
Estimating Diffusion Network Structures: Recovery Conditions, Sample Complexity & Soft-thresholding Algorithm

Hadi Daneshmand¹
Manuel Gomez-Rodriguez¹
Le Song²
Bernhard Schölkopf¹

HADI.DANESHMAND@TUE.MPG.DE
MANUELGR@TUE.MPG.DE
LSONG@CC.GATECH.EDU
BS@TUE.MPG.DE

¹MPI for Intelligent Systems and ²Georgia Institute of Technology

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 ℓ_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 inference algorithm, which we use to illustrate the consequences of our theoretical results, and show that our framework outperforms other alternatives in practice.

1. Introduction

Diffusion of information, behaviors, diseases, or more 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 & 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; Snowsill et al., 2011; Du et al., 2012a), 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 a paucity of work along this direction (Netrapalli & Sanghavi, 2012; Abrahao et al., 2013) which provide only partial views of the problem. None of them is able to identify the recovery

condition relating to the interaction between the network structure and the cascade sampling process, which we will make precise in our paper.

Overview of results. We consider the network inference problem under the continuous-time diffusion model recently introduced by [Gomez-Rodriguez et al. \(2011\)](#). We identify a natural incoherence 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 ℓ_1 -regularized maximum likelihood estimator and $O(d^3 \log N)$ cascades, and the probability of success is 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 ℓ_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.

Related work. [Netrapalli & 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 is rather restricted: essentially, 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](#)).

[Abraham 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 only in

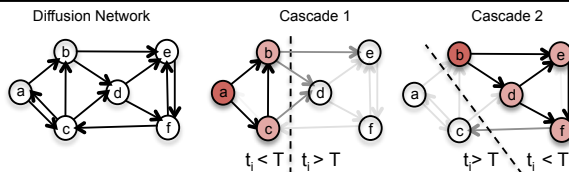


Figure 1. The diffusion network structure (left) is unknown and we only observe cascades, which are N -dimensional vectors recording the times when nodes get infected by contagions that spread (right). Cascade 1 is $(t_a, t_b, t_c, \infty, \infty, \infty)$, where $t_a < t_c < t_b$, and cascade 2 is $(\infty, t_b, \infty, t_d, t_e, t_f)$, 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.

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 incoherence condition, our method outperforms the First-Edge algorithm and the algorithm for bounded-degree networks in terms of rate and sample complexity.

[Gripone & 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, unrealistic 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 $j \rightarrow i$ with a transmission function, $f(t_i | t_j; \alpha_{ji}) = f(t_i - t_j; \alpha_{ji})$, a density over time parameterized by α_{ji} . This is 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* 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, $\tau = t_j - t_i$, drawn from an associated transmission function $f(\tau; \alpha_{ji})$. We assume transmission times are independent, possibly dis-

Table 1. Functions.

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

tributed differently across edges, and, in some cases, can be arbitrarily large, $\tau \rightarrow \infty$. 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*. Figure 1 illustrates the process.

Observations from the model are recorded as a set C^n of cascades $\{\mathbf{t}^1, \dots, \mathbf{t}^n\}$. Each cascade \mathbf{t}^c is an N -dimensional vector $\mathbf{t}^c := (t_1^c, \dots, t_N^c)$ recording when nodes are infected, $t_k^c \in [0, T^c] \cup \{\infty\}$. Symbol ∞ 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 cascade. We assume $T^c = T$ for all cascades; the results generalize trivially.

2.2. 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

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

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. 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 $h(\mathbf{t}; \boldsymbol{\alpha}_i) := \sum_{k:t_k \leq t_i} H(t_i|t_k; \alpha_{ki})$ 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 $\{\alpha_{ji}^*\}$. 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 $\mathbb{P}(s)$. 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)

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

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

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

where $\boldsymbol{\alpha}_i := \{\alpha_{ji} \mid j = 1, \dots, N, i \neq j\}$ 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 $\boldsymbol{\alpha}_i$ (also see Table 1 for the definition of $g(\cdot; \boldsymbol{\alpha}_i)$). 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 $\mathcal{V} \setminus \mathcal{V}_i$ 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 $n \rightarrow \infty$, 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 ℓ_1 -regularizer to the objective function and solving instead the following optimization problem

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

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 λ_n , the solution of Eq. 4 successfully re-

covers 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 $\alpha = \alpha_i$, $\alpha_j = \alpha_{ji}$, $\mathcal{N}^- = \mathcal{N}^-(i)$, $\mathcal{R} = \mathcal{R}_i$, $\mathcal{U} = \mathcal{U}_i$, $d = d_i$, $p_i = p$ 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 α if the survival functions are log-concave and the hazard functions are concave in α . Under these conditions, the Hessian matrix, $\mathcal{Q}^n = \nabla^2 \ell^n(\alpha)$, can be expressed as the sum of a nonnegative diagonal matrix D^n and the outer product of a matrix $X^n(\alpha)$ with itself, *i.e.*,

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

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

$$X^n(\alpha) = [X(\mathbf{t}^1; \alpha) | X(\mathbf{t}^2; \alpha) | \dots | X(\mathbf{t}^n; \alpha)], \quad (6)$$

with each column $X(\mathbf{t}^c; \alpha) := h(\mathbf{t}^c; \alpha)^{-1} \nabla_\alpha h(\mathbf{t}^c; \alpha)$. Intuitively, the Hessian matrix captures the co-occurrence information of nodes in cascades. Then, we can prove

Theorem 1 *If the source probability $\mathbb{P}(s)$ is strictly positive for all $s \in \mathcal{R}$, then, the maximum likelihood estimator $\hat{\alpha}$ 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 & McFadden, 1994). Continuity is obvious. For compactness, since $L \rightarrow -\infty$ for both $\alpha_{ij} \rightarrow 0$ and $\alpha_{ij} \rightarrow \infty$ for all i, j so we lose nothing imposing upper and lower bounds thus restricting to a compact subset. For the identification condition, $\alpha \neq \alpha^* \Rightarrow \ell^n(\alpha) \neq \ell^n(\alpha^*)$, we use Lemma 9 and 10 (refer to Appendices A and B), which establish that $X^n(\alpha)$ has full row rank as $n \rightarrow \infty$, 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 incoherence 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, $\mathbb{E}_c [\ell^n(\alpha)] = \mathbb{E}_c [\log g(\mathbf{t}^c; \alpha)]$, 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 $\mathbb{E}_c [\log g(\mathbf{t}^c; \alpha)]$ evaluated at the true model parameter α^* as \mathcal{Q}^* . Then, we place two conditions on the Lipschitz continuity of $X(\mathbf{t}^c; \alpha)$, and the boundedness of $X(\mathbf{t}^c; \alpha^*)$ and $\nabla g(\mathbf{t}^c; \alpha^*)$ at the true model parameter α^* . 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 α_S^* the set of parameters indexed by S .

Condition 1 (Dependency condition): There exists constants $C_{min} > 0$ and $C_{max} > 0$ such that $\Lambda_{min}(\mathcal{Q}_{SS}^*) \geq C_{min}$ and $\Lambda_{max}(\mathcal{Q}_{SS}^*) \leq C_{max}$ 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 (Incoherence condition): There exists $\varepsilon \in (0, 1]$ such that $\| \mathcal{Q}_{S^c S}^* (\mathcal{Q}_{SS}^*)^{-1} \|_\infty \leq 1 - \varepsilon$, 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.

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

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

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

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

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

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

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 α^* , is bounded,

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

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 incoherence 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 $T \rightarrow \infty$. Then, it is easy to show that the incoherence 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 incoherence 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 $T \rightarrow \infty$. Then, it can be shown that the incoherence 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 incoherence condition reduces to the inequalities $\left(1 - \frac{\alpha_{0j}}{\alpha_{0i} + \alpha_{0j}}\right) e^{-(\alpha_{0i} + \alpha_{0j})T} + \frac{\alpha_{0j}}{\alpha_{0i} + \alpha_{0j}} < 1 - \varepsilon(1 + e^{-\alpha_{0i}T})$, $j = 1, \dots, p : j \neq i$, which always holds for some $\varepsilon > 0$. If $T \rightarrow \infty$, 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 ε for which the incoherence condition holds. To summarize, as long as $P_0 > 0$, there is always some $\varepsilon > 0$ for which the condition holds, and such ε value depends on the time window and the parameters α_{0j} .

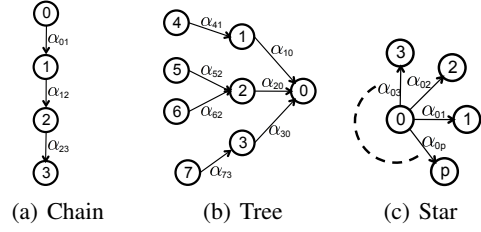


Figure 2. Example networks.

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 α_{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 λ_n is selected to satisfy

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

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

$$n > Ld^3 \log p, \quad (9)$$

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 ℓ_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 recove-

ring 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 sparsity 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 D):

Lemma 3 *Suppose that there exists an optimal primal-dual solution $(\hat{\alpha}, \hat{\mu})$ 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 $\tilde{\alpha}$ must have $\tilde{\alpha}_{S^c} = 0$. Moreover, if the Hessian sub-matrix \mathcal{Q}_{SS}^n is strictly positive definite, then $\hat{\alpha}$ is the unique optimal solution.*

Next, we will construct a primal-dual vector $(\hat{\alpha}, \hat{\mu})$ 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 α^* , i.e.,

$$\hat{\alpha}_j > 0, \forall j : \alpha_j^* > 0, \quad (10)$$

$$\hat{\alpha}_j = 0, \forall j : \alpha_j^* = 0. \quad (11)$$

Then, based on Lemma 3, we can deduce that the optimal solution to Eq. 4 correctly recovers the sparsity pattern of α^* , and thus the incoming edges to node i .

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

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

$$\tilde{\mu}_j \tilde{\alpha}_j = 0, \quad (13)$$

$$\tilde{\mu}_j \geq 0, \quad (14)$$

$$\tilde{z}_j = 1, \forall \tilde{\alpha}_j > 0, \quad (15)$$

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

where $\ell^n(\tilde{\alpha}) = -\frac{1}{n} \sum_{c \in C^n} \log g(\mathbf{t}^c; \tilde{\alpha})$ and $\tilde{\mathbf{z}}$ denotes the subgradient of the ℓ_1 -norm.

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

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

maximum likelihood problem

$$\hat{\alpha}_S = \underset{(\alpha_S, 0), \alpha_S \geq 0}{\operatorname{argmin}} \{ \ell^n(\alpha) + \lambda_n \|\alpha_S\|_1 \}. \quad (17)$$

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

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

$$\begin{aligned} & \underset{\alpha}{\operatorname{minimize}} \quad \mathbb{E}_c [\ell^n(\alpha)] \\ & \text{subject to} \quad \alpha_j \geq 0, j = 1, \dots, N, i \neq j. \end{aligned} \quad (18)$$

Thus, our construction satisfies condition (14).

3. We obtain $\hat{\mathbf{z}}_{S^c}$ from (12) by substituting in the constructed $\hat{\alpha}$, $\hat{\mu}$ 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 incoherence 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 show that our constructed solution satisfies condition (10):

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

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

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

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

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 $\hat{\mu}_S = \mu_S^* = 0$, and hence $\hat{\mu} = \mu^*$.

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,

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

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

$$R_j^n = [\nabla^2 \ell^n(\tilde{\alpha}_j) - \nabla^2 \ell^n(\alpha^*)]_j^T (\hat{\alpha} - \alpha^*),$$

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

$$\begin{aligned} & \begin{pmatrix} \mathcal{Q}_{SS}^n & \mathcal{Q}_{SS^c}^n \\ \mathcal{Q}_{S^cS}^n & \mathcal{Q}_{S^cS^c}^n \end{pmatrix} \begin{pmatrix} \hat{\alpha}_S - \alpha_S^* \\ \hat{\alpha}_{S^c} - \alpha_{S^c}^* \end{pmatrix} \\ &= - \begin{pmatrix} \nabla_S \ell^n(\alpha^*) \\ \nabla_{S^c} \ell^n(\alpha^*) \end{pmatrix} - \lambda_n \begin{pmatrix} \hat{\mathbf{z}}_S \\ \hat{\mathbf{z}}_{S^c} \end{pmatrix} + \begin{pmatrix} \hat{\mu}_S \\ \hat{\mu}_{S^c} \end{pmatrix} - \begin{pmatrix} \mathbf{R}_S^n \\ \mathbf{R}_{S^c}^n \end{pmatrix} \end{aligned}$$

and, after some algebraic manipulation, we have

$$\begin{aligned} \lambda \hat{\mathbf{z}}_{S^c} &= -\nabla_{S^c} \ell^n(\boldsymbol{\alpha}^*) + \hat{\boldsymbol{\mu}}_{S^c} - \mathbf{R}_{S^c}^n \\ &- \mathcal{Q}_{S^c S}^n (\mathcal{Q}_{SS}^n)^{-1} (-\nabla_S \ell^n(\boldsymbol{\alpha}^*) - \lambda \hat{\mathbf{z}}_S + \hat{\boldsymbol{\mu}}_S - \mathbf{R}_S^n). \end{aligned}$$

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

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

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 F and G):

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

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

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 incoherence condition, if conditions 3 and 4 holds, λ_n is selected to satisfy*

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

where $K = k_1 + k_4 k_1 + k_1^2 + k_1 \sqrt{C_{\max}}$, 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 incoherence 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 λ_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 $n > 288^2 k_3^2 (2-\varepsilon)^4 C_{\min}^{-4} \varepsilon^{-4} d^2 \log p + (48k_3(2-\varepsilon)C_{\min}^{-1}(\alpha_{\min}^*)^{-1}\varepsilon^{-1})^2 d \log p$.

Last, we lift the dependency and incoherence 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 H and I).

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

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

Algorithm 1 ℓ_1 -regularized network inference

Require: C^n, λ_n, K, L

for all $i \in \mathcal{V}$ **do**

$k = 0$

while $k < K$ **do**

$$\boldsymbol{\alpha}_i^{k+1} = (\boldsymbol{\alpha}_i^k - L \nabla_{\boldsymbol{\alpha}_i} \ell^n(\boldsymbol{\alpha}_i^k) - \lambda_n L)_+$$

$k = k + 1$

end while

$$\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i^{K-1}$$

end for

return $\{\hat{\boldsymbol{\alpha}}_i\}_{i \in \mathcal{V}}$

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

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

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

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

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.

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 & Boyd, 2013). Moreover, they are easy to understand, derive, and implement. We first rewrite Eq. 4 as an unconstrained optimization problem:

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

where the non-smooth convex function $g(\boldsymbol{\alpha}) = \lambda_n \|\boldsymbol{\alpha}\|_1$ if $\boldsymbol{\alpha} \geq 0$ and $+\infty$ otherwise. Here, the general recipe from Parikh & 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 1) and the proximal operator $\text{prox}_{L^k g}(v)$, where L^k is a step size that we can set to a constant value L or find using a simple line search (Beck & Teboulle, 2009). Using Moreau's decomposition and the conjugate function g^* , it is easy to show that the proximal operator for our particular function $g(\cdot)$ is a soft-thresholding operator, $(v - \lambda_n L^k)_+$, 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

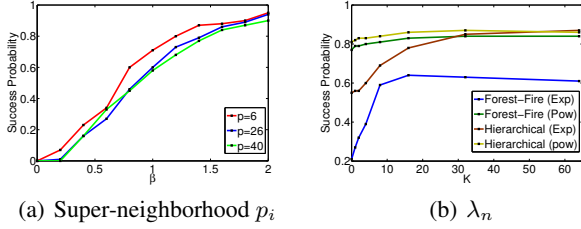


Figure 3. Success probability vs. # of cascades.

λ_n . Then, we compare our algorithm to two different state-of-the-art algorithms: NETRATE (Gomez-Rodriguez et al., 2011) and First-Edge (Abraham et al., 2013).

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 & 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 $U(0.5, 1.5)$. 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 infer the incoming links of nodes of a hierarchical Kronecker network with the same in-degree ($d_i = 3$) but different super-neighborhood set sizes p_i under different scalings β of the number of cascades $n = 10\beta d \log p$ and choose the regularization parameter λ_n as a constant factor of $\sqrt{\log(p)/n}$ as suggested by Th. 2. 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. As predicted by Th. 2, very different p values lead to curves that line up with each other quite well.

Regularization parameter λ_n Our main result indicates that the regularization parameter λ_n should be a constant

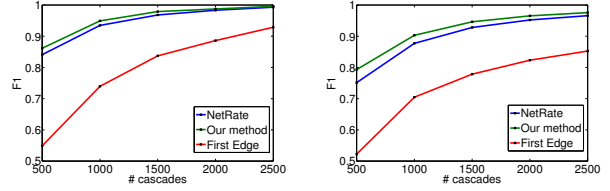


Figure 4. F_1 -score vs. # of cascades.

factor of $\sqrt{\log(p)/n}$. Fig. 3(b) 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 λ_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. 4 compares the accuracy of our algorithm, NETRATE and First-Edge against number of cascades for a hierarchical Kronecker network with power-law transmission model and a Forest Fire network with exponential transmission model, with an observation window $T = 10$. Our method outperforms both competitive methods, finding especially striking the competitive advantage with respect to First-Edge.

9. Conclusions

Our work contributes towards establishing a theoretical foundation of the network inference problem. Specifically, we proposed a ℓ_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 incoherence 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, 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 incoherence 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.

Acknowledgement

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

References

- Abrahamo, B., Chierichetti, F., Kleinberg, R., and Panconesi, A. Trace complexity of network inference. In *KDD*, 2013.
- Adar, E. and Adamic, L. A. Tracking Information Epidemics in Blogspace. In *Web Intelligence*, pp. 207–214, 2005.
- Barabási, A.-L. and Albert, R. Emergence of Scaling in Random Networks. *Science*, 286:509–512, 1999.
- Beck, A. and Teboulle, M. Gradient-based algorithms with applications to signal recovery. *Convex Optimization in Signal Processing and Communications*, 2009.
- Boyd, S. P. and Vandenberghe, L. *Convex optimization*. Cambridge University Press, 2004.
- Du, N., Song, L., Smola, A., and Yuan, M. Learning Networks of Heterogeneous Influence. In *NIPS*, 2012a.
- Du, N., Song, L., Woo, H., and Zha, H. Uncover Topic-Sensitive Information Diffusion Networks. In *AISTATS*, 2012b.
- Gomez-Rodriguez, M., Leskovec, J., and Krause, A. Inferring Networks of Diffusion and Influence. In *KDD*, 2010.
- Gomez-Rodriguez, M., Balduzzi, D., and Schölkopf, B. Uncovering the Temporal Dynamics of Diffusion Networks. In *ICML*, 2011.
- Gomez-Rodriguez, Manuel. *Ph.D. Thesis*. Stanford University & MPI for Intelligent Systems, 2013.
- Gripon, V. and Rabbat, M. Reconstructing a graph from path traces. *arXiv:1301.6916*, 2013.
- Kempe, D., Kleinberg, J. M., and Tardos, É. Maximizing the Spread of Influence Through a Social Network. In *KDD*, 2003.
- Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., and Ghahramani, Z. Kronecker Graphs: An Approach to Modeling Networks. *JMLR*, 2010.
- Mangasarian, O. L. A simple characterization of solution sets of convex programs. *Operations Research Letters*, 7(1):21–26, 1988.
- Netrapalli, P. and Sanghavi, S. Finding the Graph of Epidemic Cascades. In *ACM SIGMETRICS*, 2012.
- Newey, W. K. and McFadden, D. L. Large Sample Estimation and Hypothesis Testing. In *Handbook of Econometrics*, volume 4, pp. 2111–2245. 1994.
- Parikh, Neal and Boyd, Stephen. Proximal algorithms. *Foundations and Trends in Optimization*, 2013.
- Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. High-dimensional ising model selection using l_1 -regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.
- Rogers, E. M. *Diffusion of Innovations*. Free Press, New York, fourth edition, 1995.
- Saito, K., Kimura, M., Ohara, K., and Motoda, H. Learning continuous-time information diffusion model for social behavioral data analysis. *Advances in Machine Learning*, pp. 322–337, 2009.
- Snowsill, T., Fyson, N., Bie, T. De, and Cristianini, N. Refining Causality: Who Copied From Whom? In *KDD*, 2011.
- Wainwright, M. J. Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.
- Wang, L., Ermon, S., and Hopcroft, J. Feature-enhanced probabilistic models for diffusion network inference. In *ECML PKDD*, 2012.

A. 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 $\mathbf{X}^n(\boldsymbol{\alpha})$ 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. ■

B. 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 $n \rightarrow \infty$, there exists an ordering of the nodes and cascades within the cascade set such that the hazard matrix $\mathbf{X}^n(\boldsymbol{\alpha})$ 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 $[TB]$, 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 $[TB]$, and thus it is non-singular. ■

C. Proof of Eq 7.

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

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

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

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

D. Proof of Lemma 3

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

$$\begin{aligned} & \text{minimize}_{\boldsymbol{\alpha}_i} \quad \ell^n(\boldsymbol{\alpha}_i) \\ & \text{subject to} \quad \alpha_{ji} \geq 0, j = 1, \dots, N, i \neq j, \\ & \quad \|\boldsymbol{\alpha}_i\|_1 \leq C(\lambda_n) \end{aligned} \quad (20)$$

where $C(\lambda_n) < \infty$ is a positive constant. In this alternative formulation, λ_n is the Lagrange multiplier for the second constraint. Since λ_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. 20, 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 ℓ_1 -norm.

Consider the solution $\hat{\boldsymbol{\alpha}}$ such that $\|\hat{\mathbf{z}}_{S^c}\|_\infty < 1$ and thus $\nabla_{\alpha_{S^c}} \ell^n(\hat{\boldsymbol{\alpha}}) = -\lambda_n \hat{\mathbf{z}}_{S^c} + \hat{\boldsymbol{\mu}}_{S^c}$. 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 $\hat{\mu}_{ji} = 0$ 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 $\tilde{\alpha}_{S^c} = 0$ for the gradient to be constant across optimal solutions.

Finally, since $\alpha_{S^c} = 0$ 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.

E. 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 $\hat{\mathbf{u}}_S := \hat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_S^*$. That is $G(\hat{\mathbf{u}}_S) \leq G(\mathbf{u}_S), \forall \mathbf{u}_S \neq \hat{\mathbf{u}}_S$.
3. $G(\hat{\mathbf{u}}_S) \leq G(\mathbf{0}) = 0$.

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

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

Next, we will find a sphere centered at $\mathbf{0}$ with strictly positive radius B , $\mathbb{S}(B) := \{\mathbf{u}_S : \|\mathbf{u}_S\|_2 = B\}$, such that function $G(\mathbf{u}_S) > 0$ (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 $\hat{\mathbf{u}}_S := \hat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_S^*$ 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 λ_n and the neighborhood size d . More specifically, we will start by applying a Taylor series expansion and

the mean value theorem,

$$G(\mathbf{u}_S) = \nabla_S \ell^n(\boldsymbol{\alpha}_S^*)^\top \mathbf{u}_S + \mathbf{u}_S^\top \nabla_{SS}^2 \ell^n(\boldsymbol{\alpha}_S^* + b\mathbf{u}_S) \mathbf{u}_S + \lambda_n(\|\boldsymbol{\alpha}_S^* + \mathbf{u}_S\|_1 - \|\boldsymbol{\alpha}_S^*\|_1), \quad (21)$$

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, $\nabla_S \ell(\cdot)$,

$$|\nabla_S \ell^n(\boldsymbol{\alpha}_S^*)^\top \mathbf{u}_S| \leq \|\nabla_S \ell\|_\infty \|\mathbf{u}_S\|_1 \leq \|\nabla_S \ell\|_\infty \sqrt{d} \|\mathbf{u}_S\|_2 \leq 4^{-1} \lambda_n B \sqrt{d}. \quad (22)$$

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

$$\lambda_n \|\boldsymbol{\alpha}_S^* + \mathbf{u}_S\|_1 - \|\boldsymbol{\alpha}_S^*\|_1 \leq \lambda_n \|\mathbf{u}_S\|_1 \leq \lambda_n \sqrt{d} \|\mathbf{u}_S\|_2. \quad (23)$$

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

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

Now, we introduce two additional quantities,

$$\Delta \mathbf{D}_{SS}^n = \mathbf{D}_{SS}^n(\boldsymbol{\alpha}_S^* + b\mathbf{u}_S) - \mathbf{D}_{SS}^n(\boldsymbol{\alpha}_S^*) \\ \Delta \mathbf{X}_S^n = \mathbf{X}_S^n(\boldsymbol{\alpha}_S^* + b\mathbf{u}_S) - \mathbf{X}_S^n(\boldsymbol{\alpha}_S^*),$$

and rewrite q as

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

Next, we use dependency condition,

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

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

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

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

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

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 ,

$$q \geq C_{\min} B^2 - (k_2 + 2k_1 \sqrt{C_{\max}}) B^3. \quad (24)$$

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 λ_n to satisfy $\lambda_n \sqrt{d} \leq 0.5 C_{\min} / K (k_2 + 2k_1 \sqrt{C_{\max}})$. Then,

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

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

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

F. 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

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

and then,

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

G. Proof of Lemma 6

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

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

where,

$$\omega_j^n = [\mathbf{D}^n(\bar{\boldsymbol{\alpha}}_j) - \mathbf{D}^n(\boldsymbol{\alpha}^*)]_j^\top (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)$$

$$\delta_j^n = \frac{1}{n} \mathbf{V}_j^n (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)$$

$$\mathbf{V}_j^n = [\mathbf{X}^n(\bar{\boldsymbol{\alpha}}_j)]_j \mathbf{X}^n(\bar{\boldsymbol{\alpha}}_j)^\top - [\mathbf{X}^n(\boldsymbol{\alpha}^*)]_j \mathbf{X}^n(\boldsymbol{\alpha}^*)^\top.$$

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 ω_j^n .

$$\begin{aligned} |\omega_j^n| &\leq k_1 \|\bar{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*\|_2 \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2 \\ &\leq k_1 \theta_j \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2^2 \\ &\leq k_1 \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2^2. \end{aligned} \quad (25)$$

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

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

where,

$$\boldsymbol{\Lambda}_1 = [\mathbf{X}^n(\boldsymbol{\alpha}^*)]_j (\mathbf{X}^n(\bar{\boldsymbol{\alpha}}_j)^\top - \mathbf{X}^n(\boldsymbol{\alpha}^*)^\top)$$

$$\boldsymbol{\Lambda}_2 = \{[\mathbf{X}^n(\bar{\boldsymbol{\alpha}}_j)]_j - [\mathbf{X}^n(\boldsymbol{\alpha}^*)]_j\} (\mathbf{X}^n(\bar{\boldsymbol{\alpha}}_j)^\top - \mathbf{X}^n(\boldsymbol{\alpha}^*)^\top)$$

$$\boldsymbol{\Lambda}_3 = ([\mathbf{X}^n(\bar{\boldsymbol{\alpha}}_j)]_j - [\mathbf{X}^n(\boldsymbol{\alpha}^*)]_j) \mathbf{X}^n(\boldsymbol{\alpha}^*)^\top.$$

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

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

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

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

For the second term, we also start by applying Cauchy in-

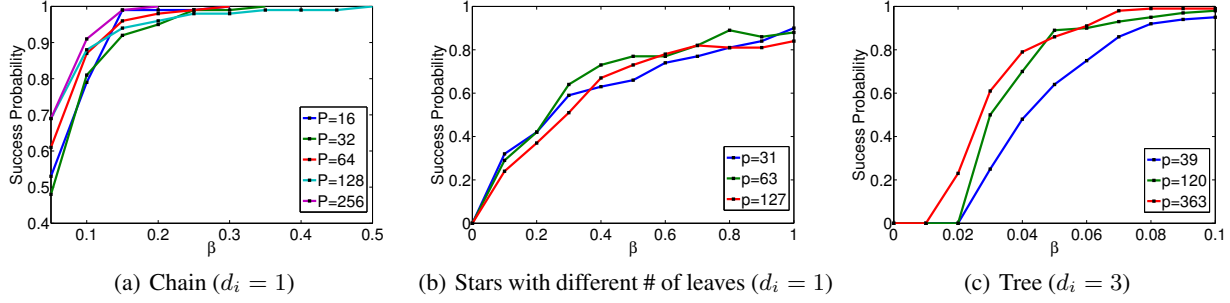


Figure 5. Success probability vs. # of cascades. Different super-neighborhood sizes p_i .

equality,

$$|\Lambda_2(\hat{\alpha} - \alpha^*)| \leq \|[\mathbf{X}^n(\bar{\alpha}_j)]_j - [\mathbf{X}^n(\alpha^*)]_j\|_2 \times \|[\mathbf{X}^n(\bar{\alpha}_j)]^\top - [\mathbf{X}^n(\alpha^*)]^\top\|_2 \|\hat{\alpha} - \alpha^*\|_2,$$

and then use condition 3 (Lipschitz Continuity),

$$|\Lambda_2(\hat{\alpha} - \alpha^*)| \leq nk_1^2 \|\hat{\alpha} - \alpha^*\|_2^2.$$

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

$$|\Lambda_3(\hat{\alpha} - \alpha^*)| \leq \|[\mathbf{X}^n(\bar{\alpha}_j)]_j - [\mathbf{X}^n(\alpha^*)]_j\|_2 \times \|[\mathbf{X}^n(\alpha^*)]^\top\|_2 \|\hat{\alpha} - \alpha^*\|_2,$$

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

$$|\Lambda_3(\hat{\alpha} - \alpha^*)| \leq nk_1 \sqrt{C_{\max}} \|\hat{\alpha} - \alpha^*\|_2^2$$

Now, we combine the bounds,

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

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 λ_n to satisfy $\lambda_n d \leq C_{\min}^2 \frac{\varepsilon}{36K(2-\varepsilon)}$ yields:

$$\begin{aligned} \|\mathbf{R}^n\|_\infty / \lambda_n &\leq 3K \lambda_n d / C_{\min}^2 \\ &\leq \frac{\varepsilon}{4(2-\varepsilon)} \end{aligned}$$

H. 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 $z_{jk}^c = [\nabla^2 g(\mathbf{t}^c; \alpha^*) - \nabla^2 \ell^n(\alpha^*)]_{jk}$ 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 sam-

ple mean cascade log-likelihood \mathcal{Q}_{SS}^n as:

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

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

$$P(|z_{jk}| \geq \beta) \leq 2 \exp\left(-\frac{\beta^2 n}{8k_5^2}\right), \quad (26)$$

and further,

$$\begin{aligned} &P(\| \mathcal{Q}_{SS}^n(\alpha^*) - \mathcal{Q}_{SS}^*(\alpha^*) \|_2 \geq \delta) \\ &\leq 2 \exp\left(-K \frac{\delta^2 n}{d^2} + 2 \log d\right) \quad (27) \end{aligned}$$

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,

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

Reasoning in a similar way, we bound the minimum eigen-

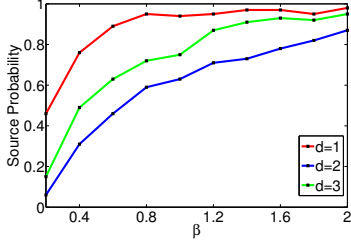


Figure 6. Success probability vs. # of cascades. Different in-degrees d_i .

value of \mathcal{Q}_{SS}^n :

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

I. Proof of Lemma 8

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

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

where,

$$\begin{aligned} A_1 &= \mathcal{Q}_{S^c S}^*(\mathcal{Q}_{S^c S}^n)^{-1} - (\mathcal{Q}_{S^c S}^*)^{-1}, \\ A_2 &= [\mathcal{Q}_{S^c S}^n - \mathcal{Q}_{S^c S}^*][(\mathcal{Q}_{S^c S}^n)^{-1} - (\mathcal{Q}_{S^c S}^*)^{-1}] \\ A_3 &= [\mathcal{Q}_{S^c S}^n - \mathcal{Q}_{S^c S}^*](\mathcal{Q}_{SS}^*)^{-1}, \\ A_4 &= \mathcal{Q}_{S^c S}^*(\mathcal{Q}_{SS}^*)^{-1}, \end{aligned}$$

$\mathcal{Q}^* = \mathcal{Q}^*(\alpha^*)$ and $\mathcal{Q}^n = \mathcal{Q}^n(\alpha^*)$. 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:

$$\begin{aligned} P[\|\mathcal{Q}_{S^c S}^n - \mathcal{Q}_{S^c S}^*\|_{\infty} \geq \delta] \\ \leq 2 \exp\left(-K \frac{n\delta^2}{d^2} + \log d + \log(p-d)\right) \quad (28) \end{aligned}$$

$$\begin{aligned} P[\|\mathcal{Q}_{SS}^n - \mathcal{Q}_{SS}^*\|_{\infty} \geq \delta] \\ \leq 2 \exp\left(-K \frac{n\delta^2}{d^2} + 2 \log d\right) \quad (29) \end{aligned}$$

$$\begin{aligned} P[\|(\mathcal{Q}_{SS}^n)^{-1} - (\mathcal{Q}_{SS}^*)^{-1}\|_{\infty} \geq \delta] \\ \leq 4 \exp\left(-K \frac{n\delta}{d^3} - K' \log d\right) \quad (30) \end{aligned}$$

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

$$\begin{aligned} P[\|\mathcal{Q}_{S^c S}^n - \mathcal{Q}_{S^c S}^*\|_{\infty} \geq \delta] \\ = P\left[\max_{j \in S^c} \sum_{k \in S} |z_{jk}| \geq \delta\right] \\ \leq (p-d)P\left[\sum_{k \in S} |z_{jk}| \geq \delta\right], \end{aligned}$$

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,

$$\begin{aligned} P\left[\sum_{k \in S} |z_{jk}| \geq \delta\right] &\leq P[\exists k \in S \mid |z_{jk}| \geq \delta/d] \\ &\leq dP[|z_{jk}| \geq \delta/d]. \end{aligned}$$

Thus,

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

At this point, we can obtain the first confidence bound by using Eq. 26 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:

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

Next, we bound each term of the final expression in the above equation separately. The first term can be bounded using Eq. 27:

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

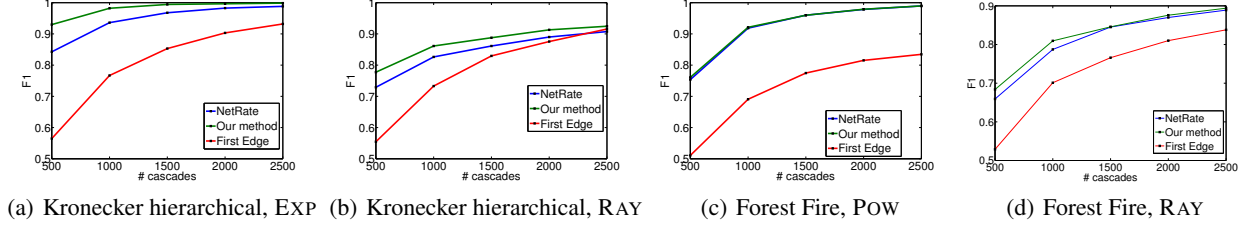
The second term can be bounded using Lemma 6:

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

Then, the third confidence bound follows. \blacksquare

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

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


 Figure 7. F_1 -score vs. # of cascades.

and further,

$$\begin{aligned} \|\|A_1\|\|_\infty &\leq \|\|Q_{S^c S}^*(Q_{SS}^*)^{-1}\|\|_\infty \\ &\quad \times \|\|(Q_{SS}^*) - (Q_{SS}^n)\|\|_\infty \|\|(Q_{SS}^n)^{-1}\|\|_\infty. \end{aligned}$$

Next, using the incoherence condition easily yields:

$$\begin{aligned} \|\|A_1\|\|_\infty &\leq (1 - \varepsilon) \|\|(Q_{SS}^*) - (Q_{SS}^n)\|\|_\infty \\ &\quad \times \sqrt{d} \|\|(Q_{SS}^n)^{-1}\|\|_2 \end{aligned}$$

Now, we apply Lemma 6 with $\delta = C_{\min}/2$ to have that $\|\|(Q_{SS}^n)^{-1}\|\|_2 \leq \frac{2}{C_{\min}}$ with probability greater than $1 - \exp(-Kn/d^2 + K' \log d)$, and then use Eq. 30 with $\delta = \frac{\varepsilon C_{\min}}{12\sqrt{d}}$ to conclude that

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

Control of A_2 . We rewrite the term A_2 as

$$\|\|A_2\|\|_\infty \leq \|\|Q_{S^c S}^n - Q_{S^c S}^*\|\|_\infty \|\|(Q_{SS}^n)^{-1} - (Q_{SS}^*)^{-1}\|\|_\infty,$$

and then use Eqs. 28 and 29 with $\delta = \sqrt{\varepsilon/6}$ to conclude that

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

Control of A_3 . We rewrite the term A_3 as

$$\begin{aligned} \|\|A_3\|\|_\infty &= \sqrt{d} \|\|(Q_{SS}^*)^{-1}\|\|_2 \|\|Q_{S^c S}^n - Q_{S^c S}^*\|\|_\infty \\ &\leq \frac{\sqrt{d}}{C_{\min}} \|\|Q_{S^c S}^n - Q_{S^c S}^*\|\|_\infty. \end{aligned}$$

We then apply Eq. 28 with $\delta = \frac{\varepsilon C_{\min}}{6\sqrt{d}}$ to conclude that

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

and thus,

$$\begin{aligned} P[\|\|Q_{S^c S}^n (Q_{SS}^n)^{-1}\|\|_\infty \geq 1 - \frac{\varepsilon}{2}] & \\ &= \mathcal{O}\left(\exp\left(-K \frac{n}{d^3} + \log p\right)\right). \end{aligned}$$

J. Additional experiments

Parameters (n, p, d) . Figure 5 shows the success probability at inferring 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-neighborhood set sizes p_i and experiment with different scalings β of the number of cascades $n = 10\beta d \log p$. We set the regularization parameter λ_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.

Figure 6 shows the success probability at inferring 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 β of the number of cascades $n = 10\beta d \log p$ and choose the regularization parameter λ_n as a constant factor of $\sqrt{\log(p)/n}$ as suggested by Theorem 2. We used an exponential transmission model and time window $T = 5$. As predicted by Theorem 2, in this case, different d values lead to noticeably different curves.

Comparison with NETRATE and First-Edge. Figure 7 compares the accuracy of our algorithm, NETRATE and First-Edge against number of cascades for different type of networks and transmission models. Our method typically outperforms both competitive methods. We find 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.