

2016-04-19

# Predicting the epidemic threshold of the susceptible-infected-recovered model

---

Wei Wang, Quan-Hui Liu, Lin-Feng Zhong, Ming Tang, Hui Gao, H Eugene Stanley. 2016.  
"Predicting the epidemic threshold of the susceptible-infected-recovered model." SCIENTIFIC  
REPORTS, Volume 6, 12 pp. <https://doi.org/10.1038/srep24676>

<https://hdl.handle.net/2144/39930>

*"Downloaded from OpenBU. Boston University's institutional repository."*

# SCIENTIFIC REPORTS



OPEN

## Predicting the epidemic threshold of the susceptible-infected-recovered model

Received: 21 December 2015

Accepted: 31 March 2016

Published: 19 April 2016

Wei Wang<sup>1,2,3</sup>, Quan-Hui Liu<sup>1,2</sup>, Lin-Feng Zhong<sup>1,2</sup>, Ming Tang<sup>1,2</sup>, Hui Gao<sup>1,2</sup> & H. Eugene Stanley<sup>3</sup>

Researchers have developed several theoretical methods for predicting epidemic thresholds, including the mean-field like (MFL) method, the quenched mean-field (QMF) method, and the dynamical message passing (DMP) method. When these methods are applied to predict epidemic threshold they often produce differing results and their relative levels of accuracy are still unknown. We systematically analyze these two issues—relationships among differing results and levels of accuracy—by studying the susceptible-infected-recovered (SIR) model on uncorrelated configuration networks and a group of 56 real-world networks. In uncorrelated configuration networks the MFL and DMP methods yield identical predictions that are larger and more accurate than the prediction generated by the QMF method. As for the 56 real-world networks, the epidemic threshold obtained by the DMP method is more likely to reach the accurate epidemic threshold because it incorporates full network topology information and some dynamical correlations. We find that in most of the networks with positive degree-degree correlations, an eigenvector localized on the high  $k$ -core nodes, or a high level of clustering, the epidemic threshold predicted by the MFL method, which uses the degree distribution as the only input information, performs better than the other two methods.

Because many real-world phenomena incorporate spreading dynamics on complex networks, the topic has received much attention over the last decade<sup>1,2</sup>. Notable examples include the spread of sexually-transmitted diseases through contact networks<sup>3</sup>, the spread of malware on wireless networks<sup>4</sup>, and the spread of computer viruses through email networks<sup>5</sup>. In each case the spreading dynamics are strongly affected by network topology, and this complicates the task of understanding their behavior. Existing studies of spreading dynamics have focused on both theoretical aspects (e.g., nonequilibrium critical phenomena<sup>6,7</sup>) and practical issues (e.g., proposing efficient immunization strategies<sup>8,9</sup>). Researchers have focused on developing ways of accurately identifying epidemic thresholds because of their important ramifications in many real-world scenarios. Theoretically speaking, an epidemic threshold characterizes the critical condition above which a global epidemic occurs<sup>7</sup>. Being able to predict an epidemic threshold allows us to determine the critical exponents<sup>10</sup> and Griffiths effects<sup>11</sup>, which are important in research on nonequilibrium phenomena<sup>6</sup>. Practically speaking, quantifying an epidemic threshold allows us to determine the effectiveness of a given immunization strategy<sup>8</sup>. A proposed immunization strategy is effective if it increases the epidemic threshold. In addition, knowing the epidemic threshold enables us to more accurately determine the optimum source node<sup>12</sup>.

Researchers have put much effort into developing a theory for quantifying the thresholds in epidemic spreading models such as the susceptible-infected-recovered (SIR) model<sup>1</sup>. The best-known theoretical methods fall into three categories based on the topology information that they use. The first is the mean-field like (MFL) approach, which uses the degree distribution as the sole input parameter. This category includes the heterogeneous mean-field theory<sup>7,13</sup>, the percolation theory<sup>14</sup>, the edge-based compartmental approach<sup>15–18</sup>, and the pairwise approximation method<sup>19,20</sup>. The second type is the quenched mean-field (QMF) method that describes network topology in terms of the adjacent matrix. Examples include the discrete-time Markov chain<sup>21</sup> and the  $N$ -intertwined approach<sup>22</sup>. The third type is the dynamical message passing (DMP) method<sup>23</sup> that describes network topology in terms of the non-backtracking matrix. This approach is accurate in the case of tree-like

<sup>1</sup>Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 610054, China. <sup>2</sup>Big data research center, University of Electronic Science and Technology of China, Chengdu 610054, China. <sup>3</sup>Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA. Correspondence and requests for materials should be addressed to M.T. (email: tangminghan007@gmail.com)

networks. Researchers have used these three approaches to uncover the macroscopic statistical characteristics (e.g., degree<sup>7</sup> and weight distributions<sup>17</sup>), mesoscale structure (e.g., degree-degree correlations<sup>24</sup>, clustering<sup>25</sup> and community<sup>26</sup>), and microcosmic characteristics (e.g., node degree<sup>27</sup> and edge weight<sup>17</sup>) that strongly affect the epidemic threshold. For example, uncorrelated or correlated networks with a strongly heterogeneous degree distribution can, under certain conditions, reduce or even eliminate the epidemic threshold<sup>7,24</sup>.

The theoretical approaches always assume (i) that an epidemic can spread on a large, sparse network<sup>7,14,16,28</sup>, (ii) that dynamical correlations among the neighbors do not exist<sup>7</sup>, and (iii) that all the nodes or edges within a given class are statistically equivalent<sup>7,17</sup>. These three methods also usually focus on a class of networks, such as uncorrelated networks, clustering networks, and community networks. In any given network, the three theoretical methods usually predict different epidemic thresholds<sup>29</sup>. To determine the relationships among the three differing outcomes of the MFL method, the QMF method, and the DMP method and to determine which more closely describes real-world epidemic thresholds, we use a comprehensive study of the SIR model on uncorrelated configuration networks and of a group of 56 real-world networks. We find that the MFL and DMP methods predict the same epidemic threshold value for uncorrelated configuration networks and that this value is larger and more accurate than the value predicted by the QMF method. The relationships among the three theoretical predictions for real-world networks, however, remain unclear. In the 56 real-world networks studied, the DMP method performs the best in most cases because it considers the full topology and many of the dynamical correlations among the states of the neighbors, but due to the localized eigenvector of the adjacent matrix the QMF method often deviates from accurate epidemic threshold values. For networks with an eigenvector localized on the high  $k$ -core nodes, positive degree-degree correlations, or high clustering, the prediction by MFL method is more likely to be accurate than the predictions from other two methods, even though the MFL method uses the degree distribution as the sole input parameter. For networks with an eigenvector localized on the hubs, negative degree-degree correlations, or low clustering, the DMP method performs the best in most occasions. Finally, we note that the performances of the three predictions do not exhibit an obvious regularity versus the modularity, and in most cases the DMP method performs better than other two.

## Results

**Theoretical predictions of epidemic threshold.** In the SIR pattern of the spread of disease through a network, at any given time each node is either susceptible, infected, or recovered. A susceptible node does not transmit the disease. Infected nodes contract the disease and spread it to their neighbors. A recovered node has returned to health and no longer spreads the disease. The synchronous updating method<sup>30</sup> is applied to renew the states of nodes. To initiate the epidemic, we randomly select a “seed” node and designate all other nodes susceptible. At each time step, infected nodes transmit the disease to susceptible neighbors with a probability  $\beta$ . Infected nodes can also recover with a probability  $\gamma$ . The spreading terminates when all infected nodes have recovered. The spreading dynamics can be characterized by the effective spreading rate  $\lambda = \beta/\gamma$ . More details are shown in the Supporting Information. When  $\lambda$  is below the epidemic threshold  $\lambda_c$  (i.e.,  $\lambda \leq \lambda_c$ ), the disease spreads locally (i.e., only a tiny fraction of nodes transmit the disease). Epidemics can occur when  $\lambda > \lambda_c$  (i.e., when a finite fraction of nodes transmit the disease).

The mean-field like (MFL) method, the quenched mean-field (QMF) method, and the dynamical message passing (DMP) method are commonly-used theoretical methods of predicting an epidemic threshold. In this section we clarify the relationships among these epidemic thresholds predicted by the three theoretical methods.

The mean-field like (MFL) method incorporates the heterogeneous mean-field theory, percolation theory, the edge-based compartmental approach, and the pairwise approximation method. Here the epidemic threshold is predicted by using only the degree distribution, and it is assumed that (i) all the nodes and edges in a given class are statistically equivalent, (ii) the states of nodes among neighbors are independent, and (iii) the network size is infinite. Using the degree distribution  $P(k)$  as the only input parameter, the theoretical epidemic threshold prediction using the MFL method is

$$\lambda_c^{\text{MFL}} = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}, \quad (1)$$

where  $\langle k \rangle$  and  $\langle k^2 \rangle$  are the first and second moments of the degree distribution, respectively. Although  $\lambda_c^{\text{MFL}}$  is a good predictor of the epidemic threshold in uncorrelated networks, the prediction may fail in real-world networks because of their complex structure (e.g., degree-degree correlations, clustering, and community) and the strong dynamical correlations among the states of neighbors<sup>27,31</sup>.

The quenched mean-field (QMF) method<sup>21,32,33</sup> takes into account the complete network structure by using the adjacent matrix  $A$ . This distinguishes it from the MFL method, which simply uses the degree distribution. The adjacent matrix  $A$  is also used to describe network topology by the discrete-time Markov chain<sup>21</sup>, the  $N$ -intertwined method<sup>22</sup>, and other similar methods, and thus they fall into the same class as the QMF method. The QMF method is unable to capture the dynamical correlations among the states of neighbors and uses only the correlation between the theoretical epidemic threshold and the leading eigenvalue of the adjacent matrix to predict the epidemic threshold, i.e.,

$$\lambda_c^{\text{QMF}} = \frac{1}{\Lambda_A}, \quad (2)$$

where the leading eigenvalue of the adjacent matrix is<sup>22</sup>

$$\Lambda_A = \max_{\vec{v}} \left( \frac{\vec{v}^T A \vec{v}}{\vec{v}^T \vec{v}} \right), \tag{3}$$

where  $\vec{v}$  is a column vector with  $N$  elements, and  $N$  is the network size. Note that the epidemic threshold predicted by Eq. (2) is the same with the lower bound of epidemic threshold of SIS model<sup>33</sup>. Since the epidemic threshold of SIS model is smaller than that of SIR model<sup>34</sup>, we know that  $\lambda_c^{\text{QMF}}$  is precise a lower bound of epidemic threshold of SIR model.

The dynamical message passing (DMP) method was recently developed and used to study nonreversible epidemic spreading dynamics in an SIR modeled finite-sized network<sup>23,28,35</sup>. The DMP method uses the non-backtracking matrix to determine the complete network structure. This method can both describe the complete network structure and capture some of the dynamical correlations among the states of neighbors that are neglected in the MFL and QMF methods. In large sparse networks the DMP method provides a good estimation of the epidemic threshold, i.e.,

$$\lambda_c^{\text{DMP}} = \frac{1}{\Lambda_M}, \tag{4}$$

where

$$\Lambda_M = \max_{\vec{w}} \left( \frac{\vec{w}^T M \vec{w}}{\vec{w}^T \vec{w}} \right) \tag{5}$$

is the leading eigenvalue of the non-backtracking matrix<sup>36-39</sup>

$$M = \begin{pmatrix} A & \mathbf{1} - D \\ \mathbf{1} & \mathbf{0} \end{pmatrix}, \tag{6}$$

and  $\mathbf{1}$  is a  $N \times N$  unit matrix,  $D$  is the diagonal matrix with the vertex degrees along its diagonal, and  $\mathbf{0}$  is a  $N \times N$  null matrix. From Eqs (1, 2 and 4), we know that the predicted epidemic threshold of SIR model has the same formula with the bond percolation model<sup>36</sup>. Since the SIR spreading is a dynamical evolution process, the interplay between complex structures and dynamical correlations may result in a distinct accurate critical point from the bond percolation model<sup>40,41</sup>. Therefore, how the above three classical theoretical methods perform in predicting the epidemic threshold of SIR model in complex networks is worth pursuing.

The three theoretical predictions of epidemic threshold are closely correlated. In any given network they distinct, e.g.,  $\lambda_c^{\text{QMF}}$  is less than  $\langle k \rangle / \langle k^2 \rangle^{1/2}$ . To determine other relationships among the three theoretical thresholds, we assume that  $\kappa$  is a eigenvalue of non-backtracking matrix  $M$  and that  $w = (\vec{w}_1, \vec{w}_2)^T$  is the corresponding eigenvector of  $\kappa$ , where  $\vec{w}_1$  and  $\vec{w}_2$  are the first and last  $N$  elements of vector  $w$ , respectively. Using Eq. (6), the eigenvalue problem is written

$$\begin{cases} A\vec{w}_1 + (\mathbf{1} - D)\vec{w}_2 = \kappa\vec{w}_1, \\ \vec{w}_1 = \kappa\vec{w}_2. \end{cases} \tag{7}$$

Multiplying the left vector  $\vec{u} = (1, \dots, 1)$  on the first line of (7) and combining the second line of (7) yields

$$\kappa = \frac{\vec{d}^T \vec{w}_1}{\vec{u} \vec{w}_1} - 1, \tag{8}$$

where  $\vec{d} = (d_1, \dots, d_N)^T$  and  $d_i$  is the degree of node  $i$ . In uncorrelated networks the nonbacktracking centrality of a node is proportional to its degree<sup>37</sup>, i.e.,  $w_i \sim d_i$ . Here the theoretical prediction  $\lambda_c^{\text{DMP}}$  using the DMP method is the same as  $\lambda_c^{\text{MFL}}$  using the MFL method.

To examine the eigenvalue relationships between the adjacent matrix and non-backtracking matrix, we insert the second equation of (7) into the first equation and obtain

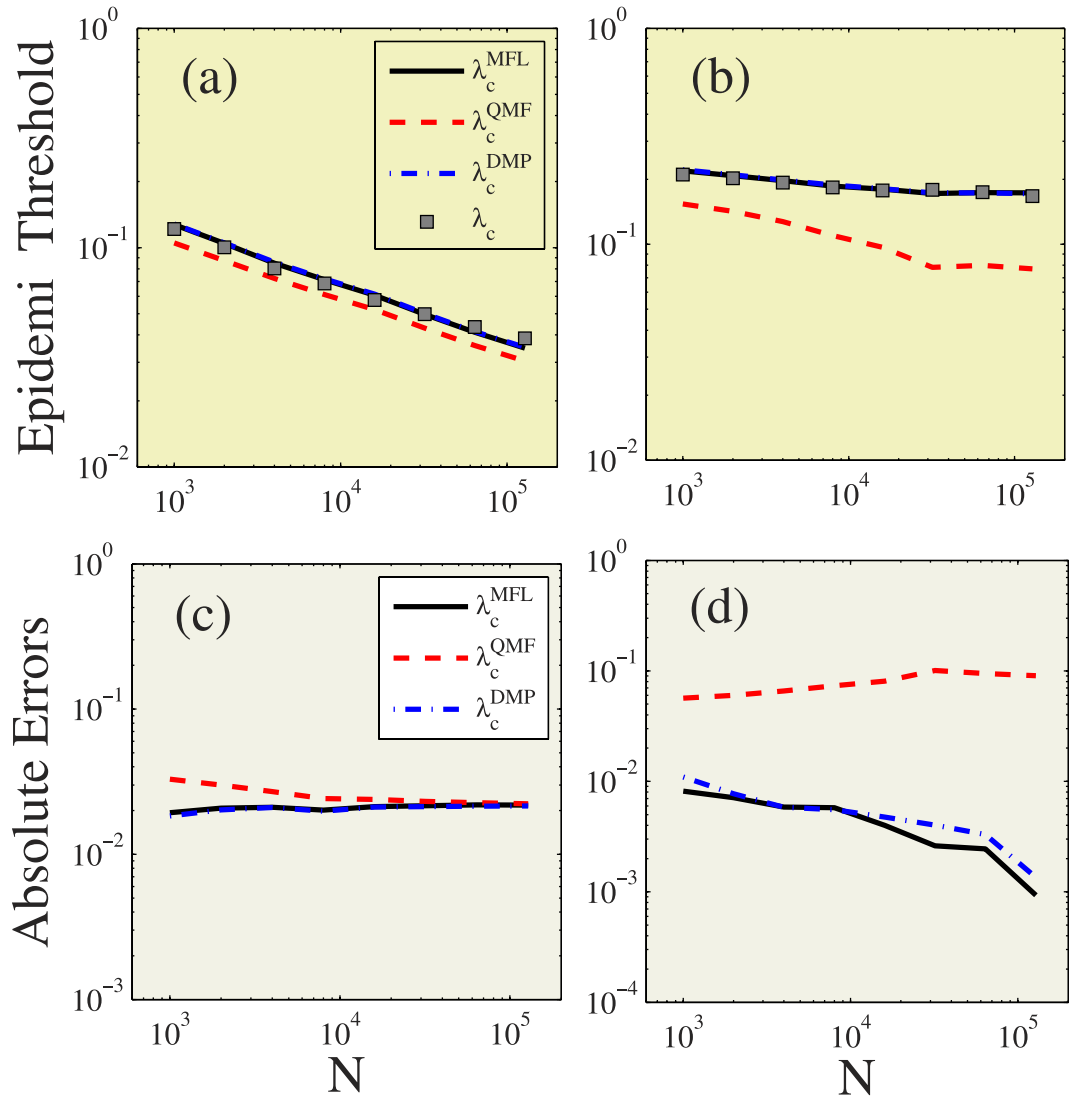
$$\kappa A \vec{w}_2 + (\mathbf{1} - D) \vec{w}_2 = \kappa^2 \vec{w}_2. \tag{9}$$

Multiplying  $\vec{w}_2^T$  on both sides of Eq. (9) and dividing  $\vec{w}_2^T \vec{w}_2$ , we get

$$\frac{\kappa \vec{w}_2^T A \vec{w}_2}{\vec{w}_2^T \vec{w}_2} + \frac{\vec{w}_2^T (\mathbf{1} - D) \vec{w}_2}{\vec{w}_2^T \vec{w}_2} = \kappa^2. \tag{10}$$

Using matrix theory<sup>22</sup> we know that the eigenvalue  $\epsilon$  and its corresponding eigenvector  $\vec{h}$  of a matrix  $\mathcal{X}$  satisfy  $\epsilon = \frac{\vec{h}^T \mathcal{X} \vec{h}}{\vec{h}^T \vec{h}}$ . We assume that  $\xi_1$  and  $\xi_2$  are the eigenvalue of  $A$  and  $\mathbf{1} - D$ , respectively, i.e.,  $\xi_1 = \frac{\vec{w}_2^T A \vec{w}_2}{\vec{w}_2^T \vec{w}_2}$  and  $\xi_2 = \frac{\vec{w}_2^T (\mathbf{1} - D) \vec{w}_2}{\vec{w}_2^T \vec{w}_2}$ . Thus Eq. (10) can be written as

$$\kappa^2 = \kappa \xi_1 + \xi_2. \tag{11}$$



**Figure 1. Predicting epidemic threshold for uncorrelated configuration networks under different network sizes.** Theoretical predictions of  $\lambda_c^{\text{MFL}}$  (black solid lines),  $\lambda_c^{\text{QMF}}$  (red dashed lines),  $\lambda_c^{\text{DMP}}$  (blue dash-dotted lines) and numerical prediction (gray squares) versus network size  $N$  for degree exponent  $\nu_D = 2.1$  (a) and  $\nu_D = 3.5$  (b). The absolute errors between  $\lambda_c$  and  $\lambda_c^{\text{MFL}}$  (black solid lines),  $\lambda_c^{\text{QMF}}$  (red dashed lines) and  $\lambda_c^{\text{DMP}}$  (blue dash-dotted lines) versus  $N$  for  $\nu_D = 2.1$  (c) and  $\nu_D = 3.5$  (d).

Because the minimum eigenvalue of  $\mathbf{1} - D$  is  $1 - k_{\max}$ , we find that

$$\kappa^2 \leq \kappa \xi_1 + 1 - k_{\max}. \tag{12}$$

Rewriting Eq. (12) we get

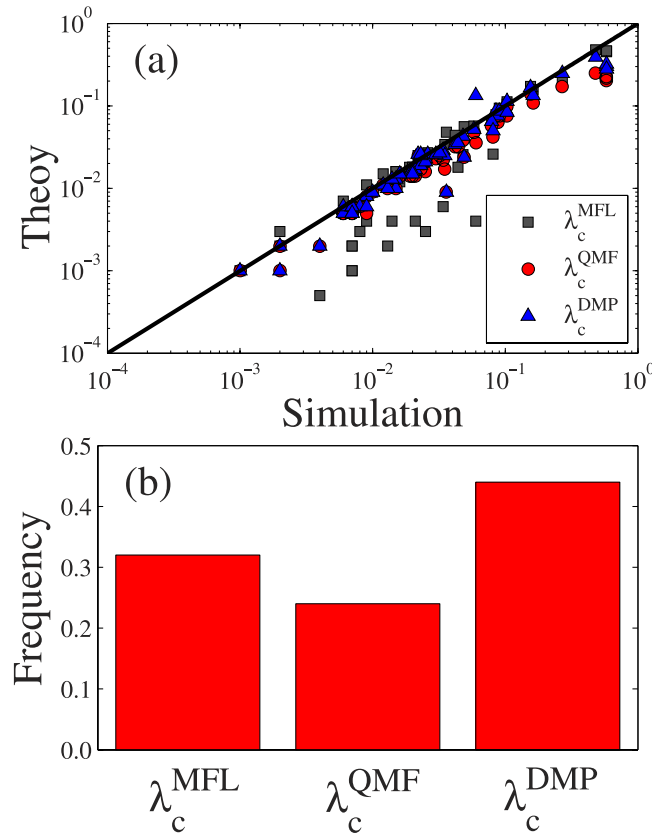
$$\kappa + \frac{k_{\max} - 1}{\kappa} \leq \xi_1. \tag{13}$$

Note that  $\kappa$  and  $\xi_1$  are the eigenvalues of matrixes  $M$  and  $A$  respectively, and we get

$$\lambda_c^{\text{DMP}} \geq \lambda_c^{\text{QMF}}. \tag{14}$$

With similar arguments in ref. 42 and combining Eq. (14), we know that  $\lambda_c^{\text{DMP}}$  is a tight lower bound of the accurate epidemic threshold  $\lambda_c$  for local tree-like networks. For real-world networks, the basic assumption (i.e., local tree-like) can not always be satisfied, thus,  $\lambda_c^{\text{DMP}}$  is possible larger than  $\lambda_c$ .

Many real-world networks have a heterogeneous degree distribution, e.g., a power-law degree distribution  $P(k) \sim k^{-\nu_D}$ , where  $\nu_D$  is the degree exponent. In uncorrelated scale-free networks,  $\lambda_c^{\text{MFL}}$  vanishes in the thermodynamic limit when  $\nu_D < 3$  because  $\langle k^2 \rangle$  diverges. When  $\nu_D > 3$ ,  $\lambda_c^{\text{MFL}}$  is a finite value. Using the QMF method, the epidemic threshold  $\lambda_c^{\text{QMF}}$  is determined by the maximum degree  $k_{\max}$ . When the degree exponent  $\nu_D > 2.5$ ,



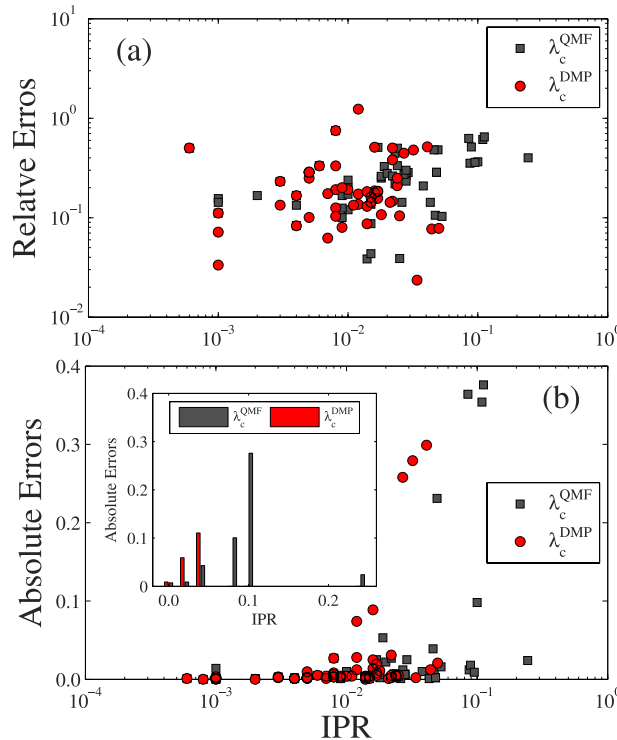
**Figure 2. Comparing the accuracy between three types of theoretical and numerical predictions of the epidemic threshold on 56 real-world networks. (a)** Theoretical predictions of  $\lambda_c^{\text{MFL}}$  (gray squares),  $\lambda_c^{\text{QMF}}$  (red circles) and  $\lambda_c^{\text{DMP}}$  (blue up triangles) versus numerical predictions  $\lambda_c$  of the epidemic threshold. **(b)** In all the entire sample of real-world networks, the fraction of  $\lambda_c^{\text{MFL}}$  [ $\lambda_c^{\text{QMF}}$  or  $\lambda_c^{\text{DMP}}$ ] is the closest value to  $\lambda_c$ .

and  $\lambda_c^{\text{QMF}} \propto 1/\sqrt{k_{\text{max}}}$ . When  $\nu_D < 2.5$ , we have  $\lambda_c^{\text{QMF}} \propto \langle k \rangle / \langle k^2 \rangle^{43}$ , which indicates that  $\lambda_c^{\text{QMF}} < \lambda_c^{\text{MFL}}$ . Note that  $\lambda_c^{\text{DMP}} = \langle k \rangle / (\langle k^2 \rangle - \langle k \rangle)$  for uncorrelated networks<sup>38</sup> is the same with  $\lambda_c^{\text{MFL}}$ . According to Eq. (14),  $\lambda_c^{\text{DMP}}$  is always larger than  $\lambda_c^{\text{QMF}}$ . Unfortunately, the complex topology of the real-world networks makes the relationships among the three types of prediction unclear.

**Simulation results.** Increasing the amount of network topology information utilized in any predictive method, the intuitional understanding tells us that the better performance of the method. Using the assumptions listed in previous section, we expect the DMP method to outperform the QMF method and the QMF method to outperform the MFL method. We next evaluate the performance of the three types of method using a large number on SIR studies of (i) uncorrelated configuration networks, and (ii) 56 real-world networks. We employ the estimators supplied in previous section to determine the theoretical epidemic threshold, and use the relative variance to determine the accurate epidemic threshold (see details in Method).

To better understand the performance of the three types of method, we further classify the networks into two classes according to the distinct eigenvector localizations of the leading eigenvalue of the adjacent matrix<sup>44</sup>, i.e., (i) localized hub networks (LHNs) in which the leading eigenvalue of the adjacent matrix  $\Lambda_A$  is closer to  $\sqrt{k_{\text{max}}}$  than  $\langle k^2 \rangle / \langle k \rangle$ , where  $k_{\text{max}}$  is the maximum degree of the network (the eigenvector is localized on the hub nodes), and (ii) localized  $k$ -core networks (LKNs) in which  $\Lambda_A$  is closer to  $\langle k^2 \rangle / \langle k \rangle$  than  $\sqrt{k_{\text{max}}}$  (the eigenvector is localized on nodes with a high  $k$ -core index).

**Uncorrelated configuration networks.** Figure 1 shows a systematic study of the SIR model on uncorrelated configuration networks. We focus on size  $N$  scale-free networks with power-law degree distributions, i.e.,  $P(k) \sim k^{-\nu_D}$ , where  $\nu_D$  is the degree exponent. The minimum degree is  $k_{\text{min}} = 3$ , and the maximum degree  $k_{\text{max}}$  is set at  $\sqrt{N}$ , which ensures that there will be no degree-degree correlations in the thermodynamic limit. Without lack of generality, we can set  $\gamma = 1$  in simulations. Two values,  $\nu_D = 2.1$  and  $\nu_D = 3.5$ , are considered. According to definition<sup>44</sup>, networks with  $\nu_D = 2.1$  are LKNs and networks with  $\nu_D = 3.5$  are LHNs. Figure 1 shows that predictions from the MFL ( $\lambda_c^{\text{MFL}}$ ) and DMP ( $\lambda_c^{\text{DMP}}$ ) methods in general produce similar theoretical values and perform better than the prediction from the QMF ( $\lambda_c^{\text{QMF}}$ ) method. When  $\nu_D = 2.1$ , the absolute errors in the epidemic threshold from the MFL and DMP methods are very small for all values of  $N$ , and the absolute errors from the QMF method decrease with  $N$ . The absolute error for method  $u \in \{\text{MFL}, \text{QMF}, \text{DMP}\}$  is  $\Delta(\lambda_c^u) = |\lambda_c^u - \lambda_c|$ .



**Figure 3. The effects of inverse participation ratio (IPR) of the adjacency and the nonbacktracking matrices on the accuracy of theoretical predictions.** (a) The relative errors and (b) absolute errors as a function of IPR of the principal eigenvectors of the adjacency (black squares) and the nonbacktracking matrices (red circles). The inset of (b) is the average absolute errors as a function of IPR.

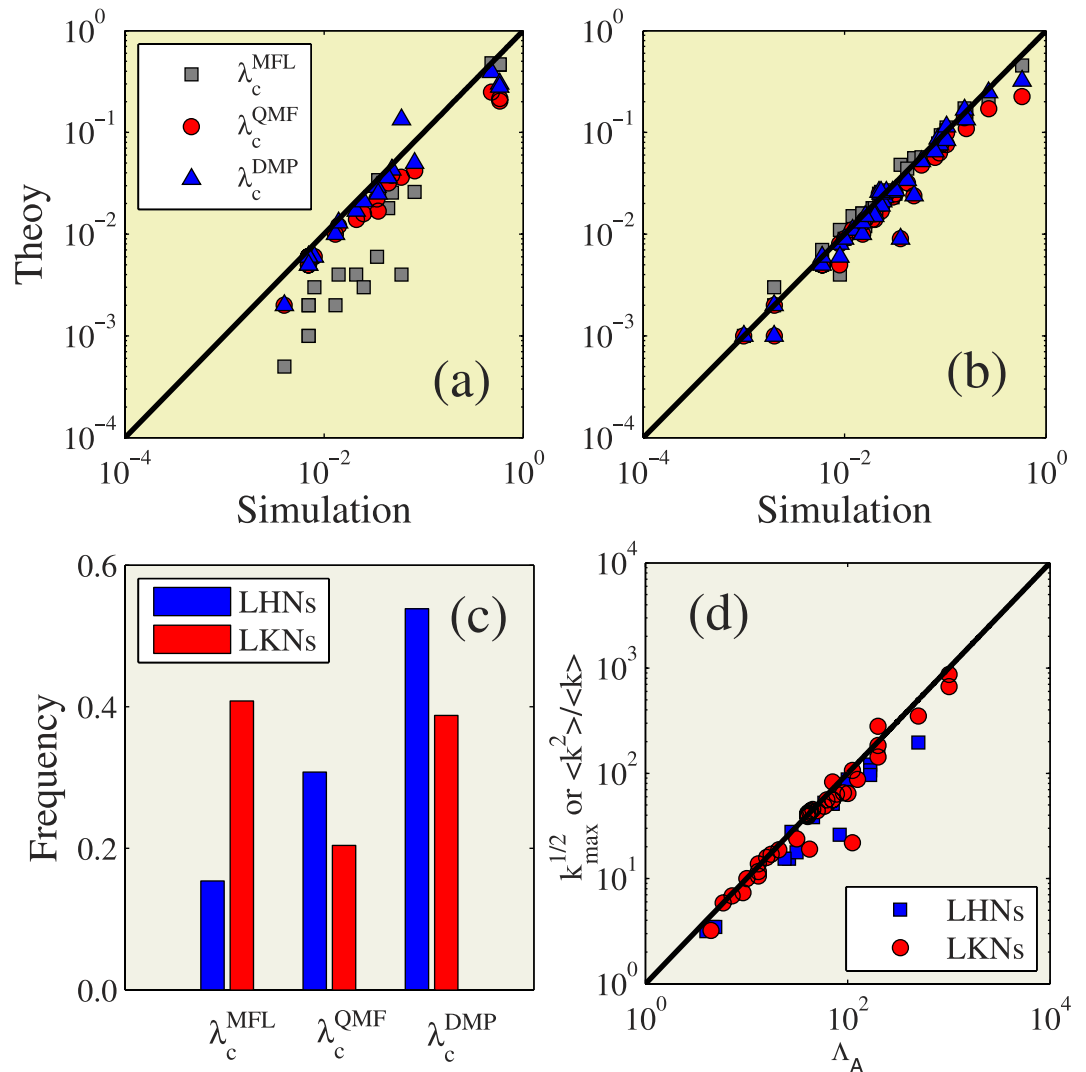
When  $\nu_D = 3.5$ , the absolute error from the QMF method stabilizes to finite values even in infinitely large networks, and the absolute errors for the MFL and DMP methods decrease with  $N$ . From these results we find that the performance of the QMF method is counterintuitive, i.e., that its performance is even worse than the MFL method. At the same time, all of these results confirm the relationships among the three theoretical predictions for uncorrelated networks previously discussed.

**Real-world networks.** We now examine the performances of the three theoretical predictions  $\lambda_c^{MFL}$ ,  $\lambda_c^{QMF}$  and  $\lambda_c^{DMP}$  on a group of 56 real-world networks of various types, e.g., social networks, citation networks, infrastructure networks, computer networks, and metabolic networks. The Supporting Information supplies additional statistical information about these real-world networks. Note that spreading processes are performed on giant connected clusters. At times, for the sake of simplicity, we treat the directed networks as undirected and the weighted networks as unweighted.

Figure 2(a) shows the accuracy of  $\lambda_c^{MFL}$ ,  $\lambda_c^{QMF}$ , and  $\lambda_c^{DMP}$  when applied to the 56 networks. Each symbol marks a theoretical prediction versus a numerical network prediction. We compute the relative frequency of  $\lambda_c^{MFL}$ ,  $\lambda_c^{QMF}$ , and  $\lambda_c^{DMP}$  to determine which one produces a value closest to  $\lambda_c$  [see Fig. 2(b)]. Because the DMP method considers the full information of network topology and also some dynamical correlations,  $\lambda_c^{DMP}$  is the best prediction in more than 40% of the networks. The  $\lambda_c^{QMF}$  value is the closest to the actual epidemic threshold in 25% of the networks, and the epidemic threshold predicted by the MFL method, which uses the degree distribution as the only input parameter, is closest to the real epidemic threshold in about one-third of the real-world networks. Comparing these three predictions we find that the DMP method outperforms the other two, i.e., when determining the epidemic threshold in a general network, the DMP method is more frequently accurate than the other two.

Theoretical predictions  $\lambda_c^{MFL}$  given by the MFL method often fail because it neglects much structural information and also all dynamical correlations. The performance of the QMF method is counterintuitive because of the localized eigenvector of the leading eigenvalue of the adjacent matrix [see Fig. 3(a)]. Figure 3 shows the effects of the inverse participation ratios (IPR)<sup>39,45</sup> of the adjacent and non-backtracking matrixes. We find that the relative and absolute errors between the theoretical and numerical predictions increase with IPR, i.e., the QMF and DMP methods deviate from the accurate epidemic threshold more easily when IPR is large because the eigenvector centralities of adjacent and non-backtracking matrixes are localized on hub nodes or high  $k$ -core index nodes<sup>44</sup>. The relative error of method  $u \in \{MFL, QMF, DMP\}$  can be  $\Delta'(\lambda_c^u) = |\lambda_c - \lambda_c^u|/\lambda_c$ .

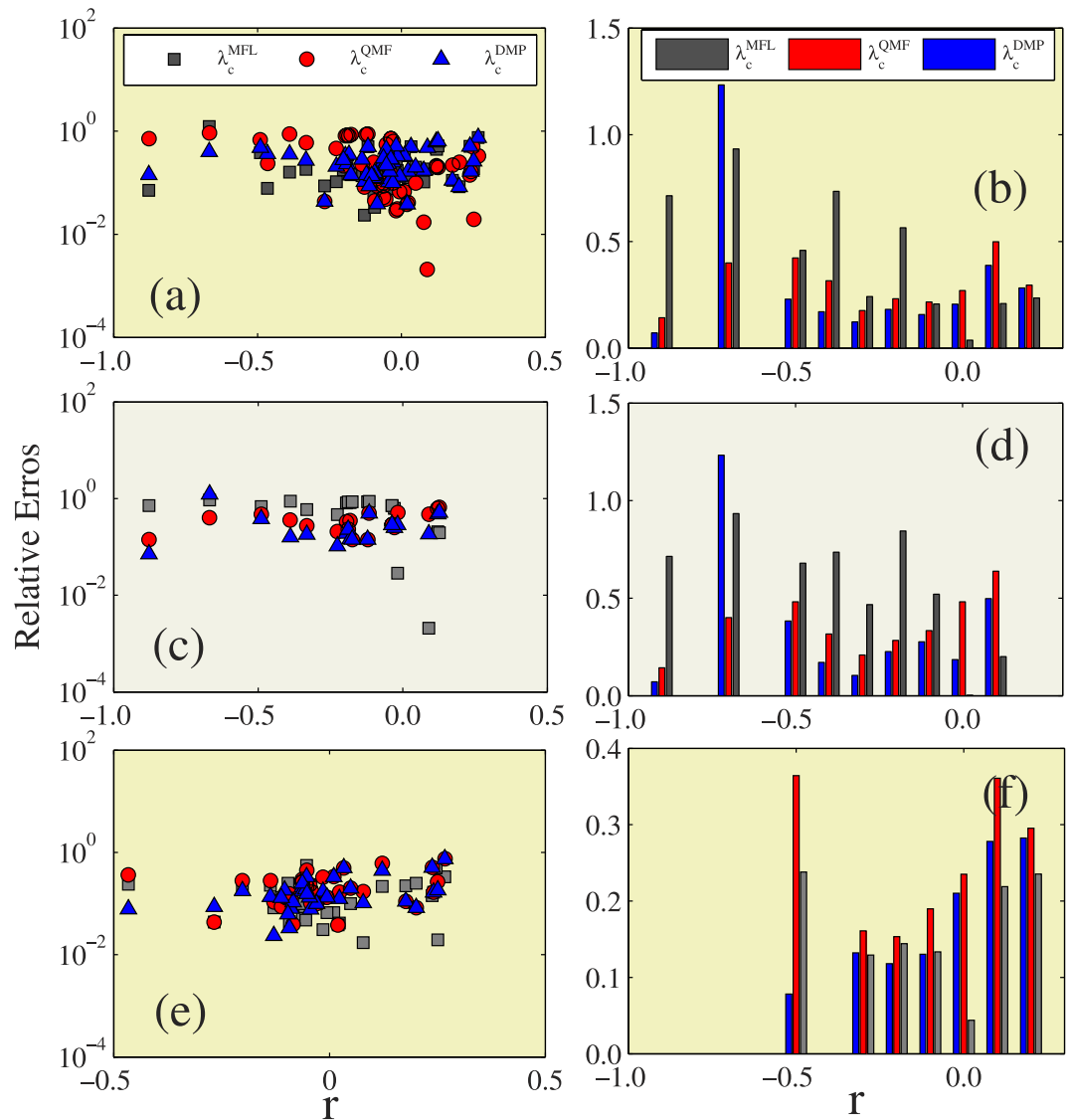
Recent research results indicate that networks have distinct eigenvector localizations<sup>44</sup>. In real-world networks they are either localized on hubs networks (LHNs) or localized on  $k$ -core networks (LKNs). Depending on the



**Figure 4.** Verify the accuracy for three types of theoretical epidemic threshold on real-world networks. The theoretical predictions of  $\lambda_c^{\text{MFL}}$  (gray squares),  $\lambda_c^{\text{QMF}}$  (red circles) and  $\lambda_c^{\text{DMP}}$  (blue up triangles) versus numerical predictions  $\lambda_c$  of the epidemic threshold on (a) LHNs and (b) LKNs. (c) In the collective of LHNs and LKNs of real-world networks, the fraction of  $\lambda_c^{\text{MFL}}$  [ $\lambda_c^{\text{QMF}}$  or  $\lambda_c^{\text{DMP}}$ ] is the closest value to  $\lambda_c$ . (d) The values of  $k_{\text{max}}^{1/2}$  for LHNs and  $\langle k^2 \rangle / \langle k \rangle$  for LKNs versus the leading eigenvalue  $\Lambda_A$  of the adjacent matrix.

localization of the eigenvector of adjacent matrix, there are 19 LHNs and 37 LKNs among the 56 real-world networks. Figure 4(d) shows that the values  $\Lambda_A$  of LHNs are close to  $k_{\text{max}}^{1/2}$  (blue squares), and the values  $\Lambda_A$  of LKNs are close to  $\langle k^2 \rangle / \langle k \rangle$  (red circles). In LHNs [see Fig. 4(a,c)] the three methods perform as we would expect. The DMP method is the best predictor and the MFL method the worst because it neglects much detailed network structure information. In contrast, in the LKNs [see Fig. 4(b,c)], the simple MFL method performs the best, and it is slightly accurate than the DMP method.

We now compare the accuracy between the three theoretical epidemic thresholds under different microscopic and mesoscale topologies of real-world structures, including degree-degree correlations  $r$ , clustering  $c$ , and modularity  $Q$ . To measure the accuracy of the three methods in each theoretical prediction, we compute the average relative errors in the interval  $(x - \Delta x/2, x + \Delta x/2)$ , where  $x$  is  $r$ ,  $c$ , and  $Q$ . Here we set  $\Delta x = 0.1$  unless otherwise specified. Figure 5(a,b) show that in all cases except the Facebook (NIPS) network the DMP method has a lower relative error when the Pearson correlation coefficient value is  $r < 0$ . The Facebook (NIPS) network may be an exception because the IPR value of its non-backtracking matrix is relatively large, i.e., 0.012. When  $r < 0$ , we can conclude that the DMP method performs the best and the MFL method performs the worst. When  $r > 0$ , the MFL method is the most accurate and the QMF method is the least. Figure 5(c–f) show the 56 real-world networks, separating them according to eigenvector localization. In LHNs we see a phenomenon similar to that shown in Fig. 5(a,b), i.e., when  $r < 0$  the DMP method is the most accurate and the MFL method is the least, but when  $r > 0$  the MFL method is the most accurate and the QMF method is the least. In LKNs, when  $r < 0$  the DMP method is the most accurate, when  $r > 0$  the MFL method is the most accurate, and the QMF method is always the least.

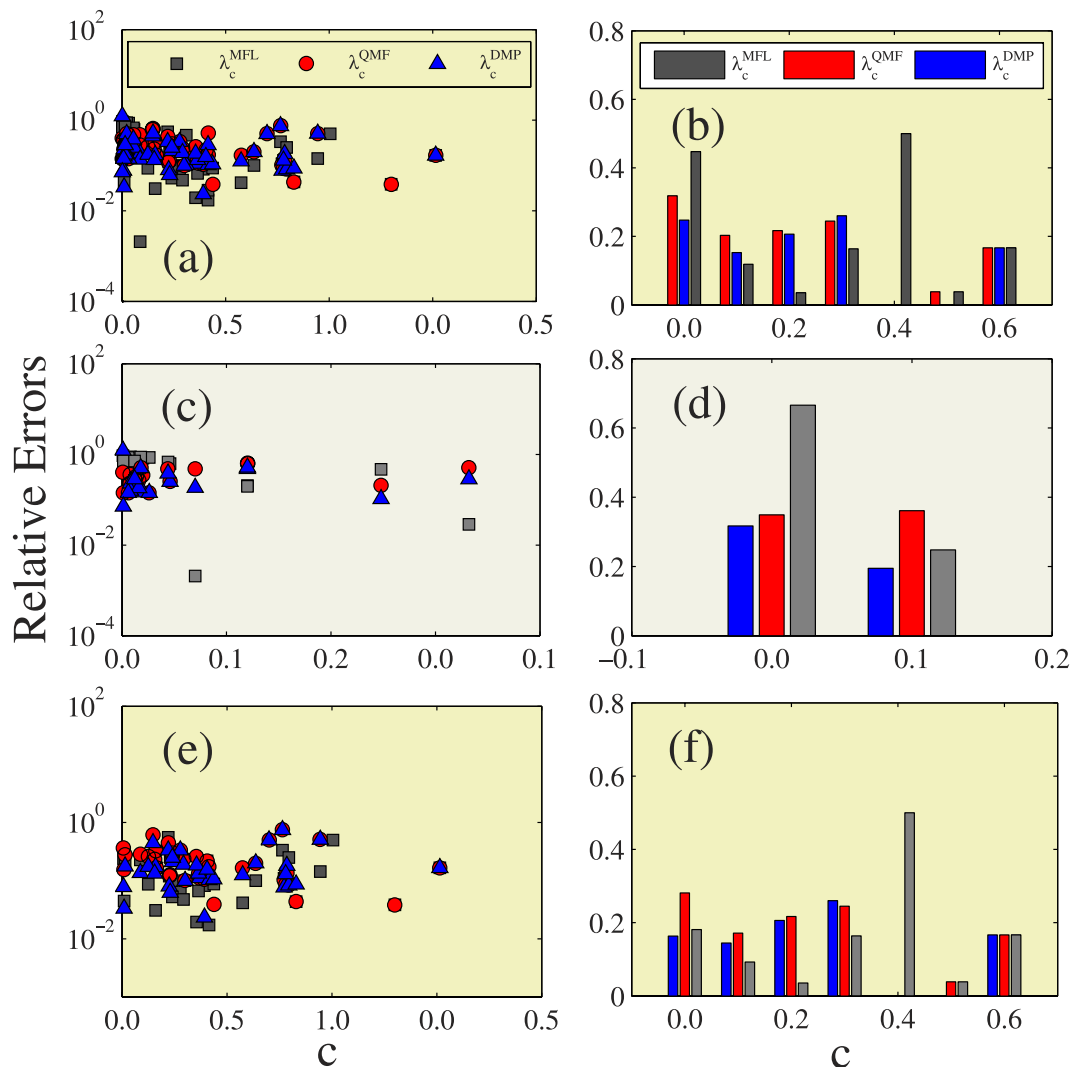


**Figure 5. Effects of degree-degree correlations on the relative errors of different theoretical predictions.** In the first column, figures (a,c,e) are the the relative errors of the three different theoretical predictions versus degree-degree correlations  $r$ . In the second column, figures (b,d,f) are the the average relative errors for the three different theoretical predictions versus  $r$ . The first row exhibits the results of 56 real-world networks, the second row shows the results of LHNs, the third row performs the results of the LKNs.

accurate. This suggests that the MFL method is the best for predicting epidemic thresholds in networks with positive degree-degree correlations, but that the DMP method is better in all other cases.

Using an analytic framework similar to that shown in Fig. 5, we compare the accuracy among the three theoretical predictions under different clustering coefficient  $c$  in Fig. 6. Figure 6(a,b) show that when  $c < 0.1$ , the relative error of the DMP method is the lowest and the relative error of the MFL method is the largest. When  $c > 0.1$ , the relative error of the MFL method is the lowest and the relative error of the QMF method is, in most cases, the largest. Thus when  $c < 0.1$  the DMP method is the most accurate in predicting the epidemic threshold, but when  $c > 0.1$  the MFL method is the most accurate. In LHNs, we find the same phenomena as shown in Fig. 6(a,b). The DMP method is the best predictor when  $c < 0.1$ , and the MFL method the best when  $c > 0.1$  [see Fig. 6(c,d)]. Figure 6(e,f) show that in LKNs the DMP method performs the best for small  $c$  and the MFL method the best for large  $c$ .

Finally, Fig. 7 compares the effectiveness between the three predictions under different modularity  $Q$ . Note that in real-world networks the relative errors increase with  $Q$ . In the 56 networks, in LHNs, and in LKNs, we note that the performances of the three predictions do not exhibit an obvious regularity versus the modularity, and in most cases the DMP method performs better than other two.



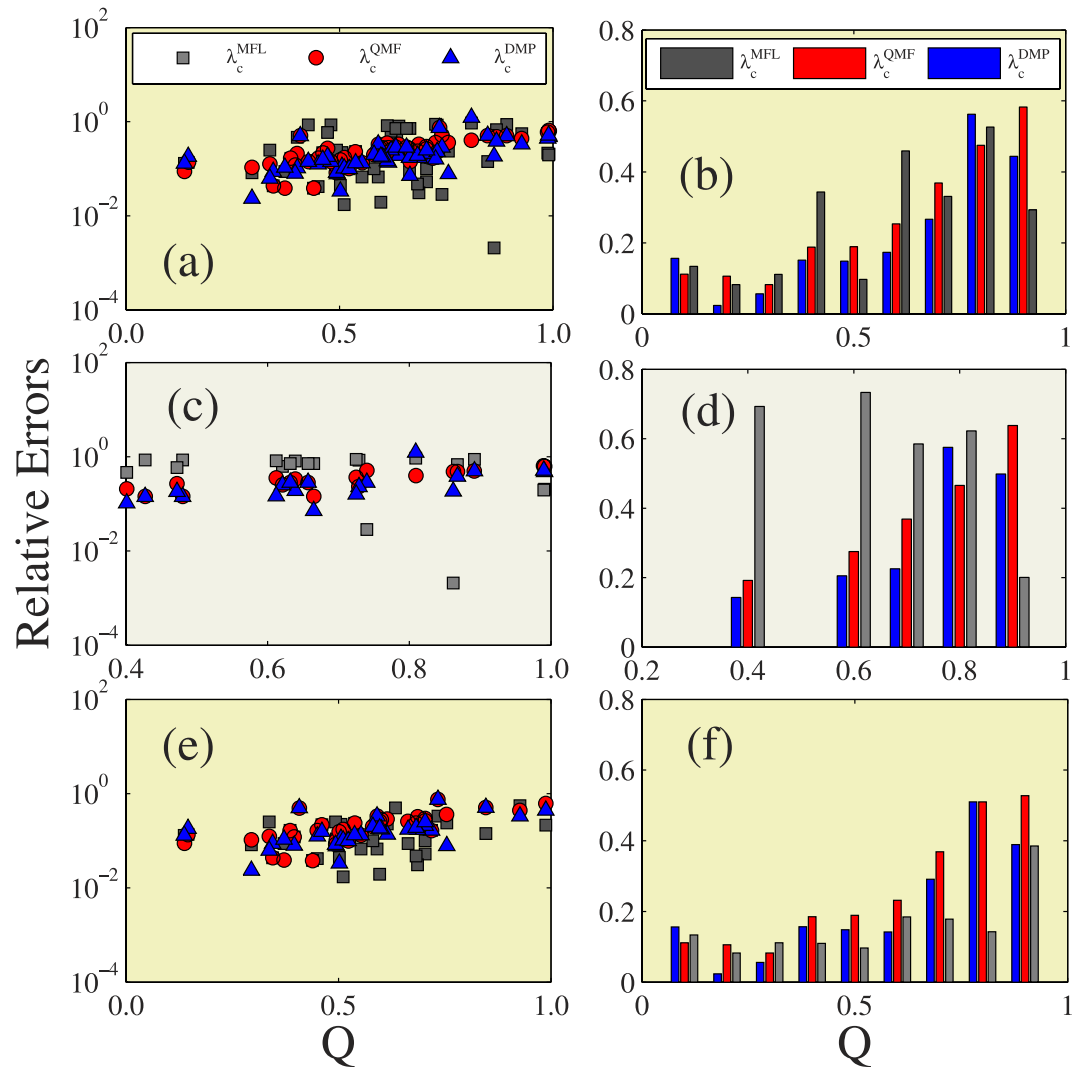
**Figure 6.** Effects of clustering on the relative errors of different theoretical prediction. In the first column, figures (a,c,e) are the the relative errors of the three different theoretical predictions versus clustering  $c$ . In the second column, figures (b,d,f) are the the average relative errors for the three different theoretical predictions versus  $c$ . The first row exhibits the results of 56 real-world networks, the second row shows the results of LHNs, the third row performs the results of the LKNs.

## Conclusions

In this study we have systematically examined the accuracies and relationships among the MFL, QMF, and DMP methods for predicting the epidemic threshold in the SIR model. To do this we have focused on a large number of artificial network simulations and on 56 real-world networks. We first analyzed the differences and correlations among the three theoretical epidemic threshold predictions. Generally speaking, the three predictions differ, and the epidemic threshold predicted by the DMP method is often larger than that predicted by the QMF method. In uncorrelated networks, the DMP and MFL methods produce the same epidemic threshold prediction, which is larger than the prediction produced by the QMF method. When applied to real-world networks, however, the relationships among the three predictions are still unclear.

We then checked the accuracies of the three predictive methods using uncorrelated configuration networks, and found that the MFL and DMP methods perform well, but that the QMF method does not. In the group of 56 real-world networks we found that the DMP method performs the best in most occasions, and that the epidemic threshold predicted by the MFL method is more accurate than the one predicted by the QMF method in most of the networks. In networks with an eigenvector localized on high  $k$ -core nodes, i.e., LKNs, the MFL method performs the best and the QMF method the worst, but in networks with an eigenvector localized on hubs, i.e., LHNs, the DMP method performs the best and the MFL method the worst.

Finally we measured the performances of the three methods versus the microscopic and mesoscale topologies in the 56 real-world networks, including degree-degree correlations  $r$ , clustering  $c$ , and modularity  $Q$ . For this purpose, we compute the average relative errors between theoretical thresholds and accurate thresholds for the networks in the interval  $(x - \Delta x/2, x + \Delta x/2)$ , where  $x$  is  $r$ ,  $c$ , and  $Q$ . The smaller value of the relative error



**Figure 7. Effects of modularity on the relative errors of different theoretical prediction.** In the first column, figures (a,c,e) are the the relative errors of the three different theoretical predictions versus modularity  $Q$ . In the second column, figures (b,d,f) are the the average relative errors for the three different theoretical predictions versus  $Q$ . The first row exhibits the results of 56 real-world networks, the second row shows the results of LHNs, the third row performs the results of the LKNs.

indicates the better performance of the theory. In networks with negative degree-degree correlations, we found that the DMP method performs the best, and the QMF method performs than the MFL method. In the networks with positive degree-degree correlations, the MFL method is the most accurate, and the QMF method is the least. In networks with low clustering, the DMP method is the most accurate, and the MFL method is the least. In networks with high clustering, the MFL method is the most accurate, and the QMF method is the least. The relative accuracies of the three predictions versus the modularity are, unfortunately, irregular.

Predicting accurate epidemic thresholds in networks is profoundly significant in the field of spreading dynamics. Our results present a counterintuitive insight into the use of network information in theoretical methods, i.e., the performance level of a method is not only proportional to the topological information used, but also correlates with the dynamical correlations among the states of neighbor nodes. Our results expand our understanding of epidemic thresholds and provide ways of determining which method of theoretical prediction is best in a variety of given situations. Our results also indicate directions for further research into the development of more accurate theoretical methods of predicting epidemic thresholds. It should be noted that we just considered the SIR spreading dynamics with synchronous updating method, whether or not the results apply to the case with asynchronous updating method needs to be further studied. Some further investigations about the effects of network structural characteristics (e.g., degree-degree correlations) on the accuracy of the theoretical methods are still called for. For instance, one can study the effect of degree-degree correlations on the accuracy of the three theoretical methods by changing the degree-degree correlations<sup>46</sup> of the configuration model gradually (see details in Supporting Information).

## Methods

**Predicting numerical threshold.** To determine the theoretical epidemic threshold, we employ the estimators supplied by the MFL, QMF and DMP methods and use the relative variance  $\chi$  to numerically determine the size-dependent epidemic threshold<sup>47</sup>,

$$\chi = \frac{\langle r - \langle r \rangle \rangle^2}{\langle r^2 \rangle}, \quad (15)$$

where  $r$  denotes the final epidemic size and  $\langle \dots \rangle$  is the ensemble averaging. We use at least  $10^5$  independent dynamic realizations on a network to calculate the average value of  $\chi$ , which exhibits a maximum value at the epidemic threshold  $\lambda_c$ . This numerical prediction  $\lambda_c$  obtained by observing  $\chi$  we consider the accurate epidemic threshold<sup>47</sup>. The Supporting Information supplies illustrations of numerically locating the epidemic threshold by observing  $\chi$ . There are also other ways of determining  $\lambda_c$ , e.g., susceptibility<sup>27</sup> and variability methods<sup>48</sup>.

## References

- Pastor-Satorras, R. *et al.* Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**, 925 (2015).
- Castellano, C., Fortunato, S. & Fortunato, S. Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 0034 (2009).
- Rocha, L. E. C., Liljeros, F. & Holme, P. Information dynamics shape the sexual networks of Internet-mediated prostitution. *Proc. Natl. Acad. Sci.* **107**, 5706 (2010).
- Hu, H. *et al.* WiFi networks and malware epidemiology. *Proc. Natl. Acad. Sci.* **106**, 1318 (2009).
- Newman, M. E. J. Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Phys. Rev. E* **66**, 035101(R) (2002).
- Dorogovtsev, S. N., Godtsev, A. V. & Mendes, J. F. F. Critical phenomena in complex networks. *Rev. Mod. Phys.* **80**, 1275 (2008).
- Moreno, Y., Pastor-Satorras, R. & Vespignani, A. Epidemic outbreaks in complex heterogeneous networks. *Eur. Phys. J. B* **26**, 521 (2002).
- Cohen, R., Havlin, S. & ben-Avraham, D. Efficient Immunization Strategies for Computer Networks and Populations. *Phys. Rev. Lett.* **91**, 247901 (2003).
- Wang, W. *et al.* Asymmetrically interacting spreading dynamics on complex layered networks. *Sci. Rep.* **4**, 5097 (2014).
- Mata, A. S. *et al.* Lifespan method as a tool to study criticality in absorbing-state phase transitions. *Phys. Rev. E* **91**, 052117 (2015).
- A. Muñoz, M. *et al.* Griffiths Phases on Complex Networks. *Phys. Rev. Lett.* **105**, 128701 (2010).
- Kitsak, M. *et al.* Identification of influential spreaders in complex networks. *Nat. Phys.* **6**, 888 (2010).
- Pastor-Satorras, R. & Vespignani, A. Epidemic dynamics and endemic states in complex networks. *Phys. Rev. E* **63**, 066117 (2001).
- Newman, M. E. J. Spread of epidemic disease on networks. *Phys. Rev. E* **66**, 016128 (2002).
- Volz, E. M. *et al.* Effects of Heterogeneous and Clustered Contact Patterns on Infectious Disease Dynamics. *PLoS Comput. Biol.* **7**, e1002042 (2011).
- Miller, J. C., Slim, A. C. & Volz, E. M. Edge-based compartmental modelling for infectious disease spread. *J. R. Soc. Interface* **9**, 890 (2012).
- Wang, W. *et al.* Epidemic spreading on complex networks with general degree and weight distributions. *Phys. Rev. E* **90**, 042803 (2014).
- Wang, W. *et al.* Dynamics of social contagions with memory of nonredundant information. *Phys. Rev. E* **92**, 012820 (2015).
- Eames, K. T. D. & Keeling, M. J. Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *Proc. Natl. Acad. Sci.* **99**, 13330 (2002).
- Gross, T., D'Lima, C. J. D. & Blasius, B. Epidemic Dynamics on an Adaptive Network. *Phys. Rev. Lett.* **96**, 208701 (2006).
- Gómez, S. *et al.* Discrete-time Markov chain approach to contact-based disease spreading in complex networks. *Europhys. Lett.* **89**, 38009 (2010).
- Van Mieghem, P. *Graph spectral for complex networks* (Cambridge University Press, 2011).
- Karrer, B. & Newman, M. E. J. *Phys. Rev. E* Message passing approach for general epidemic models. **82**, 016101 (2010).
- Boguñá, M., Pastor-Satorras, R. & Vespignani, A. Absence of Epidemic Threshold in Scale-Free Networks with Degree Correlations. *Phys. Rev. Lett.* **80**, 028701 (2003).
- Serrano, M. Á. & Boguñá, M. Percolation and Epidemic Thresholds in Clustered Networks. *Phys. Rev. Lett.* **97**, 088701 (2006).
- Newman, M. E. J. Random Graphs with Clustering. *Phys. Rev. Lett.* **103**, 058701 (2009).
- Ferreira, S. C., Castellano, C. & Pastor-Satorras, R. Epidemic thresholds of the susceptible-infected-susceptible model on networks: A comparison of numerical and theoretical results. *Phys. Rev. E* **86**, 041125 (2012).
- Shrestha, M., Scarpino, S. V. & Moore, C. A message-passing approach for recurrent-state epidemic models on networks. *Phys. Rev. E* **92**, 022821 (2015).
- Li, C., van de Bovenkamp, R. & Van Mieghem, P. Susceptible-infected-susceptible model: A comparison of N-intertwined and heterogeneous mean-field approximations. *Phys. Rev. E* **86**, 026116 (2012).
- Schonfisch, B. & De Roos, A. Synchronous and asynchronous updating in cellular automata. *Bio. Syst.* **51**, 123 (1999).
- Castellano, C. & Pastor-Satorras, R. Thresholds for Epidemic Spreading in Networks. *Phys. Rev. Lett.* **105**, 218701 (2010).
- Chakrabarti, D. *et al.* Epidemic Thresholds in Real Networks. *ACM Trans. Inf. Syst. Secur.* **10**, 1 (2008).
- Van Mieghem, P., Omic, J. & Kooij, R. Virus Spread in Networks. *IEEE ACM Trans. Netw.* **17**, 1 (2009).
- Parshani, R., Carmi, S. & Havlin, S. Epidemic Threshold for the Susceptible-Infectious-Susceptible Model on Random Networks. *Phys. Rev. Lett.* **104**(25), 258701 (2010).
- Lokhov, A. Y., Mézard, M. & Zdeborová, L. Dynamic message-passing equations for models with unidirectional dynamics. *Phys. Rev. E* **91**, 012811 (2015).
- Radicchi, F. Predicting percolation thresholds in networks. *Phys. Rev. E* **91**, 010801(R) (2015).
- Martin, T., Zhang, X. & Newman, M. E. J. Localization and centrality in networks. *Phys. Rev. E* **90**, 052808 (2014).
- Krzakala, F. *et al.* Spectral redemption in clustering sparse networks. *Proc. Natl. Acad. Sci.* **110**, 20935 (2013).
- Karrer, B., Newman, M. E. J. & Zdeborová, L. Percolation on Sparse Networks. *Phys. Rev. Lett.* **113**, 208702 (2014).
- Lagorio, C. *et al.* Effects of epidemic threshold definition on disease spread statistics. *Physica A* **388**, 755–763 (2009).
- Kenah, E. & Robins, J. M. Second look at the spread of epidemics on networks. *Phys. Rev. E* **76**, 036113 (2007).
- Van Mieghem, P. & van de Bovenkamp, R. Non-Markovian infection spread dramatically alters the SIS epidemic threshold in networks. *Phys. Rev. Lett.* **110**(10), 108701 (2013).
- Chung, F., Lu, L. & Vu, V. Spectra of random graphs with given expected degrees. *Proc. Natl. Acad. Sci. USA* **100**, 6313–6318 (2003).
- Pastor-Satorras, R. & Castellano, C. Distinct types of eigenvector localization in networks. arXiv:1505.06024v1 (2015).
- Goltsev, A. V. *et al.* Localization and Spreading of Diseases in Complex Networks. *Phys. Rev. Lett.* **109**, 128702 (2012).
- Van Mieghem, P. *et al.* Influence of assortativity and degree-preserving rewiring on the spectra of networks. *Eur. Phys. J. B* **76**(4), 643–652 (2010).
- Chen, W., Schroder, M. & D'Souza, M. R. Microtransition Cascades to Percolation. *Phys. Rev. Lett.* **112**, 155701 (2014).
- Shu, P. *et al.* Numerical identification of epidemic thresholds for susceptible-infected-recovered model on finite-size networks. *Chaos*, **25**, 063104 (2015).

## Acknowledgements

This work was partially supported by the National Natural Science Foundation of China under Grants Nos 11105025, 11575041 and 61433014, and the Program of Outstanding Ph. D. Candidate in Academic Research by UESTC under Grand No. YXBSZC20131065.

## Author Contributions

W.W. and M.T. devised the research project. W.W., Q.-H.L. and L.-F.Z. performed numerical simulations. W.W., Q.-H.L., L.-F.Z., M.T., H.G. and H.E.S. analyzed the results. W.W., Q.-H.L., L.-F.Z., M.T., H.G. and H.E.S. wrote the paper.

## Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Wang, W. *et al.* Predicting the epidemic threshold of the susceptible-infected-recovered model. *Sci. Rep.* **6**, 24676; doi: 10.1038/srep24676 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>