Identifying Asymptomatic Nodes in Network Epidemics using Betweenness Centrality

Conrado C. Pinto[1], **Daniel R. Figueiredo**[1]

[1]Programa de Engenharia de Sistemas e Computação (PESC)
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brazil

conrado@cos.ufrj.br, daniel@cos.ufrj.br

Abstract. *Epidemics of certain viruses in a population can have major impact effects, as is the case in the recent global pandemic caused by the COVID-19 virus. Identifying infected individuals during the course of an epidemic is extremely important for measuring spread and designing more effective control measures. However, in some epidemics infected individuals do not exhibit clear symptoms despite being infected and contributing to the contagion of others (called asymptomatic). This work addresses the problem of identifying asymptomatic individuals in network epidemics based on the observation of infected (symptomatic) individuals. The main contribution of this work is the evaluation of different centrality measures to identify asymptomatic individuals when a fraction of the infected nodes in a network epidemic is observed at a given moment in time. In particular, a variation of the betweenness centrality measure is proposed in this work. An evaluation using different network models and different asymptomatic rates shows that the proposed centrality measure outperforms other centrality measures in many scenarios. Furthermore, the performance of centrality measures increases as the fraction of asymptomatic decreases, showing an interesting trade-off.*

1. Introduction

Large-scale epidemics of novel viruses in the human population are often generated by the contact of humans with uninhabited and isolated environments [Quammen 2012]. Such epidemics can spread fast through the population, triggering serious health and economic crises, such as the recent COVID-19 pandemic. To contain the spread of such epidemics, it is imperative to accurately identify infected individuals, either for quarantine or treatment. Moreover, public policies are also often crafted based on the observation and number (fraction) of infected individuals.

However, in many epidemics - including COVID-19 - it is customary that a fraction of the infections do not exhibit any symptoms and, without testing, such asymptomatic individuals remain hidden [Inui et al. 2020]. Despite not showing symptoms, such individuals continue to be a health risk, as they continue to infect other individuals, and their invisible epidemic state only worsens the situation [Arons et al. 2020]. Thus, identifying asymptomatic individuals is fundamental to building more effective epidemic containment and public policies.

One approach to identifying asymptomatic individuals is through mass testing, where individuals without symptoms are tested regularly for the virus. However, such

policies can be expensive [Seguí et al. 2021], as tests must be performed continuously across the population. Therefore, it is important to develop mechanisms that can accurately infer asymptomatic individuals without testing.

Network epidemic models offer a powerful framework to model social contacts and epidemic processes in order to design and evaluate mechanisms for identifying asymptomatic individuals from partial observations [Keeling and Eames 2005]. In particular, the classic SI (Susceptible-Infected) epidemic model in networks considers the propagation of the epidemic through network nodes where susceptible nodes (S) can get infected by neighboring infected nodes (I). In this model, nodes are either susceptible (S) or infected (I), and once infected the node remains infected.

Using this classic model, this work defines a simple and random observation process that reveals a random set of the infected population. Such individuals have symptoms that can be directly observed. Unobserved infected individuals correspond to asymptomatic cases.

However, there are no differences between susceptible nodes and asymptomatic nodes. Therefore, the objective of this work is to develop an algorithm capable of identifying asymptomatic nodes using the network structure and the set of observed infected nodes. The main idea is to use centrality measures for nodes that have not been observed as infected where a high-ranked node should correspond to asymptomatic nodes. The main proposal is an adaptation of the betweenness centrality measure, which considers only infected vertices as final vertices (more details in Section 3). Intuitively, the metric explores the properties of the random infection process, which generally spreads over shortest paths.

The proposed method is evaluated through simulation of the epidemic process on different network models and different scenarios (such as the asymptomatic probability). The results are also compared with other centrality measures, such as the vertex degree and the fraction of neighbors observed as infected. The true positive and false negative rates are used to evaluate and compare the different centrality measures. The results indicate that the proposed method is more efficient than the others in both metrics and in many different scenarios, showing that the identification of asymptomatic individuals is possible and effective.

This paper is organized as follows: Section 2 presents the fundamental concepts of network epidemics and centrality measures, and also some related works; Section 3 describes the method used to generate asymptomatic individuals and the proposed methods to infer them; Section 4 describes the evaluation methodology and presents the results; finally, Section 5 offers a discussion of what we have done.

2. Fundamentals and related works

2.1. Networks

Traditionally, there are several methods to mathematically describe real-world systems. Translating natural phenomena into equations is essential for us to understand our surroundings. But there is one method in particular that allows us to capture the intrinsic and hidden relationships between individual objects: networks. Deeply and increasingly studied and developed since the second half of the 90s, there are network models that are

very well described mathematically. Some of them will be used in this work and therefore deserve a presentation.

A classic model for generating random graphs developed by Erdős and Rényi in the 1950s is also used in the study of complex networks. Given a fixed number of vertices n, each possible edge between pairs of vertices occurs with probability p. Its structural properties have already been deeply analyzed mathematically based on its parameters. A shortcoming of this model when we intend to represent real networks is that its degree distribution follows a binomial distribution while many real systems have a heavy-tailed degree distribution.

Barabási and Albert's revolutionary model allowed the construction of a growing network through the gradual introduction of vertices [Albert and Barabási 2002]. Extremely important to explain the appearance of hubs (highly connected nodes) and the lack of normality between the vertices' degrees, it has a power-law degree distribution, which is heavy-tailed. The model starts with a small network and at each instant of time, a new vertex is added with m edges (m is a model parameter) that will connect to the previously existing vertices. But this choice of vertices to connect is not uniform, it follows a preferential attachment that prioritizes vertices of higher degree and that's how it allows the existence of hubs.

With the aim of studying short distances between two individuals and sparsity in real complex systems - a phenomenon we call small-world - Watts and Strogatz developed a network model that manages to harmonize topological aspects that seem antagonistic: high clustering, short paths, and high sparsity [Watts and Strogatz 1998]. We start with a regular lattice of n vertices in which each vertex has edges to its neighbors at a distance k or less, and each of these edges has its endpoints repositioned uniformly with probability p.

2.2. Network centrality

Different notions of importance can be assigned to nodes as a function of the network structure, and this is known as node centrality. The degree centrality is arguably the most simple metric: the centrality of a vertex is given by the number of neighbors it has. Thus, $C_v = d_v$ for a node $v \in V$. This metric can be normalized by the maximum degree of any node, and thus $C_v = \frac{d_v}{n-1}$ yields a centrality metric that can be used to compare nodes in different networks. Note that C_v captures a local property of v with respect to the network structure (namely, its relative degree).

The betweenness centrality [Freeman 1977] seeks to assign greater importance to vertices that are in the shortest paths between pairs of vertices. For each pair of vertices of a connected component of the network, there is one or more shortest paths between them. The betweenness of a vertex reflects the number of times this vertex is on the shortest path of all pairs of nodes. Vertices with a high betweenness have a greater influence on the flow of information through the network, as many shortest paths traverse them. Mathematically, betweenness is defined as:

$$C_v = \sum_{s,t \in V; s,t \neq v} \frac{\sigma(s,t|v)}{\sigma(s,t)}, \tag{1}$$

where $\sigma(s,t)$ is the number of shortest paths between vertices s e t, and $\sigma(s,t|v)$ is the number of such shortest paths that pass through v.

2.3. Network epidemics

In recent years, networks have been used to model the spread of contagious diseases. Taking the vertices as a representation of individuals, the network's edges can model different forms of contact, which allows us to use this structure for the most diverse diseases.

Compartmental epidemic models are utilized jointly. Each individual is assigned to a single epidemic state: Susceptible, Infected, Recovered, Exposed, etc., and as the epidemic progresses, individuals transition through classes. Here, we focus on the SI model, composed only of the Susceptible (S) and Infected (I) compartments, and in which the only transition occurs from S to I, that is, healthy individuals can be infected, and once infected they no longer transition through epidemic states.

Consider a discrete-time SI model with $t = 0, 1, \ldots$ in a connected and undirected network (graph) $G = (V, E)$, where V and E denote the vertices set and the edges set. Let $S(t)$ and $I(t)$ be the set of susceptible vertices and infected vertices at time t, respectively. Note that $S(t) \cap I(t) = \emptyset$ and $S(t) \cup I(t) = V$ because every vertex has exactly one epidemic state at each time step.

We will consider a probabilistic epidemic model, with parameter $\beta > 0$ that denotes the infection probability through an edge, that is, the probability that an infected vertex infects a susceptible neighbor. The contagion event is independent for each edge and each time step. More precisely, the epidemic evolves as follows: for each vertex $u \in S(t)$, the probability that u belongs to $I(t+1)$ is given by

$$P\left[u \in I(t+1) | u \in S(t)\right] = 1 - (1-\beta)^{r_u(t)}, \tag{2}$$

where $r_u(t)$ is the number of infected neighbors of u at time t,

$$r_u(t) = \sum_{v \in N_u} \mathbb{1}(v \in I(t)), \tag{3}$$

where N_u is the set of neighbors of u and $\mathbb{1}(\cdot)$ is the indicator function. Notice that the probability in Eq. 2 is the complement to the event of vertex u not being infected in time t, that is, none of its neighbors infect it.

Notice that all the vertices that have at least one infected neighbor also have a positive probability of being infected at each time step. In that way, eventually, the epidemic will spread through the network and all vertices become infected for some sufficiently (but finite) large t, considering that the network is connected.

We will assume that the epidemic starts with only one infected vertex, chosen randomly and uniformly among all network vertices. That is, $I(0) = \{v\}$ with v uniformly chosen from V, and $S(0) = V \setminus \{v\}$.

While the techniques proposed in this work focuses on the SI epidemic model, a brief discussion of other epidemic models is presented in Section 5.

2.4. Related works

Networks have been widely used to trace infection routes and contacts in epidemic scenarios as well as in finding epidemic sources, and it is a main challenge to propose social interventions through the analysis of network epidemics [Pellis et al. 2015]. Many works have already taken advantage of vertex centrality measures to infer the origin of dissemination in networks. [Comin and da Fontoura Costa 2011] uses degree, betweenness, closeness, and eigenvector centrality for this purpose. [Shah and Zaman 2011] developed a 'rumor centrality' to find the initial vertex of an SI epidemic.

[Huang et al. 2023] utilize dynamical equations to model transitions between different epidemic compartments and combines it with contact network topology and data on the observed infection history to infer asymptomatic hidden nodes. The state transition of each node is modeled as a Markov process in which the probability transitions are determined by the infection status of known nodes at previous time steps. Infection histories of infected nodes are known, helping in tracking the asymptomatic nodes.

[Chen et al. 2023] proposes a prediction algorithm based on a Machine Learning algorithm: TrustRank. This algorithm is a semi-automatic classification method that ranks objects based on their trust level, and thus it is a good framework to infer asymptomatic individuals. They use the information obtained from machine learning as a key metric to rank the nodes and determine how much propagation a node has. The prediction results are then used to propose isolation measures.

The Degree and Contact methods (detailed in Section 3) are used as comparative methods in works with the same proposal as ours: finding asymptomatic individuals in a network epidemic. [Zhang et al. 2023] use Bayesian inference methods to infer unobserved cases of infection, with backward temporal propagation processes and cross-ensemble covariability. An important difference in the process is that the techniques are used by them at each time step, while here the entire inference method is carried out only at the end of the epidemic dynamics.

3. Asymptomatic individuals identification

Consider an epidemic unfolding on a network as described by the model presented in Section 2.3. Consider also that the epidemic is observed only at a time step t. For this time step, the epidemic state of each vertex is revealed to the observer. However, asymptomatic vertices will be observed as susceptible (healthy).

We will consider the following model to define which vertices are asymptomatic. At time step t, each infected vertex belonging to the set $I(t)$ will be asymptomatic with probability p, independently from the other vertices. Notice that p determines the fraction of asymptomatic vertices in the population.

In this way, we can define the set of observed vertices $O(t)$ in the following way: for each vertex $v \in I(t)$, $v \in O(t)$ with probability $1-p$. Finally, the set of asymptomatic vertices is defined by $A(t) = I(t) \setminus O(t)$. Notice that the set $V \setminus O(t)$ are all the vertices observed as susceptible.

Figure 1 illustrates a network epidemic at time t and the vertices observed as infected at this time step. Notice that the asymptomatic vertices (those not observed as

infected) are indistinguishable from susceptible vertices. The aim of this work is exactly to identify the asymptomatic vertices from this observation, that is, to recover the set $A(t)$.

In what follows, we will define centrality measures to rank network vertices that are not observed as infected with the aim of revealing the asymptomatic ones. Intuitively, vertices that appear at the top of the ranking should have higher chances of being asymptomatic.

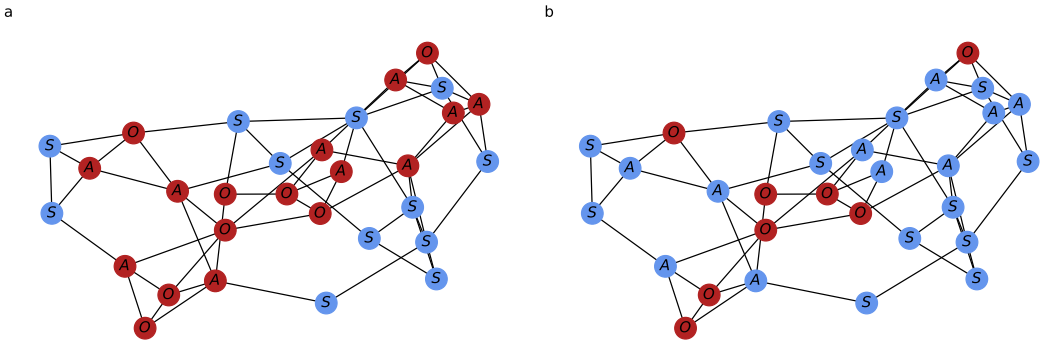

Figure 1. Network diagrams with true nodes' labels. Figure (a) shows in blue the susceptible nodes and in red the infected ones. In Figure (b) we have this same color classification but from an observation point of view, so S and A are indistinguishable.

3.1. Degree-based method

The higher the degree of a vertex, the higher the chance of having an infected neighbor and, thus, of becoming infected. This centrality measure does not take into account the epidemic state of any vertex in the network and assigns to the vertices their degree centrality.

$$c_d(v) = |N_v|, \tag{4}$$

N_v is the set of neighbor nodes of v.

3.2. Contact method

This centrality measure modifies the degree centrality to add information on vertices observed as infected. The importance of a vertex is no longer given only by its degree; but by the fraction of observed infected neighbors, that is:

$$c_c(v,t) = \frac{|N_v \cap O(t)|}{|N_v|}. \tag{5}$$

Intuitively, the more infected neighbors a vertex has, the greater the chance of this vertex also being infected.

3.3. Betweenness method

The change we make here regarding betweenness centrality seeks to focus the analysis on the shortest paths between pairs of vertices that have been observed as infected. The

centrality of a vertex v (that has not been observed as infected) will be the fraction of shortest paths between vertices belonging to $O(t)$ that passes through v.

$$c_b(v,t) = \sum_{x,y \in O(t)} \frac{\sigma(x,y|v)}{\sigma(x,y)}, \tag{6}$$

where $\sigma(x,y)$ is the number of shortest paths between vertices x and y, and $\sigma(x,y|v)$ is the number of these paths that pass through v.

The above betweenness centrality can be computed by running a breadth-first search (BFS) for each node observed as infected. Assuming a connected graph, the BFS has running time complexity $\Theta(|E|)$. Assuming a total of $|O(t)|$ observed infected nodes at time t, the total running time of this centrality metric is $\Theta(|O(t)||E|)$.

Given that the infection process occurs through the network edges, a vertex that is midway between several pairs of infected vertices will have a greater chance to be infected than another vertex that belongs to just a few of these shortest paths, intuitively. This method aims to take advantage of the network's structural information (shortest paths) and the infection process.

3.4. Estimating the asymptomatic individuals

Using the centrality measures, we will define the set of vertices considered asymptomatic. In particular, vertices not observed as infected will be ranked decreasingly according to each centrality measure defined above. We will consider the first k vertices from the ranking as being asymptomatic, defined as

$$\hat{A}_x(k,t) = \{v \mid v \in \text{ top-}k \text{ nodes in } x \text{ centrality ranking given } O(t)\}, \tag{7}$$

where x is the degree, contact, or betweenness centrality. Notice that k is a parameter from the model to identify asymptomatic individuals. Intuitively, the higher the value of k, the greater the number of asymptomatic vertices at the set, however, the greater the number of susceptible vertices as well. In this way, there is a trade-off to the value of k, as discussed hereafter.

4. Metrics and results

4.1. TPR, FPR, and AUC

Traditional classification metrics will be considered to evaluate the accuracy of different centrality measures in identifying asymptomatic vertices. The true positive rate (TPR) represents the fraction of correct positives returned by the centrality measure, given by

$$TPR_x(k,t) = \frac{|\hat{A}_x(k,t) \cap A(t)|}{|A(t)|}, \tag{8}$$

where x is the degree, contact, or betweenness measure. The false positive rate (FPR) represents the fraction of incorrect positives returned by the centrality measure, given by

$$FPR_x(k,t) = \frac{|\hat{A}_x(k,t) \cap S(t)|}{|S(t)|}. \tag{9}$$

For each time step t and parameter value k, we have a set of true positive nodes (those belonging to $\hat{A}_x(k,t)$ that are asymptomatic) and false positive cases (those belonging to $\hat{A}_x(k,t)$ that are susceptible) and thus the above metrics can be directly calculated.

Given a fixed t, for each value of k, we have a value for TPR and another one for FPR. By increasing k, we generate a curve with these pairs of values (that increase with respect to k) that characterizes the performance of the centrality measure. Note that when k assumes its largest possible value, namely $k = |S(t)| + |A(t)|$ both TPR and FPR are equal to one.

In order to measure and compare different centralities without using the value of k, we compute the area under this curve, called AUC. The closer this value is to 1, the better the efficiency of the centrality in identifying correctly the asymptomatic individuals.

4.2. Evaluation methodology

With the aim of evaluating the centrality measures in networks with different structures, we use different network models with the following parameters to generate networks with the same expected degree:

- Watts-Strogatz: $k = 8$ e $p = 0.3$;
- Barabási-Albert: $m = 4$;
- Erdős-Rényi: $p = 8/3000$.

The three networks have $n = 3000$ vertices and all have average degree equals to 8 (see parameters details in Sec 2.1).

To start the network epidemic, a single node is chosen uniformly at random to be infected. Then, the epidemic process is simulated on the generated network by randomly infecting neighboring nodes at each time step according to the SI epidemic model presented in Section 2.3. The simulation ends when the fraction of infected vertices reaches 20% of all vertices, at time step t_{20}, or 60%, at time step t_{60} (different simulations for each case). At this time, the infected vertices are chosen to be observed, with the following observation rates, p: 5%, 10%, 25%, 50%, 75%, 90%. The infected vertices not chosen by the observation process are the asymptomatic nodes.

For each network model and each observation rate, we rank the vertices that were not observed as infected (i.e., asymptomatic and susceptible) using the three centrality measures. Using the top-k ranking vertices as asymptomatic, we evaluate the performance of each centrality measure with the AUC metric.

Since both the network model and the epidemic model are random, a total of ten independent simulations have been performed for each network and each observation time. The performance metrics have been calculated for each simulation, but in what follows the mean and standard deviation over the ten runs are presented.

4.3. Results

Figure 2 shows the FPRxTPR curves for the three methods in different scenarios. In these graphs, the plotted diagonal represents what the performance of a random inference model would be, whose AUC is equal to 0.5.

In a network generated by the Watts-Strogatz (WS) model, for any observation rate in both the scenario with 20% infected nodes and the scenario with 60% infected

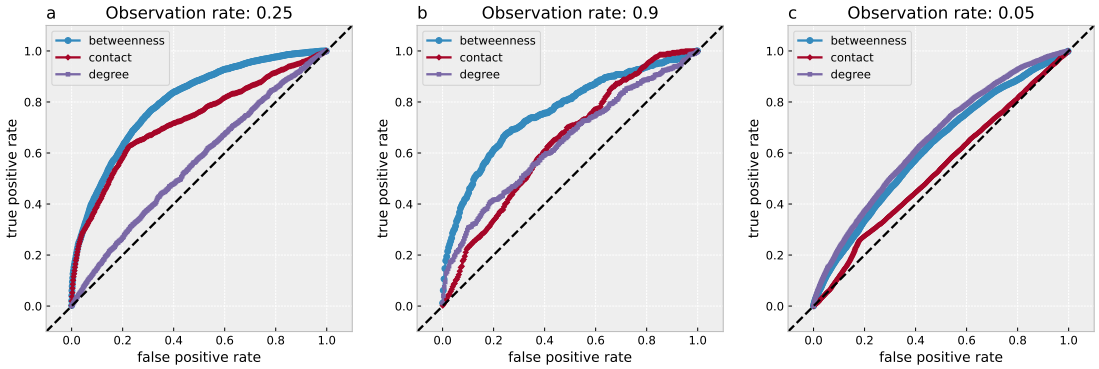

Figure 2. (a) WS network with 20% of infected nodes; (b) BA network with 20% of infected nodes; (c) ER network with 60% of infected nodes.

nodes, the Degree method presents very little difference in performance when compared to a random model that cannot distinguish between positive cases and negative cases. This can be checked by the AUC values very close to 0.5 in Table 1. The Degree method was unable to achieve AUC metrics greater than 0.6 in any configuration for this network, which indicates a low-quality performance.

When we have 20% of infected individuals, the Betweenness method already performs acceptably well from an observation rate of 10% (AUC greater than 0.7) and presents excellent results with AUC greater than 0.8 for observation rates greater or equal to 50%. Also for 50% or higher observations, the Contact method has a notable performance improvement and can be competitive with Betweenness, even slightly outperforming the latter.

Stop at t_{20}						
Obs.	5%	10%	25%	50%	75%	90%
Betw.	$\mathbf{0.63 \pm 0.02}$	$\mathbf{0.71 \pm 0.02}$	$\mathbf{0.78 \pm 0.02}$	$\mathbf{0.83 \pm 0.01}$	$\mathbf{0.85 \pm 0.01}$	0.86 ± 0.02
Cont.	0.56 ± 0.03	0.62 ± 0.02	0.73 ± 0.02	0.8 ± 0.02	$\mathbf{0.85 \pm 0.01}$	$\mathbf{0.88 \pm 0.03}$
Degr.	0.57 ± 0.01	0.56 ± 0.01	0.55 ± 0.02	0.56 ± 0.02	0.56 ± 0.02	0.55 ± 0.03
Stop at t_{60}						
Obs.	5%	10%	25%	50%	75%	90%
Betw.	$\mathbf{0.63 \pm 0.02}$	$\mathbf{0.66 \pm 0.01}$	$\mathbf{0.7 \pm 0.01}$	$\mathbf{0.73 \pm 0.02}$	0.74 ± 0.02	0.74 ± 0.02
Cont.	0.6 ± 0.03	0.61 ± 0.02	0.66 ± 0.01	$\mathbf{0.73 \pm 0.02}$	$\mathbf{0.78 \pm 0.02}$	$\mathbf{0.81 \pm 0.02}$
Degr.	0.58 ± 0.02	0.58 ± 0.02	0.58 ± 0.02	0.58 ± 0.02	0.58 ± 0.02	0.59 ± 0.02

Table 1. AUC results for the three methods in each epidemic scenario for a WS network (average and standard deviation).

For a BA network, Degree and Contact present very similar AUC metrics, and in some cases Degree can even achieve slightly better results, but neither of them reaches any AUC value greater than 0.7. In this network, the Betweenness method is constantly better than the other two and is also the only one to return results that are completely distant from a random methodology, with AUC values above 0.7. It is very interesting to note that, except for the Betweenness method, the others do not show a performance improvement when we increase the observation rate. Between their lowest and highest efficiency values, they vary no more than 0.08 for Contact and 0.02 for Degree, while Betweenness improves

its values up to 0.09 more than what we have for the lowest observation rate (5%).

Stop at t_{20}						
Obs.	5%	10%	25%	50%	75%	90%
Betw.	**0.69 ± 0.02**	**0.69 ± 0.01**	**0.72 ± 0.01**	**0.76 ± 0.01**	**0.77 ± 0.02**	**0.78 ± 0.02**
Cont.	0.66 ± 0.03	0.63 ± 0.03	0.63 ± 0.02	0.63 ± 0.03	0.65 ± 0.02	0.65 ± 0.03
Degr.	0.65 ± 0.01	0.65 ± 0.01	0.65 ± 0.01	0.66 ± 0.01	0.65 ± 0.03	0.67 ± 0.03
Stop at t_{60}						
Obs.	5%	10%	25%	50%	75%	90%
Betw.	**0.67 ± 0.01**	**0.67 ± 0.01**	**0.7 ± 0.01**	**0.73 ± 0.02**	**0.75 ± 0.02**	**0.76 ± 0.01**
Cont.	0.63 ± 0.01	0.59 ± 0.02	0.55 ± 0.02	0.55 ± 0.02	0.58 ± 0.01	0.59 ± 0.02
Degr.	0.66 ± 0.01	0.66 ± 0.01	0.66 ± 0.01	0.66 ± 0.01	0.66 ± 0.02	0.66 ± 0.03

Table 2. AUC results for the three methods in each epidemic scenario for a BA network (average and standard deviation).

In an ER network, for low observation values (5% and 10%), the Degree method can perform better even than Betweenness (see (c) in Figure 2), but not on its own merits. In fact, the first one exhibits almost constant behavior as we increase the observation rate, and never shows AUC values above 0.7. Betweenness, on the other hand, starts with slightly worse results at low observation rates but quickly starts to perform well and even reaches excellent results, AUC close to 0.8.

It is interesting the difference of velocity in improvement as we increase the observation rate for Betweenness and Contact in both scenarios. The Betweenness method has very similar values when we look at the results with a 5% observation rate (0.62 in both cases), but if we look at the values with a 90% observation rate, the method reaches considerably better results in the 20% infected nodes scenario - when compared to the 60% of infected nodes (0.79 and 0.73). The same happens with Contact, 0.55 and 0.53 for the lowest observation rate, and 0.77 and 0.69 for 90% observation rate.

Stop at t_{20}						
Obs.	5%	10%	25%	50%	75%	90%
Betw.	**0.62 ± 0.02**	**0.66 ± 0.02**	**0.73 ± 0.02**	**0.76 ± 0.01**	**0.78 ± 0.01**	**0.79 ± 0.03**
Cont.	0.55 ± 0.01	0.58 ± 0.01	0.65 ± 0.02	0.71 ± 0.02	0.75 ± 0.01	0.77 ± 0.03
Degr.	0.6 ± 0.01	0.6 ± 0.02	0.6 ± 0.02	0.6 ± 0.01	0.61 ± 0.02	0.62 ± 0.03
Stop at t_{60}						
Obs.	5%	10%	25%	50%	75%	90%
Betw.	0.62 ± 0.01	**0.66 ± 0.01**	**0.69 ± 0.02**	**0.72 ± 0.01**	**0.72 ± 0.01**	**0.73 ± 0.02**
Cont.	0.53 ± 0.01	0.56 ± 0.01	0.58 ± 0.01	0.63 ± 0.02	0.67 ± 0.02	0.69 ± 0.02
Degr.	**0.66 ± 0.01**	**0.66 ± 0.01**	0.66 ± 0.01	0.66 ± 0.01	0.65 ± 0.01	0.65 ± 0.02

Table 3. AUC results for the three methods in each epidemic scenario for a ER network (average and standard deviation).

It is worth noting that Betweenness and Contact are not as efficient in the scenario with the largest spread of the epidemic (60% of infected nodes) as they are with 20% of infected nodes, having a drop in efficiency of up to 14% when compared to the stop scenario at t_{20}. However, Degree is not affected by this. It is also very little affected by the increase in the observation rate, and as already mentioned, it has very low variation in

its AUC values. The same happens with Contact on the BA network. The Betweenness method, on the other hand, shows continuous improvement as we increase the observation rate in any network and any epidemic scenario.

5. Conclusion

This work considered the problem of identifying asymptomatic nodes in a network epidemic when a fraction of the infected population is randomly observed (the non-observed infected nodes correspond to the asymptomatic nodes). The classic SI epidemic model is considered and the random observation of infected individuals occurs at some instant of the epidemic process (all at once). Different centrality metrics were used to rank the nodes that had not been observed as infected and top-ranked nodes were taken as asymptomatic.

The proposed method based on betweenness centrality consistently outperforms other centrality metrics, especially in scenarios with low observation of infected individuals (observation rate equal to or less than 25%). Moreover, the performance of this metric improves as the fraction of unobserved individuals decreases, for all network models.

This suggests an advantage in using it in scenarios where it is difficult to carry out mass testing or when there is some suggestion of a predominance of asymptomatic individuals over those who show symptoms. As it is a method that takes advantage of the contagion logic of the epidemic in a network, it can achieve high AUC values, even though it is based only on the topological structure of the network and is much simpler than other proposals that use probabilistic tools such as Bayesian inference.

However, in scenarios where the epidemic has spread more (when 60% of nodes are reached), there is a drop in performance in the betweenness-based estimator with respect to its performance at 20% of infected nodes. The large number of observed nodes generates a considerable number of shortest paths that end up encompassing the majority of the remaining nodes. This phenomenon increases the number of false positive cases and the betweenness centrality does not obtain the best results in this scenario.

It is surprising that degree centrality, even without having any epidemic information, manages to perform better than contact centrality in some scenarios, especially on BA networks. But if we think about the case of a hub, for the degree method it will be positioned very high in the ranking, but given its large number of connections, the fraction of infected neighbors ends up being dissipated. If this hub is indeed asymptomatic, degree centrality will identify it more easily.

5.1. Extension to other epidemic models

While this work has focused on the SI epidemic model, other classic models like SIR or SIS could be considered. The methodology here proposed could be directly applied to the SIR epidemic model. In this case, a node observed in the "Recovered" state indicates that it was previously infected. Moreover, a node in R that was not observed can be treated as an asymptomatic node. Evaluations of this scenario are beyond the scope of this work.

As for the SIS model, the proposed methodology is not directly applicable since only infected nodes at time t are observed. An infected node that has returned to the Susceptible state but possibly infected others would never be observed, and this could make the problem of identifying asymptomatic nodes more difficult. A possible approach

would be to observe the epidemic multiple times during its evolution in different instants of time since this would allow us to observe more infected nodes (including ones that later became susceptible). However, evaluations of this scenario are beyond the scope of this work.

References

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97.

Arons, M., Hatfield, K., Reddy, S., et al. (2020). Presymptomatic sars-cov-2 infections and transmission in a skilled nursing facility. *New England Journal of Medicine*, 382(22).

Chen, Y., He, H., Liu, D., Zhang, X., Wang, J., and Yang, Y. (2023). Prediction of asymptomatic covid-19 infections based on complex network. *Optimal Control Applications and Methods*, 44(3):1602–1616.

Comin, C. and da Fontoura Costa, L. (2011). Identifying the starting point of a spreading process in complex networks. *Physical Review E*, 84(5).

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41.

Huang, S., Sun, J., Feng, L., et al. (2023). Identify hidden spreaders of pandemic over contact tracing networks. *Scientific Reports*, 13(11621).

Inui, S. et al. (2020). Chest ct findings in cases from the cruise ship diamond princess with coronavirus disease (covid-19). *Radiology Cardiothorac Imaging*, 2.

Keeling, M. J. and Eames, K. T. (2005). Networks and epidemic models. *Journal of The Royal Society Interface*, 2(4):295–307.

Pellis, L., Ball, F., Bansal, S., Eames, K., House, T., Isham, V., and Trapman, P. (2015). Eight challenges for network epidemic models. *Epidemics*, 10:58–62. Challenges in Modelling Infectious Disease Dynamics.

Quammen, D. (2012). *Spillover: Animal Infections and the Next Human Pandemic*. W. W. Norton Company.

Seguí, F., Cuxart, O., i Villar, O. M., et al. (2021). A cost-benefit analysis of the covid-19 asymptomatic mass testing strategy in the north metropolitan area of barcelona. *International Journal of Environmental Research and Public Health*, 18(13).

Shah, D. and Zaman, T. (2011). Rumors in a network: Who's the culprit? *IEEE Transactions on Information Theory*, 57(8):5163–5181.

Watts, D. and Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393:440–442.

Zhang, R., Tai, J., and Pei, S. (2023). Ensemble inference of unobserved infections in networks using partial observations. *PLOS Computational Biology*, 19(8):1–18.