Locating Patient-Zero In a Network

Pablo A. Curiel

Department of Computer Science, California State University, Chico, USA

Abstract

This paper aims to provide an introduction to spread in networks with a focus on identifying the source of a spread, commonly referred to as patient-zero when related to disease spread. A literature review is presented to provide a high-level overview of past methods used to locate the source of a spread in a network. Experimental methods of identifying patient-zero are described, and their results are shown.

1. Introduction

The modeling of spread using networks is an increasingly popular area of networks research. This is largely due to how networks capture important details necessary for spread, as well as the many real-world applications associated to spread. While there is extensive research related to modeling the future state of a spread, literature related to modeling the past state of a spread is scarce. This lack of backwards-time spread research has been attributed to an increased difficulty and computational cost. However, studying the past-states of a spread can have a broad impact on many real-world applications.

Some real-world examples of spread include the spread of: infectious diseases in a physical-contact network, rumors, information, or memes in a social network, malware attacks or digital viruses in a computer-systems network, and rolling blackouts in a power grid network.

In recent years, determining the source of a spread (often referred to as **patient-zero**) has become a popular area of research in network science. This is largely due to how networks capture important features necessary for spread, in contrast to some traditional models of spread. Determining the source of a spread can have many critical real-world applications, such as contact tracing, optimizing the allocation of immunization resources, mitigating the spread of rumors or misinformation, and defending against malware attacks. This paper focuses on popular methods of identifying the source of a spread.

2. Spread overview

As previously mentioned, spread in networks have many real-world applications.

While modeling spread using networks is a very useful tool, many newer models are based on some traditional methods of modeling spread. The most popular being the continuous-time models of spread from epidemiology, often referred to as SIR-models [1]. Due to the continous-time nature of these models, they are most commonly described using a system of differential equations.

SIR-models are thought of as compartmentalized models, as they assume the individuals in a population are separated into three compartments, or states. These states are: **S**usceptible, **I**nfected, and **R**ecovered (or removed). In these models, an individual's state follows a specific ordering: $S \rightarrow I \rightarrow R$. Once in a state, an individual can not go back to a previous state.

A more important, and unrealistic, assumption of these continuous-time models is that of homogeneous mixing, or well-mixed. This means that every individual has equal probability to either infect or become infected by any other other individual within the population.

Networks have allowed the modeling of spread to become more realistic by not requiring the assumption of homogenous mixing [2]. In network models of spread, individuals can only get infected or infect nodes they are directly connected to. This helps provide a more realistic scenario of spread.

Network models of spread often rely analogs of the continuous-time models. These models include the SIR-model described above, as well as variations of the SIR-model. Some popular variations include the SI-model ($S \rightarrow I$) and the SEIR-model ($S \rightarrow E \rightarrow I \rightarrow R$), where E represents an exposed state. Also, an important characteristic of spread is known as the *basic reproductive number* $R_0 = \frac{\beta\langle k\rangle}{\mu}$, where β is a rate of infection, μ is a rate recovery, and $\langle k\rangle$ is the average degree of the network. The basic reproductive number is defined as the number of secondary infections produced from a single infected individual in a fully susceptible population. Notably, when $R_0 > 1$, we can expect the spread to increase throughout the network, and when $R_0 < 1$, we can expect the spread

to die out.

3. Related work

This section provides a high-level overview of four prominent papers [3,4,5,6] related to identifying the source of a spread in networks. The papers are presented in chronological order based on their publish date.

3.1. Shah and Zaman (2011)

To my knowledge, the first paper dedicated to locating the source of a spread in a network was by Shah and Zaman (2011). They model the spread using the SI-model described above through the lens of a rumor spreading through a social network. In their model, infected nodes describe those that have heard the rumor.

Their primary contribution is the creation of a source estimator that they refer to as "rumor centrality". They prove that on tree graphs rumor centrality acts as a maximum likelihood estimator for inferring the source node and is equivalent to distance (or closeness) centrality.

Shah and Zaman apply their estimator to a variety of network structures, including small-world networks, scale-free networks, and real-world networks. They measure the accuracy of their method using distance (number of hops) from the true source node. Their method is often within five hops of the true source for most network structures, which could be useful on large graphs. However, their best accuracy for correctly identifying exact the source node is 16% on small-world networks.

3.2. Pinto, Thiran, and Vetterli (2012)

A year after Shah and Zaman's paper, Pinto, Thiran, and Vetterli created a new method of locating the source of a spread that made use of so-called "observer" nodes. Their model of spread is similar to the SI-model in the sense that they have "informed" (infected) nodes and "uninformed" (susceptible) nodes.

The observer nodes track at what time they become infected and which node infected them. They state some common difficulties with this method include choosing the placement and density of the observer nodes in the network. Also, their method has a complexity of $O(N)$ for arbitrary trees and $O(N^3)$ for arbitrary graphs.

3.3. Lokhov et al. (2014)

This paper Lokhov et al. was important in the literature for identifying patient-zero, as it showed increased accuracy in comparison to past methods. Also, it was proven to be exact on tree graphs. This method makes use of an algorithm called Dynamic Message Passing, or DMP. While their method showed increase accuracy, it was very computationally expensive. This method had a complexity of $O(tN^2\langle k\rangle)$ for t time-steps.

3.4. Shah et al. (2020)

The paper by Shah et al. is unpublished, but it is arguably the most interesting. They make use of a variety of graph neural network (GNN) architectures to identify the source of spread. They argue that their method produces increased accuracy with increased efficiency in comparison to past methods, such as DMP. Interestingly, they claim their method is model agnostic, which means the GNN does not to take into account the model of spread or its parameters. Their method is said to have a complexity of $O(N^2 logN)$, and their results show accuracies of $\geq\approx 20\%$, depending on the network structure and spread-time simulated.

4. Methods

My methods of inferring the source node make use of various centrality measures, clustering coefficients, and random guessing. These measures are calculated for all Recovered nodes, or all possible patient-zero nodes. The centrality measures were calculated over either the entire graph (global) or sub-graph of Recovered nodes (local). These measures were then normalized over their sum, and in some cases aggregated using an average of the normalized values. The centrality measures calculated were degree centrality, harmonic centrality, eigenvector centrality, and closeness (or distance) centrality. All measures were calculated using built-in functions in NetworkX.

The model of spread simulated on the network is similar to the SIR-model described above. At time $t = 0$, a node is chosen at random to be the sources node, or patient-zero, and labeled as Infected. Then, at each following time step, all infected nodes have equal probability of infecting each of their neighbors. This probability is known as the spread rate $\lambda = \frac{R_0}{\langle k\rangle}$. The spread is simulated for $t \in [0, 15]$ time-steps to show how the accuracy of determining patient-zero depends on time. In the majority of experiments, the basic reproductive number $R_0 = 1.5$, which implies the disease should spread throughout the network since $R_0 > 1$. In one experiment, the measures are calculated over varying R_0-values in the range $[0, 4]$ in increments of 0.5. This is done to show how locating the source of spread is dependent on R_0.

In all experiments, the spread is simulated for 100 iterations, and the accuracy of correctly identifying patient-zero is calculated as the average of correct identifications per 100 iterations. The graph types experimented on include: Barabási-Albert scale-free graphs ($m = 3$), Erdős-Rényi random graphs ($p = 0.2$), and a real-world social network

of Russian troll-bot accounts on Twitter ($N = 1245, |E| = 2,974$. All graphs, except the real-world network, have $N = 1,000$ nodes.

4.1. Various centrality measures (Global)

In this method, the various centrality measures described above were calculated for all nodes in the pool of Recovered nodes. This is referred as a "global" measure because it describes a node's centrality over the entire network. The centralities of the Recovered nodes are then normalized and aggregated. Patient-zero is inferred to be the maximum of the resulting values.

4.2. Various Centrality measures (Local)

This method of inferring the source node follows a similar structure to the one described above. However, a Recovered node's centrality is not calculated over the entire network. Instead, the centrality values are calculated over the sub-graph of Recovered nodes.

4.3. Clustering coefficients

This method calculates the clustering coefficients of all nodes in the network over the entire network. The values are then normalized over the sum of clustering coefficinets. Patient-zero is inferred to be the maximum of these values.

4.4. Centrality and clustering coefficients

This method is a combination of the global centrality measures and clustering coefficients. The centrality measures and clustering coefficients are normalized and aggregated together. Patient-zero is inferred to be the maximum of these aggregated values.

4.5. Random guessing

This method is, as expected, the least accurate method of identifying patient-zero. A node is chosen at random from the pool of Recovered nodes. This node is then inferred to be the patient-zero. Due to its extremely low accuracy, this method was almost omitted from this paper entirely.

5. Results and discussion

This section displays figures generated by the results of each method. The results of the experiments on Barabási-Albert scale-free networks are shown in Figure 1. The results of the experiments on Erdős-Rényi random networks are shown in Figure 2. The results of the experiments on the real-world Twitter Russian trolls network are shown in Figure 3. Notably, the average accuracy of each method, except random guessing, are all in range of $30 - 50\%$. In all experiments, the method of random guessing displayed mostly 0% accuracy, with a single spike at $t = 11$ in Figure 3e. This spike is likely due a single random guess correctly identifying patient-zero.

As an aside, the figure placement in this paper is very poor. The template used for this two-column structure kept placing the figures at the top of the next page, and I was unable to find a solution to this in time.

6. Future work

There is much work that can be done to extend on the methods of this paper. For example, the centrality measures were all averaged giving them an equal contribution towards identifying the source. This can be improved by weighing the measures differently based on how each contributes to identifying the source. Also, most experiments only simulated the spread for $t \in [0, 15]$. It would be beneficial to simulate the spread for longer time to see how this affects the accuracy of identifying the source. Another metric that could have been used on the methods is the distance (in number of hops) from the inferred source to the true source.

Disregarding its algorithmic complexity, the method of Dynamic Message Passing by Lokhov et al. still has some of the highest accuracies of current methods found in the literature. It would be interesting to expand on this method, by either increasing its accuracy or decreasing its computational cost.

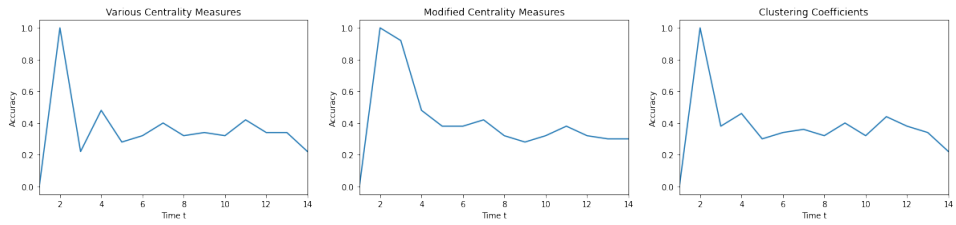
Also, the work presented in the paper by Shah et al. seems remarkably powerful for identifying the source of a spread. It would be interesting to expand on this method, specifically their section about future work. However, this paper lacks important details related to their method that would be helpful for expanding on it. These include details about their GNN architecture, input, and implementation.

7. Conclusion

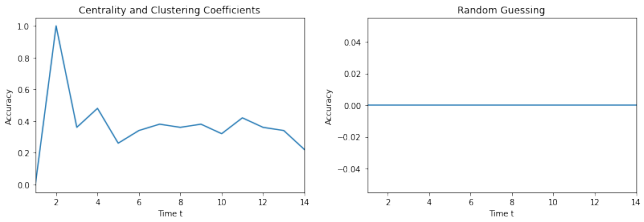
The modeling of spread in networks has many important real-world applications. Perhaps the most important being identifying the source of a spread, especially in relation to a disease spreading through a population. However, this problem is often difficult and computationally expensive. While current methods of identifying the source are improving, there is much more work to be done in this area.

Acknowledgments

Thank you to Dr. Richard Carter Tillquist for his frequent willingness to meet and provide guidance.

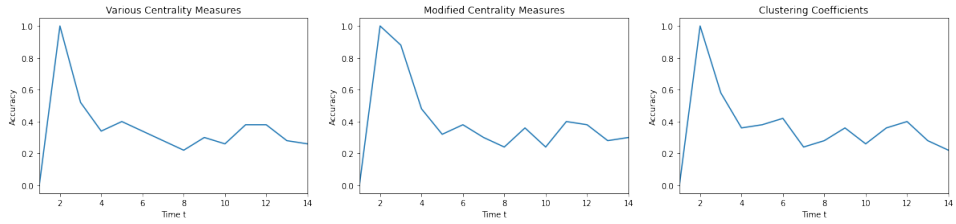


(a) Centrality (global) (b) Centrality (local) (c) Clustering coefficients

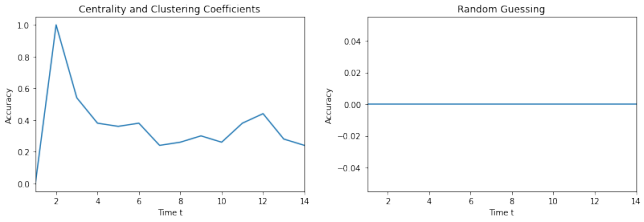


(d) Centrality and clustering (e) Random guessing

Figure 1. Barabási-Albert scale-free network: Average accuracy of each method over $t = 15$ time-steps. Accuracies in **a.) - d.)** range between $30 - 40\%$, while **e.)** was $0\% \; \forall$ time t.



(a) Centrality (global) (b) Centrality (local) (c) Clustering coefficients



(d) Centrality and clustering (e) Random guessing

Figure 2. Erdős-Rényi random network: Average accuracy of each method over $t = 15$ time-steps. Accuracies in **a.) - d.)** range between $30 - 40\%$, while **e.)** was $0\% \; \forall$ time t.

References

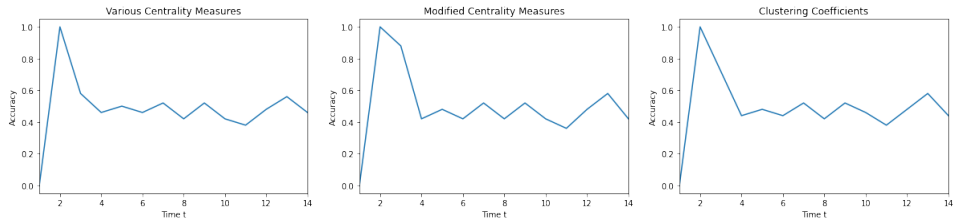
[1] J.D. Murray. Mathematical Biology I: An Introduction. Interdisciplinary Applied Mathematics Springer, vol. 17, 2002.

[2] Albert-László Barabási. Network science. Cambridge University Press, 2016.

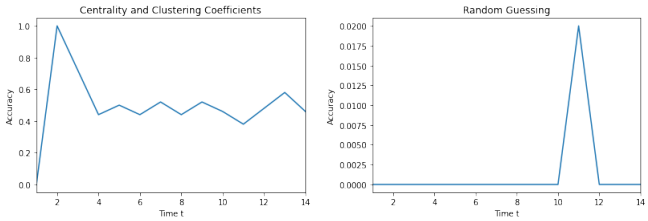
[3] Devavrat Shah, Tauhid Zaman. Rumors in a Network: Who's the Culprit? IEEE Transcactions on Information Theory, vol. 57, no. 8, August 2011. https://doi.org/10.48550/arxiv.0909.4370

[4] Pedro C. Pinto, Patrick Thiran, Martin Vetterli. Locating the Source of Diffusion in Large-Scale Networks. Phys. Rev. Lett., vol. 109, no. 6, August 2012. 10.1103/PhysRevLett.109.068702

[5] Andrey Y. Lokhov, Marc Mézard, Hiroki Ohta, Lenka Zdeborová. Inferring the Origin of an Epidemic with a Dynamic Message-Passing Algorithm. Phys. Rev. E, vol. 90, no. 1, July 2014. 10.1103/PhysRevE.90.012801

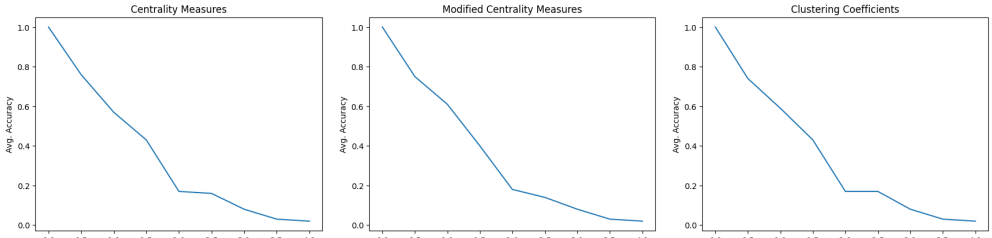


(a) Centrality (global) (b) Centrality (local) (c) Clustering coefficients

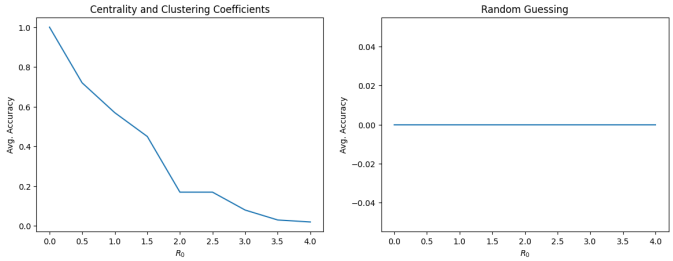


(d) Centrality and clustering (e) Random guessing

Figure 3. Real-world Twitter Russian trolls network: Average accuracy of each method over $t = 15$ time-steps. Accuracies in **a.) - d.)** are higher than on randomly generated networks, while **e.)** shows a random spike at $t = 11$.



(a) Centrality (global) (b) Centrality (local) (c) Clustering coefficients



(d) Centrality and clustering (e) Random guessing

Figure 4. Barabási-Albert scale-free network: Average accuracy of each method for varying R_0 values in range $[0, 4]$.

[6] Chintan Shah, Nima Dehmamy, Nicola Perra, Matteo Chinazzi, Albert-László Barabási, Alessandro Vespignani, Rose Yu. Finding Patient Zero: Learning Contagion Source with Graph Neural Networks. June 2020. https://doi.org/10.48550/arXiv.2006.11913