Published in "Physical Review E - Statistical, nonlinear, and
soft matter physics 85(3): 036101 , 2012"
which should be cited to refer to this work.

Removing spurious interactions in complex networks

An Zeng and Giulio Cimini
Department of Physics, University of Fribourg, Chemin du Musée 3, CH-1700 Fribourg, Switzerland

Identifying and removing spurious links in complex networks is meaningful for many real applications and is crucial for improving the reliability of network data, which, in turn, can lead to a better understanding of the highly interconnected nature of various social, biological, and communication systems. In this paper, we study the features of different simple spurious link elimination methods, revealing that they may lead to the distortion of networks' structural and dynamical properties. Accordingly, we propose a hybrid method that combines similarity-based index and edge-betweenness centrality. We show that our method can effectively eliminate the spurious interactions while leaving the network connected and preserving the network's functionalities.

I. INTRODUCTION

Many social, biological, and information systems are naturally described by networks where nodes represent individuals, proteins, genes, computers, web pages, and so on, and links denote the relations or interactions between nodes. Hence, network analysis has become a crucial focus in many fields including biology, ecology, technology, and sociology [1]. However, the reliability of network data is not always guaranteed: Biological networks that are inferred from experiments or social networks that result from spontaneous human activity may contain inaccurate and misleading information, resulting in missing and spurious links [2,3].

The problem of identifying missing interactions, known as *link prediction*, consists of estimating the likelihood of the existence of a link between two nodes according to the observed links and node's attributes [4]. Link prediction has already attracted much attention from disparate research communities due to its broad applicability. For instance, in many biological networks (such as food webs, protein-protein interactions, and metabolic networks), the discovery of interactions is often difficult and expensive, hence, accurate predictions can reduce the experimental costs and speed the pace of uncovering the truth [5,6]. Applications in social networks include the prediction of the actors co-starring in acts [7] and of the collaborations in coauthorship networks [8], the detection of the underground relationships between terrorists [5], and many others. In addition, the process of recommending items to users can be considered as a link prediction problem in a user-item bipartite graph [9] so that similarity-based link prediction techniques have been applied to personalized recommendations [10]. Moreover, the link prediction approach can be used to solve the classification problem in partially labeled networks, such as predicting protein functions [11], detecting anomalous email [12], distinguishing the research areas of scientific publications [13], and finding out the fraud and legitimate users in cell phone networks [14]. For a review of the field, see Ref. [15].

On the other hand, the problem of identifying spurious interactions has received less attention despite its numerous potential applications. For instance, the identification of inactive connections in social networks or spam hyperlinks on the World Wide Web (WWW) may improve the efficiency of link-based ranking algorithms [16], and the detection of redundant interactions in biological, communication, or citation networks may find applications in community detection, in constructing networks' backbones [17] or in other connection optimization problems. A possible reason for the lack of effective methods to deal with this problem is that a spurious link removal error has far more serious consequences than a missing link addition one. If some unexpected links are incorrectly identified as spurious and are removed from the network, the system's structure and function may be altered significantly or may even be compromised. For instance, the network may break up into separate components so that the system's functionality is destroyed. In power grids, only the power plants in the giant component (GC) can work [18]. In traffic systems, only the cities in the GC can mutually communicate [19]. In neural systems, only neurons in the GC can reach a synchronized state in order to effectively process signals [20]. Hence, the main challenge for a spurious link detection method is to identify the spurious interactions and, at the same time, to construct a network with close functionalities to the original one.

In this paper, we show that many simple spurious link detection methods indeed have a serious drawback to remove real and important links, which causes the networks' structure to be altered significantly. Hence, we propose a hybrid algorithm that combines a similarity-based index known as common neighbors (CN) with the edge-betweenness (EB) centrality. We show that this method cannot only effectively identify and remove spurious links, but also preserve the size of the GC and many important structural and dynamical properties of the network at the same time.

II. METHOD

In this section, we describe our procedure to study the features and to evaluate the performance of a spurious link detection algorithm. We make use of six empirical undirected networks: the *Caenorhabditis elegans* (CE) neural network [21], an email (Email) network [22], a scientists' coauthorship (SC) network [23], the U.S. political blogs' (PB) network [24], a protein-protein interaction (PPI) network [25], and the U.S. air transportation (USAir) network [26]. We only consider the GC of these real networks. Some properties of these systems are reported in Table I. All of these networks are widely used in the literature as model systems, hence, we assume that they are true networks (i.e., without spurious interactions),

TABLE I. Features of empirical networks: number of nodes (N) and edges (E), average degree ($\langle k\rangle$), average shortest path length ($\langle d\rangle$), clustering coefficient (C), degree assortativity (r), degree heterogeneity ($H = \langle k^2\rangle/\langle k\rangle^2$), and traffic congestability ($B_{\max}$).

	N	E	$\langle k\rangle$	$\langle d\rangle$	C	r	H	$B_{\max}$
CE	297	2148	14.46	2.46	0.308	−0.163	1.801	2.65×10^4
Email	1133	5451	9.62	3.61	0.220	0.078	1.942	5.06×10^4
SC	379	914	4.82	4.93	0.798	−0.082	1.663	5.66×10^4
PB	1222	16 717	27.36	2.51	.360	−0.221	2.970	1.46×10^5
PPI	2375	11 693	9.85	4.59	0.388	0.454	3.476	8.98×10^5
USAir	332	2126	12.81	2.46	0.749	−0.208	3.464	2.28×10^4

which we denote as A^t. We then add a fraction f of spurious random connections to these true networks to obtain observed networks, which we denote as A^o, and evaluate the ability of the spurious link detection algorithm to recover the features of the true networks.

To quantify the accuracy of the algorithm in identifying the spurious interactions, we use the standard metric of the area under the receiver operating characteristic curve (AUC) [27]. Since the algorithm returns an ordered list of links (or equivalently, gives each link a score to quantify its reliability), the AUC represents the probability that a spurious link is ranked lower than a true link. To obtain the value of the AUC, we pick a spurious link and a true link in the observed network A^o and compare their scores. If, among all possible pairs n [28], the real link has a higher score than the spurious link n' times and equal score n'' times, the AUC value is as follows:

$$\mathrm{AUC} = \frac{n' + n''/2}{n}.$$

Note that, if links were ranked at random, the AUC value would be equal to 0.5.

As stated in Sec. I, high accuracy is not sufficient for a spurious link detection algorithm: If just a few real important links are removed, the structural and dynamical properties of the network may change dramatically. A simple example can be seen in Fig. 1. If the dashed link is removed, the network will break into two separated components. To study the robustness of the algorithm in this respect, we remove the fraction f' of the bottom-ranked links from the observed network to obtain the reconstructed network, which we denote as A^r. We then compare the structure and functionality of true and reconstructed networks. We will focus mainly on the GC's size, which is of great importance for the functionality of many real systems. Then, we will consider the clustering coefficient [29], average shortest path length, traffic congestability [30] (i.e., the maximum betweenness centrality in the network), and other dynamical properties. We will first study the case of A^t and A^r having the same number of links ($f' = f$). However, as, in general, one does not know how many spurious links there are in a given network, we finally consider the situation where $f' \neq f$.

III. RELIABILITY INDICES

In this section, we describe some representative spurious link detection methods. These algorithms assign a reliability index (denoted as R_{ij} for the link connecting nodes i and j) to each link in A^o, which quantifies the likelihood of its true existence and allows for link ranking.

Similarity-based indices use the network's structure to assign a score for each pair of connected nodes ij, which is directly defined as their similarity with the underlying assumption that a connection between similar nodes is likely to be a true one. These algorithms can be classified as local, quasilocal, and global according to the amount of information they need. Here are some examples:

(1) Common neighbors (CN): $R_{ij}^{\mathrm{CN}} = \|\Gamma_i \cap \Gamma_j\|$, where Γ_i is the set of neighbors of node i and $\|\cdots\|$ indicates the number of nodes in a set.

(2) Resource allocation (RA): $R_{ij}^{\mathrm{RA}} = \sum_{k\in\Gamma_i\cap\Gamma_j} \frac{1}{\|\Gamma_k\|}$.

(3) Local path (LP): $R_{ij}^{\mathrm{LP}} = (A^2)_{ij} + \epsilon\,(A^3)_{ij}$, where A is the network's adjacency matrix and $\epsilon < 1$ is a free parameter.

(4) Katz index (Katz): $R_{ij}^{\mathrm{Katz}} = \sum_{l=1}^{\infty}[(\beta\,A)^l]_{ij}$, where β is a free parameter that must be lower than the reciprocal of the largest eigenvalue of A.

Centrality-based indices measure the importance of a link in the network, assuming that the higher the link's centrality, the higher its reliability. We consider two simple indices:

(5) Preferential attachment (PA): $R_{ij}^{\mathrm{PA}} = \|\Gamma_i\| \times \|\Gamma_j\|$.

(6) Edge-betweenness (EB): $R_{ij}^{\mathrm{EB}} = \sum_{m>n} \frac{C_{mn}^{(ij)}}{C_{mn}}$, where C_{mn} is the number of shortest paths from node m to node n and $C_{mn}^{(ij)}$ is the number of such shortest paths passing through link ij.

Clearly, CN, RA, and PA are local indices. CN is the simplest possible measure of neighborhoods' overlap, whereas, RA [31] is the best performing local index for the purpose of link prediction. PA is the algorithm that requires less information. Instead, LP [32] is a quasilocal method, as it considers local paths with wider horizons than CN (it also counts the number of different paths with length 3 connecting i and j). Finally, Katz [33] and EB methods are global indices as they are based on the ensemble of all paths in the network. Specifically, Katz counts the paths between two nodes and weights them according to their length l, whereas, EB is built with the number of shortest paths from all vertices to all others that pass through the given link.

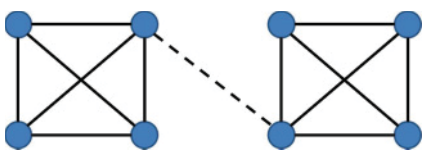

FIG. 1. (Color online) A simple example to illustrate how an improper spurious link removal method can disconnect a network.

IV. HYBRID INDEX

We now introduce a hybrid index that combines the similarity-based and the centrality-based approaches. The underlying idea is that we consider a link to be a true one either if it connects similar nodes or if it has a central position in the network. Even if this assumption is not necessarily true, as we show later, it avoids the removal of important links so that the network's properties and functions are preserved with the small drawback of failing to identify a few spurious interactions.

To construct the hybrid index, we combine the simple CN with EB centrality as

$$R_{ij}^{\mathrm{hyb}} = \lambda \frac{R_{ij}^{\mathrm{CN}}}{\max_{mn}\left(R_{mn}^{\mathrm{CN}}\right)} + (1-\lambda)\frac{R_{ij}^{\mathrm{EB}}}{\max_{mn}\left(R_{mn}^{\mathrm{EB}}\right)},$$

where $\lambda \in [0,1]$ is the hybridization parameter. In what follows, we set $\lambda = 0.9$ because we want to exploit mainly the CN, and a small contribution from EB suffices for our purposes (however, see Sec. VI for a study of the index behavior for different λ's). Note that this is only one possibility for defining such an index. We made use of the CN because it is the most well known of the similarity-based indices. However, one could use, e.g., RA or Katz instead, although the qualitative features of the hybrid method would not change.

V. RESULTS

In this section, we compare the features of the spurious link detection approaches that have been previously introduced. We start by adding a fraction f of random connections to the true networks A^t to obtain the observed networks A^o. For each particular index, we rank the links according to their reliability values and measure the accuracy of the method in identifying spurious interactions by the AUC (Fig. 2). Generally, we observe that the similarity-based methods perform better than the centrality-based ones. Among the first category, Katz and LP [34] perform slightly better than CN and RA as they take advantage of using more information. Among the second, EB is the worst performing, with AUC even lower than 0.5. Instead, the performance of the hybrid method is very close to that of the pure similarity-based indices. Hence, having a contribution from EB in the hybridization does not result in worse spurious link detection (as one might expect).

We already argued that accuracy is not the only criterion to assess the performance of these methods. The other important aspect is that the removal of putative spurious links should not alter the GC's size as well as other properties of the networks. To investigate this aspect, we remove the fraction f' of the bottom-ranked links from A^o to obtain the reconstructed networks A^r, whose features we compare with the ones of the relative true networks A^t. We start with the simple case $f' = f$, and we first focus on the GC's size, which is of great relevance in many contexts. As shown in Fig. 3, the GC's size significantly decreases with f' when using any similarity-based method (as well as PA): In these cases, many nodes become disconnected from the networks' core and end up losing their function. On the contrary, EB always keeps the networks connected. This is not surprising as it has already been pointed out [35] that similarity indices and EB are highly anticorrelated, meaning that removing links between nonsimilar nodes causes links with high betweenness to be cut and vice versa. What is remarkable is that the hybrid method also can effectively preserve the connectedness of the original networks in most of the cases and, in general, much better than any other

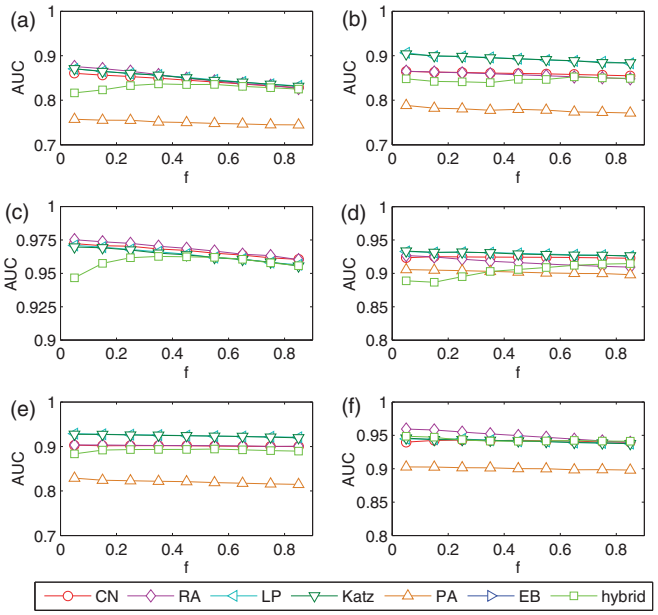

FIG. 2. (Color online) AUC for various indices and for different values of f. The true networks are (a) CE, (b) Email, (c) SC, (d) PB, (e) PPI, and (f) USAir. Results are averaged over 100 independent realizations. Note that the curves for EB are not shown as the respective AUC values are too low. The same holds for PA in panel (c).

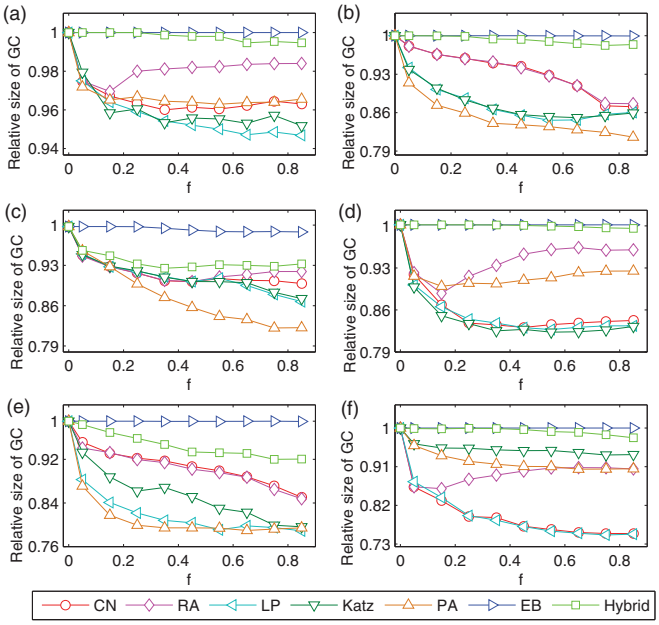

FIG. 3. (Color online) Relative size of the GC in A^r with respect to the one in A^t when various indices are used to build A^r (here, $f' = f$) and for different f. The true networks are (a) CE, (b) Email, (c) SC, (d) PB, (e) PPI, and (f) USAir. Results are averaged over 100 independent realizations. Refer to Appendix A for the nonmonotonic trend in the RA case.

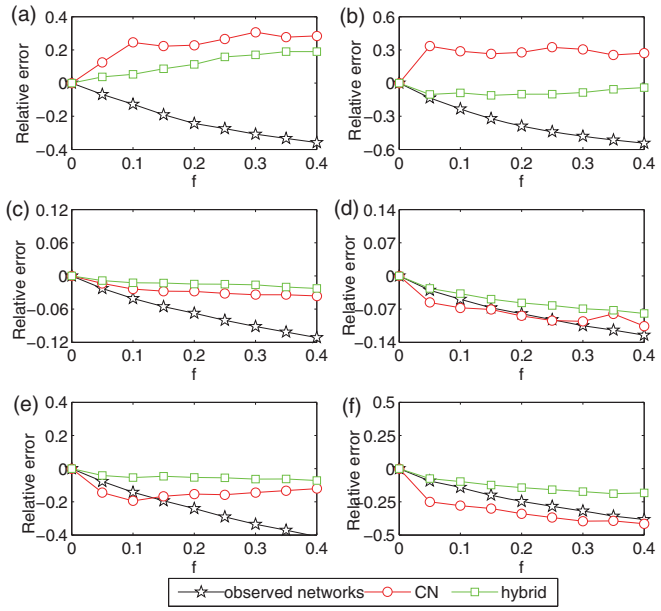

FIG. 4. (Color online) Relative errors of clustering coefficient [(a) and (b)], average shortest path length [(c) and (d)], and transportation congestability [(e) and (f)] for different f. The different lines correspond to the errors in A^o and in the two A^r built by the CN and hybrid methods, respectively, with $f' = f$. Left plots refer to PB, whereas, right plots refer to USAir. Results are averaged over 100 independent realizations.

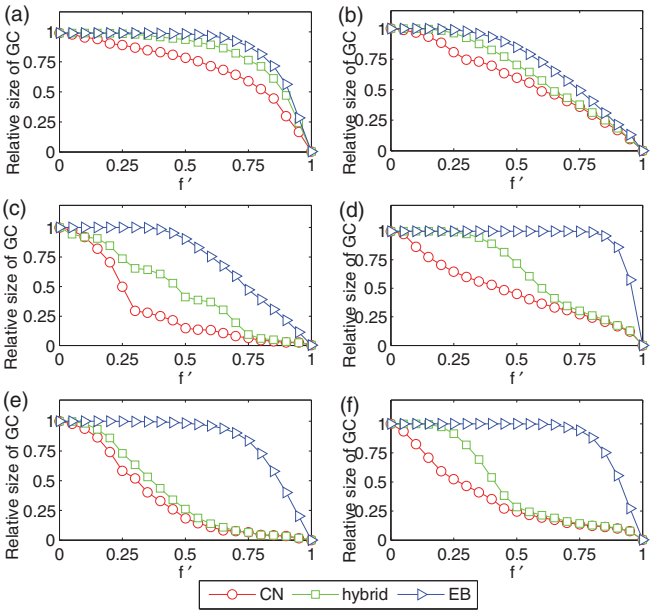

FIG. 5. (Color online) Relative size of the GC when different fractions of links f' are removed from A^o by the CN, hybrid, and EB methods. The true networks are (a) CE, (b) Email, (c) SC, (d) PB, (e) PPI, and (f) USAir. Results are averaged over 100 independent realizations.

similarity-based method, despite the small contribution it receives from EB. Hence, it is sufficient to modestly increase the reliability of central and important links to avoid removing them.

We move further by considering other network properties. In order to compare the true and the reconstructed networks under a given property X, we compute the relative error of X as $[X(A^r) - X(A^t)]/X(A^t)$. As a benchmark, we also compute the relative error of X in the observed networks as $[X(A^o) - X(A^t)]/X(A^t)$. For an effective spurious link removal method, which is able to reproduce the properties of the true network, the absolute value of the relative error for A^r should be smaller than the absolute value of the relative error for A^o (meaning that A^r is a better estimate of A^t than A^o) and as close as possible to zero (meaning that X has approximately the same value in A^t and A^r). Figure 4 shows the relative errors made by the CN and hybrid methods for the clustering coefficient, average shortest path length, and traffic congestability (i.e., the maximum betweenness centrality in the network) in the GC. We only report the results for the PB and USAir networks as these are the cases in which the GC's size is relatively more affected when using pure similarity-based methods (Fig. 3). We observe that, in these cases, the hybrid method is always able to restore the properties of the true network with respect to the observations, whereas, this is not always true for the CN. Moreover, the hybrid method always preserves the networks' properties better than the CN, at the small cost of achieving smaller AUC values. This is because the CN and other similarity-based methods alter the GC, which is much more harmful for the networks' properties and functions than keeping fewer more spurious links. Note, however, that, if the CN method does not cause serious enough damage to the GC—as happens for the CE and SC networks—then, the situation may be reversed: The CN can preserve some of the network properties better than the hybrid method due to its higher accuracy (see Figs. 8 and 9 in Appendix B).

There are plenty of other network static and dynamical properties that can be considered, such as synchronization, spreading threshold, and so on. As these dynamics can only take place in the GC, similarity-based methods, which break the network into pieces, alter them seriously. For example, the nodes out of the GC can never reach the global synchronized state, and the signal from the GC can never spread to these nodes. Again, these methods eventually destroy the system's functions.

As in real applications of spurious link removal where one does not know the exact number of spurious links in a network, we finally consider the case $f' \neq f$. To do so, we fix the number of random connections added to A^t at $f = 10\%$. We then study the properties of networks A^r reconstructed by different methods by removing different fractions f' of links from A^o.

Figure 5 shows the GC's size for varying f'. We observe that the GC's size naturally decreases with the fraction of removed links. Such a decrease is very fast when using the CN and very slow when using the EB—in the latter case, the GC's size is preserved in any network even when half the links are removed. The hybrid method lies between these two, and remarkably, it performs like the EB when the fraction of removed links is not too big (in many cases, the GC's size has a plateau that may last up to large f'). Another interesting aspect would be to investigate how many of the original f spurious links are left in the networks for various f'. Results are shown in Fig. 6. We again observe that the more we remove links, the higher the probability to remove a spurious link. Due to its low

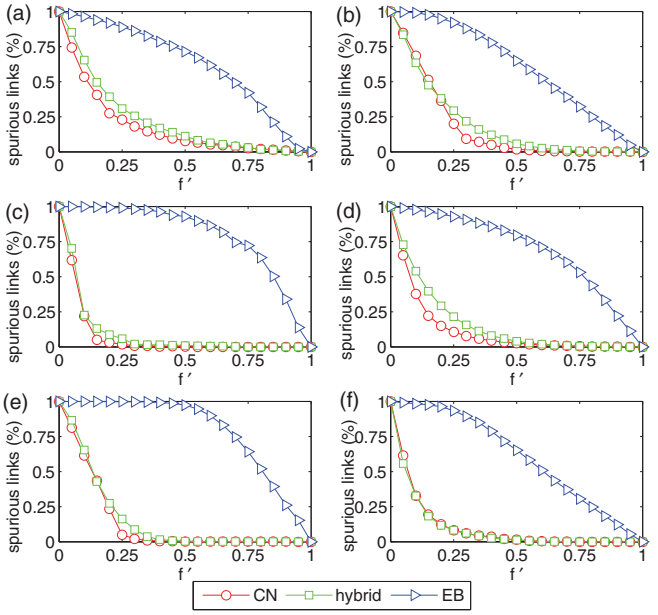

FIG. 6. (Color online) The residual fraction of spurious links in A^r when different fractions of links f' are removed from A^o by the CN, hybrid, and EB methods. The true networks are (a) CE, (b) Email, (c) SC, (d) PB, (e) PPI, and (f) USAir. Results are averaged over 100 independent realizations.

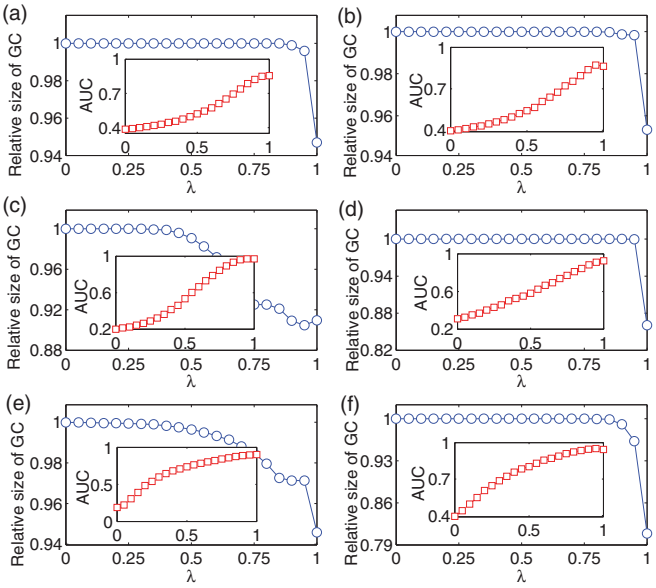

FIG. 7. (Color online) Relative size of the GC in the networks reconstructed by the hybrid method with different values of λ. (Insets) The AUC for different λ's. The respective true networks are (a) CE, (b) Email, (c) SC, (d) PB, (e) PPI, and (f) USAir. Results are averaged over 100 independent realizations.

accuracy, the EB must remove almost all links in order to get rid of the spurious ones. On the contrary, the CN can eliminate all the spurious links quite soon ($f' \simeq 25\%$). Interestingly, the hybrid performs as well as the CN, and their curves almost overlap. These results again indicate that the hybrid method represents an effective approach to both preserve the GC's size and to achieve high accuracy. Moreover, it is also more robust than other methods when considering the intrinsic uncertainty of the number of spurious interactions in a system.

VI. THE HYBRIDIZATION PARAMETER

At last, we show how the hybrid index behaves by varying the value of the parameter λ. In order to do that, we consider the particular case in which the observed networks A^o are obtained from the true networks A^t with the addition of $f = 20\%$ of spurious links. Figure 7 shows the AUC and GC size of networks A^r reconstructed by the hybrid method (with $f' = f$) for different values of λ. We observe that, whereas, the AUC decreases for decreasing λ (but this decrease is always slower at the beginning), the GC remains almost in one piece except when λ becomes too close to 1. Therefore, it is sufficient to have a small contribution from the EB in the hybrid method to keep the network connected at the cost of being slightly less accurate. This is the reason why we have previously set $\lambda = 0.9$. Note that one can use a bigger value of λ if accuracy is the main goal or a smaller value if the GC's integrity is a major issue.

VII. DISCUSSION

How to detect and to remove spurious interactions in networks is a significant problem, which may find applications in almost any field of complex science. Still, it has not yet attracted much attention as the consequences of a removal error can heavily harm the system under investigation. In the literature, many similarity-based methods, for the purpose of link prediction, have been proposed. In this paper, we showed that, when applied to spurious link detection, all these methods achieved high accuracy but suffered from the important drawback of decreasing the size of the GC and distorting other static and dynamic properties of the network. This harmful effect may cause a system to lose its functions as nodes that are disconnected from the GC cannot communicate with the network's core. In order to overcome these flaws, we proposed a hybrid method that combined the similarity-based CN's index with EB centrality. We showed that this approach can effectively eliminate the spurious links and, at the same time, keep the network connected. Moreover, important properties, such as clustering coefficient, average shortest path length, and traffic congestability can be generally preserved better. This method is still more advantageous when the number of spurious interactions within a system is unknown.

In the literature, there are other important examples of spurious link detection approaches (e.g., hierarchical random graph [5] and stochastic block model [36]), which, however, do not focus on preserving the GC's size. Moreover, these methods are based on global algorithms that can be prohibitive to use for large-scale systems. Instead, our method would be easily applicable for large networks. This is because it combines the CN index, which requires only local information of a link and EB centrality, whose computational complexity is now as low as $O(NE)$, where N and E are respectively the number of nodes and edges in the network [37].

Finally, we remark that the problem of identifying spurious interactions is much more difficult to deal with than predicting missing interactions. We already pointed out how serious a removal error may be. In addition, whereas, in link prediction

studies, there is a true network from which some existing links are removed to generate the observation and to test the algorithm, for spurious link detection, how to add spurious interactions to the true network is generally unknown. In this paper, we explored the simplest situation in which spurious links were just random connections between nodes. This approach can be suitable for describing some systems (for instance, biological networks obtained from measurements prone to random errors or social networks in which some links result from once in a lifetime interactions between people) but may result inadequately for others (such as biological systems when measurements are prone to systematic errors or the WWW where spam hyperlinks always start from the same set of pages). Hence, the effectiveness of a spurious link detection method in these systems deserves further validation, which will be the subject of future studies.

ACKNOWLEDGMENTS

We would like to thank Yi-Cheng Zhang, Matúš Medo, Chi Ho Yeung, and Stanislao Gualdi for helpful suggestions. This work was partially supported by the Swiss National Science Foundation under Grant No. 200020-132253 and by the Future and Emerging Technologies program of the European Commission FP7-COSI-ICT (project QLectives, Grant No. 231200).

APPENDIX A: NONMONOTONIC DECREASE IN GC's SIZE

To explain the nonmonotonic trend of the GC's size with f when the RA index is used (Fig. 3), we use the following argument. In the spurious link removal process by similarity-based methods, the decrease in the GC's size is due to two effects: (1) formation of communities separated from the GC and (2) formation of isolated nodes (i.e., with no connections). The first effect is caused by links between communities always connecting node pairs with few common neighbors, which are likely to be removed. The second effect is due to nodes with few connections, which cannot have many neighbors in common with any other node—take, for instance, the case of nodes with $\|\Gamma\| = 1$, for which the CN score with any other node is zero. Again, the links connecting these nodes with the network's core are likely to be removed. Generally, the higher the fraction of spurious links added and then removed to the true network, the higher the chances for these effects to manifest and, thus, the more serious the damage to the GC. However, when the RA index is used, we observe a different behavior, which can be explained as follows. When some spurious links randomly are added to the true network, pairs of connected nodes may eventually acquire new common neighbors. Because of the broad degree distribution in the network, these new neighbors are likely to have low degree. Since RA favors links between node pairs with low-degree common neighbors, these links are now unlikely to be removed also if one of the two nodes in the pair has a small degree: In this case, the formation of isolated nodes is hindered, and this effect is stronger when many links are added to the true networks (more chances for a node pair to obtain new common neighbors). This explains the nonmonotonic trend of the GC's size observed in this case. We also remark that this phenomenon does not appear for networks in which the node's average degree is small (Email,

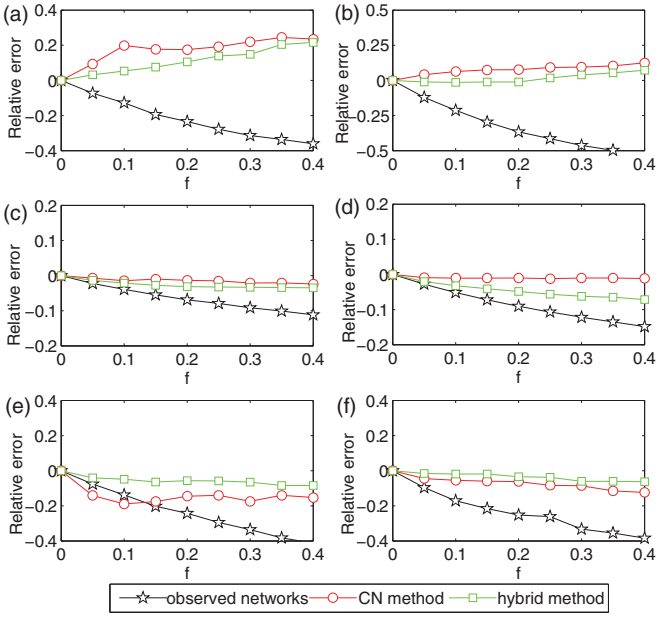

FIG. 8. (Color online) Relative errors of clustering coefficient [(a) and (b)], average shortest path length [(c) and (d)], and transportation congestability [(e) and (f)] for different f's. The different lines correspond to the relative errors in A^o and in the two A^r built by the CN and hybrid methods, respectively, with $f' = f$. The left plots refer to the CE, whereas, the right plots refer to Email. Results are averaged over 100 independent realizations.

SC, and PPI—see Table I) because, in these cases, the number of spurious links added is low as it is the chance for node pairs to acquire new common neighbors.

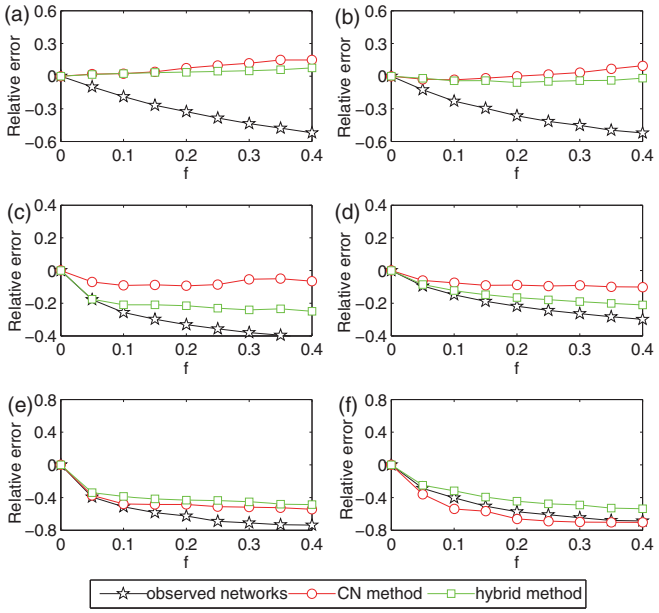

FIG. 9. (Color online) Relative errors of clustering coefficient [(a) and (b)], average shortest path length [(c) and (d)], and transportation congestability [(e) and (f)] for different f's. The different lines correspond to the relative errors in A^o and in the two A^r built by the CN and hybrid methods, respectively with $f' = f$. The left plots refer to the SC, whereas, the right plots refer to the PPI. Results are averaged over 100 independent realizations.

APPENDIX B: NETWORK'S PROPERTIES WHEN THE GC IS NOT HEAVILY HARMED

In this section, we report the relative errors made by the CN and hybrid methods for the clustering coefficient, average shortest path length, and traffic congestability for the networks that were not discussed in Sec. V—namely, the CE, Email (Fig. 8) and SC, PPI (Fig. 9). In these cases, the GC's size is not affected much when using pure similarity-based methods, hence, the CN can restore some of the true network's properties better than the hybrid method. This happens, for instance, for the average shortest path length in the GC. Contrary to the hybrid method, the CN is likely to remove links between different communities, destroying many shortest paths in this way but also decreasing the size of the GC. If such a decrease is not dramatic, the two effects can balance out so that the shortest path length may not change significantly from A^t to A^r. However, the CN is still disadvantageous for other features. With respect to clustering, the CN hardly can cut a link that belongs to a closed triangle, even if such a link is a spurious one because its ending nodes have at least one common neighbors. Hence, the number of closed link triangles (and, consequently, the clustering coefficient) will be generally greater in A^r than in the original A^t. In addition, the CN can rarely preserve the links from A^t with the highest betweenness, causing a significant decrease in traffic congestability from A^t to A^r.

[1] L. A. N. Amaral and J. M. Ottino, Eur. Phys. J. B **38**, 147 (2004).
[2] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Field, and P. Bork, Nature (London) **417**, 399 (2002).
[3] C. T. Butts, Soc. Networks **25**, 103 (2003).
[4] L. Getoor and C. P. Diehl, ACM SIGKDD Explor. Newsl. **7**, 3 (2005).
[5] A. Clauset, C. Moore, and M. E. J. Newman, Nature (London) **453**, 98 (2008).
[6] S. Redner, Nature (London) **453**, 47 (2008).
[7] J. O'Madadhain, J. Hutchins, and P. Smyth, ACM SIGKDD Explor. Newsl. **7**, 23 (2005).
[8] D. Liben-Nowell and J. Kleinberg, J. Am. Soc. Inf. Sci. Technol. **58**, 1019 (2007).
[9] J. Kunegis, E. W. De Luca, and S. Albayrak, in *IPMU Proceedings of the Computational Intelligence for Knowledge-Based Systems Design* (Springer-Verlag, Berlin, 2010), pp. 380–389.
[10] Q.-M. Zhang, M.-S. Shang, W. Zeng, Y. Chen, and L. Lü, Phys. Procedia **3**, 1887 (2010).
[11] P. Holme and M. Huss, J. R. Soc., Interface **2**, 327 (2005).
[12] Z. Huang and D. D. Zeng, in *SMC IEEE International Conference on Systems, Man and Cybernetics* (Taipei, 2006), pp. 1131–1136.
[13] B. Gallagher, H. Tong, T. Eliassi-Rad, and C. Faloutsos, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2008), p. 256.
[14] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, and A. Joshi, in *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology* (ACM, New York, 2008), p. 668.
[15] L. Lü and T. Zhou, Physica A **390**, 1150 (2011).
[16] Y. Wang and J. Chu, in *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia* (ACM, New York, 2009), p. 377.
[17] D.-H. Kim, J. D. Noh, and H. Jeong, Phys. Rev. E **70**, 046126 (2004).
[18] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, Nature (London) **464**, 1025 (2010).
[19] R. Guimerà, S. Mossa, A. Turtschi, and L. A. N. Amaral, Proc. Natl. Acad. Sci. USA **102**, 7794 (2005).
[20] A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, and C. Zhou, Phys. Rep. **469**, 93 (2008).
[21] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[22] R. Guimerà, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, Phys. Rev. E **68**, 065103 (2003).
[23] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).
[24] R. Ackland [http://incsub.org/blogtalk/images/robertackland.pdf]
[25] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, Nature (London) **417**, 399 (2002).
[26] V. Batageli and A. Mrvar [http://vlado.fmf.uni-lj.si/pub/networks/data/default.htm]
[27] J. A. Hanely and B. J. McNeil, Radiology **143**, 29 (1982).
[28] When the total number of pairs n is large, it may become prohibitive to go over all possible comparisons. Hence, for analyzing big data sets, one usually only considers a random subset of n, which is big enough to reliably estimate the true AUC value.
[29] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas, Adv. Phys. **56**, 167 (2007).
[30] R. Guimerà, A. Diaz-Guilera, F. Vega-Redondo, A. Cabrales, and A. Arenas, Phys. Rev. Lett. **89**, 248701 (2002).
[31] T. Zhou, L. Lü, and Y.-C. Zhang, Eur. Phys. J. B **71**, 623 (2009).
[32] L. Lü, C.-H. Jin, and T. Zhou, Phys. Rev. E **80**, 046122 (2009).
[33] L. Katz, Psychometrika **18**, 39 (1953).
[34] For LP and Katz, we set parameters ϵ and β to the values that maximize the respective AUC.
[35] S. Fortunato, Phys. Rep. **486**, 75 (2010).
[36] R. Guimerà and M. Sales-Pardo, Proc. Natl. Acad. Sci. USA **106**, 22073 (2009).
[37] U. Brandes, J. Math. Sociol. **25**, 163 (2001).