A comparison of centrality measures and their role in controlling the spread in epidemic networks

Ekaterina Dudkina[a], Michelangelo Bin[b], Jane Breen[c], Emanuele Crisostomi[a], Pietro Ferraro[d], Steve Kirkland[e], Jakub Mareček[f], Roderick Murray-Smith[g], Thomas Parisini[b,h,i], Lewi Stone[j,k], Serife Yilmaz[d] and Robert Shorten[d]

[a]Department of Energy, Systems, Territory and Constructions Engineering, University of Pisa, Pisa, Italy; [b]Department of Electrical and Electronic Engineering, Imperial College London, London, UK; [c]Ontario Tech University, Canada; [d]Dyson School of Design Engineering, Imperial College London, London, UK; [e]Department of Mathematics, University of Manitoba, Canada; [f]Czech Technical University, Prague, Czech Republic; [g]School of Computing Science, University of Glasgow, Glasgow, Scotland; [h]Department of Engineering and Architecture, University of Trieste, Trieste, Italy; [i]KIOS Research and Innovation Center of Excellence, University of Cyprus, Nicosia, Cyprus; [j]The George S. Wise Faculty of Life Sciences, Tel Aviv University, Israel; [k]Mathematics, School of Science, RMIT University, Melbourne, Australia.

ARTICLE HISTORY
Compiled February 13, 2023

ABSTRACT
The ranking of nodes in a network according to their centrality or "importance" is a classic problem that has attracted the interest of different scientific communities in the last decades. The current COVID-19 pandemic has recently rejuvenated the interest in this problem, as it informs the selection of which individuals should be tested in a population of asymptomatic individuals, or which individuals should be vaccinated first. Motivated by these issues, in this paper we review some popular methods for node ranking in undirected unweighted graphs, and compare their performance in a benchmark realistic network that takes into account the community-based structure of society. In particular, we use the information of the relevance of individuals in the network to take a control decision, i.e., which individuals should be tested, and possibly quarantined. Finally, we also review the extension of these ranking methods to weighted graphs, and explore the importance of weights in a contact network by exhibiting a toy model and comparing node rankings for this case in the context of disease spread.

KEYWORDS
Network science; graph theory; centrality measures; dynamic systems.

1. Introduction

1.1. Motivation

The outbreak of COVID-19 and the various attempts to find effective non-pharmaceutical interventions to mitigate the impact of this virus have highlighted

CONTACT Ekaterina Dudkina. Email: ekaterina.dudkina@phd.unipi.it

both the potential value of network science, and also the pressing need to further understand complex interaction networks to support the development of efficient control measures for graphs. Among these interventions, smart testing strategies are one of the most effective measures (Manabe et al., 2020; Salathé et al., 2020). To assess the impact of different proposed strategies, a range of forecasting models has been applied, of which the most popular are compartmental models (Rahimi et al., 2021). Lyng et al. (2021) modeled testing strategies for schools and business considering various parameters (such as test sensitivity, testing frequency, results lag, sample pooling, etc.) based on a Susceptible, Infectious, Removed (SIR) model. Choi and Shim (2021) defined the testing and social isolation strategies based on a Susceptible, Exposed, Infectious and Recovered (SEIR) model. They concluded that in order to reduce the disease burden, in combination with social distancing, the testing should be more intense in the early stage of the infection wave and after the peak. A modified SEIR model was also used by Amaku et al. (2021) to compare two test-trace-and-quarantine strategies to control the COVID-19 outbreak in Brazil. They concluded that total population testing allows reducing the number of cases by 90%, while the testing of symptomatic individuals and their contacts still reached the same number of cases (90% reduction), but at a significantly lower cost for the healthcare system.

Another widely used class of methods consider machine learning algorithms. Bastani et al. (2021) used reinforcement learning to increase the testing efficiency of the traversers of Greek border. The algorithm uses data about passengers' country and region of origin, age and sex to target individuals who should be tested. This measure almost doubled the testing efficiency in terms of number of cases detected. Goodman-Meza et al. (2020) tested seven machine learning techniques and further merged them to create a single model. The input parameters include a set of basic demographic characteristics and blood biomarkers obtained in hospitals, and are used as a proxy for testing.

Alternatively, network science could be applied to deal with the distribution of tests, in the case where tests (or adherence to public health testing requirements) are limited. Indeed, the population in which the virus spreads can be parsed as a network, where individuals are represented by nodes, the registered interactions among individuals are the links between nodes, while relevant parameters, such as contact duration or distance during the contact, may be considered as link weights. Then, naturally, graph theory techniques may be applied to analyze and possibly mitigate the spread of the disease. The current study highlights both the potential value of network science and also the pressing need to further understand complex interaction networks to support the development of efficient control measures for graphs.

Several diverse scientific communities have already developed different techniques to better analyze and understand complex networks of interactions. For instance, in the study of electrical grids, a graph with vertices and edges can represent a power system, with its vertices representing system components and its edges representing interactions or dependencies among those components. Analyses of this grid using graph theory can assess the vulnerability of grid components during cascade failure (a model called extended betweenness combines network structure with electrical characteristics of the power grid (Yan et al., 2014)) or help to integrate renewable energy sources (for example, networks with the small-world property might be more resistant to small variations in load and generation (Koeth, 2019)). In neuroscience, networks may represent connections between neurons and help to understand the architecture and development of the brain. In this context, the detection of communities and bridges helps to "separate functionally related neural elements" and to

study the flow of neural signals and information (Sporns, 2018). Also, important or central nodes may help to identify key players in social networks (Borgatti, 2006), which may ultimately help in finding spread blockers in dynamic networks (Yu et al., 2008), with possibly applications in the mitigation of the spreading of fake news (Bovet and Makse, 2019). Obviously, applications in social networks are also useful for understanding the propagation of epidemics over graphs (Hufnagel et al., 2004).

There are many other research areas in which the application of graph theory has played an important role. However, despite such a great interest from different communities, COVID-19 has illustrated both the importance of graph theory and its enormous potential, but also its limitations in the eyes of policy makers. For example, as we write this note, there are more than 7000 results for the combination of "COVID-19" and "graph theory" on Google Scholar. Yet few of these results, to the best of our knowledge, have actually influenced policy in response to COVID-19. This is despite the fact that all networks arising in studies prior to the pandemic share many important features with networks in which disease propagates. All are "driven by common organizing principles" and obey fundamental laws (Barabási, 2016), and in fact common mathematical methods can be applied to these.

As we have mentioned, the field of network science has benefited from the fact that it is interdisciplinary and applicable in many research areas. Consequently, overlapping contributions have been made in diverse areas. This is both good and bad; good as progress has been fast, but also bad as it makes parsing available results difficult, both from the perspective of understanding how they connect with each other, and from the perspective of applying these results to real-world problems. Indeed, while many comparisons between more classical graph theory indicators, and other community-aware centrality measures, have been already performed in the existing literature (see for instance Ghalmane et al. (2019) and Rajeh et al. (2021)) usually the existing comparisons focus on the correlations between different indicators, and observe qualitative properties of networks (e.g., that community-aware centrality measures are more distinct in situations where the community structure is strong). Conversely, in this work we try to evaluate more quantitative results, and in particular what could be the actual advantage (e.g., in terms of the number of infected individuals) if one ranking measure is used over another to recommend which individuals should be tested, in the case where testing must be limited (i.e. only a certain proportion of the population can be tested per day). For this purpose, and motivated by the recent world interest in the spread of COVID-19, and building on our recent paper on this topic (Yilmaz et al., 2020), the specific objective of the present manuscript is to review the many existing indicators that have been proposed to rank nodes in graphs, and to compare them in terms of their effectiveness and efficiency (i.e., their ability to mitigate the infection spread, considering computational effort for large-scale graphs) in a simplified COVID-19 case study. While the case study is indeed quite simple and considers a limited amount of individuals, it nonetheless allows one to appreciate that different indicators may lead to different results, under certain conditions on the properties of the network. While we focus on the classic case of undirected and unweighted graphs, we also provide an interesting comparison for the weighted case as well, for which very few results can be found in the literature (e.g., see Kamp et al., 2013; Zhang et al., 2013).

1.2. Contributions

This paper provides the following main contributions:

(1) First, we review the most popular centrality measures that have been used in the existing literature to rank nodes in networks. While many other review papers have been published on this topic (see Gómez, 2019; Oldham et al., 2019; Ronqui and Travieso, 2015), in our interdisciplinary review we also include some centrality measures taken from different scientific communities (e.g., epidemiology, operations research, spectral graph theory, control theory, communications and computer science communities) and compare their ability to mitigate the spreading of the virus in the same network. Some of them have never been included in similar comparisons before, see for instance the Kemeny index (Yilmaz et al., 2020).
(2) Second, we propose a benchmark case study which can be used for comparison purposes. This includes a description of how to create a realistic network of contacts, which is adapted from a work on epidemiological studies (Salathé and Jones, 2010), and a simplified model to mimic the COVID-19 spreading dynamics.
(3) As opposed to other comparisons, we do not just compare the correlations of the indicators in ranking the nodes of a single static network. Rather, we make an attempt at evaluating the *cumulative* impact of using a given indicator to make a control decision (testing and quarantining) and compare how the use of different indicators may affect the outcome in disease spread.
(4) Finally, we produce a simple example of a weighted network that motivates the consideration of the duration of a contact between individuals in a network, and allows us to observe what indicators may be more appropriate to use in the case of weighted networks.

Remark: It is important to note that the objective of this paper is to compare and contrast control measures based on different node rankings. We use a simplified model for the disease, and our objective is thus not to propose practical solutions for testing individuals for the COVID-19 pandemic. We are aware that more sophisticated and realistic models than those that are actually deployed in this manuscript, should be used for this purpose to validate any result we obtain. Conversely, this manuscript only uses simplified COVID-19 transmission models as a motivating case study to showcase that *not all nodes are equally important*, and not all ranking methods provide equally useful indications in some circumstances. Accordingly, for the mere purpose of comparison, we did not consider it particularly important that realistic models of COVID-19 transmissions are used, but that exactly the same model is used every time a different ranking indicator is adopted.

1.3. Organization of the paper

The paper is organized as follows: firstly, we briefly describe the indicators that will be later considered in the comparisons, surveying existing literature on this topic. In the following section we present the benchmark case study of undirected and unweighted network and the indicators' comparison. Next, we describe the revisited network to compare indicators in the weighted framework, and show the corresponding obtained ranks of the nodes. Finally, we conclude our manuscript and briefly outline possible extensions of our work.

2. Centrality and connectivity measures of a network

In this section we provide an extensive review of methods for ranking nodes by how "central" they are to the network. While many surveys of this type exist (see, for example Gómez, 2019; Oldham et al., 2019; Ronqui and Travieso, 2015), the aim of the current work is to compare and contrast these and their effectiveness in the context of disease spread in a contact network. We also include several centrality measures which are currently underrepresented in network science, that play a more prominent role in other disciplines. Furthermore, we consider alternative methods for ranking node centrality by considering how much each node's removal contributes to a change in some global measure of connectivity of the network. As such, we also review some network connectivity measures, and their interpretation in the context of disease spread.

2.1. Some mathematical preliminaries

A simple undirected graph is denoted $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with vertex set $\mathcal{V} = \{1, \ldots, N\}$ and edge set $\mathcal{E} \subseteq \{\{i, j\} \mid i, j \in \mathcal{V}\}$. We say that i and j are *adjacent*, and denote this as $i \sim j$, if there is an edge between i and j (i.e. $\{i, j\} \in \mathcal{E}$). If $i \sim j$, we say that j is a *neighbour* of i. The number of neighbours of i is referred to as the *degree* of vertex i, and denoted $\deg(i)$. The adjacency matrix of $\mathcal{G}$ is the symmetric matrix A defined entrywise as:

$$a_{ij} = \begin{cases} 1, & \text{if } i \sim j; \\ 0, & \text{otherwise.} \end{cases}$$

A directed graph with vertex set $\mathcal{V}$ as above is one in which the edges are *ordered pairs* of vertices (i, j); i.e. $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. The edge (or arc) (i, j) is considered to represent a connection from i to j, sometimes denoted $i \to j$, to indicate that the arc's initial vertex is i and terminal vertex is j. The adjacency matrix is defined in the same way as above, though $a_{ij} = 1$ if and only if $(i, j) \in \mathcal{E}$, and $a_{ij} \neq a_{ji}$ in general.

A weighted graph arises when each edge $\{i, j\}$ is assigned a weight w_{ij}. This is taken into account in the adjacency matrix by simply allowing the (i, j) entry of the matrix to be the weight w_{ij}.

Directed and weighted graphs provide generalizations of the simple undirected graph which may be very useful in the context of disease spread in a contact network. Differing weights of edges may account for differing strengths of transmission between pairs of individuals, due to, for example, different circumstances of the interaction—length of time, distance, etc. Asymmetric values allow for transmission to be more likely in one direction than the other between two individuals, for reasons more pertinent to the individual; for example, different levels of susceptibility, preventative measures, waning immunity, age differences, etc.

Centrality measures and connectivity measures can be defined for simple undirected graphs, and while some definitions may allow extensions to the weighted or directed case, they may lose some aspect of their interpretation; some do not generalize at all. In what follows, we attempt to indicate for each metric listed whether or not they do generalize in this way. The ones which extend most naturally, we find, are those which are derived from random walks, and so we include some mathematical preliminaries pertaining to these here.

A random walk on a connected undirected graph G (and here we shall consider

strongly connected graphs in the directed case) is a discrete-time stochastic process in which, at any given time, a "random walker" occupies one vertex of the graph, and in a subsequent time-step, moves to an adjacent vertex j of his/her current vertex i, according to some transition probability p_{ij}. For a simple random walk on an undirected graph, the transition probability p_{ij} is simply $\frac{1}{\deg(i)}$; that is, the random walker chooses his/her next position uniformly at random from among the neighbours of his/her current vertex. This process is a Markov chain whose state space is the vertex set of G, since the state in any time-step depends only on the state of the chain in the previous time-step. The probability transition matrix $P = [p_{ij}]$ for this Markov chain is easily determined from the adjacency matrix A by normalizing the rows so that they sum to 1. In the case of weighted graphs, the transition matrix may be determined in exactly the same way from the weighted adjacency matrix, although a novel alternative will be later described in this manuscript; it follows similarly in the case of directed graphs, although if there are any vertices with no outgoing edges, the random walk is no longer well-defined.

For an ergodic Markov chain with states indexed $1, \ldots, n$ and $n \times n$ transition matrix P, the stationary distribution vector of P, denoted w, is a left eigenvector of P corresponding to the eigenvalue 1, normalized so that the entries of w sum to 1 and thus represents a probability distribution across the states. In particular, w_i represents the long-term probability that the Markov chain occupies the i^{th} state. Note that in the case of a simple random walk on a connected undirected graph, the stationary distribution vector entries are proportional to the vertex degrees;

$$w_i = \frac{\deg(i)}{\sum_k \deg(k)}.$$

For any two states i and j, the *mean first passage time from i to j*, denoted m_{ij}, is the expected time it takes to reach the j^{th} state, given that the chain starts in the i^{th} state.

2.2. Node centrality measures

2.2.1. Degree centrality

The first and simplest proposal for ranking nodes in a network is by considering each node's degree (sometimes called valency), or the number of neighbours of each node. The degree of node i can be easily computed via the i^{th} row sum of the adjacency matrix:

$$\deg(i) = \sum_{j=1}^{N} a_{ij}. \tag{1}$$

The degree is a simple centrality measure for undirected networks, where it identifies the most connected nodes, in the sense that highly-ranked nodes under this metric will have the most neighbours. In the context of disease spread in a contact network, these highly-ranked nodes correspond to individuals with the most contacts; as such, a high-degree node infected with the disease has more opportunity to spread the disease to other individuals.

In the case of directed graphs, one has to distinguish between incoming and outgoing

edges, defining the *in-degree* and *out-degree*, respectively, as

$$\deg_-(i) = \sum_j a_{ij} \qquad \deg_+(i) = \sum_k a_{ki}.$$

Given this dual definition, it is more difficult to consider degree as a measure of centrality in the directed case; some options are to consider the sum or the average of the two (as in Gómez, 2019).

For weighted graphs, the degree of a vertex is easily extended by defining $\deg(i)$ as the sum of the weights of incident edges. This is sometimes referred to as the *strength* of a vertex (see, for example Barrat et al., 2004).

2.2.2. Closeness centrality

The *distance* between nodes i and j, denoted $\text{dist}(i, j)$, is defined as the minimum number of consecutive edges needed to move from node i to j or, equivalently, as the length of a shortest path between them. The *closeness centrality* of a node i is computed by taking the inverse of the average distance from i to any other node:

$$CC(i) = \frac{N-1}{\sum_{j \neq i} \text{dist}(i, j)}. \tag{2}$$

The larger the value of $CC(i)$, the more central node i is in the network, in the sense that it is, on average, close to many other nodes.

As with degree centrality, the definition of closeness centrality can be extended to directed networks, though a distinction must be made on whether the distances are computed from, or to, the reference node i, respectively. Note also that "distances" are not symmetric in directed networks. For weighted graphs, one could consider the edge-weights as a "cost" to traversing the edge, and thus define shortest-distance between u and v as the minimum weight of any path from u to v (where the weight of a path is the sum of the weights of edges in the path). With such a definition for distance in hand, it is reasonable to generalize the closeness centrality for weighted graphs; see Opsahl et al. (2010) for some limited discussion.

In a disease spread context, ranking vertices in a contact network by their closeness centrality would, in theory, highlight individuals for whom there is (on average) low degree of separation between the individual and all other members of the community. If this individual were infected, then, it takes fewer secondary infections on average to infect others.

2.2.3. Betweenness centrality

The *betweenness centrality* of a node i is computed in terms of how many shortest paths pass through that node (Freeman, 1977, 1978). In particular, fix a source node s and target node t (distinct from i), and let σ_{st} denote the total number of shortest paths from s to t (or geodesics, i.e. paths of length $dist(s, t)$). Letting $\sigma_{st}(i)$ denote the number of those paths that include node i, we take the ratio of these and then average over all choices for $s, t \neq i$:

$$BC(i) = \frac{2}{(N-1)(N-2)} \sum_{\substack{s,t=1 \\ s,t \neq i}}^{N} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{3}$$

Accordingly, the betweenness centrality of a node corresponds to the fraction of shortest paths that pass across that node, and this expression is valid for both directed and undirected networks. In principle, betweenness centrality is expected to rank highly the nodes that behave as bridges between clusters in the network (Gómez, 2019). In the context of disease spread, these highly-ranked nodes would correspond to individuals who bridge multiple communities.

We encounter similar difficulties with the extension of betweenness centrality to weighted and directed graphs as with closeness centrality. Some limited work exists; see for example H. Wang et al. (2008), that focuses on the betweenness centrality of an edge in a weighted network, rather than betweenness centrality of nodes, and White and Borgatti (1994), that discusses several possible extensions to directed graphs and their limitations.

2.2.4. PageRank centrality

The PageRank algorithm computes a ranking for every web page based on the graph of the World Wide Web. PageRank has applications in search engines and traffic estimation (Page et al., 1999). The PageRank algorithm is best understood via a random walk on the network.

PageRank can be thought of as a model of a "random surfer" who starts on a random webpage and keeps clicking on links randomly, never hitting "back"; that is, he/she takes a random walk on the World Wide Web, a directed graph in which (i, j) is an edge if webpage i has a hyperlink to webpage j. At any point, the surfer may get bored and "teleport" to a random page in the network. The long-term probability that the random surfer occupies webpage i in this stochastic process also corresponds to the stationary distribution of the corresponding Markov chain. Note that the transition matrix for this Markov chain is

$$P = (1-\alpha)D^{-1}A + \alpha \tfrac{1}{N} J, \tag{4}$$

where D is the diagonal matrix of vertex out-degrees, A is the adjacency matrix, and J is the $N \times N$ matrix of all ones (Langville and Meyer, 2006). The parameter α is related to the *damping factor* and represents the probability at each step that the "random surfer" will request another random page; as a matter of convention, this value is usually set to 0.15 (Brin and Page, 1998). We note that in the case of the World Wide Web, the underlying directed graph is not strongly-connected; indeed, there may be many webpages with no outgoing hyperlinks. This causes immediate problems with the random walk being well-defined, but is usually fixed by replacing any zero row of the adjacency matrix with a row of all 1s. Thus when the random surfer ends up on a webpage with no outgoing links, he/she chooses a webpage uniformly at random in the next step. This allows computation of the PageRank vector in the setting that some nodes have outdegree 0.

The interpretation of this measure of centrality is interesting in the context of disease spread. In the case that we work with a simple undirected graph, if $\alpha = 0$ (i.e. a simple

random walk with no teleportation), then the PageRank centrality ranking corresponds exactly to the node degree centrality ranking. However, including a damping factor may allow for the possibility that a person contracts the disease not from another individual they have contact with, but rather by chance (e.g. touching a surface with traces of the virus), or by some mechanism not accounted for in the contact network model. PageRank centrality ranking of the nodes can essentially be thought of as strongly related to the node degree ranking in this context, with some relaxation that may actually reflect the features of the disease more accurately.

If one wishes to consider a weighted or directed graph, PageRank centrality is a measure which can be applied along with its interpretability, simply by considering a random walk on the given weighted or directed graph.

2.2.5. Random walk betweenness centrality

A measure of betweenness centrality based on random walks, called *random walk betweenness (RWB)* was introduced by Newman (2005). The idea is to calculate the centrality of a given node i by the proportion of random walks that pass through node i. This is very similar to betweenness centrality, but does not consider shortest paths. In particular, Newman (2005) describes it as "the expected net number of times a random walk passes through vertex i on its way from a source s to a target t, averaged over all s and t". This measure was shown to better rank the importance of nodes in graphs with existing communities, and to be less correlated with vertex degree in most networks (Newman, 2005). Interestingly, the method for computing this measure is strongly dependent on considering the graph in question as an electrical network and considering current flow through a vertex. The author then proves that this is equivalent to the "flow" of a random walk; however, this means that this particular definition of random walk betweenness does not extend to directed or weighted graphs, and there is no literature which attempts this.

Random-walk betweenness centrality is obtained by averaging the current flow ($f_i^{(st)}$) through vertex i over all possible origins (s) and destinations (t),

$$RWB(i) = \frac{2}{N(N-1)} \sum_{s<t} f_i^{(st)}. \tag{5}$$

See Newman (2005) for further information on how this quantity is calculated.

While it is not necessarily obvious from the electrical network description above, there is a large body of literature on analogy between electrical networks and random walks (see Doyle and Snell, 1984), and this value does indeed compute the expected *net* number of times a random walk would pass through node i before reaching a target t, given that it starts at a source s, averaged over all pairs s, t. This intuitively seems to indicate "pivotal" individuals in a contact network in which a disease is spreading at random, and as we will see in the next section, this metric is one which is particularly effective in controlling the disease when used exclusively to determine testing protocols. It is also known that in networks with strong community structure, immunization interventions targeted at individuals bridging communities (e.g., using random walk betweenness) are more effective than those simply targeting highly connected individuals (Salathé and Jones, 2010).

Note that the definition 5 can be directly applied only to simple undirected graphs; there does not exist a generalization to weighted or directed graphs.

2.2.6. Random walk centrality (RWC)

Random walk centrality is introduced by Noh and Rieger (2004), and is said to "quantify how central a node i is located regarding its potential to receive information which is randomly diffusing over the network". Given a graph G, consider a random walk on the graph with transition matrix P. The stationary distribution vector for P is denoted w, and the k^{th} entry of w is denoted w_k. The characteristic relaxation time of vertex k is introduced as $\tau_k \equiv \sum_{j=0}^{\infty}((P^j)_{k,k} - w_k)$. This quantity converges whenever the transition matrix P is primitive, which is the case for a random walk on a connected non-bipartite graph. The *random walk centrality* (Noh and Rieger, 2004) of vertex k is then calculated as $C_k \equiv \frac{w_k}{\tau_k}$.

Kirkland (2016) observes that the random walk centrality is the reciprocal of the measure of centrality known as the *accessibility index*. For a given vertex k, the accessibility index of vertex k is defined

$$\alpha_k = \sum_{j \neq k} w_j m_{jk},$$

where m_{jk} is the mean first passage time from j to k. This definition allows the interpretation of the accessibility index as the expected time to reach vertex k for the first time, given that we start in any randomly-chosen initial state. It is shown by Kirkland (2016) that $C_k = 1/\alpha_k$. This is particularly useful as it admits a definition of this as a measure of centrality not just for simple undirected graphs (as is introduced by Noh and Rieger (2004)) but for any Markov chain, as a measure of state centrality. In particular, this is a useful metric for both weighted and directed graphs. Further work on estimating the accessibility index using statistical techniques may be found in Johnson and Kirkland (2019).

In the disease spread context, this measure ranks highly the vertices which are "easily accessed" from other vertices of the graph. This is a useful way to consider how "central" an individual is in a community. However, in the context of the spread of disease, and in particular when considering testing protocols in order to control the disease, this may not be appropriate. Controlling the disease in our simulations means determining individuals who are most instrumental in the disease spreading through the whole graph, and while those individuals with high random walk centrality may have increased likelihood of being infected, this may not coincide with individuals who ought to be tested and isolated as quickly as possible.

2.3. Network connectivity measures

Many connectivity measures exist which provide a single, numerical value describing the "connectedness" of a network in some way. Such measures can be easily extended to measure the criticality of a single node i in a graph $\mathcal{G}$ by inferring the criticality of the i^{th} node from the change in criticality of the network after the i^{th} node is removed from the network. This is described as follows in Fouss et al. (2016): given some graph invariant $cr(G)$ measuring the connectedness of the graph G in some manner, define

$$cr_i(\mathcal{G}) := cr(\mathcal{G}) - cr(\mathcal{G} \setminus i).$$

We include several such measures of node centrality here for consideration in later simulations.

2.3.1. Effective graph resistance (Kirchhoff index)

The effective graph resistance is interpreted as a "robustness measure" of a network (X. Wang et al., 2017). To formulate the effective graph resistance (also known as the total graph resistance, or the Kirchhoff index of the graph), the (undirected and connected) graph G is seen as an electrical circuit, where an edge $\{i, j\}$ corresponds to a resistor of r_{ij} Ohm, and the effective graph resistance is the sum of the effective resistances over all pairs of vertices. These effective resistances R_{st} can be computed using the Laplacian matrix of the graph.

The Laplacian matrix of a graph is defined as the adjacency matrix subtracted from the diagonal matrix of vertex degrees: $L = D - A$. Note that in the context of the electrical network, an undirected, unweighted graph is considered to have resistors of $r_{ij} = 1$ Ohm on each edge. However, this can be extended to weighted networks, where resistors have $r_{ij} = 1/w_{ij}$, where w_{ij} denotes the edge weight between vertices i and j. In this case, the weighted Laplacian matrix can easily be defined, and used to extend the following results.

Through applications of Kirchhoff's current and circuit laws, one can show that the effective resistance R_{st} between vertices s and t may be calculated using the Moore-Penrose pseudoinverse $L^{\dagger}$ of the Laplacian matrix of a graph:

$$R_{st} = (e_s - e_t)^T L^{\dagger}(e_s - e_t),$$

where e_i denotes the standard unit vector, i.e., a vector consisting of all zeros except a single 1 in the i^{th} position. Given this expression for the effective resistance between two vertices, the total graph resistance (or resistance distance - RD) can be expressed in terms of the sum of the reciprocals of the nonzero eigenvalues of the Laplacian matrix (Klein and Randić, 1993):

$$RD(G) = \frac{1}{N} \sum_{j=2}^{N} \frac{1}{\rho_j}. \tag{6}$$

As previously mentioned, this measure can be easily extended to weighted graphs, though it is not clear that a generalization for directed graphs is possible.

2.3.2. Kemeny's constant

Given an ergodic Markov chain with transition matrix P, stationary vector w, and mean first passage times m_{ij}, one can define the following quantity:

$$\kappa_i = \sum_{j=1}^{N} m_{ij} w_j.$$

This can be interpreted as the expected time to reach a randomly-chosen state j, given that the chain starts in state i. Astonishingly, this quantity is independent of the initial state i (Kemeny and Snell, 1960). Thus *Kemeny's constant*, denoted as $K(P)$, can be computed as

$$K(P) = \sum_{j=1}^{N} m_{ij} w_j. \tag{7}$$

An interpretation of this result is that the expected time to reach a randomly-selected destination state j from a fixed initial state i (where state j is selected randomly according to the stationary distribution w) does not depend on the starting point i (Doyle, 2009). Furthermore, since the w_j sum to 1, we can write Kemeny's constant as a double-sum:

$$K(P) = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i m_{ij} w_j,$$

admitting the interpretation of Kemeny's constant as the expected length of a random trip in the chain, where both initial and destination states are chosen randomly according to the stationary distribution. Therefore, Kemeny's constant is an intrinsic measure of a Markov chain. If the transition matrix P has eigenvalues $\lambda_1 = 1, \lambda_2, ..., \lambda_n$, then another way of computing $K(P)$ is (Levene and Loizou, 2002),

$$K(P) = \sum_{j=2}^{N} \frac{1}{1 - \lambda_j}. \tag{8}$$

Note that this expression furnishes a computationally useful method for determining $K(P)$ in practice; $K(P)$ is computed as the trace of a given generalized inverse of the singular matrix $I - P$.

Kemeny's constant is a proxy for the global "connectedness" of a network, given the interpretation as the expected length of a random trip between states in the chain. Thus, networks characterised by small values of Kemeny's constant should be more efficient in terms of flow (Crisostomi et al., 2011). In a contact network, small values of Kemeny's constant for the random walk on the graph indicate that the individuals in the network are well-connected, while large values of Kemeny's constant may be indicative of clustered behaviour, and that it is difficult to traverse the graph. If one can identify the vertex whose removal causes the largest increase in the value of Kemeny's constant, this could be interpreted as a "central" vertex.

Since Kemeny's constant can be computed for any ergodic Markov chain, it extends most usefully to weighted, directed graphs.

2.3.3. Subdominant Eigenvalues

For an ergodic Markov chain with transition matrix P, the eigenvalues may be listed $\lambda_1 = 1, \lambda_2, \ldots, \lambda_n$, and by Perron-Frobenius theory, $|\lambda_j| \leq 1$ for all $j = 2, \ldots, n$. As is evident from our discussion of Kemeny's constant above, much information regarding the dynamics of the Markov chain may be extracted from these eigenvalues. We are frequently interested in the eigenvalue λ_j for which $|\lambda_j|$ has second-largest modulus (SLEM) after 1. Without loss of generality, suppose this is λ_2. We outline here several ways in which the value of λ_2 can be used to infer how well-connected the states of a Markov chain are (or the vertices of a graph; where the chain in question represents a random walk on a graph). In general, if λ_2 is bounded away from 1, the states of the Markov chain are considered to be "well-connected".

For a Markov chain with transition matrix P, the probability distribution after k time-steps is given by $u_k^T = u_0^T P^k$, where u_0 is the initial probability distribution vector. It is well-known that if the Markov chain is ergodic (i.e. the matrix P is primitive), then u_k^T converges to the stationary distribution vector w^T of the chain

as $k \to \infty$, independently of the initial distribution. The asymptotic rate of this convergence is clearly dictated by the moduli of the eigenvalues of P. If all eigenvalues have modulus bounded away from 1, convergence happens quickly. This convergence is often referred to as the "mixing rate" of the chain, and is framed in terms of how quickly the initial information is lost.

In the context of a random walk on an undirected connected graph, the transition matrix for the random walk on G is $P = D^{-1}A$; denote its eigenvalues by $1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. There is the following bound on the total variation distance after k time-steps have passed (see Chung and Graham, 1997):

$$\|u_0^T P^k - w^T\| \leq \max_{j \neq 1} |\lambda_j|^k \frac{\max_i \sqrt{\deg(i)}}{\min_j \sqrt{\deg(j)}}.$$

As P has real eigenvalues, issues arise when λ_j is close to 1 or close to -1. We note that there is some difference in the dynamics of the Markov chain in these cases—a subdominant eigenvalue close to 1 in value is indicative of clustering behaviour, or of near-decoupling (Breen et al., 2018; Crisostomi et al., 2011; Hartfiel and Meyer, 1998), while -1 occurs as an eigenvalue of P if the undirected graph is bipartite, causing periodic behaviour in the Markov chain. For this reason, in the context of disease spread in a network, we are more interested in a subdominant eigenvalue whose value is close to 1 (not whose modulus is close to 1). See Breen et al. (2018) for more discussion.

The subdominant eigenvalue λ_2 and its relationship to graph structure is also well-studied in the context of the normalized Laplacian matrix, defined as

$$\mathcal{L} = D^{-1/2}(D - A)D^{-1/2}$$

when G has no isolated vertices. It is easily seen that $\mathcal{L}$ is similar to the matrix $I - D^{-1}A$, so that the normalized Laplacian eigenvalues $0 \leq \mu_1 \leq \cdots \leq \mu_{n-1}$ are in one-to-one correspondence with the eigenvalues of $D^{-1}A$, where $\mu_j = 1 - \lambda_{j+1}$. In particular, the eigenvalue $\mu_1 = 1 - \lambda_2$ is often referred to as the *algebraic connectivity* of the graph (Chung and Graham, 1997) (not to be confused with Fiedler's algebraic connectivity, defined in the next section as an eigenvalue of the combinatorial Laplacian $L = D - A$).

The eigenvalue μ_1 (or the spectral gap $1 - \lambda_2$) is well-known to be related to isoperimetric numbers of the graph. In particular, we have the following relationship with Cheeger's constant $h(G)$ (Chung and Graham, 1997):

$$\frac{h(G)^2}{2} \leq \mu_1 \leq 2h(G).$$

The Cheeger's constant is a measure of how much a vertex set can expand in the graph; formally,

$$h(G) = \min_{\substack{S \subset V(G) \\ vol(S) \leq \frac{1}{2} vol(G)}} \frac{e(S, \overline{S})}{vol(S)},$$

where $vol(S)$ is twice the number of edges between vertices in S, and $e(S, \overline{S})$ denotes the number of edges in the graph between vertices in S and in $\overline{S}$, i.e. vertices outside

S. We note that graphs with low values of Cheeger's constant typically have "bottleneck" vertices, or clusters/communities with very few edges between them. Finally, the quantity $1/\mu_1$ is referred to as the *relaxation time* of the random walk, and is studied in itself as a measure of the asymptotic rate of convergence to the stationary distribution (see Aldous and Fill, 2002). Note that the relaxation time $1/(1-\lambda_2)$ is the first and largest term in the expression (8) of Kemeny's constant.

In the context of disease spread in a graph G, the value of the second-largest eigenvalue λ_2 of the transition matrix for the random walk (and how close it is to 1) may be used to indicate how quickly the disease may disperse through the graph, as a measure of clustering behaviour in the graph, or as a stand-in for Cheeger's constant, indicating whether the graph has good expansion properties (which would imply that the disease spreads quickly).

Since subdominant eigenvalues can be calculated for any Markov chain, this measure is easily extended to weighted and directed networks. We remark that in the case that a subdominant eigenvalue is a complex number, some work has been done in that area to show how clustering behaviour may be derived from the value of λ_2 in that case (Breen et al., 2018).

2.3.4. Algebraic Connectivity (AC)

The eigenvalues of the Laplacian matrix of a graph G, $L = D - A$, can be listed $\rho_0 = 0 \leq \rho_1 \leq \cdots \leq \rho_{n-1}$, and the second-smallest eigenvalue ρ_1 is referred to as the *algebraic connectivity* of the graph and denoted $a(G)$ (Fiedler, 1989). This eigenvalue is strictly greater than 0 if and only if $\mathcal{G}$ is a connected graph. The magnitude of this value reflects how well connected the overall graph is. It has been used in analysing the robustness and synchronizability of networks. There are many papers and results relating the value of $a(G)$ to many other structural quantities in the graph, such as vertex- and edge-connectivity, diameter, minimum degree, isoperimetric constants, and many others. For an overview, see the excellent survey by De Abreu (de Abreu, 2007). There is some research in the area of extending the definition of algebraic connectivity to directed graphs (Wu, 2005) and how it may be interpreted.

2.3.5. The Basic Reproduction Number ($\mathcal{R}_0$)

The basic reproduction number $\mathcal{R}_0$ is perhaps the most popular indicator in the epidemiology community, and it may be defined as the expected number of individuals that a randomly infected individual can infect during his/her infection period in a fully-healthy susceptible population (Mei et al., 2017). This value depends on the specific disease (e.g., its intrinsic infectivity), and on the topology of the network of contacts. In particular, it can be shown that in a wide class of network-SIS models (a susceptible-infected-susceptible model) $\mathcal{R}_0$ is proportional to λ_{max}, where λ_{max} denotes the dominant eigenvalue of the adjacency matrix (Castellano and Pastor-Satorras, 2017; Mei et al., 2017; Y. Wang et al., 2003). Accordingly, in this paper we shall use the dominant eigenvalue of the adjacency matrix as a proxy for $\mathcal{R}_0$, with the ultimate goal of testing individuals whose removal gives rise to the lowest value of $\mathcal{R}_0$. We note that this eigenvalue is shown to reflect the basic reproduction number in Mei et al., 2017 under certain modelling assumptions which may not hold in our simplified model. Nevertheless, the largest eigenvalue of the adjacency matrix is well-known as a key quantity in determining properties of dynamical systems on graphs; see Restrepo et al., 2006 for similar work in the criticality of nodes and edges with respect to λ_{max}.

2.4. Modularity

The last measure we discuss here is the *modularity* of a network, which is considered to quantify the degree of community structure of the network. In order to define this measure (as defined by Newman and Girvan (2004)), we assume there exist some predetermined communities to which the vertices of a graph G belong, and partition the vertex set accordingly; say $V(G) = S_1 \cup S_2 \cup \cdots S_r$. We define e_{ij} to be the fraction of all edges in the network that join vertices from S_i with vertices from S_j, with e_{ii} considered as the fraction of edges within the community S_i (i.e. $vol(S_i)/vol(G)$). Define $a_i = \sum_{j=1}^{r} e_{ij}$, considered as the fraction of edges that connect to vertices in S_i. Then the modularity is defined

$$Q = \sum_{i=1}^{r} (e_{ii} - a_i^2).$$

This quantifies the degree of community structure in the network by comparing the fraction of all edges that are within a community to the fraction of the edges connecting that community to the other communities, and observing that if there was no community structure, one could expect that $e_{ij} = a_i a_j$, giving a modularity of $Q = 0$.

In this present manuscript we do not consider modularity as one of our criticality measures used to rank nodes by the effect of their removal on this quantity. Indeed, we note that due to how we construct our contact networks, modularity is unlikely to be sensitive to the removal of a single node, and would not produce meaningful results if considered as a means of ranking nodes. Instead, modularity plays a more vital role in our simulations, as we construct contact networks with varying levels of community structure as described by modularity. We follow the method outlined in Salathé and Jones (2010) to construct these; see section *Case study: Creation of the network of daily contacts*.

3. Comparison for the COVID-19 case study

3.1. Case study: Creation of the network of daily contacts

Our case study is created according to the procedure outlined by Salathé and Jones (2010), and later reused also by Yilmaz et al. (2020). The procedure to create the network can be summarized in the following steps:

(1) 6 small-world communities of 40 nodes are first created using the Watts-Strogatz algorithm (Watts and Strogatz, 1998), so that each node has exactly N_d^0 edges connecting to nodes of the same community;
(2) We then add $30 \cdot N_d^0$ edges in a random way to connect different communities; after this step, the average node degree becomes $N_d = \frac{5}{4} N_d^0$;
(3) We then rewire *between-communities* edges (i.e., edges that connect nodes belonging to two different communities) so that they become *within-community* edges (i.e., edges that connect nodes belonging to the same community). In doing this, the modularity of the graph increases, and we stop the procedure once a desired level of modularity M is achieved.

In particular, we compare the different indicators for different values of N_d, that correspond to the number of daily contacts of individuals, and for different values of

modularity M, that allows one to evaluate what happens for societies with weaker or stronger community structures. We create a new network for each day in the simulation according to the procedure previously given, and we assume that it corresponds to a network of daily contacts. Individuals remain in the same communities each day, but inter- and intra-community edges change from day to day.

3.2. Case study: simulation of epidemic spread on a graph

The previously-described procedure had been proposed in Salathé and Jones, 2010 to mimic the network of contacts of individuals in a society, with the final objective of evaluating the impact of different testing/vaccination campaigns, to better mitigate the spreading of epidemics. While such networks could be only guessed in that context, contact tracing applications are now giving the unprecedented advantage of indeed knowing when individuals meet at a close distance for a sufficient time to infect a new individual (e.g., at least 15 minutes at a distance of 1 meter). An implicit assumption here is that individuals allow the sharing of the information of their daily contacts to some centralized data center that aggregates this kind of information and knows the network (e.g., in terms of an adjacency matrix).

We then model the spreading of the virus, and the testing procedure in the following simplified way: On the first day, we randomly label two individuals as "infected", and they correspond to our initial condition. Every single day, we assume that a susceptible individual who gets in contact with an infectious individual, has a probability of 10 % of being infected. Also, every day, we test N_{test} individuals according to the different indicators introduced in Section 2: if a tested individual is found infected, then he/she is quarantined for the following 14 days, after which we assume the individual is fully recovered and not susceptible anymore (i.e., they can not be infected again). Also all his/her contacts of the same day are further tested (but not those of the previous days). During the quarantine, we assume that the individual is fully compliant with the rules, and does not have contact with any other individual. It may also happen that infectious individuals are never tested, and obviously may infect other individuals as well as they are not quarantined, and we assume that they also recover after 14 days, after which they are not susceptible anymore.

The previous model is obviously a simplification of the COVID-19 dynamics, and may be seen as a simplified agent-based SIQR model, where each individual in the population may belong to only one of the disjoint compartments of Susceptible – Infected – Quarantined – and Recovered individuals. While more accurate and sophisticated models have been recently introduced in this context, e.g., the popular SIDARTHE model (Giordano et al., 2020), the core of most compartmental models is the same as the one analyzed by Bin et al. (2021). Since we only test individuals on the basis of the topology of the network, depending on the specific indicator of interest, this method may be applied in practice to target-testing individuals in a population of asymptomatic individuals, and should be obviously complemented with the good practices of testing individuals who, independently from the topology of the network, exhibit COVID-19 symptoms. However, as we have already mentioned, simulations results are only intended to be used in a comparative fashion, and actual or realistic applications of any indicator is beyond the purpose of this work.

3.3. Discussion of the simulation results

Simulation results are described in Fig. 1, where the final number of susceptible individuals at the end of the simulation (i.e., after 30 days) is shown. Note that due to the adopted COVID-19 spreading model, the final number of susceptible individuals corresponds to 240 (size of the considered population) minus the number of infected individuals (so a larger number of susceptibles individuals corresponds to a smaller number of infected individuals). Due to the stochasticity of the considered model, results shown in Fig. 1 are averaged over 100 simulations, and every time two randomly chosen individuals are supposed to be infectious. The evolution of the number of infected individuals throughout the 30 days of simulations is shown in Fig. 2. As can be seen on the graph, the infection spread starts from 2 individuals, and is constrained by testing and quarantining individuals according to different indicators. It is possible to observe that the curve of infected individuals has a similar shape in all cases, but by testing individuals using some of the indicators (e.g., RWB, SLEM, BC, and Kemeny) allows one to significantly reduce the peak of the curve.

In the simulations we have considered networks of modularity equal to 0.8, average node degree N_d equal to 7.5 and the possibility of testing 20 different individuals every day. Then such default values are perturbed to observe how results consequently change, as summarized in Fig. 1. In particular, from the figure above it is possible to appreciate that better results can be obtained if networks with stronger community structure are considered (values of modularity greater than 0.8). This result also highlights the importance of the indicator choice depending on the graph structure: while in the absence of communities all indicators perform in a similar manner, with the increase of modularity significant differences may be noticed, and the indicators that are able to identify bridges between communities (such as AC, SLEM, RWB, Kemeny) are associated with a higher number of healthy people by the end of the simulation. The second panel of Fig. 1, demonstrates that the most relevant quantity is the average node degree N_d. This could be expected since it corresponds to the number of daily relevant contacts. Finally, the last panel shows the obvious result that as the number of daily tests is increased, then a reduced number of individuals gets infected.

3.4. Computational burden for different indicators

With the increase of the network size, the computational burden could become too high for some indicators, and thus limit their scalability. Table 1 reports the computation time for the different indicators with a personal computer, equipped with a 6-core i7-8700 CPU @ 3.20GHz, RAM 16 GB. The values correspond to an average time for ten runs of the whole 30-day simulation and are sorted accordingly.

The indicators that can be computed fastest are those measuring the node centrality, i.e., Page Rank, Closeness centrality, Node degree, Betweenness centrality, and RWC. As expected, node connectivity measures are more resource-demanding and result at least one order of magnitude slower. Among them, the most computationally expensive are Second Largest Eigenvalues in Modulus (SLEM) and Kemeny constant. Finally, it is noteworthy that although RWB still measures the node centrality, it is the most burdensome measure among all indicators (almost 300 times slower than Page rank).

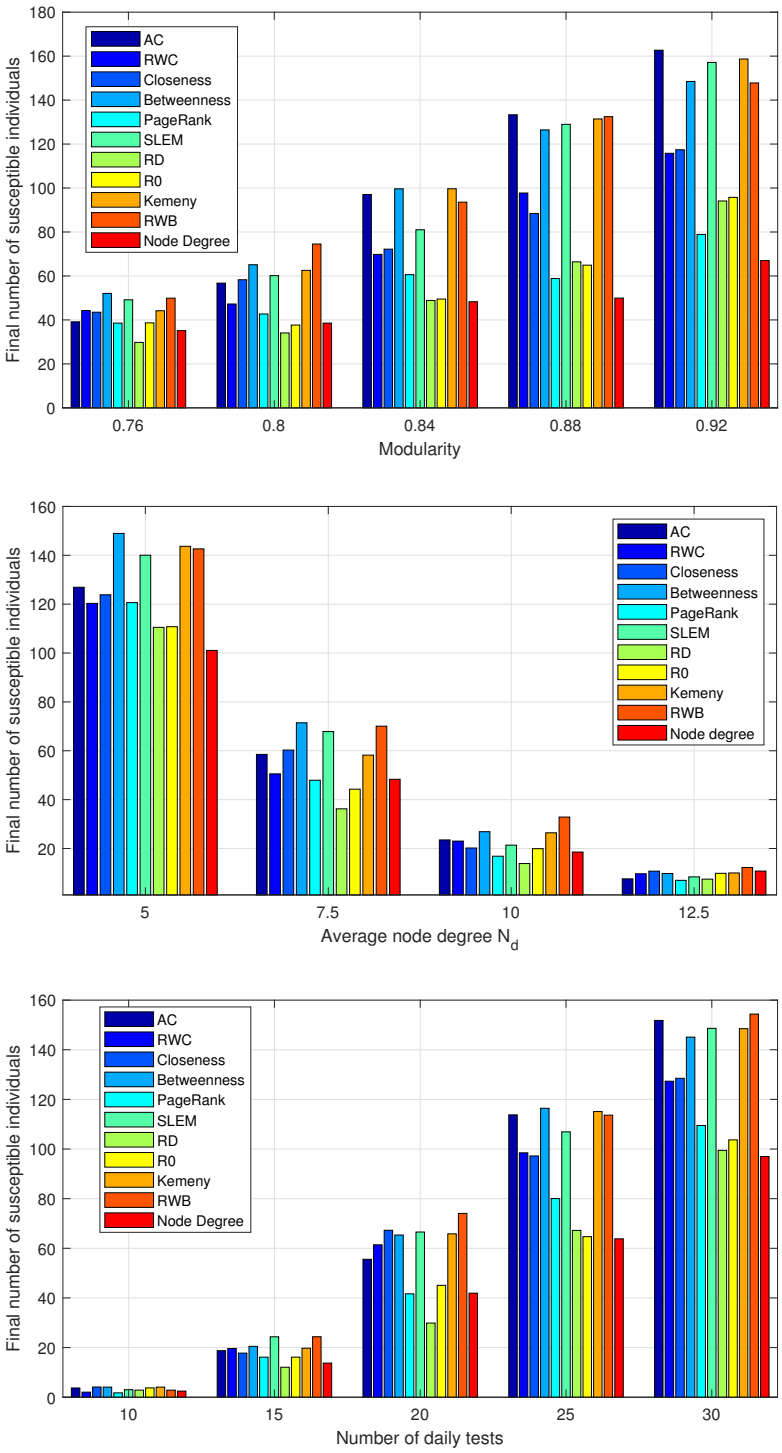

Figure 1. Final number of susceptible individuals at the end of the simulation, for networks of different modularities (above), different average node degree (middle), and for different values of daily tests.

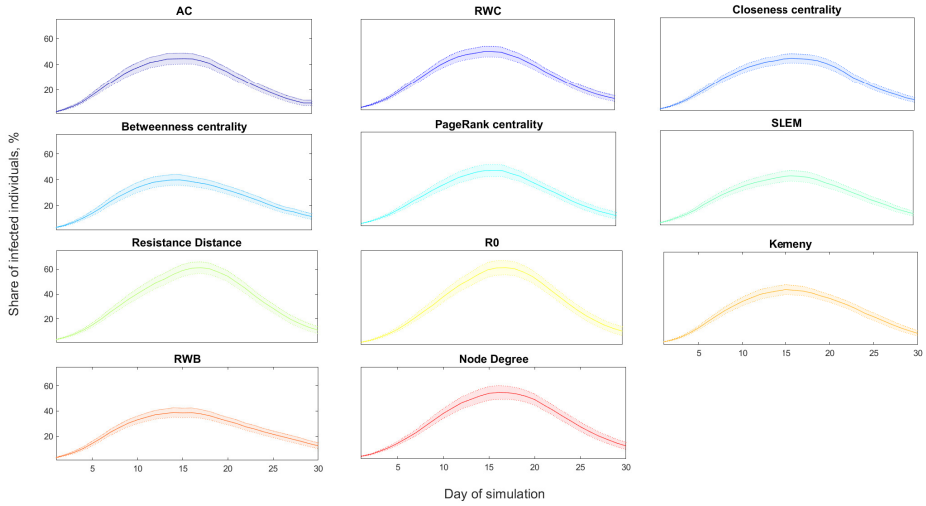

Figure 2. Daily number of infected individuals for different indicators. The shaded area corresponds to 95% confidence interval.

4. Case study: simulation on weighted graphs

The primary contribution of this article is the survey and comparison of various node ranking procedures in the control of disease spread in a contact network. However, throughout the review in Section II of existing centrality indicators, we have stressed the importance of considering metrics which generalize to weighted or asymmetric (i.e., directed) graphs. This is motivated in no small part by the nature of the current pandemic and the emphasis on increased risk of transmission caused by a lengthier interaction, or a contact event with no personal protective equipment. In this section, we present a small example to motivate further consideration of this issue.

In Newman's paper (see Newman, 2005), the idea of random walk betweenness was introduced, but first motivated by the example shown in Fig. 3. This example indicates that while a node may not sit on any shortest path, intuition would still indicate that it is still somehow more "central" than others. Finding existing measures of centrality (which depended largely on shortest paths) falling short in appropriately categorizing such nodes, Newman introduced the idea of random walk betweenness. Inspired by this key example, and its natural appeal to the reader's intuition, we present a comparable example here which highlights the importance of considering the weights of connections between nodes; first by an intuitive argument, then backed up by computations.

In Fig. 4, we have a network consisting of 10 nodes, and edges of different weights.

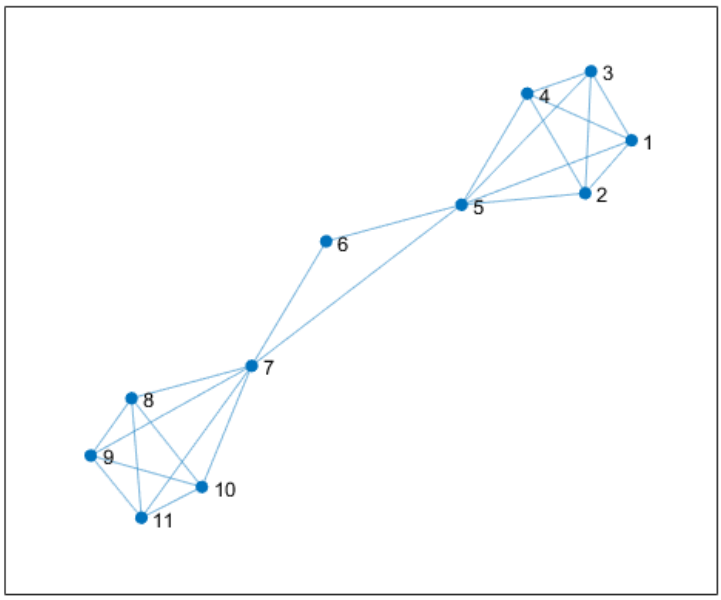

Figure 3. Newman's classic example to indicate the shortcomings of shortest-path centrality measures; node 6 receives the lowest centrality ranking.

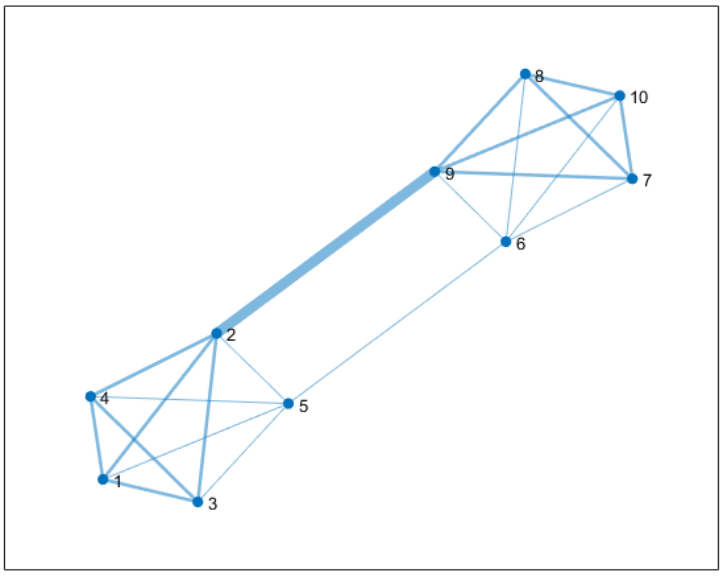

Figure 4. A weighted network on 10 vertices, where weights correspond to varying probabilities of infection between individuals.

Indicator	Average computation time (seconds)	Computation method
PageRank centrality	3.3	Matlab function (centrality, 'pagerank')
Closeness centrality	3.5	Matlab function (centrality, 'closeness')
Node degree	3.6	Matlab function (degree)
Betweenness centrality	3.8	Matlab function (centrality, 'betweenness')
Random walk centrality (RWC)	8.3	Based on generalized inverse (Theorem 1.1 in (a) (Kirkland, 2016))
Algebraic connectivity (AC)	102.5	Second smallest eigenvalue of Laplacian matrix
$\mathcal{R}_0$	111.2	Highest eigenvalue of adjacency matrix
Resistance distance RD	154.8	Based on eigenvalues of Laplacian matrix - Equation (6)
Kemeny's constant	179.5	Based on generalized inverse
SLEM	194.8	Second largest eigenvalue modulus of a Markov chain transition matrix
Random walk betwenness (RWB)	944.0	Eq. (5) (the complete algorithm to compute it is described in Newman, 2005, section 2.2)

Table 1. Computational burden to compute each indicator. Also, for replicability purposes, we explain which equations, or which Matlab function, has been used to compute it.

The weights are given by the following matrix:

$$W = \begin{bmatrix} 0 & 0.3 & 0.3 & 0.3 & 0.1 & 0 & 0 & 0 & 0 & 0 \\ 0.3 & 0 & 0.3 & 0.3 & 0.1 & 0 & 0 & 0 & 0.95 & 0 \\ 0.3 & 0.3 & 0 & 0.3 & 0.1 & 0 & 0 & 0 & 0 & 0 \\ 0.3 & 0.3 & 0.3 & 0 & 0.1 & 0 & 0 & 0 & 0 & 0 \\ 0.1 & 0.1 & 0.1 & 0.1 & 21 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.1 & 0 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0 & 0 & 0 & 0 & 0 & 0.1 & 0 & 0.3 & 0.3 & 0.3 \\ 0 & 0 & 0 & 0 & 0 & 0.1 & 0.3 & 0 & 0.3 & 0.3 \\ 0 & 0.95 & 0 & 0 & 0 & 0.1 & 0.3 & 0.3 & 0 & 0.3 \\ 0 & 0 & 0 & 0 & 0 & 0.1 & 0.3 & 0.3 & 0.3 & 0 \end{bmatrix}.$$

For our purposes, we will interpret the edge weight w_{ij} as the probability of transmission from individual i to j (or vice versa) based on their contact time and conditions, if one of them is already infected. Note that in Fig. 4, the weights are represented proportionally by thinner or thicker lines. It is a reasonable argument that based on these infection probabilities, nodes 2 and 9 should play a larger role in the spread of the disease through the entire community than nodes 5 and 6. However, if we were to assume that the graph is unweighted, or that all infection probabilities are equal, then we would assume that all four nodes (2, 5, 6, and 9) have equal importance. If testing/quarantine strategies are based on this assumption, we may fail to prevent or slow the spread of the disease.

For the remainder of this section, we analyse this example in a few ways. First, we compare the rankings of the nodes of this example according to each centrality indicator from the previous section, where we consider the network to be unweighted (i.e. all weights of contacts/edges are either 0 or 1). Then we will compute rankings when the weights of the edges are taken into account. To do so, we will need to elaborate further on how these will be computed in the weighted case. Finally, we run some simulations to compare the effectiveness of each indicator (both unweighted and weighted) in controlling the disease. These simulations differ from those in the previous section since this network is too small for those same simulations to be sensible.

For unweighted indicators (node degree, betweenness centrality, random walk betweenness, second-largest eigenvalue, and Kemeny's constant), we compute as before, taking the weight of every edge to be 1, and working with either the $0-1$ adjacency matrix A, or the simple random walk on the graph with transition matrix $D^{-1}A$. For weighted indicators, we make a few adjustments:

- For weighted node degree, we use W as the weighted adjacency matrix, and determine the weight of node i as the sum of the weights of incident edges, or the sum of row i.
- In computing betweenness centrality of a node in a weighted graph, typically the edge-weights correspond to a "cost", and minimum distance corresponds to paths or walks of minimal cost. As such, it does not make sense in this context to compute betweenness centrality using infection probabilities as weights, since if the value of w_{ij} is high, it is *easier* for the disease to spread, not harder (which we would associate with high cost). We choose to simply replace the weight w_{ij} of the edge between i and j by the cost $1 - w_{ij}$, and then compute betweenness centrality for each node in the usual way.
- There is no known formulation of random walk betweenness for weighted graphs, so we do not include a weighted version here.
- For the second-largest eigenvalue of a probability transition matrix, and for Kemeny's constant, there are a few choices we could make for how to represent a random walk on the weighted network. One option is to simply normalize the rows of the matrix W, producing a stochastic matrix. However, in the disease-spread context, this may not be appropriate. Instead, to account for individuals with less-risky contacts overall, we implement an adjusted random walk, in which for certain nodes, there is a nontrivial probability that the random walker stays in place in the next time-step (in our application, this would correspond to assuming that an infectious individual will not infect another individual). To do this, we find the maximum row-sum of W, and normalize the entire matrix by that value. Then we replace the diagonal entry in each row by the difference between 1 and the i^{th} row-sum to produce a stochastic matrix. See matrix P

$$P = \begin{bmatrix} 0.487 & 0.154 & 0.154 & 0.154 & 0.051 & 0 & 0 & 0 & 0 & 0 \\ 0.154 & 0 & 0.154 & 0.154 & 0.051 & 0 & 0 & 0 & 0.487 & 0 \\ 0.154 & 0.154 & 0.487 & 0.154 & 0.051 & 0 & 0 & 0 & 0 & 0 \\ 0.154 & 0.154 & 0.154 & 0.487 & 0.051 & 0 & 0 & 0 & 0 & 0 \\ 0.051 & 0.051 & 0.051 & 0.051 & 0.744 & 0.051 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.051 & 0.744 & 0.051 & 0.051 & 0.051 & 0.051 \\ 0 & 0 & 0 & 0 & 0 & 0.051 & 0.487 & 0.154 & 0.154 & 0.154 \\ 0 & 0 & 0 & 0 & 0 & 0.051 & 0.154 & 0.487 & 0.154 & 0.154 \\ 0 & 0.487 & 0 & 0 & 0 & 0.051 & 0.154 & 0.154 & 0 & 0.154 \\ 0 & 0 & 0 & 0 & 0 & 0.051 & 0.154 & 0.154 & 0.154 & 0.487 \end{bmatrix}. \tag{9}$$

in Eq. (9) for the stochastic matrix representing the adjusted random walk for this example with these weights. The advantage of this normalization is that all contacts having the same duration are eventually associated with the same probability of spreading the virus in the transition matrix, while without the diagonal correction this is not guaranteed to occur, due to the normalization steps required to make transition matrices row-stochastic. With this transition matrix in hand, we compute the weighted versions of these indicators using this representation of the Markov chain. For the sake of comparison, we also compute Kemeny's constant for the weighted random walk (without the diagonal correction).

Due to the symmetry of the graph, some nodes of this network are associated with the same values of criticality. Accordingly, such nodes can be grouped and analyzed together. In particular, there are at most three groups: $\{2, 9\}$, $\{5, 6\}$ and $\{1, 3, 4, 7, 8, 10\}$, where nodes in the same group receive the same rank. The results are reported in Table 2 and can be summarized as follows:

- All unweighted indicators group $\{2, 5, 6, 9\}$ together as being equally the most critical nodes, while the remaining nodes rank second.
- Both weighted node degree (ND W) and the weighted version of Kemeny's constant (Kemeny WA) for the adjusted random walk rank $\{2, 9\}$ as most critical, $\{1, 3, 4, 7, 8, 10\}$ next, and $\{5, 6\}$ as least critical.
- The weighted betweenness centrality (BC W) measure ranks $\{2, 9\}$ as the most central nodes in the weighted network, and weights all other nodes equally.
- The second-largest eigenvalue measure for the adjusted random walk (SLEM W) ranks $\{2, 9\}$ as most critical, then $\{5, 6\}$, then the remaining nodes $\{1, 3, 4, 7, 8, 10\}$. Interestingly, Kemeny's constant for the weighted random walk without the diagonal correction (Kemeny W) ranks the nodes the same way.

Intuition and visual inspection of the graph are sufficient to determine that unweighted indicators fail to properly rank nodes because nodes $\{2, 9\}$ are more critical than $\{5, 6\}$ (the virus is spread more likely from one community to the other through the $\{2, 9\}$-link as the duration of their contact is longer than that of $\{5, 6\}$). Indeed, all indicators based on the weighted graph reach a consensus on indicating nodes $\{2, 9\}$ as the most critical. Conversely, there is a discrepancy on which one should be the second most critical set of nodes (i.e., Kemeny's constant indicates $\{1, 3, 4, 7, 8, 10\}$, while the weighted betweenness centrality measure indicates $\{5, 6\}$).

Accordingly, we run disease-spread simulations with this network so to establish

Ranking strategy	First group	Second group	Third group
Correct ranking	2,9	1,3,4,7,8,10	5,6
Unweighted indicators	2,5,6,9	1,3,4,7,8,10	-
Kemeny W	2,9	5,6	1,3,4,7,8,10
Kemeny WA	2,9	1,3,4,7,8,10	5,6
ND W	2,9	1,3,4,7,8,10	5,6
SLEM W	2,9	5,6	1,3,4,7,8,10
BC W	2,9	1,3,4,5,6,7,8,10	-

Table 2. Nodes ranking on the weighted graph

which one is the correct response. As mentioned already, this network is too small to obtain meaningful results from the heuristic outlined in the previous section. Instead, we do the following: For each node, remove it from the network, infect one other node chosen at random, and run an infection simulation according to the probabilities given in W. Take note of the number of days it takes for the disease to spread to all the remaining 9 nodes in the network. Run this same simulation 1000 times and record the average number of days for the disease to spread to the entire network. Thus for each node in the network, we have a simulated measure of how "critical" it is in the spread of disease through the network - the higher the number of days for the disease to spread with node i removed, the more critical node i must be to the spread in the underlying network shown in Fig. 4. The average numbers of days are shown for each node as follows:

Node	Avg. days
1	7.55
2	17.41
3	7.39
4	7.39
5	6.84
6	6.83
7	7.63
8	7.63
9	17.37
10	7.39

Note that in 1000 simulations on the full weighted network, the average number of days for the disease to spread to all ten nodes is 7.86 days. Removing some nodes might either speed up or slow down the spreading of the disease. We consider the outcome of this simulation as the evidence of which nodes are most critical, and we report it accordingly in Table 2.

4.1. Discussion of the Results for the Weighted Case

Table 2 shows through a simple example several interesting results:

- Unweighted indicators fail to work properly in weighted graphs. Thus, indicators that cannot be directly used in weighted graphs should not be used when weights are known and available;
- Some unweighted indicators may be actually used in the weighted case as well, but they may not be as effective as they used to be in the unweighted case, see

for instance betweenness centrality in Table 2;

- The indicator based on the Kemeny's constant (and the transition matrix with diagonal scaling) appears to work equally well for weighted and unweighted networks, thanks to its ability to identify bridges of communities, which are indeed critical in spreading a virus in a population Yilmaz et al., 2020. The previous comments suggest that indicators like the Kemeny's constant should be used to analyse weighted networks. Conversely, the usage of unweighted indicators corresponds to neglecting potentially useful information that drives the spreading of the infection (e.g., different duration of contacts that leads to different probabilities of infection).

5. Conclusions

Inspired by the recent problems arising in the context of the COVID-19 pandemic, and most notably in terms of who should be tested, and who should be vaccinated first, this manuscript reviews some of the most popular ranking methodologies to identify the importance of nodes in networks of individuals. While the dynamics of the COVID-19 have been simplified, still it is possible to observe that significantly different results may be obtained if different ranking methods are adopted to select the most suitable individuals for tests or vaccination. In particular, three main parameters were considered to assess the actual effectiveness of the indicators, i.e., the network modularity, the average number of daily contacts and the number of available tests. The simulations showed that for many combinations of these parameters, indicators like the algebraic connectivity, betweenness centrality, second largest eigenvalue modulus, Kemeny's constant and random walk betweenness, may actually be twice as effective as other indicators in abating the number of infected individuals (i.e., with respect to PageRank or node degree).
The comparison becomes even more interesting under the assumption that duration of contacts may be measured and shared. In this case, this information can be encapsulated as a weight in edges of the graph, but unweighted indicators, i.e., indicators based on the corresponding unweighted graph, fail to take into consideration this piece of information and its impact in the virus spread. In addition, indicators that appeared to perform very well in the unweighted case, most notably, betweenness centrality and the second largest eigenvalue, fail to correctly rank nodes in the proposed simple network, which is a weighted revisitation of the classic Newman's network. Other indicators that explicitly take into account the (weighted) transition matrix, as Kemeny's constant appear to be the most suitable in correctly ranking the nodes.

Funding

J.M. acknowledges support of the OP RDE funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics". S.Y. acknowledges support by TUBITAK (The Scientific and Technological Research Council of Turkey). S.K. acknowledges research support by NSERC Discovery Grant RGPIN-2019-05408. J.B. acknowledges research support by NSERC Discovery Grant RGPIN-2021-03775. M.B., P.F., R.M.-S., T.P., R.S. & S.S. acknowledge support from EPSRC project EP/V018450/1. M.B. and T.P. also acknowledge funding support from the European Union's Horizon 2020 Research and Innovation Programme under

Grant Agreement No 739551 (KIOS CoE), E.C. & E.D acknowledge support by the Italian Ministry of Education, Universities and Research through the Research Project PRIN 2017 "Advanced Network Control of Future Smart Grids" (project VECTORS).

References

Aldous, D., & Fill, J. A. (2002). Reversible Markov Chains and Random Walks on Graphs [Unfinished monograph, recompiled 2014, available at http://www.stat.berkeley.edu/~aldous/RWG/book.html].

Amaku, M., Covas, D. T., Coutinho, F. A. B., Neto, R. S. A., Struchiner, C., Wilder-Smith, A., & Massad, E. (2021). Modelling the test, trace and quarantine strategy to control the COVID-19 epidemic in the state of São Paulo, Brazil. *Infectious Disease Modelling*, *6*, 46–55.

Barabási, A.-L. (2016). *Network Science*. Cambridge University Press.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, *101*(11), 3747–3752. https://doi.org/10.1073/pnas.0400087101

Bastani, H., Drakopoulos, K., Gupta, V., Vlachogiannis, I., Hadjicristodoulou, C., Lagiou, P., Magiorkinis, G., Paraskevis, D., & Tsiodras, S. (2021). Efficient and targeted COVID-19 border testing via reinforcement learning. *Nature*, *599*(7883), 108–113.

Bin, M., Cheung, P. Y., Crisostomi, E., Ferraro, P., Lhachemi, H., Murray-Smith, R., Myant, C., Parisini, T., Shorten, R., Stein, S., et al. (2021). Post-lockdown abatement of COVID-19 by fast periodic switching. *PLoS Computational Biology*, *17*(1), e1008604.

Borgatti, S. P. (2006). Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory*, *12*(1), 21–34.

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, *10*(1), 1–14.

Breen, J., Crisostomi, E., Faizrahnemoon, M., Kirkland, S., & Shorten, R. (2018). Clustering behaviour in Markov chains with eigenvalues close to one. *Linear Algebra and its Applications*, *555*, 163–185. https://doi.org/10.1016/j.laa.2018.06.013

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

Castellano, C., & Pastor-Satorras, R. (2017). Relating topological determinants of complex networks to their spectral properties: Structural and dynamical effects. *Physical Review X*, *7*(4), 041024.

Choi, W., & Shim, E. (2021). Optimal strategies for social distancing and testing to control COVID-19. *Journal of theoretical biology*, *512*, 110568.

Chung, F. R. K., & Graham, F. C. (1997). *Spectral Graph Theory*. American Mathematical Soc.

Crisostomi, E., Kirkland, S., & Shorten, R. (2011). A Google-like model of road network dynamics and its application to regulation and control. *International Journal of Control*, *84*(3), 633–651. https://doi.org/10.1080/00207179.2011.568005

de Abreu, N. M. M. (2007). Old and new results on algebraic connectivity of graphs. *Linear Algebra and its Applications*, *423*(1), 53–73. https://doi.org/10.1016/j.laa.2006.08.017

Doyle, P. G. (2009). The Kemeny constant of a Markov chain. *arXiv:0909.2636 [math]*. Retrieved May 17, 2021, from http://arxiv.org/abs/0909.2636
Comment: Version 1.0 dated 14 September 2009; GNU FDL

Doyle, P. G., & Snell, J. L. (1984). *Random Walks and Electric Networks* (Vol. 22). American Mathematical Soc.

Fiedler, M. (1989). Laplacian of graphs and algebraic connectivity. *Banach Center Publications*, *25*(1), 57–70. Retrieved May 17, 2021, from https://eudml.org/doc/267812

Fouss, F., Saerens, M., & Shimbo, M. (2016). *Algorithms and Models for Network Data and Link Analysis*. Cambridge University Press.

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, *40*(1), 35–41. https://doi.org/10.2307/3033543

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239. https://doi.org/10.1016/0378-8733(78)90021-7

Ghalmane, Z., El Hassouni, M., & Cherifi, H. (2019). Immunization of networks with non-overlapping community structure. *Social Network Analysis and Mining*, *9*(1), 1–22.

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., & Colaneri, M. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in italy. *Nature Medicine*, *26*(6), 855–860.

Gómez, S. (2019). Centrality in Networks: Finding the Most Important Nodes. In P. Moscato & N. J. de Vries (Eds.), *Business and Consumer Analytics: New Ideas* (pp. 401–433). Springer International Publishing. https://doi.org/10.1007/978-3-030-06222-4_8

Goodman-Meza, D., Rudas, A., Chiang, J. N., Adamson, P. C., Ebinger, J., Sun, N., Botting, P., Fulcher, J. A., Saab, F. G., Brook, R., et al. (2020). A machine learning algorithm to increase COVID-19 inpatient diagnostic capacity. *PLoS ONE*, *15*(9), e0239474.

Hartfiel, D. J., & Meyer, C. D. (1998). On the structure of stochastic matrices with a subdominant eigenvalue near 1. *Linear Algebra and its Applications*, *272*(1), 193–203. https://doi.org/10.1016/S0024-3795(97)00333-9

Hufnagel, L., Brockmann, D., & Geisel, T. (2004). Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences*, *101*(42), 15124–15129. https://doi.org/10.1073/pnas.0308344101

Johnson, B. C., & Kirkland, S. (2019). Estimating random walk centrality in networks. *Computational Statistics & Data Analysis*, *138*, 190–200. https://doi.org/10.1016/j.csda.2019.04.009

Kamp, C., Moslonka-Lefebvre, M., & Alizon, S. (2013). Epidemic spread on weighted networks. *PLoS Computational Biology*, *9*(12), e1003352.

Kemeny, J., & Snell, L. (1960). *Finite Markov Chains*. Van Nostrand.

Kirkland, S. (2016). Random walk centrality and a partition of Kemeny's constant. *Czechoslovak Mathematical Journal*, *66*(3), 757–775. https://doi.org/10.1007/s10587-016-0291-9

Klein, D. J., & Randić, M. (1993). Resistance distance. *Journal of Mathematical Chemistry*, *12*(1), 81–95. https://doi.org/10.1007/BF01164627

Koeth, F. (2019). Graphs, Networks and Power Systems. Retrieved May 17, 2021, from https://www.incite-itn.eu/blog/graphs-networks-and-power-systems/

Langville, A. N., & Meyer, C. D. (2006). *Google's PageRank and Beyond - the science of search engine rankings*. Princeton University Press.

Levene, M., & Loizou, G. (2002). Kemeny's Constant and the Random Surfer. *The American Mathematical Monthly*, *109*(8), 741–745. https://doi.org/10.1080/00029890.2002.11919905

Lyng, G. D., Sheils, N. E., Kennedy, C. J., Griffin, D. O., & Berke, E. M. (2021). Identifying optimal COVID-19 testing strategies for schools and businesses: Balancing testing frequency, individual test technology, and cost. *PLoS ONE*, *16*(3), e0248783.

Manabe, Y. C., Sharfstein, J. S., & Armstrong, K. (2020). The need for more and better testing for COVID-19. *JAMA*, *324*(21), 2153–2154.

Mei, W., Mohagheghi, S., Zampieri, S., & Bullo, F. (2017). On the dynamics of deterministic epidemic propagation over networks. *Annual Reviews in Control*, *44*, 116–128. https://doi.org/10.1016/j.arcontrol.2017.09.002

Newman, M. E. J. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, *27*(1), 39–54. https://doi.org/10.1016/j.socnet.2004.11.009

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113

Noh, J. D., & Rieger, H. (2004). Random Walks on Complex Networks. *Physical Review Letters*, *92*(11), 118701. https://doi.org/10.1103/PhysRevLett.92.118701

Oldham, S., Fulcher, B., Parkes, L., Arnatkevičiūtė, A., Suo, C., & Fornito, A. (2019). Consistency and differences between centrality measures across distinct classes of networks. *PLoS ONE*, *14*(7), e0220061.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, *32*(3), 245–251. https://doi.org/10.1016/j.socnet.2010.03.006

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web* (Technical Report No. 1999-66) [Previous number = SIDL-WP-1999-0120]. Stanford InfoLab. Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/

Rahimi, I., Chen, F., & Gandomi, A. H. (2021). A review on COVID-19 forecasting models. *Neural Computing and Applications*, 1–11.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2021). Characterizing the interactions between classical and community-aware centrality measures in complex networks. *Scientific Reports*, *11*(1), 1–15.

Restrepo, J. G., Ott, E., & Hunt, B. R. (2006). Characterizing the dynamical importance of network nodes and links. *Physical Review Letters*, *97*(9), 094102.

Ronqui, J. R. F., & Travieso, G. (2015). Analyzing complex networks through correlations in centrality measurements. *Journal of Statistical Mechanics: Theory and Experiment*, *2015*(5), P05030. https://doi.org/10.1088/1742-5468/2015/05/P05030

Salathé, M., Althaus, C. L., Neher, R., Stringhini, S., Hodcroft, E., Fellay, J., Zwahlen, M., Senti, G., Battegay, M., Wilder-Smith, A., et al. (2020). COVID-19 epidemic in switzerland: On the importance of testing, contact tracing and isolation. *Swiss medical weekly*, (11).

Salathé, M., & Jones, J. H. (2010). Dynamics and Control of Diseases in Networks with Community Structure. *PLOS Computational Biology*, *6*(4), e1000736. https://doi.org/10.1371/journal.pcbi.1000736

Sporns, O. (2018). Graph theory methods: Applications in brain networks. *Dialogues in Clinical Neuroscience*, *20*(2), 111–121.

Wang, H., Hernandez, J. M., & Van Mieghem, P. (2008). Betweenness centrality in a weighted network. *Physical Review E*, *77*(4), 046105. https://doi.org/10.1103/PhysRevE.77.046105

Wang, X., Dubbeldam, J. L. A., & Van Mieghem, P. (2017). Kemeny's constant and the effective graph resistance. *Linear Algebra and its Applications*, *535*, 231–244. https://doi.org/10.1016/j.laa.2017.09.003

Wang, Y., Chakrabarti, D., Wang, C., & Faloutsos, C. (2003). Epidemic spreading in real networks: An eigenvalue viewpoint. *22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings.*, 25–34.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

White, D. R., & Borgatti, S. P. (1994). Betweenness centrality measures for directed graphs. *Social Networks*, *16*(4), 335–346. https://doi.org/10.1016/0378-8733(94)90015-9

Wu, C. W. (2005). Algebraic connectivity of directed graphs. *Linear and Multilinear Algebra*, *53*(3), 203–223. https://doi.org/10.1080/03081080500054810

Yan, J., He, H., & Sun, Y. (2014). Integrated Security Analysis on Cascading Failure in Complex Networks. *IEEE Transactions on Information Forensics and Security*, *9*(3), 451–463. https://doi.org/10.1109/TIFS.2014.2299404

Yilmaz, S., Dudkina, E., Bin, M., Crisostomi, E., Ferraro, P., Murray-Smith, R., Parisini, T., Stone, L., & Shorten, R. (2020). Kemeny-based testing for COVID-19. *PLOS ONE*, *15*(11), e0242401. https://doi.org/10.1371/journal.pone.0242401

Yu, Y., Berger-Wolf, T. Y., Saia, J., et al. (2008). Finding spread blockers in dynamic networks. *International Workshop on Social Network Mining and Analysis*, 55–76.

Zhang, Z., Shan, T., & Chen, G. (2013). Random walks on weighted networks. *Physical Review E*, *87*(1), 012112.