MAXIMUM LIKELIHOOD RUMOR SOURCE DETECTION IN A STAR NETWORK

Sam Spencer and R. Srikant

Coordinated Science Laboratory and
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

ABSTRACT

Here we examine the problem of rumor source identification in star networks. We assume the SI model for rumor propagation with exponential waiting times. We consider the case where a rumor originates from a single source, and find an explicit, non-iterative, maximum likelihood estimate for the source given the observed infection pattern. The theoretical derivation is supported by computational data. We contrast this estimator with the "rumor center" estimator of Shah and Zaman. Unlike rumor centrality, our ML estimator admits the possibility of more than two equiprobable maxima for a given infection pattern, and while a unique rumor center is always equivalent to the distance center, we show that this is not the case for our ML estimator.

Index Terms— Infection source identification, SI model, star network, rumor source identification, maximum likelihood

1. INTRODUCTION

Rumor propagation and source detection problems (and their mathematical analogs) arise in a variety of contexts, including cybersecurity, information assurance, privacy, epidemiology, and social network analysis. We consider the problem of estimating the source of a rumor that spreads in a star network. Given a snapshot of the infected region at a given point in time, we wish to characterize the maximum likelihood estimate of the source of the rumor.

A true ML estimator is very difficult to produce for general graphs. Accordingly, it is necessary either to restrict the classes of graphs of interest, or to use an approximation to the ML estimator. In this paper, we will consider a specific topology (the star), but attempt to characterize it more precisely than most of the approaches discussed below.

Similar problems and approaches have been considered previously. In [1], the authors present the problem of ML estimation of a single source of a rumor which propagates according to the SI model. They propose a metric known as "rumor centrality", and show that it is an ML estimator for regular trees. When generalized beyond regular trees, the computation of the original metric becomes intractable, so a breadth-first search-based heuristic is incorporated into the metric for non-regular trees, which provides asymptotically good results for many cases.

In [2] and [3], the authors consider approaches to more complex propagation models (SIS and SIRI, respectively), but since they are iterative in nature, and operate on individual instances of the problem (rather than our general, closed-form solution), they do not lend themselves to the types of characterizations we offer here.

In [4], the authors consider the problem of multiple sources in a tree under the SIR model, and where the infection process occurs in discrete time. Their method is applied to regular trees of degree greater than two, whereas our trees are irregular, and almost all nodes are of degree two.

Our approach builds upon the work in [5], but expands the development from a line graph to a star topology.

Other current work in this area includes [6], which looks at this problem from the opposite side — trying to conceal the source of the rumor by manipulating the propagation model. This type of approach could be used to provide anonymity for whistle-blowers or political dissidents, for example.

2. MODEL

For present purposes, we define a star network as a point O, along with m "arms" of nodes proceeding outward from O. The nodes of each arm will be numbered starting with 1 (for the node adjacent to O) and increasing from there. We use the SI infection model with edge-based propagation in continuous time to describe the spreading of the rumor. That is, nodes are either "susceptible" (have not yet heard the rumor) or "infected" (have already heard it). If a susceptible node shares an edge with an infected neighbor, then the infection will "traverse" that edge and infect the susceptible node with a waiting time that is exponentially distributed with mean T. Once infected, a node remains that way indefinitely. An important consequence of this model is that we can invoke the

Research supported in part by AFOSR MURI FA 9550-10-1-0573 and DTRA Grant HDTRA1-15-1-0003.

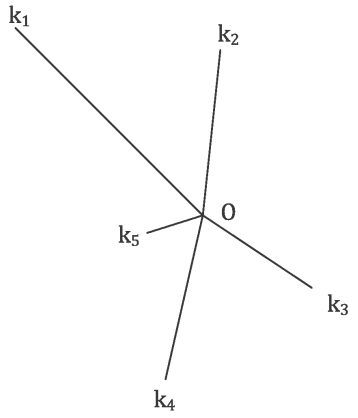

Fig. 1. A sample infection on a star network for $m = 5$.

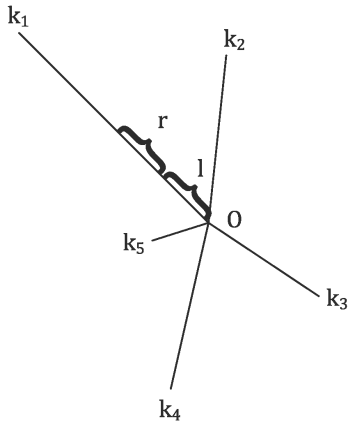

Fig. 2. Illustrating the calculation in (3). Note that r can range from 0 up to $k_1 - l$.

memoryless property of the system to state that at any given time, the next infection is equally likely to occur along any outgoing edge from the current infected set. For a given observed infection pattern (the subgraph of infected nodes at some point in time), we wish to find the maximum likelihood estimate of the source giving rise to that infection pattern.

Our infection pattern consists of O, along with the closest k_i nodes along each arm i. If the infection were confined to a single arm, we could simply consider the problem on a line graph, and the ML solution is well-known to be the midpoint of the infection (in fact, for a uniform prior, the likelihood function follows a binomial distribution on the infected nodes [1], [5]). Since the infection arises from a single source, it must be contiguous, so any infection which spans multiple arms must also include O. This is illustrated in Fig. 1.

3. ANALYSIS

Accordingly, we start by computing the likelihood of the observed pattern occurring at some point in time given that the rumor originated at O. In this case, each subsequent infection can spread along each of the m arms with probability $1/m$. Let $K = \sum_{i=1}^{m} k_i$. Then the probability of observing k_1 infections along arm 1, k_2 infections along arm 2, etc. is given by a multinomial distribution.

$$P(O; k_1, k_2, \ldots, k_m) = \frac{K!}{k_1!k_2! \ldots k_m! m^K} \tag{1}$$

If instead, the rumor source is located along one of the arms (let us assume, without loss of generality, that the source is located on arm 1) at node l. Then the propagation of the rumor occurs in two phases: At first, the infection spreads along arm 1 in either direction, until the inward propagation reaches O. At that point, it can spread outward along any of the m arms. Accordingly, we decompose the set of possibilities according to the extent that the infection proceeds outward along arm 1 before the inward propagation reaches O. Suppose the infection reaches an additional r nodes beyond l before reaching O, as shown in Fig. 2. Then the probability of r outward infections and $l-1$ inward infections (in any order) followed the last inward infection reaching O is $\frac{\binom{r+l-1}{l-1}}{2^{r+l}}$. Afterwards, the probability of fulfilling the remaining infections exactly can be computed using (1), replacing k_1 with $k_1 - (r+l)$. Multiplying these two probabilities, we obtain

$$\begin{aligned} &P(l, r; k_1, k_2, \ldots, k_m) \\ &= \frac{\binom{r+l-1}{l-1}}{2^{r+l}} \frac{(K-(r+l))!}{(k_1-(r+l))!k_2! \ldots k_m! m^{K-(r+l)}} \\ &= \frac{(r+l-1)!}{r!(l-1)!2^{r+l}} \frac{(K-(r+l))!}{(k_1-(r+l))!k_2! \ldots k_m! m^{K-(r+l)}}. \end{aligned} \tag{2}$$

Summing over all possible values of r, we obtain

$$\begin{aligned} &P(l; k_1, k_2, \ldots, k_m) \\ &= \sum_{r=0}^{k_1-l} \frac{\binom{r+l-1}{l-1}}{2^{r+l}} \frac{(K-(r+l))!}{(k_1-(r+l))!k_2! \ldots k_m! m^{K-(r+l)}} \\ &= \sum_{r=0}^{k_1-l} \frac{(r+l-1)!}{r!(l-1)!2^{r+l}} \frac{(K-(r+l))!}{(k_1-(r+l))!k_2! \ldots k_m! m^{K-(r+l)}}. \end{aligned} \tag{3}$$

In order to further analyze the situation, we will use the method of types. For a source with a uniform distribution on X, the probability of observing a type T of n samples with empirical distribution Q satisfies

$$\frac{1}{(n+1)^{|\chi|}} 2^{-nD(Q||U_X)} \leq P(T^n(Q)) \leq 2^{-nD(Q||U_X)} \tag{4}$$

where $|\chi|$ is the size of the set of choices [7]. Note that the lower and upper bounds are the same except for the leading term in the lower bound. Therefore, we will work with the upper bound for now, and consider the effect of the leading

term in the lower bound afterwards.

$$P(T^n(Q)) \leq 2^{-nD(Q||U_X)} = 2^{-n(\log|X| - H(Q))} = 2^{-n(\log|X| + \sum_X Q(x)\log Q(x))}. \quad (5)$$

Applying this to each of the phases of the infection yields

$$P(l; k_1, k_2, \ldots, k_m) = \sum_{r=0}^{k_1 - l} [2^{-(r+l)(1 - H(\frac{r}{r+l}, \frac{l}{r+l}))} * 2^{-(K-(r+l))(\log m - H(\frac{k_1-(r+l)}{K-(r+l)}, \frac{k_2}{K-(r+l)}, \ldots, \frac{k_m}{K-(r+l)}))}] \quad (6)$$

Remember that we are interested in finding the value of l that maximizes this expression, and observe that the value of the sum is asymptotically dominated by the term with the largest (or least negative) exponent. Furthermore, notice that the terms of the sum depend only on $r + l$ rather than r or l individually, with the exception of the $H(\frac{r}{r+l}, \frac{l}{r+l})$ in the first exponent. This value is maximized when $r = l$. Therefore, we can conclude that the dominant term of the sum for the maximizing value of l occurs when $r = l$. If this were not the case, we could replace r with r' and l with l', where $r' = l' = \frac{r+l}{2}$ and obtain a more dominant term with a different value of l. Accordingly, we will replace r with l going forward, and in the process, we eliminate the first part of the dominant term.

$$P(l; k_1, k_2, .., k_m) \leq 2^{-(K-2l)(\log m - H(\frac{k_1-2l}{K-2l}, \frac{k_2}{K-2l}, \ldots, \frac{k_m}{K-2l}))} \quad (7)$$

$$\leq 2^{-(K-2l)\log m + (K-2l)H(\frac{k_1-2l}{K-2l}, \frac{k_2}{K-2l}, \ldots, \frac{k_m}{K-2l})}$$

$$\leq 2^{-(K-2l)\log m - (K-2l)(\frac{k_1-2l}{K-2l}\log\frac{k_1-2l}{K-2l} + \sum_{i=2}^m \frac{k_i}{K-2l}\log\frac{k_i}{K-2l})}$$

$$\leq 2^{-(K-2l)\log m - (k_1-2l)\log\frac{k_1-2l}{K-2l} - \sum_{i=2}^m k_i \log\frac{k_i}{K-2l}} \quad (8)$$

Taking the exponent, and setting the derivative with respect to l to zero, we obtain

$$0 = 2\log m + 2\log(k_1 - 2l) + 2 - 2\log(K - 2l) - 2$$

which leads to

$$l = \frac{k_1 - \frac{\sum_{i=2}^m k_i}{m-1}}{2} \quad (9)$$

In other words, l is chosen so that the remaining length of arm 1 once O is reached is equal to the arithmetic mean of the other arms. Since we can apply this reasoning to any arm, we have a local maximum for any arm whose length is above average. However, the form of the exact solution in (3) makes it clear that the global maximum is attained when the longest arm is chosen to be arm 1. Let us denote this choice of l as l^*.

Having chosen l^* to optimize the upper bound, let us consider the effect of the leading coefficient in the lower bound. While it is different for each term of the sum in (6), it can be bounded from below by $\frac{1}{(K+1)^m}$. We can then use our earlier reasoning with (7) to show that

$$P(l; k_1, k_2, \ldots, k_m) \geq \frac{1}{(K+1)^m} 2^{-(K-2l)(\log m - H(\frac{k_1-2l}{K-2l}, \frac{k_2}{K-2l}, \ldots, \frac{k_m}{K-2l}))}. \quad (10)$$

Consider what happens if we allow the pattern to grow larger, but maintain the relative sizes of the k's (in other words, replace each k by nk, and let n go to infinity). Then we have

$$P(l; nk_1, nk_2, \ldots, nk_m) \leq 2^{-(nK-2l)(\log m - H(\frac{nk_1-2l}{nK-2l}, \frac{nk_2}{nK-2l}, \ldots, \frac{nk_m}{nK-2l}))} \quad (11)$$

and

$$P(l; nk_1, nk_2, \ldots, nk_m) \geq \frac{1}{(nK+1)^m} 2^{-(nK-2l)(\log m - H(\frac{nk_1-2l}{nK-2l}, \frac{nk_2}{nK-2l}, \ldots, \frac{nk_m}{nK-2l}))}. \quad (12)$$

Remember that l^* was chosen (proportional to the k's) in such a way as to minimize the (negated) exponent in (8). Therefore, letting $l = nl^*$ will yield the minimum exponent in (11) and (12) (n times the old optimal exponent). If, instead, we were to choose $l = l' \neq nl^*$ (relative to the k's), then the higher (negated) exponent will eventually cause the upper bound in (11) evaluated at l' to drop below the lower bound in (12) evaluated at l^*. Therefore, for sufficiently large instances, our choice of $l = l^*$ must be optimal. In fact, our empirical results suggest that this is the case even for smaller instances.

4. PROPERTIES AND CONTRAST WITH RUMOR CENTRALITY

In order to most easily see contrasts with rumor centrality, let us consider a special case of a star network. Based on the visual similarity to a biological structure, let us define a *neuron* as a region of a star network that includes O, and whose arms only take on two distinct lengths. The shorter arms, called *dendrites*, have length L_0, while the longer arms, called *axons*, have length $L_0 + L_1$, and we stipulate that both L_0 and L_1 are strictly greater than 0. This is illustrated in Fig. 3.

Suppose our infection pattern is a neuron that has a single axon. If there is only a single dendrite as well (meaning $m = 2$), then effectively this is the same as a line graph of length $2L_0 + L_1$. In this case, the distance center and the ML center would both be located on the axon at node $L_1/2$.

Now, consider what happens if the number of dendrites (and the corresponding m) increases. (Note that this different than choosing a larger m to begin with and allowing the

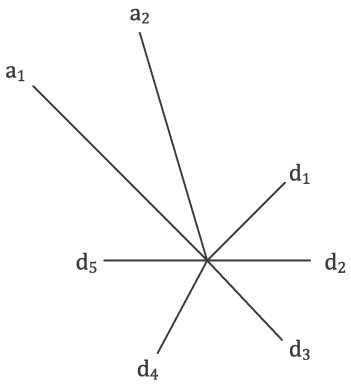

Fig. 3. A neuron with two axons and five dendrites.

"extra" arms to have length 0 at the outset.) The ML center remains at $2L_0 + L_1$, because the second term in the numerator of (9) remains constant. However, the distance center would begin to move towards O, and eventually reach it and stay there. Thus, the ML center and the distance center will no longer be the same. In contrast, [1] tells us that the distance center and rumor center are the same if the latter is unique, so we know that the ML center (which is unique in this case) and the rumor center cannot be equivalent. (This is not inconsistent with the findings in [1], since they do not claim the rumor center to be optimal in the ML sense under these conditions, but we have identified a fairly simple yet clear example where these two centers may differ significantly.) To illustrate a second significant difference between these two centrality measures, consider a neuron with n dendrites and $n + 1$ axons, where $n > 2$. It can be easily shown that the (unique) rumor center in this case is at O (otherwise there would be at least n equivalent rumor centers by symmetry, while [1] guarantees us that a tree can have at most 2 rumor centers). However, (9) tells us that there are $n + 1$ equiprobable ML centers, located at node $L_1/4$ on each of the axons.

5. COMPUTATIONAL RESULTS

Since the derivation in section 3 relies on large deviation theory, we include some computational results in Table 1. These results were derived using the exact combinatorial expression in (3), not the subsequent approximations.

The early examples show how the results scale for different sized regions with the same proportions, and show that the formula for the ML estimator in (9) works exactly, even for very small cases (despite the fact that we used large deviation methods to derive it). The later ones show the results to hold for larger m and more diverse arm lengths. We computed even more varied examples, but space does not permit including them here. However, all of them agree with (9) exactly.

Table 1. Computational results for several examples.

m	Arm Lengths	ML Estimate (Assume Arm 1)
3	4, 2, 2	1
3	20, 10, 10	5
3	200, 100, 100	50
3	40, 40, 20	5 (Arms 1 & 2)
3	200, 200,100	25 (Arms 1 & 2)
3	20, 20, 0	5 (Arms 1 & 2)
3	100, 100, 0	25 (Arms 1& 2)
3	300, 200, 100	75
3	5, 1, 1	2
3	50, 10, 10	20
3	500, 100, 100	200
4	500, 100, 100, 100	200
5	500, 100, 100, 100, 100	200
5	500, 500, 100, 100, 100	150 (Arms 1 & 2)
5	160, 120, 80, 0, 0	55
5	1000, 800, 600, 400, 200	250

6. CONCLUSIONS

A true maximum likelihood estimator is very difficult to produce for general graphs. Accordingly, it is necessary either to restrict the classes of graphs of interest, or to use an approximation to the ML estimator. The rumor centrality measure of [1] is indeed ML for regular trees, and performs well for many cases of non-regular trees. We set out to investigate a relatively simple class of non-regular trees with different conditions (geometric trees with sources of degree 2). We were able to derive an expression for the ML estimator, prove it for large networks (which are our primary interest), and empirically demonstrate it for smaller cases. We also showed that unlike the rumor center, there can be more than two equiprobable ML centers, and that the ML center need not coincide with the distance center.

7. REFERENCES

[1] D. Shah and T. Zaman, "Rumors in a network: Who's the culprit?," *Information Theory, IEEE Transactions on*, vol. 57, no. 8, pp. 5163–5181, Aug 2011.

[2] W. Luo and W. P. Tay, "Finding an infection source under the sis model," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2930–2934.

[3] W. Hu, W. P. Tay, A. Harilal, and G. Xiao, "Network infection source identification under the siri model," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 1712–1716.

[4] K. Zhu and L. Ying, "Information source detection in the sir model: A sample path based approach," in *Information Theory and Applications Workshop (ITA), 2013*, Feb 2013, pp. 1–9.

[5] S. Spencer and R. Srikant, "On the impossibility of localizing multiple rumor sources in a line graph," *SIGMETRICS Perform. Eval. Rev.*, vol. 43, no. 2, pp. 66–68, Sept. 2015.

[6] G. Fanti, P. Kairouz, S. Oh, and P. Viswanath, "Spy vs. spy: Rumor source obfuscation," in *Proceedings of the 15th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, 2015, SIGMETRICS '15, to appear.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York, NY, USA, 1991.