Approximating nonbacktracking centrality and localization phenomena in large networks

G. Timár,* R. A. da Costa, S. N. Dorogovtsev, and J. F. F. Mendes
Departamento de Física da Universidade de Aveiro & I3N, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal

(Received 16 March 2021; revised 18 October 2021; accepted 28 October 2021; published 18 November 2021)

Message-passing theories have proved to be invaluable tools in studying percolation, nonrecurrent epidemics, and similar dynamical processes on real-world networks. At the heart of the message-passing method is the nonbacktracking matrix, whose largest eigenvalue, the corresponding eigenvector, and the closely related nonbacktracking centrality play a central role in determining how the given dynamical model behaves. Here we propose a degree-class-based method to approximate these quantities using a smaller matrix related to the joint degree-degree distribution of neighboring nodes. Our findings suggest that in most networks, degree-degree correlations beyond nearest neighbor are actually not strong, and our first-order description already results in accurate estimates, particularly when message-passing itself is a good approximation to the original model in question, that is, when the number of short cycles in the network is sufficiently low. We show that localization of the nonbacktracking centrality is also captured well by our scheme, particularly in large networks. Our method provides an alternative to working with the full nonbacktracking matrix in very large networks where this may not be possible due to memory limitations.

DOI: 10.1103/PhysRevE.104.054306

I. INTRODUCTION

A lot of recent scientific effort has been aimed at understanding how dynamical processes running on top of complex networks are affected by the underlying network structure. A large and relevant subset of dynamical processes can be accurately approximated by the message-passing method (also known as the cavity method), where it is assumed that the contribution of a node j to the behavior of a neighboring node i is completely determined by the contribution of the neighbors of j, excluding i. This approximation is appropriate to study percolation [1,2] and spreading processes where a node may be activated at most once, as is the case in the SIR (susceptible, infected, recovered or removed) model of nonrecurrent epidemics [3]. Message-passing methods disregard backtracking propagation, therefore their linearized version is described by the nonbacktracking (or Hashimoto) matrix [4,5] instead of the adjacency matrix. The nonbacktracking (NB) matrix $\mathbf{H}$ is a $2L \times 2L$ nonsymmetric matrix (L being the number of links in the network) whose elements are indexed by directed links $i \leftarrow j$ instead of nodes. It is defined as $H_{i\leftarrow j,k\leftarrow l} = \delta_{j,k}(1-\delta_{i,l})$, where δ is the Kronecker symbol. In Ref. [6] it was shown that message-passing equations treat any finite loopy network as a well-defined infinite locally treelike network that preserves all local structures of the original, as seen by a nonbacktracking walker. This structure is encoded in the nonbacktracking matrix of the graph.

The key advantage of the NB matrix compared to the adjacency matrix is that it suffers to a much lesser degree from localization of the eigenvectors on large hubs [5], due to the prohibition of backtracking. This circumstance has made it a useful tool in spectral community detection methods [7,8]. The NB matrix has been used to design optimal percolation and node immunization strategies [9–11], identify influential spreaders [12,13], and estimate the time an epidemic takes to reach individual nodes in a network [14]. The relevant quantity in most of these applications is the nonbacktracking centrality (NBC) of a node, defined as $x_i = \sum_{j\in\mathcal{N}_i} v_{i\leftarrow j}$, where $\mathcal{N}_i$ denotes the set of node i's neighbors, and $v_{i\leftarrow j}$ is the component of the principal eigenvector (PEV) of the NB matrix corresponding to the directed link $i \leftarrow j$. The largest eigenvalue (LEV) of the NB matrix plays the role of an effective "branching number," which determines the percolation or SIR epidemic threshold in the message-passing approximation of these processes [1,3,6]. The NBC of a given node is proportional to the probability of belonging to the giant component (or of being infected in an SIR epidemic) close to the transition threshold. For this reason, it is of great importance to know which nodes have the highest NBC, i.e., what group of nodes contribute most to the PEV of the NB matrix. Localization of the adjacency matrix PEV and its consequences for dynamical models such as the SIS (susceptible, infected, susceptible) epidemic model have been studied in detail; see, for example, Refs. [5,15,16]. In Refs. [17,18] it has been suggested that the PEV of the NB matrix may also become localized, although not on individual hubs, but rather on densely connected small subgraphs such as the highest k-core of the network or a group of "overlapping hubs." Using these findings, an estimate for the LEV of the NB matrix was given (see Ref. [18]), which was found to be a strong improvement over the mean branching $\langle k^2\rangle/\langle k\rangle - 1$.

We explore the possibility of approximating the LEV of the NB matrix and the NBC of nodes in a network considering only nearest-neighbor degree-degree correlations, i.e., substituting a given network with an infinite random network that has a joint degree-degree distribution $P(k,k')$ identical to the original network in question. Such an approximation is

*gtimar@ua.pt

described by a matrix whose number of rows is equal to the number of different degrees in the network, which may be significantly smaller than the number of links. Therefore, if found to be accurate, such a degree-class-based approximation may be useful to estimate the percolation or epidemic threshold and individual NBC values of nodes in cases where the network in question is too large to be easily studied using the full NB matrix. We find that such a degree-based approximation indeed works well, and the relevant matrix is closely related to the branching matrix used in Refs. [19,20]. Additionally we observe that localization of the NBC, quantified by the inverse participation ratio (IPR), is also reproduced fairly accurately in our method, lending more credence to the validity of a degree-based approximation.

II. TWO MATRICES DESCRIBING CORRELATED NETWORKS

We discuss two related matrices that represent the same infinite maximally random network with given nearest-neighbor degree-degree correlations described by the joint degree-degree distribution $P(k,k')$. We will use these matrices in Sec. III to write approximations for the LEV of the NB matrix and the mean NBC of nodes of degree k.

A. The branching matrix: Percolation and SIR epidemics in correlated networks

The branching matrix $\mathbf{B}$, defined as

$$B_{k,k'} = (k'-1)P(k'|k), \tag{1}$$

has been used to study SIR epidemics [19] and percolation [20] in random networks with only nearest-neighbor degree-degree correlations. This matrix emerges by considering the probability y_k that a random edge emanating from a node of degree k leads to a finite component. Using the locally treelike property of an infinite random correlated network, we can write the recursive equation

$$y_k = \sum_{k'} P(k'|k) y_{k'}^{k'-1}, \tag{2}$$

where $P(k'|k)$ is the probability that a randomly chosen link has an end node of degree k' given that the other end node is of degree k. Assuming that $P(k'|k)$ is such that we are close to the percolation threshold, we can write $a_k = 1 - y_k \ll 1$ and keep only terms linear in a_k,

$$a_k = \sum_{k'} (k'-1)P(k'|k)a_{k'}, \tag{3}$$

or in vector form,

$$\mathbf{a} = \mathbf{B}\mathbf{a}, \tag{4}$$

where the matrix $\mathbf{B}$ is the branching matrix defined in Eq. (1). Equation (4), with the Perron-Frobenius theorem, implies that the LEV of matrix $\mathbf{B}$, at the percolation threshold, is $\lambda_1^{(\mathbf{B})} = 1$. Thus the quantity $\lambda_1^{(\mathbf{B})}$ is an effective branching in such correlated networks [20]. From Eq. (4), we also learn that close to the percolation threshold, the probability a_k that a random link emanating from a node of degree k leads to infinity is proportional to v_k, the appropriate component of the PEV of matrix $\mathbf{B}$. The probability that a node of degree k belongs to the giant component is thus proportional to kv_k, and the sum of this probability over all nodes of degree k is proportional to $kP(k)v_k$.

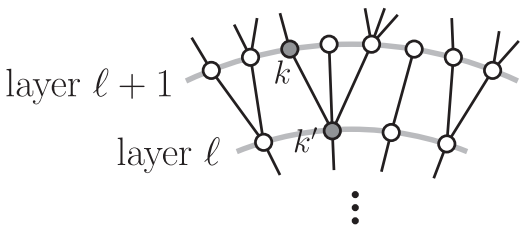

FIG. 1. Schematic representation of two successive layers in the expansion of a random network with nearest-neighbor degree-degree correlations specified by a joint degree-degree distribution $P(k,k')$.

B. The expansion matrix: Nonbacktracking expansion of correlated networks

A different way of obtaining a description of correlated networks is by following the ideas of Ref. [6] according to which the nonbacktracking expansion is defined. Let us build the expanding neighborhood (a tree, layer by layer) of a random node in a network with only nearest-neighbor degree-degree correlations. Below the percolation threshold, when $\lambda_1^{(\mathbf{B})} < 1$, such a construction will always end within a finite number of steps (since all nodes belong to finite components, i.e., they have finite neighborhoods). When $\lambda_1^{(\mathbf{B})} > 1$, however, there is a nonzero probability S that such a construction continues to infinity. (This happens when the randomly chosen starting node belongs to the giant component.) Using the quantities y_k from Sec. II A, we can write S as

$$S = 1 - \sum_k P(k) y_k^k, \tag{5}$$

where $P(k) = (\langle k\rangle/k)\sum_{k'} P(k,k')$ is the degree distribution of the given correlated network. Assuming that we can build an infinite expansion, let us calculate the relative frequency of nodes of degree k on its "boundary" at infinity. In other words, we are interested in the relative frequency of nodes of degree k, i.e., the degree distribution, in an infinite local neighborhood of a correlated network. Note that this is not the same as the distribution of degrees at the end of a random link, which is, in general, $kP(k)/\langle k\rangle$. (The two distributions are identical for uncorrelated networks, but not for correlated ones.) Let $n_k(\ell)$ denote the mean number of nodes of degree k in layer ℓ of the expansion, and let $n(\ell)$ be the mean total number of nodes in layer ℓ (see Fig. 1). [The averages are taken over the randomness of the expansion process, governed by the joint degree-degree distribution $P(k,k')$.]

Relating the quantities of layers ℓ and $\ell+1$, we can write

$$n_k(\ell+1) = \sum_{k'} n_{k'}(\ell)(k'-1)P(k|k'). \tag{6}$$

Introducing $\eta(\ell) = n(\ell+1)/n(\ell)$, we get

$$\eta(\ell) f_k(\ell+1) = \sum_{k'} f_{k'}(\ell)(k'-1)P(k|k'), \tag{7}$$

where $f_k(\ell) = n_k(\ell)/n(\ell)$ is the relative mean number of nodes of degree k on layer ℓ. Assuming that the branching factor $\eta(\ell)$ and the relative mean numbers $f_k(\ell)$ converge to some constants for $\ell \to \infty$, we have the eigenvector equation

$$\eta \mathbf{f} = \mathbf{E}\mathbf{f}, \tag{8}$$

with the *expansion matrix* $\mathbf{E}$ defined as

$$E_{k,k'} = (k'-1)P(k|k'). \tag{9}$$

The Perron-Frobenius theorem implies that η and $\mathbf{f}$ are the LEV and PEV of the expansion matrix, respectively.

The branching and expansion matrices ($\mathbf{B}$ and $\mathbf{E}$) are two different ways of describing the same correlated network. The spectrum and eigenvectors of the two matrices are closely related, as is shown below in Sec. II C.

C. Spectrum and eigenvectors of the two matrices

Let λ and $\mathbf{v}$ be an eigenvalue and corresponding eigenvector of matrix $\mathbf{B}$. Then

$$\lambda v_k = \sum_{k'} B_{k,k'} v_{k'} \tag{10}$$

$$= \sum_{k'} (k'-1)P(k'|k) v_{k'}. \tag{11}$$

Multiplying both sides by $kP(k)$ and using the relation $P(k'|k)kP(k) = P(k|k')k'P(k')$ [21], we get

$$\lambda k P(k) v_k = \sum_{k'} (k'-1)P(k'|k)kP(k)v_{k'} \tag{12}$$

$$= \sum_{k'} (k'-1)P(k|k')k'P(k')v_{k'}. \tag{13}$$

Introducing $\tilde{v}_k = kP(k)v_k$, we finally have

$$\lambda \tilde{v}_k = \sum_{k'} (k'-1)P(k|k')\tilde{v}_{k'} \tag{14}$$

$$= \sum_{k'} E_{k,k'} \tilde{v}_{k'}. \tag{15}$$

This means that the entire spectrum of the two matrices is identical, and if a vector with components v_k is an eigenvector of $\mathbf{B}$ with eigenvalue λ, then the vector with components $kP(k)v_k$ is an eigenvector of $\mathbf{E}$ with the same eigenvalue λ.

III. APPROXIMATING NONBACKTRACKING CENTRALITY

To approximate the NBC values of nodes in a given network, we construct the expansion matrix $\mathbf{E}$ (or the branching matrix $\mathbf{B}$) using the joint degree-degree distribution measured in the original network. To be precise, in a network consisting of L links, we count the number $L(k,k')$ of links connecting nodes of degrees k and k'. The joint degree-degree distribution is then given as $P(k,k') = L(k,k')/L$ if $k = k'$ and $P(k,k') = L(k,k')/(2L)$ if $k \neq k'$. We may be more cautious and attempt to estimate the "actual" joint degree-degree distribution, assuming that the observed network is a single given realization of a certain stochastic generative process. This would, however, lead outside the scope of this paper, therefore we simply count the number of links connecting nodes of given degrees. Consequently, in sparse networks the time complexity of our method is linear in system size.

We saw in Sec. II A that the components $v_k^{(\mathbf{B})}$ of the PEV of matrix $\mathbf{B}$ are proportional (close to the percolation threshold) to the probabilities that a link emanating from a node of degree k leads to the giant component. The components $v_k^{(\mathbf{E})} = kP(k)v_k^{(\mathbf{B})}$ of the PEV of matrix $\mathbf{E}$, on the other hand, are proportional to the sum of probabilities of nodes of degree k belonging to the giant component. In the message-passing scheme, the NBC of a node is proportional to the probability of that node belonging to the giant component. Alternatively, in the nonbacktracking expansion of an arbitrary network, the NBC of a given node is equal to the relative frequency of replicas of that node on the boundary of the expansion at infinity (see Ref. [6]). Similarly, in the expansion of a correlated network (Sec. II B), the relative frequency of nodes of degree k on the boundary at infinity was found to be $v_k^{(\mathbf{E})}$.

The appropriate comparison is, therefore, between $v_k^{(\mathbf{E})}$ and the sum of NBC values of nodes of degree k,

$$v_k^{(\mathbf{E})} \approx \sum_{i:k_i=k} x_i, \tag{16}$$

where x_i is the NBC of node i, k_i denotes the degree of node i, and we assume the normalization $\sum_i x_i = \sum_k v_k^{(\mathbf{E})} = 1$. Equivalently, we can write the approximation

$$\langle x \rangle_k \approx \frac{v_k^{(\mathbf{E})}}{NP(k)} = \frac{k v_k^{(\mathbf{B})}}{N}, \tag{17}$$

where $\langle x \rangle_k$ denotes the mean NBC of nodes of degree k. We make the assumption that a node's NBC is sufficiently well approximated by the mean NBC of its degree class, i.e.,

$$x_i \approx \langle x \rangle_k \approx \frac{v_{k_i}^{(\mathbf{E})}}{NP(k_i)}. \tag{18}$$

This (heterogeneous mean-field) approximation, arrived at by considering the nonbacktracking expansion of a correlated network, is derived more rigorously in the Appendix, using a methodology similar to that of Ref. [22]. We show below, using a varied set of real-world example networks, that Eq. (18) works fairly well for most and exceptionally well in certain cases, particularly when message-passing is itself a valid approximation of percolation-type phenomena.

The LEV of the NB matrix $\mathbf{H}$ plays the role of an effective branching in the given network, and it determines the percolation (or SIR epidemic) threshold in the message-passing theory of these models. In our approximation, we replace a given network with an infinite random network that has the same joint degree-degree distribution $P(k,k') = P(k|k')k'P(k')/\langle k \rangle$ as the original. Such an infinite random network is described by either matrix $\mathbf{E}$ or $\mathbf{B}$. An approximation to the LEV of the NB matrix is therefore simply given as

$$\lambda_1^{(\mathbf{H})} \approx \lambda_1^{(\mathbf{E})} = \lambda_1^{(\mathbf{B})}; \tag{19}$$

see the Appendix for a derivation of this approximation. (If the original network is connected, then both matrices $\mathbf{E}$ and $\mathbf{B}$ are irreducible, which means that their largest eigenvalue is real and positive according to the Perron-Frobenius theorem.)

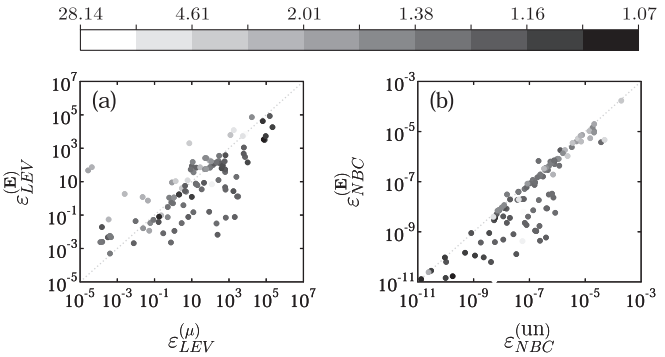

FIG. 2. (a) Error of the expansion matrix approximation to the LEV of the NB matrix as a function of the error of the approximation of Eq. (20). (b) Error of the expansion matrix approximation to the NBC as a function of the error of the local approximation [Eq. (24)]. Each point on the panels shows the errors for one of 109 real-world networks. The errors are defined in Eqs. (22),(23) and (25),(26). Points below the dashed gray lines in panels (a) and (b) have $\varepsilon^{(\mathbf{E})}_{\mathrm{LEV}} < \varepsilon^{(\mu)}_{\mathrm{LEV}}$ and $\varepsilon^{(\mathbf{E})}_{\mathrm{NBC}} < \varepsilon^{(\mathrm{un})}_{\mathrm{NBC}}$, respectively. The color code in both panels corresponds to the ratio $\lambda_1^{(\mathbf{H})}/p_c^{-1}$, indicating the quality of the message-passing approximation to percolation.

From here onwards, we will refer to the approximations of Eqs. (17) and (19) as degree-based or expansion matrix approximations, although they could also be attributed to the branching matrix **B**, as the two matrices contain the same information.

Recently, in Ref. [18], it was shown that the LEV of the NB matrix could be well approximated by the expression

$$\mu = \max(\mu^{\mathrm{un}}, \mu^{\mathrm{oh}}, \mu^{\mathrm{core}}), \tag{20}$$

where

$$\mu^{\mathrm{un}} = \frac{\sum_{ij}(k_i - 1)A_{ij}(k_j - 1)}{\sum_j k_j(k_j - 1)} \tag{21}$$

is an estimate based on the assumption that the network is uncorrelated. (The adjacency matrix is denoted by **A**, and k_i denotes the degree of node i.) The quantities μ^{oh} and μ^{core} are the LEVs associated with the strongest "overlapping hubs" subgraph and the highest k-core, respectively. The reason why these contributions must be dealt with separately is, as pointed out in Ref. [18], that such subgraphs are particularly sensitive to the given correlation patterns in a network, and their contribution is, in general, not included implicitly in μ^{un}. Equation (20) was found to be a significant improvement over the mean branching $\langle k^2\rangle/\langle k\rangle - 1$ in approximating the LEV of the NB matrix.

In Fig. 2(a), we compare our approximation [Eq. (19)] with that of Eq. (20) for 109 real-world networks (see Table I in the Supplemental Material [23]) also considered in Ref. [18], featuring a variety of different sizes, clustering, and correlation patterns. For the comparison, we use, as a measure of the approximation error, the squared distances from the actual LEV of the NB matrix,

$$\varepsilon^{(\mathbf{E})}_{\mathrm{LEV}} = \left(\lambda_1^{(\mathbf{H})} - \lambda_1^{(\mathbf{E})}\right)^2, \tag{22}$$

$$\varepsilon^{(\mu)}_{\mathrm{LEV}} = \left(\lambda_1^{(\mathbf{H})} - \mu\right)^2. \tag{23}$$

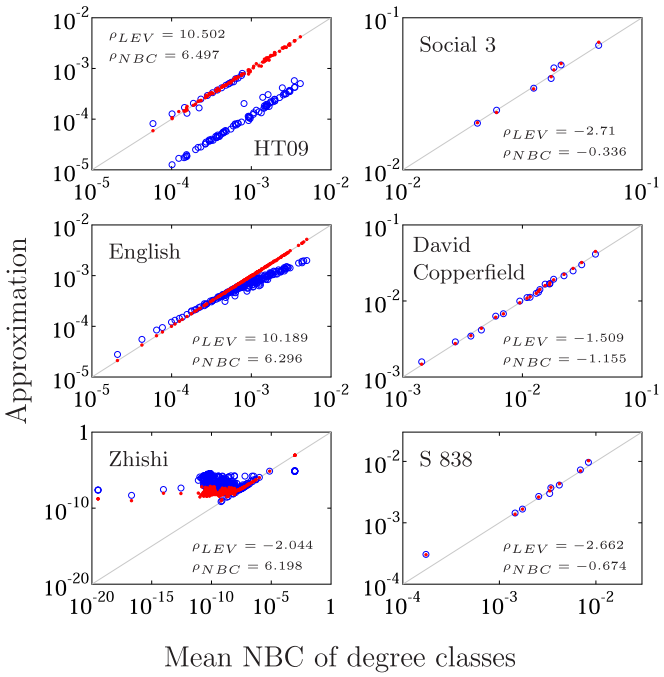

FIG. 3. Approximation of the mean NBC of nodes within degree classes using the expansion matrix approximation of Eq. (17) (red dots) and the local estimate of Ref. [18], Eq. (24) (open blue circles), averaged over the given degree class. (In each plot, one marker corresponds to one degree class.) Left panels show results for the three highest, right panels for the three lowest values of the quantity ρ_{NBC}, Eq. (27), indicating the three best and three worst cases (out of the 109) from the expansion matrix approximation viewpoint.

The expansion matrix LEV provides a better approximation in 77 of the 109 cases. More importantly, the expansion matrix approximation is consistently better when message-passing itself is a good approximation to percolation: the color code in Fig. 2 corresponds to the ratio of the LEV $\lambda_1^{(\mathbf{H})}$ of the NB matrix to the inverse percolation threshold p_c^{-1} estimated via simulations (see Ref. [18]). According to Ref. [1], $\lambda_1^{(\mathbf{H})}$ is a good approximation to p_c^{-1} in many empirical networks, and $\lambda_1^{(\mathbf{H})} \geqslant p_c^{-1}$ is strictly true in infinite networks. (The inequality was found to be true in all 109 empirical networks considered.)

These results imply that whatever structure is responsible for the LEV of the NB matrix of the network, it is also captured implicitly in the correlated, degree-based, expansion matrix approximation. To what extent this holds true may be checked by looking at how well the mean NBC is approximated for individual degree classes. In Fig. 3, we plot the expansion matrix approximation for the mean NBC of degree classes, as a function of the actual mean NBC (red dots), for six example networks.

The expansion matrix approximation of the mean NBC is compared with the uncorrelated approximation of Ref. [18], where a local estimate is given for the NBC of individual nodes,

$$x_i^{\mathrm{un}} \approx \frac{\sum_j A_{ij}(k_j - 1)}{\sum_j k_j(k_j - 1)}. \tag{24}$$

The mean value of $x_i^{\rm un}$ for degree classes is shown in Fig. 3 as a function of the actual mean NBC (open blue circles). To quantify the quality of the two approximations, we use the errors

$$\varepsilon_{\rm NBC}^{(\mathbf{E})} = \sum_{k}^{k_{\max}} \langle x\rangle_k \left(\langle x\rangle_k - \frac{v_{k_i}^{(\mathbf{E})}}{NP(k_i)} \right)^2, \tag{25}$$

$$\varepsilon_{\rm NBC}^{(\rm un)} = \sum_{k}^{k_{\max}} \langle x\rangle_k \left(\langle x\rangle_k - \langle x\rangle_k^{\rm un} \right)^2, \tag{26}$$

which are the weighted sums of squared differences for the two approximations. Degree classes of higher mean NBC generally play a bigger role in the underlying dynamics as described by the message-passing theory. It is therefore appropriate to use the mean NBC $\langle x\rangle_k$ as the weight in the above measure. To compare the two approximations, we use the logarithm of the ratio of the respective errors,

$$\rho_{\rm NBC} = \ln\left(\varepsilon_{\rm NBC}^{(\rm un)}/\varepsilon_{\rm NBC}^{(\mathbf{E})}\right). \tag{27}$$

$\rho_{\rm NBC} > 0$ indicates a smaller error for the expansion matrix approximation, while $\rho_{\rm NBC} < 0$ indicates a smaller error for the approximation based on Eq. (24). The six sample networks in Fig. 3 were chosen to contain the three networks where the expansion matrix approximation worked best compared to the local one (highest $\rho_{\rm NBC}$ values, left panels in Fig. 3) and the three networks where it performed the worst (lowest $\rho_{\rm NBC}$ values, right panels in Fig. 3). We can observe only small differences between the two approximations in the worst cases, but striking differences in the best. Importantly, the NBCs of high degree nodes, which play a more important role, tend to be much better approximated by the expansion matrix. Equivalent figures for all 109 networks (showing similar trends) are presented in the Supplemental Material.

Figure 2(b) shows the error $\varepsilon_{\rm NBC}^{(\mathbf{E})}$ as a function of $\varepsilon_{\rm NBC}^{(\rm un)}$ for all 109 networks. The expansion matrix approximation is better in all but six cases. Importantly, it is often markedly better when the message-passing approximation is itself valid. It is interesting to note that the approximation to the NBC works well in almost all cases, even when the LEV is badly approximated [see Fig. 2(a)].

It is worth analyzing how the performance of the two approximations depends on particular degree-degree correlation patterns. The expansion matrix approximation accounts for nearest-neighbor degree-degree correlations completely but assumes that there are no finite loops, which may be more problematic for certain types of networks than others. Analogously to Eq. (27), we define the following quantity to compare the two approximations to the LEV of the NB matrix:

$$\rho_{\rm LEV} = \ln\left(\varepsilon_{\rm LEV}^{(\mu)}/\varepsilon_{\rm LEV}^{(\mathbf{E})}\right), \tag{28}$$

which, again, is positive when the expansion matrix approximation is better. Figure 4 shows the quantities $\rho_{\rm NBC}$ and $\rho_{\rm LEV}$ as functions of the Pearson correlation coefficient σ (of nearest-neighbor degrees) for the 109 networks considered. Considering the LEV approximation [Fig. 4(a)], a clear trend is seen according to which the expansion matrix approximation works better for disassortative networks, while the approximation of Eq. (20) favors assortative networks. The reason why the latter works better for assortative networks

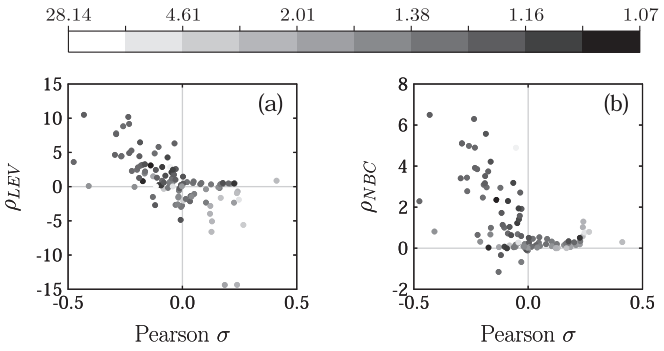

FIG. 4. Logarithmic approximation error ratios (a) $\rho_{\rm LEV}$ and (b) $\rho_{\rm NBC}$ as functions of the Pearson correlation coefficient σ for nearest-neighbor degree-degree correlations. The color code in both panels corresponds to the ratio $\lambda_1^{(\mathbf{H})}/p_c^{-1}$, indicating the quality of the message-passing approximation to percolation.

can be mostly attributed to the LEV of the highest k-core, which is explicitly included in Eq. (20). Strongly assortative networks are expected to contain many short loops among the highest-degree nodes, most of which belong to the highest k-core, which in turn tends to dominate the LEV of the NB matrix. This feature is missed in the expansion matrix approximation, where a locally treelike structure is assumed. It is important to note, however, that for the very same reason, message-passing theory itself is not a valid approximation in most of these assortative networks. Conversely, in the cases where message-passing is valid, the expansion matrix approximation is generally better, often markedly better. A similar trend can be seen in the case of the NBC approximation [Fig. 4(b)], where the expansion matrix approximation tends to strongly dominate for disassortative networks. For the NBC approximation, interestingly, also for assortative networks the expansion matrix approximation appears to be better, or at least as good as the local approximation.

The failure of the local approximation to correctly estimate the NBC in disassortative networks stems from the fact that the NBC of low degree nodes tends to be overestimated due to hubs in their immediate neighborhood. The NBC of higher degree nodes is underestimated as a consequence. The expansion matrix method provides a reliable approximation for all degree classes in such networks. The obvious advantage of this method, compared to the local one, is that it is self-referential, i.e., it takes the entire network into account to determine the estimate for the mean NBC of degree classes, similarly to the message-passing algorithm, only it does so in a course-grained manner, circumventing the necessity to have access to the full NB matrix.

The quality of these findings indicates that whatever structural property of a network is responsible for determining the LEV of the NB matrix, nodes of identical degrees generally have similar roles, therefore they can be treated parsimoniously as a degree-class if nearest-neighbor degree-degree correlations are taken into account. For large networks, this may be a significant simplification and reduction in computer memory requirement. The NBCs in a given network can be

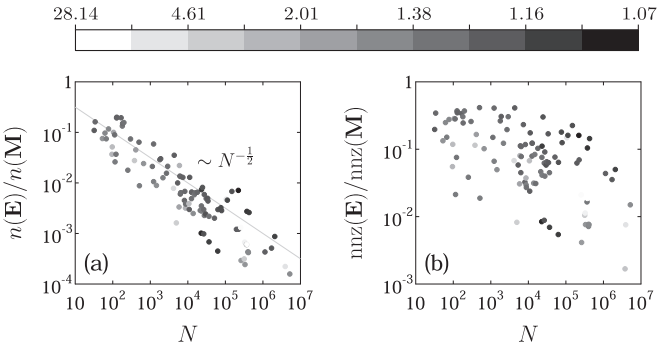

FIG. 5. Comparison of the size of the matrices **E** and **M** as a function of network size N. The ratio of the number of rows as a function of N is presented in panel (a); the ratio of the number of nonzero elements as a function of N is presented in panel (b). The color code in both panels corresponds to the ratio $\lambda_1^{(\mathbf{H})}/p_c^{-1}$, indicating the quality of the message-passing approximation to percolation.

calculated by first obtaining the PEV of the $2N \times 2N$ matrix,

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{I}-\mathbf{D} \\ \mathbf{I} & \mathbf{0} \end{pmatrix}, \tag{29}$$

where **A** is the adjacency matrix, **I** is the identity matrix, and **D** is the "degree matrix" whose elements are $D_{ij} = \delta_{ij} k_i$. The NBC values x_i correspond to the first N components of the PEV of matrix **M** [5]. The number of rows in matrix **M** is $2N$, whereas the number of rows in the expansion matrix **E** is the number n of different node degrees present in the network. The latter can be much smaller than the former for large networks. Figure 5(a) shows that the ratio of the number of rows for the two matrices tends to decay with network size slightly faster than $N^{-1/2}$. This is a consequence of the fact that $n(\mathbf{E})$ is upper bounded by $k_{\max}$, which is typically of the order of $N^{1/2}$ [24]. For computational purposes, however, what matters more than the number of rows is the number of nonzero elements in these matrices, $\mathrm{nnz}(\mathbf{E})$ and $\mathrm{nnz}(\mathbf{M})$, respectively. As can be seen in Fig. 5(b), the ratio of this quantity for the two matrices does not decay as strongly as the ratio of the number of rows, but a decay is still evident.

IV. LOCALIZATION OF NONBACKTRACKING CENTRALITY

It is well established that the PEV of the adjacency matrix may become localized on hubs and their neighboring nodes [5,15,16,25]. In particular, if the highest degree, $k_{\max}$, in the network is larger than $(\langle k^2\rangle/\langle k\rangle)^2$, then the PEV of the adjacency matrix is localized on this hub and the LEV is given by $\sqrt{k_{\max}}$. (Otherwise the PEV is effectively localized on the highest k-core [26].) This has significant consequences for recurrent epidemic models such as the SIS model, where it has been shown that in the quenched mean-field approximation, the epidemic threshold coincides with the inverse of the LEV of the adjacency matrix [27]. Hubs in the SIS model, therefore, have a special role in initiating disease spreading and maintaining an endemic state. For percolation and nonrecurrent epidemics, for which the NB matrix and the NBC are the relevant quantities, hubs lose their special role, although not completely. Contrary to the case of the adjacency matrix PEV, independent hubs cannot be centers of localization of the NBC [5]. However, as shown recently in Refs. [17,18], the NBC may still become localized on high-degree nodes when they are supported by other high-degree nodes, either directly (in a densely connected subgraph, e.g., the highest k-core) or indirectly (in an "overlapping hubs" structure, where a group of high-degree nodes share the same neighbors).

Here we demonstrate that the expansion matrix approximation can also capture this localization phenomenon, which is consistent with the high quality of the NBC and LEV estimates. We quantify the localization of the NBC using the inverse participation ratio (IPR),

$$Y_4 = \frac{\sum_i x_i^4}{\left(\sum_i x_i^2\right)^2}. \tag{30}$$

(The normalization $\sum_i x_i^2 = 1$ is often used.) The quantity Y_4 may be approximated by replacing each x_i with $\langle x\rangle_{k_i}$, the mean NBC value of nodes of degree k_i,

$$\tilde{Y}_4 = \frac{\sum_k NP(k)\langle x\rangle_k^4}{\left(\sum_k NP(k)\langle x\rangle_k^2\right)^2}. \tag{31}$$

In the expansion matrix approximation [Eq. (17)], we have $\langle x\rangle_k \approx v_k^{(\mathbf{E})}/[NP(k)]$, where $v_k^{(\mathbf{E})}$ are the components of the PEV of the expansion matrix. Our approximation to the IPR of the NBC is then

$$Y_4^{(\mathbf{E})} = \frac{\sum_k \left(v_k^{(\mathbf{E})}\right)^4/[NP(k)]^3}{\left[\sum_k \left(v_k^{(\mathbf{E})}\right)^2/[NP(k)]\right]^2}. \tag{32}$$

We will compare this estimate with the one obtained by using the uncorrelated, local approximation of the NBC, Eq. (24),

$$Y_4^{\mathrm{un}} = \frac{\sum_i \left(x_i^{\mathrm{un}}\right)^4}{\left[\sum_i \left(x_i^{\mathrm{un}}\right)^2\right]^2}. \tag{33}$$

Note that Eq. (32) is a degree-based, course-grained estimate, where each node is represented by the mean value of its degree class. Equation (33), on the other hand, is a node-based estimate, i.e., the estimate of each node's individual NBC makes a contribution. It should be noted that Eq. (33) is meant to be a good approximation primarily when the NBC is not localized on the highest k-core or overlapping hubs, i.e., when μ^{un} dominates in Eq. (20). Data points comparing the approximations of Eqs. (32) and (33) for all such cases are shown in Fig. 6(a). The expansion matrix approximation is better for 43 out of 55 networks. In the 12 cases where it is not, there is very little difference between the two approximations.

These findings indicate that, for the purposes considered here, the role of a node is largely determined by its degree, and the interaction between nodes of different degrees may be substituted by an averaged interaction between the respective degree classes.

Another important thing to consider is that although the NBC is not localized on independent hubs, the IPR is still dominated by high-degree nodes, due to their positions in densely connected subgraphs. Degree classes of high degrees

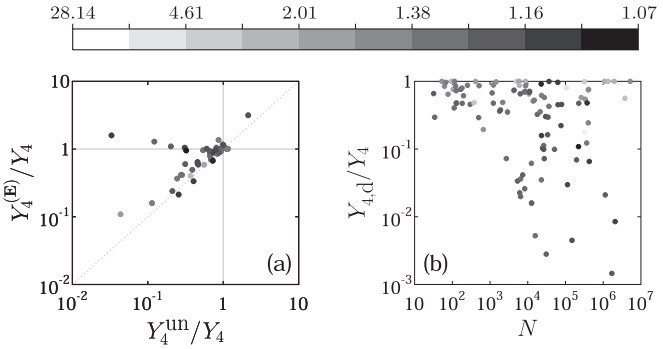

FIG. 6. (a) Comparison of the inverse participation ratio using the expansion matrix scheme and the local approximation. Values relative to the true inverse participation ratio are shown on both axes. The dashed gray line corresponds to $Y_4^{(\mathbf{E})} = Y_4^{\mathrm{un}}$. (b) Relative contribution of nodes of degenerate degrees to the inverse participation ratio Y_4, Eq. (30), as a function of network size. The color code in both panels corresponds to the ratio $\lambda_1^{(\mathbf{H})}/p_c^{-1}$, indicating the quality of the message-passing approximation to percolation.

have lower "degeneracy" (have fewer nodes in the degree class), therefore the degree-based approximation can be expected to work better. In particular, for nodes of the highest degrees, the degree classes are typically nondegenerate, and the expansion matrix describes their interconnections without any loss of information (compared to the NB matrix). Most of the information in the degree-based approximation is lost on low-degree classes that are highly degenerate, but these degree classes play a much smaller role in the localization phenomenon. We define $Y_{4,\mathrm{d}}$ as the contribution of nodes of degenerate degree classes to the IPR,

$$Y_{4,\mathrm{d}} = \frac{\sum\limits_{i:\mathrm{degen.}} x_i^4}{\left(\sum_i x_i^2\right)^2}, \tag{34}$$

where the sum in the numerator is taken over nodes whose degrees are not unique in the network, that is, nodes that belong to degenerate degree classes. In Fig. 6(b), we plot the relative contribution of such degree classes to the IPR in all 109 networks. This contribution is often quite small, particularly for larger networks where the message-passing approximation is valid, meaning that the IPR is to a large extent dominated by nodes of unique degrees, whose interconnections are correctly described by the expansion matrix.

V. DISCUSSION AND CONCLUSIONS

In this paper, we propose an approximation to the NBC of nodes in a network and the related LEV of the NB matrix. Our approximation relies on the assumption that the given network behaves similarly to an infinite random network, without finite loops, that has the same nearest-neighbor degree-degree correlations. Such correlated but otherwise uniformly random networks are described by a branching matrix—or equivalently, an expansion matrix—whose elements are related with the joint degree-degree distribution. The number of rows and columns in these matrices is given by the number of different degrees in the network, and hence they are generally much smaller than the NB matrix. The method we propose is degree-based, i.e., it is assumed that the NBCs of nodes are well approximated by the estimate of the mean NBC of their degree class. The estimates of the mean NBCs of degree classes are obtained by calculating the PEV of the branching or expansion matrix, and the estimate for the LEV of the NB matrix is given by the (identical) LEV of these two small matrices.

In spite of the fact that this method does not distinguish between nodes of identical degrees, the approximation for the mean NBC of degree classes and that for the LEV of the NB matrix are consistently better than existing local approximations to these quantities in networks where message-passing is a valid approximation to percolation-type phenomena. Importantly, our method tends to approximate the NBC of high degree nodes better, which play a more important role in determining the LEV of the NB matrix. The small matrices in our method may be thought of as a "compression" of the NB matrix, where most of the information lost is on low-degree classes, which are generally strongly degenerate (contain many nodes). High-degree classes have much lower degeneracy, so the description of their interconnections remains true to the information contained in the original NB matrix. The connections between degree-classes that are nondegenerate (contain a single node) are described without any loss of information. In light of this fact, it is understandable that the localization of the NBC, on subgraphs consisting of densely connected high-degree nodes, is also captured well in our approximation. That is, the localization of the PEV of the NB matrix is well traceable in the degree-based PEVs of the corresponding expansion or branching matrices. Our estimate of the inverse participation ratio is consistently better than that of the local approximation, despite the fact that in our method the NBCs of all nodes of identical degrees are considered equal.

The quality of our results also demonstrates that in most real-world networks, considering only nearest-neighbor degree-degree correlations is already sufficient for an accurate description—a description that is much better than the one resulting from assuming that the network is uncorrelated. An even more potent approximation may be achieved in principle by considering also tripletwise (or even higher order) correlations, as opposed to only pairwise. (A triplet is defined as a set of three nodes occupying the ends of two adjacent edges.) To construct the corresponding branching or expansion matrices, however, one would need to search through all the triplets (or higher-order structures) in the network. The number of triplets is dominated by the second moment of the degree distribution, therefore it may be very large for networks that possess fat-tailed degree distributions. Specifically, for scale-free networks the number of triplets (and of higher-order structures) is superlinear in system size if the degree distribution exponent is less than 3. Thus computational complexity will, in most cases, constitute a barrier to considering tripletwise (or higher-order) correlations in large networks.

In social networks, various characteristics of people may be correlated with degree, such as, e.g., age, profession, and income status. Assuming that such correlations are known, our method provides a simple means of estimating the contribution of different groups of people in dynamical processes

such as epidemics. These findings may help design preventative measures and vaccination strategies.

ACKNOWLEDGMENTS

We are grateful to Romualdo Pastor-Satorras and Claudio Castellano for providing us with the data for the 109 real-world networks. This work was developed within the scope of the project No. i3N, UIDB/50025/2020 & UIDP/50025/2020, financed by national funds through the FCT/MEC–Portuguese Foundation for Science and Technology. G.T. and R.A.d.C. were supported by FCT Grants No. CEECIND/03838/2017 and No. CEECIND/04697/2017.

APPENDIX: DERIVATION OF THE EXPANSION MATRIX APPROXIMATION OF NONBACKTRACKING CENTRALITY

The basic equations from which the NBC can be obtained are for directed links $i \leftarrow j$:

$$v_{i\leftarrow j} = \lambda_1^{-1} \sum_{k\leftarrow l} H_{i\leftarrow j,k\leftarrow l} v_{k\leftarrow l}, \tag{A1}$$

where $\mathbf{H}$ is the nonbacktracking matrix, and λ_1 and $\mathbf{v}$ are its largest eigenvalue and principal eigenvector, respectively. Let us define the quantity $\bar{v}(\vec{k})$ for directed links of degree class $\vec{k} = (k_1, k_2)$ as

$$\bar{v}(\vec{k}) = \frac{1}{2LP(\vec{k})} \sum_{i\leftarrow j\in\vec{k}} v_{i\leftarrow j}, \tag{A2}$$

where $2L$ is the number of directed links in the network, and $P(\vec{k})$ is the probability that a uniformly randomly chosen directed link has end- and start-node degrees k_1 and k_2, respectively. The quantity $\bar{v}(\vec{k})$ is the mean $v_{i\leftarrow j}$ value over all links $i \leftarrow j$ where the degree of node i is k_1 and the degree of node j is k_2. Summing Eq. (A1) over all directed links of degree class $\vec{k}$ and dividing by $2LP(\vec{k})$, we have

$$\begin{aligned}\bar{v}(\vec{k}) &= \frac{\lambda_1^{-1}}{2LP(\vec{k})} \sum_{i\leftarrow j\in\vec{k}} \sum_{k\leftarrow l} H_{i\leftarrow j,k\leftarrow l} v_{k\leftarrow l} \\ &= \frac{\lambda_1^{-1}}{2LP(\vec{k})} \sum_{i\leftarrow j\in\vec{k}} \sum_{\vec{k}'} \sum_{k\leftarrow l\in\vec{k}'} H_{i\leftarrow j,k\leftarrow l} v_{k\leftarrow l}.\end{aligned} \tag{A3}$$

In the last equation, we split the sum over directed links $k \leftarrow l$ into a sum over degree classes and a sum over links within degree classes. Now we make the mean-field assumption that all $v_{k\leftarrow l}$ values can be approximated by the corresponding mean value for the degree class,

$$\begin{aligned}\bar{v}(\vec{k}) &= \frac{\lambda_1^{-1}}{2LP(\vec{k})} \sum_{i\leftarrow j\in\vec{k}} \sum_{\vec{k}'} \sum_{k\leftarrow l\in\vec{k}'} H_{i\leftarrow j,k\leftarrow l} \bar{v}(\vec{k}') \\ &= \frac{\lambda_1^{-1}}{2LP(\vec{k})} \sum_{\vec{k}'} \bar{v}(\vec{k}') \left(\sum_{i\leftarrow j\in\vec{k}} \sum_{k\leftarrow l\in\vec{k}'} H_{i\leftarrow j,k\leftarrow l} \right).\end{aligned} \tag{A4}$$

The quantity in large parentheses has a well-defined meaning: this is the number of $i_1 \leftarrow i_2 \leftarrow i_3$ directed triplets in the network where nodes i_1 and i_2 have degrees k_1 and k_2, and also the nodes i_2 and i_3 have degrees k_1' and k_2'. Or, more succinctly, this is the number of directed link junctions of type $\vec{k} \leftarrow \vec{k}'$. Let us approximate this by the number of such directed link junctions in a network with nearest-neighbor degree-degree correlations described by a joint degree-degree distribution. This number is

$$J_{\vec{k}\leftarrow\vec{k}'} = 2LP(\vec{k}')(k_2 - 1)P(k_1|k_2)\delta_{k_2,k_1'}, \tag{A5}$$

where $\delta_{k,k'}$ is the Kronecker delta. Plugging Eq. (A5) into Eq. (A4) and rearranging, we have

$$2LP(\vec{k})\bar{v}(\vec{k}) = \lambda_1^{-1} \sum_{\vec{k}'} 2LP(\vec{k}')\bar{v}(\vec{k}')(k_2 - 1)P(k_1|k_2)\delta_{k_2,k_1'}. \tag{A6}$$

Recall that $\bar{v}(\vec{k})$ is the mean $v_{i\leftarrow j}$ value of directed links of degree class $\vec{k}$. Then the left-hand side of Eq. (A6) is just the sum of such values. Denoting this sum by $g(\vec{k}) = g(k_1 \leftarrow k_2)$, we can write

$$g(k_1 \leftarrow k_2) = \lambda_1^{-1} \sum_{k_1'\leftarrow k_2'} g(k_1' \leftarrow k_2')(k_2 - 1)P(k_1|k_2)\delta_{k_2,k_1'}, \tag{A7}$$

where we have switched from vectorial to componentwise notation for degree classes of directed links. Rearranging, we get

$$\begin{aligned}g(k_1 \leftarrow k_2) &= \lambda_1^{-1}(k_2 - 1)P(k_1|k_2) \sum_{k_1'\leftarrow k_2'} g(k_1' \leftarrow k_2')\delta_{k_2,k_1'} \\ &= \lambda_1^{-1}(k_2 - 1)P(k_1|k_2) \sum_{k_2'} g(k_2 \leftarrow k_2').\end{aligned} \tag{A8}$$

Recall that $g(k \leftarrow k')$ is the sum of $v_{i\leftarrow j}$ values of directed links within the degree class $k \leftarrow k'$. The sum of $v_{i\leftarrow j}$ values over all directed links incoming to nodes of degree class k is then written as

$$h_k = \sum_{k'} g(k \leftarrow k'). \tag{A9}$$

Using Eq. (A9) and summing both sides of Eq. (A8) over k_2, we have

$$h_{k_1} = \lambda_1^{-1} \sum_{k_2} (k_2 - 1)P(k_1|k_2)h_{k_2}. \tag{A10}$$

Introducing the matrix $E_{k_1,k_2} = (k_2 - 1)P(k_1|k_2)$, we finally have

$$h_{k_1} = \lambda_1^{-1} \sum_{k_2} E_{k_1,k_2} h_{k_2}, \tag{A11}$$

or in vector form,

$$\lambda_1 \mathbf{h} = \mathbf{E}\mathbf{h}. \tag{A12}$$

Equation (A12) is an eigenvector equation for the principal eigenvector of $\mathbf{E}$ (by the Perron-Frobenius theorem). The components of the vector $\mathbf{h}$ approximate the sum of NBCs of nodes of the corresponding degree class, and the LEV of $\mathbf{E}$ is an approximation to λ_1, the LEV of the NB matrix.

[1] B. Karrer, M. E. J. Newman, and L. Zdeborová, Percolation on Sparse Networks, Phys. Rev. Lett. **113**, 208702 (2014).

[2] F. Radicchi, Predicting percolation thresholds in networks, Phys. Rev. E **91**, 010801(R) (2015).

[3] B. Karrer and M. E. J. Newman, Message passing approach for general epidemic models, Phys. Rev. E **82**, 016101 (2010).

[4] K. Hashimoto, *Automorphic Forms and Geometry of Arithmetic Varieties*, edited by K. Hashimoto and Y. Namikawa (Elsevier, Amsterdam, 1989), p.211.

[5] T. Martin, X. Zhang, and M. E. J. Newman, Localization and centrality in networks, Phys. Rev. E **90**, 052808 (2014).

[6] G. Timár, R. A. da Costa, S. N. Dorogovtsev, and J. F. F. Mendes, Nonbacktracking expansion of finite graphs, Phys. Rev. E **95**, 042322 (2017).

[7] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, Spectral redemption in clustering sparse networks, Proc. Natl. Acad. Sci. (USA) **110**, 20935 (2013).

[8] C. Bordenave, M. Lelarge, and L. Massoulié, Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs, in *2015 IEEE 56th Annual Symposium on Foundations of Computer Science* (IEEE, Piscataway, NJ, 2015), p. 1347.

[9] F. Morone and H. A. Makse, Influence maximization in complex networks through optimal percolation, Nature (London) **524**, 65 (2015).

[10] F. Morone, B. Min, L. Bo, R. Mari, and H. A. Makse, Collective influence algorithm to find influencers via optimal percolation in massively large social media, Sci. Rep. **6**, 1 (2016).

[11] L. Torres, K. S. Chan, H. Tong, and T. Eliassi-Rad, Nonbacktracking Eigenvalues under Node Removal: X-Centrality and Targeted Immunization, SIAM J. Math. Data Sci. **3**, 656 (2020).

[12] F. Radicchi and C. Castellano, Leveraging percolation theory to single out influential spreaders in networks, Phys. Rev. E **93**, 062314 (2016).

[13] B. Min, Identifying an influential spreader from a single seed in complex networks via a message-passing approach, Eur. Phys. J. B **91**, 1 (2018).

[14] S. Moore and T. Rogers, Predicting the Speed of Epidemics Spreading in Networks, Phys. Rev. Lett. **124**, 068301 (2020).

[15] A. V. Goltsev, S. N. Dorogovtsev, J. G. Oliveira, and J. F. F. Mendes, Localization and Spreading of Diseases in Complex Networks, Phys. Rev. Lett. **109**, 128702 (2012).

[16] R. Pastor-Satorras and C. Castellano, Eigenvector localization in real networks and its implications for epidemic spreading, J. Stat. Phys. **173**, 1110 (2018).

[17] T. Kawamoto, Localized eigenvectors of the non-backtracking matrix, J. Stat. Mech.: Theor. Exp. (2016) 023404.

[18] R. Pastor-Satorras and C. Castellano, The localization of non-backtracking centrality in networks and its physical consequences, Sci. Rep. **10**, 1 (2020).

[19] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, Epidemic spreading in complex networks with degree correlations, in *Statistical Mechanics of Complex Networks* (Springer, Berlin, 2003), p. 127.

[20] A. V. Goltsev, S. N. Dorogovtsev, and J. F. F. Mendes, Percolation on correlated networks, Phys. Rev. E **78**, 051105 (2008).

[21] M. Boguñá and R. Pastor-Satorras, Epidemic spreading in correlated complex networks, Phys. Rev. E **66**, 047104 (2002).

[22] S. Fortunato, M. Boguñá, A. Flammini, and F. Menczer, Approximating PageRank from in-degree, in *International Workshop on Algorithms and Models for the Web-graph* (Springer, Berlin, 2006), p. 59.

[23] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.104.054306 for a set of plots comparing approximations of the non-backtracking centrality in 109 empirical networks and a table with key characteristics of these networks and numbers quantifying the accuracy of different approximations to the nonbacktracking centrality.

[24] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Critical phenomena in complex networks, Rev. Mod. Phys. **80**, 1275 (2008).

[25] C. Castellano and R. Pastor-Satorras, Competing activation mechanisms in epidemics on networks, Sci. Rep. **2**, 1 (2012).

[26] R. Pastor-Satorras and C. Castellano, Distinct types of eigenvector localization in networks, Sci. Rep. **6**, 1 (2016).

[27] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos, Epidemic spreading in real networks: An eigenvalue viewpoint, in *Proceedings of the 22nd International Symposium on Reliable Distributed Systems* (IEEE, Piscataway, NJ, 2003), p. 25.