Open Problem: Finding *Good* Cascade Sampling Processes for the Network Inference Problem

Manuel Gomez-Rodriguez MANUELGR@TUE.MPG.DE
MPI for Intelligent Systems

Le Song LSONG@CC.GATECH.EDU
Georgia Institute of Technology

Bernhard Schölkopf BS@TUE.MPG.DE
MPI for Intelligent Systems

Abstract

Information spreads across social and technological networks, but often the network structures are hidden and we only observe the traces left by the diffusion processes, called *cascades*. It is known that, under a popular continuous-time diffusion model, as long as the model parameters satisfy a natural incoherence condition, it is possible to 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. However, the incoherence condition depends, in a non-trivial way, on the source (node) distribution of the cascades, which is typically *unknown*. Our open problem is whether it is possible to design an *active* algorithm which samples the source locations in a sequential manner and achieves the same or even better sample complexity, *e.g.*, $o(d_i^3 \log N)$, than previous work.

1. Introduction

Diffusion of information can be naturally modeled as a stochastic process that occur over the edges of an underlying network (Rogers, 1995). In this context, 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). 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 et al., 2010).

1.1. Diffusion Model

A sequence of recent work has argued that modeling information diffusion using *continuous-time* diffusion networks can provide significantly more accurate models than discrete-time models (Gomez Rodriguez et al., 2011; Du et al., 2012, 2013b,a). For our open problem, we build on this line of work, which models diffusion as follows: Given a *directed* contact network, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with N nodes, each diffusion 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, τ, drawn from an associated transmission function $f(\tau; \alpha_{ji})$, a density over $\mathbb{R}_+$. A concrete and common example of transmission function is $f(\tau; \alpha_{ji}) = \alpha_{ji} \exp(-\alpha_{ji}\tau)$. Transmission times are sampled independently, possibly from different distributions, across edges. Then, the infected neighbors transmit the contagion to their respective neighbors, and the process continues. Here, 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*.

Observations 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 $\mathbf{t}^c := (t^c_1, \ldots, t^c_N)$ recording when nodes are infected, $t^c_k \in [0, T^c] \cup \{\infty\}$. Symbol ∞ labels nodes that are not infected during the observation window $[0, T^c]$ – it does not imply they are never infected. Under these settings, the likelihood of a cascade $\mathbf{t}$ is (Gomez Rodriguez et al., 2011):

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

where $\mathbf{A} = \{\alpha_{ji}\}$ denotes the collection of parameters, $S(t_i|t_j; \alpha_{ji}) = 1 - \int_0^{t_i - t_j} f(\tau; \alpha_{ji})\, d\tau$ 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. Term I accounts for the probability that uninfected nodes survive to all infected nodes in the cascade up to T; and term II accounts 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.

2. The 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 $\mathbf{A}^*$. 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 ℓ_1-regularized maximum likelihood estimation problem (Daneshmand et al., 2014), which decouples into a set of independent smaller subproblems, one per node, where we infer the incoming edges of each node and the parameters associated with these edges:

$$\begin{array}{ll} \text{minimize}_{\mathbf{A}} & \ell^n(\boldsymbol{\alpha}_i) + \lambda_n ||\boldsymbol{\alpha}_i||_1 \\ \text{subject to} & \alpha_{ji} \geq 0,\ j = 1, \ldots, N, i \neq j, \end{array} \quad (2)$$

where $\boldsymbol{\alpha}_i := \{\alpha_{ji} \,|\, j = 1, \ldots, N, i \neq j\}$ 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 involving $\boldsymbol{\alpha}_i$ in the cascades log-likelihood $\log f(\mathbf{t}^c, \boldsymbol{A})$. Furthermore, d_i denotes the number of *true* parents for node i.

Under some technical conditions, including an incoherence condition on the Hessian, $\mathcal{Q}^*$, of the population log-likelihood, $\mathbb{E}\left[\ell^n(\boldsymbol{\alpha}_i)\right] = \mathbb{E}\left[\log g_i(\mathbf{t}^c; \boldsymbol{\alpha}_i)\right]$, which states that there exists $\varepsilon \in (0, 1]$ such that $|||\mathcal{Q}^*_{S^c S}\left(\mathcal{Q}^*_{SS}\right)^{-1}|||_\infty \leq 1 - \varepsilon$, where $|||A|||_\infty = \max_j \sum_k |A_{jk}|$ and S denote the subset of indexes associated to node i's true parents, the following result holds:

Theorem 1 (Daneshmand et al. (2014)) *Consider an instance of the continuous-time diffusion model with parameters α^*_{ji} and associated edges $\mathcal{E}^*$, 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*

$$\lambda_n \geq 8k_3 \frac{2-\varepsilon}{\varepsilon}\sqrt{\frac{\log N}{n}}. \tag{3}$$

Then, under some technical conditions, there exist positive constants L and K, independent of (n, N, d_i), such that if

$$n > L d_i^3 \log N, \tag{4}$$

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 l1-regularized network inference problem defined in Eq. 2 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.*

3. Active Source Sampling

The success of the network inference algorithm in equation (2) relies on the fulfillment of the above mentioned incoherence 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 incoherence condition depends, in a non-trivial way, on the network structure, diffusion parameters, and the source distribution $\mathbb{P}(s)$ (Daneshmand et al., 2014), 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. Therefore, we propose the following open problem:

> **Open Problem:** Suppose there exists an unknown $\mathbb{P}(s)$ where the incoherence 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 1, or even better sample complexity, *e.g.*, $o(d_i^3 \log N)$?

References

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

H. Daneshmand, M. Gomez-Rodriguez, L. Song, and B. Schölkopf. Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm. In *Proc. of the 31st International Conference on Machine Learning (ICML)*, 2014.

N. Du, L. Song, A. Smola, and M. Yuan. Learning Networks of Heterogeneous Influence. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

Nan Du, Le Song, Manuel Gomez-Rodriguez, and Hongyuan Zha. Scalable influence estimation in continuous-time diffusion networks. In *Advances in Neural Information Processing Systems*, pages 3147–3155, 2013a.

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, 2013b.

M. Gomez Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the Temporal Dynamics of Diffusion Networks. In *Proc. of the 28th International Conference on Machine Learning (ICML)*, 2011.

M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring Networks of Diffusion and Influence. In *Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2010.

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