From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks

Carlo Vittorio Cannistraci, Gregorio Alanis-Lobato and Timothy Ravasi

Methods and Supplementary Information

CONTENTS

I. INTRODUCTION

This document serves as Methods and Supplementary Information for the article *From Link Prediction in Brain Connectomes and Protein Interactomes to the Local-community Paradigm in Complex Networks* and is organised as ten sections. Subsequent to the current Introduction, Section II briefly summarises the fundamental ideas behind link prediction by exclusive means of network topology; it also describes the applicability of this process to the network representation of different systems. Section III describes the most commonly used prediction techniques (indices). Some of these predictors are general purpose (Section III.A) and some are tailored for biological networks (Section III.B), another more sophisticated group is based on statistical inference (Section III.C), and we also introduce variations of the classical prediction approach so that they adopt a link/community strategy to predict (Section III.D). Section IV provides an in-depth description of the methodology applied to the destruction-prediction experiment in brain connectomes while Section V gives a detailed explanation of how prediction in protein interactomes is performed. At comparable levels of detail, Sections VI and VII examine the concepts of Performance and Time Robustness and of the In-silico validation of the best prediction tools as applied to protein-interaction networks. Section VIII expands the main article's explanation of the Local Community Paradigm (LCP) Decomposition Plot and the way in which the LCP-correlation is computed. Finally, Section IX lists the network data analysed in this study and depicts the LCP-Decomposition Plots for networks that are and are not characterised by LCP organisation within their topology.

II. PREDICTION OF LINKS IN NETWORKS

The idea of predicting links on the mere basis of network topology originates from the fact that part of the information associated with the organisation of the topology itself was generated gradually and in accordance with the growth process of the network. Links between nodes exist because two people have similar interests (social networks), two proteins are bound together to perform a specific function (protein interaction networks), or two cities are connected through a direct flight (flight maps); nodes and edges are only abstractions of the dynamics that exist within complex systems, and the information contained within these abstractions can definitely be exploited for the prediction of new links.

How to take advantage of existing network topologies to predict missing edges is still not clear. However, techniques have been developed to do so, both in unspecific networks (here referred to as General Purpose and Maximum Likelihood techniques, see Sections III.A and III.C) and in very specific ones, such as Protein-protein Interaction Networks (PPINs, see Section III.B).

Taking the network of interest as the input, these prediction techniques proceed as follows:

1. Assign a likelihood score to every pair of nodes that is not directly connected in the current network topology (each of these pairs is known as 'candidate interaction').
2. Sort the list of candidate interactions by score (for some approaches, the larger the score, the more the interaction is likely to be real, for others the scoring works in the opposite way).

Performance assessment for any given prediction tool can avail of several methods; specific choices will depend mainly on the type of network and the available information about the nodes and links. The most common approaches are:

a. Some systems allow access to their network representations at different timestamps. One example is Facebook; the number of links at time t_{i+1} is (almost certainly) larger than at time t_i. Thus, one can apply a prediction technique to the network topology at t_i and verify whether the links at the top of the candidate interaction list appear in the network at t_{i+1}.

b. When networks at different timestamps are not available, the procedure is to remove *n* links randomly from the available network, obtain a candidate interaction list that is sorted by likelihood using a prediction technique, and take *n* links from the top of this list to verify if they match those removed from the network (see Section IV for a detailed example of this approach).
c. For some systems, it is possible to access reliable node property information that serves as a criterion to decide whether two nodes can interact or not. An example is Gene Ontology (a vocabulary for processes, functions and localization of genes and proteins). If, for instance, two proteins perform the same function and are located in the same cellular compartment, they are likely to interact. Thus, the proportion of interactions at the top of the candidate list that fulfils the given criterion constitutes a prediction technique performance measure (see Section V for a detailed example of this approach).

III. TOPOLOGICAL TECHNIQUES (INDICES) FOR LINK PREDICTION

A. General purpose

Numerous indices have been proposed to predict links in differing types of networks on the exclusive basis of the given network's topology. For examples of indices, see the surveys by (Liben-Nowell and Kleinberg, 2007) and (Lü and T. Zhou, 2011). We decided to use the most reliable and referential indices, which have the merit of being fast and parameter-free, and we measured their prediction performance. The indices we selected are Common Neighbours (CN), Jaccard (JC), Preferential Attachment (PA), Adamic and Adar (AA), and Resource Allocation (RA).

CN (M. Newman, 2001) follows the basic idea that two people are more likely to meet if they have common acquaintances. If $\Gamma(x)$ is the set of neighbours of *x,* and $\Gamma(y)$ is the set of neighbours of *y*, the CN score is defined as in Equation S1.

$$CN(x,y) = |\Gamma(x) \cap \Gamma(y)| \tag{S1}$$

JC (Jaccard, 1912) measures the probability that both *x* and *y* share a common neighbour (see Equation S2).

$$JC(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{S2}$$

PA (M. Newman, 2001) is based on a process of the same name, where a certain quantity is distributed according to the number of entities it already has. In networks, these entities are nodes and they prefer to connect with others that have a high number of connections. Therefore, this index is given as in Equation S3.

$$PA(x,y) = |\Gamma(x)| \cdot |\Gamma(y)| \tag{S3}$$

AA (Adamic and Adar, 2003) and RA (T. Zhou et al., 2009) are two similar indices that assign better scores to candidate interactions whose common neighbours have very few other neighbours (see Equations S4 and S5 respectively).

$$AA(x,y) = \sum_{s \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log_2(|\Gamma(s)|)} \tag{S4}$$

$$RA(x,y) = \sum_{s \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(s)|} \tag{S5}$$

B. Bio-inspired indices

When prior functional biological knowledge about the proteins that form a PPIN is not available, the only biological resource we are left with is the information allocated in the PPIN topology itself. Several techniques have been proposed to exploit the topology of a PPIN in order to assess the reliability of the network interactions or to predict protein function.

The pioneers of Protein-protein Interaction (PPI) reliability assessment are Saito and his team. In 2002, they proposed the so-called Interaction Generality (IG1) Index which, given that partners of 'sticky' proteins and self-activators do not interact with anything else in the network (J. Chen, Chua, et al., 2006), assigns high index values to potential false positives (interactions whose seed proteins x and y have a lot of neighbours that do not interact with anything else) and low values to more reliable PPIs (see Equation S7, where *G* is a network, and *x*, *y*, *x*', and *y*' are nodes in *G*).

$$IG1(x,y) = 1 + |\{\{x',y'\} \in G \big| \boldsymbol{x}' \in \{\boldsymbol{x}',\boldsymbol{y}'\}, \boldsymbol{y}' \notin \{\boldsymbol{x},\boldsymbol{y}\}, \boldsymbol{\Gamma}(\mathbf{y}') = \mathbf{1}\}| \tag{S7}$$

Later, further indices were proposed, such as Interaction Generality Two (IG2) (Saito et al., 2003) or Interaction Reliability by Alternative Path (IRAP) (J. Chen et al., 2005), but their minimal comparative performance makes them very computationally expensive (J. Chen et al., 2005). On the other hand, indices to predict functions such as Czekanowski-Dice Dissimilarity (CDD) (Brun et al., 2003), Functional Similarity Weight (FSW) (Chua et al., 2006), and Adjust Czekanowski-Dice Dissimilarity (ACDD) (Liu et al., 2009) were additionally and successfully employed for reliability assessment (J. Chen, Chua, et al., 2006; Liu et al., 2009) and link prediction in PPINs (Liu et al., 2009; You et al., 2010). Equations S8-S10 respectively represent the formulae of these last three indices . In these equations $\gamma(x)$ is the set of neighbours of *x* including itself, and n_{avg} is the average node degree of the network.

$$CDD(x,y) = \frac{|\gamma(x)\Delta\gamma(y)|}{|\gamma(x) \cup \gamma(y)| + |\gamma(x) \cap \gamma(y)|} \tag{S8}$$

$$FSW(x,y) = \frac{2|\gamma(x) \cap \gamma(y)|}{|\gamma(x) - \gamma(y)| + 2|\gamma(x) \cap \gamma(y)| + \lambda_{x,y}} \cdot \frac{2|\gamma(x) \cap \gamma(y)|}{|\gamma(y) - \gamma(x)| + 2|\gamma(x) \cap \gamma(y)| + \lambda_{y,x}} \tag{S9}$$

$$ACDD(x,y) = \frac{|\gamma(x)\Delta\gamma(y)| + \lambda_{x,y} + \lambda_{y,x}}{|\gamma(x) \cup \gamma(y)| + |\gamma(x) \cap \gamma(y)|} \tag{S10}$$

where $\lambda_{x,y} = \max(0, n_{avg} - (|\gamma(x) - \gamma(y)| + |\gamma(x) \cap \gamma(y)|))$.

In 2009, Kuchaiev et al. proposed a method for geometric denoising of PPINs. The algorithm is based on multidimensional scaling (MDS) which is used to preserve the network shortest-path (SP) lengths between nodes in a low dimensional space, where the network after embedding is proven to be denoised. The new candidate links are scored according to their Euclidean distance (ED) in the low dimensional space, following the principle that the closer two proteins are, the higher the likelihood that they interact (Kuchaiev et al., 2009). Although it is not explicitly mentioned in the article, the embedding method adopted by Kuchaiev et al. is equivalent to Isomap (Tenenbaum, et al., 2000).

In 2010, in an independent study, You et al. proposed a hybrid strategy based on network embedding to assign prediction scores to candidate interactions. They exploited the notion that a PPIN - like theoretically any network - lies on a low dimensional manifold shaped in a high-dimensional space. The shape of the manifold and the associated topology is a result of the constraints imposed on the protein interactions by the biological evolution. You et al. used a renowned algorithm for manifold embedding, Isomap (Tenenbaum, et al., 2000), to embed the PPIN in a space of reduced dimensionality. Then, they applied FSW to the embedded

network (pruned according to a cutoff on the ED) to assign the likelihood scores to the candidate interactions

As shown in (J. Chen, Chua, et al., 2006) and mentioned in (You et al., 2010), FSW and CDD outperform IG1, IG2 and IRAP. Accordingly, and because IG2 and IRAP are computationally expensive (J. Chen et al., 2005), we limited our choice of indices to: IG1 (which is a baseline for improvement in the field), FSW, CDD, ACDD, ISOMAP and ISOMAP+FSW.

C. Maximum likelihood indices

As mentioned in sections III.A and III.B, we decided to compare CAR (our proposed approach, see main article and the following sections for details) against parameter-free and efficient prediction techniques. There are, however, two recently published approaches to link prediction that are based on the underlying community structure of real world networks and we decided to take them into account as well.

In the Hierarchical Random Graph or HRG (Clauset et al., 2008), the space of all possible dendrograms of a network is searched to get the ones that best fit the hierarchical structure of such a network. Non-adjacent pairs of nodes that have high average probability of being connected within these dendrograms are then considered good candidates for interaction. In the Stochastic Block Model or SBM (Guimerà and Sales-Pardo, 2009), a general network model in which a network is partitioned into groups, the probability that two nodes are connected only depends on the groups they belong to. For further details and implementations of HRG and SBM, look at the referenced articles and the following websites: http://tuvalu.santafe.edu/~aaronc/hierarchy/ for HRG and http://etseq.urv.cat/seeslab/downloads/network-c-libraries-rgraph/ for SBM.

One important issue with these approaches is that they are very time-consuming and they are not parameter-free (because of this, we used them with their default parameters). The HRG for instance, requires a parameter that indicates the number of dendrogram models that a Markov Chain Monte Carlo algorithm will sample. This sampling process usually requires $O(N^2)$ steps (where N is the number of nodes) and for large N it might require exponential time in the worst case (Lü and T. Zhou, 2011). Similarly, the SBM relies on all possible network partitions to assign link likelihood scores but since the number of partitions grows extremely fast as N grows, a Metropolis algorithm is needed to estimate these scores. Even then, the process is computationally expensive and can only manage networks with a few thousand nodes (Lü and T. Zhou, 2011). For the sake of computational time, we did not perform parameter tuning when applying this techniques.

D. CAR-based variations of general purpose indices

The ranking offered by CAR deviates from the ranking offered by CN, because of the penalization of the candidate links whose common neighbours do not present a community/link structure (the value of their Local Community Links (LCL) is low or equal to zero). Conceptually it means that we penalize the candidate links that deviate from the community structure behaviour, which we proved to be a general characteristic of most real networks. In conclusion, the product of CN and LCL should be interpreted as a logical *AND* operator. Pairs of nodes with very few (or without) LCL attain low scores and migrate to the bottom of the candidate-ranking list.

As mentioned in the first paragraph of the Results section in the main article, more than a new index, our interest is to introduce a new philosophy in the formulation of parameter-free/neighbourhood-based indices. We advocate a shift in perspective from nodes to links, and in particular from nodes to community links. The use of the LCL in CAR's formulation can be seen as an attempt to propose a variation of the CN index that offers more resolution for candidate interactions with equal number of CNs. This boosting in resolution is given by the use of the link/community perspective, which is introduced adopting LCL in the formula.

Thus, if the LCL is seen as an index enhancer, it can be plugged into PA, AA, RA and JC indices so that these techniques also shift to the link/community perspective. In the main body of the article we extensively prove the value of this idea.

Let us consider k_x as the degree of node x, CN the common neighbour index for nodes x and y, and e_x as the number of neighbours of x not shared with y (i.e. $e_x = k_x - CN$). We can write that:

$$PA = k_x \cdot k_y = (e_x + CN)(e_y + CN) = e_x \cdot e_x + e_x \cdot CN + e_y \cdot CN + CN^2$$

and if in this expression we multiply CN by LCL (using the CAR-trick), we obtain the PA's reformulation based on the link/community perspective, which results in:

$$CPA = e_x \cdot e_x + e_x \cdot CAR + e_x \cdot CAR + CAR^2$$

Applying the CAR-trick to the JC formulation, we can modify also JC as below:

$$CJ = \frac{CN}{U} \xrightarrow{CAR-trick} CJC = \frac{CAR}{U}$$

where U indicates the number of elements contained in the union set of the first neighbours of the nodes x and y. On the other hand, an important observation is that the LCL can be written also as:

$$LCL = \sum_{s \in CN} \frac{d_s}{2}$$

where d_s is the inner-community degree of the common neighbour (i.e. the degree of s in the subnetwork formed only by the common neighbours of x and y). This observation allows for the modification of AA, RA according to the link perspective as follows:

$$CAA = \sum_{s \in CN} \frac{d_s}{\log_2(k_s)}$$

$$CRA = \sum_{s \in CN} \frac{d_s}{k_s}$$

In the following we prove that CAR can be considered a reasonable community-based variation of AA, and that CAR is proportional to CAA under certain assumptions.

$$CAR = CN \times LCL = CN \times \sum_{s'' \in CN} \frac{d_{s''}}{2}$$

If we exploit the assumption that our networks are very sparse, it is reasonable to write that the average node degree $\hat{k}$ in the network tends to be equal to two ($\hat{k} \to 2$) with a very small standard deviation. Under this assumption $k_{s'}$ will be a value very close to $\hat{k} \to 2$, and we can approximate the formulation of CAR:

$$CAR \approx \sum_{s' \in CN} \frac{1}{\log_2(k_{s'})} \times \sum_{s'' \in CN} \frac{d_{s''}}{2} =$$

$$= \frac{1}{2} \cdot \left(\sum_{(s' = s'' = s \in CN)} \frac{d_s}{\log_2(k_s)} + \sum_{(s' \neq s'' \in CN)} \frac{d_{s'}}{\log_2(k_{s''})} \right) = \frac{1}{2} \cdot (CAA + R)$$

Considering that in very sparse networks the average node degree tends to two, $\widehat{k_s} \to 2$, with a very small standard deviation, the average node degree of the common neighbours

considering only the internal community links $\widehat{d_s}$ also tends to $\varepsilon < 2$ ($\widehat{d_s} \to \varepsilon < 2$) with a very small standard deviation. Thus we can say that, under the hypothesis of very sparse networks with conserved community structure, although $s' \neq s''$ the probability that $d_{s'} \approx d_{s''}$ and $k_{s'} \approx k_{s''}$ is very high. A consequence of this assumption is that the R term will be proportional to CAA, and:

$$\mathrm{CAR} \approx \frac{1}{2} \cdot (\mathrm{CAA} + k \cdot \mathrm{CAA}) = K \cdot \mathrm{CAA}$$

where K indicates a constant. The conclusion is that under the assumption of very sparse networks both CAR and CAA give the same ranking because, according to the approximations that we used, they end up being proportional. On the other hand, when we largely deviate from this assumption the two indices will start to produce different performances.

Figure S2 (see Section IV) confirms the above claimed and in addition shows that the performance of CAR, LCL, CAA, CRA and R practically overlap in eight different networks, especially when the hypothesis of very sparse network is forced in the simulations. A further confirmation is that the performances of CAR, LCL, CAA, CRA and R practically overlap also in the link prediction tests on the protein interactomes reported in the Fig. 3 of the article, and in Figure S3 (see Section V).

The Matlab code to compute the classical and CAR-based indices is provided at the following link: https://sites.google.com/site/carlovittoriocannistraci/5-datasets-and-matlab-code.

IV. REDISCOVERY OF REMOVED LINKS IN CONNECTOMES, SOCIAL AND ECOLOGICAL NETWOKS

This experiment was carried out in three connectomes and five social networks, two of which were ecological networks belonging to the class of food webs. The three connectomes are: Mouse Visual Cortex Neuro-synaptic Connectome, Macaque Cortical Connectome, and *C. elegans* Rostral Ganglia Neuro-synaptic Connectome. The social networks are: Dolphin Associations, College Football, Zachary's Karate Club, Food Web of Tuesday Lake, and Food Web of Grassland Species (see Tables S3 and S4 for details).

The aim of the experiment was to measure the prediction (rediscovery) power of each general-purpose neighbourhood-based technique (including CAR and the CAR-based variations) and each maximum likelihood approach, if links are removed from each network uniformly at random. The methodology followed was to destroy a specific amount of links in the networks (uniformly at random, as mentioned above) ranging from 10% to 90%, 1000 times per percentage. This procedure requires that 10% of the links in the network are chosen randomly 1000 times, and thus generate 1000 different topologies to which each prediction technique, along with a random predictor, are applied. At the second step, a number equal to the 10% of the original links is removed (i.e. 20% of links are removed in total at the second deletion step) from each of the 1000 networks obtained at the first step, and the prediction techniques are applied again. The process continues up to the point where 90% of the links are removed.

Every time a prediction technique is applied to a network topology, a ranked list of candidate interactions is generated and a number of links equivalent to the number of those removed is taken from the top of this list. The proportion of candidate interactions that matches the removed links is the performance measure for the prediction technique used. Notice that this measure is equivalent to Precision. Since this process was repeated 1000 times for each sparsification level, in practise mean precision and standard error were considered for each stage.

To characterise the deviation of each predictor from randomness, we transformed the indices' mean precisions at each sparsification level into decibels (dB); as a reference, we used the mean precision of the random predictor (a dB indicates a logarithm scaled ratio between a

quantity and reference level). We were thus able to measure the deviation of each predictor from randomness, as depicted in Equation S11.

$$Predictive\ Power_{Index} = 10\ \log_{10} \frac{\overline{Precision}_{Prediction\ Technique}}{\overline{Precision}_{Random\ Predictor}} \qquad \text{(S11)}$$

The Predictive Power, comprising its standard error at each percentage of links destroyed, generates a Prediction Power Curve, and the area under this curve (Area Under the Prediction Power Curve or AUPPC) summarises the power of each predictor.

As it is clear from Figure 2A in the main body of the article, link removal from the initial network topology impacts the community structure of network and all common-neighbours-based approaches have an important reduction in performance (Feng et al., 2012). To have a fair comparison between all different predictors, we decided to focus our analysis on their prediction power when only 10% of the network links have been destroyed (i.e. 90% of the original edges remain). While in the main text we show results for parameter-free and efficient techniques only, here we include HRG and SBM in Figure S1.
The above-mentioned supplementary figure shows that CAR and the CAR-based variations of the general-purpose indices are, in general, the best and most robust techniques. While the HRG approach also shows its robustness and adaptability to different sparse topologies, its computational time is prohibitive for large networks. Our Matlab implementation of CAR was able to perform all the 1000 realizations for 10% of the links removed in only 7 minutes in the largest of the 8 networks used (*C. elegans* Rostral Ganglia Neuro-synaptic Connectome). HRG's C++ implementation spent 81 hours to complete the simulation and, as stated in section IIIC, no parameter tuning was carried out. The poor performance of SBM may be due to this lack of parameter tuning, however this is also prohibitive since 814 hours were needed to complete our analyses for the *C. elegans* connectome with the technique's C implementation (all these experiments were carried out in a Dell Precision T7500 Workstation, 24 GB RAM and 2 Intel® Xeon® X5550 @ 2.67GHz quad-core processors).

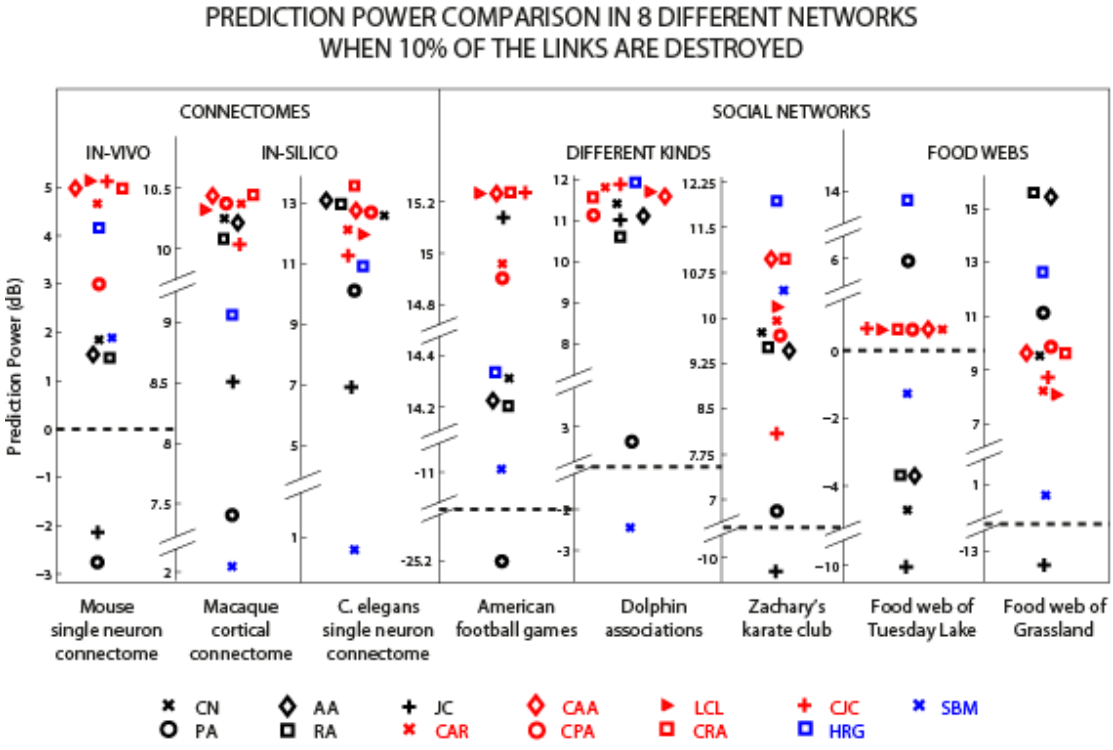

Figure S1. Prediction power comparison in 8 different networks when only 10% of the links are destroyed. Techniques in black represent neighbourhood-based, general-purpose approaches; techniques in red are all CAR-based indices; and techniques in blue correspond to maximum likelihood algorithms.

In Figure 2C we also include the LCP-correlation of each network and a p-value indicating that there is a statistically significant difference between the mean performance of the CAR-based index family and that of the general-purpose approaches. To compute these p-values a permutation test with 1000 resampling realizations was performed between the prediction power of the CAR-based indices and the prediction power of the general-purpose techniques.

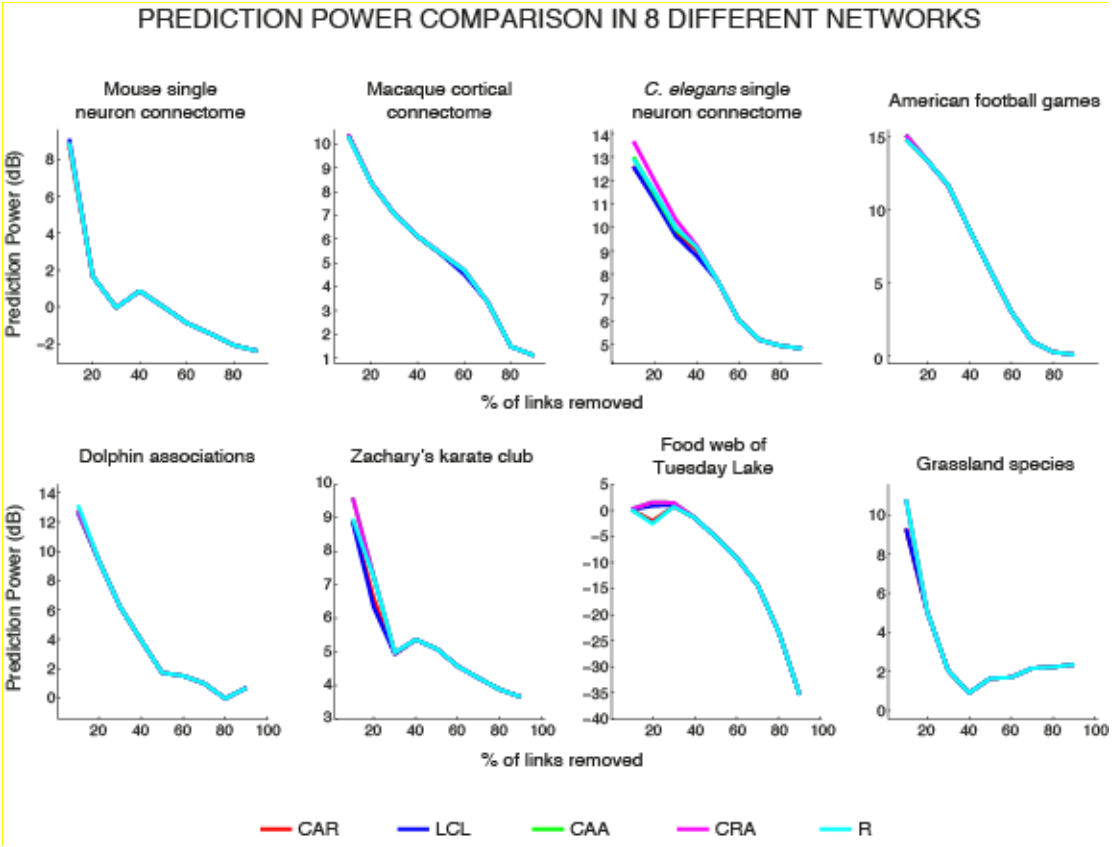

Figure S2. Prediction power comparison in 8 different networks when up to 90% of the links are destroyed. The CAR-based variations shown in these plots, as it is clear, offer a similar result. In particular, the performance-curves of CAR, CAA and R overlap, and their difference is practically indistinguishable.

V. PREDICTION OF LINKS IN PROTEIN-PROTEIN INTERACTION NETWORKS

Saito's research team established a schema for the validation of PPI reliability indices (Saito et al., 2002, 2003). This schema has been used in every work on PPI reliability and prediction assessment (Saito et al., 2002, 2003; J. Chen et al., 2005; J. Chen, Chua, et al., 2006; You et al., 2010). With specific regard to prediction, the procedure consists (i) in annotating every protein in the original network topology with their respective Gene Ontology (GO) terms from the three existent gene ontologies (Molecular Function, or MF, Biological Process, or BP, and Cellular Compartment, or CC) and (ii) looking for link similarities between all possible protein pairs in relation to Functional Homogeneity (two proteins are involved in the same function or process) or Localization Coherence (two proteins are located in the same cellular compartment). Subsequently a recursive procedure is applied to an increasing fraction of PPIs taken from the top of the list of candidate ranked interactions, and a curve that reports the proportion of pairs with Functional Homogeneity and Localization Coherence for each fraction is computed. It is important to note that this procedure measures a *precision* curve for the index: for each step, it evaluates the ratio between the number of interactions that are very likely to occur and the total number of interactions taken from top of the list.

Conventionally, a number of top-ranked candidate links equivalent to 10% of the links of the original network is used to compute the entire precision curve.

Our study followed the same approach to measure the predictive performance of techniques applied to PPINs (referred to as Network1, Network2, and Network3, see Table S3 for details). We take 10% of the links in the original network from the candidate interaction list generated by a prediction technique. As previously mentioned, we used GO, which is a reference ontology for discrimination between candidate interactions that are highly likely to be real (because GO confirms their Functional Homogeneity or Localization Coherence) and candidate interactions that are not likely to occur (because according to GO their proteins do not perform the same functions, are not involved in the same process or are not in the same cellular location).

To annotate the protein pairs and measure the similarity between GO terms, we used the R-GOSemSim package (G. Yu et al., 2010) and Wang GO Semantic Similarity (J. Wang et al., 2007). The GOSemSim function that we used takes the list of proteins that form the PPIN as its input, annotates them, computes the Wang GO semantic similarity between proteins, and outputs a matrix whose entries are the GO similarities for every possible PPI.

Most of the GO semantic similarity indices (Jiang and Conrath, 1997; Lin, 1998; Resnik, 1999) were originally developed for natural language taxonomies, and it is not known whether they are 100% suitable for GO. Wang's measure was created from the ground up, especially for the GO. Thus, Wang's index is more consistent with human perspective and with manual gene clustering into GO terms (J. Wang et al., 2007). This index ranges within 0 (no GO information available for one or both proteins, or no similarity in MF, BP or CC terms) and 1 (proteins share one or more identical GO terms). For these reasons Wang's measure is generally preferred for the evaluation of index performance in PPI prediction, and we too selected the measure for the current study.

In Wang et al. (J. Wang et al., 2007), it is mentioned that whenever Wang similarity is at the high end of the range, the proteins being analysed can be considered as analogous in their MF, BP, or CC. On the additional grounds of previous studies (J. Chen, Hsu, M. L. Lee, et al., 2006), we decided to use the same threshold, and only those pairs with Wang similarity above 0.5 were evaluated as good candidates. Wang's semantic similarity is a sort of confidence value, because the closer it is to 1, the greater the chance the interaction is real.

Finally, to obtain the precision curve that measures the performance of the indices, we removed protein pairs in chunks that were multiples of 100 interactions from the top of the candidate list to the previously mentioned 10% ceiling of the number of links in the original network (up to 1100 PPIs in Network1, 1200 PPIs in Network2 and 1300 PPIs in Network3 respectively). The proportion of interactions with relevant GO similarity (>= 0.5) was computed for each chunk, and each point generated a precision curve (see Figure S3). The area under this curve (Area Under the Precision curve or AUP) summarises the performance of each predictor.

To better depict the difference in performance between techniques, we report performance robustness (PF) for each technique. Of the three networks analysed, PF is the minimum area under the precision curve (AUP). This value puts each index at a disadvantage and measures its performance in the worst-case scenario. If the technique remains better than the others, it is considered stable and robust. PF values for each technique are listed in Table S1.

Times reported in the same table correspond to the maximum time the indices took to score the candidate interactions deriving from all the given networks. This is once again a measure of Time Robustness, because it depicts the worst-case scenario of the technique in terms of computational time. These times correspond to Matlab implementations running in a Dell Precision T7500 Workstation, with 24 GB of RAM and 2 Intel® Xeon® X5550 @ 2.67GHz quad-core processors.

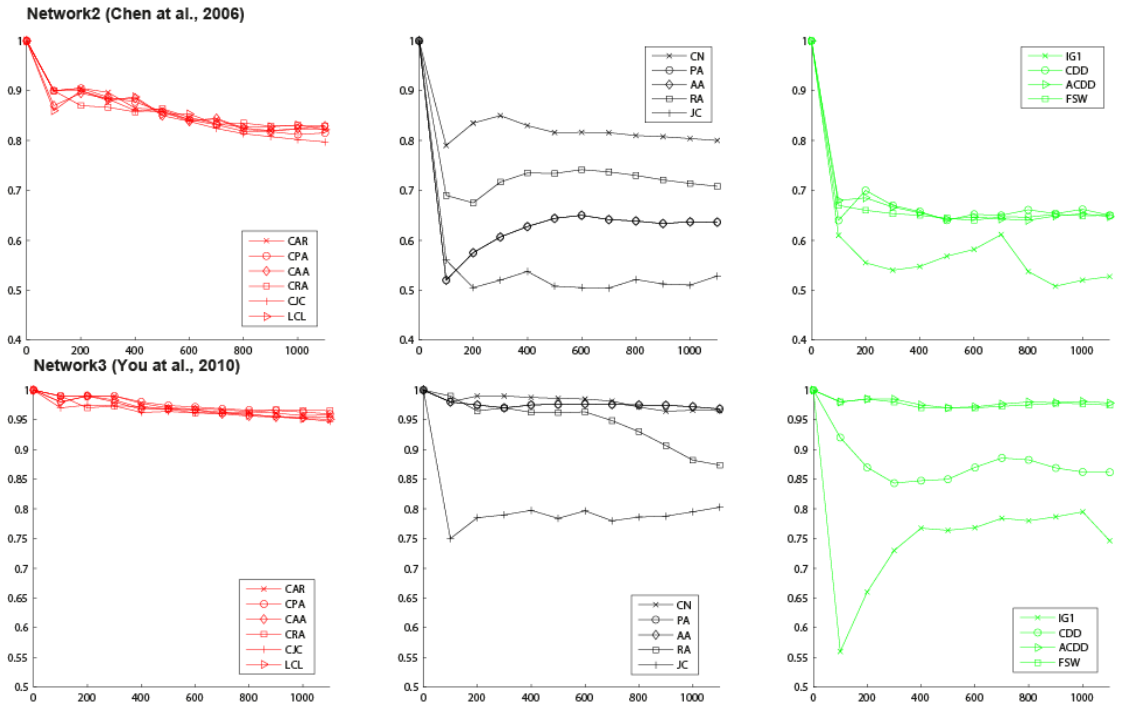

Figure S3. Precision curves for Yeast PPINs Network2 and Network3 (the precision curves for Network1 appear in the main article). The X-axis indicates how many protein pairs are taken from the top of the candidate list and the Y-axis indicates the precision of the prediction technique for that amount of PPIs. The AUP values for all the above-depicted techniques are listed in Table S1. For brevity, the performance of ISOMAP and ISOMAP+FSW is reported only as AUP in the Table S1.

Technique	*Area Under the Precision Curve (AUP)*				*Time cost (minutes)*
	Network1	*Network2*	*Network3*	*Performance Robustness*	
CAR	**0.86*** **	**0.86***	**0.97****	**0.86***	**17*****
CAA	**0.87****	**0.86***	**0.97****	**0.86***	20
LCL	**0.87****	**0.86***	**0.97****	**0.86***	**17*****
CPA	**0.86*** **	**0.86***	**0.98***	**0.86***	**17*****
CRA	**0.86*** **	**0.86***	**0.97****	**0.86***	19
CJC	**0.88***	**0.85****	**0.96*****	**0.85****	30
CN	0.82	**0.82*****	**0.98***	**0.82*****	**16****
RA	0.71	0.73	0.95	0.71	18
FSW	0.74	0.67	**0.98***	0.67	84
CDD	0.75	0.67	0.88	0.67	71
CDDP	0.75	0.67	**0.97****	0.67	80
PA	0.60	0.64	**0.98***	0.60	**1***
AA	0.60	0.64	**0.98***	0.60	18
IG1	0.55	0.57	0.75	0.55	29
JC	0.55	0.54	0.80	0.54	29
ISO	0.53	0.50	0.80	0.50	37
ISO+FSW	0.62	0.59	0.93	0.59	42

Table S1. Comparative table of performance robustness (minimum AUP) and time cost (in minutes, considering the maximum value between the time spent in each network) for the bio-inspired, general-purpose, CAR variations, and embedding techniques (coloured green, black, red, and blue respectively). Values in bold indicate the top three techniques for the characteristic measured in the respective column. The number of stars indicates the position occupied in the top three.

VI. IN-SILICO VALIDATION OF CANDIDATE LINKS IN STRING DATABASE

To verify the quality of the interactions proposed by the best techniques, namely CAR and FSW (from the general-purpose and the bio-inspired categories respectively), we took the top 100 candidate interactions from each technique's list and intersected them with the entire STRING Database (Szklarczyk et al., 2011) version 9.0, which was queried in February 2012. STRING is a compendium of PPIs found in the literature, experiments, coexpression, etc. Given a list of proteins, it finds the links between them and assigns them a confidence value based on the available evidence to the effect that the interactions exist. The number of interactions confirmed in STRING is a conservative estimate, because an interaction that does not appear in STRING might in fact be a non-detected interaction.

We queried the list of proteins involved in the top 100 candidate PPIs in STRING; Table S2 reports how many protein pairs per 100 were validated for each network, the average STRING confidence and its standard deviation, the average GO confidence (Wang's semantic similarity) and its standard deviation, and the Robustness of each one of these indicators (i.e. the minimum value that emerged from all three networks). Figure S4 presents the subnetworks formed by the top 100 PPIs and the interactions found in STRING are labelled in red.

Technique		*Network1*	*Network2*	*Network3*	*Robustness*
(G) CAR	Number of PPIs validated	**71**	**67**	64	**64**
	STRING confidence	0.75 ± 0.26	0.76 ± 0.25	**0.99 ± 0.06**	0.75 ± 0.26
	GO confidence	0.8 ± 0.17	0.79 ± 0.17	**0.99 ± 0.05**	0.79 ± 0.17
(B) FSW	Number of PPIs	55	38	**92**	38
	STRING confidence	**0.91 ± 0.2**	**0.92 ± 0.16**	0.97 ± 0.11	**0.91 ± 0.2**
	GO confidence	**0.93 ± 0.12**	**0.89 ± 0.16**	0.97 ± 0.06	**0.89 ± 0.16**

Table S2. Comparative table of the In-silico Validation performed for the best prediction techniques. CAR is a general-purpose technique and is marked with a (G), while FSW is bio-inspired and is marked with a (B). Values in bold indicate which predictor is better in each network for each indicator (Number of PPIs validated, STRING confidence and GO confidence). As is evident, the number of valid interactions detected by CAR is almost twice the number revealed by FSW. However, the STRING and GO confidence of the valid interactions for FSW is higher: this implies that CAR is able to predict additional interactions, which are related to novel and more heterogeneous biological knowledge that is still not strongly validated in STRING (confidence 0.75) but that is well-characterised in GO (confidence almost 0.80).

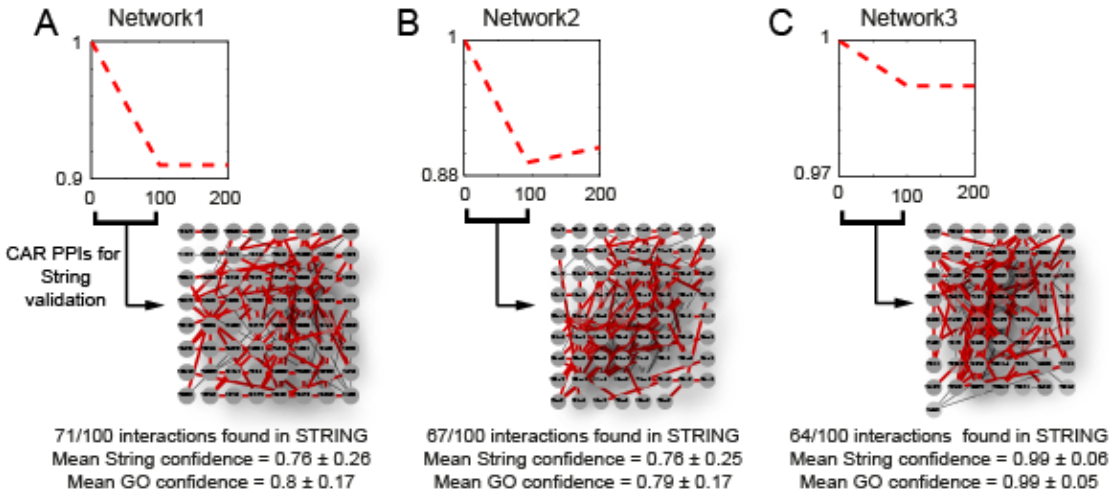

Figure S4. The dashed-red lines depict CAR's GO-precision-curve for the first 100 candidate PPIs. The top 100-ranked candidate interactions proposed by CAR were intersected with the STRING Database. This database provides each interacting pair with a confidence value. Below, the sub-networks formed by these candidate links are displayed, and the protein pairs found in STRING are shown in red.

VII. NETWORK SPARSIFICATION EXPERIMENTS ON PROTEIN INTERACTOMES

To investigate the robustness of the given prediction techniques against link deletion, and to study their behaviour in very sparse networks, we carried out a sparsification experiment on Network1, Network2 and Network3.

The experiment consisted in removing 10% of the network links uniformly at random, and the same procedure was repeated 10 times (over the same original network) to generate 10 different topologies. All prediction techniques were applied to each network, and their AUP was computed, after which another 10% of links was removed from each network (again randomly); this procedure was repeated to the point where the network lost connectivity.

VIII. LCP-DECOMPOSITION PLOT AND LCP-CORRELATION

The Local-community Paradigm Decomposition Plot (LCP-DP) is the new form of network visualisation that this study proposes. A point in the LCP-DP is a link from the network, and its coordinates are specified by the number of neighbours the interacting nodes have in common (CN) and the number of local community links between them (LCL).

The LCP-DP provides important information about the number and size of the existing communities within a network. The larger the number of common neighbours two interacting nodes share, the larger the number of LCLs they can have. Equation S12 shows the upper bound. Since this upper bound is a quadratic function of CN, we decided to take the square root of the number of LCL to linearize the representation shown in the LCP-DP.

$$Maximum\ number\ of\ LCL = \frac{CN(CN-1)}{2} = \frac{CN^2 - CN}{2} \qquad \text{(S12)}$$

In practice we found that a strong linear dependence between the number of CNs and the number of LCLs exists in many real networks. In order to quantify this linear dependency, we define the LCP-correlation (LCP-corr), which is the Pearson correlation coefficient of CN (considering only CN values greater than one, because a single CN cannot generate LCL) and LCL.

The formula to compute LCP-corr is given in Equation S13, where $cov(X,Y)$ is the covariance of the variables *X* and *Y,* and σ_X is the standard deviation of the variable *X*. The Matlab code to compute the LCP-corr for a given network is provided at the following link: https://sites.google.com/site/carlovittoriocannistraci/5-datasets-and-matlab-code.

$$LCP\text{-}corr = \frac{cov(CN, LCL)}{\sigma_{CN} \cdot \sigma_{LCL}}, \ when\ CN > 1 \tag{S13}$$

In conclusion, we show that most networks representing dynamic systems have an LCP-correlation greater than 0.80 (see Figures S5-S7 and Tables S3-S5).

IX. NETWORK DATA

This study used five types of real interaction networks, for a total of 45 networks, which represent differing systems and have their own meaning and characteristics: biological networks, social networks, atomic-level networks, power grids and road networks. Detailed information about each network is listed in Tables S3-S4 and their LCP-DPs appear in Figures S5-S7. The adjacency matrices of the networks for which we computed LCP-corr are provided at the following link: https://sites.google.com/site/carlovittoriocannistraci/5-datasets-and-matlab-code.

Almost all the categories mentioned above presented a high LCP-corr, ranging from 0.84 to 0.99. The Power Grid and the Karate Club network are borderlines with LCP-corr = 0.78 and LCP-corr = 0.75 respectively, as well as the Grassland Species Food Web with LCP-corr = 0.42. The road network values range from 0 to 0.16. Other networks analysed (most of them representing atomic interactions), present a clear LCP-corr of 0 (there are either no common neighbours between their interactors or there are no links between common neighbours if they exist).

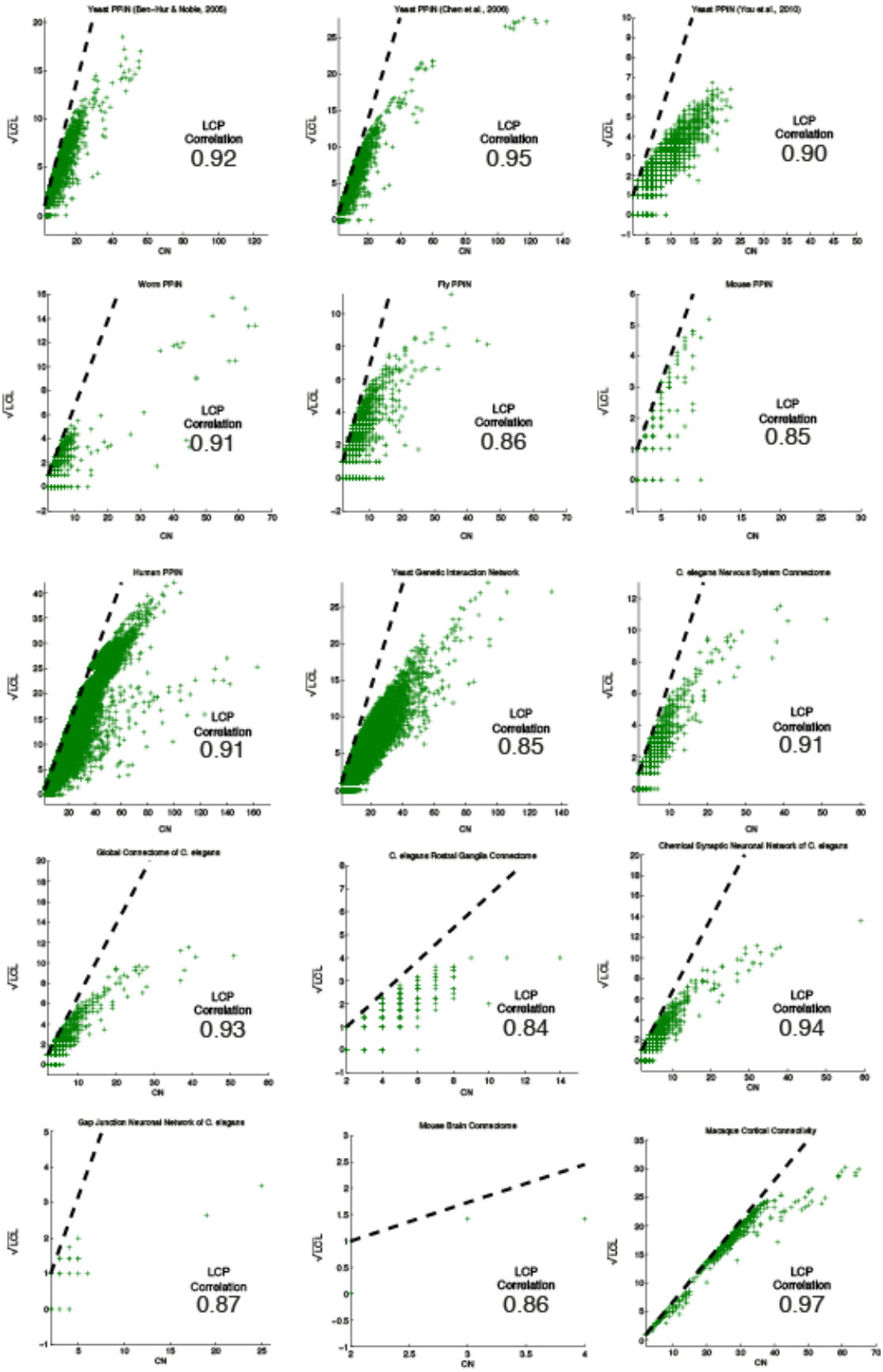

Figure S5. LCP-DP for networks of biological origin. The X-axis indicates the number of common neighbours and the Y-axis the square root of the number of links between them. Each point in the plot is an interaction from the network. The black, dashed line represents the LCL's upper bound.

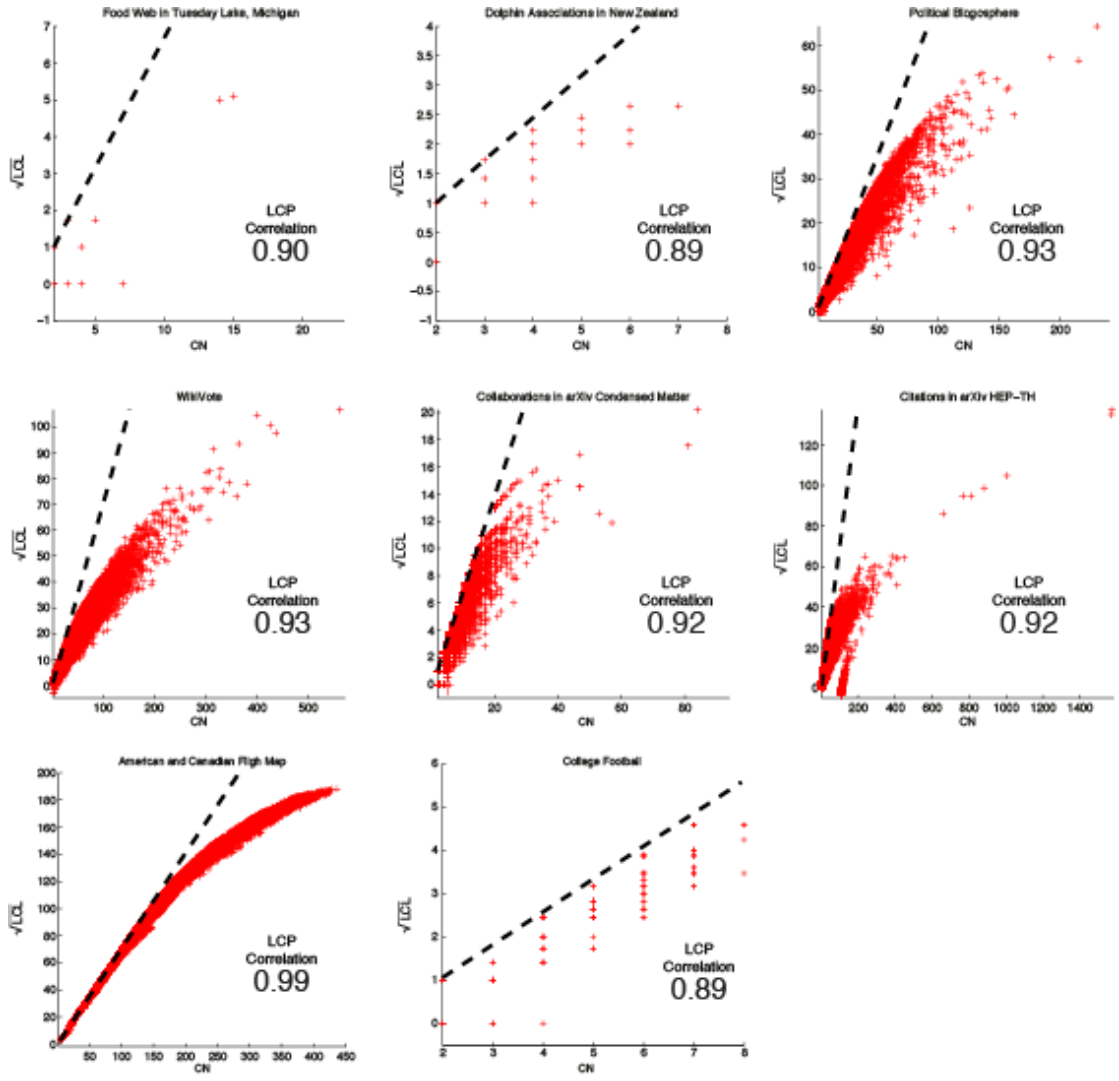

Figure S6. LCP-DP for social networks. The X-axis indicates the number of common neighbours and the Y-axis the square root of the number of links between them. Each point in the plot is an interaction from the network. The black, dashed line represents the LCL's upper bound.

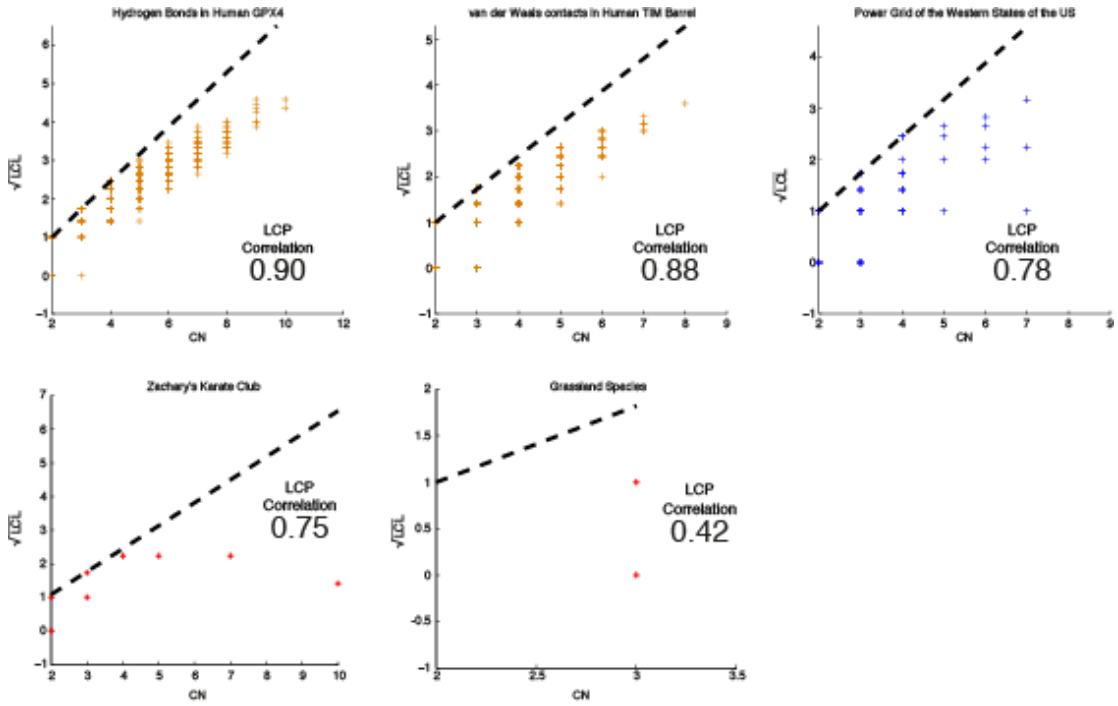

Figure S7. LCP-DP for atomic-level networks (top-left and top-centre) with significant LCP-corr. The Power Grid network (top-right), the Karate Club network (bottom-left), and the Grassland Species Food Web (bottom-centre) represent the borderline cases that we identified. The X-axis indicates the number of common neighbours and the Y-axis the square root of the number of links between them. Each point in the plot is an interaction from the network. The black, dashed line represents the LCL's upper bound.

Network	N	L	LCP-corr	Directionality	Description
Yeast Protein-protein Interactions (PPI) – Network1 (Ben-Hur and Noble, 2005)	4036	10411	0.92	Undirected	Interactions between proteins in yeast.
Yeast PPI – Network2 (J. Chen, Hsu, M.-L. Lee, et al., 2006)(Ben-Hur and Noble, 2005)	4385	12234	0.95	Undirected	Interactions between proteins in yeast.
Yeast PPI – Network3 (You et al., 2010)	3645	12934	0.90	Undirected	Interactions between proteins in yeast.
Worm PPI (Razick et al., 2008)	4743	18752	0.91	Undirected	Interactions between proteins in worm.
Fly PPI (Razick et al., 2008)	7809	71211	0.86	Undirected	Interactions between proteins in fly.
Mouse PPI (Razick et al., 2008)	2969	4033	0.85	Undirected	Interactions between proteins in mouse.
Human PPI (Razick et al., 2008)	11816	83422	0.91	Undirected	Interactions between proteins in human.
Yeast Genetic Interaction Network (Costanzo et al., 2010)	4319	74984	0.85	Undirected	Interactions between genes in yeast.
C. elegans Nervous System Connectome(Watts and Strogatz, 1998)	297	2345	0.91	Directed	Synaptic interactions between neurons in *C. elegans*.
C.elegans Global Connectome (Kaiser and C. C. Hilgetag, 2006)	277	1918	0.93	Directed	Global connectome of *C. elegans*.
C.elegans Rostral Ganglia Neuro-synaptic Connectome(Kaiser and C. C. Hilgetag, 2006)	131	687	0.84	Undirected	Rostral ganglia (anterior, dorsal, lateral, and ring) synaptic interactions in *C. elegans.*
C.elegans Chemical Synaptic (Varshney et al., 2011)	279	1961	0.94	Directed	Chemical synapse network of *C. elegans* neurons.
C.elegans Gap Junction (Varshney et al., 2011)	279	514	0.87	Directed	Gap junction network of *C.elegans* neurons.
Macaque Brain Connectome (Kötter, 2004)	94	1515	0.97	Directed	Macaque cortical connectivity network

					within one hemisphere.
Mouse Visual Cortex Neuro-synaptic Connectome(Bock et al., 2011)	18	42	0.86	Directed	Synaptic interactions between neurons in the primary visual cortex (layers 1, 2/3 and upper 4) in Mouse.
Food Web (Cohen et al., 2009)	51	241	0.90	Directed	Which taxon eats which in Tuesday Lake, Michigan
Dolphin Associations (Lusseau et al., 2003)	62	318	0.89	Undirected	Frequent associations between dolphins in New Zealand.
Political Blogosphere US Elections (Adamic and Glance, 2005)	1490	19025	0.93	Directed	Incoming and outgoing links and posts on blogs at the time of the 2004 US presidential election
WikiVote(Leskovec et al., 2010)	7115	103689	0.93	Directed	Who votes who to be a Wikipedia administrator
Scientific arXiv Collaborations (M. E. Newman, 2001)	16726	47594	0.92	Undirected	Scientific Collaborations on the Condensed Matter archive of arXiv from 1995 to 1999.
Citations (Leskovec et al., 2005)	27770	352807	0.92	Directed	Citation between papers on the High Energy Physics Theory archive of arXiv from 1993 to 2003.
American/Canadian Flight Map (Frey and Dueck, 2007)	456	37947	0.99	Undirected	Flights between pairs of American and Canadian cities.
College Football (Girvan and M. E. J. Newman, 2002)	115	613	0.89	Undirected	Network representation of the schedule of American football games.
Hydrogen Bonds between residues in a protein (A. J. M. Martin et al., 2011)	164	876	0.90	Undirected	Hydrogen bond network of human GPX4.
van der Waals Contacts between residues in a protein (A. J. M. Martin et al., 2011)	248	1979	0.88	Undirected	van der Waals contact network of human TIM barrel.

Table S3. The set of 25 LCP networks used in this work, along with their number of nodes N, the number of links between them L, the Pearson correlation between CN and LCL (LCP-corr), their directionality, and a brief description of what they represent. The background colour indicates the type of network: green for LCP networks of biological origin, red for LCP networks of social origin, orange for LCP atomic-level networks.

Power Grid (Watts and Strogatz, 1998)	4941	13188	0.78	Undirected	Power Grid of the Western States of the USA.
Zachary's Karate Club (Zachary, 1977)	34	78	0.75	Undirected	Friendship between members of a karate club in the US.
Grassland Species (Dawah et al., 1995)	75	113	0.42	Directed	Food web of grassland species.

Table S4. The set of 3 borderline networks used in this work, along with their number of nodes N, the number of links between them L, the Pearson correlation between CN and LCL (LCP-corr), their directionality, and a brief description of what they represent. The background colour indicates the type of network: blue for grids, and red for social networks.

San Joaquin Road Network (Brinkhoff, 2002)	18263	23797	0	Undirected	Road map of San Joaquin County, California
San Francisco Road Network (Brinkhoff, 2002)	174956	221802	0.16	Undirected	Road map of San Francisco, California
California Road Network (Feifei Li et al., 2005)	21048	21693	0	Undirected	Road map between cities in California, USA
Oldenburg Road Network (Brinkhoff, 2002)	6105	7035	0	Undirected	Road map of Oldenburg, Germany
North America Road Network (http://www.cs.fsu.edu/~lifeifei/SpatialDataset.htm)	175813	179102	0	Undirected	Road map of North America
California Road Network (Leskovec et al., 2009)	1965206	2766607	0.13	Undirected	Road map of California, USA where nodes are both cities and road intersections
Texas Road Network (Leskovec et al., 2009)	1379917	1921660	0.16	Undirected	Road map of Texas, USA
Pennsylvania Road Network (Leskovec et al., 2009)	1088092	1541898	0.15	Undirected	Road map of Pennsylvania, USA
German highway (Kaiser and C. Hilgetag, 2004)	1168	1243	0	Undirected	German highway system
Ice	-	-	0	Undirected	Bonds between atoms in ice.
Diamond	-	-	0	Undirected	Bonds between atoms in diamond.
Graphite	-	-	0	Undirected	Bonds between atoms in graphite.
Fullerene	-	-	0	Undirected	Bonds between atoms in fullerene.
DNA	-	-	0	Undirected	Bonds between bases in DNA.
β-sheet	-	-	0	Undirected	Bonds between residues in a β-sheet like protein structure.
α-helix	-	-	0	Undirected	Bonds between residues in a α-helix like protein structure.
Abiraterone Drug	-	-	0	Undirected	Bonds between atoms in the chemical structure of the Abiraterone drug.

Table S5. The set of 17 non-LCP networks used in this work, along with their number of nodes N, the number of links between them L, the Pearson correlation between CN and LCL (LCP-corr), their directionality, and a brief description of what they represent. The background colour indicates the type of network: grey for road maps, and orange for atomic networks.

X. LICENCE INFORMATION ON THE INDIVIDUAL ELEMENTS OF FIGURE 5

Here, we provide a table with a reference to each of the images/elements adopted in Figure 5, a link to the website where they were taken (or/and indication of our authorship) and a brief description of their terms of use.

IMAGE	SOURCE	TERMS OF USE
Protein folding tertiary network	Made by us using http://commons.wikimedia.org/wiki/File:Alpha_helix.png	GNU Free Documentation License 1.2 or later and Creative Commons Attribution-Share Alike 3.0 Unported
Dolphin associations	Made by us using http://www.clker.com/clipart-dolphin-silhouette.html	Public domain
Scientific collaborations	Made by us using http://commons.wikimedia.org/wiki/File:Silhouette.svg	Public domain
WikiVote	Made by us	Creative Commons Attribution-NonCommercial-ShareALike 3.0 Unported License

Political blogosphere	Made by us	Creative Commons Attribution-NonCommercial-ShareALike 3.0 Unported License
Human protein interaction network	Made by us	Creative Commons Attribution-NonCommercial-ShareALike 3.0 Unported License
Yeast genetic interaction network	Made by us	Creative Commons Attribution-NonCommercial-ShareALike 3.0 Unported License
Mouse connectome	Made by us	Creative Commons Attribution-NonCommercial-ShareALike 3.0 Unported License
C. elegans connectome	Made by us	Creative Commons Attribution-NonCommercial-ShareALike 3.0 Unported License
Plane	http://www.clker.com/clipart-15304.html	Public domain
Power grid	Based on modifications to http://en.wikipedia.org/wiki/File:Pylon_ds.jpg	Creative Commons Attribution-Share Alike 2.5 Generic
Grassland species food web	Made by us based on: 1. http://www.clker.com/search/grasshopper/1 2. http://www.clker.com/clipart-25717.html	1. Public domain 2. Public domain
Ice	http://commons.wikimedia.org/wiki/File:%E5%86%B0%E6%99%B6%E7%BB%93%E6%9E%84.png	Creative Commons Attribution-Share Alike 3.0 Unported and GNU Free Documentation License 1.2 or later.
Diamond and Graphite	http://commons.wikimedia.org/wiki/File:Diamond_and_graphite2.jpg	Creative Commons Attribution-Share Alike 3.0 Unported and GNU Free Documentation License 1.2 or later.
DNA	http://commons.wikimedia.org/wiki/File:ADN_static.png	Public domain
Beta-sheet	http://en.wikipedia.org/wiki/File:1gwe_antipar_betaSheet_both.png	Creative Commons Attribution-NonCommercial-ShareALike 3.0 Unported License
Alpha-helix	http://bioinsilico.blogspot.com/2008/11/secondary-structure-prediction_25.html	Public domain
Abiraterone	http://en.wikipedia.org/wiki/File:Abiraterone-3D-balls.png	Public domain
Road maps	Made by us using: 1. http://www.clker.com/clipart-16724.html 2. http://commons.wikimedia.org/wiki/File:Avenue_Road_map.png	1. Public domain 2. Creative Commons Attribution-ShareAlike 3.0 Unported Licence

REFERENCES

Adamic,L.A. and Adar,E. (2003) Friends and neighbors on the Web. *Social Networks*, **25**, 211–230.

Adamic,L.A. and Glance,N. (2005) The Political Blogosphere and the 2004 U . S . Election : Divided They Blog. *Proceedings of the 3rd international workshop on Link discovery*, 36–43.

Ben-Hur,A. and Noble,W.S. (2005) Kernel methods for predicting protein-protein interactions. *Bioinformatics*, **21 Suppl 1**, i38–46.

Bock,D.D. et al. (2011) Network anatomy and in vivo physiology of visual cortical neurons. *Nature*, **471**, 177–82.

Brinkhoff,T. (2002) A Framework for Generating Network-Based Moving Objects. *Geoinformatica*, **6**, 153–180.

Brun,C. et al. (2003) Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome biology*, **5**, R6.

Chen,J. et al. (2005) Discovering reliable protein interactions from high-throughput experimental data using network topology. *Artificial intelligence in medicine*, **35**, 37–47.

Chen,J., Hsu,W., Lee,M.L., et al. (2006) Increasing confidence of protein interactomes using network topological metrics. *Bioinformatics (Oxford, England)*, **22**, 1998–2004.

Chen,J., Chua,H.N., et al. (2006) Increasing confidence of protein-protein interactomes. *Genome informatics. International Conference on Genome Informatics*, **17**, 284–97.

Chen,J., Hsu,W., Lee,M.-L., et al. (2006) NeMoFinder : Dissecting genome-wide protein-protein interactions with meso-scale network motifs. *KDD '06 Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 106–115.

Chua,H.N. et al. (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, **22**, 1623–30.

Clauset,A. et al. (2008) Hierarchical structure and the prediction of missing links in networks. *Nature*, **453**, 98–101.

Cohen,J.E. et al. (2009) Food webs are more than the sum of their tritrophic parts. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 22335–40.

Costanzo,M. et al. (2010) The genetic landscape of a cell. *Science*, **327**, 425–31.

Dawah,H.A. et al. (1995) Structure of the Parasitoid Communities of Grass-Feeding Chalcid Wasps. *Journal of Animal Ecology*, **64**, 708–720.

Feng,X. et al. (2012) Link prediction in complex networks: a clustering perspective. *The European Physical Journal B*, **85**, 3.

Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–6.

Girvan,M. and Newman,M.E.J. (2002) Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7821–6.

Guimerà,R. and Sales-Pardo,M. (2009) Missing and spurious interactions and the. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 1–6.

Jaccard,P. (1912) The distribution of flora in the alpine zone. *The New Phytologist*, **11**, 37–50.

Jiang,J.J. and Conrath,D.W. (1997) Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings of 10th International Conference on Research In Computational Linguistics*.

Kaiser,M. and Hilgetag,C. (2004) Spatial growth of real-world networks. *Physical Review E*, **69**, 1–5.

Kaiser,M. and Hilgetag,C.C. (2006) Nonoptimal component placement, but short processing paths, due to long-distance projections in neural systems. *PLoS computational biology*, **2**, e95.

Kuchaiev,O. et al. (2009) Geometric de-noising of protein-protein interaction networks. *PLoS computational biology*, **5**, e1000454.

Kötter,R. (2004) Online retrieval, processing, and visualization of primate connectivity data from the CoCoMac database. *Neuroinformatics*, **2**, 127–44.

Leskovec,J. et al. (2009) Community Structure in Large Networks : Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. *Internet Mathematics*, **6**, 29–123.

Leskovec,J. et al. (2005) Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Leskovec,J. et al. (2010) Signed Networks in Social Media. *CHI*.

Li,Feifei et al. (2005) On Trip Planning Queries in Spatial Databases. *Proceedings of SSTD'05*, 273–290.

Liben-Nowell,D. and Kleinberg,J. (2007) The Link-Prediction Problem for Social Networks. *Journal of the American Society for Information Science*, **58**, 1019–1031.

Lin,D. (1998) An Information-Theoretic Definition of Similarity. *Proceedings of the 15th International Conference on Machine Learning*, 296–304.

Liu,G. et al. (2009) Complex discovery from weighted PPI networks. *Bioinformatics*, **25**, 1891–7.

Lusseau,D. et al. (2003) The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, **54**, 396–405.

Lü,L. and Zhou,T. (2011) Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, **390**, 1150–1170.

Martin,A.J.M. et al. (2011) RING: networking interacting residues, evolutionary information and energetics in protein structures. *Bioinformatics*, **27**, 2003–5.

Newman,M. (2001) Clustering and preferential attachment in growing networks. *Physical Review E*, **64**, 1–4.

Newman,M.E. (2001) The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 404–9.

Razick,S. et al. (2008) iRefIndex: a consolidated protein interaction database with provenance. *BMC bioinformatics*, **9**, 405.

Resnik,P. (1999) Semantic Similarity in a Taxonomy : An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, **11**, 95–130.

Saito,R. et al. (2003) Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics*, **19**, 756–763.

Saito,R. et al. (2002) Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic acids research*, **30**, 1163–8.

Szklarczyk,D. et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, **39**, D561–8.

Varshney,L.R. et al. (2011) Structural properties of the Caenorhabditis elegans neuronal network. *PLoS computational biology*, **7**, e1001066.

Wang,J. et al. (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274–81.

Watts,D.J. and Strogatz,S.H. (1998) Collective dynamics of “small-world” networks. *Nature*, **393**, 440–2.

You,Z.-H. et al. (2010) Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioinformatics*, **26**, 2744–51.

Yu,G. et al. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, **26**, 976–8.

Zachary,W.W. (1977) An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, **33**, 452–473.

Zhou,T. et al. (2009) Predicting missing links via local information. *The European Physical Journal B*, **71**, 623–630.