction, $G(\mathbf{u}_S)$ has the following properties

1. It is convex with respect to $\mathbf{u}_S$.
2. Its minimum is obtained at $\hat{\boldsymbol{u}}_S := \hat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}^*_S$. That is $G(\hat{\boldsymbol{u}}_S) \leq G(\mathbf{u}_S)$, $\forall \mathbf{u}_S \neq \hat{\boldsymbol{u}}_S$.
3. $G(\hat{\boldsymbol{u}}_S) \leq G(\mathbf{0}) = 0$.

Based on the properties 1 and 3 above, we deduce that any point in the segment, $\mathbb{L} := \{\tilde{\boldsymbol{u}}_S : \tilde{\boldsymbol{u}}_S = t\hat{\boldsymbol{u}}_S + (1-t)\mathbf{0}, t \in [0,1]\}$, connecting $\hat{\boldsymbol{u}}_S$ and $\mathbf{0}$ has $G(\tilde{\boldsymbol{u}}_S) \leq 0$. That is

$$G(\tilde{\boldsymbol{u}}_S) = G(t\hat{\boldsymbol{u}}_S + (1-t)\mathbf{0}) \leq tG(\hat{\boldsymbol{u}}_S) + (1-t)G(\mathbf{0}) \leq 0.$$

Next, we will find a sphere centered at $\mathbf{0}$ with strictly positive radius $B$, $\mathbb{S}(B) := \{\boldsymbol{u}_S : \|\boldsymbol{u}_S\|_2 = B\}$, such that function $G(\boldsymbol{u}_S) > 0$ (strictly positive) on $\mathbb{S}(B)$. We note that this sphere $\mathbb{S}(B)$ can not intersect with the segment $\mathbb{L}$ since the two sets have strictly different function values. Furthermore, the only possible configuration is that the segment is contained inside the sphere entirely, leading us to conclude that the end point $\hat{\boldsymbol{u}}_S := \hat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}^*_S$ is also within the sphere. That is $\|\hat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}^*_S\|_2 \leq B$.

In the following, we will provide details on finding such a suitable $B$ which will be a function of the regularization parameter $\lambda_n$ and the neighborhood size $d$. More specifically, we will start by applying a Taylor series expansion and the mean value theorem,

$$G(\mathbf{u}_S) = \nabla_S \ell^n(\boldsymbol{\alpha}^*_S)^\top \mathbf{u}_S + \mathbf{u}_S^\top \nabla^2_{SS} \ell^n(\boldsymbol{\alpha}^*_S + b\mathbf{u}_S)\mathbf{u}_S + \lambda_n(\|\boldsymbol{\alpha}^*_S + \mathbf{u}_S\|_1 - \|\boldsymbol{\alpha}^*_S\|_1), \tag{25}$$

where $b \in [0, 1]$. We will show that $G(\mathbf{u}_S) > 0$ by bounding below each term of above equation separately.

We bound the absolute value of the first term using the assumption on the gradient, $\nabla_S \ell(\cdot)$,

$$|\nabla_S \ell^n(\boldsymbol{\alpha}^*_S)^\top \mathbf{u}_S| \leq \|\nabla_S \ell\|_\infty \|\mathbf{u}_S\|_1 \leq \|\nabla_S \ell\|_\infty \sqrt{d} \|\mathbf{u}_S\|_2 \leq 4^{-1}\lambda_n B\sqrt{d}. \tag{26}$$

We bound the absolute value of the last term using the reverse triangle inequality.

$$\lambda_n |\|\boldsymbol{\alpha}^*_S + \mathbf{u}_S\|_1 - \|\boldsymbol{\alpha}^*_S\|_1| \leq \lambda_n \|\mathbf{u}_S\|_1 \leq \lambda_n \sqrt{d} \|\mathbf{u}_S\|_2. \tag{27}$$

Bounding the remaining middle term is more challenging. We start by rewriting the Hessian as a sum of two matrices, using Eq. 5,

$$\begin{aligned} q &= \min_{\mathbf{u}_S} \mathbf{u}_S^\top \mathbf{D}^n_{SS}(\boldsymbol{\alpha}^*_S + b\mathbf{u}_S)\mathbf{u}_S + n^{-1}\mathbf{u}_S^\top \mathbf{X}^n_S(\boldsymbol{\alpha}^*_S + b\mathbf{u}_S)\mathbf{X}^n_S(\boldsymbol{\alpha}^*_S + b\mathbf{u}_S)^\top \mathbf{u}_S \\ &= \min_{\mathbf{u}_S} \mathbf{u}_S^\top \mathbf{D}^n_{SS}(\boldsymbol{\alpha}^*_S + b\mathbf{u}_S)\mathbf{u}_S + \|\mathbf{u}_S^\top \mathbf{X}^n_S(\boldsymbol{\alpha}^*_S + b\mathbf{u}_S)\|_2^2. \end{aligned}$$

Now, we introduce two additional quantities,

$$\Delta \mathbf{D}^n_{SS} = \mathbf{D}^n_{SS}(\boldsymbol{\alpha}^*_S + b\mathbf{u}_S) - \mathbf{D}^n_{SS}(\boldsymbol{\alpha}^*_S) \quad \text{and} \quad \Delta \mathbf{X}^n_S = \mathbf{X}^n_S(\boldsymbol{\alpha}^*_S + b\mathbf{u}_S) - \mathbf{X}^n_S(\boldsymbol{\alpha}^*_S),$$

and rewrite $q$ as

$$\begin{aligned} q = \min_{\mathbf{u}_S} \big[ &\mathbf{u}_S^\top \mathbf{D}^n_{SS}(\boldsymbol{\alpha}^*_S)\mathbf{u}_S + n^{-1}\|\mathbf{u}_S^\top \mathbf{X}^n_S(\boldsymbol{\alpha}^*_S)\|_2^2 + n^{-1}\|\mathbf{u}_S^\top \Delta\mathbf{X}^n_S\|_2^2 + \mathbf{u}_S^\top \Delta\mathbf{D}^n_{SS}\mathbf{u}_S \\ &+ 2n^{-1}\langle \mathbf{u}_S^\top \mathbf{X}^n_S(\boldsymbol{\alpha}^*_S), \mathbf{u}_S^\top \Delta\mathbf{X}^n_S\rangle \big]. \end{aligned}$$

Next, we use dependency condition,

$$q \geq C_{\min} B^2 - \max_{\mathbf{u}_S} |\underbrace{\mathbf{u}_S^\top \Delta\mathbf{D}^n_{SS}\mathbf{u}_S}_{T_1}| - \max_{\mathbf{u}_S} 2|\underbrace{n^{-1}\langle \mathbf{u}_S^\top \mathbf{X}^n_S(\boldsymbol{\alpha}^*_S), \mathbf{u}_S^\top \Delta\mathbf{X}^n_S\rangle}_{T_2}|,$$

and proceed to bound $T_1$ and $T_2$ separately. First, we bound $T_1$ using the Lipschitz condition,

$$|T_1| = |\sum_{k\in S} u_k^2 [\mathbf{D}^n_k(\boldsymbol{\alpha}^*_S + b\mathbf{u}_S) - \mathbf{D}^n_k(\boldsymbol{\alpha}^*_S)]| \leq \sum_{k\in S} u_k^2 k_2 \|b\mathbf{u}_S\|_2 \leq k_2 B^3.$$

Then, we use the dependency condition, the Lipschitz condition and the Cauchy-Schwartz inequality to bound $T_2$,

$$\begin{aligned} T_2 &\leq \frac{1}{\sqrt{n}}\|\mathbf{u}_S^\top \mathbf{X}_S^n(\boldsymbol{\alpha}_S^*)\|_2 \frac{1}{\sqrt{n}}\|\mathbf{u}_S^\top \Delta\mathbf{X}_S^n\|_2 \leq \sqrt{C_{\max}} B \frac{1}{\sqrt{n}}\|\mathbf{u}_S^\top \Delta\mathbf{X}_S^n\|_2 \\ &\leq \sqrt{C_{\max}} B \|\mathbf{u}_S\|_2 \frac{1}{\sqrt{n}}|||\Delta\mathbf{X}_S^n|||_2 \leq \sqrt{C_{\max}} B^2 k_1 \|b\mathbf{u}_S\|_2 \\ &\leq k_1\sqrt{C_{\max}} B^3, \end{aligned}$$

where we note that applying the Lipschitz condition implies assuming $B < \frac{\alpha_{min}}{2}$. Next, we incorporate the bounds of $T_1$ and $T_2$ to lower bound $q$,

$$q \geq C_{\min} B^2 - (k_2 + 2k_1\sqrt{C_{\max}})B^3. \tag{28}$$

Now, we set $B = K\lambda_n\sqrt{d}$, where $K$ is a constant that we will set later in the proof, and select the regularization parameter $\lambda_n$ to satisfy

$$\lambda_n\sqrt{d} \leq \frac{C_{\min}}{2K(k_2 + 2k_1\sqrt{C_{\max}})}. \tag{29}$$

Then,

$$\begin{aligned} G(\mathbf{u}_S) &\geq -4^{-1}\lambda_n\sqrt{d}B + 0.5C_{\min}B^2 - \lambda_n\sqrt{d}B \geq B(0.5C_{\min}B - 1.25\lambda_n\sqrt{d}) \\ &\geq B(0.5C_{\min}K\lambda_n\sqrt{d} - 1.25\lambda_n\sqrt{d}). \end{aligned}$$

In the last step, we set the constant $K = 3C_{\min}^{-1}$, and we have

$$G(\mathbf{u}_S) \geq 0.25\lambda_n\sqrt{d} > 0,$$

as long as

$$\begin{aligned} \sqrt{d}\lambda_n &\leq \frac{C_{\min}^2}{6(k_2 + 2k_1\sqrt{C_{\max}})} \\ \alpha_{\min}^* &\geq \frac{6\lambda_n\sqrt{d}}{C_{\min}}. \end{aligned}$$

Finally, convexity of $G(\mathbf{u}_S)$ yields

$$\|\hat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_S^*\|_2 \leq 3\lambda_n\sqrt{d}/C_{\min} \leq \frac{\alpha_{\min}^*}{2}.$$

### 12.6 Proof of Lemma 5

Define $z_j^c = [\nabla g(\mathbf{t}^c; \boldsymbol{\alpha}^*)]_j$ and $z_j = \frac{1}{n}\sum_c z_j^c$. Now, using the KKT conditions and condition 4 (Boundedness), we have that $\mu_j^* = \mathbb{E}_c\{z_j^c\}$ and $|z_j^c| \leq k_3$, respectively. Thus, Hoeffding's inequality yields

$$P\left(|z_j - \mu_j^*| > \frac{\lambda_n\varepsilon}{4(2-\varepsilon)}\right) \leq 2\exp\left(-\frac{n\lambda_n^2\varepsilon^2}{32k_3^2\,(2-\varepsilon)^2}\right),$$

and then,

$$P\left(\|\boldsymbol{z} - \boldsymbol{\mu}^*\|_\infty > \frac{\lambda_n\varepsilon}{4(2-\varepsilon)}\right) \leq 2\exp\left(-\frac{n\lambda_n^2\varepsilon^2}{32k_3^2\,(2-\varepsilon))^2} + \log p\right).$$

### 12.7 Proof of Lemma 6

We start by factorizing the Hessian matrix, using Eq. 5,

$$R_j^n = \left[\nabla^2 \ell^n(\bar{\boldsymbol{\alpha}}_j) - \nabla^2 \ell^n(\boldsymbol{\alpha}^*)\right]_j^\top (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) = \omega_j^n + \delta_j^n,$$

where,

$$\begin{aligned}
\omega_j^n &= \left[\mathbf{D}^n(\bar{\boldsymbol{\alpha}}_j) - \mathbf{D}^n(\boldsymbol{\alpha}^*)\right]_j^\top (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) \\
\delta_j^n &= \frac{1}{n} \boldsymbol{V}_j^n (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) \\
\boldsymbol{V}_j^n &= [\mathbf{X}^n(\bar{\boldsymbol{\alpha}}_j)]_j \mathbf{X}^n(\bar{\boldsymbol{\alpha}}_j)^\top - [\mathbf{X}^n(\boldsymbol{\alpha}^*)]_j \mathbf{X}^n(\boldsymbol{\alpha}^*)^\top.
\end{aligned}$$

Next, we proceed to bound each term separately. Since $[\bar{\boldsymbol{\alpha}}_j]_S = \theta_j \hat{\boldsymbol{\alpha}}_S + (1-\theta_j)\boldsymbol{\alpha}_S^*$ where $\theta_j \in [0,1]$, and $\|\hat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_S^*\|_\infty \leq \frac{\alpha_{\min}^*}{2}$ (Lemma 4), it holds that $[\bar{\boldsymbol{\alpha}}_j]_S \geq \frac{\alpha_{\min}^*}{2}$. Then, we can use condition 3 (Lipschitz Continuity) to bound $\omega_j^n$.

$$|\omega_j^n| \leq k_1 \|\bar{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*\|_2 \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2 \leq k_1 \theta_j \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2^2 \leq k_1 \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2^2.$$

However, bounding term $\delta_j^n$ is more difficult. Let us start by rewriting $\delta_j^n$ as follows.

$$\delta_j^n = (\boldsymbol{\Lambda}_1 + \boldsymbol{\Lambda}_2 + \boldsymbol{\Lambda}_3)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*),$$

where,

$$\begin{aligned}
\boldsymbol{\Lambda}_1 &= [\mathbf{X}^n(\boldsymbol{\alpha}^*)]_j (\mathbf{X}^n(\bar{\boldsymbol{\alpha}}_j)^\top - \mathbf{X}^n(\boldsymbol{\alpha}^*)^\top) \\
\boldsymbol{\Lambda}_2 &= \{[\mathbf{X}^n(\bar{\boldsymbol{\alpha}}_j)]_j - [\mathbf{X}^n(\boldsymbol{\alpha}^*)]_j\} (\mathbf{X}^n(\bar{\boldsymbol{\alpha}}_j)^\top - \mathbf{X}^n(\boldsymbol{\alpha}^*)^\top) \\
\boldsymbol{\Lambda}_3 &= \left([\mathbf{X}^n(\bar{\boldsymbol{\alpha}}_j)]_j - [\mathbf{X}^n(\boldsymbol{\alpha}^*)]_j\right) \mathbf{X}^n(\boldsymbol{\alpha}^*)^\top.
\end{aligned}$$

Next, we bound each term separately. For the first term, we first apply Cauchy inequality,

$$|\boldsymbol{\Lambda}_1(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)| \leq \|[\mathbf{X}^n(\boldsymbol{\alpha}^*)]_j\|_2 \times |||\mathbf{X}^n(\bar{\boldsymbol{\alpha}}_j)^\top - \mathbf{X}^n(\boldsymbol{\alpha}^*)^\top|||_2 \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2,$$

and then use condition 3 (Lipschtiz Continuity) and 4 (Boundedness),

$$|\boldsymbol{\Lambda}_1(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)| \leq n k_4 k_1 \|\bar{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*\|_2 \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2 \leq n k_4 k_1 \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2^2.$$

For the second term, we also start by applying Cauchy inequality,

$$|\boldsymbol{\Lambda}_2(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)| \leq \|[\mathbf{X}^n(\bar{\boldsymbol{\alpha}}_j)]_j - [\mathbf{X}^n(\boldsymbol{\alpha}^*)]_j\|_2 \times |||\mathbf{X}^n(\bar{\boldsymbol{\alpha}}_j)^\top - \mathbf{X}^n(\boldsymbol{\alpha}^*)^\top|||_2 \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2,$$

and then use condition 3 (Lipschtiz Continuity),

$$|\boldsymbol{\Lambda}_2(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)| \leq n k_1^2 \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2^2.$$

Last, for third term, once more we start by applying Cauchy inequality,

$$|\boldsymbol{\Lambda}_3(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)| \leq \|[\mathbf{X}^n(\bar{\boldsymbol{\alpha}}_j)]_j - [\mathbf{X}^n(\boldsymbol{\alpha}^*)]_j\|_2 \times |||\mathbf{X}^n(\boldsymbol{\alpha}^*)^\top|||_2 \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2,$$

and then apply condition 1 (Dependency Condition) and condition 3 (Lipschitz Continuity),

$$|\boldsymbol{\Lambda}_3(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)| \leq n k_1 \sqrt{C_{\max}} \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2^2$$

Now, we combine the bounds,

$$\|\boldsymbol{R}^n\|_\infty \leq K \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2^2,$$

where

$$K = k_1 + k_4 k_1 + k_1^2 + k_1 \sqrt{C_{\max}}.$$

Finally, using Lemma 4 and selecting the regularization parameter $\lambda_n$ to satisfy $\lambda_n d \leq C_{\min}^2 \frac{\varepsilon}{36K(2-\varepsilon)}$ yields:

$$\frac{\|\boldsymbol{R}^n\|_\infty}{\lambda_n} \leq \frac{3K\lambda_n d}{C_{\min}^2} \leq \frac{\varepsilon}{4(2-\varepsilon)}$$

### 12.8 Proof of Lemma 7

We will first bound the difference in terms of nuclear norm between the population Fisher information matrix $\mathcal{Q}_{SS}$ and the sample mean cascade log-likelihood $\mathcal{Q}_{SS}^n$. Define $z_{jk}^c = [\nabla^2 g(\mathbf{t}^c; \boldsymbol{\alpha}^*) - \nabla^2 \ell^n(\boldsymbol{\alpha}^*)]_{jk}$ and $z_{jk} = \frac{1}{n}\sum_{c=1}^n z_{jk}^c$. Then, we can express the difference between the population Fisher information matrix $\mathcal{Q}_{SS}$ and the sample mean cascade log-likelihood $\mathcal{Q}_{SS}^n$ as:

$$|||\mathcal{Q}_{SS}^n(\boldsymbol{\alpha^*}) - \mathcal{Q}_{SS}^*(\boldsymbol{\alpha^*})|||_2 \leq |||\mathcal{Q}_{SS}^n(\boldsymbol{\alpha^*}) - \mathcal{Q}_{SS}^*(\boldsymbol{\alpha^*})|||_F = \sqrt{\sum_{j=1}^d \sum_{k=1}^d (z_{ik})^2}.$$

Since $|z_{jk}^{(c)}| \leq 2k_5$ by condition 4, we can apply Hoeffding's inequality to each $z_{jk}$,

$$P(|z_{jk}| \geq \beta) \leq 2\exp\left(-\frac{\beta^2 n}{8k_5^2}\right), \tag{30}$$

and further,

$$P(|||\mathcal{Q}_{SS}^n(\boldsymbol{\alpha^*}) - \mathcal{Q}_{SS}^*(\boldsymbol{\alpha^*})|||_2 \geq \delta) \leq 2\exp\big(-K\frac{\delta^2 n}{d^2} + 2\log d\big) \tag{31}$$

where $\beta^2 = \delta^2/d^2$. Now, we bound the maximum eigenvalue of $\mathcal{Q}_{SS}^n$ as follows:

$$\begin{aligned}\Lambda_{\max}(\mathcal{Q}_{SS}^n) &= \max_{\|x\|_2=1} x^\top \mathcal{Q}_{SS}^n x = \max_{\|x\|_2=1} \{x^\top \mathcal{Q}_{SS}^* x + x^\top(\mathcal{Q}_{SS}^n - \mathcal{Q}_{SS}^*)x\} \\ &\leq y^\top \mathcal{Q}_{SS}^* y + y^\top(\mathcal{Q}_{SS}^n - \mathcal{Q}_{SS}^*)y,\end{aligned}$$

where $y$ is unit-norm maximal eigenvector of $\mathcal{Q}_{SS}^*$. Therefore,

$$\Lambda_{\max}(\mathcal{Q}_{SS}^n) \leq \Lambda_{\max}(\mathcal{Q}_{SS}^*) + |||\mathcal{Q}_{SS}^n - \mathcal{Q}_{SS}^*|||_2,$$

and thus,

$$P\big(\Lambda_{\max}(\mathcal{Q}_{SS}^n) \geq C_{\max} + \delta\big) \leq \exp\left(-K\frac{\delta^2 n}{d^2} + 2\log d\right).$$

Reasoning in a similar way, we bound the minimum eigenvalue of $\mathcal{Q}_{SS}^n$:

$$P\big(\Lambda_{\min}(\mathcal{Q}_{SS}^n) \leq C_{\min} - \delta\big) \leq \exp\left(-K\frac{\delta^2 n}{d^2} + 2\log d\right)$$

### 12.9 Proof of Lemma 8

We start by decomposing $\mathcal{Q}_{S^cS}^n(\boldsymbol{\alpha^*})(\mathcal{Q}_{S^cS}^n(\boldsymbol{\alpha^*}))^{-1}$ as follows:

$$\mathcal{Q}_{S^cS}^n(\boldsymbol{\alpha^*})(\mathcal{Q}_{S^cS}^n(\boldsymbol{\alpha^*}))^{-1} = A_1 + A_2 + A_3 + A_4,$$

where,

$$A_1 = \mathcal{Q}^*_{S^cS}[(\mathcal{Q}^n_{S^cS})^{-1} - (\mathcal{Q}^*_{S^cS})^{-1}],$$
$$A_2 = [\mathcal{Q}^n_{S^cS} - \mathcal{Q}^*_{S^cS}][(\mathcal{Q}^n_{S^cS})^{-1} - (\mathcal{Q}^*_{S^cS})^{-1}]$$
$$A_3 = [\mathcal{Q}^n_{S^cS} - \mathcal{Q}^*_{S^cS}](\mathcal{Q}^*_{SS})^{-1},$$
$$A_4 = \mathcal{Q}^*_{S^cS}(\mathcal{Q}^*_{SS})^{-1},$$

$\mathcal{Q}^* = \mathcal{Q}^*(\boldsymbol{\alpha^*})$ and $\mathcal{Q}^n = \mathcal{Q}^n(\boldsymbol{\alpha^*})$. Now, we bound each term separately. The fourth term, $A_4$, is the easiest to bound, using simply the incoherence condition:

$$|||A_4|||_\infty \leq 1 - \varepsilon.$$

To bound the other terms, we need the following lemma:

**Lemma 11** *For any $\delta \geq 0$ and constants $K$ and $K'$, the following bounds hold:*

$$P[|||\mathcal{Q}^n_{S^cS} - \mathcal{Q}^*_{S^cS}|||_\infty \geq \delta] \leq 2\exp\left(-K\frac{n\delta^2}{d^2} + \log d + log(p-d)\right) \tag{32}$$

$$P[|||\mathcal{Q}^n_{SS} - \mathcal{Q}^*_{SS}|||_\infty \geq \delta] \leq 2\exp\left(-K\frac{n\delta^2}{d^2} + 2\log d\right) \tag{33}$$

$$P[|||(\mathcal{Q}^n_{SS})^{-1} - (\mathcal{Q}^*_{SS})^{-1}|||_\infty \geq \delta] \leq 4\exp\left(-K\frac{n\delta}{d^3} - K'\log d\right) \tag{34}$$

**Proof** We start by proving the first confidence interval. By definition of infinity norm of a matrix, we have:

$$P[|||\mathcal{Q}^n_{S^cS} - \mathcal{Q}^*_{S^cS}|||_\infty \geq \delta] = P\big[\max_{j\in S^c}\sum_{k\in S}|z_{jk}| \geq \delta\big] \leq (p-d)P\big[\sum_{k\in S}|z_{jk}| \geq \delta\big],$$

where $z_{jk} = [\mathcal{Q}^n - \mathcal{Q}^*]_{jk}$ and, for the last inequality, we used the union bound and the fact that $|S^c| \leq p - d$. Furthermore,

$$P\big[\sum_{k\in S}|z_{jk}| \geq \delta\big] \leq P[\exists k \in S||z_{jk}| \geq \delta/d] \leq dP[|z_{jk}| \geq \delta/d].$$

Thus,

$$P[|||\mathcal{Q}^n_{S^cS} - \mathcal{Q}^*_{S^cS}|||_\infty \geq \delta] \leq (p-d)dP[|z_{jk}| \geq \delta/d].$$

At this point, we can obtain the first confidence bound by using Eq. 30 with $\beta = \delta/d$ in the above equation. The proof of the second confidence bound is very similar and we omit it for brevity. To prove the last confidence bound, we proceed as follows:

$$\begin{aligned}
|||(\mathcal{Q}^n_{SS})^{-1} - (\mathcal{Q}^*_{SS})^{-1}|||_\infty &= |||(\mathcal{Q}^n_{SS})^{-1}[\mathcal{Q}^n_{SS} - \mathcal{Q}^*_{SS}](\mathcal{Q}^*_{SS})^{-1}|||_\infty \\
&\leq \sqrt{d}|||(\mathcal{Q}^n_{SS})^{-1}[\mathcal{Q}^n_{SS} - \mathcal{Q}^*_{SS}](\mathcal{Q}^*_{SS})^{-1}|||_2 \\
&\leq \sqrt{d}|||(\mathcal{Q}^n_{SS})^{-1}|||_2|||\mathcal{Q}^n_{SS} - \mathcal{Q}^*_{SS}|||_2|||(\mathcal{Q}^*_{SS})^{-1}|||_2 \\
&\leq \frac{\sqrt{d}}{C_{\min}}|||\mathcal{Q}^n_{SS} - \mathcal{Q}^*_{SS}|||_2|||(\mathcal{Q}^n_{SS})^{-1}|||_2.
\end{aligned}$$

Next, we bound each term of the final expression in the above equation separately. The first term can be bounded using Eq. 31:

$$P\left[|||\mathcal{Q}^n_{SS} - \mathcal{Q}^*_{SS}|||_2 \geq \frac{C^2_{min}\delta}{2\sqrt{d}}\right] \leq 2\exp\left(-K\frac{n\delta^2}{d^3} + 2\log d\right),$$

The second term can be bounded using Lemma 6:

$$P\left[|||(\mathcal{Q}^n_{SS})^{-1}|||_2 \geq \frac{2}{C_{\min}}\right] = P\left[\Lambda_{\min}(\mathcal{Q}^n_{SS}) \leq \frac{C_{\min}}{2}\right] \leq \exp\left(-K\frac{n}{d^2} + B\log d\right).$$

Then, the third confidence bound follows. ■

*Control of* $A_1$. We start by rewriting the term $A_1$ as

$$A_1 = \mathcal{Q}^*_{S^cS}(\mathcal{Q}^*_{SS})^{-1}[(\mathcal{Q}^*_{SS}) - (\mathcal{Q}^n_{SS})](\mathcal{Q}^n_{SS})^{-1},$$

and further,

$$|||A_1|||_\infty \leq |||\mathcal{Q}^*_{S^cS}(\mathcal{Q}^*_{SS})^{-1}|||_\infty \times |||(\mathcal{Q}^*_{SS}) - (\mathcal{Q}^n_{SS})|||_\infty |||(\mathcal{Q}^n_{SS})^{-1}|||_\infty.$$

Next, using the incoherence condition easily yields:

$$|||A_1|||_\infty \leq (1-\varepsilon)|||(\mathcal{Q}^*_{SS}) - (\mathcal{Q}^n_{SS})|||_\infty \times \sqrt{d}|||(\mathcal{Q}^n_{SS})^{-1}|||_2$$

Now, we apply Lemma 6 with $\delta = C_{\min}/2$ to have that $|||(\mathcal{Q}^n_{SS})^{-1}|||_2 \leq \frac{2}{C_{\min}}$ with probability greater than $1 - \exp(-Kn/d^2 + K'\log d)$, and then use Eq. 34 with $\delta = \frac{\varepsilon C_{\min}}{12\sqrt{d}}$ to conclude that

$$P\left[|||A_1|||_\infty \geq \frac{\varepsilon}{6}\right] \leq 2\exp\left(-K\frac{n}{d^3} + K'\log d\right).$$

*Control of* $A_2$. We rewrite the term $A_2$ as

$$|||A_2|||_\infty \leq |||\mathcal{Q}^n_{S^cS} - \mathcal{Q}^*_{S^cS}|||_\infty |||(\mathcal{Q}^n_{SS})^{-1} - (\mathcal{Q}^*_{SS})^{-1}|||_\infty,$$

and then use Eqs. 32 and 33 with $\delta = \sqrt{\varepsilon/6}$ to conclude that

$$P\left[|||A_2|||_\infty \geq \frac{\varepsilon}{6}\right] \leq 4\exp\left(-K\frac{n}{d^3} + \log(p-d) + K'\log p\right).$$

*Control of* $A_3$. We rewrite the term $A_3$ as

$$|||A_3|||_\infty = \sqrt{d}|||(\mathcal{Q}^*_{SS})^{-1}|||_2 |||\mathcal{Q}^n_{S^cS} - \mathcal{Q}^*_{S^cS}|||_\infty \leq \frac{\sqrt{d}}{C_{\min}}|||\mathcal{Q}^n_{S^cS} - \mathcal{Q}^*_{S^cS}|||_\infty.$$

We then apply Eq. 32 with $\delta = \frac{\varepsilon C_{\min}}{6\sqrt{d}}$ to conclude that

$$P\left[|||A_3|||_\infty \geq \frac{\varepsilon}{6}\right] \leq \exp\left(-K\frac{n}{d^3} + \log(p-d)\right),$$

and thus,

$$P\left[|||\mathcal{Q}^n_{S^cS}(\mathcal{Q}^n_{SS})^{-1}|||_\infty \geq 1 - \frac{\varepsilon}{2}\right] = \mathcal{O}\left(\exp(-K\frac{n}{d^3} + \log p)\right).$$

## References

B. Abrahao, F. Chierichetti, R. Kleinberg, and A. Panconesi. Trace complexity of network inference. In *KDD*, 2013.

E. Adar and L. A. Adamic. Tracking Information Epidemics in Blogspace. In *Web Intelligence*, pages 207–214, 2005.

A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286: 509–512, 1999.

A. Beck and M. Teboulle. Gradient-based algorithms with applications to signal recovery. *Convex Optimization in Signal Processing and Communications*, 2009.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

N. Du, L. Song, A. Smola, and M. Yuan. Learning Networks of Heterogeneous Influence. In *NIPS*, 2012a.

N. Du, L. Song, H. Woo, and H. Zha. Uncover Topic-Sensitive Information Diffusion Networks. In *AISTATS*, 2012b.

Nan Du, Le Song, Hyenkyun Woo, and Hongyuan Zha. Uncover topic-sensitive information diffusion networks. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 229–237, 2013.

M. Gomez-Rodriguez. *Ph.D. Thesis*. Stanford University & MPI for Intelligent Systems, 2013.

M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring Networks of Diffusion and Influence. In *KDD*, 2010.

M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the Temporal Dynamics of Diffusion Networks. In *ICML*, 2011.

M. Gomez-Rodriguez, J. Leskovec, and B. Schölkopf. Structure and Dynamics of Information Pathways in On-line Media. In *WSDM*, 2013a.

M. Gomez-Rodriguez, J. Leskovec, and B. Schölkopf. Modeling Information Propagation with Survival Theory. In *ICML '13: Proceedings of the 30th International Conference on Machine Learning*, 2013b.

M. Gomez-Rodriguez, J. Leskovec, D. Balduzzi, and B. Schölkopf. Uncovering the Structure and Temporal Dynamics of Information Propagation. *Network Science*, 2014.

V. Gripon and M. Rabbat. Reconstructing a graph from path traces. *arXiv:1301.6916*, 2013.

Martin Jaggi. Revisiting {Frank-Wolfe}: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 427–435, 2013.

D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the Spread of Influence Through a Social Network. In *KDD*, 2003.

J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker Graphs: An Approach to Modeling Networks. *JMLR*, 2010.

O. L. Mangasarian. A simple characterization of solution sets of convex programs. *Operations Research Letters*, 7(1):21–26, 1988.

S. Myers and J. Leskovec. Clash of the contagions: Cooperation and competition in information diffusion. In *Proceedings of the IEEE International Conference on Data Mining*, 2012.

S. Myers, J. Leskovec, and C. Zhu. Information Diffusion and External Influence in Networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012.

Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of $m$-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization*, 19(4):1574, 2009.

P. Netrapalli and S. Sanghavi. Finding the Graph of Epidemic Cascades. In *ACM SIGMETRICS*, 2012.

W. K. Newey and D. L. McFadden. Large Sample Estimation and Hypothesis Testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. 1994.

N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 2013.

B.A. Prakash, A. Beutel, R. Rosenfeld, and C. Faloutsos. Winner Takes All: Competing Viruses or Ideas on Fair-Play Networks. In *Proceedings of the 21st International Conference on World Wide Web*, pages 1037–1046, 2012.

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

E. M. Rogers. *Diffusion of Innovations*. Free Press, New York, fourth edition, 1995.

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

T. Snowsill, N. Fyson, T. De Bie, and N. Cristianini. Refining Causality: Who Copied From Whom? In *KDD*, 2011.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

L. Wang, S. Ermon, and J. Hopcroft. Feature-enhanced probabilistic models for diffusion network inference. In *ECML PKDD*, 2012.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 641–649, 2013.