Proceedings of Machine Learning Research vol 132:1–15, 2022

Local Central Limit Theorems for Approximate Maximum Likelihood Estimation of Network Information Spreading Models

Abram Magner AMAGNER@ALBANY.EDU
University at Albany, SUNY

Abstract

We consider infection/spreading process models on a graph in which infected nodes are associated with real-valued messages that evolve as they spread, according to a feature-based parametric probabilistic message spreading model. Estimation of the parameters of such models from an observed sample infection trajectory and associated messages presents computational and statistical challenges as a result of the fact that the set of neighbors that infected a given node is unobserved. This leads to a log likelihood function with exponentially many terms as a function of the number of infected neighbors of each node. We show that the log likelihood can be approximated using a local central limit theorem with provably accuracy and computational efficiency. We then show that under a well-posedness condition on the model, the maximum of the approximating function is close to the maximum of the true log likelihood, so that likelihood maximization can be approximately performed by maximizing the Gaussian approximation of the log likelihood.

Keywords: maximum likelihood,spreading processes,independent cascade,central limit theorem,estimation

1. Introduction

Spreading processes on networks are ubiquitous in a host of application domains. In a typical spreading process, nodes have an associated state at any given time: they may be *uninfected*, *infected*, or *active*. Active nodes are also infected, and they may infect uninfected neighbors, which themselves may become active during certain timesteps.

Frequently, additional information, such as a copy of a viral genome, an opinion, or a sentiment, is carried along with the infection. This information, which we will call a *message*, can mutate according to some statistical model as it is passed to newly infected nodes. In this paper, we are interested in developing computationally efficient statistical tools with provable accuracy and efficiency guarantees for estimation of parameters of probabilistic message spread models, given cascade data.

Estimation of message spread model parameters The general setup is as follows: fix a graph G on the vertices $[n] = \{1, ..., n\}$, a cascade model $\mathcal{C}$, and a message model $M_{\theta_{message}}$. A cascade $C \sim \mathcal{C}$ on G is generated. Here C takes the form of an *infection sequence* $S = (S_0, S_1, ..., S_N)$, where each S_j is the set of vertices infected in timestep j. When a vertex v is infected in timestep j by a set of infecting vertices $\mathcal{I}(v) \subseteq S_{j-1}$, its message M_v is sampled from $M_{\theta_{message}}(\mathcal{I}(v), M_{\mathcal{I}(v)})$. Our task is to estimate $\theta_{message}$, given observations consisting of S and the messages M_v for $v \in [n]$. Note, in particular, that in general we do *not* know the set $\mathcal{I}(v)$ for any v.

Knowledge of parameters of the message spread model is of interest in application domains where prediction or control of large-scale behavior of the messages is desired. For instance, if the message represents a sentiment, then one message spread model parameter could be a bias factor, representing a population's tendency toward some default sentiment. Estimation of this bias factor could be useful in, for example, measuring the success of some subsequent advertising campaign in modifying this bias.

Challenges inherent in the problem A few things make this problem challenging:

- The number of sample cascades need not be very large, while the *size* of each cascade may be. Thus, we cannot rely on classical guarantees for estimators such as maximum likelihood, since we only have a few independent and identically distributed (iid) sample cascades consisting of a large number of *non*-iid random variables.

- The set of infecting vertices for each vertex v is not known to us. This makes direct calculation of the log likelihood function intractable, at least by a naïve approach. In particular, conditioning on the possible values of these latent variables leads to an expression with a number of terms associated with each node that is exponential in its degree. Since this degree can be quite large, evaluating this expression is computationally cumbersome.

We will surmount these difficulties for a message model $M_{\theta_{message}}$ with a certain structure (we define it rigorously in Section 2.1): namely, we assume that each vertex is endowed with a feature vector $\boldsymbol{f}^{(v)} \in \mathbb{R}^d$ and that we are able to observe these features. When a vertex v is infected by a set of infectors $\mathcal{I}(v)$ chosen independently from the set of the active neighbors $\mathcal{NA}(v)$, its message M_v is computed according to a weighted average of the messages of its infectors. The weight of each infector w is proportional to the inner product of $\boldsymbol{f}^{(w)}$ with $\boldsymbol{f}^{(v)}$. The message then depends on a convex combination of this weighted sum with the a *bias* parameter b, where the influence of the neighbors' messages is captured by a *social influence* parameter a. The parameters to be estimated are a and b.

We will give sufficient conditions on the message spread model under which a *local central limit theorem* (LCLT) holds for the weighted average governing the message of each node. A local central limit theorem is an approximation of the (centered and normalized) probability density or mass function of the nth element of a sequence of random variables by the probability density function of the standard normal distribution. This is more precise than a standard central limit theorem, which is a convergence result for cumulative distribution functions (CDFs).

This will allow us to efficiently and accurately approximate the log likelihood function. Our estimator, then, maximizes this approximation. In the prior work section below, we discuss alternative approaches to this problem and their limitations.

1.1. Prior work/State of the art

There is a substantial literature on models involving a piece of information that is spread over a network. For instance, *opinion dynamics* models are either deterministic or stochastic dynamical systems that govern the evolution of a real-valued *opinion* for each vertex in a graph Das et al. (2014); Baccelli et al. (2015); Papachristou and Fotakis (2021); De et al. (2016). In contrast to the type of model that we consider, in almost all works cited here (e.g., De et al. (2016)), opinions are broadcast to *all* neighbors of each node. This may not be the case in many scenarios: e.g., in the case of a biological pathogen, an exposed node is only infected with some probability. In the case of information spread on a social network, an agent may only pay attention to a message with a certain probability. Furthermore, the focus of much of the literature is on the convergence properties of the models as functions of the parameters, rather than on the statistical problem of estimating parameters from observations (though this is not universally the case).

Alternative methods from the literature We are not aware of any previous works on spreading processes that use the local central limit theorem approach that we propose here. However, we review several alternative approaches and their disadvantages here.

The *expectation-maximization* algorithm and its variants provide an estimation method for parameters of a model when there are hidden variables Balakrishnan et al. (2014); Papachristou and Fotakis (2021). In our context, the relevant hidden variables would be the set of neighbors of each node v that are responsible for infecting v. The standard formulation of the EM algorithm requires an integration over all possible values of the hidden variable, which in our case is discrete, with a prohibitively large number of possibilities. Furthermore, a priori, there are no theoretical guarantees bounding the estimation error.

One could imagine apply a sampling-based approach to approximate the term of the log likelihood function corresponding to each vertex v. Specifically, supposing that the random variable corresponding to the message associated with vertex v is denoted by M_v and that we have observed the message m_v, we would need to approximate the density $p_{M_v}(m_v \mid \sigma(v))$, where $\sigma(v)$ is the set of active neighbors of v, along with their messages. We could repeatedly sample an infecting set of neighbors from the set of active neighbors, and sample a message M_v from the conditional distribution of M_v given this information. Based on repeated samples, we could then form an estimate of $p_{M_v}(m_v \mid \sigma(v))$. However, each sample would require iterating over each active neighbor of v, and the number of samples required to achieve an appropriate level of accuracy would be on the order of $1/p_{M_v}(m_v \mid \sigma(v))$. In contrast, our proposed method is more computationally efficient, with a running time that is constant with respect to the density $p_{M_v}(m_v \mid \sigma(v))$.

Algorithmic and graph-theoretic applications of local CLTs We now discuss LCLTs in other algorithmic and graph-theoretic/combinatorial domains.

An algorithmic application to approximate counting and sampling of matchings of a given size in a graph was given in Jain et al. (2021). LCLTs are also studied for parameters of random graphs Berkowitz (2017).

General theorems for LCLTs of parameters of random discrete structures based on asymptotic expansions of their probability generating functions on the unit circle are given in Flajolet and Sedgewick (2009). However, as these are generally most easily applied for integer-valued random variables (which is not the case for us), we resort to other techniques. A large number of other works provide results with similar restrictions.

Our results will depend on a general local limit theorem for sums of independent, but not necessarily identically distributed, non-lattice random variables coming from Mineka and Silverman (1970).

1.2. Our contributions

We provide a provably efficient and accurate method of approximating the log likelihood of a feature-based message spreading model in which messages of vertices are informed only by those neighbors that infected them. We do this by proving a local central limit theorem. We provide a natural sufficient well-posedness condition for the global maximum of this approximate log likelihood to be close to the true maximum likelihood. Thus, our method constitutes a computationally efficient approximation of the maximum likelihood estimator. In relation to the alternative methods that we reviewed in the prior work section, our method has the advantage of having provable guarantees in both accuracy and running time. Furthermore, we speculate that local limit theorems

can be more generally applied to similar estimation problems to produce computationally efficient estimators.

1.3. Organization of the paper

In Section 2, we formally state the problem and give the main results. We give full proofs in Section 5. We conclude in Section 3.

2. Main results

2.1. Problem statement and notation

We begin by giving notation and stating the main problem.

We fix an undirected, simple graph G on the vertices $[n] = \{1, ..., n\}$. We denote by $E(G)$ the edge set of G. We write $\mathcal{N}(v)$ for the set of neighbors of vertex v.

We further fix a *cascade model* $\mathcal{C}$. We define a generic cascade model as follows.

Definition 1 (Cascade model, infection sequence) *A cascade model $\mathcal{C}$ on a graph G on $[n]$ is a probability distribution on* infection sequences. *An* infection sequence *is a sequence $(S_0, ..., S_N)$, where each $S_j \subseteq [n]$, and $S_i \cap S_j = \emptyset$ for each $i \neq j$. We write $N(S) = \sum_{j=0}^{N} |S_j|$. We also write $|S| = N$.*

Our results will be for the *independent cascade model* Kempe et al. (2003), which we define as follows:

Definition 2 (Independent cascade model) *The independent cascade (IC) model has three parameters: an initial infection set $S_0 \subseteq [n]$, a transmission probability $p_{net}(e)$ for each directed edge e, and a probability p_{ext} of infection from an external source. Once any node is infected, it remains infected forever. At timestep 0, the nodes of S_0 are infected. At timestep j, nodes are infected as follows: for each vertex w in the* active set *S_{j-1}, for each uninfected neighbor v of w, v becomes infected by w with probability $p_{net}(w, v)$. Then, for each vertex u that is not yet infected, u becomes infected with probability p_{ext}.*

We denote by $\mathcal{TI}(v)$ the timestep at which v was infected: $\mathcal{TI}(v) = j \ : \ v \in S_j$. We call the set of active neighbors of v at the time that v was infected $\mathcal{NA}(v)$. Note that $\mathcal{NA}(v) = S_{\mathcal{TI}(v)-1}$, provided that $v \notin S_0$. We call the set of neighbors of v that chose to infect it the set of infectors *of v and denote it by $\mathcal{I}(v)$.*

We define our message spreading model as follows.

Definition 3 (Feature-based message spreading model with bias) *Fix a feature dimension $d > 0$ and associate with each node of G a fixed* feature vector *$\boldsymbol{f}^{(v)} \in \mathbb{R}^d$. Fix also two model parameters $a, b \in [0, 1]$, where a is the* social influence parameter, *and b is the* bias *parameter. Collectively, we refer to (a, b) as $\theta_{message}$.*

Finally, fix a continuous distribution $\mathcal{D}(\mu)$ supported on all of $[0, 1]$, parameterized by a mean value μ. Let the density of $\mathcal{D}(\mu)$ be denoted by $p_{\mathcal{D}}(\cdot \mid \mu)$.

Let $\sigma(v) = (\mathcal{NA}(v), M_{\mathcal{NA}(v)})$.

Conditioned on the set $\mathcal{I}(v)$ of neighbors that chose to infect v and their messages $M_{\mathcal{I}(v)}$, provided that $\mathcal{I}(v) \neq \emptyset$, the message M_v of v is

$$M_v \sim \mathcal{D}\left(ab + (1-a) \cdot \frac{1}{\sum_{w \in \mathcal{I}(v)} |\boldsymbol{f}^{(w)^T} \boldsymbol{f}^{(v)}|} \sum_{w \in \mathcal{I}(v)} |\boldsymbol{f}^{(w)^T} \boldsymbol{f}^{(v)}| \cdot M_w\right). \tag{1}$$

For brevity, we will define

$$\rho_{nbrs}(v) = \sum_{w \in \mathcal{I}(v)} |\boldsymbol{f}^{(w)^T} \boldsymbol{f}^{(v)}| \cdot M_w, \tag{2}$$

$$\hat{\rho}_{nbrs}(v) = \frac{1}{\mathbb{E}[\sum_{w \in \mathcal{I}(v)} |\boldsymbol{f}^{(w)^T} \boldsymbol{f}^{(v)}| \mid \sigma(v)]} \cdot \rho_{nbrs}(v) = \frac{1}{Z(v)} \cdot \rho_{nbrs}(v). \tag{3}$$

If the set $\mathcal{I}(v) = \emptyset$, then we set $M_v \sim \text{Uniform}([0,1])$.
We denote by $p_{M_v}(\cdot \mid \sigma(v))$ the conditional density of M_v.

Remark 4 *The model defined in Definition 3 captures scenarios with a homophily effect, in which nodes assign more weight to other nodes whose feature vectors are more similar to its own. The parameter a gives the relative strength of the latent bias b versus the impact of the messages of the infecting neighbors.*

This could be extended to a richer parametric model in which the Euclidean dot product is replaced by an inner product whose defining matrix would then form the set of parameters of the model. In this case, learning this matrix would then allow us to determine the relative importance of different features in determining how a vertex assigns weight to the messages of its neighbors. Our results could be generalized to this model, but at the cost of increasing the complexity of our theorem statements. Specifically, this added complexity would arise from the fact that the approximation objective, naïvely stated, would be ill-posed, as infinitely many parameter matrices would be equivalent and would, as a result, have equal likelihood.

By appealing to concentration of $\sum_{w \in \mathcal{I}(v)} |\boldsymbol{f}^{(w)^T} \boldsymbol{f}^{(v)}|$, it can be shown that the message log likelihood function of the above model is asymptotically equivalent to that of a simpler model in which we replace that sum with $Z(v)$, its conditional expectation. It thus suffices to consider the simpler model wherein

$$M_v \sim \mathcal{D}\left(ab + (1-a) \cdot \hat{\rho}_{nbrs}(v)\right). \tag{4}$$

For an infection sequence $S = (S_0, ..., S_N)$, we will write $M(t) = \{(v, M_v) \mid v \in S_t\}$. We will write

$$M = \bigcup_{t=0}^{N} M(t). \tag{5}$$

We now come to the estimation problem that we would like to solve.

Message spreading model parameter estimation problem Suppose that a sample infection sequence and sequence of message sets $\text{Obs} = (S, M)$ is generated according to the independent cascade model and the feature-based message spreading model with parameters $\theta_{message} = (a, b)$. We denote by $\hat{Obs} = (\hat{S}, \hat{M})$ the observed value of the random variable Obs.

Given knowledge of the graph, cascade model, and node feature vectors, as well as that $\text{Obs} = \hat{Obs}$, our task is to present an estimator $\hat{\theta}_{message}$ of $\theta_{message}$ such that

$$\|\hat{\theta}_{message} - \theta_{message}\|_\infty < \epsilon \tag{6}$$

with probability at least $1 - \delta$.

2.2. Main theoretical results

To solve the problem stated above, we take an approximate maximum likelihood approach.

Derivation of the message model log likelihood We first define and derive an exact expression for the log likelihood function for the message model parameters. We will first derive the *exact* log likelihood function, which is defined as usual, but then we will throw away terms that do not depend on $\theta_{message}$. We will call the resulting expression the message log likelihood.

The exact log likelihood function $\tilde{\mathcal{L}}(\theta_{message} \mid \text{Obs})$ of $\theta_{message}$ is as follows:

$$\tilde{\mathcal{L}}(\theta_{message} \mid \text{Obs} = \hat{Obs} = ((\hat{S}_j)_{j=0}^N, (\hat{M}_v)_{v=1}^n)) \tag{7}$$

$$= \log \Pr[S_0 = \hat{S}_0] + \log p_{M(0)}(\hat{M}(0) \mid S_0 = \hat{S}_0) + \log \left(\prod_{t=0}^{|S|} \Pr[S_t \mid S_{t-1}] p_{M(t)}(\hat{M}(t) \mid S_t, M(t-1)) \right) \tag{8}$$

$$= \sum_{t=0}^{|S|} \log \Pr[S_t \mid S_{t-1}] + \sum_{t=0}^{|S|} \log p_{M(t)}(\hat{M}(t) \mid S_t = \hat{S}_t, M(t-1) = \hat{M}(t-1)). \tag{9}$$

Here we have used the fact that the messages of vertices, as a function of time of infection, satisfy a Markov property: conditioned on $M(t-1)$ and S_t, $M(t)$ is independent of anything else. The second component is the only one that depends on $\theta_{message}$, and so we define

$$\mathcal{L}(\theta_{message} \mid \text{Obs}) = \sum_{t=0}^{|S|} \log p_{M(t)}(\hat{M}(t) \mid S_t = \hat{S}_t, M(t-1) = \hat{M}(t-1)). \tag{10}$$

We call this the message log likelihood.

We further have the following:

$$\sum_{t=0}^{|S|} \log p_{M(t)}(\hat{M}(t) \mid S_t, M(t-1)) = \sum_{t=0}^{|S|} \sum_{v \in S_t} \log p_{M_v}(\hat{M}_v \mid \sigma(v)). \tag{11}$$

By conditioning on the specific set of infectors of v, we have

$$p_{M_v}(\hat{M}_v \mid \sigma(v)) = \sum_{\iota \in 2^{\mathcal{N}_\mathcal{A}(v)} \setminus \emptyset} \Pr[\mathcal{I}(v) = \iota \mid \sigma(v)] \cdot p_{M_v}(\hat{M}_v \mid \mathcal{I}(v) = \iota, M_\iota). \tag{12}$$

Here, the sum has one term for each nonempty element ι of the power set of $\mathcal{N}(v) \cap S_{t-1} = \mathcal{NA}(v)$. It is easy to further derive an exact expression for the terms in the sum (12), using the definitions of the independent cascade model and our message model.

Thus, naïvely, to evaluate the message model log likelihood (7), we would have to evaluate the sum (12), which has a number of terms exponential in the number of active neighbors of each given vertex. This becomes intractable if the cascade and graph are dense enough.

Approximating the log likelihood via a local central limit theorem In order to derive a tractable approximation to the message log likelihood, we observe that the probability expression in (11) can be approximated via a local central limit theorem for $\rho_{nbrs}(v)$. In particular, we define

$$\mu_{nbrs}(v) = \mathbb{E}[\rho_{nbrs}(v) \mid \sigma(v)] \tag{13}$$

$$\sigma^2_{nbrs}(v) = \mathrm{Var}[\rho_{nbrs}(v) \mid \sigma(v)]. \tag{14}$$

Both of these quantities can be expressed exactly and can be calculated efficiently using at most a constant number of iterations over $\mathcal{NA}(v)$ (since the terms of $\rho_{nbrs}(v)$ are independent, conditioned on $\sigma(v)$).

Our plan now is to make an approximation to the probability expression in (11): Note that

$$p_{M_v}(\hat{M}_v \mid \sigma(v)) = \int_{z=0}^{1} \Pr[ab + (1-a)\hat{\rho}_{nbrs}(v) = z \mid \sigma(v)] \cdot p_{\mathcal{D}}(\hat{M}_v \mid z)\, dz. \tag{15}$$

Now, we will approximate $\Pr[ab + (1-a)\hat{\rho}_{nbrs}(v) = z \mid \sigma(v)]$ as follows:

$$\Pr[ab + (1-a)\hat{\rho}_{nbrs}(v) = z \mid \sigma(v)] = \Pr\left[\hat{\rho}_{nbrs}(v) = \frac{z-ab}{1-a} \mid \sigma(v)\right] \tag{16}$$

$$= \Pr\left[\rho_{nbrs}(v) = Z(v) \cdot \left(\frac{z-ab}{1-a}\right) \mid \sigma(v)\right] \tag{17}$$

$$= \Pr\left[\frac{\rho_{nbrs}(v) - \mu_{nbrs}(v)}{\sigma_{nbrs}(v)} = \frac{Z(v)\left(\frac{z-ab}{1-a}\right) - \mu_{nbrs}(v)}{\sigma_{nbrs}(v)} \mid \sigma(v)\right] \tag{18}$$

$$\approx \frac{\Delta_{min}(v)}{\sigma_{nbrs}(v)} \cdot \phi\left(\frac{Z(v)\left(\frac{z-ab}{1-a}\right) - \mu_{nbrs}(v) + \frac{1}{2}\Delta_{min}(v)}{\sigma_{nbrs}(v)}\right), \tag{19}$$

where we define $\Delta_{min}(v)$ to be

$$\Delta_{min}(v) = \min_{w \in \mathcal{NA}(v)} (M_w - p_{net}(w,v)M_w). \tag{20}$$

The value of $\Delta_{min}(v)$ is significant because it satisfies the following identity:

$$\Pr\left[\rho_{nbrs}(v) = Z(v) \cdot \left(\frac{z-ab}{1-a}\right) \mid \sigma(v)\right] = \Pr\left[\rho_{nbrs}(v) \in \left[Z(v) \cdot \left(\frac{z-ab}{1-a}\right) + \Delta_{min}(v)\right) \mid \sigma(v)\right]. \tag{21}$$

Here, we recall that $\phi(\cdot)$ is the probability density function of the standard normal distribution. We will make the approximate equality (19) formal by defining our **Gaussian approximation to the message log likelihood**:

$$\hat{\mathcal{L}}(\theta_{message} \mid \text{Obs} = (\hat{S}, \hat{M})) \tag{22}$$

$$= \sum_{t=0}^{|S|} \sum_{v \in S_t} \log \left(\int_{z=0}^{1} \frac{\Delta_{min}(v)}{\sigma_{nbrs}(v)} \cdot \phi \left(\frac{Z(v) \cdot \left(\frac{z-ab}{1-a} \right) - \mu_{nbrs}(v) + \frac{1}{2}\Delta_{min}(v)}{\sigma_{nbrs}(v)} \right) \cdot p_{\mathcal{D}}(\hat{M}_v \mid z)\, dz \right). \tag{23}$$

Our first main result, Theorem 5, gives an approximation bound on this expression for the log likelihood.

Theorem 5 (Approximation of the log likelihood function)
Suppose that the feature vectors $\boldsymbol{f}^{(w)}$ are such that all pairwise dot products are uniformly $\Omega(1)$. Suppose, further, that $a, b \neq 0$ or 1. Finally, suppose that the graph G is sampled from the Erdős-Rényi model with parameter $p = p(n) > \frac{\kappa \log n}{n}$, $\kappa > 1$.

We have the following approximation bound for the message log likelihood:

$$\left\| \frac{\mathcal{L}(\cdot \mid \text{Obs} = (S, M)) - \hat{\mathcal{L}}(\cdot \mid \text{Obs} = (S, M))}{\mathcal{L}(\cdot \mid \text{Obs} = (S, M))} \right\|_{\infty} = o(1). \tag{24}$$

This holds with probability $1 - o(1)$.

This implies that, provided that the graph is sufficiently dense (e.g., with average degree $> \kappa \log n$, with $\kappa > 1$) and provided that $p_{net}(e)$ is uniformly bounded away from 0 for all e, the relative error in approximating the message log likelihood by our Gaussian approximation is $o(1)$. We note that the Erdős-Rényi assumption may be relaxed substantially. All that is required of the graph structure is that it is sufficiently dense. Our results may be extended to graphs with sparse cuts by applying the Gaussian approximation only to the terms of the log likelihood function correpsonding to vertices with sufficiently large values for $|\mathcal{NA}(v)|$. The remaining terms, for which this set is small, may be approximated by a sampling approach.

Our next result is a consequence of Theorem 5 that says that the maximum of the true likelihood function is close to the maximum of the approximate likelihood function.

Theorem 6 (Maximum approximate likelihood is a good approximation of the maximum likelihood)
Let

$$\theta_* = \arg\max_{\theta} \mathcal{L}(\theta \mid \text{Obs} = (S, M)) \tag{25}$$

$$\hat{\theta}_{message} = \arg\max_{\theta} \hat{\mathcal{L}}(\theta \mid \text{Obs} = (S, M)). \tag{26}$$

Suppose that the following condition on $\mathcal{L}(\cdot \mid \text{Obs}(S, M))$ holds:

1. *With probability $> 1 - \delta$ over the observations, if*

$$\|\theta - \theta_*\|_{\infty} = \Omega(1), \tag{27}$$

then

$$\left| \frac{\mathcal{L}(\theta \mid \mathrm{Obs})}{\mathcal{L}(\theta_* \mid \mathrm{Obs})} \right| > 1 + \Omega(1). \tag{28}$$

We call this the well-posedness condition.

Then we have the following bound, with probability $> 1 - \delta$*:*

$$\|\theta_* - \hat{\theta}_{message}\|_\infty = o(1). \tag{29}$$

Remark 7 (Discussion of the well-posedness condition on the log likelihood function) *The well-posedness condition deserves further comment: it can be shown that provided that the log likelihood function, with* Obs *sampled from the model with the true parameter* $\theta_{message}$*, is well-concentrated around its mean, then the* failure *of this condition to hold implies a KL divergence upper bound between the distribution of* Obs *sampled from the model with parameter* $\theta_{message}$ *versus* Obs *sampled from the model with parameter* θ*. This KL divergence upper bound is of the form* $o(\mathcal{L}(\theta_* \mid \mathrm{Obs}))$*. It is likely that this can be strengthened to* $o(1)$ *by showing that the KL divergence consists of* $\Theta(\mathcal{L}(\theta \mid \mathrm{Obs}))$ *terms that are all within a fixed constant factor of one another, with the property that each term is* $\Omega(1)$ *whenever* $\|\theta - \theta_*\|_\infty = \Omega(1)$*.*

The resulting stronger KL divergence upper bound could then be translated to a nontrivial upper bound on the total variation distance between these two distributions via Pinsker's inequality. This, in turn, would imply an inapproximability result for $\theta_{message}$*. Thus, in essence, this theorem says that either it is statistically impossible to achieve an estimation error of* $o(1)$ *with probability* $1 - o(1)$*, or our approximate maximum likelihood estimator is close to the true maximum likelihood estimator.*

Our next result says that the approximate log likelihood can be computed efficiently.

Theorem 8 (Running time of our likelihood function approximation) *There exists an algorithm that computes* $\hat{\mathcal{L}}(\cdot \mid \mathrm{Obs} = (S, M))$ *in time* $O(|E(G)| + N(S))$*, where the* $O(\cdot)$ *is uniform over all parameters. Here,* $E(G)$ *is the set of edges in the graph* G*.*

This should be compared with the running time of a sampling-based estimator. The running time of such an estimator, for a given accuracy level, would have a multiplicative factor for each vertex that would on the observation.

Remark 9 *Our results extend trivially to the case of multiple iid cascade samples.*

3. Conclusions and future work

We have exhibited a message spreading model, akin to a simple model of the spread of opinions via homophily, in which a naïve implementation of the maximum likelihood estimator is computationally intractable. We proposed the use of a local central limit theorem to approximate the likelihood function and showed that the resulting approximation is both computationally tractable and provably accurate. Moreover, its maximum value is close to that of the true log likelihood of the model. Our view is that local limit theorems could be more broadly applied in similar estimation problems to yield tractable estimators with provable guarantees. In particular, our hope is to generalize our

results to a more realistic model for opinion dynamics and to apply our estimator to real data. Theoretically, it would also be desirable to provide an explicit rate of convergence for our approximation and to analyze the convergence rate of the maximum likelihood estimator for this problem as a function of graph structure. Moreover, it is both theoretically and practically of interest to analyze the optimization landscape of the approximate log likelihood function to give theoretical guarantees on optimization algorithms such as gradient ascent that could be applied to it.

4. Glossary of notation

1. G – A graph on the vertices $[n] = \{1, ..., n\}$.
2. $\mathcal{C}$ – A cascade model.
3. $\mathcal{N}(v)$ – Neighbors of v.
4. $\mathcal{TI}(v)$ – The time at which v was infected.
5. $\mathcal{NA}(v)$ – The set of neighbors of v that are active in the timestep during which v was infected.
6. $\mathcal{I}(v)$ – The set of neighbors of v that infected vertex v. $\mathcal{I}_\Omega$ is the collection of these sets (NOT the union) for every vertex $v \in \Omega$.
7. $\boldsymbol{f}^{(v)}$ – The feature vector of vertex v in $\mathbb{R}^d$.
8. $p_{\mathcal{D}}(\cdot \mid \mu)$ – The probability density function of the distribution $\mathcal{D}$ with mean parameter μ.
9. $M_v \in [0, 1]$ – The message assigned to vertex v. A vertex that was never infected has message $M_v = \emptyset$.
10. $M(t)$ – The set of messages of vertices infected at time t.
11. M_Ω – The map from $\Omega \to [0, 1]$, where Ω is a vertex subset. $M = M_{[n]}$ – The set of all vertex messages.
12. $\text{Obs} = (S, M)$.
13. $S = (S_0, S_1, ..., S_T)$ – An infection sequence, as in our fastclock paper.
14. $\theta_{message}$ – The message model parameters – e.g., (a, b). Also, could involve vertex features.
15. $\phi(z)$ – PDF of $\mathcal{N}(0, 1)$.
16. $\sigma(v) = (\mathcal{NA}(v), M_{\mathcal{NA}(v)})$.
17. $\mu_{nbrs}(v) = \mathbb{E}[\rho_{nbrs}(v) \mid \sigma_v]$.
18. $\sigma^2_{nbrs}(v) = \text{Var}[\rho_{nbrsnbrs}(v) \mid \sigma_v]$.

5. Proofs

Here we give full proofs of all results.

5.1. Proof of the main log likelihood approximation theorem, Theorem 5

Our proof relies on Theorem 1 from Mineka and Silverman (1970). It is a local limit theorem for a sum $S_n = \sum_{j=1}^{n} X_k$ of independent random variables X_k with distribution functions F_k and

$$\mathbb{E}[X_k] = 0, \qquad \mathbb{E}[X_k^2] = \sigma_k^2, \qquad \sum_{j=1}^{n} \sigma_k^2 = V_n. \tag{30}$$

To state the theorem, we need three conditions:

1. There exist $M, c > 0$ such that, for all k,

$$\sigma_k^{-2} \int_{-M}^{M} x^2 \, dF_k(x) \geq c. \tag{31}$$

Provided that the X_k are uniformly bounded, this holds trivially because we can take M equal to the bound and $c = 1$. Then we would have exact equality here.

It can be shown (details in Mineka and Silverman (1970)) that this condition implies the following two properties, which are used to state the remaining two conditions:

(a) There exists $C' > 0$ such that

$$\Pr[|X_k| < M] \geq C'. \tag{32}$$

(b) There exists a bounded sequence of numbers $\{a_k\}$ such that, for all $\delta > 0$,

$$\inf_{k \geq 1} \{\Pr[|X_k - a_k| < \delta\} > 0. \tag{33}$$

2. Let $A(t, \epsilon)$, for $t \neq 0$, $\epsilon > 0$, be defined as follows:

$$A(t, \epsilon) = \{x \mid |x| < M, |xt - \pi m| \geq \epsilon \text{ for all } m \in \mathbb{Z} \text{ with } |m| \leq M\}. \tag{34}$$

In words, this is the set of M-bounded x such that xt is sufficiently bounded away from any element of a finite set of integer multiples of π.

For each $t \neq 0$, there exists $\epsilon = \epsilon(t)$ such that

$$\frac{1}{\log V_n} \sum_{k=1}^{n} \Pr[X_k - a_k \in A(t, \epsilon)] \xrightarrow{n \to \infty} \infty, \tag{35}$$

with $\{a_k\}$ satisfying the property (33).

For us, $M = 1$, and so $A(t, \epsilon)$ is more explicitly given by

$$A(t, \epsilon) = \{x \mid |x| < 1, |x - \pi m/t| \geq \epsilon/t, m = \{-1, 0, 1\}\}. \tag{36}$$

We will define

$$\epsilon = \epsilon(t) = \begin{cases} |t|^{100} & |t| < 1 \\ 1/(|t| + 1)^{100} & |t| \geq 1 \end{cases} \tag{37}$$

Furthermore, our random variables X_k will take one of two values each, with probability uniformly bounded away from 0 and 1 over all k: ℓ_k, r_k, which are themselves random and independently drawn from continuous distributions related to $\mathcal{D}$ (specifically, they will each be a sample from $\mathcal{D}$, multiplied by a Bernoulli random variable). We can then set $a_k = \ell_k$, so that $X_k - a_k \in \{0, r_k - \ell_k\}$, so that $\Pr[X_k - a_k \in A(t, \epsilon)]$ is strictly positive provided that $r_k - \ell_k$ avoids the three points $\{-\pi/t, \pi/t, 0\}$. The probability that there is some t for which $\Theta(n)$ of the $r_k - \ell_k$ do not avoid these points (i.e., for which $\Theta(n)$ of the $r_k - \ell_k$ are all within ϵ/t of $-\pi/t, 0$, or π/t) is at most $e^{-\Omega(n)}$, since these random variables are independent, and each has a uniformly positive probability of avoiding any given value.

3. Lindeberg's condition must be satisfied: for all $\epsilon > 0$,

$$\frac{1}{V_n} \sum_{k=1}^{n} \int_{|x| > \epsilon \sqrt{V_n}} x^2 \, dF_k(x) \xrightarrow{n \to \infty} 0. \tag{38}$$

This is automatically satisfied if $V_n \to \infty$ and the X_k are uniformly bounded by M.

We then have the following theorem.

Theorem 10 (Mineka and Silverman (1970), Theorem 1) *Under the three conditions described above, we have*

$$\lim_{n \to \infty} \sqrt{2\pi V_n} \Pr[S_n \in (z, z + \Delta)] - \Delta \exp\left(-\frac{1}{2}(z + \frac{1}{2}\Delta)^2 / V_n \right) = 0. \tag{39}$$

The implication of Theorem 10 is as follows:

Corollary 11 (Approximation theorem for the probability mass function of $\hat{\rho}_{nbrs}(v)$) *Let $\phi(z) = \frac{1}{2\pi} e^{-\frac{1}{2}z^2}$ be the PDF of the standard normal distribution.*

We have that provided that $z = \Theta(1)$,

$$\Pr[ab + (1-a)\hat{\rho}_{nbrs}(v) = z \mid \sigma(v)] \tag{40}$$

$$= \frac{\Delta_{min}(v)}{\sigma_{nbrs}(v)} \cdot \phi\left(\frac{Z(v) \cdot \left(\frac{z - ab}{1-a} \right) - \mu_{nbrs}(v) + \frac{1}{2}\Delta_{min}(v)}{\sigma_{nbrs}(v)} \right) + o(1/\sigma_{nbrs}(v)). \tag{41}$$

Proof This follows directly from Theorem 10 by ordering the elements w_j of $\mathcal{NA}(v)$ from w_1 to $w_{|\mathcal{NA}(v)|}$ and defining $B_j \sim \text{Bernoulli}(p_{net}(w_j, v))$ and

$$\tilde{X}_j = B_j \cdot M_{w_j}. \tag{42}$$

The definition of X_j is then simply given by centering and normalization of $\tilde{X}_j$.

Note that, with probability 1, because of our stipulations on $\mathcal{D}$, the conditions of the theorem hold with probability 1 over $\sigma(v)$. This completes the proof. ■

Completing the log likelihood approximation proof We can complete the proof as follows: we write

$$\mathcal{L}(\theta \mid \text{Obs} = (S, M)) \tag{43}$$

$$= \sum_{t=0}^{|S|} \sum_{v \in S_t} \log \left(\int_{z=0}^{1} \Pr[ab + (1-a)\hat{\rho}_{nbrs}(v) = z \mid \sigma(v)] \cdot p_{\mathcal{D}}(\hat{M}_v \mid z)\, dz \right) \tag{44}$$

$$= \sum_{t=0}^{|S|} \sum_{v \in S_t} \log \left(\int_{0}^{1} \frac{\Delta_{min}(v)}{\sigma_{nbrs}(v)} \phi \left(\frac{Z(v) \cdot \left(\frac{z-ab}{1-a} \right) - \mu_{nbrs}(v) + \frac{1}{2}\Delta_{min}(v)}{\sigma_{nbrs}(v)} \right) \cdot p_{\mathcal{D}}(\hat{M}_v \mid z)\, dz + o(1/\sigma_{nbrs}(v)) \right) \tag{45}$$

$$= \sum_{t=0}^{|S|} \sum_{v \in S_t} \log \left(\int_{0}^{1} \frac{\Delta_{min}(v)}{\sigma_{nbrs}(v)} \phi \left(\frac{Z(v) \cdot \left(\frac{z-ab}{1-a} \right) - \mu_{nbrs}(v) + \frac{1}{2}\Delta_{min}(v)}{\sigma_{nbrs}(v)} \right) \cdot p_{\mathcal{D}}(\hat{M}_v \mid z)\, dz \right) \tag{46}$$

$$+ \sum_{t=0}^{|S|} \sum_{v \in S_t} o(1/\sigma_{nbrs}(v)) \tag{47}$$

$$= \hat{\mathcal{L}}(\theta \mid \text{Obs} = (S, M)) + o(n \cdot (\min_{v \in [n]} \sigma_{nbrs}(v))^{-1}). \tag{48}$$

Here, the first equality is by definition of the message log likelihood. The second equality is by Corollary 11. The third is by the observation that the integral is $\Theta(1/\sigma_{nbrs}(v))$, while the other term inside the logarithm is $o(1/\sigma_{nbrs}(v))$. The expression is a result of the following chain of asymptotic equalities: $\log(f + o(f)) = \log(f \cdot (1 + o(1))) = \log(f) + \log(1 + o(1)) = \log(f) + o(1)$.

To complete the proof, we show that, with high probability, $\mathcal{L}(\theta \mid \text{Obs}(S, M)) = \Omega(n \cdot (\min_{v \in [n]} \sigma_{nbrs}(v))^{-1})$. This will imply that

$$\left\| \frac{\mathcal{L}(\cdot \mid \text{Obs} = (S, M)) - \hat{\mathcal{L}}(\cdot \mid \text{Obs} = (S, M))}{\mathcal{L}(\cdot \mid \text{Obs} = (S, M))} \right\|_{\infty} \leq o(1). \tag{49}$$

To show the remaining probabilistic lower bound on the absolute value of the message log likelihood, it suffices to show that $\Omega(N(S))$ terms in its defining sum are $\Omega(1)$, regardless of the value of θ. Note that this is trivially the case because of the fact that the density of the distribution of each message M_v, conditioned on $\sigma(v)$, is supported on all of $[0, 1]$. This completes the proof of Theorem 5.

5.2. Proof of the maximum likelihood approximation theorem, Theorem 6

The proof depends on the following lemma.

Lemma 12 (Uniform approximation of functions and their extrema) *Suppose that f and g are two functions from $\mathbb{R}^k \to \mathbb{R}$, for some fixed $k > 0$. Suppose that*

$$f(\theta) = g(\theta) \cdot (1 + h_N), \tag{50}$$

for some $h_N = o(1)$ as $N \to 0$. Let

$$\theta_*^{(f)} = \arg\max_\theta f(\theta), \tag{51}$$

$$\theta_*^{(g)} = \arg\max_\theta g(\theta). \tag{52}$$

Suppose, further, that f satisfies the well-posedness condition described in Theorem 6. Then we have

$$\|\theta_*^{(g)} - \theta_*^{(f)}\|_\infty = o(1). \tag{53}$$

Proof First, note that (50) implies that the value at $\theta_*^{(f)}$ for both functions is similar:

$$f(\theta_*^{(f)}) = g(\theta_*^{(f)})(1 + h_N) \tag{54}$$

$$f(\theta_*^{(g)}) = g(\theta_*^{(g)})(1 + h_N). \tag{55}$$

Furthermore, by optimality of $\theta_*^{(g)}$ for g, we must have that

$$g(\theta_*^{(g)}) = f(\theta_*^{(g)})(1 + o(1)) \geq g(\theta_*^{(f)}) = f(\theta_*^{(f)})(1 + o(1)). \tag{56}$$

This implies that

$$f(\theta_*^{(g)}) = f(\theta_*^{(f)})(1 + o(1)). \tag{57}$$

Under the well-posedness condition, we have that if $\|\theta_*^{(g)} - \theta_*^{(f)}\| > \epsilon$, for some $\epsilon = \Omega(1)$, then we must have that (57) is violated. This implies that

$$\|\theta_*^{(g)} - \theta_*^{(f)}\|_\infty = o(1), \tag{58}$$

as desired. ■

The claim in the theorem follows directly by plugging in $f(\theta) = \mathcal{L}(\theta \mid \text{Obs})$ and $g(\theta) = \hat{\mathcal{L}}(\theta \mid \text{Obs})$ in Lemma 12.

5.3. Proof of the running time theorem, Theorem 8

Calculation of $\hat{\mathcal{L}}(\theta \mid \text{Obs} = (\hat{S}, \hat{M}))$ can be done term-by-term, for a total of at most $N(S)$ iterations. Each term requires the calculation of $\mu_{nbrs}(v)$ and $\sigma_{nbrs}(v)$, each taking time $O(|\mathcal{NA}(v)|)$. The integral itself takes $O(1)$ time, by assumption. Thus, the running time is $O(|E(G)| + N(S))$, where $E(G)$ is the number of edges in the graph. This completes the proof.

References

François Baccelli, Avhishek Chatterjee, and Sriram Vishwanath. Pairwise stochastic bounded confidence opinion dynamics: Heavy tails and stability. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 1831–1839, 2015. doi: 10.1109/INFOCOM.2015.7218565.

Sivaraman Balakrishnan, Martin Wainwright, and B. Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Ann. Stat.*, 45, 08 2014. doi: 10.1214/16-AOS1435.

Ross Berkowitz. A quantitative local limit theorem for triangles in random graphs, 2017.

Abhimanyu Das, Sreenivas Gollapudi, and Kamesh Munagala. Modeling opinion dynamics in social networks. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, page 403–412, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450323512. doi: 10.1145/2556195.2559896. URL https://doi.org/10.1145/2556195.2559896.

Abir De, Isabel Valera, Niloy Ganguly, Sourangshu Bhattacharya, and Manuel Gomez Rodriguez. Learning and forecasting opinion dynamics in social networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/f340f1b1f65b6df5b5e3f94d95b11daf-Paper.pdf.

Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009. ISBN 978-0-521-89806-5. URL http://www.cambridge.org/uk/catalogue/catalogue.asp?isbn=9780521898065.

Vishesh Jain, Will Perkins, Ashwin Sah, and Mehtaab Sawhney. Approximate counting and sampling via local central limit theorems, 2021.

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. ISBN 1581137370. doi: 10.1145/956750.956769. URL https://doi.org/10.1145/956750.956769.

J. Mineka and S. Silverman. A Local Limit Theorem and Recurrence Conditions for Sums of Independent Non-Lattice Random Variables. *The Annals of Mathematical Statistics*, 41(2):592 – 600, 1970. doi: 10.1214/aoms/1177697099. URL https://doi.org/10.1214/aoms/1177697099.

Marios Papachristou and Dimitris Fotakis. Stochastic opinion dynamics for interest prediction in social networks, 2021.