Reaction Paper

Tanvi Namjoshi, Neha Malepati, Stephanie Ginting

1 Paper Overviews

1.1 Paper 1: Containing the spread of contagion on a tree

In [1], Meister and Kleinberg model the contact tracing problem as a race between two processes, an infection process and a tracing process. Few previous papers have attempted to capture how these processes intertwine and run together in the real world. The infection process is initially unaffected by the tracing process, which has yet to begin. An individual is represented by a node v and some parameters (i.e. probability q_v that they meet a new contact sampled from distribution D_q, probability p_v that they infect said contact sampled from distribution D_p). At round $t = 0$, a random node is infected with a probability sampled independently from D_p. This node becomes the root of our infection tree. In each subsequent round, a node v meets a contact u with probability q_v. If v's status is infected, it will then infect u with probability p_v. This process generates a tree.

The tracing process adds another parameter to each node, the active status, which represents a node's ability to generate new contacts. At some time $t > 0$, the tracing process begins and a node is selected as an index case. The tracer then queries one new node at each subsequent round. If querying a node reveals that it is infected, the tracer stabilizes the node (status is inactive), and reveals its children which prevents further infections and identifies those exposed to the infection.

These two processes then run concurrently. An infection is described as contained if the tracer stabilizes all infected nodes. After describing this model, the paper then goes on to establish basic theoretical bounds, as it answers several questions using theory, and then defines policies for querying and describes a specific policy choice that affects the probability of containment.

Meister and Kleinberg conduct computational experiments implementing different policies for querying nodes such as the ascending-time policy which orders nodes by ascending time-of-arrival and the descending-time policy which orders nodes by descending time-of-arrival (if a node v joins the tree in round x, its time of arrival $\tau_v = x$). They then graph the observed probability found through the experiments using time-of-arrival heuristics and compare said heuristics to an optimal policy by training an optimal policy using reinforcement learning, in which contact tracing is formulated as a game where the tracer wins if the infection is contained. This section of the paper also gives insight into the performance of an optimal policy.

The paper's main related course topic is Cascading Behavior in Networks, particularly simple probabilistic models of contagion. This paper is interesting in relation to these topics because previous works modeling contagion spread or contact tracing have yet to delve into the in-depth dynamics of both processes occurring simultaneously, and it builds upon these previous models to study concurrent infection and tracing processes using a probabilistic model as have discussed in class. This concurrency translates well into the real world public health applications for epidemic containment. Overall, we found [1] interesting because it is an application of the cascade model that introduces a force fighting the cascade.

One weakness is that [1] does not delve as much into non-time-based querying policies for tracing as it does for time-based querying policies. It would be interesting to be able to build on the existing work in this paper and produce data visualizations for other approaches that are similar to the figures for time-based querying policies. There were also several assumptions made in constructing the network model, such as the assumption that contacts are formed in a tree-like structure (when in reality cycles may exist). This aided in simplicity but could also be a potential weakness when trying to apply this work in broader areas. Meister and Kleinberg also present a direction for future work, specifically, how to prioritize two nodes with the same time-of-arrival, where one has the higher infection probability and the other has the higher contact probability. It would be interesting to develop an mixed-parameter optimal policy based on comparing all three parameters (time-of-arrival, infection probability, and contact probability).

1.2 Paper 2: Information Spread with Error Correction

In [2], Ben-Eliezer, Mossel, and Sudan discuss a model of the spread of information (and disinformation) in society, representing the flow of communication with a directed spanning tree on a locally finite graph. Formally, the model is as follows:

Let $G = (V, E)$ be an undirected graph, T be a BFS-tree of G that directs communication between nodes, and $r \in V$ be the root vertex of T. Define two probability parameters a and b, where a is the probability a node learns incorrect information from the parent, and b is the probability a node error-corrects to its parent's opinion. Define *p(v)* as the unique parent of node v that has a path to the root, and $f_0(v) \in -1, 1$ as the opinion of v, where an opinion of 1 denotes the correct opinion and -1 is an incorrect opinion. Initially, the root node r is given information and holds the correct opinion (i.e. $f_0(r) = 1$), and all other nodes hold no opinion (denoted by $f_0(v) = \perp$ for all $v \neq r$). At each time step, a node v without an opinion forms a different opinion than its parent *p(v)*'s opinion with probability a or the same opinion with probability $1 - a$. If a node already has an opinion, it may change its existing opinion to match its parent *p(v)*'s opinion with probability b. Note a node can only change its opinion to -1 or 1 once it has formed an opinion, and cannot go back to a no opinion state.

Given this model, the paper analyzes the propagation of information in two "waves". The first wave is the *rumor frontier*, which is the set of nodes from the root that hold any opinion, and the second wave is the *truth frontier*, which is the set of nodes from the root that hold the correct opinion. An analysis of this propagation reveals the slowing growth of the truth frontier in relation to the rumor frontier, where it was found that the gap between the rumor frontier grows linearly in time. Ben-Elizer et al further scrutinize other properties of the model. First, they prove the properties of the model in terms of runs, which are a collection of consecutive nodes along a path that has the same opinion at a given time. Their analysis focused on the structures of a run and their probability of survival. Additionally, they took a look at intermediate frontiers between the truth and rumor frontiers, as well as the stabilization and convergence of opinions.

Like [1], this paper is also related to the Cascading Behavior Model, particularly the branching model discussed in class. We observe the directed spanning tree described in the model is the same as the branching tree, which involves an infection (in this case information) propagating throughout the tree with some probability and the subsequent cascading effect of infection. The model proposed in this paper, however, introduces some layers of complexity and further analysis of branching diffusion models. While the nodes in the branching model operate in two states (infected or not infected) that remain unchanged throughout the cascading process, the nodes in this model have three distinct states (correct, wrong, or no opinion) and, once a node holds an opinion, can modify its opinion with some probability. Additionally, the paper analyzes the branching process as it unfolds over time, once again echoing the model in [1], by taking a look at two frontiers of the movement of information spread and their different rates of growth. These additions of complex dynamics causes the model to more accurately reflect real world networks, specifically information networks, which makes it an interesting structure to study.

The weaknesses of this paper involve the simplistic model of information spread through a tree. The model proposed by Ben-Elizer et al assumes communication through a directed spanning tree T, which assumes information flows in one direction and that there are no cycles. However, in the real world information does not necessarily flow in one direction, and opinions may also be formed from multiple sources. Specifically, the opinion of a parent may also be influenced by the opinion of a child node, which is not considered in the error correction step in the model. One possible research question to pursue would be to apply a similar analysis on communication that could diffuse throughout the entire directed graph G that does not necessarily follow a tree-like structure.

Another weakness and potential avenue for further research is the interventions to correct opinions. In the model, the only method to change an opinion was through error correction and through a direct change of the node's opinion. This involves taking a look at a parent's opinion and updating the opinion to match it, but this process is uniformly applied for all nodes (even if they already hold the correct opinion). Other interventions may focus on slowing down the rumor frontier, such as through targeting nodes with incorrect opinions, injecting the correct opinion in certain nodes (a real-world analogy would be the presence of education or re-education), or attempting to contain the spread of disinformation in some way.

Finally, this paper provides a general theoretical framework for the spread of information in society but does not provide concrete experimental results. While useful, the application of this framework may reveal aspects about different information networks, such as the impact of the medium on the probabilities a and b for the correct opinion formation and opinion updates respectively. Another avenue for further research would be to apply this model in an existing network to see if the results and properties hold. Some existing networks may include social media networks or word-of-mouth methods of communication and information spread.

1.3 Synthesis

We recognize that while [1] is focused on a probabilistic model of contact tracing and [2] is focused on information, the two waves described in [2] are not unlike the concurrent infection and tracing processes described in [1]. If we think of the rumor frontier as an infection that is spread throughout a network and the truth frontier as a tracing process, we are able to connect the cascading models described in these papers.

Both models use trees to model spread, whether that is infection or information, and use some elements of a probabilistic model to define how the spread is modified and information is diffused. This is reflected in the branching network structure of both models, in which an infection, in the case of [1], or information, in the case of [2], is propagated from parent to child node in a directed tree structure. Both models also define two concurrent processes, where there is one process (rumor frontier for [2] and infection process for [1]) that is being countered with a secondary process (truth frontier for [2] and tracing process for [1]). The similarities between these two papers also give us insight into how to formulate optimal querying policies for the tracing process in [1], as the truth frontier in [2] is revealed to be growing slower than the rumor frontier with a gap that grows linearly in time. We can examine this to determine the effectiveness of the truth frontier and further optimize a tracing process.

At a high level, both models delve into the negative examples of the cascading model, where the spread of something throughout a network is actually detrimental to its nodes, and the effects of the spread must be countered. Additionally, few papers simulate concurrent processes affecting a network, which is why it is interesting to be able to analyze two models that focus on the importance of concurrency as it best replicates how something is spread in a real-world network.

The main difference between the models proposed by the two papers is the context of the branching cascading networks, with [1] focused on biological networks, and [2] focused on information networks. Of course, this informs the approach of analysis and properties discussed by both papers. An interesting connection, then, would be to apply the specific network structures (and properties) of one paper to the other. For example, if we consider the model proposed in [1] and apply it to the spread of information, this might suggest an alternative intervention to form the correct opinion. Instead of error correction, taking cues from the biological perspective in [1] could suggest a way of chasing down rumors (i.e. the incorrect opinion) and attempting to contain them by stopping the spread of misinformation (similar to contact tracing). Alternatively, if we consider the model proposed in [2] and apply it to the spread of infection, the error rates of information spread could be similar to a mutation in the virus. Computationally, instead of spreading the virus with a given probability, at each time step the probability could change (similar to an error) to reflect a mutation. These applications could then reveal any deviation in the properties of the models given different contexts and may give insight into the difference between biological and information networks.

2 Our Replication

Our group was particularly interested in the contagion containment paper from Section 1.1, which we will further expand upon in Section 3. Thus, as an initial exploration into what a project could look like, we started by replicating one of the simulations described in [1]. Specifically, we conducted 100 trials for each instance (p, q) for both the ascending-time and descending-time querying policies, where both the infection probability p and contact probability q range from 0 to 1 in increments of .01. For each instance (p, q) we computed the rate of containment and plotted a heat map similar to those in Figure 6 of the original paper.

One significant difference between our replication and the original analysis is that Meister and Kleinberg ran $N = 7.5 * 10^6$ trials per instance for each policy. Due to computational and time restrictions from running these simulations on our computers, we ran significantly fewer trials per instance. Nonetheless, we produced a very similar result. We would be happy to provide the simulation code if desired. The results from this can be found in Figure 1. As demonstrated in the paper, a high p and q value will lead to low levels of containment with both policies.

(a) Ascending-Time Policy Results

(b) Descending-Time Policy Results

Figure 1: The observed probability of containment for ascending-time and descending-time, respectively, from our first round of computational experiments. Although not plotted in the graph, the x-axis is p and the y-axis is q.

The second computational experiment in the paper was to find when one policy dominated over another. Although we did not fully replicate this analysis, we did plot the probability of containment with ascending-time policy - the probability of containment with descending-time policy. Again each (p, q) instance of the problem was run 100 times. We see that it is not true that one policy is always superior. In Figure 2 there are instances when this plot is both negative and positive, indicating that different policies can lead to better outcomes in specific solutions.

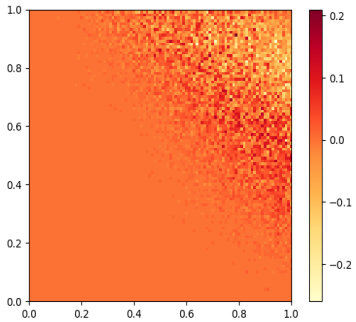

Figure 2: The difference in containment probabilities.

3 Our Proposal

Our plan for the final project is to build on the contagion model in [1]. After reading the paper, we had a few ideas for natural extensions to this paper. The first was to analyze other heuristics for contact tracing beyond time-based heuristics described in [1]. The majority of the analysis described in [1] relies on a fixed constant distribution over all p_vs and q_vs. However, it is reasonable to assume that not everyone transmits the disease to their contacts at the same rate, or generates new contacts at the same rate. The former could come from different rates of transmission between asymptomatic and symptomatic infected people, the latter could come from people who socialize to varying degrees. Thus, one heuristic, in a model where p and q are

not fixed, could be to try choosing who to contact trace based on their given probability for creating a new contact q_v. The question here is: is the chance of containment better if we test a node that is more likely to meet others first? Similarly, we could have each node v sample from two different D_p for p_v, representing the transmission rates of people who do and do not show symptoms of the disease. Then, a new question to ask is: is contagion more likely to stop if we prioritize tracing symptomatic or asymptomatic people?

Another heuristic we discussed was the number of contacts a given node has, but this would require the network structure to be known in advance rather than revealed when a node signals its contacts. As of right now, the model in [1] makes no assumptions about the structure of the network, and only uses the tree structure as a way to model spread. We could test people who have made the most new contacts (have the most children in the tree) first, but this is very similar to the ascending-time query policy as those who arrive at an earlier time have had more chances to make contacts if q is fixed across all nodes.

The extension that our group was particularly interested in exploring is the idea of a different contact tracing process. Meister's and Kleinberg's paper models a scenario with an independent contact tracer. However, in many real-world scenarios, when someone realizes they are infected they contact their network to signal that they are infected. Then, each contact can independently test and contact others who they have interacted with. We can think of Cornell Daily Check or other workplace COVID-era mechanisms as real-life examples of this model. Our project proposal is to analyze a model that simulates this individual-based contact tracing model instead of one with an assigned tracer.

The infection process would be as described in [1]. The tracing process would then run concurrently, and have each node v in the frontier set test itself at every round t. We add an additional variable r_v that represents the likelihood node v takes a test. This models the real-world phenomenon of people not taking tests, either because they refuse or because they are asymptomatic. If the test is positive, the node v signals its contacts (at this point the children would be revealed similar to the tracing process in [1]), and each contact u gets added to the frontier set. Then each node v is represented by the triple (p_v, q_v, r_v).

We can add further complexity by introducing a fourth node parameter s_v, which represents the probability a given node quarantines itself and becomes stable instead of active. The nodes that quarantine would no longer generate new contacts, while those that do not quarantine could infect more people. However, adding this fourth parameter might unnecessarily complicate the model and its analysis. If we think of Cornell's system during the 2020-2021 academic year, those who tested positive were mandated to quarantine. Thus, we think the model with the parameters p, q, r will be sufficient. The contact tracing will start at time $t = k$ as described in [1], thus allowing for a few initial rounds of infection to be inhibited by contact tracing.

Although we will attempt to prove a theoretical result, it is likely that the defined model is too complex to analyze rigorously. Thus we plan to focus our project on showing results through simulation. As in the paper we read, we will initially fix the model to have a constant distribution over the node-specific variables. This is in line with Meister's and Kleinberg's analysis that we replicated where we analyzed the probability of contagion over all possible instances of the contact tracing problem where $p_v = p$ and $q_v = q$ for all v. Similarly, we shall start by analyzing the probability of contagion over instances of the contact tracing problem defined by (p, q, r) where $p_v = p$, $q_v = q$, and $r_v = r$ for all v.

Now, rather than a problem of choosing the best heuristic-based policy for the contact tracer seen in [1], our proposal aims to determine the individual testing rate that is best for containing the spread of the infection. This can inform institution-level policy of mandated testing during pandemics. Formally, for a given p and q, what testing probability r maximizes the probability of containment?

We recognize that a (p, q, r) model is much heavier in computation than a (p, q) model, but we believe our initial replication is promising and that we will be able to add in the additional parameter r_v in our final project to simulate this proposal.

References

[1] M. Meister and J. Kleinberg, "Containing the spread of a contagion on a tree," 2022.

[2] O. Ben-Eliezer, E. Mossel, and M. Sudan, "Information spread with error correction," 2021.