Journal of Machine Learning Research 27 (2026) 1-46 Submitted 11/25; Revised 9/26; Published 9/26

Flexible Modeling of Information Diffusion on Networks with Statistical Guarantees

Alexander Kagan AMKAGAN@UMICH.EDU
Elizaveta Levina ELEVINA@UMICH.EDU
Ji Zhu JIZHU@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI, 48109

Editor: Jie Peng

Abstract

Modeling information spread through a network is one of the key problems of network analysis. Most approaches assume the spread is governed by some probabilistic diffusion model, often parameterized by the strength of connections between network members (edge weights), highlighting the need for methods that can accurately estimate them. Multiple prior works suggest estimators for particular diffusion models, but lack a rigorous statistical analysis establishing their asymptotic properties and allowing for uncertainty quantification. In this paper, we develop a likelihood-based approach to estimate edge weights from the observed information diffusion paths under the proposed General Linear Threshold (GLT) model, a broad class of discrete-time information diffusion models that includes both the well-known linear threshold (LT) and independent cascade (IC) models. We derive a finite sample error bound for the estimator and demonstrate that it is asymptotically normal under mild conditions. We conclude by studying the GLT model in the context of the Influence Maximization (IM) problem. Extensive experiments on synthetic and real-world networks demonstrate that the proposed estimation and inference frameworks for the GLT model parameters can significantly improve estimation of spread from a given subset of nodes, prediction of node activation, and the quality of the IM problem solutions.

Keywords: social networks, information diffusion, linear threshold model, independent cascade model, influence maximization

1. Introduction

The emergence of large-scale online social networks has led to the appearance of rich datasets which include not only connections between users and user features, but also the paths of information propagation (also called information diffusion) between users. The term "information" here is interpreted broadly and can refer to anything that spreads from node to node, be it a news item, a tweet, or a virus. Information propagation paths, also known as *propagation traces*, are especially valuable in modeling information spread, since they provide direct data on the influence users have on their network neighbors. For example, Liu and Wu (2018) used propagation traces for fake news detection, while Saito et al. (2008) used them to estimate the probabilities of information transmission between users, i.e., edge weights, assuming that the propagation follows the Independent Cascade (IC) model (Goldenberg

©2026 Alexander Kagan, Elizaveta Levina, and Ji Zhu.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-2716.html.

et al., 2001), a simple and arguably the most popular information diffusion model. Goyal et al. (2011) used this data to assign influence credits to users and subsequently solve the influence maximization (IM) problem (Richardson and Domingos, 2002; Kempe et al., 2003), that is, a task of identifying a fixed-size subset of users to "seed" the propagation that would result in the highest number of nodes eventually reached. While all these papers use the propagation traces to solve important problems related to influence propagation, the solutions are usually obtained assuming a very specific diffusion model and with no further uncertainty quantification. This calls for developing a general statistical framework enabling flexible modeling, estimation, and uncertainty quantification of a network diffusion process based on these traces, which is the main contribution of this paper.

We introduce a natural but surprisingly rich class of models we call General Linear Threshold (GLT) diffusion models, which includes both the popular Linear Threshold (LT) (Granovetter, 1978) and Independent Cascade (IC) models. Unlike other flexible generalizations, it is not over-parametrized and thus allows for robust estimation of its parameters from propagation traces and uncertainty quantification for various downstream tasks. Similarly to the LT model, the GLT model assumes that each edge is assigned a deterministic weight, and a node activates when the sum of incoming weights from its currently active parents crosses a node-specific random threshold. Unlike the LT model, which assumes that all the thresholds are uniformly distributed on $[0, 1]$, the GLT model allows the threshold distribution to vary between nodes. This seemingly straightforward generalization turns out to bring a lot of new flexibility to the model, since it allows for heterogeneity in how readily different nodes accept new information; some may be much more easily influenced than others. The popular alternative to the LT, the IC model, which we show corresponds to the GLT model with exponentially distributed thresholds (Proposition 6), is also not able to allow for this heterogeneity, leaving the GLT model the only option that can.

Importantly, the proposed class of GLT models comes with a convenient likelihood form, allowing for efficient parameter estimation via constrained convex optimization. Under mild regularity conditions, we establish a finite-sample error bound for this estimator which guarantees consistency, and derive its asymptotic distribution, allowing for construction of asymptotic confidence intervals for the GLT edge weights and their smooth transformations. In section 4.3, we demonstrate several applications of this result, including uncertainty quantification for predicted node activation probabilities, testing for the difference in parents' influence on a child node, and solving the Robust Influence Maximization problem, an extension of the standard IM problem where edge weights are not available but are known to lie in a given confidence interval.

While there have been several papers (He et al., 2016; Narasimhan et al., 2015) establishing Probably Approximately Correct (PAC) learnability guarantees for nodes' activation probabilities under the IC, LT, and several similar models, there has been very little work establishing theoretical guarantees for the diffusion model parameters. The only papers we are aware of that address this are Gomez-Rodriguez et al. (2014), which derived sufficient identifiability conditions and established consistency for the parameters of several continuous-time diffusion models, and Pouget-Abadie and Horel (2015), which introduced a General Cascade model (which we show can be viewed as a variant of GLT as well) and focused on inferring the unobserved graph structure from the observed cascades. While they

established a bound implying $\sqrt{n}$-consistency for their estimator, they acknowledged their framework cannot be extended to the LT model and leave this for future work. To the best of our knowledge, this paper is the first to provide a theoretical guarantee for parameter estimates from propagation traces for any threshold-based diffusion model. We conclude our theoretical analysis with a study of the GLT model in the context of the IM problem in Section 5, showing that, under the GLT model, the IM problem can be solved using the natural greedy strategy with standard optimality guarantees if all threshold distributions are concave, and that, for a class of graphs, the error rate of the IM solution mirrors the error rate of weight estimates.

The rest of this manuscript is organized as follows. In Section 2, we fix notation and define important concepts related to information propagation. In Section 3, we introduce the General Linear Threshold model and show how it relates to other diffusion models. In Section 4, we establish identifiability conditions for the GLT, derive a likelihood approach to weight estimation, and establish conditions for the estimator's consistency and asymptotic normality. We further extend the estimation procedure to the case of partially observed propagation traces and unknown threshold distributions. Section 5 presents a short study of the IM problem under the GLT model. Finally, Section 6 presents experiments on synthetic and real-world networks showing how the flexibility of the GLT model, coupled with the proposed weight estimation and uncertainty quantification procedure, can significantly improve performance in various downstream tasks, including the IM problem, prediction of node activation probabilities, and spread estimation.

2. A statistical framework for information diffusion models

In this section, we set up a general framework for modeling discrete-time influence propagation on a network. We present all models from the point of view of statistical models dependent on parametric distributions, which may differ from other standard treatments of such models. We start by setting up notation.

Let $G = (V, E)$ be a graph where V is the set of nodes and $E \subseteq V \times V$ is the set of edges. Unless otherwise stated, we assume throughout this manuscript that G is a simple directed graph, that is, a graph with no self-loops and no repeated edges. For a directed edge $u \to v$, we refer to u as the parent node and v as the child node. We denote the sets of parent and children nodes of a node v as, respectively,

$$P(v) = \{u : (u, v) \in E\} \quad \text{and} \quad C(v) = \{u : (v, u) \in E\}.$$

Similarly, for any set of nodes $S \subset V$, we write

$$P(S) := \bigcup_{v \in S} P(v) \setminus S \quad \text{and} \quad C(S) := \bigcup_{v \in S} C(v) \setminus S.$$

In words, $P(S)$ and $C(S)$ consist of nodes outside of S which have at least one child or parent in S, respectively. Since we will often additionally assume that each edge is associated with a weight or a transmission probability, it is convenient to fix some edge order for this correspondence to be uniquely defined. For that, we assume that the nodes are enumerated from 1 to $|V|$ and the edges are sorted lexicographically, by child node first and then by

parent node. For example, a triangle with each edge going in both directions corresponds to the ordered set $E = [(2,1),(3,1),(1,2),(3,2),(1,3),(2,3)]$.

A *diffusion model* associated with G is a probabilistic model that governs the spread of information through the graph, usually with some edge- or node-dependent transmission parameters $\boldsymbol{\theta}$. Our primary goal is estimation and inference for these parameters from the observed information diffusion paths. In general, the discrete-time information diffusion path starts with a given non-empty set $A_0 \subset V$ of initially activated (influenced) nodes, also known as the *seed set*. Then, at each time step $t = 1, 2 \ldots$, the currently active nodes A_{t-1} try to activate their network children while possibly deactivating themselves, making the set of all active nodes change to A_t. The process stops when no node can change its activation status or when the maximal time horizon is reached. In this work, we focus on the (quite general) subtype of discrete-time diffusion processes satisfying the following assumptions.

1. Information propagation is *progressive*, which means that once a node is activated, it remains activated. This implies that $A_{t-1} \subset A_t$ for every time step t. We define the disjoint sets of *newly activated nodes* at times $t \geq 0$ as $D_t := A_t \setminus A_{t-1}$ with the convention that $A_{-1} = \emptyset$ (no nodes are active before time 0).

2. A child node v can be activated at a time $t \geq 1$ only if it has a newly activated parent, that is, if $D_{t-1} \cap P(v)$ is not empty. Intuitively, this means that if a parent's influence was not enough to activate a child as soon as the parent become active, the only way for the child to be activated later is if another one of its parents becomes active.

Assumption 2 implies that propagation stops at time $T = \arg\min_{t \geq 0}\{t : D_{t+1} = \emptyset\}$, that is, the first time when no new nodes were activated. The entire diffusion process can then be encoded by the set sequence $\mathcal{D} := (D_0, \ldots, D_T)$, which we refer to as the *information diffusion path* or the *propagation trace*. We write $A(\mathcal{D})$ to denote the set of all nodes activated along the entire path.

The assumptions imply that not every sequence of node subsets can be a feasible propagation trace. The following definition formalizes what makes a propagation trace feasible. We denote the set of all feasible traces on G by $\mathcal{F}(G)$.

Definition 1 (Feasible trace) *We say that a set sequence $\mathcal{D} = (D_0, \ldots, D_T)$ is a feasible propagation trace if*

1. *$D_0 \neq \emptyset$;*

2. *All sets $D_t,\ t = 0, \ldots, T$ are disjoint;*

3. *For every $t = 1, \ldots, T$, if $v \in D_t$ then $v \in C(D_{t-1}) \setminus A_{t-1}$, i.e., each node newly activated at time t has at least one parent that was newly activated at time $t-1$.*

Now, we are ready to state the formal definition of the diffusion model on the graph, which can be thought of as a collection of rules governing the transmission of information from parents to children. These rules are not necessarily Markovian and can depend on the entire previous propagation history.

Definition 2 (Diffusion model) *A diffusion model on $G = (V, E)$ with a possibly graph-dependent parameter space Θ is a collection $\mathcal{M}_G(\Theta) = \{M_{G,\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ where $M_{G,\boldsymbol{\theta}}$ is a mapping from a feasible trace $\mathcal{D}_t = (D_0, \ldots, D_{t-1}) \in \mathcal{F}(G)$ to a probability distribution on feasible sets of newly activated nodes,*

$$\mathbb{P}^t_{\boldsymbol{\theta}}(D_t = S \mid \mathcal{D}_t),\ S \subset C(D_{t-1}) \setminus A_{t-1}. \tag{1}$$

We omit the subscript G in $M_{G,\boldsymbol{\theta}}$ and $\mathcal{M}_G(\Theta)$ whenever it is clear from the context.

The most general form of Definition 2 requires specifying the activation probability for each possible subset of $C(D_{t-1}) \setminus A_{t-1}$, which is impractical to work with. A standard simplifying assumption is that node activations are independent conditionally on the preceding propagation history, implying that the right-hand side in (1) can be decomposed as follows:

$$\mathbb{P}^t_{\boldsymbol{\theta}}(D_t = S \mid \mathcal{D}) = \prod_{v \in S} \mathbb{P}^t_{\boldsymbol{\theta}}(v \in D_t \mid \mathcal{D}) \prod_{v \in C(D_{t-1}) \setminus (A_{t-1} \sqcup S)} \Big(1 - \mathbb{P}^t_{\boldsymbol{\theta}}(v \in D_t \mid \mathcal{D})\Big).$$

With this additional assumption, defining a diffusion model is equivalent to defining the activation probability of a node given an arbitrary feasible diffusion history.

Remark 3 *Sometimes, it may be convenient to refer to a class of diffusion models without a reference to a specific graph G. We define a diffusion model class as the collection $\mathcal{M} = \{\mathcal{M}_G(\Theta_G) : G \in \mathcal{G}\}$, where $\mathcal{G}$ is the set of simple directed graphs.*

Next, we introduce the three arguably most popular diffusion model classes – the Linear Threshold (LT), the Independent Cascade (IC), and the Triggering model – and show how they fit into our general framework. We note that by construction, node activation events are independent conditionally on propagation history for all of these model classes.

Example 1 (Linear Threshold (LT) Model) *Assume each edge in the graph $G = (V, E)$ is assigned a weight parameter $b_{u,v}$, and we arrange the weights into a vector $\boldsymbol{\theta}$. The parameter space for the LT model is given by*

$$\Theta_{LT} = \Big\{\boldsymbol{\theta} \in \mathbb{R}^{|E|} \text{ s.t. for all } v \in V,\ \sum_{u \in P(v)} b_{u,v} \le 1 \ \text{ and } \ b_{u,v} \ge 0 \ \text{ for all } \ (u,v) \in E\Big\}. \tag{2}$$

Each node $v \in V$ gets a random activation threshold U_v sampled i.i.d. from the Unif$[0,1]$ *distribution at the outset. At every time step $t \ge 1$, a non-active node v becomes activated if the sum of the edge weights from all its previously activated parents exceeds its threshold U_v, that is,*

$$\mathbb{P}^t_{\boldsymbol{\theta}}(v \in D_t \mid D_0, \ldots, D_{t-1}) = \frac{\sum_{u \in D_{t-1} \cap P(v)} b_{u,v}}{1 - \sum_{u \in A_{t-2} \cap P(v)} b_{u,v}}.$$

This expression gives the probability of exceeding the threshold at time t given that it was not yet exceeded at time $t-1$.

Example 2 (Independent Cascade (IC) Model) *Assume that each edge $(u, v) \in E$ is associated with a propagation probability $p_{u,v} \in [0, 1]$, arranged into a vector $\boldsymbol{\theta} \in \Theta_{IC} = [0, 1]^{|E|}$. At every time step $t \geq 1$, each newly active node $u \in D_{t-1}$ independently tries to activate all of its not yet active children $v \in C(u) \setminus A_{t-1}$ with probability $p_{u,v}$, that is,*

$$\mathbb{P}^t_{\boldsymbol{\theta}}(v \in D_t \mid D_0, \ldots, D_{t-1}) = 1 - \prod_{D_{t-1} \cap P(v)} (1 - p_{u,v}).$$

Example 3 (Triggering Model) *At the outset, each node v independently chooses a random triggering set Γ_v according to some distribution over subsets of its parents, with probability of a node v choosing $S \subset P(v)$ as its triggering set denoted by $p_{v,S}$. With $\mathcal{P}^m$ denoting the set of all discrete distributions on m points, the parameter space is then $\Theta_{TR} = \{\mathcal{P}^{|P(v)|},\ v \in V\}$. An inactive node v becomes active at time $t \geq 1$ if its triggering set Γ_v contains a node in D_{t-1}, with probability of activation given by*

$$\mathbb{P}^t_{\boldsymbol{\theta}}(v \in D_t \mid D_0, \ldots, D_{t-1}) = \frac{\sum_{S \subset P(v) \cap D_{t-1}} p_{v,S}}{\sum_{S \subset P(v) \setminus A_{t-2}} p_{v,S}}.$$

As shown by (Kempe et al., 2003), the IC and LT model classes are special cases of the Triggering model class. Conveniently, within our parametric framework, we can formalize this in a very general form.

Definition 4 (Diffusion model subclass) *Consider two diffusion model classes $\mathcal{M} = \{\mathcal{M}_G(\Theta_G) : G \in \mathcal{G}\}$ and $\tilde{\mathcal{M}} = \{\tilde{\mathcal{M}}_G(\tilde{\Theta}_G) : G \in \mathcal{G}\}$. We say that $\mathcal{M}$ is a subclass of $\tilde{\mathcal{M}}$ if for any $G \in \mathcal{G}$ and any instance $M_{G,\boldsymbol{\theta}} \in \mathcal{M}_G(\Theta_G)$, there exists an instance $\tilde{M}_{G,\boldsymbol{\theta}} \in \tilde{\mathcal{M}}_G(\tilde{\Theta}_G)$, such that they coincide on each feasible trace $\mathcal{F}(G)$. If $\tilde{\mathcal{M}}$ and $\mathcal{M}$ are subclasses of each other, we say that the two model classes are equivalent.*

Note that Definition 2 does not specify the distribution that generates the seed set D_0; to describe the full data-generating distribution, we need a distribution $\mathbb{P}^0$ over the subsets of V from which the seed set is generated. We will assume that $\mathbb{P}^0$ does not depend on $\boldsymbol{\theta}$, and $\mathbb{P}^0(\emptyset) = 0$. We refer to the pair $(\mathcal{M}_G(\Theta), \mathbb{P}^0)$ as a *seeded diffusion model* with the seed distribution $\mathbb{P}^0$. The seeded diffusion model corresponds to a distribution on all feasible propagation traces:

$$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{D}) := \mathbb{P}^0(D_0) \prod_{t=1}^{T} \mathbb{P}^t_{\boldsymbol{\theta}}(D_t | D_0, \ldots, D_{t-1}) \mathbb{P}^{T+1}_{\boldsymbol{\theta}}(\emptyset | \mathcal{D}), \qquad \mathcal{D} \in \mathcal{F}(G). \tag{3}$$

In turn, the trace distribution also uniquely defines the seeded diffusion model.

Assuming one observes a collection of traces from a seeded diffusion model $(\mathcal{M}_G(\Theta), \mathbb{P}^0)$, the established equivalence with the trace-generating distribution allows directly using the standard statistical parameter estimation techniques, such as the maximum likelihood estimation (MLE). Before proceeding to estimation, a natural question statistician would address is when the family $\{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ of trace distributions is identifiable, that is, $\mathbb{P}_{\boldsymbol{\theta}_1} = \mathbb{P}_{\boldsymbol{\theta}_2}$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. While specific conditions will depend on the form of the distribution, it is clear that all nodes in the graph must be reachable with a positive

probability; otherwise, parameters associated with these nodes, such as the incoming edge weights, have no influence on the trace distribution and thus cannot be identified. The necessary and sufficient identifiability conditions for our proposed class of models will be stated formally in Theorem 11.

3. The General Linear Threshold model

In this section, we introduce the General Linear Threshold (GLT) model and establish its relationship to the IC, LT, and Triggering models. One may ask why there is a need for a new model, when there are already several, and in particular the triggering model seems quite general, encompassing both IC and LT models. The problem with the triggering model is its number of unknown parameters, $\sum_{v \in V} \left(2^{|P(v)|} - 1\right)$, which is not feasible to fit in most cases. On the other extreme, the LT and IC models have only $|E|$ parameters, but are often insufficiently flexible in practice. In particular, both these models assume that all nodes behave identically when receiving equal amounts of influence from their neighbors, which does not allow for any node heterogeneity. In many contexts, some individuals will need a lot more influence than others to become "activated", and this can be estimated from data on propagation traces. The proposed GLT model allows us to account for differences in users' willingness to accept new information while not significantly increasing the number of unknown parameters.

Definition 5 (General Linear Threshold (GLT) model) *Assume that each node $v \in V$ has a random threshold U_v drawn independently from a distribution supported on $[0, h_v]$ with $h_v \leq \infty$, with cumulative distribution function (cdf) F_v. Further assume that edges have weights $b_{u,v} \geq 0$, such that the in-degree of each node $v \in V$ satisfies $\sum_{u \in P(v)} b_{u,v} \leq h_v$. As in the LT model, a node v activates at time $t \geq 1$ if $\sum_{u \in P(v) \cap A_{t-1}} b_{u,v} \geq U_v$.*

If we treat all threshold distributions as fixed, the parameter space of the GLT model is the set of all possible edge weights satisfying the model constraints:

$$\Theta = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{|E|} \text{ s.t. for all } v \in V,\ \|\boldsymbol{\theta}_v\|_1 \leq h_v \text{ and } \boldsymbol{\theta}_v \geq 0 \right\}, \tag{4}$$

where the sub-vector of $\boldsymbol{\theta}$ corresponding to the parent edges of v is denoted by $\boldsymbol{\theta}_v = \{b_{u,v} : u \in P(v)\}$. If we also model the threshold distributions $F_v, v \in V$ in some way, there may be additional parameters. We will discuss this option in Section 4.5; for now, we will assume they are fixed.

Under the GLT model, the activation probability of a node v given a feasible history $(D_0, \ldots, D_{t-1})$ equals the probability that the sum of weights from A_{t-1} is higher than U_v conditional on the event that it had not exceeded U_v at time $t-2$:

$$\mathbb{P}^t_{\boldsymbol{\theta}}(v \in D_t \mid D_0, \ldots, D_{t-1}) = \frac{F_v\Big(B_v(A_{t-1}; \boldsymbol{\theta}_v)\Big) - F_v\Big(B_v(A_{t-2}; \boldsymbol{\theta}_v)\Big)}{1 - F_v\Big(B_v(A_{t-2}; \boldsymbol{\theta}_v)\Big)}, \tag{5}$$

where

$$B_v(S; \boldsymbol{\theta}_v) = \sum_{S \cap P(v)} b_{u,v} \tag{6}$$

is the influence node v with parent edge weights $\boldsymbol{\theta}_v$ receives from a node set S. Plugging the cdf $F_v(x) = x$ of the uniform distribution into (5) gives the activation probability under the LT model stated in Example 1, confirming the LT model class is equivalent to the GLT model class with uniformly distributed thresholds.

The following proposition establishes the less obvious relationship between the IC and the GLT models.

Proposition 6 *For any $\mu > 0$, the class of IC models is equivalent to the class of GLT models with all node thresholds distributed as $U_v \sim \text{Exponential}(\mu)$, that is, with $F_v(x) = 1 - e^{-\mu x}$.*

Proof By Definition 4, it is enough to construct a bijective mapping between each IC edge probability $p_{u,v} \in [0,1]$, and the GLT edge weight $b_{u,v} \in [0,\infty]$ for each $(u,v) \in E$. Let $b_{u,v} := -\log(1-p_{u,v})/\mu$. Then, if we model node threshold distributions as exponential with parameter μ, by the memoryless property of the exponential distribution, the activation probability of a node $v \in V$ given a feasible history $(D_0, \ldots, D_{t-1}) \in \mathcal{F}(G)$ can be written under the GLT model as

$$\begin{aligned}\mathbb{P}^t_{\boldsymbol{\theta}}(v \in D_t |\ D_0, \ldots, D_{t-1}) &= \mathbb{P}\Big(U_v \le B_v(A_{t-1}; \boldsymbol{\theta}_v) \mid\ U_v > B_v(A_{t-2}; \boldsymbol{\theta}_v)\Big) \\ &= \mathbb{P}\Big(U_v \le B_v(D_{t-1}; \boldsymbol{\theta}_v)\Big) = 1 - \prod_{u \in D_{t-1} \cap P(v)} (1 - p_{u,v}),\end{aligned}$$

which coincides with the transition kernel of the IC model. ■

We have now established that both IC and LT models are subclasses of the GLT model with identically distributed node thresholds. As we argued, a more interesting scenario is allowing this distribution to vary from node to node. A natural approach to make the number of GLT parameters manageable in this case is to choose threshold distributions from a parametric family. For example, if $U_v \sim \text{Beta}(\alpha_v, \beta_v)$, the model has only $|E| + 2|V|$ parameters. If we further assume that the network can be partitioned into communities and nodes within one community follow the same distribution (see Figure 1), we can further reduce the number of parameters. The following proposition demonstrates more generally that even if each node has an individual set of r parameters that is not too large compared to the average node in-degree, the GLT has negligibly more parameters than the IC and LT models, and many fewer than the Triggering model.

Proposition 7 *Consider a directed graph $G = (V, E)$ with the average node in-degree d. Then for any $0 \le r \le 2^d - d - 1$, it holds that*

$$|E| + r|V| \le \sum_{v \in V} \left(2^{|P(v)|} - 1\right).$$

In particular, the inequality is strict with $r = 1$ and $d > 2$ and with $r = 2$ and $d > 2.45$.

Proof Noting that $|E| = d|V|$ and applying Jensen's inequality to the convex function $f(x) = 2^x$, we have

$$\frac{|E| + r|V|}{\sum_{v \in V}\left(2^{|P(v)|} - 1\right)} \le \frac{(d+r)|V|}{(2^d - 1)|V|} \le 1.$$

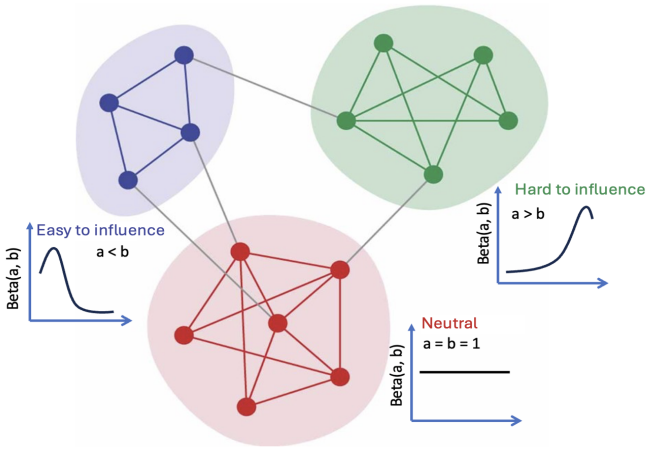


Figure 1: An example network with three communities with different levels of receptiveness to new information, all modeled with the Beta distribution.

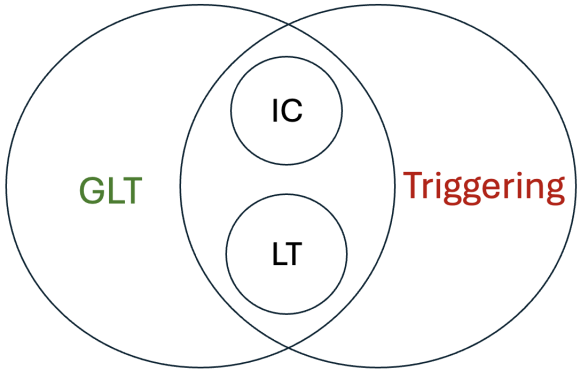


Figure 2: Relationship between different diffusion models.

■

The next proposition shows that, despite a much smaller number of parameters, the GLT model is not a subclass of the Triggering model. Unsurprisingly, the Triggering model is also not a subclass of the GLT. The proof can be found in Section A.1 of the Appendix.

Proposition 8 *The GLT model is not a subclass of the Triggering model and vice versa.*

Figure 2 summarizes the relationships between all the diffusion models discussed. We can think of the GLT and the Triggering model as two alternative flexible generalizations, both encompassing the popular IC and LT models, with GLT being preferable for statistical estimation and inference due to its better balance between flexibility and the number of parameters.

Although the estimation procedure and theory we further develop for the GLT model can be applied to a wide range of threshold distributions, in all our simulations, we use the Beta-GLT model with $U_v \sim \text{Beta}(\alpha_v, \beta_v)$. We believe Beta distribution is sufficiently rich to model various types of users' receptiveness to new information (see Figure 1) while encompassing all properties needed for efficient estimation, such as the bounded distribution support and convexity of the arising optimization problem (this will be discussed in detail in Section 4.2). Specifically, the Beta family includes the most popular uniform distribution (LT model) for $\alpha_v = \beta_v = 1$ and can closely mimic the exponential distribution with rate μ (IC model) for $\alpha_v = 1$ and $\beta_v = \mu - 1$ if μ is sufficiently large. Indeed, direct computation shows that the Kullback-Leibler (KL) divergence between the two distributions goes to zero as $\mu \to \infty$:

$$D_{KL}\big(\text{Beta}(1, \mu - 1) \;||\; \text{Exponential}(\mu)\big) = \log\Big(1 - \frac{1}{\mu}\Big) + \frac{1}{\mu - 1} = \frac{1}{2\mu^2} + O(\mu^{-3}).$$

This computation demonstrates that the IC model essentially assumes an "easy-to-influence" type of users, as it corresponds to the Beta distribution with $\alpha \ll \beta$, and points out the expressiveness of the Beta family, allowing us to model conservative activation mechanisms due to the additional $\alpha > \beta$ regime.

4. Estimation and theoretical properties under the GLT model

In this section, we study the GLT model on a given fixed network $G = (V, E)$. We present the necessary and sufficient conditions for identifiability of the weights, derive a constrained maximum likelihood estimator (MLE) of the weights from fully observed propagation traces and a finite sample bound for its error, and show that the estimator is asymptotically normal. We then extend the estimation procedure to the case of partially observed traces and unknown node threshold distributions.

4.1 Identifiability for the GLT model

We begin by studying identifiability of the GLT edge weights with respect to the trace distribution family $\{\mathbb{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ induced by the GLT model. Plugging the GLT transition probability from (5) into the general trace likelihood in (3), we can conveniently express the trace distribution as follows

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{D}) = \mathbb{P}^0(D_0) &\prod_{v \in C(A_T)} \{1 - F_v\left[B_v\left(A_T; \boldsymbol{\theta}_v\right)\right]\} \\ &\times \prod_{t=0}^{T-1} \prod_{v \in D_{t+1}} \{F_v\left[B_v\left(A_t; \boldsymbol{\theta}_v\right)\right] - F_v\left[B_v\left(A_{t-1}; \boldsymbol{\theta}_v\right)\right]\}.\end{aligned} \tag{7}$$

Here, the first term does not depend on $\boldsymbol{\theta}$ by definition of the seeded diffusion model. The second term represents nodes that were not activated but have at least one active parent in the trace. The third term captures activated nodes, i.e., nodes in $A_T \setminus D_0$.

For reasons that we will elaborate on later, it is hard to establish identifiability unless the probability of any feasible trace with $\mathbb{P}^0(D_0) > 0$ in (7) is bounded away from zero. To

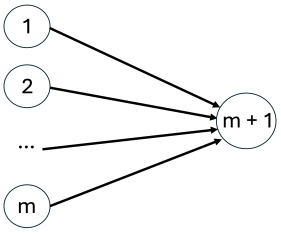


Figure 3: A star graph of in-degree m: $V = \{1, \ldots, m+1\}$, $E = \{(1, m+1) \ldots, (m, m+1)\}$.

guarantee that, we first need to assume that the threshold cdf F_v is strictly monotone for the nodes v that may appear in the trace likelihood, that is, the (child) nodes having at least one parent:

$$V_c = \{v \in V : P(v) \neq \emptyset\}. \tag{8}$$

Assumption 1 (Invertible cdf) *The threshold cdf F_v of every node $v \in V_c$ is strictly monotone (and thus invertible) on its support $[0, h_v]$.*

We also need to ensure that edge weights are strictly positive and that the parent weights of every node $v \in V_c$ sum to less than h_v. Thus we truncate the parameter space as follows:

$$\tilde{\Theta} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{|E|} \text{ s.t. for all } v \in V,\ \|\boldsymbol{\theta}_v\|_1 \leq \gamma \text{ and } \boldsymbol{\theta}_v \geq \varepsilon \right\}, \tag{9}$$

where $\varepsilon > 0$ and $\max_{v \in V_c} |P(v)|\varepsilon < \gamma < \min_{v \in V_c} h_v$ are some universal constants. Together with Assumption 1, $\boldsymbol{\theta}_v \geq \varepsilon$ will ensure that any node has a positive chance to activate its child even if none of the other parents are activated, and $\|\boldsymbol{\theta}_v\|_1 \leq \gamma$ will ensure that even if all parents of a node are activated, there is a positive probability that the node will not be activated.

The following lemma states these restrictions on the parameter space and Assumption 1 are sufficient for the likelihood in (7) to be positive for every feasible trace starting with a seed set from the support of $\mathbb{P}^0$.

Lemma 9 *Under Assumption 1, any feasible trace $\mathcal{D} \in \mathcal{F}(G)$ with $\mathbb{P}^0(D_0) > 0$ satisfies $\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{D}) > 0$ for each $\boldsymbol{\theta} \in \tilde{\Theta}$.*

The proof of this lemma and all other results in this section can be found in Section A.2 of the Appendix.

Lemma 9 does not guarantee that any node can appear in a feasible trace with positive probability. Indeed, if the seed sets from the support of $\mathbb{P}^0$ are not sufficiently rich, some nodes may be not reachable by any trace, making their parent weights non-identifiable. We say a node $u \in V$ is *reachable* if either $\mathbb{P}^0(u \in D_0) > 0$ or there is a directed path to u from at least one $v \in V$ with $\mathbb{P}^0(v \in D_0) > 0$. The following proposition formalizes this intuition as a necessary identifiability condition.

Proposition 10 *If $\{\mathbb{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \tilde{\Theta}\}$ is identifiable, all non-isolated nodes in V are reachable.*

The following example demonstrates that reachability by itself is not sufficient for identifiability.

Example 4 *Consider the star graph in Figure 3 with $m = 2$. Its edge set $E = \{(1,3),(2,3)\}$ has weights b_{13} and b_{23}, respectively. Fix the seed set $D_0 = \{1,2\}$, so that $\mathbb{P}^0(\{1,2\}) = 1$. Any seeded GLT model with this $\mathbb{P}^0$ induces a trace distribution that is a function of $b_{13} + b_{23}$:*

$$\mathbb{P}(\mathcal{D}) = \begin{cases} F_3(b_{13} + b_{23}), & \text{node } 3 \in D_1, \\ 1 - F_3(b_{13} + b_{23}), & \text{node } 3 \notin D_1, \end{cases}$$

and therefore b_{13} and b_{23} are not individually identifiable, only their sum is. It is easy to verify that if the support of $\mathbb{P}^0$ includes at least two distinct subsets of $\{1,2\}$, the weights are identifiable. More formally, there needs to exist $S_j \subseteq \{1,2\}, j = 1,2$ with $\mathbb{P}^0(S_j) > 0$ such that the 2×2 matrix $X = [\mathbf{1}(i \in S_j)]_{i,j=1}^2$ has full rank.

The condition on $\mathbb{P}^0$ can be directly extended to a star graph of arbitrary in-degree m, requiring the existence of $S_1, \ldots, S_m \subseteq \{1, \ldots, m\}$ such that an analogous $m \times m$ matrix X is of full rank. For graphs with more than one child node, it turns out that it is necessary and sufficient to require a similar condition for each child node v, with the only difference that the corresponding matrix X_v is now constructed using parent subsets S_j that can appear within any active set D_t preceding activation of v, not just the seed D_0. We formally state this condition in the following theorem.

Theorem 11 *Under Assumption 1, $\{\mathbb{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \tilde{\Theta}\}$ is identifiable if and only if for each child node $v \in V_c$ with $P(v) = \{u_1, \ldots, u_m\}$, there exist $S_1, \ldots, S_m \subseteq P(v)$ such that*

1. *For each $j = 1, \ldots, m$, there is a feasible trace $(D_0^{(j)}, \ldots, D_{t_j}^{(j)})$ with $\mathbb{P}^0(D_0^{(j)}) > 0$, $v \notin A_{t_j}^{(j)}$, and $D_{t_j}^{(j)} \cap P(v) = S_j$.*
2. *The matrix $\tilde{X}_v = [\mathbf{1}(u_i \in S_j)]_{i,j=1}^m$ is invertible.*

Remark 12 *The identifiability condition in Theorem 11 implies the necessary reachability condition from Proposition 10: if a source node u of an edge (u,v) is unreachable, the matrix $\tilde{X}_v$ is not invertible, as it has a row of zeros $[\mathbf{1}(u \in S_j)]_{j=1}^m$ for any choice of S_j.*

Intuitively, Theorem 11 states that identifiability holds if and only if, for each child node of in-degree m, it is possible to observe its m "essentially distinct" newly active parent subsets, that is, subsets with linearly independent binary activation vectors. Unfortunately, if this condition were to be used in practice, for example to determine if a given seed set distribution is sufficiently rich for weight identifiability, it would require finding all subsets of parent nodes that can simultaneously become activated in a trace. For complex network topologies, this can be a non-trivial combinatorial task, so we further propose a slightly less general but more easily verifiable sufficient condition, which ensures that all parent nodes can appear in a trace as singleton newly active sets. Notice that they are clearly "distinct," since the corresponding $\tilde{X}_v$ is a full-rank identity matrix.

Corollary 13 *Under Assumption 1, $\{\mathbb{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \tilde{\Theta}\}$ is identifiable if for every child node $v \in V_c$ and each of its parents $u \in P(v)$, either u can form a singleton seed set, that is, $\mathbb{P}^0(\{u\}) > 0$, or there is a seed set D_0 with $\mathbb{P}^0(D_0) > 0$ and $u, v \notin D_0$, such that u can be reached from a node in D_0 along a directed path not containing v.*

To conclude this section, we would like to stress the novelty of the result in Theorem 11. Most previously proposed methods for estimating the parameters of a diffusion model (Narasimhan et al., 2015; Gomez-Rodriguez et al., 2014; Daneshmand et al., 2014) assume that any node can appear in the seed set with a positive probability, which is much more restrictive, as in many applications not all nodes can be directly influenced. To the best of our knowledge, this paper is the first to derive both necessary and sufficient identifiability conditions for the parameters of a diffusion model.

4.2 Weight estimation under the GLT model

Next, we derive a maximum likelihood estimator for the weights in the GLT model, given a collection $\mathbb{D}$ of N observed (and therefore feasible) propagation traces,

$$\mathbb{D} = \{\mathcal{D}_n := (D_1^{(n)}, \dots, D_{T_n}^{(n)}) \mid n = 1, \dots, N\}, \tag{10}$$

where T_n is the number of time steps in trace $\mathcal{D}_n$. For now, we assume that all threshold distributions F_v are known and postpone the discussion of estimating the threshold distribution to Section 4.5.

We assume that the trace collection $\mathbb{D} = \{\mathcal{D}_n |\ n = 1, \dots, N\}$ is i.i.d., by which we mean

(a) Seed sets $\{D_0^{(n)} |\ n = 1, \dots, N\}$ are generated independently from the seed distribution $\mathbb{P}^0$;

(b) Node thresholds are generated independently for each trace and for each node, with

$$\mathcal{U}_n := \left(U_1^{(n)}, \dots, U_{|V|}^{(n)}\right) \overset{\text{iid}}{\sim} (F_1, \dots, F_{|V|}), \quad n = 1, \dots, N.$$

For an i.i.d. trace collection, the parameters can be estimated by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \tilde{\Theta}} \sum_{n=1}^{N} L(\mathcal{D}_n | \boldsymbol{\theta}), \tag{11}$$

where $L(\mathcal{D}_n|\boldsymbol{\theta})$, the log-likelihood of the trace $\mathcal{D}_n$, by (7), takes the form

$$\begin{aligned} L(\mathcal{D}_n|\boldsymbol{\theta}) = &\sum_{v \in C(A_{T_n}^{(n)})} \log\left\{1 - F_v\left[B_v\left(A_{T_n}^{(n)}; \boldsymbol{\theta}_v\right)\right]\right\} \\ &+ \sum_{t=1}^{T_n} \sum_{v \in D_t^{(n)}} \log\left\{F_v\left[B_v\left(A_{t-1}^{(n)}; \boldsymbol{\theta}_v\right)\right] - F_v\left[B_v\left(A_{t-2}^{(n)}; \boldsymbol{\theta}_v\right)\right]\right\}. \end{aligned} \tag{12}$$

Here, we omitted the $\log \mathbb{P}^0(D_0)$ term as it does not depend on $\boldsymbol{\theta}$. Note that in (11), we optimize over the truncated space $\tilde{\Theta}$, for which we need to know the "slack" variables ε and γ from Assumption 2. As their values are inaccessible in practice, in our implementation, we set $\varepsilon = 10^{-6}$ and $\gamma = h_v - \varepsilon$ for distributions with $h_v < \infty$ and $\gamma = 10$, otherwise. These constraints empirically improved numerical stability compared to optimization over the untruncated space Θ.

Examining (12), we see that the trace log-likelihood only involves weights $b_{u,v}$ for nodes v that were either activated after time 0 or failed to become activated but had an active parent. We will denote this set of "informative" nodes in trace $\mathcal{D}$ as

$$V_c(\mathcal{D}) := \left[A(\mathcal{D}) \setminus D_0^{(n)}\right] \cup C(A(\mathcal{D})). \tag{13}$$

Similarly, we can define the set of all informative nodes in the observed trace collection $\mathbb{D}$ as $V_c(\mathbb{D}) := \bigcup_{n=1}^{N} V_c(\mathcal{D}_n)$. In principle, this set may still be smaller than $|V_c|$, meaning that the parent weights of some nodes cannot be estimated from the data. But even in this case, with a sufficiently rich trace collection, we expect the total number of parameters in (11) to be close to $|E|$, which creates a major computational problem for large networks.

Fortunately, the optimization problem has a block structure we can use to speed up computations. By changing the order of summation, we can rewrite its objective as a sum of terms, each depending only on the parent edges of a single child node:

$$\sum_{n=1}^{N} L(\mathcal{D}_n|\boldsymbol{\theta}) = \sum_{v \in V_c(\mathbb{D})} L_v(\boldsymbol{\theta}_v)$$

with

$$\begin{aligned} L_v(\boldsymbol{\theta}_v) = &\sum_{n:\ v \in C(A_{T_n}^{(n)})} \log\left\{1 - F_v\left[B_v\left(A_{t(v,n)}^{(n)}; \boldsymbol{\theta}_v\right)\right]\right\} \\ &+ \sum_{n:\ v \in A_{T_n}^{(n)} \setminus D_0^{(n)}} \log\left\{F_v\left[B_v\left(A_{t(v,n)}^{(n)}; \boldsymbol{\theta}_v\right)\right] - F_v\left[B_v\left(A_{t(v,n)-1}^{(n)}; \boldsymbol{\theta}_v\right)\right]\right\}, \end{aligned} \tag{14}$$

where we denote the last time that node v is *not* active in trace n by $t(v,n) := \arg\max\{t \leq T_n : v \notin A_t^{(n)}\}$. Importantly, it is not just the likelihood that conveniently separates into blocks with independent variables, but also the feasibility set $\tilde{\Theta}$ that can be rewritten as a Cartesian product of child node-specific individual parameter spaces $\tilde{\Theta} = \bigotimes_{v \in V_c} \tilde{\Theta}_v$ defined as

$$\tilde{\Theta}_v = \{\boldsymbol{\theta}_v \in \mathbb{R}^{|P(v)|} :\ \boldsymbol{\theta}_v \geq \varepsilon,\ \|\boldsymbol{\theta}_v\|_1 \leq \gamma\}. \tag{15}$$

Therefore, solving (11) is equivalent to maximizing $L_v(\boldsymbol{\theta}_v)$ over $\boldsymbol{\theta}_v \in \tilde{\Theta}_v$ for each $v \in V_c(\mathbb{D})$:

$$\hat{\boldsymbol{\theta}}_v = \arg\max_{\boldsymbol{\theta}_v \in \tilde{\Theta}_v} L_v(\boldsymbol{\theta}_v), \quad v \in V_c(\mathbb{D}). \tag{16}$$

Each optimization problem in (16) now has only $|P(v)|$ variables and $|P(v)| + 1$ affine constraints, allowing for efficient parallelized optimization.

The next natural question is whether a node-specific optimization problem in (16) is convex. The feasible set $\tilde{\Theta}_v$ is a convex simplex, and the arguments of F_v in (14) depend linearly on $\boldsymbol{\theta}_v$. Thus, L_v is a concave function of $\boldsymbol{\theta}_v$ if $\log[F_v(x) - F_v(y)]$ is a concave function on $h_v \geq x > y \geq 0$. For example, if U_v is uniformly distributed on $[0,1]$, as in the standard LT model, then $\log[F_v(x) - F_v(y)] = \log(x - y)$ is concave. The following proposition demonstrates that it is true for all distributions with log-concave densities, and in particular when F_v is the Beta distribution with parameters $\alpha_v \geq 1$ and $\beta_v \geq 1$. The proof is given in Section A.3 of the Appendix.

Proposition 14 *The function $L_v(\boldsymbol{\theta}_v)$ in (14) is concave in $\boldsymbol{\theta}_v$ if F_v has a log-concave density.*

Note that for a child node v with just one parent u, (14) reduces to the log-likelihood of a Bernoulli$\big(F_v(b_{u,v})\big)$ random variable, which is concave in $b_{u,v}$ only if F_v is log-concave. Therefore, the log-concavity of the cdf F_v is clearly necessary as well as sufficient for concavity of $L_v(\boldsymbol{\theta}_v)$ for a node v with an arbitrary in-degree.

Remark 15 *Some special cases of the GLT model may have a different natural parametrization; for example, for the IC model the natural parameters are the edge transmission probabilities $p_{u,v}$, which can be expressed in terms of the weights of the GLT model with $U_v \sim \text{Exponential}(\mu)$ as $p_{u,v} = 1 - \exp(-\mu b_{u,v})$ according to Proposition 6. We can always estimate a reparametrized set of parameters by using the plug-in estimators, such as $\hat{p}_{u,v} = 1 - \exp(-\mu \hat{b}_{u,v})$ for the IC model.*

4.3 Theoretical properties of the GLT weight estimator

In this section, we derive a finite sample bound on the error of the MLE in (16) which holds with high probability, and also the asymptotically normal distribution of the estimator. Since the weight estimation procedure separates into a collection of optimization problems, each involving only the parent weights $\boldsymbol{\theta}_v$ of a given node $v \in V_c$, we will establish the theoretical properties of the estimate $\hat{\boldsymbol{\theta}}_v$ for a given fixed node $v \in V_c$ with $P(v) = \{u_1, \dots, u_m\}$ and the corresponding ground-truth parameter $\boldsymbol{\theta}_v^*$. The proofs of all the results in this section can be found in Section A.4 of the Appendix.

We begin by establishing the finite sample result. First, we introduce additional notation to conveniently encode the data used to fit the subproblem of node v in (16). Denote the trace indices where node v is informative by $\mathcal{I}_v := \{1 \le n \le N : \ v \in V_c(\mathcal{D}_n)\}$ and for each $\mathcal{D}_n, n \in \mathcal{I}_v$, denote the time points when v acquired at least one new active parent node by $\mathcal{T}_v^{(n)} := \{0 \le \tau \le T_n : D_\tau^{(n)} \cap P(v) \neq \emptyset\}$. For each time point $t \in \mathcal{T}_v^{(n)}$, let $x_t^{(n)} := [\mathbf{1}(u_i \in D_t^{(n)}]_{i=1}^m$ be the indicator vector of v's newly active parents at time t. Define also the matrix of $x_t^{(n)}$ stacked over $t \in \mathcal{T}_v^{(n)}$ as $X_v^{(n)} \in \{0,1\}^{|\mathcal{T}_v^{(n)}| \times m}$ and the further stacked matrix of $X_v^{(n)}, n \in \mathcal{I}_v$ as $X_v \in \{0,1\}^{N_v \times m}$ with $N_v := \sum_{n \in \mathcal{I}_v} |\mathcal{T}_v^{(n)}|$.

Importantly, the role of the "sample size" in our finite sample bound will be played not by the number of informative traces $|\mathcal{I}_v|$, but by N_v, the number of times v had a non-empty newly active parent set across all traces. Note that $1 \le |\mathcal{T}_v^{(n)}| \le m$ for any $n \in \mathcal{I}_v$, since v can have no more than m different newly active parent sets throughout a progressive propagation, but should have at least one such set, since the trace is informative. This means that in theory, N_v can be up to m times larger than $|\mathcal{I}_v|$; however, our simulations (see Section B.2 of the Appendix) demonstrate that in practice the multiplicative factor fluctuates between 1 and 3, depending on the user conservativeness (GLT distribution and weight magnitudes) and network topology (node in-degrees). The dependence of the estimation error on the cumulative number of informative observations rather than traces aligns with other work, for example, with the finite sample bound for the parent weight estimator derived for the General Cascade model in Theorem 1 of (Pouget-Abadie and Horel, 2015). Similarly to their framework, we assume that within each informative trace $n \in \mathcal{I}_v$,

we observe the propagation history up to the last time point $t(v,n)$ when v is not active, and that the only randomness is in the activation event of the node v. Since the matrix X_v essentially encodes all such histories, we condition on X_v whenever we need to emphasize that these histories are observed.

Before we state the main result, we make a couple of additional mild regularity assumptions. First, we require that the threshold cdfs are not only invertible as in Assumption 1, but also sufficiently smooth, and, secondly, that the ground-truth parent weights of every node lie in the interior of the corresponding parameter space:

Assumption 2 *For each $v \in V_c$, the threshold cdf F_v is strictly monotone and three times continuously differentiable, and the ground-truth weights $\boldsymbol{\theta}_v^*$ satisfy $\boldsymbol{\theta}_v^* > \varepsilon$ and $\|\boldsymbol{\theta}_v^*\|_1 < \gamma$.*

We also require that the negative log-likelihood is almost surely non-strictly convex and strictly convex on average for any $\boldsymbol{\theta}_v$ in the parameter space:

Assumption 3 (Convexity) *The density of F_v is log-concave to ensure that, per Proposition 14, $-L_v(\boldsymbol{\theta}_v)$ is non-strictly convex. Moreover, the expected Hessian of $-L_v(\boldsymbol{\theta}_v)$ should be positive definite everywhere on $\tilde{\Theta}_v$. With compactness of $\tilde{\Theta}_v$ and continuity of the Hessian guaranteed by Assumption 2, this means that there exists $\lambda_{\min} > 0$ such that for every $\boldsymbol{\theta}_v \in \tilde{\Theta}_v$, it holds conditional on X_v:*

$$-\mathbb{E}\big[\frac{1}{N_v}\nabla^2 L_v(\boldsymbol{\theta}_v)\big] \succeq \lambda_{\min} I_{|P(v)|}.$$

For the LT, IC, and Beta-GLT models, Assumption 3 can be replaced by a much more intuitive sufficient condition requiring non-degeneracy of X_v:

Proposition 16 *For $F_v \sim \text{Exponential}(\mu)$ with $\mu > 0$ or $F_v \sim \text{Beta}(\alpha, \beta)$ with $\alpha, \beta \geq 1$, Assumption 3 is satisfied if X_v has a full column rank. Moreover, $\lambda_{\min}$ can be set as the smallest eigenvalue of a Gram matrix $c_\lambda X_v^\top X_v / N_v$ where c_λ is a constant that depends only on $(F_v, \varepsilon, \gamma)$.*

Notice that the condition in Proposition 16 is guaranteed to hold asymptotically if the identifiability condition of Theorem 11 is satisfied. Indeed, with positive probability, X_v includes each row of the $m \times m$ invertible identifiability matrix $\tilde{X}_v$ defined in Theorem 11, and thus has the full column rank itself.

Now, we are ready to state the main theoretical result.

Theorem 17 *Consider the MLE $\hat{\boldsymbol{\theta}}_v$ obtained by solving the optimization problem* (16) *and fix an arbitrary $\delta \in (0,1)$. Under Assumptions 2, 3, and the assumption of i.i.d. traces, $\hat{\boldsymbol{\theta}}_v$ satisfies the following concentration bound conditional on X_v as long as $N_v \geq \frac{c_0 m}{\lambda_{\min}} \log \frac{2m}{\delta}$:*

$$\mathbb{P}\Big[\|\hat{\boldsymbol{\theta}}_v - \boldsymbol{\theta}_v^*\|_2 \leq \frac{C_0}{\lambda_{\min}} \sqrt{\frac{m}{N_v} \log \frac{4m}{\delta}}\Big] \geq 1 - \delta. \tag{17}$$

Here, $C_0, c_0 > 0$ are constants depending only on $(F_v, \varepsilon, \gamma)$.

Remark 18 *Since in our framework, the underlying network and thus the node's indegree $m = |P(v)|$ are fixed, the established result presents sufficient conditions for the estimator $\hat{\boldsymbol{\theta}}_v$ to be $\sqrt{N_v}$-consistent. However, in principle, we could consider a sequence of networks with associated GLT models on them and apply this result to each network to explore how the estimator's error changes when the sample size N_v and the indegree m grow simultaneously. For this scenario, the derived concentration bound suggests that consistency holds as long as N_v asymptotically dominates $m \log m$. Unfortunately, Proposition 16 shows that in the final sample case, the necessary Assumption 3 may not be satisfied when m is too close to N_v, as this would lead to the singularity of the log-likelihood Hessian.*

In addition to a finite sample result, we derive an asymptotic distribution for the MLE error. Contrary to a finite sample statement of Theorem 17, establishing an asymptotic result requires considering the full trace-generating distribution in (7) with the total number of traces N being a more natural candidate for the sample size than N_v. Another difference is that we need much weaker conditions on the log-likelihood convexity – now it is sufficient to require it only locally in the neighborhood of $\boldsymbol{\theta}^*$. We state this as a separate assumption:

Assumption 4 (Local convexity) *The negative expectation of the trace log-likelihood Hessian is positive definite at $\boldsymbol{\theta}^*$:*

$$\mathbb{E}_{\mathcal{D}\sim\mathbb{P}_{\boldsymbol{\theta}^*}}\left[-\nabla^2 \log \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{D})\right]|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \succ 0.$$

The following proposition characterizes the asymptotic behavior of the MLE.

Proposition 19 *Consider the MLE $\hat{\boldsymbol{\theta}}_v$ obtained by solving the optimization problem* (16)*. Then, under Assumptions 2, 4, identifiability condition on the seed distribution $\mathbb{P}^0$ in Theorem 11, and the assumption of i.i.d. traces, it holds*

$$\hat{\Sigma}_v(\hat{\boldsymbol{\theta}}_v)^{-1/2}(\hat{\boldsymbol{\theta}}_v - \boldsymbol{\theta}_v^*) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}^*}} \mathcal{N}(0, I_m) \quad as \quad N \to \infty, \tag{18}$$

where the limit is taken with respect to the ground-truth trace distribution $\mathbb{P}_{\boldsymbol{\theta}^*}$ defined in (3) *and the estimated covariance matrix is $\hat{\Sigma}_v(\boldsymbol{\theta}_v) = [-\nabla^2 L_v(\boldsymbol{\theta}_v)]^{-1}$.*

There are many downstream tasks for which the asymptotic distribution of the GLT weights may be useful. For example, Chen et al. (2016) propose a method for solving the Influence Maximization problem (defined formally in Section 5) assuming that the edge weights are only known to lie in some intervals. This method can be naturally coupled with estimation of these confidence intervals from the data using the normal approximation in (18).

Uncertainty quantification for the estimated weights may also be of interest. As a simple example, consider comparing the effects of two parent nodes u and w on a child node v. We may then want to test the hypothesis of no difference between the corresponding edge weights,

$$H_0 : \; b_{u,v} - b_{w,v} = 0 \quad \text{vs} \quad H_a : \; b_{u,v} - b_{w,v} \neq 0,$$

and use the difference of their estimates as the test statistic. Then, the reference distribution is normal with zero mean and variance that can be derived from the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_v$ in (18):

$$\mathrm{Var}[\hat{b}_{u,v} - \hat{b}_{w,v}] \;\approx\; \hat{\Sigma}_{v,uu} + \hat{\Sigma}_{v,ww} - 2\hat{\Sigma}_{v,uw},$$

where $\hat{\Sigma}_{v,ij}$ denotes the (i,j)-th entry of $\hat{\Sigma}_v(\hat{\boldsymbol{\theta}}_v)$.

In other applications, knowing the asymptotic distribution of the weights can help quantify uncertainty in predicting various quantities of interest; for example, in epidemiology we may be interested in predicting the probability of node activation (infection) in the next time step given the propagation history. This is a complicated function of the weights but one can still compute a confidence interval for it using the delta method.

4.4 Extension to partially-observed traces.

In many applications, we do not observe a full propagation trace, but we know which of the node's parents were active before it was activated. We write each such observation for node v as a pair (A_v, y), referred to as *pseudo-trace*, where $A_v \subset P(v)$ is a set of v's active (infected) parents and $y \in \{0,1\}$ is an indicator of the event that A_v together activate (infect) v.

Suppose that for each child node $v \in V_c$, we observe a possibly empty collection of pseudo-traces

$$\mathbb{D}_v = \{(A_v^{(n)}, y_v^{(n)}),\ n = 1, \dots, N_v\} \tag{19}$$

that we would like to use to estimate the GLT weights $\boldsymbol{\theta}_v$. If we could specify a pseudo-trace generating distribution for $\mathbb{D}_v$, we could apply the likelihood approach. One way to do this is to treat a pseudo-trace (A_v, y) as a trace seeded at A_v and propagating in a star graph G_v attached to the child node v. Given that $A_v \subset P(v)$, the feasible traces on G_v can only be of two types: those that stopped immediately at the seed set A_v (corresponding to the pseudo-trace $(A_v, 0)$) and those that activated v at $t = 1$ and then stopped (pseudo-trace $(A_v, 1)$). Assuming that the sets $A_v^{(n)},\ n = 1, \dots, N_v$ are independently generated from a parameter-free seed distribution $\mathbb{P}_v^0$ supported on the subsets of $P(v)$, the pseudo-trace likelihood has the form

$$\mathbb{P}_{\boldsymbol{\theta}_v}(A_v, y) = \mathbb{P}_v^0(A_v) \left\{1 - F_v\left[B_v(A_v; \boldsymbol{\theta}_v)\right]\right\}^{1-y} F_v\left[B_v(A_v; \boldsymbol{\theta}_v)\right]^y . \tag{20}$$

Aggregation of these terms across all pseudo-traces in $\mathbb{D}_v$ results in the log-likelihood

$$L_v^{pt}(\boldsymbol{\theta}_v) = \sum_{n:\ y^{(n)}=0} \log\big\{1 - F_v\big[B_v\big(A_v^{(n)}; \boldsymbol{\theta}_v\big)\big]\big\} + \sum_{n:\ y^{(n)}=1} \log F_v\big[B_v\big(A_v^{(n)}; \boldsymbol{\theta}_v\big)\big], \tag{21}$$

where we omitted the parameter-free terms $\log \mathbb{P}_v^0(A_v)$. This pseudo-trace likelihood coincides with the that of the General Cascade model (see Section 2.3 of (Pouget-Abadie and Horel, 2015)), which assumes that the activation probability of a node is an increasing function f mapping the sum of incoming edge weights from active parents to $[0,1]$. This implies that the model is the analogue of the GLT model for pseudo-trace case with f set as the cdf F_v.

The assumption that the seed distribution $\mathbb{P}_v^0$ does not depend on any diffusion model parameters may seem strong. However, by carefully comparing (21) with its counterpart for fully observed traces in (14), we observe that the only difference is that, in the latter case, we always subtract $F_v\big[B_v(A_{t(v,n)-2}^{(n)})\big]$ under the logarithm for traces where v was activated. This term represents the probability that the active parent set preceding the one that eventually activated v was *not* enough for v's activation. Thus, the only information lost in a pseudo-trace, compared to a fully observed trace, is which parent subset of an influenced node v was *not* sufficient to activate it. Importantly, since pseudo-traces are assumed as

traces propagating on a star graph G_v, both consistency and asymptotic normality results of Theorem 17 and Proposition 19 still hold in the pseudo-trace case.

4.5 Estimation of threshold parameters

So far, we have treated threshold distributions as known, which is unlikely in reality. While it would be challenging to estimate these distributions fully nonparametrically given we typically only observe a limited number of traces concerning any given node, we could easily obtain an estimate if we model each $F_v, v \in V$ as a member of some parametric family with parameters $\boldsymbol{\varphi}_v \in \Phi_v \subset \mathbb{R}^{r_v}$. For example, if we model $F_v \sim \text{Beta}(\alpha_v, \beta_v)$, we can define $\boldsymbol{\varphi}_v = (\alpha_v, \beta_v)$ with $\Phi_v = [1, +\infty)^2$ to satisfy the condition of Proposition 14. Then we can estimate $(\boldsymbol{\theta}_v, \boldsymbol{\varphi}_v)$ for each $v \in V_c(\mathbb{D})$ by solving the following optimization problem:

$$\max_{\boldsymbol{\varphi}_v \in \Phi_v,\ \boldsymbol{\theta}_v \in \Theta_v} L_v\left(\mathbb{D} | \boldsymbol{\theta}_v, \boldsymbol{\varphi}_v\right) , \tag{22}$$

where the individual node likelihood L_v is obtained from (14) with the Beta distribution cdfs plugged in. Allowing F_v to vary within the feasible set makes the optimization problem non-convex even in the simplest case of a one-parameter Beta family. Thus finding even a local optimum of (22) requires careful tuning of the gradient steps since $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ might have very different magnitudes. A natural way to deal with that problem is to switch to coordinate gradient descent, alternating between fixing one set of variables ($\boldsymbol{\theta}_v$ or $\boldsymbol{\varphi}_v$) and optimizing over the other one. Conveniently, according to Proposition 14, optimization over $\boldsymbol{\varphi}_v \in \Phi_v$ with fixed $\boldsymbol{\theta}_v$ is also a convex problem, since the Beta distribution density is log-concave with respect to the parameters (α_v, β_v). Although in general such biconvex problems are not guaranteed to converge to the global optimum or even to the same local optimum for different initializations, we empirically verify that this is nearly always the case for (22).

More generally, unless the dimension r_v of Φ_v is very high, we can choose Φ_v from a discrete grid, optimize (22) over $\boldsymbol{\theta}$ for each $\boldsymbol{\varphi}_v \in \Phi_v$ and choose the one resulting in the highest log-likelihood. While in further numerical experiments we do assume all nodes' thresholds follow the Beta distribution, it is important to note that due to the node-wise separability of the optimization problem in (22), the parametric family as well as the parameter grid Φ_v are not required to be the same across $v \in V_c(\mathbb{D})$.

5. Influence maximization under the GLT model

In this section, we study the GLT model in the context of the Influence Maximization (IM) problem, that is, the task of choosing a seed set of a given size that maximizes the expected spread of information through the network. We start with a brief review of the IM problem, which was introduced by Richardson and Domingos (2002) and further formalized by Kempe et al. (2003). Formally, we define the *influence function*, for a given simple directed graph $G = (V, E)$ and a diffusion model M, as the function that maps any subset $S \subset V$ to the expected number of nodes influenced if M is seeded by $D_0 = S$,

$$\sigma_M(S) = \mathbb{E}_{\mathcal{D}|D_0=S}|A(\mathcal{D})|, \qquad S \subset V.$$

The IM problem is then to find a subset $S^* \subset V$ which maximize the influence function over all subsets of V of a given size k,

$$S^* = \underset{S \subset V:\ |S| \leq k}{\operatorname{argmax}}\ \sigma_M(S).$$

Kempe et al. (2003) showed that the IM problem is NP-hard under the LT, IC, and Triggering models. However, if the influence function has certain properties, the optimal solution can be well approximated by a greedy strategy (see Algorithm 1). These properties are monotonicity and submodularity.

Definition 20 (Monotonicity) *An influence function $\sigma(\cdot)$ is monotone if $\sigma(S') \leq \sigma(S)$ for any $S' \subset S \subseteq V$. In words, increasing the size of the seed set cannot decrease the value of the influence function.*

Definition 21 (Submodularity) *An influence function $\sigma(\cdot)$ is submodular if $\sigma(\{v\} \cup S') - \sigma(S') \geq \sigma(\{v\} \cup S) - \sigma(S)$ for any $S' \subset S \subseteq V$ and $v \in V \backslash S$. In words, adding a node to a given seed set increases the influence function by at least as much as adding it to a superset of this seed set.*

These are both reasonable and mild assumptions, reflecting the intuitive meaning of information propagation. Though Kempe et al. (2003) were the first to study the greedy algorithm behavior in the context of the IM problem, worst-case performance under monotonicity and submodularity assumptions dates back to the following classical result:

Theorem 22 (Cornuéjols et al. (1977) and Nemhauser et al. (1978)) *Let $\sigma : 2^V \to \mathbb{R}+$ be a a monotone and submodular influence function. Let $\hat{S} \subset V$ of size k be the set obtained by selecting elements from V one at a time, when at each step one chooses an element that provides the largest marginal increase in the value of σ. Let S^* be the true maximizer of σ over all k-element subsets of V. Then*

$$\sigma(\hat{S}) \geq \left(1 - \frac{1}{\mathrm{e}}\right) \sigma(S^*), \tag{23}$$

that is, $\hat{S}$ provides a $1 - \frac{1}{\mathrm{e}}$ approximation to the optimal S^*.

Kempe et al. (2003) showed that under the LT, IC, and Triggering models, the influence function has the monotonicity and submodularity properties, thus showing that the greedy algorithm can solve the IM problem under these models with the optimality guarantee (23). A natural question about our new GLT model is whether it enjoys similar properties. Monotonicity is trivially satisfied for the GLT model class, and the following theorem gives a sufficient condition for submodularity, which is also necessary in a particular sense. The proof of this theorem and all other results in this section can be found in Section A.5 of the Appendix.

Theorem 23 *A GLT model on a graph $G = (V, E)$ has a submodular influence function if the threshold cdf F_v is concave for each child node $v \in V_c$. Moreover, for any non-concave cdf F with a non-negative support, there exists an instance of the GLT model on a star graph as in Figure 3 with F as the threshold cdf of the only child node, such that the corresponding influence function is not submodular.*

Algorithm 1 The greedy algorithm for the IM problem

Input: graph $G = (V, E)$, diffusion model instance M, and seed budget k
$S \leftarrow \emptyset$
while $|S| < k$ **do**
  $v \leftarrow \arg\max_{v \in V \setminus S} (\sigma_M(S \cup \{v\}) - \sigma_M(S))$
  $S \leftarrow S \cup \{v\}$
return S

Theorem 23 implies submodularity of the IC and LT models since both uniform and exponential distributions have concave cdfs. For GLT models with beta-distributed thresholds, we can use this general statement to derive a simpler submodularity condition.

Corollary 24 *A GLT model on a graph $G = (V, E)$ with $U_v \sim \text{Beta}(\alpha_v, \beta_v)$ has a submodular influence function if $\alpha_v \leq 1$ and $\beta_v \geq 1$ for all $v \in V_c$.*

Proof The second derivative of the cdf of the $\text{Beta}(\alpha, \beta)$ distribution is given by

$$F''(x|\alpha, \beta) \ = C(\alpha, \beta)\ x^{\alpha-2}(1-x)^{\beta-2}[(\alpha-1)(1-x) - (\beta-1)x]$$

where $C(\alpha, \beta) > 0$ does not depend on x, and thus $F''(x|\alpha, \beta) \ \leq 0$ for all $x \in (0, 1)$ if $\alpha \leq 1$ and $\beta \geq 1$. ■

Since we estimate the parameters of GLT model from data, the estimation error in $\hat{\boldsymbol{\theta}}$ can affect the IM problem solution. To study this, let $\sigma_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} \in \Theta$ denote the influence function of a diffusion model $M_{G,\boldsymbol{\theta}}$. Let $S^*(\boldsymbol{\theta}) = \arg\max_{|S| \leq k} \sigma_{\boldsymbol{\theta}}(S)$ denote the solution of the IM problem under $M_{G,\boldsymbol{\theta}}$, where we omit the dependency on k and treat it as fixed throughout this section. Then the question is to relate the difference $|\sigma_{\boldsymbol{\theta}}(S^*(\hat{\boldsymbol{\theta}})) - \sigma_{\boldsymbol{\theta}}(S^*(\boldsymbol{\theta}))|$ to the parameter estimation error $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|$. Unfortunately, general results which hold for an arbitrary graph topology and any choice of the true parameters usually imply loose and impractical bounds (see, for example, Lemma 3 and subsequent discussion in Chen et al. (2016)). Therefore, we present a less general but more illustrative result for a family of directed bipartite graphs, that is, graphs with two disjoint sets of nodes and all edges going from a node in the first set to a node in the second. The bipartite graph structure can be thought of as the most general graph topology that ensures the propagation traces are at most of unit length. The following proposition essentially states that in this setting, the discrepancy between the spreads from the IM solution under the true and estimated GLT models is governed by the ℓ_1 error of the weight estimates.

Proposition 25 *Consider a directed bipartite graph and a GLT model on it, and assume every child node $v \in V_c$ has a L-Lipschitz cdf, that is, $|F_v(x) - F_v(y)| \leq L|x - y|$ for any $0 \leq x, y \leq h_v$. Then with the notations above, it holds:*

$$|\sigma_{\boldsymbol{\theta}}(S^*(\hat{\boldsymbol{\theta}})) - \sigma_{\boldsymbol{\theta}}(S^*(\boldsymbol{\theta}))| \leq 2L\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1.$$

In particular, if all threshold cdfs of a GLT model are differentiable and concave, and therefore satisfy the submodularity condition in Theorem 23, the Lipschitz constant can be taken as $L = \max_{v \in V_c} F_v'(0)$.

Combining this result with the finite-sample error bound derived in (17), we can conclude that the spread of the IM solution obtained under the estimated model converges in probability to the spread of the ground-truth model solution at a rate of $\min_{v \in V_c} \sqrt{N_v}$.

6. Experiments

In this section, we present numerical results on both simulated and real-world data. The code for these analyses is available at `https://github.com/AlexanderKagan/gltm_experiments`. The Python package `InfluenceDiffusion`, available at `https://github.com/AlexanderKagan/InfluenceDifusion`, includes the convex optimization method and the greedy algorithm to fit the GLT model, as well as code for trace sampling and spread estimation. Whenever we fit the GLT model by solving the optimization problem in equation (16), we use the SciPy implementation of the SLSQP solver.

6.1 Simulation settings

We generate synthetic networks from the connected Watts-Strogatz model (Watts and Strogatz, 1998). As the original model is for an undirected graph, we double each sampled undirected edge to go both ways. We denote the distribution of a directed graph G generated this way by $G \sim \mathrm{CWS}(n, k, p)$, where n is the number of nodes, k is the initial degree of each node, and $p \in [0, 1]$ is the probability of edge rewiring, controlling the randomness of the graph. By construction, the number of edges in G is fixed to kn and therefore the edge density is fixed at k/n. The imposed connectedness of G is not strictly necessary for modeling information diffusion, but it is convenient in simulations, ensuring that enough nodes are reachable.

To generate edge eights for G, we independently and uniformly sample parent edge weights of each child node from a simplex scaled by a given positive constant $d_{\max}$, which upper bounds the node's weighted in-degree:

$$\boldsymbol{\theta}_v \sim \mathrm{Unif}\{\boldsymbol{w} \in \mathbb{R}^{|P(v)|} : \boldsymbol{w} \geq 0,\ \|\boldsymbol{w}\|_1 \leq d_{\max}\}. \tag{24}$$

To generate seed sets for the traces, we independently and uniformly sample them from node sets of sizes between 1 and $s_{\max}$:

$$D_0 \sim \mathrm{Unif}\{S \subset V : 1 \leq |S| \leq s_{\max}\}. \tag{25}$$

Unless otherwise stated, we use $d_{\max} = 1$ and $s_{\max} = 5$ as default values.

The difference between two vectors $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ (the truth and the estimator) will be measured by Relative Mean Absolute Error (RMAE), defined as $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_1 / \|\boldsymbol{y}\|_1$.

6.2 Estimation of edge weights

In this section, we study how the quality of weight estimation depends on key parameters of the underlying graph and the ground-truth GLT model. For simplicity, we use the original LT model here, with uniform thresholds, as the observed trends in weight estimation are very similar across different threshold distributions.

In the first experiment, we fix the number of traces at $N = 2000$ while varying the number of graph nodes n and the Watts-Strogatz model average node in-degree k. The

weights are sampled as in (24) with $d_{\max} = 1$. In the second experiment, we fix the graph size to $n = 100$ nodes and the in-degree to $k = 10$, and vary the number of traces N and the maximum weighted in-degree of the nodes $d_{\max}$, sampling the weights from (24) with $d_{\max} \in \{0.2, 0.4, 0.6, 0.8, 1\}$. In both scenarios, we use $p = 0.2$ as the probability of edge rewiring.

The results are presented in Figure 4. In the left panel of Figure 4, we observe that the estimation error increases as the density k/n or the network size n grow. This is expected, as both higher density and larger size increase the number of edge weights to estimate, thus requiring more traces for accurate estimation. The right panel of Figure 4 shows that lower weights lead to higher estimation errors. This is because larger weights result in higher node activation probabilities, producing longer traces with more data that can be used in estimation.

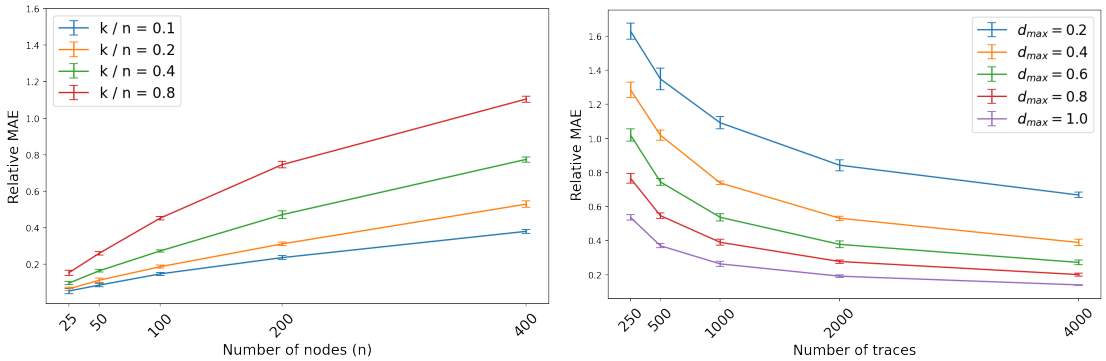


Figure 4: Left: Relative MAE of the LT estimator as a function of the number of nodes n for different densities k/n, with $N = 2000$ traces and $d_{\max} = 1$. Right: Relative MAE as a function of the number of traces N for different maximum node in-degrees $d_{\max}$, with $n = 100$ and density $k/n = 0.1$. The error bars represent two standard errors and are calculated from 10 repetitions of each experiment.

6.3 Uncertainty quantification and robustness to model misspecification

Here, we focus on node activation probabilities as the main object of interest and present experiments that use asymptotic theory developed in Section 4.3 to quantify the uncertainty in their estimation. We also study how node activation probabilities behave under misspecification of the threshold distribution.

For this experiment, we sample a 100-node CWS network with $p = 0.2$ and $k = 5$. The GLT model thresholds all have Beta$(2, 1)$ distribution, and the weights are sampled as in (24) with $d_{\max} = 1$. With seed sets generated as in (25) with $s_{\max} = 10$, we sample $N = 1500$ traces from this model for training and generate additional 500 traces for testing. Then, for each candidate distribution Beta$(2, 1)$, Beta$(3, 1)$, Unif$[0, 1]$ (LT model), and Exponential(1) (IC model), we use the training set to estimate the weights of the GLT model under the assumption of the candidate distribution for the node thresholds. Then, we run through each test trace $\mathcal{D}_n$, $n = 1, \ldots, 500$ and do the following:

1. For each "informative" node $v \in V_c(\mathcal{D}_n)$, as defined in (13), extract the last time step it is not activated, that is, $t(v,n) = \arg\max\{t \le T_n : \ v \notin A_t^{(n)}\}$,

2. Under each of the estimated GLT models, compute the probability v is activated at $t(v,n)+1$ conditional on the history of $\mathcal{D}_n$ as defined in (5).

We also do this with the ground-truth GLT model to obtain the true conditional activation probabilities. Finally, for each of the four candidate models, we plot the true probabilities against their predicted values along with the corresponding 95% asymptotic confidence intervals computed using the Delta method, as described at the end of Section 4.3, shown in Figure 5. We also evaluate RMAE between the predicted and true probabilities, as well as the confidence interval average length and coverage, that is, the proportion of times it contains the ground truth probability. According to RMAE and coverage metrics, the best performance is clearly obtained by using the true model, Beta$(2,1)$, while the average confidence interval length roughly equals 0.04 for all four models. When misspecified as the "hard to influence" threshold model Beta$(3,1)$ (shown in Figure 1), activation probabilities close to 0 get underestimated, and those close to 1 overestimated. With the "easy to influence" threshold models given by the uniform and the exponential, which correspond to LT and IC models, all activation probabilities tend to be underestimated. Confidence interval coverage meets the nominal target of 95% under the true model, and is considerably lower under misspecified models.

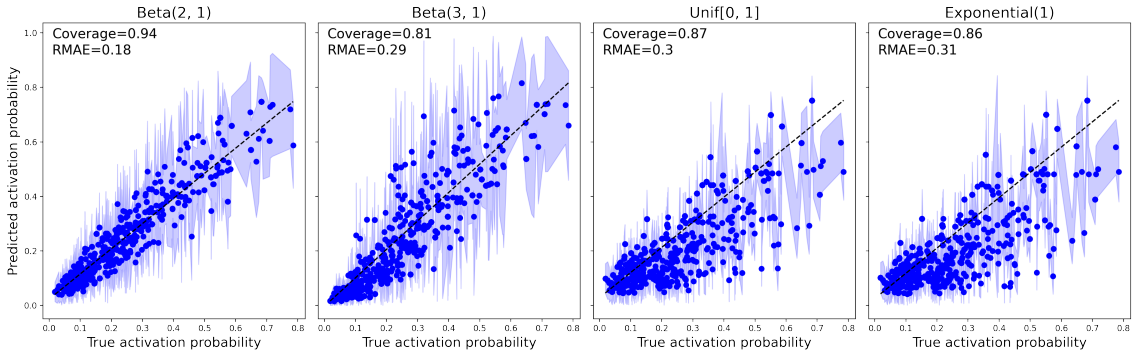


Figure 5: Estimated node activation probabilities together with the corresponding Delta method confidence intervals computed under different GLT models with Beta$(2,1)$-GLT as the ground truth.

6.4 The GLT model in the Influence Maximization problem

In this section, we demonstrate that using an appropriate diffusion model, which can be learned from trace data, can significantly improve the quality of the seed set obtained by Algorithm 1. To explore the behavior of IM solutions across different network instances, we sample 10 networks $G_\ell, \ell = 1,\dots,10$ from the CWS model with $n = 100$, $p = 0.2$, and $k = 10$. For each network, we generate $N = 2000$ traces from the ground-truth GLT model

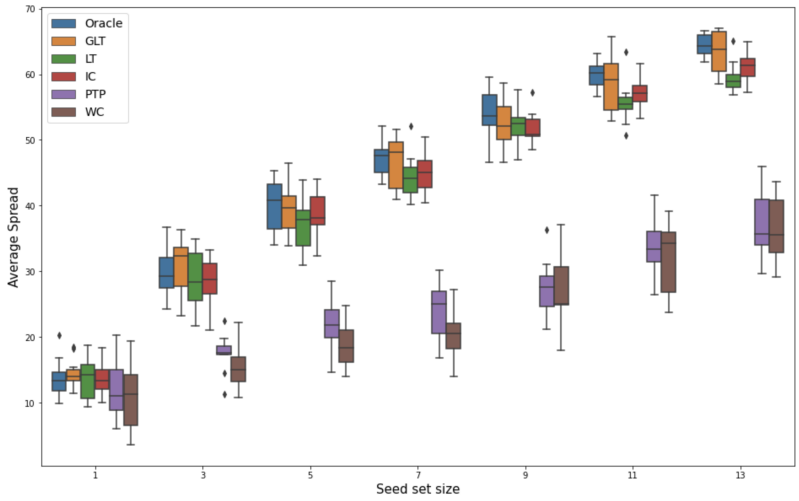


Figure 6: Comparison of the average spread across different seed set sizes, where the seeds are selected by a greedy algorithm under five candidate diffusion models and the ground truth. The ground truth threshold distributions are $F_v \sim \text{Beta}(1, \beta_v)$, where $\beta_v \sim \text{Unif}\{1, 2, 3, 4, 5\}$. Each box plot represents the estimated spread across 10 networks, where the spread is averaged over five fitted models (each trained on a separate set of 2000 traces). All values of the spread function $\sigma(\cdot)$ are estimated using 1000 Monte Carlo simulations.

with weights sampled according to (24) with $d_{max} = 1$, and the threshold distribution is set to $F_v \sim \text{Beta}(1, \beta_v)$, with $\beta_v \sim \text{Uniform}\{1, 2, 3, 4, 5\}$.

To examine how misspecification of the diffusion model impacts IM solutions, we compare the following methods: the LT model and the IC model fitted by solving (16), and the GLT model with both weights and threshold distributions estimated by solving problem (22), where the threshold distribution for each node v is assumed to be $\text{Beta}(1, \beta_v)$, with β_v estimated from data. As benchmarks, we also include the LT model with weights assigned via the following heuristics from Goyal et al. (2011):

1. Weighted Cascade (**WC**): The weight of an edge $(u, v) \in E$ is estimated as the inverse of the in-degree of v, i.e., $\hat{b}_{u,v} = 1/|P(v)|$.

2. Propagated Trace Proportion (**PTP**): The weight of edge (u, v) is estimated based on the ratio between the number of traces where u is activated before v and the number of traces where u is activated. Normalization is used to ensure that the in-degree of each node equals 1:

$$\hat{b}_{u,v} \propto \frac{|\{n : u \in D_{t_u}^{(n)}, v \in D_{t_v}^{(n)},\ t_u < t_v\}|}{|\{n : u \in A(\mathcal{D}_n)\}|}, \qquad \sum_{u \in P(v)} \hat{b}_{u,v} = 1.$$

For each of the diffusion models described above and each generated network, we follow these steps. First, we estimate the model weights and, in the case of GLT, also the β_v values via a grid search over $\{1, 2, \ldots, 10\}$. Then we

1. Run Algorithm 1 with seed set size $k = \{1, 4, 7, 10, 13\}$ under the fitted diffusion model to obtain a seed set $\hat{S}_k$. The influence function σ in Algorithm 1 is approximated using 1000 Monte Carlo simulations.

2. Run the diffusion 1000 times from the seed set obtained in Step 1 and obtain the average spread $\hat{\sigma}_k$.

As a benchmark, we also compute the average spread under the ground truth GLT model, referred to as the *oracle*.

Figure 6 presents boxplots (across the 10 networks) of $\hat{\sigma}_k$ for different values of k, ranging from 1 to 13. As the seed set size k increases, the choice of model has a greater impact, likely because for very small seed set size the greedy algorithm tends to select the most connected nodes under any model. However, significant differences in spread emerge as the seed set size increases. As expected, LT and IC are inferior to the ground truth and the GLT model, and the heuristics work poorly. The GLT model with estimated threshold distributions achieves performance comparable to that of the oracle.

6.5 Spread estimation

In the previous experiment, we only evaluated the spreads from seed sets that were selected by the IM algorithm. In some applications, we may also be interested in assessing spread, for example, of fake news or a virus, initiated from a given seed set that has not been optimally selected. Here we show that, for propagation under the GLT model, selecting an accurate threshold distribution can significantly improve the accuracy of spread estimation from any seed set.

As in Section 6.3, we consider a CWS network with $n = 100$, $p = 0.2$, and $k = 10$, and the associated GLT model with weights sampled according to (24). This time, we set the ground-truth threshold distributions as $F_v \sim \text{Beta}(2, 2)$ for all child nodes $v \in V_c$. With seed sets generated as in (25) with $s_{\max} = 20$, we generate a set of 1000 train traces and additional 500 test seed sets. We pick a more "easy-to-influence" model and a higher than usual $s_{\max}$ to make the observed traces spread farther, and so allow studying the effect of model misspecification on the full range of trace lengths. Then, for each candidate distribution Beta$(2, 2)$, Beta$(1, 2)$, Beta$(2, 1)$, and Unif$[0, 1]$ (LT model), we use the training set to estimate the weights of the GLT model under the assumption of the candidate distribution for the node thresholds. We then compute the predicted spread by running the estimated GLT models 1000 times from each test seed set. We also do this with the ground-truth GLT model to obtain the true spreads. In Figure 7, we plot the estimated spreads against the ground truth for each candidate GLT model and report the RMAE between them. Similarly to the node activation probabilities, the spread from a given seed set is also a function of the GLT model weights; however, it is a complicated implicit function that does not lend itself to a delta method calculation, so instead we estimated the spread empirically, by repeating the simulation 10 times. The variability was negligible on the scale of the plots, and is thus not shown; this is expected since spread estimators are known to

be robust to small perturbations in diffusion model weights (see, for example, Goyal et al. (2011)).

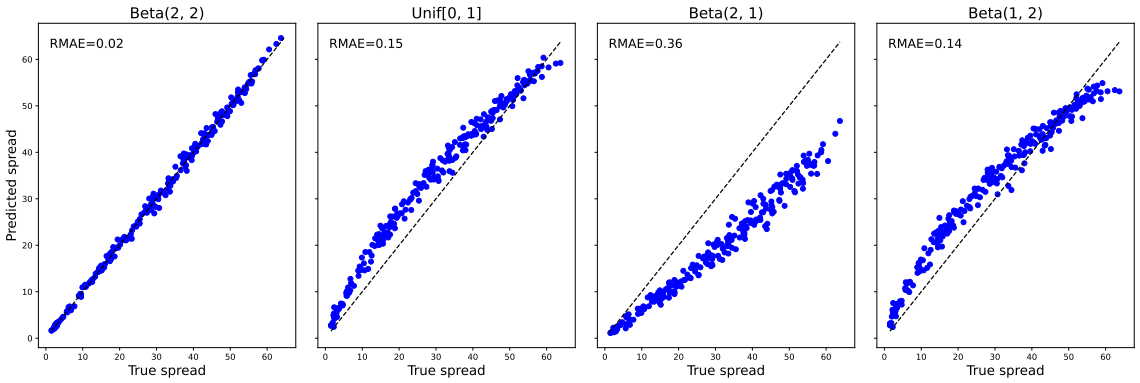


Figure 7: Comparison between the ground-truth spreads from 500 test trace sets and the estimated spreads for several candidate GLT models with a fixed Beta threshold distribution for all nodes. The trace-generating model is the GLT with $F_v \sim \text{Beta}(2, 2)$ for all nodes. Weights for each candidate GLT model are estimated using 1000 training traces.

In Figure 7, we observe that the ground-truth model clearly outperforms the other candidates, even the LT model with the same mean of $1/2$, suggesting that getting the shape of the threshold distribution right is important. As a more general conclusion, we can notice that the more "easy-to-influence" models, such as $\text{Beta}(1, 2)$, tend to overestimate the spread and the more "hard-to-influence" models, such as $\text{Beta}(2, 1)$, tend to underestimate it.

6.6 The movie ratings example

In this section, we apply the proposed weight estimation procedures to the Flixster dataset, collected from www.flixster.com, a popular social media platform for movie ratings. The dataset contains an undirected, unweighted social network of approximately 1 million users and over 8 million time-stamped ratings of movies by the users, in the time period from 2005 to 2009. Following Goyal et al. (2011), we represent these ratings as *action logs*, which is a collection of triples (u, a, t), where u represents the user ID, a the movie ID, and t the time when the user rated the movie; the value of the rating is disregarded. Goyal et al. (2011) analyzed this dataset under the assumption that if user u rated movie a before user v, and u and v are connected, then v was influenced by u. There is of course the possibility that v rated the movie independently of u, but they showed that even without accounting for this possibility, learning from the action logs can significantly improve the estimation of influence propagation. Here, our goal is to compare the accuracy of different diffusion models for predicting node activation events.

To process the data, we first removed all users who rated fewer than 20 movies, then applied the algorithm (Kojaku and Masuda, 2018) to extract the core sub-graph of the

remaining users, and finally extracted the largest connected component of the core. This resulted in a network of 8174 nodes, approximately 50K undirected edges (which we doubled to create directed edges), and approximately 2.1M action logs.

The next step is to transform the action logs into trace or pseudo-trace data. Inferring full propagation traces from action logs is challenging, even in simple scenarios. For example, consider a graph of three connected users who rated the same movie at distinct times $t_1 < t_2 < t_3$. There are already five possible ways to construct the corresponding trace – $(\{1,2,3\}), (\{1\},\{2,3\}), (\{1,2\},\{3\}), (\{1\},\{2,3\})$, or $(\{1\},\{2\},\{3\})$ – because of the inherent ambiguity of converting the continuous time stamps into discrete propagation events. While one could pick a threshold to decide whether two time stamps should be considered the same or not in the discrete time space, it would be arbitrary, and the resulting analysis can be sensitive to this arbitrary choice. Instead, we use the pseudo-trace framework described in Section 4.4.

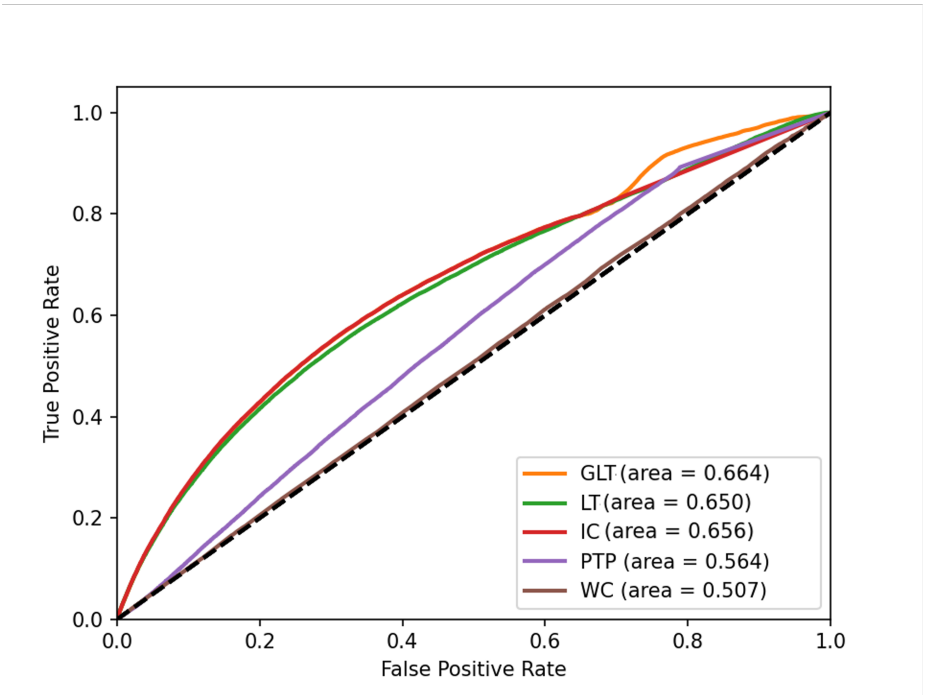


Figure 8: ROC curves and AUC scores for the activation probabilities computed on the test pseudo-trace set, from candidate diffusion models estimated on the training pseudo-trace set.

For each user (node) v, we extract the pseudo-traces where v was activated by identifying all movies a they rated and noting the set $A_v^{(a)} \subset P(v)$ of v's parents who rated a before v did. For pseudo-traces where v was not activated, we consider all movies a that v did not rate but at least one of their parents did, noting the set $A_v^{(a)}$ of all its active parents at the last recorded time point in the action log. We then randomly split the trace data into training (80%) and test (20%) sets, stratified by the activation status of the node. Using

the same candidate diffusion models as in Section 6.4, we fit the models on the training pseudo-traces and compute predicted node activation probabilities on the test pseudo-traces. The only difference from the settings in Section 6.4 is a larger parameter grid for estimation of the Beta distribution parameters, where we allows both α_v and β_v to range from 1 to 10. Note that for the fitting problem (22), we only require F_v to be log-concave, which is satisfied when both $\alpha \geq 1$ and $\beta \geq 1$ for the Beta distribution; it is only for solving the IM problem that F_v is required to be concave.

Figure 8 presents the resulting ROC curves and AUC scores for the estimated diffusion models. These results are for one random split into training and test; variability in AUC scores under different random splits was less than 0.001. The results show that GLT performs the best, followed fairly closely by IC and LT, and these three models significantly outperform the heuristic PTP and WC. While this dataset does not represent an ideal test case for our model, since the traces are not observed directly, it demonstrates that flexibility in modeling the thresholds can help even when neither the data collection mechanism nor, presumably, the true propagation model exactly match the GLT framework.

7. Discussion

In this paper, we have proposed a new flexible framework for information diffusion on networks, the general linear threshold (GLT) model. We derived identifiability conditions for edge weights which are weaker than previously available, developed a statistically principled likelihood-based method to estimate the edge weights from fully or partially observed traces, and proved that these estimates are $\sqrt{n}$-consistent and asymptotically normal when threshold distributions are known. We also proposed a parametric approach to estimating threshold distributions, which saves computational time but is relatively inflexible compared to nonparametric distribution estimators; we leave that for future work. In the parametric setting, it would also be of interest to establish identifiability and consistency conditions for the GLT model where both the parent edge weights $\boldsymbol{\theta}_v$ and the threshold distribution parameters $\boldsymbol{\varphi}_v$ vary with v.

In Section 5, we illustrated the application of the GLT to the IM problem, and established the relationship between the quality of the IM solution under the GLT model errors for the case of bipartite graphs; establishing this relationship for more general graph classes is a topic for future work. Another important question to study is stability of the IM solution to misspecification of node threshold distributions.

To allow for easy use with arbitrary distribution of thresholds, our implementation of the likelihood optimization problem (16) uses the SLSQP solver, which accommodates both convex and non-convex problems. This choice was made to allow for fitting the GLT model with non-log-concave threshold densities, which may violate the convexity condition in Proposition 14 but be desirable in practice. When convexity is guaranteed, however, optimization efficiency can be significantly improved by using convex solvers, such as Gurobi or Mosek. Improving optimization speed and quality is another goal for future work.

Acknowledgments

Authors thank Jonathan Terhorst for useful discussions and advice. This work was partially supported by the US National Science Foundation DSM award 2210439.

References

Wei Chen, Tian Lin, Zihan Tan, Mingfei Zhao, and Xuren Zhou. Robust influence maximization. 2016. arXiv:1601.06551.

Gérard Cornuéjols, Marshall L. Fisher, and George L. Nemhauser. Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management Science*, 23:789–810, 1977. doi: 10.1287/mnsc.23.8.789.

Hadi Daneshmand, Manuel Gomez-Rodriguez, Le Song, and Bernhard Schölkopf. Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm. 2014. arXiv:1405.2936.

Eleftherios Gkioulekas. On equivalent characterizations of convexity of functions. *International Journal of Mathematical Education in Science and Technology*, 44(3):410–417, 2013. doi: 10.1080/0020739X.2012.703336.

Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, 2001. doi: 10.1023/A:1011122126881.

Manuel Gomez-Rodriguez, Jure Leskovec, David Balduzzi, and Bernhard Schölkopf. Uncovering the structure and temporal dynamics of information propagation. *Network Science*, 2:26–65, 2014. doi: 10.1017/nws.2014.3.

Amit Goyal, Francesco Bonchi, and Laks V. S. Lakshmanan. A data-based approach to social influence maximization. *Proc. VLDB Endow.*, 5(1):73–84, 2011. doi: 10.14778/2047485.2047492.

Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978. doi: 10.1086/226707.

Xinran He, Ke Xu, David Kempe, and Yan Liu. Learning influence functions from incomplete observations. 2016. arXiv:1611.02305.

David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, page 137–146, New York, NY, USA, 2003. Association for Computing Machinery. doi: 10.1145/956750.956769.

Sadamori Kojaku and Naoki Masuda. Core-periphery structure requires something else in the network. *New Journal of Physics*, 20(4):043012, 2018. doi: 10.1088/1367-2630/aab547.

Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer Science & Business Media, New York, 2 edition, 1998.

Yang Liu and Yi-Fang Brook Wu. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2018. doi: 10.1609/aaai.v32i1.11268.

Elchanan Mossel and Sebastien Roch. Submodularity of influence in social networks: From local to global. *SIAM Journal on Computing*, 39:2176–2188, 2010. doi: 10.1137/080714452.

Harikrishna Narasimhan, David C. Parkes, and Yaron Singer. Learnability of influence in networks. In *Proceedings of the 29th International Conference on Neural Information Processing Systems*, NIPS'15, page 3186–3194, Cambridge, MA, USA, 2015. MIT Press.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14: 265–294, 1978. doi: 10.1007/BF01588971.

Jean Pouget-Abadie and Thibaut Horel. Inferring graphs from cascades: A sparse recovery framework. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 977–986, Lille, France, 2015. PMLR.

András Prékopa. On logarithmic concave measures and functions. 1973.

Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 61–70, New York, NY, USA, 2002. Association for Computing Machinery. doi: 10.1145/775047.775057.

Kazumi Saito, Ryohei Nakano, and Masahiro Kimura. Prediction of information diffusion probabilities for independent cascade model. In *Knowledge-Based Intelligent Information and Engineering Systems*, volume 5179 of *Lecture Notes in Computer Science*, page 67–75, Berlin, Heidelberg, 2008. Springer-Verlag. doi: 10.1007/978-3-540-85567-5_9.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z.

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998. doi: 10.1038/30918.

Appendix A. Proofs

In this section, we present all the proofs not present in the main body of the manuscript.

A.1 Section 3 Proofs

Proof [Proof of Proposition 8] Consider a GLT model on the star graph of in-degree 3 in Figure 3 with equal weights $b_{u,v} = 1/3$. Let node 4 have a threshold cdf F satisfying

$$F(0) = 0,\ F(1/3) = 0.5,\ F(2/3) = 0.85,\ F(1) = 1.$$

Threshold distributions of nodes 1, 2, and 3 do not affect the diffusion model as other nodes cannot activate them. Any triggering model on this graph will have only 8 relevant parameters, representing probabilities of each possible subset of $\{1, 2, 3\}$ to be the triggering set for node 4. We will denote this probability distribution as

$$\mathcal{P} = \{P_\emptyset, P_1, P_2, P_3, P_{12}, P_{13}, P_{23}, P_{123}\}.$$

If the GLT model was a special case of the Triggering model, a distribution $\mathcal{P}$ would exist such that the activation probability of node 4 is the same for the two models given any seed set, i.e., the following linear system should have a solution:

$$\begin{cases} P_1 + P_{12} + P_{13} + P_{123} & = F(b_{1,4}) = 0.5 \\ P_2 + P_{12} + P_{23} + P_{123} & = F(b_{2,4}) = 0.5 \\ P_3 + P_{13} + P_{23} + P_{123} & = F(b_{3,4}) = 0.5 \\ P_1 + P_2 + P_{12} + P_{13} + P_{23} + P_{123} & = F(b_{1,3} + b_{2,4}) = 0.85 \\ P_1 + P_3 + P_{12} + P_{13} + P_{23} + P_{123} & = F(b_{1,3} + b_{3,4}) = 0.85 \\ P_2 + P_3 + P_{12} + P_{13} + P_{23} + P_{123} & = F(b_{1,3} + b_{2,4}) = 0.85 \\ P_1 + P_2 + P_3 + P_{12} + P_{13} + P_{23} + P_{123} & = F(b_{1,4} + b_{2,4} + b_{3,4}) = 1 \end{cases}$$

Solving this system, we can verify that a solution exists, but it has $P_{123} = -0.05$, and thus is not a valid probability distribution.

Now, assume that the Triggering model is a special case of the GLT model. Consider an instance of the Triggering model on the same star graph of in-degree 3 with $\mathbb{P}(\Gamma_4 = \{1\}) = 0.4$ and $\mathbb{P}(\Gamma_4 = \{2, 3\}) = 0.6$. Then, there exist weights $b_{1,4}, b_{2,4}, b_{3,4} \geq 0$ and a cdf F, such that

$$F(b_{1,4}) = 0.4, \quad F(b_{2,4}) = 0.6, \quad F(b_{2,4} + b_{3,4}) = 0.6, \quad \text{and} \quad F(b_{1,4} + b_{3,4}) = 1.$$

By monotonicity of the cdf, the first two equations imply $b_{1,4} < b_{2,4}$. Adding $b_{3,4}$ on both sides and applying a non-decreasing F implies the following contradiction:

$$1 = F(b_{1,4} + b_{3,4}) \leq F(b_{2,4} + b_{3,4}) = 0.6.$$

■

A.2 Section 4.1 Proofs

Proof [Proof of Lemma 9] Note that the trace probability in (7) consists of at most $|V_c|$ terms (excluding the positive seed set probability), as only child nodes can be activated at $t \geq 1$ and each node in a trace can be activated at most once. In turn, for each node $v \in V_c$, the corresponding term is either $F_v(B_v(A_t; \boldsymbol{\theta}_v)) - F_v(B_v(A_{t-1}; \boldsymbol{\theta}_v))$ or $1 - F_v(B_v(A_T; \boldsymbol{\theta}_v)$. By trace feasibility, $D_t \cap P(v)$ is non-empty. Coupled with the strict monotonicity of F_v, it implies that for any $\boldsymbol{\theta}_v \in \tilde{\Theta}_v$ that

$$F_v\left[B_v(A_{t-1}; \boldsymbol{\theta}_v)\right] < F_v\left[B_v(A_t; \boldsymbol{\theta}_v)\right] \quad \text{and} \quad F_v\left[B_v(A_T; \boldsymbol{\theta}_v)\right] \leq F_v\left[B_v(P(v); \boldsymbol{\theta}_v)\right] < 1.$$

Thus, both types of terms are positive, which completes the proof. ■

Before we proceed to prove Proposition 10, we establish a useful lemma, showing the equivalence between the node reachability and its appearance in a trace.

Lemma 26 *The following conditions on node $u \in V$ are equivalent:*

(a) For all $\boldsymbol{\theta} \in \tilde{\Theta}$, it holds $\mathbb{P}_{\boldsymbol{\theta}}(u \in A(\mathcal{D})) > 0$;

(b) There exists $\boldsymbol{\theta} \in \tilde{\Theta}$ such that $\mathbb{P}_{\boldsymbol{\theta}}(u \in A(\mathcal{D})) > 0$;

(c) Node u is reachable.

Proof [Proof of Lemma 26] Statement (a) trivially implies (b). To show (b) implies (c), note that (b) implies that there is a feasible trace $\mathcal{D} = (D_0, \ldots, D_t, \ldots, D_T)$ with $\mathbb{P}^0(D_0) > 0$ and $u \in D_t$. We prove by induction over $\tau \geq 1$ that there is a path to u from a node $w_{t-\tau} \in D_{t-\tau}$. First, for $\tau = 1$, since $\mathcal{D}$ is feasible, u should have a parent w_{t-1} in D_{t-1}. Now suppose the induction hypothesis holds for τ. If there is a path to u from $w_{t-\tau} \in D_{t-\tau}$, there is also a path from $D_{t-\tau-1}$ since $w_{t-\tau}$ should have a parent in $D_{t-\tau-1}$ by feasibility. This results in a path $(w_0, \ldots, w_{t-1}, u)$ connecting D_0 and u which implies that u is reachable.

To prove (c) implies (a), consider an arbitrary reachable u. If $\mathbb{P}^0(u \in D_0) > 0$, then (a) holds. Assume now there is a sequence of nodes $(w_0, w_1, \ldots, w_T = u)$ such that $(w_{t-1}, w_t) \in E$ for all $t = 1, \ldots, T$ and $w_0 \in D_0$ with $\mathbb{P}^0(D_0) > 0$. Then, $\mathcal{D} = (D_0, \{w_1\}, \ldots, \{w_T\})$ is a feasible trace that has a positive probability for any $\boldsymbol{\theta} \in \tilde{\Theta}$ according to Lemma 9. ■

Proof [Proof of Proposition 10] Suppose there is an unreachable non-isolated node $u \in V$. Then, by Lemma 26, there is $\boldsymbol{\theta} \in \tilde{\Theta}$ such that $\mathbb{P}_{\boldsymbol{\theta}}(u \in A(\mathcal{D})) = 0$. If u has a child v, the edge (u, v) will never participate in a trace and thus changing $b_{u,v}$, while keeping all other weights fixed, preserves $\mathbb{P}_{\boldsymbol{\theta}}$. If u has a parent v, then v should also be unreachable, as otherwise, there would be a directed path from some positive probability seed set to u passing through v. But then changing $b_{v,u}$ alone again does not change $\mathbb{P}_{\boldsymbol{\theta}}$. In both cases, we get a contradiction with identifiability. ■

Proof [Proof of Theorem 11] (Sufficiency) Suppose there are distinct vectors of parameters $\boldsymbol{\theta} = \{b_{u,v} : (u, v) \in E\}$ and $\tilde{\boldsymbol{\theta}} = \{\tilde{b}_{u,v} : (u, v) \in E\}$ in $\tilde{\Theta}$ for which $\mathbb{P}_{\boldsymbol{\theta}} = \mathbb{P}_{\tilde{\boldsymbol{\theta}}}$. Since $\boldsymbol{\theta} \neq \tilde{\boldsymbol{\theta}}$, there is an edge $(u, v) \in E$ such that $b_{u,v} \neq \tilde{b}_{u,v}$. Consider the subsets $S_j, j = 1, \ldots, m$ of $P(v)$ together with the corresponding traces $\mathcal{D}_j = (D_0^{(j)}, \ldots, D_{t_j}^{(j)})$ satisfying conditions of the theorem. For each $\mathcal{D}_j$, let $t_{jk}, k = 1, \ldots, r_j$ with $t_{jr_j} = t_j$ denote all time points before t_j when $D_{t_{jk}} \cap P(v) \neq \emptyset$. As the equality of distributions $\mathbb{P}_{\boldsymbol{\theta}}$ and $\mathbb{P}_{\tilde{\boldsymbol{\theta}}}$ implies equality of the corresponding seeded diffusion model instances, we have for any time point t_{jk}:

$$\mathbb{P}_{\boldsymbol{\theta}}(v \notin D_{t_{jk}+1} | D_0, \ldots, D_{t_{jk}}) = \mathbb{P}_{\tilde{\boldsymbol{\theta}}}(v \notin D_{t_{jk}+1} | D_0, \ldots, D_{t_{jk}}),$$

which implies by (5) that

$$\frac{1 - F_v\left(B_v(A_{t_{jk}}^{(j)}; \boldsymbol{\theta}_v\right)}{1 - F_v\left(B_v(A_{t_{jk}-1}^{(j)}; \boldsymbol{\theta}_v))\right)} = \frac{1 - F_v\left(B_v(A_{t_{jk}}^{(j)}; \tilde{\boldsymbol{\theta}}_v)\right)}{1 - F_v\left(B_v(A_{t_{jk}-1}^{(j)}; \tilde{\boldsymbol{\theta}}_v))\right)}. \tag{26}$$

Taking the product of these equalities across t_{jk} for $k \le r_j$ and $k < r_j$, we obtain by telescoping

$$F_v\left(B_v(A_{t_j}^{(j)};\boldsymbol{\theta}_v)\right) = F_v\left(B_v(A_{t_j}^{(j)};\tilde{\boldsymbol{\theta}}_v)\right) \quad \text{and} \quad F_v\left(B_v(A_{t_j-1}^{(j)};\boldsymbol{\theta}_v)\right) = F_v\left(B_v(A_{t_j-1}^{(j)};\tilde{\boldsymbol{\theta}}_v)\right),$$

where for the second equality we used $A_{t_{j,r_j-1}}^{(j)} \cap P(v) = A_{t_j-1}^{(j)} \cap P(v)$. By the monotonicity of F_v, we can apply its inverse to both sides of the equations above to deduce

$$B_v(A_{t_j}^{(j)};\boldsymbol{\theta}_v) = B_v(A_{t_j}^{(j)};\tilde{\boldsymbol{\theta}}_v) \quad \text{and} \quad B_v(A_{t_j-1}^{(j)};\boldsymbol{\theta}_v) = B_v(A_{t_j-1}^{(j)};\tilde{\boldsymbol{\theta}}_v).$$

By definition of $\mathcal{D}_j$, we have $D_{t_j}^{(j)} \cap P(v) = S_j$, so, by subtracting the above equations we obtain $B_v(S_j;\boldsymbol{\theta}_v) = B_v(S_j;\tilde{\boldsymbol{\theta}}_v)$. Aggregating the resulting equalities across $j = 1,\dots,m$, we obtain $\tilde{X}_v^\top \boldsymbol{\theta}_v = \tilde{X}_v^\top \tilde{\boldsymbol{\theta}}_v$. But according to our assumption, $\tilde{X}_v$ is invertible, thus $\boldsymbol{\theta}_v = \tilde{\boldsymbol{\theta}}_v$ and $b_{u,v} = \tilde{b}_{u,v}$ in particular. Contradiction.

(Necessity) Suppose $\{\mathbb{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \tilde{\Theta}\}$ is identifiable but there is $v \in V_c$ with $P(v) = \{u_1,\dots,u_m\}$ for which conditions of the theorem do not hold. Take an arbitrary $\boldsymbol{\theta}_v$ in the interior of $\tilde{\Theta}_v$, which is non-empty by definition of ε_0. Our further goal is to obtain a contradiction by constructing $\tilde{\boldsymbol{\theta}} \in \tilde{\Theta}$ coinciding with $\boldsymbol{\theta}$ everywhere except for $\boldsymbol{\theta}_v$ so that $\mathbb{P}_{\tilde{\boldsymbol{\theta}}} = \mathbb{P}_{\boldsymbol{\theta}}$. Consider all possible subsets $S_j \subset P(v), j = 1,\dots,k$ which satisfy condition 1 of the theorem. Then, our assumption implies that the matrix $\tilde{X}_v = [\mathbf{1}(u_i \in S_j)] \in \{0,1\}^{m\times k}$ has $\operatorname{rank}(\tilde{X}_v) < m$. In other words, there is a non-zero vector $z \in \mathbb{R}^m$ such that $\tilde{X}_v^\top z = \mathbf{0}_k$. As $\boldsymbol{\theta}_v$ lies in the interior of $\tilde{\Theta}_v$ by definition, we can find a sufficiently small scalar $\delta > 0$, such that $\tilde{\boldsymbol{\theta}}_v = \boldsymbol{\theta}_v + \delta z$ still lies in $\tilde{\Theta}_v$ while preserving

$$\tilde{X}_v^\top \boldsymbol{\theta}_v = \tilde{X}_v^\top \tilde{\boldsymbol{\theta}}_v. \tag{27}$$

Note that by changing $\boldsymbol{\theta}_v$ alone, we could only change the probability of a feasible trace $\mathcal{D} = (D_0,\dots,D_T)$ with $\mathbb{P}^0(D_0) > 0$ and $P(v) \cap D_t \neq \emptyset$ for some $t \le t_v := \max_{\tau \le T}\{\tau : v \notin A_\tau\}$ where t_v is the last time v is not activated. Indeed, if $\mathbb{P}^0(D_0) = 0$ then $\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{D}) = \mathbb{P}_{\tilde{\boldsymbol{\theta}}}(\mathcal{D}) = 0$ and if $P(v) \cap D_t = \emptyset$ for any $t \le t_v$, then by (7), trace probability does not depend on $\boldsymbol{\theta}_v$. Take such a trace and consider all times $s_j \le t_v, j = 1\dots,r$ for which $P(v) \cap D_{s_j} \neq \emptyset$. Note that for each time s_j, the trace $(D_0,\dots,D_{s_j})$ is also feasible and satisfies condition 1 of the theorem. Therefore, there is a corresponding column $x_j := [\mathbf{1}(u_i \in D_{s_j})]_{i=1}^m$ in matrix $\tilde{X}_v$ which according to (27) satisfies

$$B_v(D_{s_j};\boldsymbol{\theta}_v) = \langle x_j, \boldsymbol{\theta}_v\rangle = \langle x_j, \tilde{\boldsymbol{\theta}}_v\rangle = B_v(D_{s_j};\tilde{\boldsymbol{\theta}}_v).$$

Summing these equations over $j \le r$ and $j < r$, we obtain, respectively,

$$B_v(A_{t_v};\boldsymbol{\theta}_v) = B_v(A_{t_v};\tilde{\boldsymbol{\theta}}_v) \quad \text{and} \quad B_v(A_{t_v-1};\boldsymbol{\theta}_v) = B_v(A_{t_v-1};\tilde{\boldsymbol{\theta}}_v).$$

But from (7), trace probability is either a function of $B_v(A_T;\boldsymbol{\theta}_v) = B_v(A_{t_v};\boldsymbol{\theta}_v)$ if $v \notin A_T$ or of $B_v(A_{t_v};\boldsymbol{\theta}_v)$ and $B_v(A_{t_v-1};\boldsymbol{\theta}_v)$ if $v \in A_T$. Therefore, we deduce $\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{D}) = \mathbb{P}_{\tilde{\boldsymbol{\theta}}}(\mathcal{D})$ and arrive at a contradiction with identifiability of $\tilde{\Theta}$. ■

Proof [Proof of Corollary 13] Fix an arbitrary child node $v \in V_c$ with parents $P(v) = \{u_1, \ldots, u_m\}$. Define the singleton parent subsets $S_j := \{u_j\},\ j = 1, \ldots, m$. Since the corresponding $\tilde{X}_v$ is an identity matrix, it is sufficient to verify that $S_1, \ldots, S_m$ satisfy the first condition of Theorem 11. Take arbitrary S_j and corresponding u_j. If $\mathbb{P}^0(S_j) > 0$, the required condition trivially holds. If not, the formulated condition guarantees that there is a seed set D_0 with $\mathbb{P}^0(D_0) > 0$, such that u_j is reachable from some $w_0 \in D_0$ via a directed path $(w_0, \ldots, w_{t-1}, u_j)$ and $D_0 \cup \{w_1, \ldots, w_{t-1}\}$ does not contain u_j or v. Without loss of generality, we may also assume that $w_1, \ldots, w_{t-1}$ are not contained in D_0, as otherwise we can start this path from w_{τ^*} with $\tau^* = \max\{\tau \leq t-1 : w_\tau \in D_0\}$. But then $\mathcal{D} = (D_0, \{w_1\}, \ldots, \{w_{t-1}\}, S_j)$ is a feasible trace. Noticing that $v \notin A(\mathcal{D})$ by construction completes the proof. ■

A.3 Section 4.2 Proofs

Proof [Proof of Proposition 14] Let F be an arbitrary cumulative distribution function with density f. We need to show that

$$F(x) - F(y) = \int_{\mathbb{R}} \mathbf{1}(y < t \leq x) f(t) dt \tag{28}$$

is concave on $\{(x, y) : x > y\}$. Note that $g(x, y, u) := \mathbf{1}(y < t \leq x)$ is a log-concave function since for any $\lambda \in [0, 1]$ and points $A_1 = (x_1, y_1, u_1), A_2 = (x_2, y_2, u_2)$ with $x_1 > y_1, x_2 > y_2$, it holds:

$$\begin{aligned} g(\lambda A_1 + (1-\lambda)A_2) &= \mathbf{1}\left[\lambda x_1 + (1-\lambda)x_2 < \lambda u_1 + (1-\lambda)u_2 \leq \lambda y_1 + (1-\lambda)y_2\right] \\ &\geq \mathbf{1}\left[x_1 < u_1 \leq y_1\right] \mathbf{1}\left[x_2 < u_2 \leq y_2\right] \\ &= \mathbf{1}\left[x_1 < u_1 \leq y_1\right]^{\lambda} \mathbf{1}\left[x_2 < u_2 \leq y_2\right]^{1-\lambda} \\ &= g(A_1)^{\lambda} g(A_2)^{1-\lambda}. \end{aligned}$$

Thus, the expression under the integral in (28) is log-concave as a product of log-concave functions. Finally, by Theorem 6 in Prékopa (1973), the integral of a multivariate log-concave function w.r.t. any of its arguments is also log-concave. This completes the proof. ■

A.4 Section 4.3 Proofs

Along with the data matrix X_v, it will be convenient for this section to define its cumulative row-sum version. More formally, for $n \in \mathcal{I}_v$ and $t \in \mathcal{T}_v^{(n)}$, denote the indicator vector of all the parent nodes active at time t within trace n as $z_t^{(n)} = [\mathbf{1}(u_i \in A_t^{(n)})]_{i=1}^m$, its concatenation over $t \in \mathcal{T}_v^{(n)}$ as $Z_v^{(n)}$, and the further concatenation over $n \in \mathcal{I}_v$ as $Z_v \in \{0, 1\}^{N_v \times m}$.

Before we proceed to the proof of Proposition 16, we present and prove a technical lemma relating the minimum eigenvalues of $X_v^{(n)}$ and $Z_v^{(n)}$. Consider an arbitrary $X \in \mathbb{R}^{k \times m}$ and let $R \in \{0, 1\}^{k \times k}$ denote a lower triangular matrix with $R_{ij} = \mathbf{1}[j \leq i]$. Then the rows of $Z = RX$ are cumulative sums of the rows of X. Let $\lambda_{\min}(\cdot)$ denote the minimal eigenvalue of a symmetric matrix.

Lemma 27 *With X and Z defined as above,*

$$\lambda_{\min}(XX^\top) \le 4\lambda_{\min}(ZZ^\top).$$

Proof [Proof of Lemma 27] Since$[RR^\top]^{-1}$ is a tri-diagonal Toeplitz matrix with conveniently computed characteristic polynomial, one can show that $\lambda_{\min}(RR^\top) = 1/[4\cos^2(\pi/(4k+2))] \ge 1/4$. Therefore, we obtain

$$\lambda_{\min}(ZZ^\top) = \lambda_{\min}(RXXR^\top) \ge \lambda_{\min}(RR^\top)\lambda_{\min}(XX^\top) \ge \frac{1}{4}\lambda_{\min}(XX^\top).$$

■

Proof [Proof of Proposition 16] With $n \in \mathcal{I}_v$ and $t \in \mathcal{T}_v^{(n)}$, denote $d_{n,t+1}(\boldsymbol{\theta}_v) := F_v(\boldsymbol{\theta}_v^\top z_t^{(n)}) - F_v(\boldsymbol{\theta}_v^\top z_{t-1}^{(n)})$. Additionally, let $d_{n,T_n+1}(\boldsymbol{\theta}_v) := 1 - F_v(\boldsymbol{\theta}_v^\top z_{T_n}^{(n)}) = 1 - F_v(\boldsymbol{\theta}_v^\top z_{t(v,n)}^{(n)})$. Observe that for any $t \in \mathcal{T}_v^{(n)}$, we have

$$\begin{aligned}
&\mathbb{P}[y_{t+1}^{(n)} = 1 \mid X_v] = \mathbb{P}[B_v(A_{t-1}^{(n)}; \boldsymbol{\theta}_v^*) < U_v \le B_v(A_t^{(n)}; \boldsymbol{\theta}_v^*) \mid X_v] = d_{n,t+1}(\boldsymbol{\theta}_v^*) \quad \text{and} \\
&\mathbb{P}[y_{t+1}^{(n)} = 0 \text{ for all } t \in \mathcal{T}_v^{(n)} \mid X_v] = \mathbb{P}[B_v(A_{T_n}^{(n)}; \boldsymbol{\theta}_v^*) < U_v \mid X_v] = d_{n,T_n+1}(\boldsymbol{\theta}_v^*).
\end{aligned}$$

Note that by strict monotonicity and continuity of F_v, we can lower bound $d_{n,t+1}(\boldsymbol{\theta}_v), t \in \mathcal{T}_v^{(n)} \cup \{T_n\}$ uniformly for all $\boldsymbol{\theta}_v \in \tilde{\Theta}_v$ and for $\boldsymbol{\theta}_v^*$ in particular as

$$d_{\min} = \min_{\substack{z_1, z_2 \in [0,\gamma], \\ z_1 \ge z_2 + \varepsilon}} \big[F_v(z_1) - F_v(z_2)\big] > 0.$$

Moreover, with all considered threshold distributions having a log-concave density, it holds $\nabla^2 \log d_{n,t+1} \succeq 0$ for any $n \in \mathcal{I}_v, t \in \mathcal{T}_v^{(n)} \cup \{T_n\}$ per Proposition 14. So, we can lower-bound the expected Hessian as:

$$\begin{aligned}
\mathbb{E}[-\nabla^2 L_v(\boldsymbol{\theta}_v) \mid X_v] &= -\sum_{n \in \mathcal{I}_v} \sum_{t \in \mathcal{T}_v^{(n)} \cup \{T_n\}} d_{n,t+1}(\boldsymbol{\theta}_v^*) \nabla^2 \log d_{n,t+1}(\boldsymbol{\theta}_v) \\
&\succeq -d_{\min} \sum_{n \in \mathcal{I}_v} \sum_{t \in \mathcal{T}_v^{(n)}} \nabla^2 \log d_{n,t+1}(\boldsymbol{\theta}_v),
\end{aligned} \tag{29}$$

where each term in the final expression can be expressed using the chain rule in terms of $H_v(z_1, z_2) = \nabla^2_{z_1,z_2} \log(F_v(z_1) - F_v(z_2)) \in \mathbb{R}^{2\times 2}$ and $L_{n,t} = [z_t^{(n)}; z_{t-1}^{(n)}] \in \mathbb{R}^{m \times 2}$ as:

$$\nabla^2 \log d_{n,t+1}(\boldsymbol{\theta}_v) = L_{n,t} \cdot H_v(\boldsymbol{\theta}_v^\top z_t^{(n)}, \boldsymbol{\theta}_v^\top z_{t-1}^{(n)}) \cdot L_{n,t}^\top.$$

One can show by checking $\det(H_v) > 0$ and $H_{v,11} < 0$ that for $F_v \sim \text{Beta}(\alpha, \beta)$ with $\alpha \ge 1, \beta \ge 1$ except for the case $\alpha = \beta = 1$ (uniform distribution), the Hessian $H_v(z_1, z_2)$ is strictly negative definite for any $z_1 > z_2$. In this case, by continuity of the Hessian and the mapping $A \mapsto \lambda_{\min}(A)$ taking the minimum eigenvalue of a negative definite matrix, we can define

$$c_1 = \min_{\substack{z_1, z_2 \in [0,\gamma], \\ z_1 \ge z_2 + \varepsilon}} \lambda_{\min}\big(H_v(z_1, z_2)\big) > 0.$$

With that, the last line of (29) can be further lower bounded by

$$d_{\min} c_1 \sum_{n \in \mathcal{I}_v} \sum_{t \in \mathcal{T}_v^{(n)}} L_{n,t} L_{n,t}^\top \succeq d_{\min} c_1 Z_v^\top Z_v,$$

where by Lemma 27, the smallest eigenvalue of the RHS is lower bounded by $\frac{d_{\min} c_1}{4} \lambda_{\min}(X_v^\top X_v)$. The result follows by substituting $c_\lambda = d_{\min} c_1 / 4$.

For $F_v \sim \text{Unif}[0,1]$ and $F_v \sim \text{Exponential}(\mu)$, the Hessian can be computed with $g(z_1, z_2) = 1/(z_1 - z_2)^2$ and $g(z_1, z_2) = e^{-\mu(z_1+z_2)}/(e^{-\mu z_2} - e^{-\mu z_1})^2$, respectively, as

$$H_v(z_1, z_2) = g(z_1, z_2) \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Note that $H_v(z_1, z_2)$ is a negative semi-definite matrix of rank 1 with eigenvector $h = (1, -1)$ and corresponding eigenvalue $\lambda_h(z_1, z_2) < 0$, satisfying $H_v(z_1, z_2) = \lambda_h(z_1, z_2) h h^\top$. By continuity of H_v and the largest eigenvalue of a matrix, we can define

$$c_2 = \min_{\substack{z_1, z_2 \in [0, \gamma], \\ z_1 \geq z_2 + \varepsilon}} -\lambda_h(z_1, z_2) > 0$$

and further lower bound the last line of (29) as follows:

$$d_{\min} c_2 \sum_{n \in \mathcal{I}_v} \sum_{t \in \mathcal{T}_v^{(n)}} (L_{n,t} h)(L_{n,t} h)^\top = d_{\min} c_2 X_v X_v^\top.$$

Letting $c_\lambda = c_2 d_{\min}$ completes the proof. ∎

Before we proceed to the proof of Theorem 17, we present its main idea. The key step is to reformulate (16) as a binary classification problem, for which a finite sample result can be conveniently derived by treating the design X_v as fixed and the response as the only source of randomness. The response here is the activation indicator of node v, defined for each trace $n \in \mathcal{I}_v$ and $t \in \mathcal{T}_v^{(n)}$ as

$$y_{t+1}^{(n)} = \mathbf{1}(v \in D_{t+1}^{(n)}).$$

Then, the log-likelihood of node v in (14) can be rewritten as

$$L_v(\boldsymbol{\theta}_v) = \sum_{n \in \mathcal{I}_v} \sum_{t \in \mathcal{T}_v^{(n)}} \mathbf{1}[t \leq t(v,n)] \Big\{ y_{t+1}^{(n)} \log\big[p_{t+1}^{(n)}(\boldsymbol{\theta}_v)\big] + (1 - y_{t+1}^{(n)}) \log\big[1 - p_{t+1}^{(n)}(\boldsymbol{\theta}_v)\big] \Big\}, \quad (30)$$

where $p_{t+1}^{(n)}(\boldsymbol{\theta}_v)$ is the GLT transition probability defined in (5) and written with new notation as

$$p_{t+1}^{(n)}(\boldsymbol{\theta}_v) = \frac{F_v(\boldsymbol{\theta}_v^\top z_t^{(n)}) - F_v(\boldsymbol{\theta}_v^\top z_{t-1}^{(n)})}{1 - F_v(\boldsymbol{\theta}_v^\top z_{t-1}^{(n)})}. \quad (31)$$

Note that the activation time $t(v,n)$ in this framework is random, and $y_{t+1}^{(n)} = 0$ for any $t < t(v,n)$ and $y_{t+1}^{(n)} = 1$ implies $t = t(v,n)$. In particular, this means that the responses

$\mathbf{y}^{(n)} = \{y_{t+1}^{(n)} : t \in \mathcal{T}_v^{(n)}, t \le t(v,n)\}$ are dependent conditionally on $X_v^{(n)}$ and their number is random, since we do not receive more observations after v is activated. This creates additional technical complexity relative to a binary classification model. Fortunately, $\mathbf{y}^{(n)}$ are independent across $n \in \mathcal{I}_v$ due to the assumed trace independence.

Given the log-likelihood of the binary classification model in (30), our next goal is to establish a high probability (with respect to the randomness in $\{\mathbf{y}^{(n)}, n \in \mathcal{I}_v\}$) error bound for the model's MLE conditional on the design X_v.

Proof [Proof of Theorem 17] Fix a node $v \in V_c$ with $m = |P(v)|$ and $\delta \in (0,1)$. Denote the negative log-likelihood term corresponding to the trace $n \in \mathcal{I}_v$ and time point $t \in \mathcal{T}_v^{(n)}$ in (30)by

$$\ell_{n,t}(\boldsymbol{\theta}_v) = -\mathbf{1}[t \le t(v,n)]\Big\{y_{t+1}^{(n)} \log\big[p_{t+1}^{(n)}(\boldsymbol{\theta}_v)\big] + (1 - y_{t+1}^{(n)}) \log\big[1 - p_{t+1}^{(n)}(\boldsymbol{\theta}_v)\big]\Big\}.$$

Then, (16) essentially solves the following problem:

$$\hat{\boldsymbol{\theta}}_v = \arg\min_{\boldsymbol{\theta}_v \in \tilde{\Theta}_v} \ell_{N_v}(\boldsymbol{\theta}_v), \quad \text{where} \quad \ell_{N_v}(\boldsymbol{\theta}_v) = \frac{1}{N_v} \sum_{n \in \mathcal{I}_v} \sum_{t \in \mathcal{T}_v^{(n)}} \ell_{n,t}(\boldsymbol{\theta}_v).$$

Our first step is to establish a high probability bound for $\|\nabla \ell_{N_v}(\boldsymbol{\theta}_v^*)\|_2$. With $f_v = F_v'$ denoting the threshold density, we can express the log-likelihood gradient of each sample as

$$\nabla \ell_{n,t}(\boldsymbol{\theta}_v) = c_{n,t}(\boldsymbol{\theta}_v) \nabla p_{t+1}^{(n)}(\boldsymbol{\theta}_v) \mathbf{1}[t \le t(v,n)], \tag{32}$$

where $c_{n,t}(\boldsymbol{\theta}_v) = \frac{p_{t+1}^{(n)} - y_{t+1}^{(n)}}{p_{t+1}^{(n)}(1 - p_{t+1}^{(n)})}$ and

$$\nabla p_{t+1}^{(n)}(\boldsymbol{\theta}_v) = \frac{[1 - F_v(\boldsymbol{\theta}_v^\top z_{t-1}^{(n)})] f_v(\boldsymbol{\theta}_v^\top z_t^{(n)}) z_t^{(n)} - [1 - F_v(\boldsymbol{\theta}_v^\top z_t^{(n)})] f_v(\boldsymbol{\theta}_v^\top z_{t-1}^{(n)}) z_{t-1}^{(n)}}{[1 - F_v(\boldsymbol{\theta}_v^\top z_{t-1}^{(n)})]^2}.$$

Since f_v is continuous, it achieves its maximum $f_{\max}$ on the interval $[0, \gamma]$ and minimum $f_{\min}$ on $[\varepsilon, \gamma]$. Note that both quantities are positive as both intervals lie in the support of f_v. Thus, by definition of $\tilde{\Theta}$ and feasibility, we have $f_v(\boldsymbol{\theta}_v^\top z_{t-1}^{(n)}) \in [0, f_{\max}]$ and $f_v(\boldsymbol{\theta}_v^\top z_t^{(n)}) \in [f_{\min}, f_{\max}]$. Then, for each coordinate $j = 1, \ldots, m$ of $\nabla p_{t+1}^{(n)}(\boldsymbol{\theta}_v)$ the triangular inequality implies

$$|[\nabla p_{t+1}^{(n)}(\boldsymbol{\theta}_v)]_j| \ \le \frac{2 f_{\max}}{[1 - F_v(\gamma)]^2}. \tag{33}$$

To bound $c(\boldsymbol{\theta}_v)$, we first show that $p_{t+1}^{(n)}$ is bounded away from 0 and 1. Note that $p_{t+1}^{(n)}(\boldsymbol{\theta}_v) \le F_v(\gamma)$. For the lower bound, if $z_{t-1}^{(n)} = \mathbf{0}_m$, then $p_{t+1}^{(n)}(\boldsymbol{\theta}_v) \ge F_v(\varepsilon)$ and otherwise, by the mean-value theorem and feasibility there is $\xi \in [\varepsilon, \gamma]$ such that $p_{t+1}^{(n)}(\boldsymbol{\theta}_v) = f_v(\xi) \boldsymbol{\theta}_v^\top (z_t^{(n)} - z_{t-1}^{(n)}) \ge f_{\min}\varepsilon$. To summarize,

$$F_v(\gamma) \ge p_{t+1}^{(n)}(\boldsymbol{\theta}_v) \ge \min\{f_{\min}\varepsilon, F_v(\varepsilon)\},$$

which implies together with $|p_{t+1}^{(n)}(\boldsymbol{\theta}_v) - y_{t+1}^{(n)}| \le 1$ that

$$|c(\boldsymbol{\theta}_v)| = \Big|\frac{p_{t+1}^{(n)}(\boldsymbol{\theta}_v) - y_n}{p_{t+1}^{(n)}(\boldsymbol{\theta}_v)(1 - p_{t+1}^{(n)}(\boldsymbol{\theta}_v))}\Big| \le \frac{1}{\min\{f_{\min}\varepsilon, F_v(\varepsilon)\}(1 - F_v(\gamma))}. \tag{34}$$

Combining (33) and (34), we obtain

$$|[\nabla \ell_{n,t}(\boldsymbol{\theta}_v^*)]_j| \leq \frac{2f_{\max}}{\min\{f_{\min}\varepsilon, F_v(\varepsilon)\}(1-F_v(\gamma))^3} = C_0'(\varepsilon, \gamma, F_v). \tag{35}$$

To bound $|\sum_{n\in\mathcal{I}_v}\sum_{t\in\mathcal{T}_v^{(n)}}[\nabla \ell_{n,t}(\boldsymbol{\theta}_v)]_j|$, we will use Azuma-Hoeffding inequality exploiting the convenient conditional structure of the subsequent terms in the sum [1]. Denote $Y_{n,t} := [\nabla \ell_{n,t}(\boldsymbol{\theta}_v^*)]_j$ and note that it is identically zero if $[\nabla p_{t+1}^{(n)}(\boldsymbol{\theta}_v^*)]_j = 0$ and otherwise satisfies

$$\begin{aligned}\mathbb{E}[Y_{n,t} \mid \{Y_{n,\tau}\}_{\tau<t}, X_v^{(n)}] &= \mathbb{E}[Y_{n,t} \mid \{y_{\tau+1}^{(n)} = 0\}_{\tau<t}, X_v^{(n)}] \\ &\propto \mathbb{E}[y_{t+1}^{(n)} - p_{t+1}^{(n)}(\boldsymbol{\theta}_v^*) \mid \{y_{\tau+1}^{(n)} = 0\}_{\tau<t}, X_v^{(n)}] = 0.\end{aligned}$$

On the other hand, $Y_{n,t}$'s with different n are independent conditional on Z_v, so partial sums of $\sum_n \sum_t Y_{n,t}$ form a martingale. Since each term in the sum is almost surely bounded by (35), we have by Azuma-Hoeffding for any $z > 0$:

$$\mathbb{P}\big(\frac{1}{N_v}\big|\sum_{n\in\mathcal{I}_v}\sum_{t\in\mathcal{T}_v^{(n)}} Y_{n,t}\big| \geq z\big) \leq 2\exp(-z^2 N_v/2C_0').$$

Using the union bound over the coordinates $j = 1, \ldots, m$ and inequality relating the ℓ_2 and ℓ_∞ norms, we have

$$\mathbb{P}\big(\|\nabla \ell_{N_v}\|_2 \geq z\big) \leq \mathbb{P}\big(\|\nabla \ell_{N_v}\|_\infty \geq z/\sqrt{m}\big) \leq 2m\exp(-z^2 N_v/2mC_0')$$

Denoting the RHS by $\delta/2$ and solving for z, we obtain:

$$\mathbb{P}\big(\|\nabla \ell_{N_v}\|_2 \geq \sqrt{\frac{2C_0' m}{N_v}\log\frac{4m}{\delta}}\big) \leq \delta/2 \tag{36}$$

Our next step is to demonstrate that the Hessian of $\ell_{N_v}(\boldsymbol{\theta}_v)$ is positive definite on the whole $\tilde{\Theta}_v$ with high probability. Denote $H_n = \nabla^2[\sum_{t\in\mathcal{T}_v^{(n)}} \ell_{n,t}(\boldsymbol{\theta}_v)]$ so that by Assumption 3, it holds $\mathbb{E}[\frac{1}{N_v}\sum_{n\in\mathcal{I}_v} H_n] = \mathbb{E}[\nabla^2 \ell_{N_v}(\boldsymbol{\theta}_v)] \succeq \lambda_{\min} I$. By telescoping, each H_n can be conveniently rewritten with $t(v,n)$ denoted by t^* for brevity:

$$H_n = -y_{t^*+1}^{(n)}\nabla^2 \log[F_v(\boldsymbol{\theta}_v^\top z_{t^*}^{(n)}) - F_v(\boldsymbol{\theta}_v^\top z_{t^*-1}^{(n)})] - (1 - y_{t^*+1}^{(n)})\nabla^2 \log[1 - F_v(\boldsymbol{\theta}_v^\top z_{t^*}^{(n)})] \tag{37}$$

By Assumption 3, the density f_v is log-concave, implying, per Proposition 14, that H_n is positive semidefinite. Our goal now is to apply a Chernoff-type concentration on the sum of independent positive semidefinite matrices H_n to show that it is close to its positive definite expectation, and thus, with high probability, is positive definite itself. For that, we also need the largest eigenvalue of each H_n to be bounded. Denote the two additive terms in the panel above by $H_{n,1}$ and $H_{n,2}$, so that $H_n = H_{n,1} + H_{n,2}$. By the chain rule, the spectral norm of the second one is dominated by

$$\|H_{n,2}\|_2 \leq \|z_{t^*}^{(n)} z_{t^*}^{(n)\top}\|_2 \sup_{z\in[\varepsilon,\gamma]} |\frac{d}{dz^2}\log[1 - F_v(z)]| \leq C_2(\varepsilon, \gamma, F_v)m,$$

1. The following derivation is inspired by the proof of Lemma 3 in (Pouget-Abadie and Horel, 2015).

where the supremum is achieved since the function inside is continuous on the interval by Assumption 2. Again, by the chain rule and submultiplicativity of the norm, we also have

$$\|H_{n,1}\|_2 \leq \|[z_{t^*}^{(n)}; z_{t^*-1}^{(n)}]\|_2^2 \sup_{\substack{z_1,z_2\in[0,\gamma]\\ z_1\geq z_2+\varepsilon}} \|\nabla^2_{z_1,z_2} \log[F_v(z_1) - F_v(z_2)]\|_2 \leq C_1(\varepsilon,\gamma,F_v)m$$

where we again used the fact that a continuous function on a compact set achieves its maximum. Therefore, with $c_0' := C_1 + C_2$, we have $\|H_n\|_2 \leq c_0' m$. So, by Corollary 5.2 in (Tropp, 2012), we have the following concentration for any $z \in [0,1]$:

$$\mathbb{P}\big(\big\|\frac{1}{N_v}\sum_{n\in\mathcal{I}_v} H_n(\boldsymbol{\theta}_v)\big\|_2 \leq \lambda_{\min} z\big) \leq m\exp[-(1-z)^2\lambda_{\min}N_v/2c_0' m].$$

Set $z := 1/2$ and $c_0 = 8c_0'$. Then, the corresponding upper bound on the probability above is dominated by $\delta/2$ when the assumed condition $N_v \geq \frac{c_0 m}{\lambda_{\min}} \log\frac{2m}{\delta}$ is satisfied:

$$\mathbb{P}\big(\big\|\nabla^2 \ell_{N_v}(\boldsymbol{\theta}_v)\big\|_2 \leq \frac{\lambda_{\min}}{2}\big) \leq m\exp[-\lambda_{\min}N_v/c_0 m] \leq \delta/2. \tag{38}$$

With that, the union bound implies that the gradient concentration in (36) and the Hessian concentration in (38) hold together with probability at least $1-\delta$. Assume that both of these events hold. Expanding ℓ_{N_v} at $\boldsymbol{\theta}_v^*$ gives for some $\tilde{\boldsymbol{\theta}}_v = \boldsymbol{\theta}_v^* + z(\hat{\boldsymbol{\theta}}_v - \boldsymbol{\theta}_v^*)$ with $z \in [0,1]$:

$$\ell_{N_v}(\hat{\boldsymbol{\theta}}_v) = \ell_{N_v}(\boldsymbol{\theta}_v^*) + \nabla\ell_{N_v}(\boldsymbol{\theta}_v^*)^\top(\hat{\boldsymbol{\theta}}_v - \boldsymbol{\theta}_v^*) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_v - \boldsymbol{\theta}_v^*)^\top\nabla^2\ell_{N_v}(\tilde{\boldsymbol{\theta}}_v)(\hat{\boldsymbol{\theta}}_v - \boldsymbol{\theta}_v^*).$$

Since $\hat{\boldsymbol{\theta}}_v, \boldsymbol{\theta}_v^* \in \tilde{\Theta}_v$ by Assumption 2 and $\tilde{\Theta}_v$ is a convex set by definition, we have $\tilde{\boldsymbol{\theta}}_v \in \tilde{\Theta}_v$, so the quadratic term above is lower bounded by $\frac{\lambda_{\min}}{4}\|\hat{\boldsymbol{\theta}}_v - \boldsymbol{\theta}_v^*\|_2$. On the other hand, $\ell_{N_v}(\boldsymbol{\theta}_v^*) \geq \ell_{N_v}(\hat{\boldsymbol{\theta}}_v)$ by optimality, which implies together with the Cauchy-Schwarz inequality

$$\|\nabla\ell_{N_v}(\boldsymbol{\theta}_v^*)\|_2\|\hat{\boldsymbol{\theta}}_v - \boldsymbol{\theta}_v^*\|_2 \geq -\nabla\ell_{N_v}(\boldsymbol{\theta}_v^*)^\top(\hat{\boldsymbol{\theta}}_v - \boldsymbol{\theta}_v^*) \geq \frac{\lambda_{\min}}{4}\|\hat{\boldsymbol{\theta}}_v - \boldsymbol{\theta}_v^*\|_2^2.$$

The needed bound is obtained by dividing through $\|\hat{\boldsymbol{\theta}}_v - \boldsymbol{\theta}_v^*\|_2$ and plugging in the gradient concentration from (36) with $C_0 = 4\sqrt{2C_0'}$. ■

Proof [Proof of Proposition 19] By Theorem 5.1 in (Lehmann and Casella, 1998), the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^*$ is consistent and satisfies

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}^*}} \mathcal{N}(0, \Sigma(\boldsymbol{\theta}^*)) \quad \text{as} \quad N \to \infty, \tag{39}$$

where $\Sigma^{-1}(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}\sim\mathbb{P}_{\boldsymbol{\theta}^*}}[-\nabla^2 \log \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{D})]$, if (a) the distribution family $\{\mathbb{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \tilde{\Theta}\}$ is identifiable, (b) the traces are independent and have common support, (c) $\boldsymbol{\theta}^*$ lies in the interior of $\tilde{\Theta}$, (d) $\mathbb{P}_{\boldsymbol{\theta}}$ is three times differentiable wrt $\boldsymbol{\theta}$ in an open neighborhood ω around $\boldsymbol{\theta}^*$, (e) the first and second order partial derivatives of $\log \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{D})$ wrt $\boldsymbol{\theta}$ are dominated by an integrable $g(\mathcal{D})$, (f) third-order partial derivatives of $\log \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{D})$ wrt $\boldsymbol{\theta}$ are bounded on ω, and (g) Σ is positive definite on ω. We can verify all of these conditions: (a) holds due to the assumed identifiability condition of Theorem 11; (b) holds because traces are assumed

independent and they have a common support of all feasible traces $\mathcal{F}(G)$ with $\mathbb{P}^0(D_0) > 0$ due to Lemma 9; (c) holds by Assumption 2; (d) holds for any $\boldsymbol{\theta} \in \tilde{\Theta}$ since, according to (7), the trace likelihood is a composition of a linear transformation of $\boldsymbol{\theta}$ and $\{F_v\}_{v \in V_c}$, which are three times differentiable by Assumption 2; (e) holds because first- and second-order partial derivatives of $\log \mathbb{P}_{\boldsymbol{\theta}}$ are continuous on the compact $\tilde{\Theta}$ by Assumption 2 and thus achieve their maximum and minimum values; (f) holds because third-order partial derivatives are also continuous by Assumption 2 and thus bounded on the compact $\tilde{\Theta}$; and finally (g) holds by Assumption 4 and continuity of the Hessian for the ω chosen as a ball of a sufficiently small radius.

Thus $\hat{\boldsymbol{\theta}}$ is consistent and by continuity of $\Sigma^{-1}(\boldsymbol{\theta})$ and the continuous mapping theorem we have that

$$\Sigma^{-1}(\hat{\boldsymbol{\theta}}) \to \Sigma^{-1}(\boldsymbol{\theta}^*) \quad \text{in probability as } N \to \infty.$$

On the other hand, by the law of large numbers, we have for any $\boldsymbol{\theta} \in \tilde{\Theta}$ that

$$\hat{\Sigma}^{-1}(\boldsymbol{\theta}) := -\frac{1}{N} \sum_{n=1}^{N} \nabla^2 \log \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{D}_n) \to \Sigma^{-1}(\boldsymbol{\theta}) \quad \text{in probability as } N \to \infty.$$

Combining these two observations, we have $\hat{\Sigma}^{-1}(\hat{\boldsymbol{\theta}}) \to \Sigma^{-1}(\boldsymbol{\theta}^*)$ in probability. So, by Assumption 4, $\hat{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})$ is positive definite with probability tending to one. Slutsky's theorem and (39) implies

$$\sqrt{N}\hat{\Sigma}^{-1/2}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}^*}} \mathcal{N}(0, I_{|E|}) \quad \text{as} \quad N \to \infty.$$

Finally, the node-wise asymptotic result follows from the block-diagonal structure of $\hat{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})$, which consists of the blocks $-\frac{1}{N}\nabla^2 L_v(\hat{\boldsymbol{\theta}}_v), v \in V_c$. ■

A.5 Section 5 Proofs

We start with the following technical lemma formulating a convenient equivalent definition of concavity:

Lemma 28 *Consider the cdf F of the distribution supported on $[0, h]$. Then the condition*

$$F(x + b) - F(x) \geq F(y + b) - F(y) \tag{40}$$

holds for all triples (x, y, b) with $0 \leq x \leq y \leq y + b \leq h$ if and only if F is concave on $[0, h]$.

Proof [Proof of Lemma 28] (Necessity) It is enough to verify the "midpoint" concavity condition

$$F\left(\frac{x' + y'}{2}\right) \geq \frac{F(x') + F(y')}{2}, \quad 0 \leq x' \leq y' \leq C$$

since for bounded functions (cdf is bounded between 0 and 1), it is known to be equivalent to concavity. Plugging $x = x', b = (y' - x')/2, y = (y' + x')/2$ into (40) and rearranging the terms implies the needed inequality.

(Sufficiency) Without loss of generality, assume $x + b \leq y$ (otherwise, repeat the proof for $x' = x, y' = x + b, b' = y - x$) . Consider the equivalent definition of a concave function (Lemma 2.1 in Gkioulekas (2013)), which states that for any $x_1 < x_2 < x_3$ it should hold

$$\frac{F(x_2) - F(x_1)}{x_2 - x_1} \geq \frac{F(x_3) - F(x_2)}{x_3 - x_2},$$

and use this inequality with $(x, x + b, y)$ and $(x + b, y, y + b)$ to obtain what was needed:

$$\frac{F(x+b) - F(x)}{b} \geq \frac{F(y) - F(x+b)}{y - x - b} \geq \frac{F(y+b) - F(y)}{b}.$$

■

Proof [Proof of Theorem 23] (Sufficiency) Consider a graph $G = (V, E)$ and an arbitrary GLT model on it. By Theorem 1 in Mossel and Roch (2010), it is enough to show that all threshold functions $f_v(S) = F_v\left(\sum_{u \in S} b_{u,v}\right)$ are monotone and submodular. Monotonicity holds trivially since all edge weights are nonnegative and F_v is non-decreasing. To establish submodularity, we need to check for $S' \subset S \subset P(v)$ and $w \notin S$ that

$$F_v\Big(\sum_{u \in S \cup \{w\}} b_{u,v}\Big) - F_v\Big(\sum_{u \in S} b_{u,v}\Big) \leq F_v\Big(\sum_{u \in S' \cup \{w\}} b_{u,v}\Big) - F_v\Big(\sum_{u \in S'} b_{u,v}\Big).$$

This follows by applying Lemma 28 to $b := b_{w,v}, x := \sum_{u \in S'} b_{u,v}, y := \sum_{u \in S} b_{u,v}$. The condition $0 \leq x \leq y \leq y + b \leq h_v$ follows from weights' non-negativity:

$$0 \leq \sum_{u \in S'} b_{u,v} \leq \sum_{u \in S} b_{u,v} \leq \sum_{u \in S \cup \{w\}} b_{u,v} \leq \sum_{u \in P(v)} b_{u,v} \leq h_v.$$

(Necessity) Let F be the cdf of an arbitrary distribution supported on $[0, h]$. By Lemma 28, there exist (x, y, b) with $0 \leq x \leq y \leq y + b \leq h$, such that $F(x+b) - F(x) < F(y+b) - F(y)$. Let G be a star graph of in-degree 3 as defined in Figure 3. Consider an instance of the GLT model on G with weights $\boldsymbol{\theta} = (x, y - x, b)$ and F as the cdf of node 4. Then with notations of Definition 21, submodularity is violated for $S = \{1, 2\}$, $S' = \{1\}$, and $v = 3$:

$$F(y + b) - F(y) = \sigma(S \cup \{v\}) - \sigma(S) > \sigma(S' \cup \{v\}) - \sigma(S') = F(x + b) - F(x).$$

■

Before proceeding to prove Proposition 25, we introduce a preliminary lemma that establishes a general bound on the discrepancy between the spreads from the IM solutions obtained under the ground truth and estimated models, which holds for an arbitrary graph.

Lemma 29 *For an arbitrary graph $G = (V, E)$, it holds with the notations of Proposition 25:*

$$|\sigma_{\boldsymbol{\theta}}(S^*(\boldsymbol{\theta})) - \sigma_{\boldsymbol{\theta}}(S^*(\hat{\boldsymbol{\theta}}))| \leq 2 \max_{|S| \leq k} |\sigma_{\boldsymbol{\theta}}(S) - \sigma_{\hat{\boldsymbol{\theta}}}(S)|.$$

Proof [Proof of Lemma 29] Denote for brevity $S^* = S^*(\boldsymbol{\theta})$ and $\hat{S} = S^*(\hat{\boldsymbol{\theta}})$. Then, by the triangle inequality, we have

$$|\sigma_{\boldsymbol{\theta}}(S^*) - \sigma_{\boldsymbol{\theta}}(\hat{S})| \leq |\sigma_{\boldsymbol{\theta}}(S^*) - \sigma_{\hat{\boldsymbol{\theta}}}(\hat{S})| + |\sigma_{\hat{\boldsymbol{\theta}}}(\hat{S}) - \sigma_{\boldsymbol{\theta}}(\hat{S})|. \tag{41}$$

To bound the first term, we use the definition of S^* and $\hat{S}$, followed by a standard maximum inequality

$$|\sigma_{\boldsymbol{\theta}}(S^*) - \sigma_{\hat{\boldsymbol{\theta}}}(\hat{S})| \leq |\max_{|S|\leq k} \sigma_{\boldsymbol{\theta}}(S) - \max_{|S|\leq k} \sigma_{\hat{\boldsymbol{\theta}}}(S)| \leq \max_{|S|\leq k} |\sigma_{\boldsymbol{\theta}}(S) - \sigma_{\hat{\boldsymbol{\theta}}}(S)|.$$

The second one can be bounded by observing that the difference evaluated at a given $\hat{S}$ with $|\hat{S}| \leq k$ is dominated by its maximum across all possible sets S of size not exceeding k:

$$|\sigma_{\hat{\boldsymbol{\theta}}}(\hat{S}) - \sigma_{\boldsymbol{\theta}}(\hat{S})| \leq \max_{|S|\leq k} |\sigma_{\hat{\boldsymbol{\theta}}}(S) - \sigma_{\boldsymbol{\theta}}(S)|.$$

Combining the two bounds completes the proof. ■

Proof [Proof of Proposition 25] Consider a directed bipartite graph $G = (V, E)$ with node set V consisting of child and parent node subsets, denoted respectively as V_c and V_p, so that $V = V_c \sqcup V_p$ and $E \subset \{(u, v) : u \in V_p,\ v \in V_c\}$. By Lemma 29, we only need to show that for any node subset $S \subset V$ with $|S| \leq k$, it holds

$$|\sigma_{\hat{\boldsymbol{\theta}}}(S) - \sigma_{\boldsymbol{\theta}}(S)| \leq L\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1.$$

Conveniently, we can explicitly compute the influence function of an arbitrary parameter $\boldsymbol{\theta}$ as

$$\sigma_{\boldsymbol{\theta}}(S) = |S| + \sum_{v \in V_c \setminus S} F_v(B_v(S; \boldsymbol{\theta}_v)).$$

Indeed, nodes in $V_c \cap S$ cannot activate anyone else, and nodes in $V_p \cap S$ can also propagate the influence to their children. From that, the needed bound follows by sequentially applying the triangular inequality together with the Lipschitz property of F_v:

$$\begin{aligned}
|\sigma_{\hat{\boldsymbol{\theta}}}(S) - \sigma_{\boldsymbol{\theta}}(S)| &\leq \sum_{v \in V_c \setminus S} |F_v(B_v(S; \boldsymbol{\theta}_v)) - F_v(B_v(S; \hat{\boldsymbol{\theta}}_v))| \\
&\leq L \sum_{v \in V_c \setminus S} |B_v(S; \boldsymbol{\theta}_v) - B_v(S; \hat{\boldsymbol{\theta}}_v)| \\
&\leq L \sum_{v \in V_c \setminus S} \sum_{u \in S \cap P(v)} |b_{u,v} - \hat{b}_{u,v}| \\
&\leq L\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1.
\end{aligned}$$

■

Appendix B. Additional Experiments

In this section, we present additional experiments that are not included in the main body of the manuscript.

B.1 Stability of joint estimation of weights and threshold parameters

Let G be a graph generated from the CWS model with $n = 100, k = 10$, and $p = 0.2$ (see Section 6.1 for detail). For this graph, consider an instance of the GLT model with $F_v \sim \text{Beta}(\alpha_v, \beta_v)$, where $\alpha_v, \beta_v \sim \text{Unif}\{1, \dots, 5\}$ independently for all nodes $v \in V_c$, and weights generated as in (24). From this GLT model, we generate 2000 traces with seed sets generated as in (25) with $s_{\max} = 10$. For each $v \in V_c$, we then estimate $\boldsymbol{\theta}$ and (α, β_v) using a 50-iteration alternating optimization with a constraint set $\Phi_v = [1, 5]$ for threshold parameters (see Section 4.5 for details), initialized at 10 different random weight vectors in (24). For each run, we note $(\hat{\boldsymbol{\theta}}_v^{(s)}, \hat{\alpha}_v^{(s)}, \hat{\beta}_v^{(s)})$ after each iteration $s = 1, \dots, 50$. We also run the procedure with weights initialized as the ground-truth to obtain the "gold standard" estimates $(\hat{\boldsymbol{\theta}}_v^*, \hat{\alpha}_v^*, \hat{\beta}_v^*)$ at convergence. For each initialization and each iteration of the alternating optimization, we note the Relative MAE between the gold standard and the randomly initialized estimate of $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and $\boldsymbol{\theta}$, that is, concatenations of the corresponding node-specific parameters across $v \in V_c$. Figure 9 shows the convergence plots with error bounds corresponding to 2 standard deviations computed based on the 10 initialization replicates. We can see that the randomly initialized estimates closely match the gold standard after a sufficient number of iterations, with $\boldsymbol{\beta}$ having an inferior but still relatively low error compared to $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$. We leave a careful investigation of this behavior for future work.

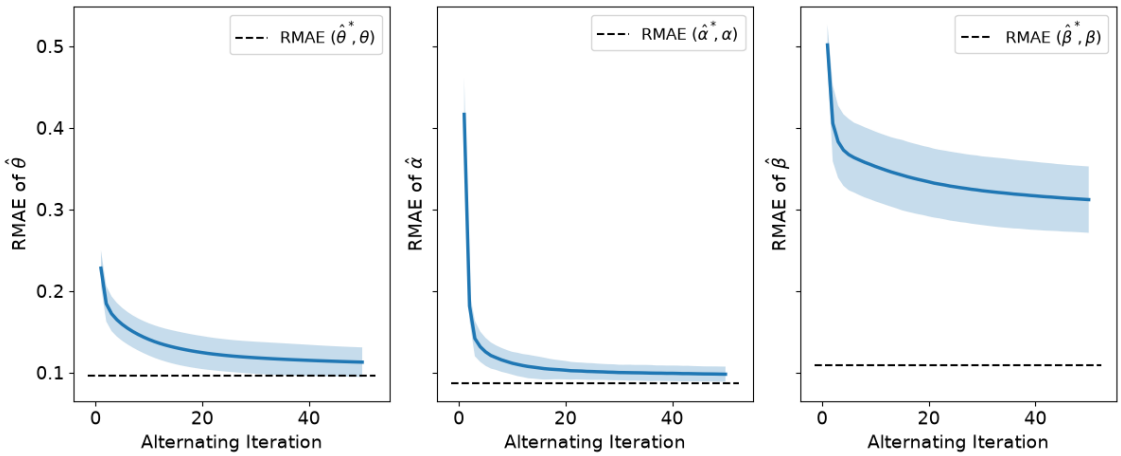


Figure 9: Error between the estimators initialized at the true and randomly sampled weights. Confidence bounds correspond to 2 standard deviations computed from 10 different weight initializations.

B.2 Dependency of the node sample size on the user's conservativeness.

Let G be a graph generated from the CWS model with $n = 100, k = 10$, and $p = 0.2$ (see Section 6.1 for details). For this graph, consider five instances of the GLT model, all with fixed weights sampled as in (24) but instance-dependent threshold distributions Beta(α, β), $(\alpha, \beta) \in \{(2, 1), (1, 1), (1, 2), (1, 3), (1, 5)\}$ (sorted from more to less conservative), assumed the same across the nodes within each instance. For each instance, we generate $N = 2000$ traces with seed sets generated as in (25) with $s_{\max} = 10$ and note the sample size N_v and the number of informative traces $|\mathcal{I}_v|$ for each node $v \in V_c$. In Figure 10, we plot the proportion of informative traces $|\mathcal{I}_v|/N$ and the average number of informative time-steps $N_v/|\mathcal{I}_v|$ per trace across the nodes, connecting the points in the box-plots corresponding to the same node. From the first plot, we can see that the proportion of traces where a node is informative reaches a plateau of roughly 80% after Beta$(1, 2)$. This shows that after some level of anti-conservativeness, a node activates in all traces where it is reached, and the proportion of these traces is governed purely by the seed set choice and the network topology. At the same time, the second plot demonstrates that after such a level is achieved, a node needs fewer active parents and thus fewer time steps to activate: right at Beta$(1, 2)$, the average number of informative time steps per trace reaches its peak of roughly 1.8 and starts gradually decreasing after that. On the other hand, before this critical level of anti-conservatism is achieved, the traces gradually increase in length, thus increasing the proportion of the reached nodes and the number of informative time steps per trace.

Specifically for the Beta$(1, 2)$ distribution, we further perform two more simulations to evaluate how the average number of informative observations per trace depends on other parameters. With all other parameters fixed, we vary the maximum GLT weight in-degree $d_{\max} \in \{0.2, 0.4, 0.6, 0.8, 1\}$ used in (24) and the average unweighted node in-degree $(k/2) \in \{2, 5, 10, 20, 40\}$, which is conveniently equal to $k/2$ by construction of the CWS model. The results in Figure 11 show that the average per-trace sample size fluctuates between 1 and 3 in all simulations, with larger values corresponding to higher weighted and unweighted node in-degrees.

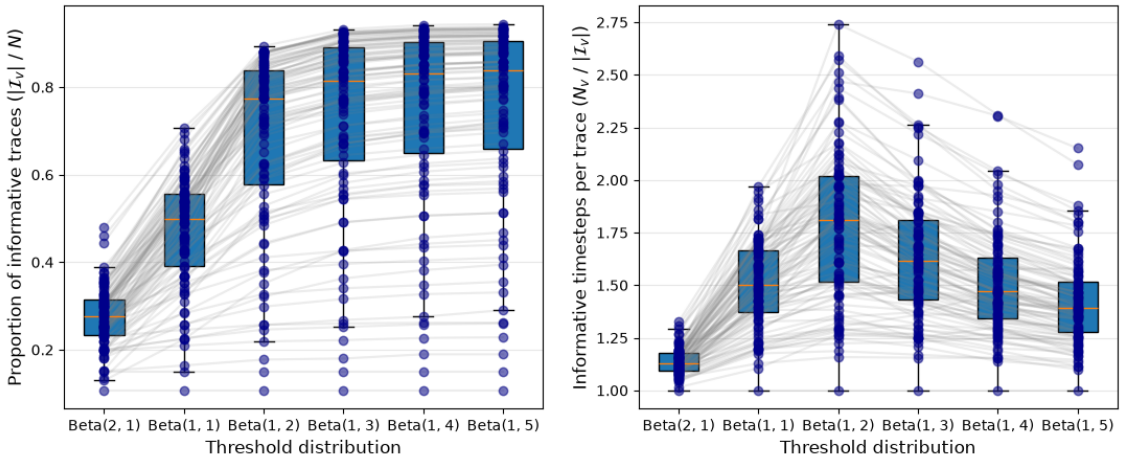


Figure 10: Distribution of the proportions of informative traces and average per-trace sample sizes across the nodes for different threshold distributions (right means less conservative).

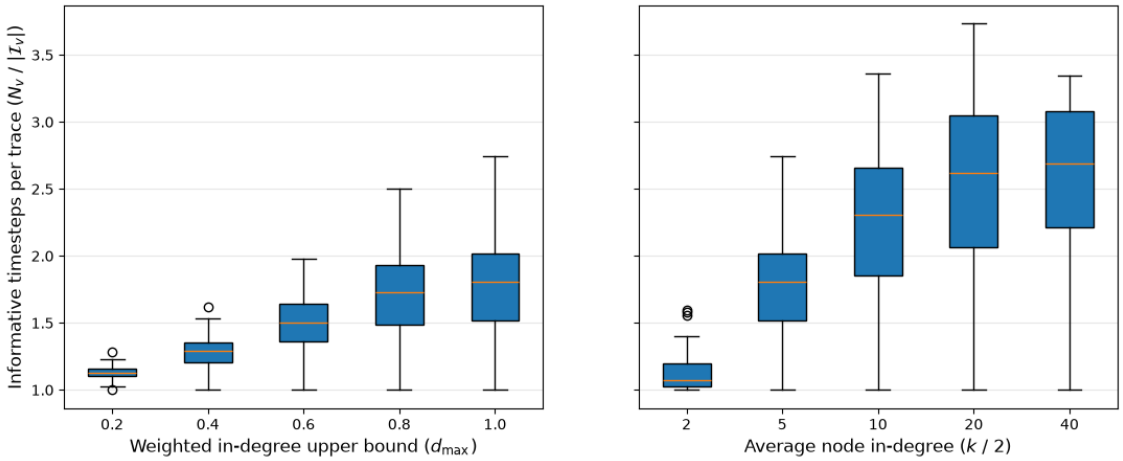


Figure 11: Distribution of average per-trace sample sizes across the nodes for the Beta$(1, 2)$-GLT model with varied maximum weight in-degrees (left) and average unweighted in-degrees (right).