Toward link predictability of complex networks

Lin Yuan Lü,a,1,2 Liming Pan a,b,1 Tao Zhou b,c,1 Yi-Cheng Zhang a,d,2 and H. Eugene Stanley a,e,2

a Alibaba Research Center for Complexity Sciences, Alibaba Business College, Hangzhou Normal University, Hangzhou 311121, China; and b Complex Lab, Web Sciences Center and Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, China; c Department of Physics, University of Fribourg, Fribourg CH-1700, Switzerland; and d Department of Physics and Center for Polymer Studies, Boston University, Boston, MA 02215

Contributed by H. Eugene Stanley, December 31, 2014 (sent for review March 10, 2014; reviewed by Giorgio Parisi and Dashun Wang)

The organization of real networks usually embodies both regularities and irregularities, and, in principle, the former can be modeled. The extent to which the formation of a network can be explained coincides with our ability to predict missing links. To understand network organization, we should be able to estimate link predictability. We assume that the regularity of a network is reflected in the consistency of structural features before and after a random removal of a small set of links. Based on the perturbation of the adjacency matrix, we propose a universal structural consistency index that is free of prior knowledge of network organization. Extensive experiments on disparate real-world networks demonstrate that (i) structural consistency is a good estimation of link predictability and (ii) a derivative algorithm outperforms state-of-the-art link prediction methods in both accuracy and robustness. This analysis has further applications in evaluating link prediction algorithms and monitoring sudden changes in evolving network mechanisms. It will provide unique fundamental insights into the above-mentioned academic research fields, and will foster the development of advanced information filtering technologies of interest to information technology practitioners.

Link prediction | complex networks | structural perturbation | predictability

Understanding the organization of real networks is a long-standing challenge in many branches of science (1). Although some mechanisms have already been accepted as primary driving forces in network organization, including homophily (2–4), triadic closure (5–7), preferential attachment (8–10), reciprocity (11), and social balance (12), one or two of these mechanisms cannot provide a complete explanation; i.e., link formation in real-world networks is usually driven by both regular and irregular factors, and only the former can be explained using mechanistic models. This intrinsic network complexity presents us with the question of how to estimate what portions of a real network can be categorized as regular, in other words, to what extent the link formation in network is explicable.

This question brings to mind the link prediction problem in which the set of observed links in a network is used to estimate the likelihood that a nonobserved link exists (13). The extent to which the network formation is explicable coincides with our capacity to predict missing links (14, 15). On the one hand, an effective link prediction algorithm provides strong evidence of the corresponding mechanism(s) of network organization, e.g., effectiveness of common-neighborhood-based methods indicates the significance of triadic closure (16, 17). On the other hand, a better understanding of network organization should be transferable to a good link prediction algorithm, e.g., the prior assumption of hierarchical or organization of network that be directly applied to the design of a prediction algorithm (18). In this sense, the precision of a link prediction algorithm tells us the extent to which the link formation in network can be explained by this algorithm. However, different algorithms provide different predictions in same network (see Table 1, the precisions of seven link prediction (LP) methods on 10 networks) and thus the prediction only reflects the link predictability associated with a specific algorithm, not the intrinsic feature of the network itself.

Predictability is usually defined as the possible maximum precision of a prediction algorithm (19). However, this kind of definition is not suitable for link prediction since a real network’s link predictability under such definition should be 1 because their nonobserved links are almost always distinguishable (see Materials and Methods). In this paper, link predictability indeed characterizes the inherent difficulty of prediction that does not depend on specific algorithms, and our fundamental hypothesis is that missing links are difficult to predict if their addition causes huge structural changes, and thus network is highly predictable if the removal or addition of a set of randomly selected links does not significantly change the network’s structural features. Accordingly, we propose a so-called “structural consistency” index that is based on the first-order matrix perturbation, which can reflect the inherent link predictability of a network and does not require any prior knowledge of the network’s organization. We also propose a structural perturbation method for link prediction that is more accurate and robust than the state-of-the-art methods.

Structural Consistency

Consider a simple undirected network $G(V, E)$ where $V$ is the set of nodes and $E$ is the set of links. The given network can be represented by an $N 	imes N (|V|)$ adjacency matrix $A$, where the element $A_{ij} = 1$ if nodes $i$ and $j$ are connected and $A_{ij} = 0$ otherwise. We randomly select a fraction $p^\text{th}$ of the links to constitute a perturbation set $\Delta E$, while the rest of the links $E - \Delta E$ constitute the set $E^\text{th}$. Denote by $A^\text{th}$ and $\Delta A$ the corresponding adjacency matrices; obviously, $A = A^\text{th} + \Delta A$. Since $A^\text{th}$ is real symmetric, it can be diagonalized as $A^\text{th} = \mathbf{V}^\text{th} \mathbf{D} \mathbf{V}^\text{th}$, where the diagonalization is unique and the $\mathbf{V}^\text{th}$ is the eigenvector matrix of $A^\text{th}$.

Significance

Quantifying a network’s link predictability allows us to (i) evaluate predictive algorithms associated with the network, (ii) estimate the extent to which the organization of the network is explicable, and (iii) monitor sudden mechanistic changes during the network’s evolution. The hypothesis of this paper is that a group of links is predictable if removing them has only a small effect on the network’s structural features. We introduce a quantitative index for measuring link predictability and an algorithm that outperforms state-of-the-art link prediction methods in both accuracy and universality. This study provides fundamental insights into important scientific problems and will aid in the development of information filtering technologies.

Author contributions: L.L., L.P., T.Z., Y.-C.Z., and H.E.S. designed research; L.L., L.P., Y.-C.Z., and H.E.S. performed research; L.L., L.P., T.Z., and Y.-C.Z. analyzed data; and L.L., T.Z., Y.-C.Z., and H.E.S. wrote the paper.

Reviewers: G.P., University of Rome; and D.W., IBM TJ Watson Research Center.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

1L.L., L.P., and T.Z. contributed equally to this work.

2To whom correspondence may be addressed. Email: hes@bu.edu, linyan.lu@hznju.edu.cn, or yi-cheng.zhang@unifr.ch.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1424644112/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1424644112

PNAS | February 24, 2015 | vol. 112 | no. 8 | 2325–2330
where $\lambda_k$ and $x_k$ are the eigenvalue and the corresponding orthogonal and normalized eigenvector for $A_R$, respectively.

We consider the set $\Delta E$ as a perturbation to the network $A_R$ and construct the perturbed matrix via first-order approximation that allows the eigenvalues to change but fixes the eigenvectors. We first consider the nondegenerated case without any repeated eigenvalues (see SI Appendix, Case of Degenerate Eigenvalues, for the case with degenerate eigenvalues). After perturbation, the eigenvalue $\lambda_k$ is corrected to be $\lambda_k + \Delta \lambda_k$ and its corresponding eigenvector is corrected to be $x_k + \Delta x_k$. Left-multiplying the eigenfunction

$$ (A_R + \Delta A)(x_k + \Delta x_k) = (\lambda_k + \Delta \lambda_k)(x_k + \Delta x_k) $$

by $x_k^T$ and neglecting second-order terms $x_k^T \Delta A \Delta x_k$ and $\Delta \lambda_k x_k^T \Delta x_k$, we obtain

$$ \Delta \lambda_k \approx x_k^T \Delta A x_k. $$

This formula is reminiscent of the expectation value of the first-order perturbation Hamiltonian in quantum mechanics. Using the perturbed eigenvalues while keeping eigenvectors unchanged, the perturbed matrix can be obtained,

$$ \tilde{A} = \sum_{k=1}^{N} (\lambda_k + \Delta \lambda_k)x_k x_k^T, $$

which can be considered as the linear approximation of the given network $A$ if the expansion is based on $A_R$.

The eigenvectors can well reflect network structural features (20). If the perturbation does not significantly change the structural features, the eigenvectors of the observed matrix $A_R$ (i.e., $x_k$) and those of the matrix $A_R + \Delta A$ (i.e., $x_k + \Delta x_k$) should be almost the same. If so, according to Eq. 4, $A$ should be very close to $A_R + \Delta A$. Therefore, given a network $A$, we first randomly remove a group of randomly selected links $\Delta E$ and then we perturb the remaining part $A_R$ by $\Delta E$ to obtain the perturbed matrix $\tilde{A}$ via Eq. 4. If the network is highly regular, the random removal $\Delta E$ will not sharply change the structure features, and thus $A$ and $\tilde{A}$ should be close to each other. To measure this quantitatively, we rank all of the links in set $U - E_R$ in descending order according to their values in $A$, where $U$ is the universal set of links. We denote $E^L$ the set of top-$L$ ranked links, where $L = |\Delta E|$, namely, the number of links in the perturbation set. Then the links in $E^L$ together with the links in $E_R$ construct the perturbed network, which is usually different from $E_R + \Delta E$. The structural consistency $\sigma_c$ is defined as the fraction of common links between $\Delta E$ and $E^L$, as

$$ \sigma_c = \frac{|E^L \cap \Delta E|}{|\Delta E|}. $$

Fig. 1 shows how to calculate the structural consistency of a simple network, with a summary of detailed procedure presented in SI Appendix, Six Steps to Calculate $\sigma_c$.

### Structural Perturbation Method

The perturbation method used to determine the structural consistency can be applied to predict missing links. Link prediction aims at estimating the existence likelihood of nonobserved links based on the observed topology (13). The simplest framework of link prediction is similarity-based algorithms (16) in which each pair of nodes, $x$ and $y$, is assigned a similarity score $s_{xy}$. All nonobserved links are ranked according to their scores, with an assumption that links with higher scores have higher existence likelihoods (21) (see mathematical description of LP problem as well as the accuracy metrics in SI Appendix, Link Prediction Problem).

Under this framework, the entries of $A$ can be considered as the similarity scores assigned to links. For example, in Fig. 1, if we want to predict one missing link of given network $A$ by using the structural perturbation method (SPM), we will rank all of the nonobserved links (i.e., the links corresponding to 0 in matrix $A$) according to their scores in $A$; then the top one is the link $(3,8)$. The feasibility of SPM is based on the strong correlation between independent perturbations (see SI Appendix, Table S1), which indicates that the missing links, which are considered as unknown information, can be recovered by perturbing the network with another set of known links (i.e., $\Delta E$).

Consider an undirected network $G(V,E)$: To test the algorithm’s accuracy, the set of links, $E$, is randomly divided into two parts: (i) a training set $E^T$, which is treated as known information, and (ii) a probe set (i.e., validation subset) $E^P$, which is used for testing and can be considered as missing links. No information in the probe set is allowed to be used for prediction. Obviously, $E^T \cup E^P = E$ and $E^T \cap E^P = \emptyset$. Our task is to uncover the links in the probe set based on the information in the training set.

Notice that, in this task, the training set $E^T$ plays a similar role to the observed network $A$, and to obtain the perturbed matrix $\tilde{A}$, we randomly select a fraction $p^T$ of links from $E^T$ as perturbation set $\Delta E$. Then, by perturbing $E^T - \Delta E$ with $\Delta E$, we obtain $\tilde{A}$ through Eq. 4. The final average prediction matrix $\tilde{A}$ is obtained by averaging over 10 independent selections of $\Delta E$. By ranking all of the nonobserved links (i.e., links in $U - E^T$) in

### Table 1. Link prediction accuracy measured by precision on the 10 real networks

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Jazz</th>
<th>Metabolic</th>
<th>Neural</th>
<th>USAir</th>
<th>Food web</th>
<th>Hamster</th>
<th>NetSci</th>
<th>Yeast</th>
<th>Email</th>
<th>Router</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPM</td>
<td>0.677</td>
<td>0.354</td>
<td>0.168</td>
<td>0.451</td>
<td>0.561</td>
<td>0.469</td>
<td>0.334</td>
<td>0.166</td>
<td>0.158</td>
<td>0.357</td>
</tr>
<tr>
<td>CN</td>
<td>0.506</td>
<td>0.137</td>
<td>0.095</td>
<td>0.374</td>
<td>0.073</td>
<td>0.061</td>
<td>0.329</td>
<td>0.109</td>
<td>0.149</td>
<td>0.027</td>
</tr>
<tr>
<td>AA</td>
<td>0.525</td>
<td>0.190</td>
<td>0.105</td>
<td>0.394</td>
<td>0.075</td>
<td>0.061</td>
<td>0.334</td>
<td>0.121</td>
<td>0.150</td>
<td>0.026</td>
</tr>
<tr>
<td>RA</td>
<td>0.541</td>
<td>0.267</td>
<td>0.104</td>
<td>0.455</td>
<td>0.076</td>
<td>0.054</td>
<td>0.541</td>
<td>0.090</td>
<td>0.148</td>
<td>0.027</td>
</tr>
<tr>
<td>Katz</td>
<td>0.546</td>
<td>0.147</td>
<td>0.107</td>
<td>0.379</td>
<td>0.181</td>
<td>0.108</td>
<td>0.370</td>
<td>0.061</td>
<td>0.149</td>
<td>0.120</td>
</tr>
<tr>
<td>HSM</td>
<td>0.326</td>
<td>0.100</td>
<td>0.073</td>
<td>0.216</td>
<td>0.249</td>
<td>0.202</td>
<td>0.303</td>
<td>0.081</td>
<td>0.134</td>
<td>0.309</td>
</tr>
<tr>
<td>SBM</td>
<td>0.410</td>
<td>0.197</td>
<td>0.143</td>
<td>0.335</td>
<td>0.460</td>
<td>0.275</td>
<td>0.177</td>
<td>0.122</td>
<td>0.094</td>
<td>0.176</td>
</tr>
</tbody>
</table>

We compare our method, SPM, to six well-known methods presented in Materials and Methods. For each real network, 10% of its links will be randomly selected to constitute the probe set, and the rest of the links constitute the training set. Prediction accuracy is measured by precision. We set $p^T = 0.1$ for SPM. For the parameter-dependent Katz index, the present results correspond to the optimal parameter subject to the highest precision. The highest value for each network is in boldface.
Figure 1. An illustration of how to calculate the structural consistency. In the first plot, the blue dashed links constitute the perturbation set \(\Delta E = ([5, 8], [6, 9])\) (corresponding to \(\Delta A\)), while the solid links constitute the set \(E^0\) (corresponding to \(A^0\)). The second plot shows the adjacency matrix \(A\) of the given network, where the number in each square is the corresponding value of the matrix element. The black and blue squares represent the links in \(E^0\) and \(\Delta E\), respectively. To calculate the consistency, we perturb \(A^0\) with \(\Delta A\). The perturbed matrix \(\tilde{A}\) is shown in the third plot, from which we derive the perturbed network in the fourth plot, where the red dashed lines are outcome links selected by ranking all links in \(U - E^0\) in descending order according to their corresponding values in \(\tilde{A}\). Since there are two links in \(\Delta E\), then \(L = 2\), and the set \(E^\Delta = ([3, 8], [6, 9])\). In this case, only one of the two blue links is recovered by perturbation; then we have \(\alpha = 0.5\).

We compare the structural perturbation method with six widely applied link prediction algorithms, including four similarity-based indices: the common neighbors (CN) index (16), the Adamic-Adar (AA) index (22), the resource allocation (RA) index (17, 23), and the Katz index (24). We also use two likelihood methods: the hierarchical structure model (HSM) (18) and the stochastic block model (SBM) (25). See Materials and Methods for the six baseline algorithms. Table 1 shows the prediction accuracy of the 10 real-world networks (see Materials and Methods and SI Appendix, Table S2, for the description and basic statistics of the data), measured by precision (see SI Appendix, Table S3 for the results measured by another metric called AUC: the area under the receiver operating characteristic curve (26); see the discussion in SI Appendix, Link Prediction Problem, Eq. 5). The highest value for each network (in each column) is in decreasing order according to their scores given by \(\langle A\rangle\), we select the top-\(|E^\Delta|\) links and see how many of them are in the probe set. This ratio is called “precision,” which is used to quantify the performance of the algorithm (21). A summary of detailed procedures can be found in SI Appendix, Five Steps to Calculate Prediction Accuracy of SPM.

We first consider the structural consistency of modeled networks and show the validity of \(\sigma_i\) as an index for link predictability. In the Erdös–Rényi (ER) network (28), each pair of nodes is connected with probability \(p\). If \(p\) is finite and the network size \(N\) goes to infinity, the spectral density adjacency matrices in ER networks obey the Wigner semicircle law and the eigenvectors are distributed isotropically at random (29, 30). The first-order perturbation of the eigenvalues is thus also random, leading to low structural consistency values. Given an ER network \(G(N, p)\), we randomly select a fraction \(p^D = 0.1\) of the links (we have tested that \(\sigma_i\) is not sensitive to the specific value of \(p^D\); see SI Appendix, Fig. S3), and determine the average structural consistency \(\langle\sigma_i\rangle\) as a function of \(N\) for different \(p\). Fig. 2A shows how the structural consistency decreases with the network size in a power-law-like relationship and tends to the random chance \(p \cdot p^D / (1 + p + p^D)\) in the thermodynamical limit, supporting the intuition that fully random networks are unpredictable, which is also in accordance with the previous results of SBM not included due to the high computational complexity.

**Table 2.** The precision of link prediction on three real-world temporal networks

<table>
<thead>
<tr>
<th>Networks</th>
<th>CN</th>
<th>AA</th>
<th>RA</th>
<th>Katz</th>
<th>HSM</th>
<th>SPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arxiv</td>
<td>0.021</td>
<td>0.022</td>
<td>0.026</td>
<td>0.033</td>
<td>0.020</td>
<td>0.085</td>
</tr>
<tr>
<td>Facebook</td>
<td>0.021</td>
<td>0.024</td>
<td>0.041</td>
<td>0.022</td>
<td>0.007</td>
<td>0.051</td>
</tr>
<tr>
<td>Enron</td>
<td>0.032</td>
<td>0.033</td>
<td>0.027</td>
<td>0.033</td>
<td>0.008</td>
<td>0.033</td>
</tr>
</tbody>
</table>

Each network is of size \(N = 4,000\), that sampled from the original networks by using the random-walk method (see SI Appendix). The best-performed entries are emphasized in bold. We set \(p^D = 0.1\) for SBM and for the parameter-dependent Katz index; the present results are obtained under the optimal parameter subject to the highest precision. The results of SBM are not included due to the high computational complexity.

**Table 3.** Pearson correlation coefficients (CC) between precision and structural consistency on the 10 real networks

<table>
<thead>
<tr>
<th></th>
<th>CN</th>
<th>AA</th>
<th>RA</th>
<th>Katz</th>
<th>HSM</th>
<th>SBM</th>
<th>SPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>0.493</td>
<td>0.476</td>
<td>0.495</td>
<td>0.698</td>
<td>0.870</td>
<td>0.819</td>
<td>0.938</td>
</tr>
</tbody>
</table>
report about the very low link prediction accuracy on ER networks (31).

We next consider the Watts–Strogatz (WS) networks (32) with controllable randomness. A WS network \( G(N,k,q) \) starts from a ring of \( N \) nodes where each node connects to its \( k \) nearest neighbors. With a probability \( q \), each link is replaced by another link that joins two randomly chosen nodes. When \( q = 0 \), the network is a deterministic ring, and when \( q = 1 \), it is fully random. Fig. 2B shows how \( \sigma_c \) decreases with the rewiring probability \( q \), indicating once again that higher irregularities (i.e., randomness) will result in lower predictability.

The above experimental results on modeled networks affirm the rationale behind the proposed index \( \sigma_c \). We next turn our attention to real-world networks, whose predictabilities cannot be controlled as WS networks. We thus compare the structural consistency with the prediction accuracy from representative link prediction algorithms. Table 3 shows how the prediction precision is positively correlated with structural consistency \( \sigma_c \) for all six baseline algorithms, indicating that \( \sigma_c \) can, to some extent, reflect the link predictability of real networks. In addition, the precision values of algorithms that account for the global organization principles are approximately linearly correlated with the structural consistency (as indicated by >0.8 Pearson correlation coefficients in Table 3; see also Fig. 3), suggesting that the structural consistency is indeed a good indicator of whether the network is organized by some perceptible regulations. The results also show us that the missing links in networks with higher structural consistencies are easier to dig out using link prediction algorithms.

**Discussion**

Throughout history, human beings, from ancient prophets to modern scientists, have attempted to make predictions. The recent development of theoretical tools and the expanding availability of massive databases have allowed scientists to predict behaviors and trends, chart emergent events, and locate missing elements of a system (33–35). In this paper, we treat predictability as an inherent measurement of the regularities in the organization of a networked system, and our hypothesis is that a missing part is predictable only when it does not significantly change the structural features of the observable part. If it does, it cannot be revealed through observation. The perspective of this hypothesis is that high predictability indicates some perceptible regulative principles in the organization of the network. Putting aside any a priori hypotheses of what this regulative principle might be, we directly measure the structural consistency of a network by perturbing its adjacency matrix and observing the change of eigenvalues provided the fixed eigenvectors, similar to the well-known first-order perturbation in quantum mechanics. Although we cannot determine the maximum precision of any given link prediction algorithm, the structural consistency \( \sigma_c \) is a good indicator of the inherent predictability of both modeled networks and real-world networks. Surprisingly, by directly applying the first-order matrix perturbation method, we achieve more-accurate link predictions than some gracefully designed methods such as HSM (18) and SBM (25). In particular, the performance of SPM for disparate networks is very robust; i.e., it is either the best or very close to the best. In contrast, other algorithms can largely fail.
For example, CN, AA, RA, and Katz indices cannot adequately manage the food web, and the HSM poorly predicts missing links in a metabolic network, which is not organized in a hierarchical way. The SPM, on the other hand, does not make any a priori assumptions about any specific organizing principles of a network, and thus its predictions are consistently more robust.

Potential applications of this work are wide and can take both theoretical and practical forms. Using structural consistency values, we can determine whether a poor prediction was caused by an inappropriate algorithm or was due to the intrinsic unpredictability of the network, and then estimate how large a space is needed to improve the algorithm. For example, the CN index does not perform well for either neural or food webs. Because the structural consistency of a food web is much larger than that of a neural web, we can infer that CN is not appropriate for a food web, while the low precision of a neural web may result from its own low predictability. Indeed, as shown in Fig. 3, the networks largely below the fitting line are those where the corresponding algorithm is not suitable to be applied. For an evolving network, the structural consistency can give a temporal estimation of whether the network becomes more elusive or not, as well as monitor the sudden changes in the evolving mechanisms (see SI Appendix, Monitor the Sudden Changes of Evolving Networks with \( \alpha \), for numerical experiments).

In addition, the structural perturbation method, as a straightforward extension of structural consistency, can be directly applied to determining the missing links in real-world networks. This work should be of interest to academic researchers in a variety of fields, to information technology practitioners, and to business practitioners.

Materials and Methods

Maximum Precision of Link Prediction. If we define link predictability as the maximum precision of any link prediction algorithm, then a network is of nearly zero predictability if all nonobserved links are completely the same (e.g., a star network). For example, a vertex-transitive (20) network is of zero predictability since all of the nodes in the observed structure are identical and thus missing links are also indistinguishable from nonexistent links. For a vertex-transitive network, given any of its two nodes \( u \) and \( v \), there is some automorphism \( f \) such that \( f(u) = v \). This extremely rigid definition from automorphism-based symmetry makes virtually all real-world networks have a predictability very close to 1, since the missing links can be distinguished from nonexistent links. Because it is approximately free of graph automorphisms, link predictability approaches 1 even in ER networks (36). Thus, this rigid approach does not allow us to obtain any a useful estimation of link predictability.

Data Description. We consider the following 10 real-world networks drawn from disparate fields: (i) Jazz (37), a collaboration network of jazz musicians consists of 198 nodes and 2742 interactions; (ii) Metabolic (38), the metabolic network of Caenorhabditis elegans; (iii) Neural (32), the neural network of C. elegans (the original network is directed and weighted; here we treat it as a simple network by ignoring the directions and weights); (iv) US Air (39), the US Air transportation network; (v) Food web (40), the food web in Florida Bay during wet season; (vi) Hamster (41), a friendship network of users on the website hamsterster.com; (vii) NetSci (42), a coauthorship network of scientists working on network theory and experiment; (viii) Yeast (43), a protein–protein interaction network in budding yeast; (ix) Email (44), a network of email interchanges between members of the University Rovira I Virgili; (x) Router (45), a symmetrized snapshot of the structure of the Internet at the level of autonomous systems; (xi) Arxiv (46), a scientific collaboration network from the arXiv's High Energy Physics C Theory (hep-th) section; (xii) Facebook (47), a network of a small group of Facebook users; and (xiii) Enron (48), an email communication network from employees of Enron between 1999 and 2003. The more detailed information and statistical features of these networks can be found in SI Appendix, Statistical Features of Experimental Networks.

Baseline Algorithms for Link Prediction. The link prediction problem has been a long-standing challenge in modern information science (13, 49). Its main goal is to estimate the existence likelihood of nonobserved links based on the known topology and node attributes. Link prediction has already found wide applications in interdisciplinary fields, including uncovering missing parts of social and biological networks (50–52) and recommending friends and products in online social networks and e-commerce web sites (53–55).

For comparison, we introduce four benchmark similarity indices (13). The simplest is the CN index (16) in which two nodes, \( x \) and \( y \), have a higher connecting probability if they have more common neighbors. Two improved indices based on CN are the AA index (22) and the RA index (17, 23), both of which assign less-connected neighbors more weight. The mathematical expressions are

\[
S_{xy}^{\text{CN}} = |\Gamma(x) \cap \Gamma(y)|,
\]

\[
S_{xy}^{\text{AA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|},
\]

\[
S_{xy}^{\text{RA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|},
\]

where \( \Gamma(x) \) denotes the set of neighbors of node \( x \). Unlike the above three local similarity indices, the Katz index (24) uses global topological information by summing over the collection of paths with exponential damping according to path lengths, i.e.,

\[
S_{xy}^{\text{Katz}} = \sum_{l=0}^{\infty} \alpha^l A^l_{xy} + \sum_{l=0}^{\infty} \alpha^l A^l_{xy} + \cdots = (I - \alpha A)^{-1} - I,
\]

which can be rewritten in the compact form, when \(|\alpha| < 1/\lambda_{\text{max}}\), as

\[
S = (I - \alpha A)^{-1} - I;
\]

where \( I \) is the identity matrix, \( A \) is the adjacency matrix, and \( \lambda_{\text{max}} \) is the largest eigenvalue of \( A \). In our experiments, we tune the parameter \( \alpha \) to optimize the performance of the Katz index. Notice that, since \( \alpha \) cannot be exactly zero, the Katz index cannot degenerate to the CN index. Even when \( \alpha \) is very close to zero, the performance of the Katz index can be different from the CN index, because, under the CN index, many nonobserved links are scored the same and thus ranked in a random way (see analysis on this so-called degeneracy phenomenon in refs. 17 and 31); therefore the very slight differences contributed by the latter items in Eq. 9 may result in considerable changes in the order of nonobserved links associated with the same number of common neighbors.

We also consider two probability methods. The HSM (18) method assumes that many real-world networks are hierarchically organized and thus nodes can be divided into groups and further divided into subgroups. The SBM (25) approach is one of the most general network models. Nodes are partitioned into groups and the connecting probability of any two nodes is only determined by the groups they belong to.

ACKNOWLEDGMENTS. We thank G. D’Agostino for helpful discussion. This work was partially supported by the National Natural Science Foundation of China (Grants 11222543, 11075031, 11205042, and 61430014), NESP Project, and CCF-Tencent Open Research Fund. L.L. acknowledges the research start-up fund of Hangzhou Normal University under Grant PE13002004039 and the EU FP7 Grant 611272 (project GROWTHCOM). T.Z. acknowledges the Program for New Century Excellent Talents in University under Grant NCET-11-0070. The work at Boston University was supported by US National Science Foundation Grants 1125290 and 0835453.

7. Yin D, Hong L, Xiong X, Davison BD (2011) Link formation analysis in microbial. Proceedings of the 34th International ACM SIGIR Conference on Re-
tering recommender systems. ACM Trans Inf Syst 22(1):5–53.
25. Guimerà R, Sales-Pardo M (2009) Missing and spurious interactions and the re-
40. Unalawaniz RE, Bondavalli C, Egnitovich MS (1998) Network Analysis of Traffic Dy-
namics in South Florida Ecosystem, FY 97: The Florida Bay Ecosystem (Chesapeake Biol Lab, Solomons, MD), Tech Rep CRL 98-123.
47. Vishwanath B, Mislove A, Cha M, Gummadi KP (2009) On the evolution of user in-
51. Cansistraci CV, Alainis-Lobato G, Ravasi T (2013) From link-prediction in brain con-