Precision as a measure of predictability of missing links in real networks

Guillermo García-Pérez,[1, 2, *] Roya Aliakbarisani,[3, *] Abdorasoul Ghasemi,[3] and M. Ángeles Serrano[4, 5, 6, †]

[1] *QTF Centre of Excellence, Turku Centre for Quantum Physics,*
Department of Physics and Astronomy, University of Turku, FI-20014 Turun Yliopisto, Finland
[2] *Complex Systems Research Group, Department of Mathematics and Statistics,*
University of Turku, FI-20014 Turun Yliopisto, Finland
[3] *Faculty of Computer Engineering, K. N. Toosi University of Technology, Tehran 1631714191, Iran*
[4] *Departament de Física de la Matèria Condensada,*
Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain
[5] *Universitat de Barcelona Institute of Complex Systems (UBICS), Universitat de Barcelona, Barcelona, Spain*
[6] *ICREA, Pg. Lluís Companys 23, E-08010 Barcelona, Spain*

Predicting missing links in real networks is an important open problem in network science to which considerable efforts have been devoted, giving as a result a vast plethora of link prediction methods in the literature. In this work, we take a different point of view on the problem and focus on predictability instead of prediction. By considering ensembles defined by well-known network models, we prove analytically that even the best possible link prediction method, given by the ensemble connection probabilities, yields a limited precision that depends quantitatively on the topological properties—such as degree heterogeneity, clustering, and community structure—of the ensemble. This suggests an absolute limitation to the predictability of missing links in real networks, due to the irreducible uncertainty arising from the random nature of link formation processes. We show that a predictability limit can be estimated in real networks, and we propose a method to approximate such bound from real-world networks with missing links. The predictability limit gives a benchmark to gauge the quality of link prediction methods in real networks.

Limits of predictability, the degree to which a system's state can be correctly forecasted, have been explored in different contexts, including weather and climate [1], human mobility [2], and biological evolution [3]. One of the causes that undermines perfect predictability in these systems—apart from lack of information, observational errors, or variability in their environmental conditions—can be found in the inherent randomness of some of the processes and phenomena that shape their organization and behavior.

In complex networks [4], randomness not only dominates the dynamical interactions between the states of nodes in many dynamical processes [5], which limits the ability to predict specific configurations of dynamical states at any given time [6], but also link formation. The structure of complex networks is far from deterministic and can be modeled in a stochastic framework where the likelihood of links to exist is characterized probabilistically. The set of link probabilities defines a network ensemble, that can be studied to gain insight into some specific network that can be considered to be an instance of such ensemble instead of an independent entity.

This uncertainty in the likelihood of connections represents an intrinsic feature of networks that affects the predictability of their structure [7]. Link prediction methods [8, 9] are able to give information about missing or future interactions in networks by exploiting non-trivial regularities in their organization—heterogeneous degree distributions, high levels of clustering, degree-degree correlations, communities—, at the local or at the global level. Different link prediction methods typically give different results on the same network and, although some methods may perform comparatively better than others, it is not clear which is the best precision that can be achieved.

The hypothesis is that, regardless of how much link prediction methods improve, they will always present an irreducible lack of accuracy in real networks as a consequence of the random nature of link-formation processes. In this work, we address the question of what is the maximal expected precision of the best possible link prediction method for a given network ensemble, that simply corresponds to ranking the likelihoods of missing links according to the corresponding connection probabilities in the ensemble. Then, we turn to real networks to show that inferred connection probabilities in well-fitted network models of fully observed real networks allow to estimate an limit to the predictability of missing links, and we propose a method to approximate this bound from real networks with missing links.

I. PREDICTABILITY OF MISSING LINKS IN NETWORK ENSEMBLES

An ensemble $\mathcal{E}_N$ is defined as a set of networks G of N nodes generated by assigning undirected links between pairs of nodes i and j with independent pairwise probabilities $\{p_{ij}\}$, where the indices run from 1 to N. Therefore, every network G in the ensemble is weighted by a

[*] These two authors contributed equally

[†] Correspondence and requests for materials should be addressed to M. A. S., marian.serrano@ub.edu

probability $P(G)$, given by

$$P(G) = \prod_{i<j} p_{ij}^{a_{ij}} (1 - p_{ij})^{1-a_{ij}}, \qquad (1)$$

where the adjacency matrix entries $\{a_{ij}\}$ take the value $a_{ij} = 1$ if i and j are connected or $a_{ij} = 0$ otherwise. Therefore, $\sum_G P(G) = 1$.

Given a graph G, we construct an *observed* graph G_{obs} by removing a fraction q of links. As it is customary in the link prediction literature, we are making the assumption that links are removed uniformly at random. Nevertheless, it would be in principle possible to extend our results to more general noise channels.

A link prediction method $\mathcal{M}$ can be regarded as a map $G_{\text{obs}} \mapsto G_{\text{inf}}$, or $G_{\text{inf}} = \mathcal{M}(G_{\text{obs}})$, that produces a new graph G_{inf} by adding predicted links to G_{obs} such that both G and G_{inf} have the same number of links (we assume that the number of missing links is known). Let $Q \equiv Q(G, G_{\text{obs}}, G_{\text{inf}})$ be the precision of the prediction, defined as the fraction of predicted links that belong to G. Thus, if $G_{\text{inf}} = G$, $Q = 1$. The expected precision can be written as $\langle Q \rangle = \sum_{G_{\text{obs}}} P(G_{\text{obs}}) \bar{Q}(G_{\text{obs}}, G_{\text{inf}})$, where $\bar{Q}(G_{\text{obs}}, G_{\text{inf}})$ stands for the expected optimal precision over all possible original graphs yielding G_{obs} upon random removal of links.

For a given ensemble $\mathcal{E}_N$, the optimal strategy for link prediction, that is, the one maximizing the expected precision $\langle Q \rangle$ in link prediction experiments over ensemble instances, is the one that generates G_{inf} from G_{obs} by adding the links according to the connection probabilities $\{p_{ij}\}$ ranked in decreasing order. We give a proof in Appendix A, from where it is easy to see that

$$\bar{Q}(G_{\text{obs}}, G_{\text{inf}}^{\text{opt}}) = \frac{1}{L} \sum_{l=1}^{L} \frac{q p_l}{1 - p_l + q p_l}, \qquad (2)$$

where L is the number of missing links and index l runs over the set of potential links of G_{obs}, with the corresponding ensemble probabilities ordered in decreasing order, $p_l \geq p_{l+1}$, $\forall l$.

From the expression above, we observe that the expected optimal precision decreases as the number of missing links decreases and it converges to the mean of the top-L_0 connection probabilities in the ensemble —where L_0 stands for the number of links in the original graph G—, when q is maximal (notice that, in the special case in which the ensemble contains a single network, all the p_l are either 0 or 1, so there would be no dependence on q whatsoever). In fact, the *precision curve is an increasing function of the number of removed links*. This apparently counter-intuitive result stems from the fact that, as the fraction of missing links q increases, the ratio of missing links over potential links in G_{obs}, given by $\frac{qL_0}{N(N-1)/2-(1-q)L_0}$, increases and so the probability of missing an actually missing link decreases. Therefore, the statistical power of the method, *i.e.* the probability that the prediction of a missing link is correct, increases with q. Notice that instead of precision one could consider AUC or specificity, which might exhibit a different behaviour as a function of q, as a measure of predictability. However, precision not only has a simple interpretation, but it better captures the performance of link prediction methods in situations where the number of links predicted is relatively small as compared to the total number of disconnected pairs in the network. This might be particularly relevant in real applications in e.g. biology, where assessing whether a link actually exists has an associated cost and, therefore, one is constrained to verifying only the top-ranked predicted links.

A. Assessing the dependence of link predictability on topological features

We give the name of *OS predictability curve* to the curve of precision values of the optimal strategy (OS) as a function of the fraction of missing links. Given an ensemble, the theoretical OS predictability curve can be estimated by computing the ratio between the expected number of correct predictions of the OS and the expected number of non-observed links of incomplete ensemble graphs, see Appendix B. This estimation allows us to quantify the effects of different topological properties — degree heterogeneity, clustering, number of communities, ...— on the ensemble predictability as a function of its parameters.

In Erdős-Rényi (ER) networks [10, 11], where all pairs of nodes have the same connection probability p, the OS predictability curve can be computed exactly (see Appendix B) and reaches very low accuracies in accordance with previous reports [7, 12], see Fig. S1 in Supplementary Information. The unpredictability of the ER model is easily understood from the uniformity of link probabilities, which leads to a lack of connectivity patterns to be exploited by link prediction algorithms.

More structured ensembles show accuracies that are parameter dependent. We considered the soft Configuration Model (sCM) [13], producing maximally random graphs with a given expected degree sequence, the $\mathbb{S}^1$ model [14], producing maximally random geometric graphs with given expected degree sequence and level of clustering, and the degree-corrected Stochastic Block Model (dc-SBM) [15], a generalized block model that accounts for heterogeneities in the degrees to generate networks with given mesoscopic structure, see details in Appendix C 3.

The results of the estimation of the theoretical OS predictability curves for these ensembles are displayed in Fig. 1. In Fig. 1**a**, the behaviour of the predictability curve in the sCM is shown for different values of the power-law degree distribution exponent γ in the typical range observed in real-world networks, $\gamma \in [2, 3]$. As the degree distribution becomes increasingly heterogeneous (for smaller values of γ), the resulting networks exhibit increasingly large predictability. In Fig. 1**b**, we study

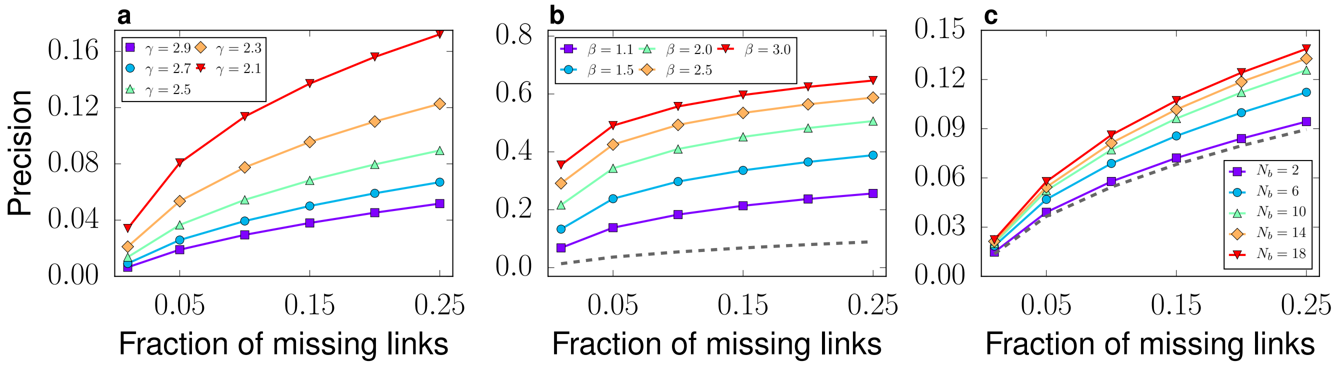

FIG. 1. **Dependence of the OS predictability curve on topological structure.** For every model, we have generated five ensembles with different model parameters and computed the corresponding theoretical OS predictability curve. **a:** Soft Configuration Model as a function of the exponent of the degree-distribution γ. To avoid unnecessary fluctuations, the degree distribution includes a natural cutoff. **b:** $\mathbb{S}^1$ model as a function of β, which regulates the mean local clustering coefficient. In this case, as well as in **c**, the degree distribution exponent has been set to $\gamma = 2.5$. **c:** Degree-corrected Stochastic Block Model as a function of the number of blocks N_b, with $\lambda = 0.5$. The black dashed curve in **b** and **c** shows the OS curve for the sCM for $\gamma = 2.5$, i.e., the green curve in **a**.

the effect of the clustering coefficient in the $\mathbb{S}^1$ model, in which the mean local clustering is a monotonically increasing function of β (see Appendix C 2). In real networks, this parameter is usually found to be in the range $\beta \in (1, 3)$. In this case, predictability grows with clustering. Notice further that this result is consistent with the fact that the overall predictability of $\mathbb{S}^1$ networks is much higher than that of the sCM, the latter exhibiting null clustering in the thermodynamic limit. Figure 1**c** shows the effect of the community structure via the number of blocks (communities), N_b, in the dc-SBM. Again, we observe an unambiguous effect on the resulting ensemble predictability, this time increasing with increasing N_b. We therefore observe a clear pattern in these results. As expected, the predictability of the network ensembles tends to decrease as they become increasingly similar to purely random graphs. This includes decreasing degree heterogeneity, clustering coefficient, or community structure.

B. The OS predictability curve as a bechmark for link prediction

We compared the OS predictability curve against the precisions of several link prediction methods on the different network ensembles analysed in this work using numerical simulations. We considered six widely applied link prediction methods. Four of them—Common Neighbors (CN) [16], Adamic-Adar (AA) [17], Resource Allocation (RA) [18], and Cannistraci-Hebb (CH) [19]—exploit local connectivity patterns, while the other two—Structural Perturbation Method (SPM) [7] and Fast probability Block Model (FBM) [20]—are global (see Supplementary Information (SI) for details). For each method, the prediction was done by computing the ranking once from G_{obs}.

The results for the different ensembles are shown in Fig. 2**a**, Fig. 2**b**, Fig. 2**c**, respectively. See also Fig. S1 in Supplementary Information, displaying the OS predictability curve in ER networks for the different link prediction methods. As expected, the optimal strategy —the best possible method— gives the best results in all cases. In Erdős-Rényi (ER) networks [10, 11], link prediction is insensitive to the method used and all of them reach the precision curve of the optimal strategy (see Appendix B for a theoretical justification in the context of our theoretical framework), reaching very low accuracies in accordance with previous reports [7, 12]. The other ensemble models show link prediction accuracies significantly above those for the ER ensemble, being the $\mathbb{S}^1$ ensemble the one with the highest predictability and, at the same time, the one in which link prediction methods perform worse.

To further illustrate this point, we infer the ensemble connection probabilities in the sCM from observed, incomplete networks and then use the resulting ranking as a link prediction method, which we name the Configuration Model Assumption (CMA). These connection probabilities are easy to estimate. In the sCM, each node i is assigned an expected degree which coincides approximately with the resulting degree k_i obtained in realizations of the model. Hence, after randomly removing a fraction q of links from an ensemble network, we expect the observed degree of every node to become $k_i^{\mathrm{obs}} \approx (1 - q)k_i$. Thus, given the observed degrees k_i^{obs} in the incomplete graph, we can estimate the original-network degrees and approximate the connection probabilities accordingly using their definition, see Appendix C 1. We then use the inferred probabilities as scores in link prediction experiments on synthetic networks belonging to all the considered ensembles. As the results in Fig. 2**a-c** show, the CMA

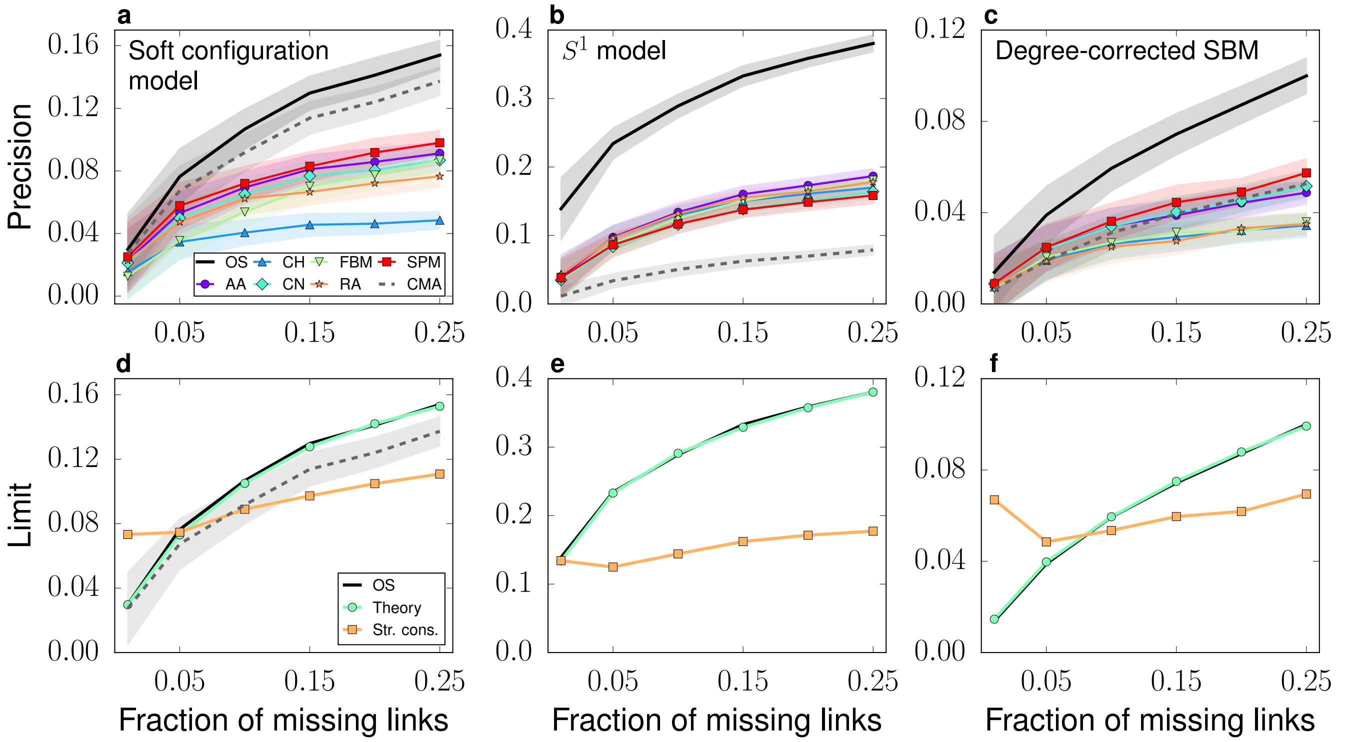

FIG. 2. **Precision as a function of the fraction of missing links for different link prediction methods on different network ensembles.** **a-c**: For each ensemble and value of the fraction of missing links q, we generated 10 networks G and, for each one of them, we generated 100 incomplete networks G_{obs} on which the link prediction methods were applied. The shaded areas represent the standard deviation of the results. The OS curves correspond to the theoretical limit computed through numerical simulations which implement the optimal strategy. In all cases, we have set $N = 1000$ nodes. Also, we have set $\langle k \rangle = 10$ and power-law degree distributions with exponent $\gamma = 2.5$. **a:** Soft Configuration Model. **b**: $\mathbb{S}^1$ model with $\beta = 1.5$. **c**: Degree-corrected Stochastic Block Model with $\lambda = 0.5$ and 7 equiprobable blocks. **d-f**: Comparison between the simulated OS predictability curve and the Structural Consistency index [7] (details in SI).

method works extremely well for networks belonging to the sCM ensemble, achieving a precision curve higher than any other link prediction method, nearly matching the theoretical maximum given by the OS predictability curve. However, the same link prediction method fails when used on completely different networks, like the $\mathbb{S}^1$-model ensemble networks. Interestingly, the results for CMA are comparable to other link prediction methods on dc-SBM networks.

We also compared the OS predictability curve with the structural consistency index [7], see Fig. 2**d-f**. Notice that the structural consistency index is not a link prediction method but it was proposed to estimate the link predictability of a network based on the assumption that removing a small subset of links at random from the given network does not change its structural features (further details in SI). However, the results for low values of q in the sCM and the dc-SBM ensembles in Fig. 2 show that the Structural Consistency index sometimes gives bounds that are impossible to achieve, see also the results for the ER ensemble in SI. Conversely, the Structural Consistency index can underestimate the limit to link predictability, actually clearly surpassed by some of the link prediction methods used here, like CMA in the sCM ensemble.

II. ESTIMATING THE OS PREDICTABILITY CURVE IN REAL NETWORKS

The inference of the ensemble connection probabilities from an observed real graph considered as incomplete is a very difficult problem for models other than the sCM. Nevertheless, we can still evaluate the predictability curve. As a preliminary step, we infer first the ensemble probabilities from the original network, before any links have been removed, and use them to apply the optimal strategy on link prediction experiments. The resulting precisions hence indicate the limits of an ideal model-based link prediction strategy, that is, in which the ensemble probabilities could be accurately inferred from the incomplete network. Second, we propose a method to estimate the OS predictability curve in real networks with missing links.

A. Inferring the OS predictability curve in fully observed real networks

We apply the aforementioned approach on eight different real networks from different domains: transitions between codewords in western modern music [21] (Music); connections between neurons within the Drosophila optic medulla [22] (Drosophila); international trade relationships in 2013 [23] (WTW); the Internet at the Autonomous Systems level [24] (Internet); a food web in the Florida Bay ecosystem [25] (Florida F. W.); adjacency between words in the novel David Copperfield by Charles Dickens [26] (Word Adjacency); the social network between members of a karate club [27] (Karate); and copurchases of books in Amazon about US politics [28] (Polbooks) (more details in Appendix D).

We inferred the connection probabilities of each of the networks—before links were randomly removed—in a suitable ensemble model. It should be stressed that, a priori, there is no obvious way to determine which network model is most adequate for a given real network. However, in this paper we focus on the dc-SBM and the $\mathbb{S}^1$, since both have proved to be able to describe real-network topologies. The network-specific choice between this two models, on the other hand, is based on the computed likelihood for the network to be generated by the model (see Appendix C 4). Once the ensemble probabilities were determined, we compared the inferred OS predictability curve, obtained by applying the optimal strategy using the inferred ensemble probabilities, with the results given by link prediction methods as a function of the number of missing links. The comparison is shown in Fig. 3. Notice that the OS predictability curve can be used to benchmark currently existing link prediction methods, as their precisions, averaged over random link removals, can now be compared against a theoretical upper-bound. The Music, Drosophila, WTW, and Internet networks are well described by the $\mathbb{S}^1$ model, due to their heterogeneous degree distributions and high levels of clustering. These networks were embedded in the underlying geometry assumed in the model by finding the parameters that maximise the likelihood for the real graphs to be generated by the model, following the same approach as in Refs. [29, 30]. Once the angular positions of the nodes in the underlying one-dimensional sphere, or circle, and the hidden degrees are found, the $\mathbb{S}^1$-model connection probabilities (see Eq. (C3) in Appendix C 2) between all pairs of nodes define an ensemble of networks which are similar to the real one. We use these probabilities to compute the OS predictability curves shown in Fig. 3**a**-**d**, which lay well above the precisions obtained by other link prediction methods.

A similar result is observed on the four datasets well described by the dc-SBM, as depicted in Fig. 3**e**-**h**, where we show the results for the Florida Food Web, Word Adjacency, Karate, and Polbooks networks. To compute the connection probabilities of a given network, it is fitted to a dc-SBM to find its community structure using a statistical inference and a Monte Carlo sampling [31]. This procedure computes the number of groups, K, and the group assignment, g, for the network to assign a connection probability to every pair of nodes, see Appendix C 3 for details. Results in Fig. 3**e**-**h** show that, even if fluctuations are more important than in the previous scenario, the precisions obtained by the different link prediction methods are still lower than the OS predictability curve.

B. Inferring the OS predictability curve in real networks with missing links

The inference of the OS predictability curve in real networks with missing links presents an evident difficulty as it requires knowledge of the original network, which is obviously inaccessible —as it is to be predicted— to compute the ensemble probabilities p_{ij}. To overcome this issue, we propose a method to estimate the OS predictability curve directly from the observed network structure by computing a set of connection probabilities that approximate those in the original ensemble. Notice that, even if this estimation could not be good enough to use the optimal strategy with the inferred probabilities as a link prediction method, the estimation of the OS predictability curve is still accurate as we show in Fig. 4. The results of the inferred OS predictability curve in real networks with missing links are compared with the inference taking into account the complete original counterparts as reported in Fig. 3. In all networks, the quality of the inferences is very good, both for the $\mathbb{S}^1$ network model ensemble and for the dc-SBM ensemble.

We explain our algorithm in what follows. Suppose that network G_{obs} has been generated by removing a fraction q_0 of links from an original network G. The following procedure allows us to estimate the OS predictability curve of the original graph G. First, we select the most suitable probabilistic network model for G_{obs} (using the likelihood criterion detailed in Appendix C 4, as in the previous subsection) to obtain the set of connection probabilities p_{ij}^{obs}, and we rank all pairs of nodes in G_{obs} in decreasing order according to their connection probabilities p_{ij}^{obs}, that we relabel as $p_{ij}^{\text{obs}} \leftrightarrow p_l$, such that $p_l > p_{l+1}, \forall l$. The next step is to estimate $\tilde{L}$, the expected number of links removed from G for a given value of q. Let E_{obs} be the number of links in G_{obs}. Since there is a fraction q_0 of missing links from G, the expected number of links in the complete network is $E = E_{\text{obs}}/(1-q_0)$. Hence, the number of missing links when a fraction q of links is removed from G and a new graph $\tilde{G}$ is produced is $\tilde{L} = qE_{\text{obs}}/(1-q_0)$.

Now, imagine we are given one such incomplete graph $\tilde{G}$ and the set of probabilities p_l. The OS prediction would then consist of the $\tilde{L}$ non-observed links of $\tilde{G}$ with highest connection probability, and the corresponding precision would simply be the fraction of those which actually exist in G. The idea behind the rest of the algorithm is therefore to estimate, by a cumulative sequen-

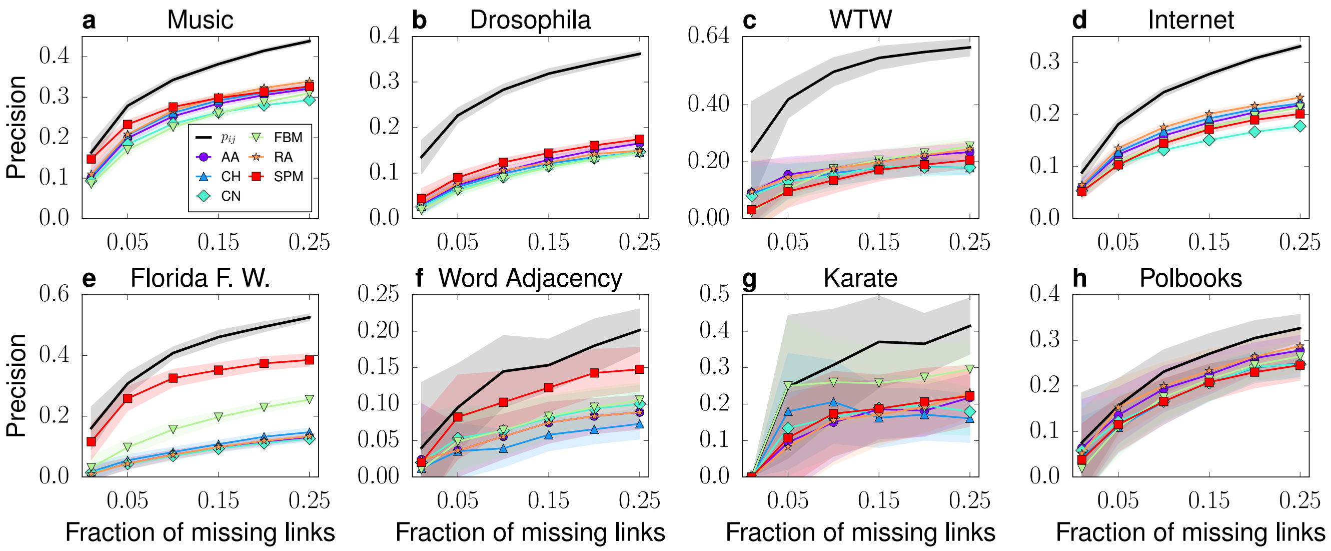

FIG. 3. **Precision as a function of the fraction of missing links for different link prediction methods on eight real-world networks.** For each network and value of the fraction of missing links q, we generated 100 incomplete networks G_{obs} on which we applied the link prediction methods. The p_{ij} curves correspond to the precisions given by the simulations of the optimal strategy using the ranking of the inferred probabilities using the original networks. **a-d**: Using the $\mathbb{S}^1$ model. **e-h**: Using the dc-SBM model. In all plots, the shaded areas represent the standard deviation of the results, typically larger for smaller networks.

tial computation using the ordered list of probabilities, both the expected number of non-links of $\tilde{G}$, H, and the expected number of non-links of $\tilde{G}$ that would exist in G, T. When $H \approx \tilde{L}$ (that is, after predicting the top-ranked $\tilde{L}$ links) the expected precision can be estimated as $\langle Q \rangle = T/H$. Hence, after initializing H and T to zero, we visit every pair of nodes $l = 1, 2, \ldots$ and compute their corresponding increments. Two different situations need to be considered differently:

1. The two nodes in pair l are connected in G_{obs}. In this case, the link must surely exist in G. Therefore, in the ensemble of $\tilde{G}$ networks, the link does not exist (and counts as a correct prediction) with probability q, so every time one such link is visited, we must update $T_{\text{new}} = T_{\text{old}} + q$ and $H_{\text{new}} = H_{\text{old}} + q$.

2. The two nodes in pair l are not connected in G_{obs}. There are two possible reasons for the link not to be observed:

   a. The link belongs to G, but has been removed from it with probability q when producing $\tilde{G}$. The probability that the link is in the original network is $q_0 p_l/(1 - p_l + q_0 p_l)$, so that the probability for it not to belong to $\tilde{G}$ is

$$\mathrm{P}\left(\tilde{a}_l = 0, a_l = 1 | a_l^{\text{obs}} = 0\right) = q \frac{q_0 p_l}{1 - p_l + q_0 p_l}. \qquad (3)$$

   b. The link does not belong to G, and therefore it cannot exist in $\tilde{G}$. Since the probability that the link does not exist in G is $(1 - p_l)/(1 - p_l + q_0 p_l)$ the corresponding probability simply reads

$$\mathrm{P}\left(\tilde{a}_l = 0, a_l = 0 | a_l^{\text{obs}} = 0\right) = \frac{1 - p_l}{1 - p_l + q_0 p_l}. \qquad (4)$$

With these two results, we can readily update T and H. Since T accounts for the expected number of correct predictions, only case *a.* contributes, that is,

$$T_{\text{new}} = T_{\text{old}} + q \frac{q_0 p_l}{1 - p_l + q_0 p_l}. \qquad (5)$$

As for H, both cases contribute, and so

$$H_{\text{new}} = H_{\text{old}} + \frac{1 + (q q_0 - 1) p_l}{1 - p_l + q_0 p_l}. \qquad (6)$$

The reason why the approximation of the OS predictability curve works even if the probabilities may not be accurate enough for a link prediction method can be understood by noting the following observation: in the algorithm, we only use the highest numerical values of the connection probabilities, without any mention whatsoever to the pair of nodes they refer to. Hence, as long as the distribution of the values of the highest probabilities is not drastically perturbed by the link removal—that is, if the highest values of the probabilities inferred on the original and incomplete networks exhibit similar distributions—the OS predictability curve can be estimated in networks with missing links, even if the specific probabilities corresponding to the removed edges change considerably and, as a result, they do not enable a good link prediction.

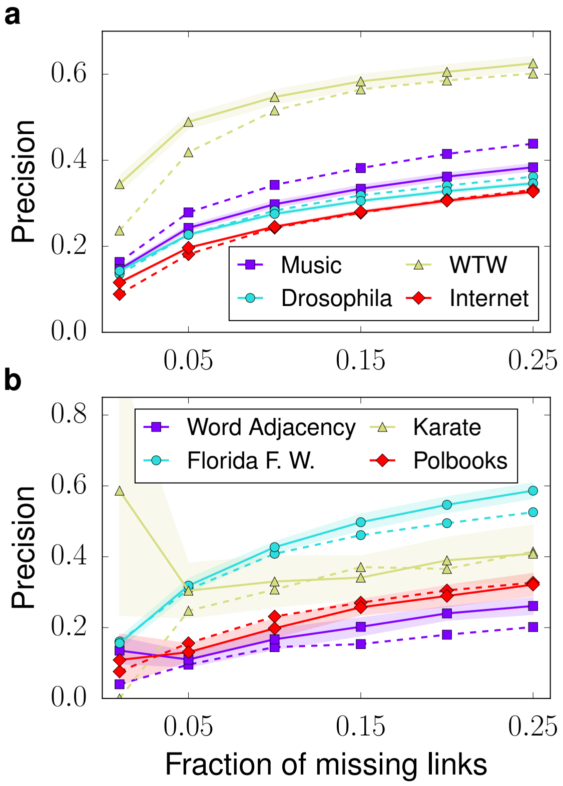

FIG. 4. **Inference of predictability on eight different real-world networks.** The dashed curves show the mean precisions given by the p_{ij}, that is, the OS predictability curves, as in Fig. 3. For each network, we considered 10 incomplete networks with $q_0 = 0.1$, and for each incomplete network, we computed its inferred predictability. The solid curves and shaded areas represent the average and standard deviation of such estimations over the 10 incomplete networks.

III. DISCUSSION

Link prediction in real networks remains a major challenge. A clear indicator is given by the current prediction accuracy of the methods in experimental tests where a part of the links is randomly removed, with precisions typically far from its absolute maximum, even for the best methods. Part of this seemingly poor performance is explained by the intrinsic unpredictability of networks, whose links are formed following processes that can be mimicked by stochastic connectivity rules determining the likelihood of interactions. Our probabilistic approach to the predictability problem makes sense as far as this assumption is fulfilled.

The optimal strategy for link prediction on networks belonging to some model network ensemble, corresponding to ranking the likelihood of missing links according to the ensemble connection probabilities, outperforms all the link prediction methods used on different network ensembles. This implies that identifying the model that best describes the connectivity of a given incomplete network and inferring the ensemble connection probabilities generating the complete network would yield the best link prediction accuracies. We have illustrated this claim by designing such link prediction strategy for soft Configuration Model networks, the CMA method, which gives, by far, the best predictions on such graphs, nearly reaching the theoretical maximum. Since the sCM misses several key properties of real networks, like the high level of clustering, the CMA method does not perform well in real situations. However, our results serve as a proof of principle motivating to pursue a similar line of model-based link prediction methods with some more realistic network models, like the $\mathbb{S}^1$ and the dc-SBM. Furthermore, the precision of the optimal strategy yields a novel indicator of the inherent predictability of network models. In particular, by calculating the predictability curves for different values of the parameters of the models hereby considered, we have quantified how several topological properties affect the resulting network predictability. As a general trend, network predictability increases as the graphs depart from purely random structures.

In real networks, we propose a method to assess predictability based on the assumption that they are well described by probabilistic network models. It should be clarified that we are not assuming that real networks are randomly sampled elements from some ensemble model. Instead, the reason why we are addressing link prediction from a model-based approach is the following. For a given observed graph G_{obs} that we deem incomplete, there generally exists a set of candidate original networks $\{G'\}$, each of them containing all the links in G_{obs} plus some others. What is more, some of these will have the same probability of yielding G_{obs} upon removal of links according to the noise channel. Hence, any link prediction method must choose among these, and it must do so by making assumptions on how the structure of the original graph must have been like. These assumptions are often based on some *expected* topological properties, like the presence of triangles on which many link prediction methods rely (or, as in the case of SPM, one assumes certain behaviour upon a stochastic perturbation). What we propose instead is to base those necessary assumptions on more elaborate expected topological properties of real networks, which have been vastly studied and are (to a higher or lesser degree) captured by network models.

The OS predictability curve can be approximated from fully observed real networks by inferring the corresponding model ensemble probabilities and by measuring the precision of the optimal strategy with them. This curve can be used as a benchmark to assess the goodness of link prediction methods, as it allows for their performances to be contrasted against the best possible performance over classes of networks which are statistically similar to the one under study. The inference of the ensemble connection probabilities is, however, a difficult task even in fully observed real networks. Its reliability is subject to the congruency between the network and the probabilistic model that best describes the network structure.

Typically, a network model can describe correctly the observed connectivity structure of a real system only to a limited extent. Nevertheless, selecting the most plausible model —in terms of its posterior probability using variations of the stochastic block model— correlates with the highest predictive performance in terms of missing links [32]. Hence, the computation of full posterior distribution might in principle make possible to achieve the OS precision in degree-corrected SBM. However, this is a difficult task and relative probabilities between individual missing links are computed instead [32].

On the other hand, it may happen that some particular link prediction method, more tailored for a single network, yields a better result than the optimal strategy. Yet, in terms of the ensemble, such a method would be overfitted and perform worse on average on the set of similar networks defined by the same set of p_{ij}. This has clear implications, for instance, in the prediction of missing links as future events in time-evolving networks. A link prediction method that is overfitted to specific realisations, like present network snapshots, will certainly fail more easily in foreseeing future connections.

A different issue is that, in situations in which one may want to assess to what extent a given incomplete real network can be predicted, the ensemble probabilities cannot be directly inferred from the original network, as it is unknown. Given that it is in general very difficult to infer the original ensemble probabilities from the incomplete network—which could be further used for actual link prediction—for models other than the sCM, we propose a method to approximate the OS predictability curve in networks with missing links, with good accuracy. We remark that a good approximation of the OS predictability curve is not a guarantee that the calculated probabilities are accurate enough to apply the optimal strategy as an efficient link prediction method. This is, for instance, the case of the $\mathbb{S}^1$ ensemble in the real-network experiments shown in Fig. 3, for which the optimal strategy works well as a link prediction method when using the inferred probabilities of the complete network but gives bad results (not shown) when using the ones calculated from the incomplete networks [33]. The reason for this phenomenon can be understood from the description of the algorithm that we use to approximate the OS predictability curve, where we only use the highest numerical values of the connection probabilities, without any mention whatsoever to the pair of nodes they refer to. Hence, as long as the distribution of the values of the highest probabilities is not drastically perturbed by the link removal—that is, if the highest values of the probabilities inferred on the original and incomplete networks exhibit similar distributions—the OS predictability curve can be estimated, even if the specific probabilities corresponding to the removed edges change considerably and, as a result, they do not enable a good link prediction. Therefore, the approximation of the OS predictability cure in real networks with missing links gives a predictability limit that can be used as a benchmark to gauge the quality of link prediction methods in real networks.

Appendix A: Optimal prediction strategy for a graph ensemble

We prove that the optimal strategy for link prediction, that is, the one maximizing the expected precision in link prediction experiments over ensemble instances, is the one that generates $G_{\rm inf}$ from $G_{\rm obs}$ by adding the links according to the connection probabilities $\{p_{ij}\}$ ranked in decreasing order. We compute the expected precision as

$$\begin{aligned}\langle Q\rangle &= \sum_{G}\sum_{G_{\rm obs}} P(G,G_{\rm obs})Q(G,G_{\rm obs},G_{\rm inf}) \\ &= \sum_{G_{\rm obs}} P(G_{\rm obs})\bar{Q}(G_{\rm obs},G_{\rm inf}),\end{aligned} \tag{A1}$$

where $P(G,G_{\rm obs})$ is the joint probability distribution for a graph G in the ensemble and an observed graph $G_{\rm obs}$. We have defined $\bar{Q}(G_{\rm obs},G_{\rm inf})$ as the expected precision of the link prediction method over all possible original graphs yielding $G_{\rm obs}$ upon random removal of links. Notice that, in the summation over original graphs G, we must take into account that $G_{\rm inf}$ is independent of G; this is the crucial property leading to a limit to the predictability of missing links. Indeed, since more than one original network G can generate the same $G_{\rm obs}$ upon random link removal, it is impossible for any link prediction method, which maps $G_{\rm obs}$ into *the same* inferred network $G_{\rm inf}$ regardless of the original G, to give a perfect prediction.

We simplify the notation by enumerating all potential links (disconnected pairs of nodes) in $G_{\rm obs}$, such that their ensemble probabilities can be written as p_l, where the index l runs from 1 to the number of potential links M. Given the corresponding adjacency matrix elements $\{a_l\}_{1\le l\le M}$ of G, for every potential link l $P(a_l=1|a_l^{\rm obs}=0)=P(a_l=1,a_l^{\rm obs}=0)/P(a_l^{\rm obs}=0)=qp_l/(1-p_l+qp_l)$, where we have used Bayes' rule. Then, the probability for any graph G compatible with the observed graph $G_{\rm obs}$, $P(G|G_{\rm obs})$, can be expressed in terms of the set of pairs as

$$P(G|G_{\rm obs})=\prod_{l=1}^{M}\frac{(1-p_l)^{1-a_l}(qp_l)^{a_l}}{1-p_l+qp_l}. \tag{A2}$$

Let us furthermore define the vector $\mathbf{v}=(a_1,\ldots,a_M)$, which characterizes the set of potential links in G, and the analogous vector $\mathbf{v}_{\rm inf}=(a_1^{\rm inf},\ldots,a_M^{\rm inf})$ for $G_{\rm inf}$. With these two vectors, we can now express the precision as

$$Q(G,G_{\rm obs},G_{\rm inf})=\frac{1}{L}\mathbf{v}\cdot\mathbf{v}_{\rm inf}, \tag{A3}$$

where $L=\sum_l a_l$ is the number of missing links in $G_{\rm obs}$ with respect to G. Hence, we can express $\bar{Q}(G_{\rm obs},G_{\rm inf})$

in Eq. A1 as

$$\begin{aligned}
&\bar{Q}(G_{\mathrm{obs}}, G_{\mathrm{inf}}) = \\
&= \sum_{G|G_{\mathrm{obs}} \in \mathcal{S}(G)} P(G|G_{\mathrm{obs}}) Q(G, G_{\mathrm{obs}}, G_{\mathrm{inf}}) \\
&= \sum_{G|G_{\mathrm{obs}} \in \mathcal{S}(G)} \prod_{l=1}^{M} \frac{(1-p_l)^{1-a_l}(qp_l)^{a_l}}{1-p_l+qp_l} \frac{1}{L} \mathbf{v} \cdot \mathbf{v}_{\mathrm{inf}} \qquad \text{(A4)} \\
&= \left(\sum_{G|G_{\mathrm{obs}} \in \mathcal{S}(G)} \prod_{l=1}^{M} \frac{(1-p_l)^{1-a_l}(qp_l)^{a_l}}{1-p_l+qp_l} \mathbf{v} \right) \cdot \frac{\mathbf{v}_{\mathrm{inf}}}{L},
\end{aligned}$$

where $\mathcal{S}(G)$ stands for the set of subgraphs of G. In the above calculation, we have used the linearity of the scalar product and neglected the fluctuations in the number of missing links (we assume that all original graphs generating the observed graph upon random link removal with probability q have approximately the same number of links, $L = q \sum_i p_i$, with the sum now taken over all pairs of nodes). One could actually give an exact result, with no assumptions or approximations about the number of missing links, by defining the precision as the fraction of inferred links actually belonging to the original graph (true positive rate). In that case, both L and the inferred vector are the outcome of the link prediction method and the above expression is exact.

Let us call the vector within the parenthesis in the equation above $\bar{\mathbf{v}}$. Its n-th component can be computed as

$$\begin{aligned}
&\sum_{G|G_{\mathrm{obs}} \in \mathcal{S}(G)} \prod_{l=1}^{M} \frac{(1-p_l)^{1-a_l}(qp_l)^{a_l}}{1-p_l+qp_l} a_n = \\
&= \sum_{a_1=0}^{1} \cdots \sum_{a_M=0}^{1} \prod_{l=1}^{M} \frac{(1-p_l)^{1-a_l}(qp_l)^{a_l}}{1-p_l+qp_l} a_n = \\
&= \frac{qp_n}{1-p_n+qp_n} \prod_{l \neq n} \left(\frac{qp_l}{1-p_l+qp_l} + \frac{1-p_l}{1-p_l+qp_l} \right) \\
&= \frac{qp_n}{1-p_n+qp_n}.
\end{aligned} \tag{A5}$$

Hence, $\bar{\mathbf{v}} = \left(\frac{qp_1}{1-p_1+qp_1}, \ldots, \frac{qp_M}{1-p_M+qp_M} \right)$.

We find that the average over graphs of the ensemble of the precision given G_{obs} can be expressed as the scalar product of two vectors,

$$\bar{Q}(G_{\mathrm{obs}}, G_{\mathrm{inf}}) = \frac{1}{L} \bar{\mathbf{v}} \cdot \mathbf{v}_{\mathrm{inf}}, \tag{A6}$$

where L is the number of missing links in G_{obs} with respect to G. The dimension of the vectors equals the number M of potential links (disconnected pairs of nodes) in G_{obs}. If we enumerate the ensemble probabilities as $\{p_l\}$, we can write $\bar{\mathbf{v}} = \left(\frac{qp_1}{1-p_1+qp_1}, \ldots, \frac{qp_M}{1-p_M+qp_M} \right)$, where each entry gives the probability that the corresponding link, missing in G_{obs}, is in G. The entries in vector $\mathbf{v}_{\mathrm{inf}} = \left(a_1^{\mathrm{inf}}, \ldots, a_M^{\mathrm{inf}} \right)$ correspond to the adjacency-matrix elements of G_{inf} for the set of potential links of G_{obs}.

The precision is then maximized when the vectors are maximally aligned. By definition of the link prediction method $\mathcal{M}$, $\mathbf{v}_{\mathrm{inf}}$ is a vector containing L values equal to one, while the rest of entries are zero. Clearly, the maximum value for the precision will be obtained if its non-zero entries are placed at the same positions where the L largest components of $\bar{\mathbf{v}}$ are located. Therefore, given that $\bar{\mathbf{v}}_i > \bar{\mathbf{v}}_j \Leftrightarrow p_i > p_j$, the best link prediction method is the one that adds the L missing links according to the highest connection probabilities in the ensemble. Moreover, the expected optimal precision for the observed graph is given by the mean of the L largest components of $\bar{\mathbf{v}}$.

Appendix B: Expected precision of the optimal strategy in network ensembles

Let ensemble $\mathcal{E}_N$ be characterised by the set of connection probabilities $\{p_l\}$, where index l runs over all possible pairs of N nodes such that $p_l \geq p_{l+1}$, $\forall l$. If links are removed with probability q, the probability for an edge l not to belong to G_{obs} is given by the sum of the probabilities for it not to belong to G, $P\left(a_l = 0, a_l^{\mathrm{obs}} = 0\right) = 1 - p_l$, and for it to belong to G and being randomly removed, $P\left(a_l = 1, a_l^{\mathrm{obs}} = 0\right) = qp_l$, that is,

$$P\left(a_l^{\mathrm{obs}} = 0\right) = 1 - p_l + qp_l. \tag{B1}$$

Now, in order to compute the expected precision of the optimal strategy, the basic idea is to compute the expected number of correct predictions, T, when following the ranking of probabilities until the expected number of non-observed links, H, matches the expected number of missing links, $L = q \sum_l p_l$. Hence, we initialise both T and H to zero and, for every link $l = 1, \ldots$, we update them as

$$T_{\mathrm{new}} = T_{\mathrm{old}} + P\left(a_l = 1, a_l^{\mathrm{obs}} = 0\right) = T_{\mathrm{old}} + qp_l \tag{B2}$$

and

$$H_{\mathrm{new}} = H_{\mathrm{old}} + P\left(a_l^{\mathrm{obs}} = 0\right) = H_{\mathrm{old}} + 1 - p_l + qp_l. \tag{B3}$$

When $H \approx L$, T approximately yields the number of correct predictions of the OS and, therefore, the expected precision can be computed as $\langle Q \rangle = T/H$. This algorithm can be repeated for different values of q in order to obtain the expected predictability curve of the ensemble.

In ER networks $p_l = p$, $\forall l$, and hence the expected precision is $\langle Q \rangle = \frac{qp}{1-p+qp}$, which agrees with the exact value. Moreover, all link prediction methods reach this precision. To understand this, all possible distributions of the L values equal to one among the (otherwise zero) different components of vector $\mathbf{v}_{\mathrm{inf}}$ in the proof of the previous section in Appendix A yield the same scalar product $\bar{\mathbf{v}} \cdot \mathbf{v}_{\mathrm{inf}}/L$ and, hence, the same precision.

Appendix C: Network Ensemble Models

1. Soft Configuration Model (sCM)

In the soft Configuration Model [13], each node i is assigned an expected degree κ_i, and each pair of nodes i and j is connected according to the ensemble connection probabilities given by

$$p_{ij} = \frac{\mu\kappa_i\kappa_j}{1+\mu\kappa_i\kappa_j}, \tag{C1}$$

with μ a free parameter controlling the number of resulting edges in the network. If one takes $\mu = 1/(\langle k\rangle N)$, then the degree of every node i in the generated networks, k_i, is approximately its expected degree, $k_i \approx \kappa_i$.

Given the degrees k_i^{obs} in an observed graph which has been produced by removing a fraction q of nodes from a complete graph in the CM ensemble, we can estimate the expected degrees and the connection probabilities in the complete graph from Eq. C1 as

$$\tilde{p}_{ij} = \frac{\frac{k_i^{\mathrm{obs}}k_j^{\mathrm{obs}}}{(1-q)\langle k^{\mathrm{obs}}\rangle N}}{1+\frac{k_i^{\mathrm{obs}}k_j^{\mathrm{obs}}}{(1-q)\langle k^{\mathrm{obs}}\rangle N}}. \tag{C2}$$

2. $\mathbb{S}^1$ model

In the $\mathbb{S}^1$ model [14], every node i is characterized by a hidden degree and an angular coordinate (κ_i, θ_i) representing the popularity and similarity dimensions. The angular coordinate is distributed at random in similarity space, which is taken to be a one-dimensional sphere, or circle, of radius R adjusted to have a density of nodes equal to 1. Every pair of nodes is connected with a probability

$$p_{ij} = \frac{1}{1+\left(\frac{R\Delta\theta_{ij}}{\mu\kappa_i\kappa_j}\right)^{\beta}}, \tag{C3}$$

where $\Delta\theta_{ij}$ stands for the angular separation between the nodes in the similarity circle, and the parameters μ and β control the average degree of the network and its level of clustering, respectively. In the limit of $N\to\infty$, and for large degrees, the expected degree $\langle k_i\rangle$ of a node i in the generated network is its hidden degree $\langle k_i\rangle = \kappa_i$.

3. Degree-corrected Stochastic Block Model (dc-SBM)

In the dc-SBM model [15], each node i is assigned an expected degree k_i and a group g_i determining the community to which it belongs, which is chosen in an arbitrary way. Then, parameter θ for every node i is computed as

$$\theta_i = \frac{k_i}{\kappa_{g_i}}, \tag{C4}$$

where κ_{g_i} is the sum of the degrees of all the nodes in group g_i. Therefore, each group g fulfills the constraint

$$\sum_{i\in g}\theta_i = 1. \tag{C5}$$

Finally, ω is a matrix of size $K\times K$ controlling the number of links between pairs of groups, where K is the total number of groups. Each element of the matrix is calculated as

$$\omega_{rs} = \lambda\omega_{rs}^{planted} + (1-\lambda)\omega_{rs}^{random}, \tag{C6}$$

where ω_{rs}^{random} corresponds to a random network with specific expected degree sequence, $\omega_{rs}^{random} = \kappa_r\kappa_s/2m$, where m is the total number of links in the network. On the other hand, $\omega_{rs}^{planted}$ generates group structure. For example, in a network with four groups, this matrix is given by

$$\omega^{planted} = \begin{bmatrix} \kappa_1 & 0 & 0 & 0 \\ 0 & \kappa_2 & 0 & 0 \\ 0 & 0 & \kappa_3 & 0 \\ 0 & 0 & 0 & \kappa_4 \end{bmatrix}. \tag{C7}$$

When $\lambda = 0$ links are placed among pairs of nodes at random considering the degree sequence, while when $\lambda = 1$ links are located within communities. Therefore, any other values for λ will result in a combination of the above extremes.

In the dc-SBM model, the number of links placed among two nodes i and j follows a Poisson distribution with mean $\theta_i\theta_j\omega_{g_i,g_j}$. However, in the sparse-network limit, the probability for multi-edges to occur is generally low, so $\theta_i\theta_j\omega_{g_i,g_j}$ is simply taken to be the connection probability. Since these amounts can be larger than 1, in this work, we consider

$$p_{ij} = \frac{\theta_i\theta_j\omega_{g_i,g_j}}{1+\theta_i\theta_j\omega_{g_i,g_j}}. \tag{C8}$$

4. Likelihood-based model selection

In this work, we consider two candidate models for each real network: the dc-SBM and the $\mathbb{S}^1$ model. We must therefore decide which one is more appropriate for each network. To do so, we simply fit both and assess which one is more congruent with the real data. Since both models considered yield pairwise connection probabilities, we can calculate the corresponding *likelihood* as

$$\mathcal{L} = \prod_{i<j} p_{ij}^{a_{ij}}\left(1-p_{ij}\right)^{1-a_{ij}}. \tag{C9}$$

In the above expression, the p_{ij} encapsulate the model connection probabilities, whereas a_{ij} are the adjacency matrix elements given by the data. The likelihood $\mathcal{L}$

hence gives the probability for the network under consideration to be generated by the model. Therefore, if the data is congruent with the model (and the fit is conducted appropriately), the likelihood will reach relatively high values. Now, in order to choose between the models, we select the one with higher likelihood (notice that, in practice, it is useful to consider the log-likelihood $\log \mathcal{L}$ instead, as it is less problematic numerically). Since higher likelihood implies higher connection probabilities on existing links and lower on non-existing ones, the OS curve will be higher in the former case. In the table below, we report the difference between the log-likelihoods of the dc-SBM and the $\mathbb{S}^1$ model for all the networks considered in the paper.

Appendix D: Data Description

Music [21] The nodes in the Music network represent codewords extracted for every single chord in a large set of songs, and directed links connecting consecutive codewords represent transitions among them. To sparsify the network, the disparity filter [34] is applied with parameter $\alpha = 0.01$. Finally, we consider an undirected version of network by replacing bidirectional links with undirected ones.

Drosophila [22] Nodes represent neurons within the Drosophila optic medulla and links represent fiber tracts connecting neurons.

WTW [23] Backbone of the international trade network in 2013, where nodes represent countries and links are placed among significant trade partners.

Internet [24] Internet topology at the level of Autonomous Systems (AS) corresponding to June 2009 and collected by the Cooperative Association for Internet Data Analysis (CAIDA). We removed nodes with degree lower than 5 to produce a reduced size version.

Florida Food Web [25] Food web in the Florida Bay ecosystem, in which every directed link connects a prey to its predator. We consider the undirected version of this network created by placing an undirected link between every pair of nodes connected by at least a single directed link.

Word Adjacency [26] Adjacency network where nodes represent a selected set of common nouns and adjectives in the novel of David Copperfield by Charles Dickens, and links are placed between adjacent pairs of words in the book.

Karate [27] Social network between members of a karate club where each link connects a pair of members who communicate outside the club.

Polbooks [28] Nodes of this network represent the books on the topic of the US politics and links represent the pairs of books bought on Amazon by the same customers.

ACKNOWLEDGMENTS

We thank M. E. J. Newman for kindly sharing with us the code to estimate the probabilities of connection in the degree corrected stochastic block model. We thank Marián Boguñá for helpful discussions. G. G.-P. acknowledges financial support from the Academy of Finland via the Centre of Excellence program (Project no. 312058 as well as Project no. 287750), and from the emmy.network foundation under the aegis of the Fondation de Luxembourg. R. A. acknowledges the support of the international affairs of K. N. Toosi University of Technology and M. A. S. and the University of Barcelona for their support during her visit. M. A. S acknowledges support from a James S. McDonnell Foundation Scholar Award in Complex Systems; Ministerio de Ciencia, Innovación y Universidades of Spain project no. FIS2016-76830-C2-2-P (AEI/FEDER, UE); and the project *Mapping Big Data Systems: embedding large complex networks in low-dimensional hidden metric spaces* – Ayudas Fundación BBVA a Equipos de Investigación Científica 2017.

AUTHOR CONTRIBUTIONS

G. G.-P., R. A., A. G. and M. A. S. contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

ADDITIONAL INFORMATION

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

[1] J. Slingo and T. Palmer, Phil Trans R Soc A **369**, 4751 (2011).

[2] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, Science **327**, 1018 (2010).

[3] P. Nosil, R. Villoutreix, C. F. de Carvalho, T. E. Farkas, V. Soria-Carrasco, J. L. Feder, B. J. Crespi, and Z. Gompert, Science **359**, 765 (2018).

[4] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, Inc., New York, NY, USA, 2010).

[5] A. Barrat, M. Barthlemy, and A. Vespignani, *Dynamical Processes on Complex Networks*, 1st ed. (Cambridge University Press, New York, NY, USA, 2008).

[6] F. Radicchi and C. Castellano, Phys Rev Lett **120**, 198301 (2018).

[7] L. Lü, L. Pan, T. Zhou, Y.-C. Zhang, and H. E. Stanley, Proc Natl Acad Sci USA **112**, 2325 (2015), http://www.pnas.org/content/112/8/2325.full.pdf.

[8] L. Lü and T. Zhou, Physica A **390**, 1150 (2011).

Network	$\log \mathcal{L}_{\mathbb{S}^1} - \log \mathcal{L}_{\mathrm{dcSBM}}$	Network	$\log \mathcal{L}_{\mathbb{S}^1} - \log \mathcal{L}_{\mathrm{dcSBM}}$
Music	7130.91	Florida F. W.	-912.88
Drosophila	5457.41	Word Adjacency	-915.44
WTW	629.37	Karate	-166.19
Internet	5574.08	Polbooks	-1600.16

TABLE I. Difference in the log-likelihoods $\log \mathcal{L}_{\mathbb{S}^1} - \log \mathcal{L}_{\mathrm{dcSBM}}$ for the two models used in this work. A positive value indicates a higher congruency with the $\mathbb{S}^1$ model, whereas a negative value, with the dc-SBM.

[9] D. Liben-Nowell and J. Kleinberg, J AM SOC INF SCI TEC **58**, 1019 (2007).

[10] P. Erdös and P. Rényi, Publ Math Debrecen **6**, 290 (1959).

[11] E. N. Gilbert, Ann Math Stat **30**, 1141 (1959).

[12] L. Lü, C.-H. Jin, and T. Zhou, Phys Rev E **80**, 046122 (2009).

[13] J. Park and M. E. J. Newman, Phys Rev E **68**, 026112 (2003).

[14] M. Á. Serrano, D. Krioukov, and M. Boguñá, Phys Rev Lett **100**, 078701 (2008).

[15] B. Karrer and M. E. J. Newman, Phys Rev E **83**, 016107 (2011).

[16] M. E. J. Newman, Phys Rev E **64**, 025102 (2001).

[17] L. A. Adamic and E. Adar, Soc Networks **25**, 211 (2003).

[18] T. Zhou, L. Lü, and Y.-C. Zhang, Eur Phys J B **71**, 623 (2009).

[19] A. Muscoloni and C. V. Cannistraci, arXiv:1707.09496. Preprint, posted Jul 29, 2017 (2017).

[20] Z. Liu, J.-L. He, K. Kapoor, and J. Srivastava, PloS one **8**, e72908 (2013).

[21] J. Serrà, Á. Corral, M. Boguñá, M. Haro, and J. L. Arcos, Sci Rep **2**, 521 (2012).

[22] S.-y. Takemura, A. Bharioke, Z. Lu, A. Nern, S. Vitaladevuni, P. K. Rivlin, W. T. Katz, D. J. Olbris, S. M. Plaza, P. Winston, *et al.*, Nature **500**, 175 (2013).

[23] G. García-Pérez, M. Boguñá, A. Allard, and M. Á. Serrano, Sci Rep **6**, 33441 (2016).

[24] K. Claffy, Y. Hyun, K. Keys, M. Fomenkov, and D. Krioukov, in *Conference For Homeland Security, 2009. CATCH'09. Cybersecurity Applications & Technology* (IEEE, 2009) pp. 205–211.

[25] R. E. Ulanowicz and D. L. DeAngelis, US Geological Survey Program on the South Florida Ecosystem **114**, 45 (2005).

[26] M. E. J. Newman, Phys Rev E **74**, 036104 (2006).

[27] W. W. Zachary, J Anthropol Res **33**, 452 (1977).

[28] M. E. J. Newman, Proc Natl Acad Sci USA **103**, 8577 (2006).

[29] M. Boguná, F. Papadopoulos, and D. Krioukov, Nat Commun **1**, 62 (2010).

[30] F. Papadopoulos, R. Aldecoa, and D. Krioukov, Phys Rev E **92**, 022807 (2015).

[31] M. A. Riolo, G. T. Cantwell, G. Reinert, and M. E. J. Newman, Phys Rev E **96**, 032310 (2017).

[32] T. Vallès-Català, T. P. Peixoto, M. Sales-Pardo, and R. Guimerà, Phys Rev E **97**, 062316 (2018).

[33] This might seem to suggest that the $\mathbb{S}^1$ model overfits the data. However, the model only has $\mathcal{O}(N)$ parameters, so it cannot overfit graphs with $\mathcal{O}(N^2)$ degrees of freedom. Instead, the reason why it fails to provide a good prediction from an incomplete network stems from the fitting procedure, i.e. the embedding, in which we find the model parameters that maximise the likelihood for the observed network to be generated by the model. Hence, we are manifestly inferring the model parameters that best explain the observed—incomplete—topology.

[34] M. Á. Serrano, M. Boguná, and A. Vespignani, Proc Natl Acad Sci USA **106**, 6483 (2009).