Leveraging percolation theory to single out influential spreaders in networks

Filippo Radicchi and Claudio Castellano

Phys. Rev. E **93**, 062314 — Published 22 June 2016

DOI: 10.1103/PhysRevE.93.062314

Leveraging percolation theory to single out influential spreaders in networks

Filippo Radicchi
Center for Complex Networks and Systems Research,
*School of Informatics and Computing, Indiana University, Bloomington, USA**

Claudio Castellano
Istituto dei Sistemi Complessi (ISC-CNR), Via dei Taurini 19, 00185 Roma, Italy,
and Dipartimento di Fisica, Sapienza Università di Roma, Roma, Italy

Among the consequences of the disordered interaction topology underlying many social, technological and biological systems, a particularly important one is that some nodes, just because of their position in the network, may have a disproportionate effect on dynamical processes mediated by the complex interaction pattern. For example, the early adoption by an opinion leader in a social network may change the fate of a commercial product, or just a few super-spreaders may determine the virality of a meme in social media. Despite many recent efforts, the formulation of an accurate method to optimally identify influential nodes in complex network topologies remains an unsolved challenge. Here, we present the exact solution of the problem for the specific, but highly relevant, case of the Susceptible-Infected-Removed (SIR) model for epidemic spreading at criticality. By exploiting the mapping between bond percolation and the static properties of SIR, we prove that the recently introduced Non-Backtracking centrality is the optimal criterion for the identification of influential spreaders in locally tree-like networks at criticality. By means of simulations on synthetic networks and on a very extensive set of real-world networks, we show that the Non-Backtracking centrality is a highly reliable metric to identify top influential spreaders also in generic graphs not embedded in space, and for noncritical spreading.

I. INTRODUCTION

Social, technological and biological systems are often characterized by underlying interaction topologies with complex features [1, 2]. In a complex network, the roles played by individual nodes are highly heterogeneous. Understanding the impact of individual vertices on the global functionality of the system is one of the most fundamental, yet not fully solved, problems of network science. Centrality measures have indeed the purpose of quantitatively gauging the importance of individual vertices [3]. Among the most natural and used ones are degree, betweenness centrality [4], k-shell (or k-core) index [5], and eigenvector centrality [6].

Spreading is at the root of a vast class of phenomena occurring on network substrates: the propagation of contagious diseases [7], the diffusion of information or memes [8], the adoption of innovations [9], etc. A large interest has been recently devoted to the identification of influential spreaders (often called super-spreaders), i.e., nodes that, if chosen as initiators, maximize the extent of a spreading process. The goal is to identify which of the many centrality metrics, that can be computed using only topological information, is most strongly correlated with the ability of a node to originate massive spreading events.

Probably, a fully universal method, able to perfectly single out the most influential nodes for arbitrary spreading dynamics on arbitrary networks, does not exist. It is in fact reasonable to expect that the predictive power of the different centralities strongly depends not only on the topology of the underlying network but also on the details of the spreading process. Numerical evidence in this sense can be found in [10, 11]. A much more reasonable goal is instead to identify a metric able to optimally solve the problem for specific types of dynamics. Here, we take this path and concentrate our attention on the Susceptible-Infected-Removed (SIR) model for epidemics. SIR is a paradigmatic model for spreading, and the vast majority of the investigations about the identification of influential spreaders in complex networks have dealt with it. In random networks, classical results on the SIR model relate the epidemic threshold to moments of the degree distribution [7]. Hence, a naive hypothesis is to assume that the spreading ability is strongly correlated to the degree of the initiator. This view has been challenged by Kitsak *et al.*, who proposed the k-core (also called k-shell) index (which singles out nodes belonging to dense, mutually interconnected, subgraphs) as a proper indicator of the spreading ability [12]. This seminal paper has been followed by an avalanche of other studies aimed at investigating the issue for the same or different dynamical processes, synthetic or real-world networks, using a wide range of centralities proposed as predictors of the spreading ability of the different vertices [11, 13–22]. Many empirical investigations have casted doubts on the ability of the k-shell index to identify influential spreaders in various topologies [16, 20]. However in a very recent work, Ferraz de Arruda *et al.* have reaffirmed the superiority of the k-shell index and degree centrality as predictors for top influencial spreaders in nonspatial networks [11]. The authors

*Electronic address: filiradi@indiana.edu

of this paper proposed also an additional centrality metric, the so-called generalized random walk accessibility, to overcome limitations of the k-shell index in spatially embedded networks. The picture emerging from all these efforts is not satisfactory: All heuristics proposed are motivated based on physical intuition but involve uncontrolled approximations; No exact result is available even for synthetic idealized but nontrivial topologies. Methods are generally validated numerically on a very limited number of networks, with no complete control of their topological properties. In this paper we fill this gap, presenting a physically grounded method which solves exactly the problem in a nontrivial case, and performs very well in a very broad spectrum of situations.

Our work is based on the connection existing between bond percolation and the static properties of the SIR model for epidemics [7, 23, 24]. Very recent results have pointed out the crucial role played by the spectral properties of the Non-Backtracking (NB) matrix in determining the properties of the bond percolation process in complex networks [25–27]. Combining these two well established facts, we therefore propose the NB centrality [28] as the quantity of choice for the identification of influential spreaders in disordered topologies. In particular we show that, on locally tree-like networks, the NB centrality provides the exact solution to the problem of finding the best single influential spreader, if critical spreading is considered. We complement this result with a thorough empirical investigation of the problem on a very large set of real-world topologies, of social, technological and biological origin, exhibiting a large variety of size, sparsity, heterogeneity and other topological features. We compare the performance, as predictors of the spreading power of single nodes, of the most important centralities proposed so far. We show that NB centrality turns out to be, in the majority of cases, the best quantity able to single out the most influential initiators of spreading processes in networks.

II. THE PROBLEM OF INFLUENTIAL SPREADERS

A. The spreading dynamics

To model the spreading dynamics, we consider the SIR model, the simplest and most studied dynamics for epidemics in the presence of acquired immunity [7]. Each vertex of a network can be either in state S (susceptible), I (infected) or R (interpreted either as recovered or removed). We consider the continuous time version of the dynamics. At each instant of time, two elementary events may occur: (i) $\mathrm{I} \xrightarrow{\nu} \mathrm{R}$, meaning that, at rate ν, a spontaneous recovery/removal event may turn a node in state I into state R; (ii) $\mathrm{I} + \mathrm{S} \xrightarrow{\beta} 2\,\mathrm{I}$, indicating the spreading of the infection, at rate β, among pairs of connected nodes in states I and S. Starting from an initial configuration where all vertices are in the state S and only node i is in state I, a connected set of contiguous vertices may be infected, but after some time all infected nodes eventually switch to the R state and the outbreak ends. The total number Q_i of nodes whose final state is R represents the extent of the spreading event originated by the single seed i. The problem of interest here is the identification of influential spreaders, i.e., finding, based only on the topology of the network, what node of the network must be selected as an initiator of the epidemics in order to maximize the average outbreak size. The asymptotic behavior of the SIR model depends on the ratio $\lambda = \beta/\nu$. If initiators are randomly chosen, then one can define a critical threshold λ_c. For λ smaller than the epidemic threshold λ_c, spreading events are of finite (subextensive) size. For $\lambda > \lambda_c$ instead the infection involves a finite fraction of the whole system. It is therefore reasonable to expect that also the identity and role of influential spreaders depends on the value of λ. See the Appendix for info about how λ_c is determined numerically.

B. Numerical simulations

For a given network, we rank the nodes on the basis of their spreading power. We numerically simulate the SIR dynamical process with a single initial seed i in state I and all other nodes in state S. After the dynamics has ended, we record the number Q_i of nodes in state R. We then repeat the procedure 10^4 times, and quantify the spreading power of node i as $\langle Q_i \rangle$, that is the average size of the outbreak generated from the initial seed i. The measure $\langle Q_i \rangle$ and its associated ranking are the benchmarks against which we compare the centralities proposed to identify influential spreaders. We consider four standard centrality metrics: degree, k-core, eigenvector centrality, and the generalized Random Walk Accessibility (RWA). Eigenvector centrality has been indicated as an effective predictor within mean field analyses, i.e. neglecting dynamical correlations between states of adjacent vertices [14]. RWA has been recently identified by Ferraz de Arruda *et al.* as the best predictor for influential spreaders in spatial networks [11]. Quantitative comparisons of performance among centrality metrics are based on two complementary measures: the imprecision function ϵ [12], and the Jaccard distance d_J. Both measures take as input two sets of nodes. The first is the list of the first ρN actual top spreaders, with $0 < \rho \leq 1$ and N size of the network, as identified from the results of numerical simulations of the SIR model, hence ranked on the basis of the score $\langle Q_i \rangle$. The second set is the list of ρN top nodes when nodes are ranked according to the centrality score x_i. Both ϵ and d_J return a value ranging between 0, for perfect matching (i.e. the centrality x perfectly predicts the spreading influence of the fraction ρ of top spreaders) and 1, for completely failed prediction. The complementarity between the two measures of performance is apparent from their definitions (see Ap-

pendix). The Jaccard distance measures the difference among the "identity" of the nodes included in the sets of true and predicted ρN top influencers. The imprecision function is completely insensitive to the identity of the nodes, and is determined instead only by their spreading power.

III. RESULTS

A. Exact solution on locally tree-like networks: Non-Backtracking centrality

The Hashimoto or Non-Backtracking (NB) matrix is a special representation of the structure of a network [29]. In an arbitrary undirected and unweighted network with E edges, the NB matrix is a $2E \times 2E$ array defined as follows. Every edge $i \leftrightarrow j$ is split in two directed edges $i \to j$ and $j \to i$. The generic entry of the NB matrix is $M_{i\to j,l\to m} = \delta_{j,l}(1 - \delta_{i,m})$, where $\delta_{i,j}$ is the Kronecker symbol. $M_{i\to j,l\to m}$ is different from zero and equals one only if the edges $i \to j$ and $l \to m$ define a non-backtracking path of length two. M is an asymmetric matrix with real and positive principal eigenvalue. The components of the principal eigenvector, namely $v_{i\to j}$, can be used to define the NB centrality of vertex i as [28]

$$n_i = \sum_j A_{ij} v_{i\to j}. \tag{1}$$

This centrality is similar to the common eigenvector centrality, but it disregards the contribution of vertex i to the centrality of its neighbors, thus avoiding the self-reinforcement effect responsible in some cases for the localization of the eigenvector centrality [28, 30]. We remark that the computation of the principal eigenpair of the matrix M can be performed using a simple power-iteration method. This allows to estimate the NB centrality of all nodes in a time that scales as $\mathcal{O}(E)$. The Ihara-Bass determinant formula may be further used to reduce memory storage in the computation of the NB centrality [31].

The NB matrix has been shown to play a crucial role in the problem of graph-clustering [32] and, more recently, in percolation [25, 33]. In particular, the percolation threshold in locally tree-like networks is exactly given by the inverse of the largest eigenvalue of the NB matrix for both bond and site percolation [25, 27, 33]. As a consequence, the probability that node i is part of the percolating cluster immediately above the threshold is given by the expression of Eq. (1).

The mapping between the static properties of the SIR model and bond percolation [7, 24, 35] reveals that SIR epidemic outbreaks coincide with the clusters of the associated bond percolation process, where the bond occupation probability p for percolation and the effective spreading rate for SIR are related by $p = \lambda/(1+\lambda)$. This connection has a very important consequence: the relative size of an epidemic outbreak started from a specific node i is proportional to the probability that i belongs to the percolating cluster. At the critical point, this probability coincides with the NB centrality, thus

$$\langle Q_i \rangle \propto n_i. \tag{2}$$

As a consequence the top spreaders are the vertices with the highest NB centrality. Note that this is an exact result, provided the network structure is locally tree-like [1]. In Fig. 1 we test numerically the validity of this connection in synthetic networks. Panel (a) confirms that Eq. (2) is generally well obeyed and tends to be more and more precise as the system size grows, thus making the network more and more locally tree-like. Panel (b) shows instead that the values of the other node centralities have a lower degree of proportionality with the average outbreak size initiated by them.

The exact equivalence between SIR outbreak size and NB centrality holds only at criticality, i.e. for $\lambda_c = p_c/(1-p_c)$. As we depart from λ_c the equivalence becomes less accurate. The probability of belonging to the percolating cluster is no more equal to n_i. Moreover, below λ_c the largest cluster does not dominate the cluster size distribution of percolation. We stress however that criticality is the regime where the identification of

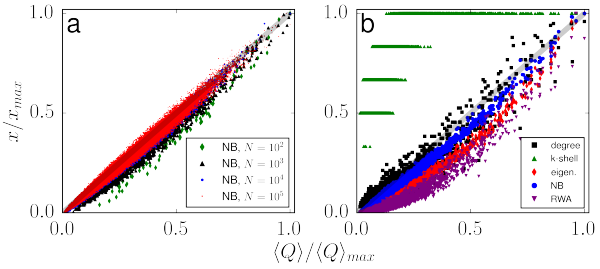

Figure 1: Impact of individual nodes at criticality. The scatter plot shows the relation between the predicted impact x_i and the actual impact in simulations $\langle Q_i \rangle$. Both measures are divided by their maximal values to obtain numbers in the interval $[0, 1]$. Predicted impact is determined on the basis of the centrality score associated to every node. Panel (a) refers to an Erdős-Rényi (ER) graph with average degree $\langle k \rangle = 4$ and varying size N. Only NB centrality is considered. Panel (b) refers to a scale-free graph with $N = 10^4$ nodes constructed according to the uncorrelated configuration model [34]. The degree sequence is composed of integer numbers selected randomly from a probability distribution $P(k) \sim k^{-\gamma}$ defined over the interval $[3, \sqrt{N}]$. We consider here $\gamma = 2.5$.

[1] The mapping is strictly valid for a SIR model where the recovery time is fixed. For the SIR version considered in simulations here, the recovery time is nondegenerate and this implies that the mapping is not strictly exact [36–38].

influential spreaders really matters: The further we move away from the critical point, the less interesting and non-trivial the problem becomes. For large values of λ in the supercritical regime, any seed will lead to large outbreaks involving a very large portion of the entire network. In the deeply subcritical regime instead, at very low values of $\lambda > 0$, all spreading events involve a very small neighborhood of the initial seed. Only around criticality the choice of the initiator may have substantial impact on the spreading event, i.e., whether the spreading phenomenon remains confined to a few nodes or it reaches an extensive fraction of the network.

B. Top spreaders in synthetic networks

We now test the implications of the results in the previous section for spreading on locally tree-like synthetic networks. We consider a network with degree distribution decaying as $P(k) \sim k^{-\gamma}$, with $\gamma = 3.5$, and compare the performance of the various centralities as predictors of the top spreaders in the network. In Fig. 2 we plot the two dissimilarity measures ϵ and d_J, for the various centralities, as a function of the fraction ρ of top-ranking nodes. The imprecision function ϵ provides a very clear picture: the outbreaks started in nodes with highest NB centrality are of the same size as those initiated by the best influential spreaders in numerical simulations. The degree, the eigenvector centrality, the generalized random walk accessibility, and, most markedly, the k-shell index perform much worse. The plot for d_J gives a similar message, with the difference that the measure does not vanish for the NB centrality. This last observation can be understood by considering that the NB centralities of distinct nodes do not differ much (see Fig. 1). Therefore, it is likely that small uncertainties in the values of $\langle Q_i \rangle$ calculated numerically may considerably alter the ranking of the nodes, leading to a nonvanishing Jaccard dissimilarity. For the very same reason, the numerical uncertainties have no appreciable effect on the imprecision function ϵ, which is very close to 0.

If the same analysis is repeated for other values of the exponent γ or for Erdős-Rényi networks, a very similar phenomenology is found [39]: NB centrality outperforms eigenvector centrality and generalized random walk accessibility in the identification of influential spreaders. Degree and k-core centrality still deliver poor performances.

We conclude that NB centrality is the optimal choice for the selection of influential spreaders on locally tree-like networks at criticality. The same considerations extend to the subcritical and supercritical regimes, provided that λ is not too far away from the critical point [39].

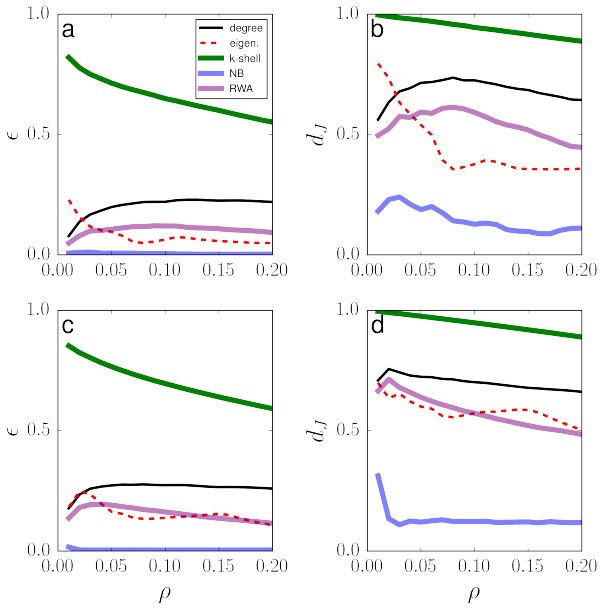

Figure 2: Identification of influential spreaders in Scale-Free (SF) graphs at criticality. The imprecision function ϵ (panels a and c) and the Jaccard distance d_J (panels b and d) are plotted against the fraction of top nodes ρ. Relative performance of the various centrality measures as a function of ρ can be deduced from the direct comparison among the curves: the lower is the value of the dissimilarity metrics, the better the centrality measure predicts true top spreaders. Results are obtained on the largest connected component of SF graphs, constructed according to the uncorrelated configuration model [34]. The pre-imposed degree sequence is composed of random integer numbers selected randomly from a probability distribution $P(k) \sim k^{-\gamma}$ defined over the interval $[3, \sqrt{N}]$. We consider the case $\gamma = 3.5$, and two distinct network sizes: $N = 10^4$ (panels a and b) and $N = 10^5$ (panels c and d). Numerical simulations of the SIR model are performed at the critical values of λ ($\lambda_c = 0.216$ in panels a and b, and $\lambda_c = 0.212$ in panels c and d).

C. Top spreaders in real-world networks

As substrates for the spreading dynamics, we now consider a very large collection of real-world topologies of diverse origin, size and topological features. Many of these networks have a sizeable clustering coefficient, so that they cannot be considered, even approximately, as tree-like. We analyze a total of 95 networks. Details can be found in [39]. In Fig. 3, we present the results for two such networks: a graph of email contacts [12], and the Gnutella peer-to-peer network [40]. It turns out very clearly that, for these structures, k-shell centrality and degree perform very badly; RWA performs slightly better, but still poorly; eigenvector and NB centralities are instead very effective in identifying influential spreaders. Among these two, NB centrality provides a slightly more effective recipe for the identification of top influential spreaders. The picture obtained for synthetic net-

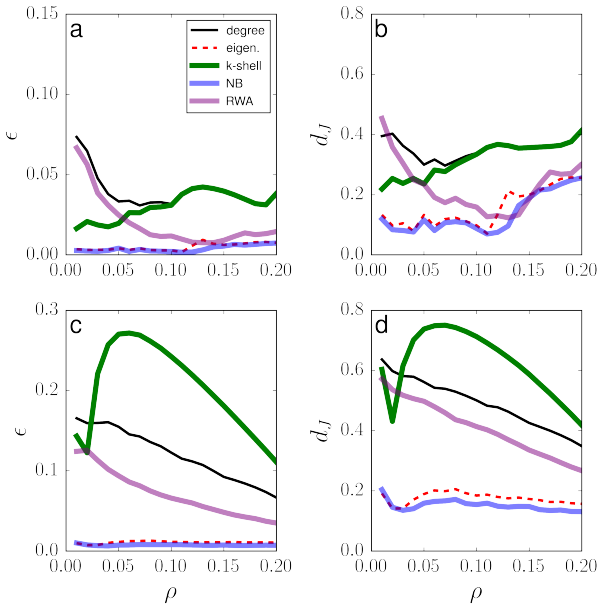

Figure 3: Identification of influential spreaders in real-world graphs at criticality. The description of the various panels is similar to those of Fig. 2. In panels a and b, we present results for the Email contact network [12]. In panels c and d, we show the results for the peer-to-peer network of Gnutella as of August 31, 2002 [40]. The clustering coefficients of the networks are 0.1088 and 0.0055, respectively. SIR simulations are performed at criticality, with $\lambda = \lambda_c$. The epidemic thresholds of the two networks are $\lambda_c = 0.031$ and $\lambda_c = 0.099$, respectively.

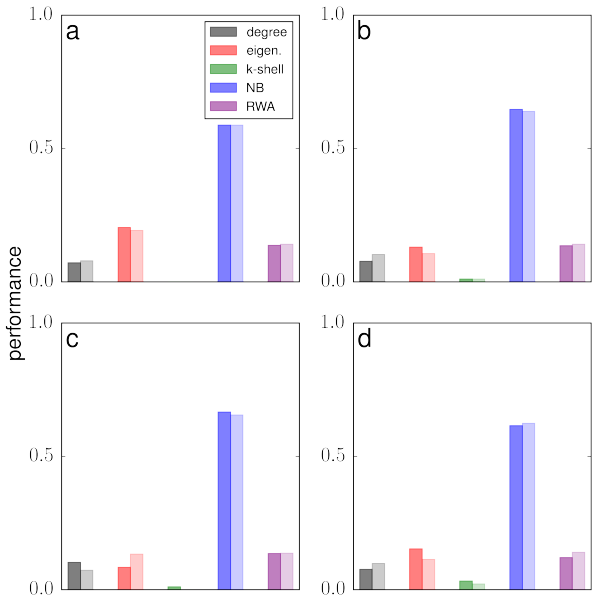

Figure 4: Comparison of predictive power of the different centralities in nonspatially embedded real-world graphs at criticality. The bars indicate the fraction of real networks where each centrality provides the best prediction for the top ρN influential spreaders. For every network in our sample, we determine the best centrality measure as the metric generating the minimal value of the imprecision function (dark-shaded bars) or the Jaccard distance (light-shaded bars) at predetermined values of ρ. We consider $\rho = 0.05$ (a), $\rho = 0.10$ (b), $\rho = 0.15$ (c), and $\rho = 0.20$ (d). In the case of ties, the score is equally split among the top metrics.

works is then essentially confirmed. We have repeated the same analysis for a very large set of networks with nonspatial embedding [39]. The set of networks include graphs of different nature (e.g., biological, technological, social) thus with large variability in their topological features (e.g., degree distribution, size, clustering coefficient). Whereas some variation exists depending on the detailed topology, overall the message is clear: the NB centrality of a node is, in about 60% of the networks analyzed, the most accurate predictor of the spreading ability of individual nodes (Fig. 4). NB centrality outperforms all other centrality metrics in real nonspatial networks. Only in graphs with spatial embedding RWA provides better performances [39].

Previous results are obtained for critical spreading, where the susceptibility of the system is maximal. We repeat the analysis for the subcritical regime by setting $\lambda = 2/3\,\lambda_c$, and the supercritical phase for $\lambda = 3/2\,\lambda_c$ [39]. The overall picture is again similar to the one observed for critical spreading: the NB centrality is in the majority of cases the most accurate predictor to identify influential spreaders.

IV. CONCLUSIONS

The present analysis provides convincing evidence that the centrality determined from the Non-Backtracking (NB) matrix of a graph represents the best predictor for the identification of SIR influential spreaders in the network. The choice of this centrality measure is motivated by recent theoretical progress in the study of percolation processes in arbitrary locally tree-like graphs, and by the equivalence between the SIR model and the bond percolation model at criticality. Even in real networks, where the locally tree-like ansatz is violated, NB centrality turns out to greatly outperform other centrality metrics in the task of identifying top influential spreaders. We remark also that NB centrality can be computed in a time that scales almost linearly with the system size, and it is thus applicable to very large networks.

An additional, interesting, result emerging from our systematic analysis of real networks is that k-shell centrality generally provides very unsatisfactory performances, not only compared to NB centrality, but also to degree, eigenvector centrality and generalized random walk accessibility. This is at odds with what claimed in the seminal paper by Kitsak and collaborators [12], and

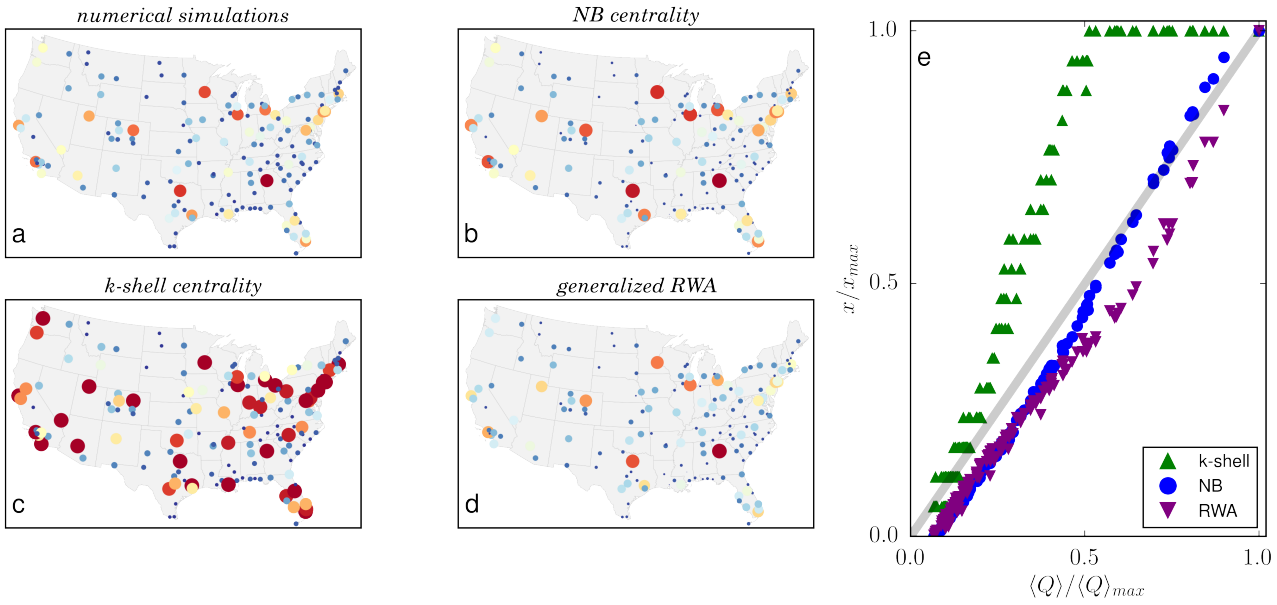

Figure 5: Influential spreaders in the US air transportation network at criticality. We consider the unweighted and undirected version of the air transportation network within the US, reconstructed by aggregating the information of all flights by major carriers (American Airlines, Delta and United) in January 2014 [41, 42]. Panel a provides a visualization of the impact of individual airports in the spreading process. The impact of airport i is measured as $\langle Q_i\rangle/\langle Q\rangle_{max}$, with $\langle Q\rangle_{max}$ maximal value of the spreading power over all airports. In the representation, the size of the circles is proportional to the impact of the corresponding airport. Colors of the circles are also proportional to the spreading power of the airport, with a continuous scale ranging from red (maximal impact) to blue (minimal impact). In panels b and c, we provide the same representation as in panel a, but replacing the spreading power with the value of their centrality. In panel d, we provide a scatter plot the centrality metrics against the value of the spreading power for individual airports. The grey line is the benchmark for perfect performance. Whereas NB centrality and generalized RWA provide a distinction for airports, k-core centrality generates identical scores for many airports.

more recently remarked by Ferraz de Arruda *et al.* [11], with the analysis of very small samples of real-world networks. Given the amount of real-world graphs considered in our study, we believe that our message is conclusive: k-shell index can be easily outperformed by other simple centrality metrics in the identification of influential nodes in dynamical processes on complex networks. One of the main reasons of the poor ability of the k-core to identify top spreaders is rooted in the very definition of k-core index, which necessarily involves a large degeneracy [16, 43]. The metric is not able to make a distinction among top vertices in the ranking, since, by definition, k nodes must be tied at the top position if k is the maximal value of the k-shell centrality measured in a network. This fact is clearly illustrated for artificial graphs in Fig. 1, where the k-core index is the same for very large groups of vertices, whereas their spreading power is highly heterogeneous. Similar considerations are valid also for real graphs. In Fig. 5, we consider SIR on a substrate given by an air transportation network within the US [41]. The spreading influence of individual nodes is well reproduced by the NB centrality and, more approximately, by the RWA. According to the k-shell centrality, several airports are ranked at the top of the list. The top tier is, however, composed of airports with fundamentally different values of their spreading power: for example, "Hartsfield-Jackson Atlanta International Airport", the actual top spreader in the network, is tied with "Wilmington International Airport", despite the latter actually has a spreading power twice smaller than the top spreader.

Beyond their applicability to relevant real situations, these results open new exciting perspectives for other, related, problems. A first question is the validity of the NB centrality solution for other types of spreading dynamics, different from the SIR class. While NB centrality is unlikely to perform well for rumor dynamics [10, 11], the question is open for more complex modeling frameworks for epidemics, such as metapopulations [44, 45]. Secondly, the problem studied here refers to individual spreaders. Substantially different results may arise in the case of optimal multiple spreaders, i.e. the identification of the subset of network vertices (of a given number of nodes) maximizing the extent of a spreading process seeded in all of them at the same time. As already noted in Ref. [12], starting the process in the best single spreaders often results in suboptimal propagation, because of the overlap among the areas of influence of the best individual spreaders. Finding the best set of multiple spreaders is a different, highly nontrivial, NP-complete opti-

mization problem [46], for which many clever approximation schemes have been proposed [46–48], but a scalable and accurate general approach is still not available. The insights provided by the mapping to percolation and the consideration of the NB centrality may pave the way for further progress also in this context. Another exciting line of research regards the identification of influential spreaders from empirical data on real-world spreading phenomena [49, 50]. In this respect, the problem is further complicated by the fact that the spreading dynamics at the microscopic level are not known a priori and may contain additional ingredients not included in the simple models usually considered.

Acknowledgments

FR acknowledges support from the National Science Foundation (Grant CMMI-1552487) and the US Army Research Office (W911NF-16-1-0104).

Appendix A: Measures of performance

1. The imprecision function

The imprecision function [12] $\epsilon(\rho)$ quantifies the difference between the average size of the spreading events initiated (as single spreaders) by the first ρN vertices according to a given centrality and the analogous size for the actual ρN most efficient spreaders in SIR simulations. N is the number of nodes in the network and $0 < \rho \leq 1$. More in detail, let us define as $\Upsilon^{(x)}(\rho)$ the set of the top ρN vertices according to the centrality x and $\Upsilon^{(eff)}(\rho)$ the actual top ρN spreaders, as measured in SIR simulations. The quantity

$$Z^{(x)}(\rho) = \frac{1}{N\rho} \sum_{i \in \Upsilon^{(x)}(\rho)} \langle Q_i \rangle \tag{A1}$$

is the average size of outbreaks originated in the most highly ranked nodes according to the centrality x. If $Z^{(eff)}(\rho)$ is the same quantity as in Eq. (A1) but computed over the set $\Upsilon^{(eff)}(\rho)$, the imprecision function is defined as

$$\epsilon^{(x)}(\rho) = 1 - \frac{Z^{(x)}(\rho)}{Z^{(eff)}(\rho)} \tag{A2}$$

If the centrality x perfectly identifies the most efficient spreaders, the imprecision function equals zero. High values of $\epsilon^{(x)}(\rho)$ indicate that the centrality is not a good predictor of the spreading power of the top ρN spreaders. To account for possible ties in the centrality metric x, we average the imprecision function over at least 10 realizations of the set $\Upsilon^{(x)}(\rho)$.

2. The Jaccard distance

The Jaccard distance $d_J(\rho)$ is a measure of the dissimilarity between two sets $\Upsilon^{(x)}(\rho)$ and $\Upsilon^{(eff)}(\rho)$. This quantity is defined as

$$d_J^{(x)}(\rho) = 1 - \frac{|\Upsilon^{(x)}(\rho) \cap \Upsilon^{(eff)}(\rho)|}{|\Upsilon^{(x)}(\rho) \cup \Upsilon^{(eff)}(\rho)|} \tag{A3}$$

where $|A|$ stands for the number of elements in the set A. Clearly, if the two sets $\Upsilon^{(x)}(\rho)$ and $\Upsilon^{(eff)}(\rho)$ coincide the distance vanishes, while if they have null intersection, then their distance equals one.

Appendix B: Centrality measures

We consider the following centrality measures.

- Degree centrality. This is the simplest centrality measure that can be defined for nodes in a network. The degree of node i equals the number of neighbors of vertex i in the network.
- k-shell centrality. A k-core is a subset of nodes composed of vertices that have at least k neighbors within the set itself. The k-shell or k-core index of a node equals the largest k value of k-cores which the node belongs to.
- Eigenvector centrality. The score assigned to each node equals the value of the component of the principal eigenvector of the adjacency matrix of the network.
- The score of node i based on the generalized Random Walk Accessibility (RWA) is defined as $\alpha_i = \exp\left(-\sum_j W_{i,j} \ln W_{i,j}\right)$, where $W_{i,j} = \sum_{q=0}^{\infty} (P^q)_{i,j}/q!$, with P^q qth power of the random walk transition matrix of the graph [11]. The exact computation of the RWA score for all nodes in the network requires the diagonalization of the matrix P, an unfeasible task for medium- and large-size networks. Good approximations of RWA scores can be obtained by means of agent-based simulations of the random walk dynamics. Our estimates of RWA are based on average values obtained over 10^6 independent walks of maximal length 20 for every node in the network.

Appendix C: Numerical determination of the epidemic thresholds

For a given network, we determine the critical value λ_c in the following way. For a given value of λ, we start from a configuration where all nodes are in state S, and one randomly chosen vertex is in state I. We run the SIR model, and measure the size of the outbreak Q. We

repeat the procedure $100,000$ times, every time choosing at random a node as initial seed of the epidemics, and compute the first and second moment of the size of the outbreak, namely $\langle Q \rangle$ and $\langle Q^2 \rangle$. The critical value value λ_c is determined from the position of the peak of the ratio $\langle Q^2 \rangle / \langle Q \rangle^2$ [51]. Values of λ_c for all networks analyzed in this paper are reported in [39].

[1] R. Albert and A.-L. Barabási, Reviews of Modern Physics **74**, 47 (2002).
[2] M. Newman, *Networks: an introduction* (OUP Oxford, 2010).
[3] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*, vol. 8 (Cambridge university press, 1994).
[4] L. C. Freeman, Sociometry **40**, 35 (1977).
[5] S. B. Seidman, Social Networks **5**, 269 (1983).
[6] P. Bonacich, Journal of Mathematical Sociology **2**, 113 (1972).
[7] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Rev. Mod. Phys. **87**, 925 (2015).
[8] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer, in *Proceedings of the 20th international conference companion on World wide web* (ACM, 2011), pp. 249–252.
[9] E. Bakshy, B. Karrer, and L. A. Adamic, in *Proceedings of the 10th ACM conference on Electronic commerce* (ACM, 2009), pp. 325–334.
[10] J. Borge-Holthoefer and Y. Moreno, Phys. Rev. E **85**, 026116 (2012).
[11] G. F. de Arruda, A. L. Barbieri, P. M. Rodríguez, F. A. Rodrigues, Y. Moreno, and L. d. F. Costa, Phys. Rev. E **90**, 032812 (2014).
[12] M. Kitsak, L. Gallos, S. Havlin, L. Liljeros, F. an d Muchnik, H. Stanley, and H. Makse, Nature Physics **6**, 888 (2010).
[13] F. Bauer and J. T. Lizier, EPL (Europhysics Letters) **99**, 68007 (2012).
[14] K. Klemm, M. Á. Serrano, V. M. Eguíluz, and M. San Miguel, Scientific reports **2**, 292 (2012).
[15] D. Chen, L. Lu, M.-S. Shang, Y.-C. Zhang, and T. Zhou, Physica A: Statistical Mechanics and its Applications **391**, 1777 (2012).
[16] R. A. P. da Silva, M. P. Viana, and L. da Fontoura Costa, Journal of Statistical Mechanics: Theory and Experiment **2012**, P07005 (2012).
[17] D.-B. Chen, R. Xiao, A. Zeng, and Y.-C. Zhang, EPL (Europhysics Letters) **104**, 68006 (2013).
[18] J.-G. Liu, Z.-M. Ren, and Q. Guo, Physica A: Statistical Mechanics and its Applications **392**, 4154 (2013).
[19] Z.-M. Ren, A. Zeng, D.-B. Chen, H. Liao, and J.-G. Liu, EPL (Europhysics Letters) **106**, 48005 (2014).
[20] Y. Liu, M. Tang, T. Zhou, and Y. Do, Scientific reports **5**, 9602 (2015).
[21] Y. Liu, M. Tang, T. Zhou, and Y. Do, Scientific reports **5**, 13172 (2015).
[22] J.-G. Liu, J.-H. Lin, Q. Guo, and T. Zhou, Scientific reports **6**, 21380 (2016).
[23] P. Grassberger, Mathematical Biosciences **63**, 157 (1983).
[24] M. E. J. Newman, Phys. Rev. E **66**, 016128 (2002).
[25] B. Karrer, M. E. J. Newman, and L. Zdeborová, Phys. Rev. Lett. **113**, 208702 (2014).
[26] K. E. Hamilton and L. P. Pryadko, Phys. Rev. Lett. **113**, 208701 (2014).
[27] F. Radicchi and C. Castellano, Nature communications **6**, 10196 (2015).
[28] T. Martin, X. Zhang, and M. E. J. Newman, Phys. Rev. E **90**, 052808 (2014).
[29] K. Hashimoto, Adv. Stud. Pure Math. **15**, 211 (1989).
[30] R. Pastor-Satorras and C. Castellano, Scientific Reports **6**, 18847 (2016).
[31] H. Bass, Int. J. Math. **3**, 717 (1992).
[32] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborov, and P. Zhang, Proceedings of the National Academy of Sciences **110**, 20935 (2013).
[33] K. E. Hamilton and L. P. Pryadko, Phys. Rev. Lett. **113**, 208701 (2014).
[34] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras, Phys. Rev. E **71**, 027103 (2005).
[35] P. Grassberger, Math. Biosci. **63**, 157 (1983).
[36] L. A. Meyers, M. E. J. Newman, and B. Pourbohloul, Journal of theoretical biology **240**, 400 (2006).
[37] J. C. Miller, Phys. Rev. E **76**, 010101 (2007).
[38] E. Kenah and J. M. Robins, Phys. Rev. E **76**, 036113 (2007).
[39] See Supplemental Material at [...] for results of the analysis of network models and real graphs in the subcritical and critical regimes. The Supplemental Material contains also the list of all real-world networks included in our analysis.
[40] J. Leskovec, J. Kleinberg, and C. Faloutsos, ACM Transactions on Knowledge Discovery from Data (TKDD) **1**, 2 (2007).
[41] *Bureau of transportation statistics*, `http://www.transtats.bts.gov`, accessed: 2015-01-18.
[42] F. Radicchi, Nature Physics **11**, 597 (2015).
[43] S. Pei and H. A. Makse, Journal of Statistical Mechanics: Theory and Experiment **2013**, P12002 (2013).
[44] B. Grenfell and J. Harwood, Trends in ecology & evolution **12**, 395 (1997).
[45] V. Colizza and A. Vespignani, Journal of Theoretical Biology **251**, 450 (2008).
[46] D. Kempe, J. Kleinberg, and E. Tardos, in *Proceedings of the ninth ACM SIGKDD international conferenc e on Knowledge discovery and data mining* (ACM, New York, NY, USA, 2003), KDD '03, pp. 137–146.
[47] F. Altarelli, A. Braunstein, L. Dall'Asta, and R. Zecchina, Journal of Statistical Mechanics: Theory and Experiment **2013**, P09011 (2013).
[48] F. Morone and H. A. Makse, Nature **524**, 65 (2015).
[49] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno, Scientific reports **1**, 197 (2011).
[50] S. Pei, L. Muchnik, J. S. Andrade Jr, Z. Zheng, and H. A. Makse, Scientific reports **4**, 5547 (2014).
[51] R. Pastor-Satorras and C. Castellano, (submitted) (2016).