(2018) Page 1 of 29
doi:10.1093/imaiai/drn000

Assessment of model fit via network comparison methods based on subgraph counts

LUIS OSPINA-FORERO*,
Alan Turing Institute and
University of Oxford
*lospina@turing.ac.uk
CHARLOTTE M. DEANE
University of Oxford
deane@stats.ox.ac.uk

AND

GESINE REINERT
University of Oxford
reinert@stats.ox.ac.uk

[Received on 4 July 2018]

While the number of network comparison methods is increasing, benchmarking of these methods is still in its infancy. The lack of understanding of complex dependencies among network characteristics makes it difficult to fully understand the meaning of the different network comparison methodologies and the relations between them.
In this paper we use a Monte Carlo framework as a way to address three general questions about the network comparison methods based on subgraph counts: (1) Can the methods differentiate between networks generated from different network generation mechanisms? (2) Are the number of nodes, or average degree, confounding factors for the comparison of networks? (3) Do all methods reach the same conclusions?
We further use the Monte Carlo framework to test the fit of ER, Chung-Lu and a duplication divergence model to the protein-protein interaction networks of Yeast, Fly, Worm, Human, E. Coli, five herpes virus networks and five social networks. In contrast to previous claims in the literature, we show that the large protein-protein interaction networks are not well modelled by the Chung-Lu model according to any of our tested methods.
We find that network comparison statistics are not completely invariant to changes in the number of nodes and edges. Some methods focus on fine grain similarities, such as GCD, while other methods, such as Netdis, can capture the similarities of networks despite them having different numbers of nodes and edges.

Keywords:

Model fit, network comparison, subgraph counts.

1. Introduction

Networks appear across a large number of application fields such as linguistics, sociology, engineering, biology and politics (Newman, 2003). Due to the increasing popularity of network analysis, network generation and network comparison methods are becoming essential tools in network analysis (e.g.

Newman, 2003; Shao et al., 2013).

Different methodologies for network comparison have been proposed in recent years (Emmert-Streib et al., 2016); including comparing summary statistics of networks, such as average degree and clustering coefficient (e.g. Shao et al., 2013; Topirceanu et al., 2013; Berlingerio et al., 2013), using network alignment methodologies (e.g. Neyshabur et al., 2013; Hashemifar and Xu, 2014), subgraph counts (e.g. Ali et al., 2014; Pržulj, 2007; Yaveroglu et al., 2014), machine learning procedures (e.g. Aliakbary et al., 2015) and latent spaces (e.g. Asta and Shalizi, 2014). However, the results obtained from these methods may not always point towards the same conclusions, as they can follow different premises, and could be accounting for characteristics not explicitly considered in the initial methodology. Currently, it is not known how some of these methods perform against other more standard methods, under what conditions they perform best, or what their general capabilities are (e.g. Emmert-Streib et al., 2016). Hence, given a particular research question, it is not always clear what network comparison measure should be used. For example, what network comparison statistic is best suited to compare the core-periphery structure in different networks? Would the same statistic detect that two networks have similar community structure? Similar problems arise in assessment of model fit, where one of the interest lay in identifying networks that share the same generation mechanisms.

This paper addresses these questions. We use a statistical framework to understand what network comparison methods, based on subgraph counts, are best suited to the task of identifying similarities between networks that come from the same network generation mechanism, even when the networks have a different number of nodes or edges. We use the network comparison methods to assess if protein-protein interaction (PPI) networks can be thought of as realisations of well known random graph models. PPI networks have been used in many biological studies, including the discovery of disease risk pathways, the investigation of genes undergoing age expression changes, and the identification of single druggable targets (e.g. Sarajlić et al., 2013; Noh et al., 2013; Zoraghi and Reiner, 2013; Higueruelo et al., 2013; West et al., 2013). The pursuit of a network model that can describe the modular structure of protein-protein interaction networks has been a focus of much research, leading to several proposals of models, such as, the Chung-Lu model (Chung and Lu, 2002; Pržulj and Higham, 2006) and duplication divergence models (Vázquez et al., 2003; Gibson and Goldberg, 2011; Ispolatov et al., 2005) among others. Similarly, network models that can describe the structure of networks in other areas, such as social networks, are also of increasing interest (Traud et al., 2012). Here we test if network models used as backgrounds in community detection of social networks, such as the ER model or Chung-Lu model (Porter et al., 2009; Newman, 2006), can also be used to describe the appearance of subgraph counts in the Facebook networks of five USA universities.

Comparing networks through small connected subgraphs is a methodology of particular interest, as small overrepresented subgraphs are thought to be building blocks of complex networks (Milo et al., 2002). These small subgraphs have been shown to be important patterns in gene regulatory networks; and there is evidence that they may be preserved in biological networks (Shen-Orr et al., 2002; Wuchty et al., 2003; Alon, 2007; Pereira-Leal et al., 2007). In this study we focus on three main network comparison methods based on such connected subgraphs: "graphlet correlation distance" (GCD) (Yaveroglu et al., 2014), the "graphlet degree distribution agreement" (GDDA) (Pržulj, 2007) and Netdis (Ali et al., 2014). We also considered one method for network alignment, i.e. comparing networks based on how can one network be mapped into the other in such a way that individual interactions on the first network match interactions on the second network. We use a network alignment method as in many biological applications, alignments are a natural and popular way of comparing networks (e.g. Mamano and Hayes, 2017; Hashemifar and Xu, 2014; Saraph and Milenković, 2014; Singh et al., 2008; Patro and Kingsford, 2012). In contrast to GDDA, GCD or Netdis, network alignments provide a more intensive structural

comparison, as they take into account network isomorphism and aim to provide a one-to-one matching between the nodes of the compared networks. Among the network alignment methodologies we choose Netal (Neyshabur et al., 2013), because it is a reference method among network alignments methodologies (e.g. Sun et al., 2015; Crawford and Milenković, 2015; Hashemifar and Xu, 2014; Malod-Dognin and Pržulj, 2015).

Here, we use an easy to apply framework to evaluate if a network can be considered as coming from a given network model. To evaluate the fit of a network model, B, to data G, through a statistic S, there are two steps. (1) Finding the null distribution of the statistic S, i.e. if the data comes from B, what should the distribution of S look like? (2) Assessing how extreme the value S is when it is observed in a given network G. Our framework addresses each step through straightforward simulation of model-vs-model and data-vs-model comparisons, followed by a Monte Carlo test (a related procedure was proposed by Rito et al. (2010) for GDDA). This procedure is then used to assess the ability of the different network comparison methods, to compare networks in relation to their generation mechanisms.

We find that the proposed statistical framework allows accurate evaluation of the different network comparison methods, and can help to gain insight into differences between apparently similar network comparisons statistics. For example, by considering the type two error of the Monte Carlo test, we find that despite the fact that GCD, GDDA and Netdis are based on subgraph counts and all aim to measure similarity between networks they may capture network similarity at different ends of the spectrum. GCD tends to focus on fine-grained distinctions at the expense of missing large scale similarities, while GDDA, and Netdis in particular, focus on capturing the common global structure of networks, irrespective of their size and/or density. We also find that Netal, despite not being developed to compare networks with a large difference in the number of nodes and edges, achieves a good performance for two out of the three random graph models considered.

Given the crucial importance of a model for PPI networks, we assess the fit of the Chung-Lu model, previously claimed as a good null model for PPI networks (Hayes et al., 2013). By applying this framework to assess the fit of the Chung-Lu and duplication-divergence (DD) models to five large PPI networks and five small (virus) PPI networks we find that the suggestion of Hayes et al. (2013) that the Chung-Lu model is a good model for PPI networks does not hold for the large PPI networks. The Chung-Lu model is rejected by all four network comparison statistics and we conclude that the Chung-Lu model does not generate networks with a subgraph structure that is similar to any of the five large PPI networks.

This paper has two main results. Firstly it provides a better understanding of the practical differences between network comparison statistics in the context of model fit. Secondly, it finds that all four network comparison statistics considered reject the Chung-Lu model for all of the five large PPI networks.

2. Materials and Methods

2.1 *Network data*

We use the large PPI networks of Yeast, Human, Fly, Worm and E. coli, downloaded from BioGRID (Stark et al., 2006) in October 2015. The E. coli PPI network was obtained from Rajagopala et al. (2014). We also use the datasets analysed by Hayes et al. (2013), for which, according to Hayes et al. (2013), "STICKY, SF-GD and GEO-GD (in that order) are the best fitting models" (see Appendix D). This claim by Hayes et al. (2013) was based on the network comparison statistic GDDA (described in the following section). In addition we consider the smaller PPI networks of the viruses EBV, VZV, mCMV, HSV-1 and KSHV obtained from Fossum et al. (2009) which are also analysed by Hayes et al. (2013). To agree with the setup in Hayes et al. (2013) and Ali et al. (2014), all self-loops, multiple edges

and interspecies interactions are removed from the PPI networks, leaving simple undirected networks for the analysis. All degree 0 nodes are also removed from the analysed PPI networks. This choice is made to enable comparison with the results in (e.g. Rito et al., 2010; Hayes et al., 2013) where the same network modifications were carried out. For an analysis of the influence which these modifications may entail, particularly for configuration models, see for example Fosdick et al. (2017).

Table 1 describes these networks; the column *density* in the table is the number of edges divided by the possible number of edges and is given here to aid comparisons.

	Nodes	Edges	Density	Avg d.	Extracted from
Worm	3189	5556	0.00109	3.49	BioGRID ver 3.4.130
Fly	7958	36322	0.00115	9.13	BioGRID ver 3.4.130
Human	15590	182701	0.00150	23.44	BioGRID ver 3.4.130
Yeast	5862	79537	0.00463	27.14	BioGRID ver 3.4.130
E. coli	2002	3574	0.00178	3.57	(Rajagopala et al., 2014)
mCMV	111	393	0.06437	7.08	(Fossum et al., 2009)
KSHV	50	115	0.09388	4.60	(Fossum et al., 2009)
VZV	57	160	0.10025	5.61	(Fossum et al., 2009)
HSV-1	47	100	0.09251	4.26	(Fossum et al., 2009)
EBV	60	208	0.11751	6.93	(Fossum et al., 2009)

Table 1. Number of nodes, edges, density, average degree (Avg d.) and source of recent Yeast, Human, Fly, Worm PPI networks (downloaded in October 2015); and previously studied mCMV, KSHV, VZV, HSV-1, EBV virus PPI networks. The dashed line separates the large PPI networks from the small virus networks.

Five social networks are also used in this work. These networks correspond to the Facebook networks of Caltech, Reed, Haverford, Simmons and Swarthmore universities. The nodes in each network represent users of Facebook who were members of the same university at September 2005. The undirected edges represent reciprocated friendship between the users. These five Facebook networks are the smallest networks of a larger set of 100 universities (Traud et al., 2012; Onnela et al., 2012). Table 2 describes these networks.

All PPI and Facebook networks are sparse, in the sense that the average node degree is much smaller than the number of nodes.

	Nodes	Edges	Density	Avg d.
Caltech	769	16656	0.05640	43.32
Reed	962	18812	0.04070	39.11
Haverford	1446	59589	0.05704	82.42
Simmons	1518	32988	0.02865	43.46
Swarthmore	1659	61050	0.04439	73.60

Table 2. Number of nodes, edges, density, average degree and source of Facebook social networks of five universities previously studied by Traud et al. (2012).

2.2 *Network Comparison methods*

In this paper we compare four network comparison statistics:

2.2.1 *Graphlet correlation distance.*

The graphlet correlation distance (GCD) (Yaveroglu et al., 2014) is a network comparison statistic that uses a generalisation of the degree distribution to the degree distribution of all 30 automorphism orbits of connected subgraphs (graphlets) on two to five nodes. An *auto-*

morphism of a graph $G = (V, E)$ is a bijection $g : V \to V$ such that $(i, j) \in E$ if and only if $(g(i), g(j)) \in E$. An *automorphism orbit* of a node $i \in V$ is the set of nodes $\{x \in V | g(x) = i\}$, where g is any automorphism of G. As an example, Figure 1 shows the 15 automorphism orbits that appear on subgraphs on two to four nodes.

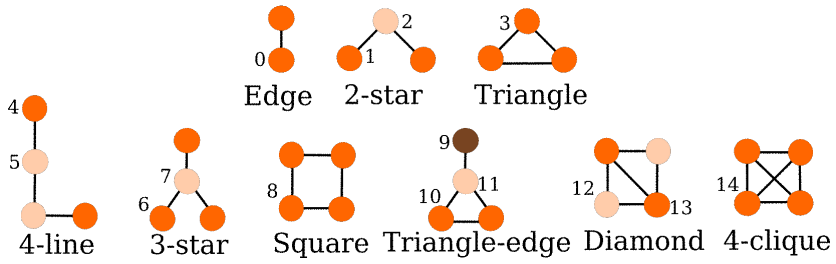

FIG. 1. Connected subgraphs in two to four nodes an their automorphism orbits. The different shades of colour within each subgraph show their different possible automorphism orbits.

GCD considers a set of 11 automorphism orbits, here-onwards called orbits, $O = \{0, 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. The first step in the construction of GCD consist of obtaining an orbit count matrix K for each individual network with number of rows equal to the number of nodes, and with 11 columns (orbits). Each orbit count matrix K, is composed of the number of times each node in the network appears at orbit $l, l \in O$, in the network. Hence, for node i and orbit l, with corresponding column $K^{(l)}$, the entry $K_i^{(l)}$ is the total number of times node i appears at orbit l among all the different induced subgraphs in the network that are isomorphic with the subgraph where orbit l is located (see Figure 1 and Appendix A for more details). As a second step, for each orbit count matrix, a correlation matrix C is formed by taking all pairwise Spearman's correlations between the columns of K. Then, for networks G_1 and G_2 with corresponding correlation matrices C^{G_1} and C^{G_2}, GCD is defined as the Euclidean distance between these two correlation matrices:

$$GCD\text{-}11 = \sqrt{\sum_{i=1}^{11}\sum_{j>i}^{11}(C_{i,j}^{G_1} - C_{i,j}^{G_2})^2}.$$

Yaveroglu et al. (2014) also proposed other graphlet correlation statistics by using different sets of orbits, for example GCD-73 where all orbits on 2-5 node subgraphs are used and GCD-15 where all subgraphs up to four nodes are used. GCD-11 uses 11 orbits of the 15 possible orbits on subgraphs of 2-4 nodes and is reported as the best choice among their alternatives. The authors mentioned that small values of GCD suggest higher similarity.

2.2.2 *Graphlet degree distribution agreement.* The graphlet degree distribution agreement (GDDA) proposed by Pržulj (2007) uses the degree distribution of all 73 orbits on connected subgraphs on two to five nodes. The distributions are represented by $d_G^j(.)$, $j = 0, 1, ..., 72$, where the frequency $d_G^j(k)$ is given by the number of nodes in the network whose j orbit count is exactly equal to k (see Appendix A for an example). The construction of GDDA is as follows: For graph G_1 on n nodes compute for every orbit j

$$N_{G_1}^j(k) = \frac{d_{G_1}^j(k)/k}{\sum_{m=1}^{n} d_{G_1}^j(m)/m},$$

and similarly for any other graph. Then, to compare graphs G_1 and G_2 take

$$D^j = \frac{1}{\sqrt{2}}\Big(\sum_{k=1}^{n}[N^j_{G_1}(k) - N^j_{G_2}(k)]^2\Big)^{1/2};$$

D^j is supposed to reflect 'how different' the orbit degree distributions are and $1 - D^j$ 'how close' they are. Finally, the GDDA is obtained as the arithmetic mean of the values $1 - D^j$. Alternatively, the geometric mean can be used, but the arithmetic mean is the standard method. Large GDDA values are thought to imply 'similarity' between the compared graphs (Hayes et al., 2013). Note that in the definition of the GDDA degree zero nodes are not taken into account as $k > 0$, in $D^j(k)$ and $N^j_{G_1}(k)$.

2.2.3 *Netal.* Netal (Neyshabur et al., 2013) is an algorithm that creates a "mapping" (alignment) between the node sets of two networks. Given two networks G_1, G_2 such that $|V(G_1)| \leqslant |V(G_2)|$, an alignment f is an injective function $f : V_1 \to V_2$, that aims to maximise the number of conserved interactions. A conserved interaction is an edge (l, l') present in G_1 for which $(f(l), f(l')))$ is an edge present in G_2. Netal uses two matrices, S and I with rows representing nodes of G_1 and columns representing nodes of G_2, to obtain an alignment score matrix, A. Here S is called a similarity score matrix and I is called an interaction score matrix. In the absence of exogenous information on the nodes, the similarity score matrix is composed only of topological scores which assess similarity between two local node neighbourhoods. Under this scenario the alignment score matrix is the weighted sum

$$A(i, j) = \lambda S(i, j) + (1 - \lambda)I(i, j),$$

where, λ is a tuning parameter between zero and one. We use $\lambda = 1/|V(G_2)|$ as suggested by Neyshabur et al. (2013).

To obtain S, initially a matrix of equal dimensions $S^0 = \mathbf{1}$ is considered prior to the construction of A, and then updated iteratively by scoring the similarity of the neighbourhood of node $i \in G_1$, $N(i)$ (the set of all nodes which are connected to i by an edge), to the neighbourhood of node $j \in G_2$, $N(j)$. By default Netal uses two iterations, thus leading to

$$S(i, j) = \max_h \frac{\sum_{u \in N(i)} \min\{1, |N(u)|/|N(h(u))|\}}{\max\{|N(i)|, |N(j)|\}},$$

where h is a one-to-one mapping between $N(i)$ and $N(j)$.

The interaction score matrix I, which is updated after each alignment step along with matrix A, and which indicates an approximation to the expected number of conserved interactions in the local neighbourhood of i and j, were nodes i and j to be aligned, is constructed as:

$$I(i, j) = \frac{min\left\{(\sum_{i' \in N(i)} 1/|N(i')|) - C_1(i), (\sum_{j' \in N(j)} 1/|N(j')|) - C_2(j)\right\}}{max_{k \in V_1 \cup V_2}\{|N(k)|\}},$$

where $C_1(i)$ and $C_2(j)$ are the number of conserved interactions, accounted by the nodes aligned up to the current alignment step, that are incident to nodes i and j, respectively.

Finally, the Netal algorithm works by selecting at each iteration a pair (i, j) for which $A(i, j)$ is largest at that iteration. $A(i, j)$ and $I(i, j)$ are updated after each pair (i, j) is selected (for more details see (Neyshabur et al., 2013)).

Now, for any two given networks such that $|V(G_1)| \leqslant |V(G_2)|$, after the alignment is created, we use the number of conserved edges in the first network as a network comparison statistic,

$$EC = \frac{|\{(u,v) \in E(G_1) : (f(u), f(v)) \in E(G_2)\}|}{|E(G_1)|}.$$

Larger values of EC would suggest similarity between the two networks. EC is known as the *edge correctness*, a standard measure used to judge network alignments (e.g. Neyshabur et al., 2013; Saraph and Milenković, 2014). However, this measure has to be interpreted with great care; for example, if G_2 is a complete graph, then $EC = 1$ for any G_1.

2.2.4 *Netdis.* Netdis (Ali et al., 2014) also uses subgraph counts as building blocks for a network comparison statistic. However, in contrast to GCD and GDDA, which start by obtaining the orbit counts for each node in the network (see Figure 1 and Appendix A); Netdis assigns to each node the total number of subgraphs that appear within that node's 2-step ego-network. In addition, Netdis also differs from GCD and GDDA in that it takes into account a background expectation for the subgraph counts, which aims to give more reliable comparisons between networks of different sizes and/or densities.

Netdis counts small subgraphs (see Figure 1) w on k nodes for all 2-step ego-networks, $k = 3,4,5$. These counts are centred by subtracting the expected number of counts E_w. The centred counts between the networks are compared to form the Netdis statistic. In detail, Netdis is constructed as follows:

Let $N_{w,i}(G)$ be the number of induced occurrences of small graphs w in the 2-step ego network of vertex i. Now, bin all 2-step ego-networks of network G according to their network density. Let $E_w(G,d)$ be the expected number of occurrences of w in an ego-network whose density falls in density bin d. For a given network G compute the centred subgraph counts as

$$S_w(G) = \sum_i \Big(N_{w,i}(G) - E_w(G, \rho(i)) \Big),$$

where i is a node in G and $\rho(i)$ the density bin of the 2-step ego-network of node i.

Now, to compare networks G_1 and G_2, set

$$netD_2^S(k) = \tfrac{1}{\sqrt{M(k)}} \sum_{w \in A(k)} \left(\tfrac{S_w(G_1) S_w(G_2)}{\sqrt{S_w(G_1)^2 + S_w(G_2)^2}} \right), \quad k = 3,4,5,$$

where $A(k)$ is the set of connected subgraphs of size k, and where $M(k)$ is a normalising constant so that $netD_2^S(k) \in [-1,1]$. $M(k)$ is equal to

$$M(k) = \sum_{w \in A(k)} \left(\tfrac{S_w(G_1)^2}{\sqrt{S_w(G_1)^2 + S_w(G_2)^2}} \right) \sum_{w \in A(k)} \left(\tfrac{S_w(G_2)^2}{\sqrt{S_w(G_1)^2 + S_w(G_2)^2}} \right).$$

The corresponding Netdis statistic is defined as

$$Netdis(k) = netd_2^S(k) = \tfrac{1}{2}(1 - netD_2^S(k)) \in [0,1].$$

Small values of Netdis suggest higher 'similarity' between the networks (Ali et al., 2014). By default Netdis uses subgraphs on $k = 4$ nodes.

Due to the lack of joint probabilistic models for subgraph counts, Ali et al. (2014) proposed to use the subgraph counts from a (fixed) gold-standard network, as an approximation to the expected

subgraph counts. In this paper, instead of considering an approximation based on an observed gold-standard network whose selection could not be straight forward, we use, following (Picard et al., 2008), a geometric Poisson approximation (also called Polya-Aeppli approximation). Hence, following (Picard et al., 2008), we compute E_w based on a geometric Poisson (GP) approximation for the distribution of the number of occurrences of subgraph w. Here, we assume $N_{w,i} \sim GP(\lambda_k^{\rho(i)}, \theta_w^{\rho(i)})$, where $\lambda_k^{\rho(i)}$ is the Poisson parameter indexed by the size of subgraph w and the density bin $\rho(i)$; and where $\theta_w^{\rho(i)}$ is the geometric parameter indexed by subgraph w and density bin $\rho(i)$. Then, we take $E_w(G,\rho(i))$ as the mean of the GP approximation, i.e. $\lambda_k^{\rho(i)}/\theta_w^{\rho(i)}$.

As $\lambda_k^{\rho(i)}$ and $\theta_w^{\rho(i)}$ are not known, they are estimated as follows: Let $x_{w,d}^j$ be the number of subgraphs w on the 2-step ego-network j of density bin d, and let

$$\bar{X}_{w,d} = \frac{1}{q}\sum_{j=1}^{q} x_{w,d}^j, \qquad V_{w,d}^2 = \frac{1}{q-1}\sum_{j=1}^{q}(x_{w,d}^j - \bar{X}_{w,d})^2,$$

where q is the number of ego-networks in density bin d. Then,

$$\hat{\lambda}_k^d = \frac{1}{l}\sum_{h\in A(k)} \frac{2(\bar{X}_{h,d})^2}{V_{h,d}^2 + \bar{X}_{h,d}}, \qquad \hat{\theta}_w^d = \frac{2\bar{X}_{w,d}}{V_{w,d}^2 + \bar{X}_{w,d}},$$

where l is the number of connected subgraphs of size k, for example, $l = 6$ for $k = 4$. These estimators are based on the moment estimators of a GP random variable and the proposal made by Picard et al. (2008), where the total count of each individual subgraph could be thought as the sum of the total subgraph counts over multiple "clumps" of edges that appear across the network.

It is important to notice that currently, contrary to the *EC* used for the network alignment method Netal, there is no clear understanding about the meaning of the values obtained from the network comparison statistics GCD, GDDA and Netdis. Whilst an *EC* value of 0.90 ($1 - EC = 0.10$), means that 10% of the edges on the first network are not mapped to edges in the second network, for GDDA, Netdis or GCD there is no clear meaning about what a value of, for example, 0.9 means in relation to the networks being compared.

2.3 *Random graph models*

We consider three random graph models:

ER MODEL: This $ER(n_v,\rho)$ model, proposed by Gilbert (1959) is a random graph on n_v nodes where single undirected edges are present independently at random, each with probability ρ. In this work all ER networks are generated without self-loops.

THE CHUNG-LU MODEL: The Chung and Lu (2002) model, also known as the Sticky model (Pržulj and Higham, 2006) is as follows: Given a sequence $\{d_1, d_2, ..., d_{n_v}\}$ such that $max_i d_i^2 < \sum_k d_k$ and $d_i > 0$ for all i, the model assigns a weight (θ_i) to each node, where

$$\theta_i = \frac{d_i}{\sqrt{\sum_{l=1}^{n_v} d_l}}.$$

Then, an edge is established between any two nodes with probability equal to the product of their weights,

$$P((i,j) \in E) = \theta_i \theta_j, \qquad i \neq j.$$

By construction, the model assumes a known sequence $\{d_1, d_2, ..., d_{n_v}\}$, which is the 'only' parameter of this model. In this work all Chung-Lu networks are generated without self-loops, i.e. $P((i,i) \in E) = 0$. We note that the probability $P((i,j) \in E) = \theta_i \theta_j$ is the same as the null model proposed by Newman (2006) for community detection based on modularity optimisation (Newman and Girvan, 2004).

A DUPLICATION DIVERGENCE MODEL: Duplication divergence (DD) models were one of the first random graph models to consider the PPI network as resulting from a process of 'evolution'. One of the first duplication divergence models was proposed by Vázquez et al. (2003). This model addresses the evolutionary process of PPI networks by two biologically related steps: 'duplication' and 'divergence' of proteins (nodes). The model is given as follows:

- Initialise the network with two connected nodes and then follow the following duplication and divergence steps, alternating between them until the desired number of nodes is reached.
- Duplication: A node v_i is randomly selected and a new node v_j is created with all the edges as the node previously selected (v_i). Then an edge between nodes v_i and v_j is created with probability p.
- Divergence: For all pair of edges $\{(v_i, v_k); (v_j, v_k)\}$ from the duplication step; one of the two edges is selected uniformly at random and then deleted with probability q.

2.4 *Monte-Carlo test*

Consider testing the null hypothesis "H_0: Network G_0 is a realisation the fully specified model B", based on a network comparison statistic S; against the alternative hypothesis "H_1: Network G_0 is not a realisation of the fully specified model B". We use a Monte Carlo test which compares data-vs-model and model-vs-model to assess such hypothesis, relative to S; see also (Rito et al., 2010) for a precursor. For simplicity assume that the model is rejected when S is large. The test is given by the following steps:

1. Generate M random graphs from the given model. For each of them generate another N random graphs and obtain the average of the comparison statistics between each of the M random graphs and the respectively generated N random graphs. This leads to a sample of M averages $\bar{S}_1, \bar{S}_2, ..., \bar{S}_M$.
2. Generate N random graphs from the given model, and for each of these random graphs calculate S comparing the random graph to the data. Take the average of this sample ($\bar{S}_0$).
3. If $\bar{S}_0$ is the k^{th} value on the ordered sample (with ties broken randomly) $\bar{S}_{(0)} \geqslant \bar{S}_{(1)} \geqslant ... \geqslant \bar{S}_{(M)}$, then the p-value of the test is $\frac{k}{M+1}$. We reject the null hypothesis when the p-value is small (typically p-value $\leqslant 0.05$).

The smallest p-value that this test report is $\frac{1}{M+1}$. Hence, p-values smaller than $\frac{1}{M+1}$ are overestimated by the test.

Such Monte Carlo tests are used in this paper to assess model fit. Consider a network G_0 generated from a model A, with the aim to test the null hypothesis "H_0 : G_0 is a realisation of the fully specified

model B" against the alternative hypothesis "H_1: Network G_0 is a realisation of the fully specified model $A, A \neq B$", with a significance level of 5%. Consider also the quantity $P(H_0 \text{ is not rejected} | H_1)$, known as the type II error of a hypothesis test.

The Monte Carlo test provides an intuitive and rigorous framework to compare the behaviour of different network comparison statistics under deviations from the null hypothesis (such as changes in the number of nodes, edge density or generation mechanism) via type II error. Here we estimate the type II error by the relative frequency of tests for which the null hypothesis is not rejected, i.e. the percentage of networks G_0 found not to be different from networks generated from a fully specified null model (B); ($A \neq B$).

Using this framework, we perform right-tailed tests with significance level of 0.05 for the network comparison statistics GCD, $1 - GDDA$, $1 - EC$ (Netal), and Netdis. It is crucial to note that for all of these test statistics, large values indicate dissimilarity. Hence it is appropriate to carry out a one-sided test only. This is in contrast to the more common practice in network analysis to carry out two-sided tests; for a recent example see Payrato Borras et al. (2017) where the fit of the Chung-Lu model to ecological networks is examined. Following Hayes et al. (2013) and Rito et al. (2010), we take $M = 99$ and $N = 30$ for ease of comparison of the results. In practice, $M = 999$ would be recommended for more precise results.

3. Results and discussion

3.1 *Model fit and assessment of network comparison statistics*

Currently the network comparison statistics based on subgraph counts *GDDA*, *GCD* and *Netdis* state that they capture structural similarities/dissimilarities between networks. However, it is not known what type of feature they best capture (e.g. edge density, density of cliques or cycles, number of connected components, etc.), nor is it clear how these methods compare to each other (e.g. Emmert-Streib et al., 2016). Thus, in order to start understanding better what features these network comparison methods capture, we consider the following general questions: (1) Can a network comparison statistic differentiate between networks generated from different network generation mechanisms? (2) Are the number of nodes, or average degree, confounding factors for the comparison of networks? (3) Do all four network comparison statistics agree in their conclusions? We address these questions based on two simulation scenarios:

(a) *Varying average degree*: ER, Chung-Lu and DD networks are generated with 1000 nodes and with approximate average degrees, (d), of 20, 15 and 11.

(b) *Varying number of nodes*: ER, Chung-Lu and DD networks are generated with 1500, 1000 and 500 nodes and with an approximate average degree of 15.

For both of these simulation scenarios, the degree sequences used in the Chung-Lu model are fixed from single realisations of a DD model in such a way that the average degree of the degree sequence is equal to the average degree in the respective scenarios, e.g. (20, 15 and 11). A detailed description of the parameterisation used for all models can be found in Appendix B.

For each of the previous simulation scenarios, Figure 2 shows 9×9 different test cases of null hypotheses of the type "H_0 : Network G_0 is a realisation of model B", for each of the network comparison statistics. Note that there are 9 possible cases for data G_0: It can be drawn from one of three different models, and with one of three possible average degrees for Scenario (a); (or for Scenario (b), three possible network sizes). Model B can also represent any of the three different models and with any of

the three average degrees for Scenario (a) (or network sizes for Scenario (b)). This leads to a total of 9×9 possible test cases for each of the four network comparison statistics. Figure 2 arranges these cases in a grid with 36×9 blocks for each simulation scenario.

For each block, 100 hypothesis tests with significance level 5% are conducted. Then, each block in Figure 2 is coloured according to the percentage of tests where the null hypothesis is not rejected. See Appendix Tables A.7 and A.8 for the exact percentage of non-rejection.

In Figure 2, we group the test cases for each network comparison statistic into larger blocks of 3×3 each. The purpose of these larger blocks is to focus on comparisons of the type "Data A" vs. "Model B" regardless of the average degree or network size. The columns of the figure represent models and the rows represent actual network realisations from those models (Data). Hence, the arrays displayed in Figure 2 are not symmetric, as a test of "Data A" vs. "Model B" is not the same as a test of "Data B" vs. "Model A".

The outcome of an ideal statistic that only captures the type of network generation process is portrayed in Figure 3. In the ideal outcome, the expected result of comparisons of "Data A" vs. "Model A", irrespective of network size and edge density, is 95%, as the significance level of the test is 5% (shown in red), while the expected result of comparisons of "Data A" vs. "Model B" is: 0%, (shown in blue), i.e. for $A \neq B$ the null hypothesis is always rejected.

In Figure 2 (and Appendix Tables A.7 and A.8), the cases where a network drawn from a fully specified model A is tested as a realisation of such fully specified model A, behave as expected from a hypothesis test with a confidence level of 5%; i.e. the number of non rejected tests mostly fall within the expected number of non-rejected tests (95) plus/minus two standard deviations (2×2.18).

In Scenario (a) of Figure 2, it can firstly be noted that Netal, the network alignment methodology, despite not being developed for comparing networks with large differences in the number of edges, performs the best for the task of identifying networks that come from the same model irrespective of having different average degrees. However, Netal seems to struggle more than the subgraph count methods at differentiating networks that come from different models, particularly differentiating DD networks from ER or Chung-Lu networks. GCD, on the contrary, is able to differentiate networks coming from different models better than Netal, but GCD is the method that struggles the most at detecting similarities between networks of different average degrees, coming from the same model. GDDA displays a performance in between that of Netal and GCD, as it identifies similarities between networks coming from the Chung-Lu and DD models, but fails at identifying similarities between the ER networks, or detecting differences between Chung-Lu and DD networks. Finally, Netdis shows a more balanced performance, as it differentiates well most cases where the data does not come from a particular model. It is also able to identify the similarities between networks that come from the same model irrespective of the average degrees considered.

Results for Scenario (b) show a similar overall behaviour as the one observed in Scenario (a), although with some minor differences: Netal's performance decreases, which is not surprising as alignment methodologies, such as Netal, are not developed to tackle this type of scenario and goal, however for the ER and the Chung-Lu model Netal's performance is better or comparable to GCD and GDDA. Results for GCD, GDDA and Netdis are also similar to what is observed in Scenario (a), although for this scenario GDDA is able to find some similarities between ER networks, and Netdis clearly displays a better performance than the others.

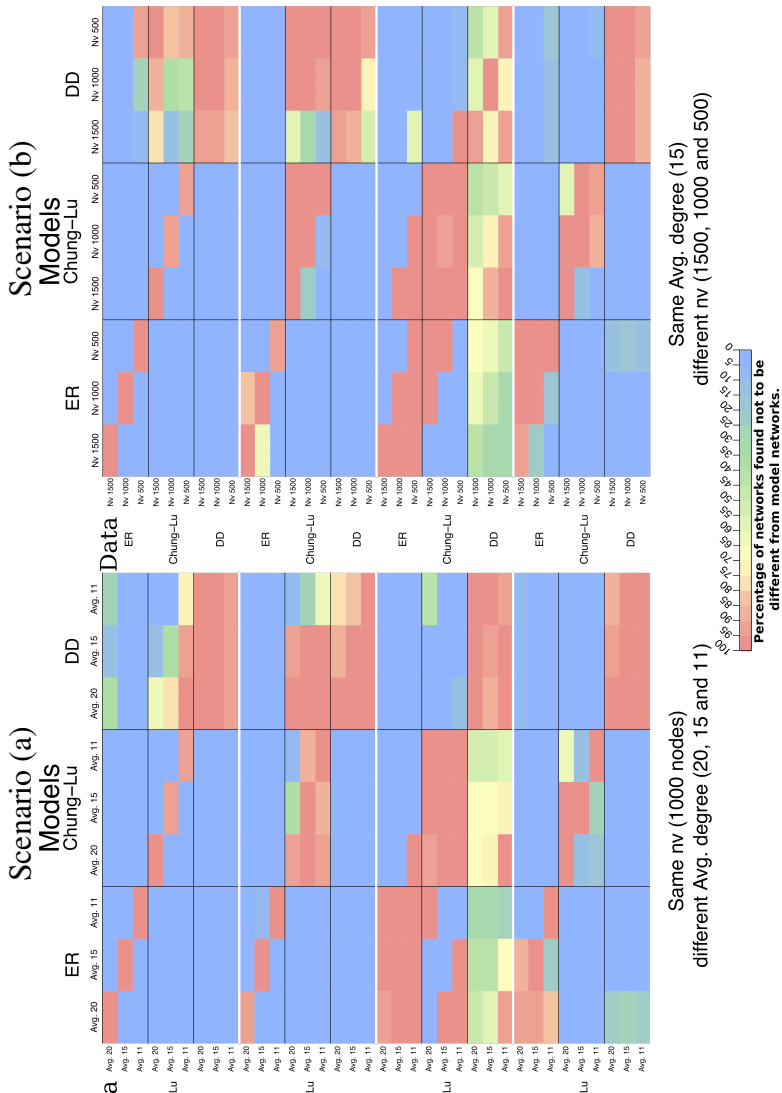

FIG. 2. This figure shows the results for Scenario (a) where all networks generated are set to have 1000 nodes and expected average degrees of 20, 15 and 11; and results for Scenario (b), where all networks generated are set to have an expected average degree of 15 and node sizes of 1500, 1000 and 500. The colour scale shows the percentage of times the null hypothesis "$H_0 : G_0$ is a realisation of model B" is not rejected, at 5%, from 100 realisations of the Monte-Carlo test using the four network comparison statistics GCD, GDDA, Netal and Netdis. An ideal result of a network comparison statistic, relative to model fit, is shown in Figure 3. Despite the fact that all methods aim to show how 'close' or 'far' two networks are from one another, they are not always able to tell when the networks come from the same or from different network generation mechanisms. Over both scenarios two opposite behaviours can be seen: On one hand methods can perform better at telling fine grained differences between networks, but struggle at detecting when networks share the same network generation mechanism, such as displayed by GCD. On the other hand, a compromise between detecting fine grain differences and the broader scale similarities can be achieved; as displayed by Netdis.

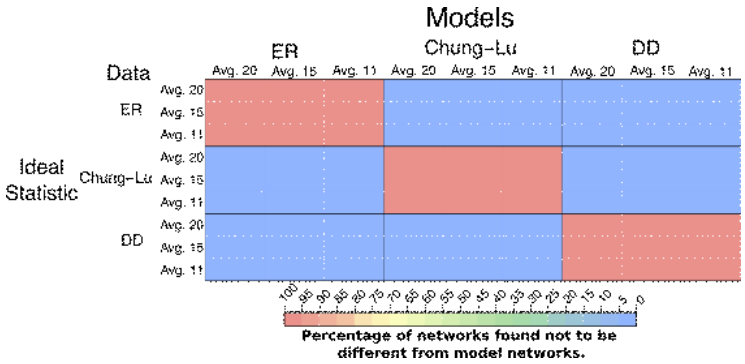

FIG. 3. The Monte Carlo result for Scenario (a) of an ideal network comparison statistic that is only able to capture the network generation mechanisms, regardless of the number of nodes or edge density of the networks being compared. The colour scale shows the percentage of times the null hypothesis "H_0 : G_0 is a realisation of model B" is not rejected, at a 5% level.

The results shown in scenarios (a) and (b) also illustrate clear differences between the network comparison statistics based on subgraph counts that are not evident from their formulation in their respective publications. GCD, GDDA and Netdis, all aim to measure "closeness" between networks and they all use similar inputs (subgraph counts). However, from Figure 2, it can be seen that GCD focuses more on fine grained differences, thus finding discrepancies more easily within networks of the ER and Chung-Lu models. In contrast, GDDA and Netdis, seem to focus on coarser differences, thus detecting similarities between networks of the same model despite having different number of nodes or edges.

Over all the four network comparison methods considered, Netdis shows the best compromise between detecting similarities between networks generated from the same model (irrespective of their size or density), and detecting differences between networks coming from different models (irrespective of their size or density). However, it can also be seen that the results of none of the methods used are completely invariant under the number of nodes or network density, as a different number of nodes and/or average degree can suffice for most methods to avoid detecting that networks share the same network generation mechanism.

3.2 *Testing the Chung-Lu and Duplication divergence models as null models for protein-protein interaction networks*

Using GDDA Hayes et al. (2013) suggested that the Chung-Lu model is a good model for the Human, Yeast, Worm, Fly and E. coli PPI networks, but provided evidence only for the five small virus networks EBV, HSV-1, KSHV, mCMV and VZV. In this section we use the proposed Monte Carlo test to assess whether the large and updated PPI networks of Human, Yeast, Worm, Fly and E. coli can be viewed as a realisation of a Chung-Lu model with the same degree sequence of the respective PPI network. In addition to the Chung-Lu model, we also consider the biologically inspired duplication divergence model from Section 2.3. We perform a Monte-Carlo test, as proposed in Section 2.4, with each of the four network comparison methods to test whether the PPI networks can be thought as realisations of the Chung-Lu or DD models.

Table 3 shows the results of the Monte-Carlo test using each of the four network comparison statis-

tics. All four statistics reject the Chung-Lu model for the five larger PPI networks. However, apart from the EBV virus, there is no unanimous rejection of the Chung-Lu model among the four network comparison statistics for all other virus networks. In addition, the DD model is only rejected on a single occasion and by a single comparison statistic; in all other cases the DD model is not rejected. Thus, both the Chung-Lu and the DD model can generate networks with a subgraph structure that is "similar" to that of the virus PPI networks. Most of the results for the five large PPI networks (Yeast, Human, Worm, Fly and E. coli) reject both the Chung-Lu and the DD model. Among the large PPI networks, only for Worm and E. coli the DD model is not rejected by all four network comparison statistics.

Model	PPI	GCD	GDDA	Netal	Netdis
CH-L	Worm	**0.01**	**0.01**	**0.01**	**0.01**
	Fly	**0.01**	**0.01**	**0.01**	**0.01**
	Human	**0.01**	**0.01**	**0.01**	**0.01**
	Yeast	**0.01**	**0.01**	**0.01**	**0.01**
	E. coli	**0.01**	**0.01**	**0.01**	**0.01**
CH-L	mCMV	**0.04**	0.35	**0.01**	0.06
	KSHV	**0.05**	**0.01**	0.14	0.98
	VZV	**0.05**	0.08	0.11	0.11
	HSV-1	0.38	0.50	**0.04**	**0.01**
	EBV	**0.02**	**0.03**	**0.02**	**0.02**
DD	Worm	**0.01**	0.35	**0.01**	0.13
	Fly	**0.01**	**0.01**	**0.01**	**0.01**
	Human	**0.01**	**0.02**	**0.01**	**0.04**
	Yeast	**0.01**	**0.03**	**0.01**	0.07
	E. coli	**0.02**	0.41	**0.01**	**0.01**
DD	mCMV	0.77	0.77	0.17	0.07
	KSHV	0.76	0.64	0.13	0.94
	VZV	0.64	0.78	0.20	0.90
	HSV-1	0.77	0.76	0.38	**0.01**
	EBV	0.64	0.24	0.23	0.29

Table 3. Table of p-values of the Monte-Carlo test using the network comparison statistics GCD, GDDA, Netal and Netdis. The test is performed for the updated E. coli, Worm, Fly, Yeast and Human PPI networks, and the small virus PPI networks. Here p-values smaller or equal to 0.05 are in bold. The smallest possible p-value is equal to $\frac{1}{99+1} = 0.01$. The DD and Chung-Lu models are rejected as models for the large PPI networks for most of the network comparison statistics. In contrast for the smaller virus PPI networks the Chung-Lu model, and the DD model are not rejected by most of the network comparison statistics. The dashed line separates large from small PPI networks.

From the Monte Carlo tests we find no evidence for the claim that the Chung-Lu model is a suitable model for the large PPI networks, even with the same statistic used by Hayes et al. (2013), GDDA. We note that Hayes et al. (2013) showed no data for their claim that large PPI networks can be modelled well by a Chung-Lu model. Hence, to understand whether our results differ because of the different data sets used, we use GDDA and perform the Monte Carlo test with the same PPI networks used by Hayes et al. (2013), see Appendix D. We find the same results as the ones obtained for the updated versions of the Yeast, Human, Worm, Fly and E. coli PPI networks, (given in Table 3). Only for some of the virus PPI networks we do obtain evidence to support the conclusion proposed by Hayes et al. (2013).

The fact that inter-species interactions are not taken into account in the virus networks, and the fact that cellular machinery highly relevant for the virus life cycle usually is from an organism (host) of a different species, could be reasons why the virus networks perform differently.

In contrast to the virus networks studied, all other organisms used in this work (except for *E. coli*) are eukaryotic organisms, and therefore have an additional diversity present in their PPI networks. This

diversity arises from the fact that eukaryotic organisms are composed of different cellular compartments, e.g. nucleus, Golgi apparatus, etc.; and the fact that the protein interaction relations between and within cellular compartments are different. Tarassov et al. (2008) showed, for yeast, that the number of protein interactions within the same cellular compartment are often larger than expected. In contrast, the number of protein interactions between different cellular compartments can in some cases be higher than expected and, in other cases be lower than expected.

In addition, even when considering prokaryote organisms, such as *E. coli*, there are still significant differences with viruses, as viruses are not able to reproduce on their own (Moreira and Lopez-Garcia, 2009), which in turn means that all protein interactions related to the reproduction of the viruses will not be present on their own PPI networks. Hence, similar complex and heterogeneous relations of the protein interactions of other organisms may only be present on the virus-host PPI networks rather than in the virus own networks. However, we note that virus-host interactions are often not taken into account when analysing PPI networks, as there are no clear ways to select which hosts to consider, (e.g. Hayes et al., 2013).

Lastly, another factor that could be leading to this difference in results is the small size of the virus PPI networks. These virus PPI networks are small networks (the largest has 111 nodes), in comparison to the Worm, Fly, Human, Yeast or E. coli PPI networks (the smallest having 2002 nodes). With a lower number of nodes there are fewer options for connections available, which may lead to a lower level of complexity than the one observed in other larger PPI networks. Hence, it may be easier for random graph models to reproduce network structures which are similar to the ones observed in the virus PPI networks.

3.3 *Testing the ER and Chung-Lu models as null models for the Facebook networks*

The edge probabilities defined in the Chung-Lu model are commonly used as a background model used to find community structures, in both PPI networks and social networks (e.g. Newman, 2006; Porter et al., 2009; Lewis et al., 2010; Onnela et al., 2012; Traud et al., 2012; Fosdick et al., 2017). Here we test if the Facebook networks of five USA universities from 2005 can also be considered as realisations of the ER or Chung-Lu models. The p-values of the Monte-Carlo test using the four network comparison methods are given in Table 4.

Model	FB	GCD	GDDA	Netal	Netdis
ER	Caltech	**0.01**	**0.01**	**0.01**	**0.01**
	Reed	**0.01**	**0.01**	**0.01**	**0.01**
	Haverford	**0.01**	**0.01**	**0.01**	**0.01**
	Simmons	**0.01**	**0.01**	**0.01**	**0.01**
	Swarthmore	**0.01**	**0.01**	**0.01**	**0.01**
CH-L	Caltech	**0.01**	**0.01**	**0.01**	**0.01**
	Reed	**0.01**	**0.01**	**0.01**	**0.01**
	Haverford	**0.01**	**0.01**	**0.01**	**0.01**
	Simmons	**0.01**	**0.01**	**0.01**	**0.01**
	Swarthmore	**0.01**	**0.01**	**0.01**	**0.01**

Table 4. Monte Carlo p-values using the network comparison statistics GCD, GDDA, Netal and Netdis. The test is performed for Facebook networks of five USA universities, p-values smaller or equal to 0.05 are in bold. The ER and Chung-Lu models are rejected as models for these five Facebook networks by all of the network comparison statistics.

The results obtained by GCD, GDDA, Netal and Netdis reject the null hypothesis in all cases. Thus we cannot recommend the Chung-Lu model as a model that recreates similar subgraph configurations

to the ones observed in these Facebook networks.

4. Discussion

We illustrated a rigorous framework to statistically assess when a given network can be considered as a realisation of a random graph model, by means of any network comparison statistic. We have also shown how the analysis of the type two errors in this framework provides a level playing field in which different network comparison statistics can be compared with regards to standard goals, such as model fit. This framework allowed us to find previously unknown differences between the four network comparison statistics used, and which could otherwise have been difficult to identify, particularly for GCD, GDDA and Netdis, which use similar comparison strategies (subgraph counts). We found that GCD focuses more on fine grained differences between networks while GDDA, Netal and Netdis focus more on coarser differences, thus making the detection of common generation mechanisms easier. We also found that changes in edge density or number of nodes imposes a challenge for the network comparison statistics. For example, for GCD this means that it might not be able to detect networks that share the same network generation mechanism, while for Netdis this means that it might not be able to perfectly differentiate between networks with different network generation mechanisms. Overall, Netdis, using the proposed geometric Poisson variation to obtain the expected counts, showed the best compromise between detecting similarities and detecting differences between networks that have different sizes or densities.

We also found that Netal, despite not being developed to compare networks of with large different number of nodes and edges, was able to relate ER and Chung-Lu networks to other ER and Chung-Lu networks with different number of nodes and edges, respectively.

In the application of the Monte Carlo test to the PPI networks, we showed that recent protein-protein interaction networks of Yeast, Human, Fly and Worm are not fitted well by the Chung-Lu model nor the DD model. In contrast, we find that all of the virus data sets can be seen as a realisation of the DD model. The difference in the results obtained between the virus PPI networks and the other larger PPI networks may stem from the level of complexity of the networks, and the fundamental differences between viruses and the other organisms.

The methodology is also applied to five small Facebook networks. Here, all network comparison methods reject both the Chung-Lu and the ER model. Hence, while the Chung-Lu model might be well suited for the task of community detection it does not provide a good representation of the small structure accounted by the occurrence of small subgraphs.

Funding

This work was supported by Colciencias [568 to L.O.]; the EPSRC [EP/K032402/1 to G.R., EP/G037280/1 and EP/L016044/1 to C.D.]. C.D. and G.R. acknowledge the support of the Alan Turing Institute (EP/NS10129/1).

Acknowledgment

We thank Amanda L. Traud for kindly providing the Facebook networks, and Andrew Barbour for his suggestions regarding the use of the geometric Poisson approximation for subgraph counts, and other useful comments. We also thank the anonymous reviewers for their helpful comments which have lead to an improvement of this paper.

Appendix

A. Example: Orbit counts

The orbit counts used in Section 2.2 are obtained considering induced subgraphs only, hence other non-induced subgraphs with n-nodes, ($n \in \{2,3,4,5\}$), which are included in the induced subgraph on n-nodes are not taken into account in the counting process. Consider the graph in Figure A.4 (which is repeated three times). The orbit counts are obtained by counting the number of times each node in the graph is "touching" each of the subgraphs on two to five nodes at the individual automorphism orbits (see Figure 1). For example, in Figure A.4, node 5 is touching the 2-star subgraph (in red) three times at orbit 1, and zero times at orbit 2, (zero times as well at orbit 3 in the triangle subgraph), since node 5 is not in the middle of any 2-star subgraph nor in any part of a triangle. On the other hand, node 1 only touches a 2-star subgraph at orbit 1 two times (corresponding to the induced subgraphs 1-3-5 and 1-3-4); if nodes 1, 2 and 3 are considered, the induced subgraph corresponds to a triangle subgraph only. Hence, node 1 touches a triangle subgraph only once.

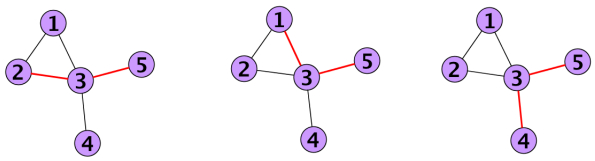

FIG. A.4. A graph on 5 nodes (repeated three times) illustrating the number of times that node 5 is "touching" the 2-star subgraph (red) at orbit 1.

Node \Orbit	1	2	3
Node 5	3	0	0
Node 4	3	0	0
Node 3	0	5	1
Node 2	2	0	1
Node 1	2	0	1

Table A.5. Number of times each node in the graph is "touching" the 2-star and triangle subgraphs at orbits 1, 2 and 3, (see Figure 1).

Once all counts of Table A.5 are compiled, the orbit degree is the number of times a node touched that orbit exactly zero, one, two, three, etc. times; see Table A.6 for an example.

Orbit \ Degree	0	1	2	3	4	5
Orbit 1	1	0	2	2	0	0
Orbit 2	4	0	0	0	0	1
Orbit 3	2	3	0	0	0	0

Table A.6. Orbit degree of the first, second and third orbit for the graph shown in Figure A.4. See Figure 1 for a graphical representation of different orbits on subgraphs with two to four nodes.

B. Setup and results of the simulation study

We use two simulation scenarios to perform a study of the type two error at 0.05 for the Monte Carlo test via different network comparison methods. These scenarios are:

(a) All model networks are generated with 1000 nodes and with approximate average degrees of 20, 15 and 11.

(b) All model networks are generated with a different number of nodes (1500, 1000 and 500), and with an approximate average degree of 15.

The aim of these scenarios is to better understand the performance of the different network comparison methods when networks have similar number of nodes and edges but come from different models. We are also interested in the ability of the network comparison methods to detect that two networks do (or do not) come from the same model, despite having different number of nodes and/or different average degrees.

To carry out the simulation we first set the parameters of the duplication-divergence model $DD(p,q)$ of Vázquez et al. (2003) on 1000 nodes (n_v) to $p = 0.2$ and $q = 0.3$. Similar parameters have been used to model some Yeast PPI networks in the past, e.g. ($p = 0.26$, $q = 0.33$) and ($p = 0.22$, $q = 0.54$) (Shao et al., 2013). This parameter selection leads to *DD* networks with an expected average degree of 20.274. Then we consider the average degrees that would be obtained if the number of nodes is reduced in half (500) and to a quarter (250) (still maintaining $p = 0.2$ and $q = 0.3$). The resulting average degrees are 15.12 and 11.226 for 500 and 250 nodes, respectively. We use these expected degrees in order to set the parameters for scenarios (a) and (b) in the following manner.

In Scenario (a) all networks are generated with 1000 nodes ($n_v = 1000$). Hence, in order to fix the expected average degrees of the DD models used in Scenario (a) to 20.274, 15.12 and 11.226, we maintain $p = 0.2$ and vary q. This procedure leads to $q = 0.3$, $q = 0.32996$ and $q = 0.36122$, respectively for the expected average degrees 20.274, 15.12 and 11.226. The parameter of a Chung-Lu model is a complete degree sequence of a graph. To that purpose, we use the degree sequence of a realisation of a DD model with our parameter choices $\{p = 0.2,\ q = 0.3\}$; $\{p = 0.2,\ q = 0.32996\}$ and $\{p = 0.2,\ q = 0.36122\}$. We select degree sequences such that their average degree is 20.274, 15.12 and 11.226, respectively. For the $ER(n_v,\rho)$ model, where ρ is the probability of connecting any two nodes, we use $\rho = 0.02029$, $\rho = 0.01513$ and $\rho = 0.01124$, respectively for each of the expected average degrees 20.274, 15.12 and 11.226.

For Scenario (b), networks are generated with a fixed expected average degree of 15.12 but with different number of nodes (1500, 1000 and 500). For the DD model, as in Scenario (a), we fix $p = 0.2$ and varied q in order to obtain the desired degree with the different number of nodes. This procedure leads to q values of 0.3442998, 0.3299654 and 0.3, for networks of sizes 1500, 1000 and 500, respectively. The parameters of the Chung-Lu model are obtained from degree sequences of realisations of DD graphs, as in Scenario (a). We select degree sequences such that their average degree is 15.12. For the $ER(n_v,\rho)$ model, where ρ is the probability of connecting any two nodes, we use $\rho = 0.01009$, $\rho = 0.01514$ and $\rho = 0.03031$, respectively for each of the network sizes.

In both simulation scenarios, we take a network G_0 generated from model A and use the Monte Carlo test to evaluate the null hypothesis "$H_0 : G_0$ is a realisation of model B" against the general alternative, with a significance level α of 5%. Models A and B can be any of the three random graph models, ER, DD and Chung-Lu. We repeat this test 100 times and record the number of times the test rejected network G_0 as a realisation of model B, for each corresponding network comparison statistic. Tables A.7 and A.8 show the results of these simulations for scenarios (a) and (b), respectively.

		Model	ER			CHL			DD		
	Data	Avg d.	20	15	11	20	15	11	20	15	11
GCD	ER	20	**0.97**	0.00	0.00	0.00	0.00	0.00	0.35	0.12	0.25
		15	0.00	**0.96**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		11	0.00	0.00	**0.95**	0.00	0.00	0.00	0.00	0.00	0.00
	CHL	20	0.00	0.00	0.00	**0.95**	0.00	0.00	**0.6**	0.13	0.00
		15	0.00	0.00	0.00	0.00	**0.94**	0.00	**0.78**	0.37	0.04
		11	0.00	0.00	0.00	0.00	0.00	**0.92**	**0.99**	**0.94**	**0.73**
	DD	20	0.00	0.00	0.00	0.00	0.00	0.00	**0.96**	**0.95**	**0.95**
		15	0.00	0.00	0.00	0.00	0.00	0.00	**0.96**	**0.95**	**0.95**
		11	0.00	0.00	0.00	0.00	0.00	0.00	**0.92**	**0.92**	**0.93**
GDDA	ER	20	**0.93**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		15	0.00	**0.95**	0.05	0.00	0.00	0.00	0.00	0.00	0.00
		11	0.00	0.00	**0.95**	0.00	0.00	0.00	0.00	0.00	0.00
	CHL	20	0.00	0.00	0.00	**0.93**	0.36	0.05	**1.00**	**0.92**	0.05
		15	0.00	0.00	0.00	**0.99**	**0.98**	**0.86**	**1.00**	**1.00**	0.28
		11	0.00	0.00	0.00	**0.91**	**0.87**	**0.96**	**1.00**	**1.00**	**0.64**
	DD	20	0.00	0.00	0.00	0.00	0.00	0.00	**0.97**	**0.89**	**0.75**
		15	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	**0.95**	**0.83**
		11	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	**1.00**	**0.98**
Netal	ER	20	**0.92**	**1.00**	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00
		15	**1.00**	**0.96**	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00
		11	**1.00**	**1.00**	**0.95**	**1.00**	0.00	0.00	0.00	0.00	0.00
	CHL	20	0.00	0.00	**1.00**	**0.94**	**1.00**	**1.00**	0.00	0.00	0.42
		15	**1.00**	0.00	0.00	**1.00**	**0.97**	**1.00**	0.00	0.00	0.02
		11	**1.00**	**1.00**	0.00	**1.00**	**1.00**	**0.97**	0.11	0.00	0.00
	DD	20	0.46	0.43	0.32	**0.65**	**0.65**	**0.53**	**0.96**	**0.96**	**1.00**
		15	**0.59**	0.42	0.3	**0.72**	**0.66**	**0.52**	**0.89**	**0.94**	**0.97**
		11	**0.9**	**0.66**	0.25	**0.95**	**0.73**	**0.55**	**0.97**	**0.95**	**0.94**
Netdis	ER	20	**0.94**	**0.86**	0.00	0.00	0.00	0.00	0.09	0.09	0.09
		15	**0.92**	**0.98**	0.00	0.00	0.00	0.00	0.02	0.01	0.02
		11	**0.81**	0.22	**0.95**	0.00	0.00	0.00	0.00	0.00	0.00
	CHL	20	0.00	0.00	0.00	**0.98**	**1.00**	**0.60**	0.01	0.00	0.00
		15	0.00	0.00	0.00	0.14	**0.95**	0.12	0.02	0.00	0.00
		11	0.00	0.00	0.00	0.28	0.19	**0.95**	0.00	0.00	0.00
	DD	20	0.22	0.00	0.00	0.00	0.00	0.00	**0.97**	**0.91**	**0.87**
		15	0.28	0.00	0.00	0.00	0.00	0.00	**0.99**	**1.00**	**0.99**
		11	0.21	0.00	0.00	0.00	0.00	0.00	**0.99**	**0.96**	**0.96**

Table A.7. Results of Scenario (a) where all networks generated are set to have 1000 nodes and approximate average degrees of 20, 15 and 11. The value in the table shows the percentage of times the null hypothesis "$H_0 : G_0$ is a realisation of model B" is not rejected against the general alternative, at 5%, from 100 realisations of the Monte Carlo test using the four network comparison statistics GCD, GDDA, Netal and Netdis. Despite the fact all methods aim to show how 'close' or 'far' two networks are from one another, they may not detect when the networks come from the same network generation mechanism. GCD performs better at telling fine grained differences between networks, but may not detect when the networks share the same network generation mechanism. Netdis shows a good compromise between fine grain differences and more broad scale similarities. Values larger than 0.50 are shown in bold.

C. Reliability of thresholds in model selection

In this section we show that the use of a universal threshold in network comparison statistics as a way to establish which networks share the same network generation mechanisms may lead to unreliable results. Here we continue using the two previous simulation scenarios defined in Appendix B.

Pržulj (2007) made GDDA comparisons of synthetic networks generated from the same model, with the same parameters and the same number of nodes. Among the different models they considered, the smallest average GDDA value (0.86) was obtained for Scale-Free Networks.

In this section we explore the behaviour of using such value as a cut off to indicate that the networks being compared are closely related. Here we use this threshold to classify networks into a given network generation mechanism. Whenever the comparison between a pair of networks passes the cut-off, they

		Model	ER			CHL			DD		
	Data	N. of nodes	1500	1000	500	1500	1000	500	1500	1000	500
GCD	ER	1500	**0.95**	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		1000	0.04	**0.96**	0.00	0.00	0.00	0.00	0.00	0.00	0.02
		500	0.00	0.00	**0.95**	0.00	0.00	0.00	0.06	0.28	**0.90**
	CHL	1500	0.00	0.00	0.00	**0.99**	0.00	0.00	**0.77**	**0.85**	**0.98**
		1000	0.00	0.00	0.00	0.00	**0.94**	0.00	0.14	0.37	**0.83**
		500	0.00	0.00	0.00	0.00	0.00	**0.94**	0.29	0.43	**0.87**
	DD	1500	0.00	0.00	0.00	0.00	0.00	0.00	**0.94**	**0.98**	**0.99**
		1000	0.00	0.00	0.00	0.00	0.00	0.00	**0.92**	**0.95**	**1.00**
		500	0.00	0.00	0.00	0.00	0.00	0.00	**0.82**	**0.85**	**0.93**
GDDA	ER	1500	**0.95**	**0.83**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		1000	**0.64**	**0.95**	0.02	0.00	0.00	0.00	0.00	0.00	0.00
		500	0.00	0.01	**0.92**	0.00	0.00	0.00	0.00	0.00	0.04
	CHL	1500	0.00	0.00	0.00	**0.96**	**1.00**	**1.00**	**0.59**	**1.00**	**1.00**
		1000	0.00	0.00	0.00	0.22	**0.98**	**0.99**	0.33	**1.00**	**1.00**
		500	0.00	0.00	0.00	0.00	0.08	**0.95**	0.12	**0.92**	**1.00**
	DD	1500	0.00	0.00	0.00	0.00	0.00	0.00	**0.92**	**0.95**	**1.00**
		1000	0.00	0.00	0.00	0.00	0.00	0.00	**0.88**	**0.95**	**1.00**
		500	0.00	0.00	0.00	0.00	0.00	0.00	**0.53**	**0.73**	**0.92**
Netal	ER	1500	**0.93**	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		1000	**1.00**	**0.96**	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00
		500	**1.00**	**1.00**	**0.96**	**1.00**	**1.00**	0.00	0.00	0.00	0.00
	CHL	1500	0.00	**1.00**	**1.00**	**0.95**	**1.00**	**1.00**	0.00	0.00	0.01
		1000	0.00	0.00	**1.00**	**1.00**	**0.97**	**1.00**	0.00	0.00	0.00
		500	0.00	0.00	0.00	**1.00**	**1.00**	**0.92**	**0.26**	0.00	0.00
	DD	1500	0.35	**0.57**	**0.71**	**0.57**	**0.55**	**0.62**	**0.96**	**0.77**	**0.53**
		1000	0.27	0.42	**0.64**	**0.65**	**0.66**	**0.54**	**0.73**	**0.94**	**0.56**
		500	0.31	0.32	0.47	**0.90**	**0.86**	**0.70**	**0.75**	**0.70**	**0.96**
Netdis	ER	1500	**0.92**	**1.00**	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00
		1000	0.21	**0.98**	**1.00**	0.00	0.00	0.00	0.00	0.01	0.02
		500	0.00	0.16	**0.97**	0.00	0.00	0.00	0.10	0.13	0.15
	CHL	1500	0.00	0.00	0.00	**0.98**	**0.95**	**0.56**	0.00	0.00	0.02
		1000	0.00	0.00	0.00	0.14	**0.95**	**0.95**	0.00	0.00	0.04
		500	0.00	0.00	0.00	0.03	**0.85**	**0.91**	0.00	0.00	0.06
	DD	1500	0.00	0.00	0.14	0.00	0.00	0.00	**0.96**	**0.97**	**0.97**
		1000	0.00	0.00	0.15	0.00	0.00	0.00	**0.99**	**1.00**	**0.99**
		500	0.00	0.00	0.10	0.00	0.00	0.00	**0.85**	**0.89**	**0.93**

Table A.8. Results of Scenario (b) where all networks generated are set to have expected average degrees of 15 but with different number of nodes 1500, 1000 and 500. The value in the table shows the percentage of times the null hypothesis "$H_0 : G_0$ is a realisation of model B" is not rejected against the general alternative, at 5%, from 100 realisations of the Monte Carlo test using the four network comparison statistics GCD, GDDA, Netal and Netdis. Despite the fact all methods aim to show how 'close' or 'far' two networks are from one another, they may not detect when the networks come from the same network generation mechanism. GCD performs better at telling fine grain differences between networks, but may not detect when the networks share the same network generation mechanism. Netdis shows a good compromise in fine grain differences to compare more broad scale similarities. Values larger than 0.50 are shown in bold.

are considered as coming from the same network generation mechanism. We show that, even with a small number of models considered (ER, Chung-Lu and DD), this type of procedure is not always able to correctly identify networks coming from the same network generation mechanism, even when the networks are generated from the same model with the same number of nodes and parameter values.

In detail, for all combinations of model and average degree (Scenario (a)), or model and number of nodes (Scenario (b)), we perform pairwise GDDA comparisons among networks generated from each of the models, thus leading to 9×9 GDDA pairwise comparisons. We repeat this procedure 100 times and record the percentage of network comparisons such that one minus their GDDA comparison (1-GDDA) was smaller than the threshold proposed in Pržulj (2007), $0.14 = 1 - 0.86$. Figure A.5 shows the results. Similarly to the plots shown for the Monte Carlo test (Figures 2 and 3), here the plots are arranged by a

grid of 9×9 blocks, which correspond to all pairwise comparisons among networks from the different models along with their respective average degrees, or number of nodes.

Note that if the threshold is used to state that two networks share the same network generation mechanism, the ideal result would show that all comparisons between networks from the same model would be smaller than the threshold considered for 1-GDDA, and all comparisons between networks from different models would be larger than such threshold.

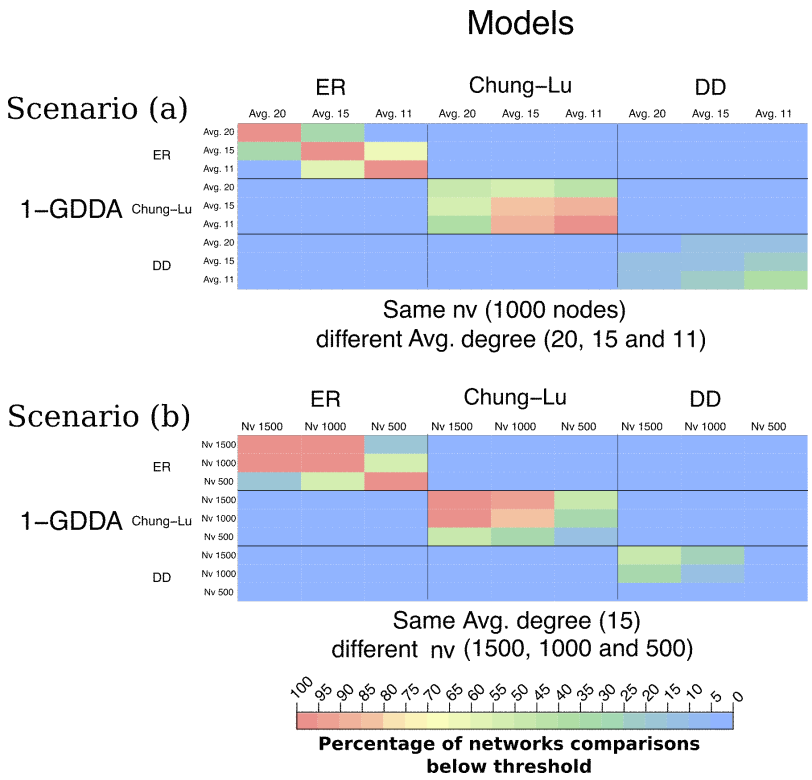

FIG. A.5. Percentage of 1-GDDA pairwise network comparisons with values smaller than $0.14 = 1 - 0.86$.

The results shown in Figure A.5 illustrate that considering the 0.14 threshold as a way to state whether two networks share similar generation mechanisms may lead to unreliable results. For example, for both scenarios (a) and (b), the GDDA was in most cases not able to detect that DD networks generated from the same model, and even with the same number of nodes and same parameters, did come from the same model. This behaviour is also observed for networks coming from the Chung-Lu model with 500 nodes and average degree 15, (Scenario (b)), where less than 20% of the comparisons detected that the networks came from the same model. Similarly, for Chung-Lu networks with 1000 nodes and average degree 20, approximately 50% of the comparisons do not detect that the networks came from the same model, even when those networks were generated with the same number of nodes and the same degree sequence.

In Figure A.5 we do not observe network comparisons with a comparison value smaller than the threshold (0.14) for networks coming from different models. However this does not mean that such

behaviour cannot occur. Take for example 1-GDDA comparisons between a complete graph and a line-like network on 10, 20, 30,...,90 and 100 nodes (Figure A.6). It can be noted that all comparison values are below the threshold $0.14 = 1 - 0.86$, despite line-like networks and complete networks being largely different.

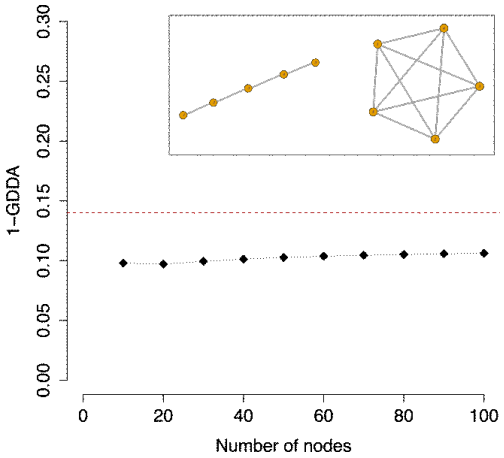

FIG. A.6. 1-GDDA comparisons of line-like networks vs. complete graphs on 10, 20, 30,...,90 and 100 nodes. The insert shows a sketch of these type of networks on five nodes. All comparison values are below the threshold $0.14 = 1 - 0.86$ (shown in red).

The results shown in Figures A.5 and A.6 illustrate that considering a threshold as a way to state that two networks share the same generation mechanisms or, are similar to one another, can lead to a wide array of results, some of which are misleading and inconsistent. In addition, there is no telling when the threshold used will be able to correctly assess the similarity between networks. Thus, for a task of model fit, a more reliable method should be used.

D. Inspecting claims of good fit

Hayes et al. (2013) wrote "... Examining the biological reasons for the good fit of the sticky model in 80% of the viral networks, and why it is a less good fit for KSHV, is a subject of future research. Similar plots (not shown because of space limitations) of three bacterial PPI networks [MZL (Shimoda et al., 2008), SPP (Sato et al., 2007) and CJJ (Parrish et al., 2007)], the functional interaction network of E. coli (Peregrin - Alvarez et al., 2009), as well as the Arabidopsis thaliana PPI network (Arabidopsis Interactome Mapping Consortium, 2011), and PPI networks of Yeast, Worm, Fly and Human from BioGRID, all indicate that STICKY, SF-GD and GEO-GD (in that order) are the best fitting models for these networks."

In this section we use the PPI datasets analysed by Hayes et al. (2013) that were publicly available, and for which "STICKY, SF-GD and GEO-GD (in that order) are the best fitting models" (Hayes et al., 2013). These datasets are shown in Table A.9. All self-loops, multiple edges and interspecies interactions are removed from PPI networks, leaving simple undirected networks for the analysis. All degree 0 nodes were also removed from the analysed PPI networks. Three of the BioGRID datasets used by Hayes et al. (2013) have a slightly different number of nodes and edges. Table A.9 describes these networks; the column *density* in the tables is the number of edges divided by the possible number of edges

and is given here to aid comparisons.

	Nodes	Edges	Density	Extracted from
Human	8920	35386	0.000890	BioGRID ver 3.1.74
Worm	2831*	4527	0.001130	BioGRID ver 3.1.74
Fly	7373*	24063	0.000885	BioGRID ver 3.1.74
Yeast	5608*	57143	0.003635	BioGRID ver 3.1.74
AT	2634	5529	0.001594	(Dreze et al., 2011)
E. coli	1941	3989	0.002119	(Peregrín-Alvarez et al., 2009)
EBV	60	208	0.117514	(Fossum et al., 2009)
HSV-1	47	100	0.092507	(Fossum et al., 2009)
KSHV	50	115	0.093878	(Fossum et al., 2009)
mCMV	111	393	0.064373	(Fossum et al., 2009)
VZV	57	160	0.100251	(Fossum et al., 2009)

Table A.9. Number of nodes, edges, density and source of the PPI networks used by Hayes et al. (2013). *The number of nodes found is slightly different from those reported in (Hayes et al., 2013). Only the Worm PPI network has a difference greater than 1 node (Worm +14.)

D.1 *The Chung-Lu model is not generally a background model for local structure in protein-protein interaction networks*

Using the GDDA, Hayes et al. (2013) indicated that the Chung-Lu model (or sticky model) was in most cases the best model for the PPI networks in Table A.9, although they only showed results for the virus PPI networks. As we do not obtain the claimed results for the more recent PPI networks of Human, Yeast, Fly and Worm, we perform the Monte Carlo test for the data used by Hayes et al. (2013) using GDDA. Table A.10 shows the p-values obtained. Except for the virus datasets EBV, HSV-1, mCMV and VZV, all p-values take the smallest possible value of the Monte Carlo test. Hence, for 8 of the 11 available datasets used by Hayes et al. (2013) we reject the hypothesis, at 5% significance level, that the Chung-Lu model is a good null model.

Data	p-value (GDDA)
Networks used by Hayes et al. (2013) ($M = 30$ $N = 30$)	
Yeast	0.0322
Human	0.0322
Worm	0.0322
Fly	0.0322
AT	0.0322
ECL	0.0322
EBV	0.0400*
HSV-1	0.4516
KSHV	0.0322
mCMV	0.3548
VZV	0.1100*

Table A.10. Monte Carlo p-values using the data vs. model and model vs. model GDDA comparisons shown in Figure A.7. Note that for this particular test the smallest p-value is equal to $\frac{1}{30+1} = 0.0322$ ($M = 30$). We reject the null hypothesis that the Chung-Lu model fits when the p-value < 0.05. Hence, except for HSV-1, mCMV and VZV, all null hypotheses are rejected. Here p-values marked with '*' were obtained using $M = 99$ since $M = 30$ was inconclusive in this two cases. The p-values obtained with $M = 30$ for VZV and EBV were 0.0645 and 0.0967, respectively.

D.2 *Issues when using histograms for comparing distributions*

Hayes et al. (2013) used comparisons of 30 networks from the random graph model to create two samples; a first sample of comparisons of PPI data vs. model networks and a second sample of GDDA comparisons of model vs. model networks. From these samples, two histograms are created and the overlap between them is used to assess the goodness of fit between the model and the PPI networks.

Figure A.7 shows the histograms we obtain from the comparisons created in the application of the Monte Carlo test, where following Hayes et al. (2013) we take $N = M = 30$. Histograms with considerable overlap are seen for three of the five virus PPI networks. In contrast, for the other larger PPI networks the amount of overlap is very low or null.

The results shown in Figure A.7 for the virus networks are expected, since a similar figure is presented by Hayes et al. (2013), (no figure was shown by Hayes et al. (2013) for the larger PPI networks).

In addition, it can also be seen from A.7 that the amount of overlap is correlated with the Monte-Carlo p-values of Table A.10. However, we believe this approach is not well suited to test whether a given network can be thought as a realisation of a given random graph model, as:

1. There is no clear indication of what amount of overlap is necessary in order to claim that the observed data comes from the model. For example, despite all five networks have considerable overlap, only for the three virus networks with the largest overlap, the Monte-Carlo test does not reject these networks as realisations of the Chung-Lu model, at a 5% significance level.

2. It is not clear what histogram building rule should be used, as different rules lead to a different number of bins and therefore to possible changes in the resulting overlap. In Figure A.7 we use the Sturges rule, a standard binning procedure to construct histograms (Venables and Ripley, 2002, p. 112), however other rules could be used.

3. In case the distributions are heavy-tailed, the histograms may struggle to assign the correct amount of mass at the tail of the distributions, thus leading to less reliable overlap statistics. Newman, in (Newman, 2005), discusses some of the problems histograms have at the tail of the distribution.

4. The histograms are not built from samples drawn from the same distribution, even when the observed network is indeed generated from the proposed model. On the one hand, one histogram is formed from one-to-many comparisons (observed network vs model networks) while, on the other hand, the second histogram is formed from a sample of many-to-many comparisons (model networks vs model networks).

5. The overlap of the histograms does not reflect the sample size, although the importance of sample size is a fundamental statistical insight. For example, when two samples of n observations are obtained from $N(0,1)$ and $N(0.5,1)$, respectively, the overlap between the histograms of those two samples will converge to the overlap between the density functions of the two distributions as n goes to infinity. However, if instead of the raw values the average is used, as done in the Monte Carlo procedure, the overlap between the corresponding distributions of the sample means $N(0,1/n)$, and $N(0.5,1/n)$, tends towards 0 as n goes to infinity.

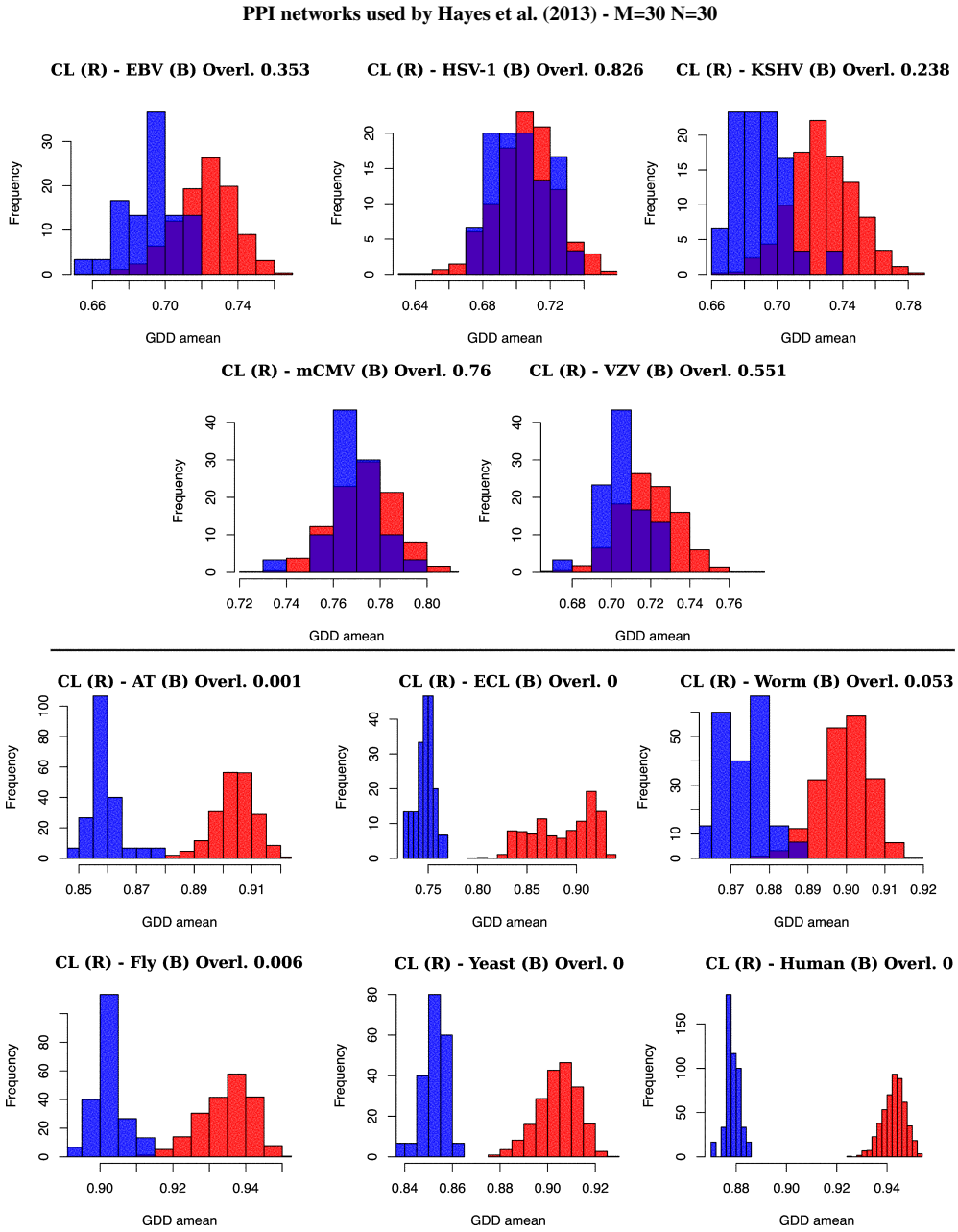

FIG. A.7. Histograms (Sturges rule) overlap of the GDDA values – data vs. model (blue) and model vs. model (red) (Chung-Lu model, CL)– obtained for the Monte Carlo test ($M = 30, N = 30$) of PPI networks shown in Table A.9.

The Sturges rule formula is $k = \lceil 1 + \log_2 N \rceil$, where k is the number of classes, $\lceil . \rceil$ is the ceiling function, and N is the size of the sample (Sturges, 1926), see also Venables and Ripley (2002, p. 112).

References

Ali, W., Rito, T., Reinert, G., Sun, F., and Deane, C. M. (2014). Alignment-free protein interaction network comparison. *Bioinformatics*, 30:i430–i437.

Aliakbary, S., Motallebi, S., Rashidian, S., Habibi, J., and Movaghar, A. (2015). Distance metric learning for complex networks: Towards size-independent comparison of network structures. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(2):023111.

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461.

Asta, D. and Shalizi, C. R. (2014). Geometric network comparison. *arXiv preprint arXiv:1411.1350*.

Berlingerio, M., Koutra, D., Eliassi-Rad, T., and Faloutsos, C. (2013). Network similarity via multiple social theories. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 1439–1440.

Chung, F. and Lu, L. (2002). Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6(2):125–145.

Crawford, J. and Milenković, T. (2015). GREAT: GRaphlet Edge-based network AlignmenT. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 220–227.

Dreze, M., Carvunis, A.-R., Charloteaux, B., Galli, M., Pevzner, S. J., Tasan, M., and Ahn (2011). Evidence for network evolution in an *Arabidopsis* interactome map. *Science*, 333(6042):601–607.

Emmert-Streib, F., Dehmer, M., and Shi, Y. (2016). Fifty years of graph matching, network alignment and network comparison. *Information Sciences*, 346-347:180 – 197.

Fosdick, B. K., Larremore, D. B., Nishimura, J., and Ugander, J. (2017). Configuring random graph models with fixed degree sequences. *arXiv preprint arXiv:1608.00607*.

Fossum, E., Friedel, C. C., Rajagopala, S. V., Titz, B., Baiker, A., Schmidt, T., Kraus, T., Stellberger, T., Rutenberg, C., Suthram, S., et al. (2009). Evolutionarily conserved herpesviral protein interaction networks. *PLoS Pathogens*, 5(9):e1000570.

Gibson, T. A. and Goldberg, D. S. (2011). Improving evolutionary models of protein interaction networks. *Bioinformatics*, 27(3):376–382.

Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144.

Hashemifar, S. and Xu, J. (2014). HubAlign: an accurate and efficient method for global alignment of protein–protein interaction networks. *Bioinformatics*, 30(17):i438–i444.

Hayes, W., Sun, K., and Pržulj, N. (2013). Graphlet-based measures are suitable for biological network comparison. *Bioinformatics*, 29(4):483–491.

Higueruelo, A. P., Jubb, H., and Blundell, T. L. (2013). Protein–protein interactions as druggable targets: recent technological advances. *Current Opinion in Pharmacology*, 13(5):791–796.

Ispolatov, I., Krapivsky, P., and Yuryev, A. (2005). Duplication-divergence model of protein interaction network. *Physical Review E*, 71(6):061911.

Lewis, A., Jones, N., Porter, M., and Deane, C. M. (2010). The function of communities in protein interaction networks at multiple scales. *BMC Systems Biology*, 4(1):100.

Malod-Dognin, N. and Pržulj, N. (2015). L-graal: Lagrangian graphlet-based network aligner. *Bioinformatics*.

Mamano, N. and Hayes, W. B. (2017). Sana: Simulated annealing far outperforms many other search algorithms for biological network alignment. *Bioinformatics*.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827.

Moreira, D. and Lopez-Garcia, P. (2009). Ten reasons to exclude viruses from the tree of life. *Nature Reviews Microbiology*, 7(4):306–311.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351.

Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74:036104.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69:026113.

Neyshabur, B., Khadem, A., Hashemifar, S., and Arab, S. S. (2013). NETAL: a new graph-based method for global alignment of protein–protein interaction networks. *Bioinformatics*, 29(13):1654–1662.

Noh, H. J., Ponting, C. P., Boulding, H. C., Meader, S., Betancur, C., Buxbaum, J. D., Pinto, D., Marshall, C. R., Lionel, A. C., Scherer, S. W., and Webber, C. (2013). Network topologies and convergent aetiologies arising from deletions and duplications observed in individuals with autism. *PLOS Genetics*, 9(6):1–12.

Onnela, J.-P., Fenn, D. J., Reid, S., Porter, M. A., Mucha, P. J., Fricker, M. D., and Jones, N. S. (2012). Taxonomies of networks from community structure. *Physical Review E*, 86:036104.

Patro, R. and Kingsford, C. (2012). Global network alignment using multiscale spectral signatures. *Bioinformatics*, 28(23):3105.

Payrato Borras, C., Hernandez, L., and Moreno, Y. (2017). Breaking the spell of nestedness. *bioRxiv*.

Peregrín-Alvarez, J. M., Xiong, X., Su, C., and Parkinson, J. (2009). The modular organization of protein interactions in *Escherichia coli*. *PLoS Computational Biology*, 5(10):e1000523.

Pereira-Leal, J., Levy, E., Kamp, C., and Teichmann, S. (2007). Evolution of protein complexes by duplication of homomeric interactions. *Genome Biology*, 8(4):R51.

Picard, F., Daudin, J.-J., Koskas, M., Schbath, S., and Robin, S. (2008). Assessing the exceptionality of network motifs. *Journal of Computational Biology*, 15(1):1–20.

Porter, M. A., Onnela, J.-P., and Mucha, P. J. (2009). Communities in networks. *Notices of the AMS*, 56(9):1082–1097.

Pržulj, N. (2007). Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183.

Pržulj, N. and Higham, D. J. (2006). Modelling protein–protein interaction networks via a stickiness index. *Journal of the Royal Society Interface*, 3(10):711–716.

Rajagopala, S. V., Sikorski, P., Kumar, A., Mosca, R., Vlasblom, J., Arnold, R., Franca-Koh, J., Pakala, S. B., Phanse, S., Ceol, A., Hauser, R., Siszler, G., Wuchty, S., Emili, A., Babu, M., Aloy, P., Pieper, R., and Uetz, P. (2014). The binary protein-protein interaction landscape of escherichia coli. *Nature Biotechnology*, 32(3):285–290.

Rito, T., Wang, Z., Deane, C. M., and Reinert, G. (2010). How threshold behaviour affects the use of subgraphs for network comparison. *Bioinformatics*, 26(18):i611–i617.

Sarajlić, A., Janjić, V., Stojković, N., Radak, D., and Pržulj, N. (2013). Network topology reveals key cardiovascular disease genes. *PloS One*, 8(8):e71537.

Saraph, V. and Milenković, T. (2014). MAGNA: Maximizing Accuracy in Global Network Alignment: Maximizing accuracy in global network alignment. *Bioinformatics*, 30(20):2931–2940.

Shao, M., Yang, Y., Guan, J., and Zhou, S. (2013). Choosing appropriate models for protein-protein interaction networks: a comparison study. *Briefings in Bioinformatics*, page 10.1093/bib/bbt014.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, 31(1):64–68.

Singh, R., Xu, J., and Berger, B. (2008). Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, 34:D535–D539.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66.

Sun, Y., Crawford, J., Tang, J., and Milenković, T. (2015). *Simultaneous Optimization of both Node and Edge Conservation in Network Alignment via WAVE*, pages 16–39. Springer Berlin Heidelberg, Berlin, Heidelberg.

Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Molina, M. M. S., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., and Michnick, S. W. (2008). An in vivo map of the yeast protein interactome. *Science*, 320(5882):1465–1470.

Topirceanu, A., Udrescu, M., and Vladutiu, M. (2013). Network fidelity: A metric to quantify the similarity and realism of complex networks. In *Cloud and Green Computing (CGC), 2013 Third International Conference on*, pages 289–296.

Traud, A. L., Mucha, P. J., and Porter, M. A. (2012). Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165 – 4180.

Vázquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003). Modeling of protein interaction networks. *Complexus*, 1(1):38–44.

Venables, W. N. and Ripley, B. D. (2002). *Modern Aapplied Statistics with S*. Statistics and Computing. Springer, New York, fourth edition.

West, J., Widschwendter, M., and Teschendorff, A. E. (2013). Distinctive topology of age-associated epigenetic drift in the human interactome. *Proceedings of the National Academy of Sciences*, 110(35):14138–14143.

Wuchty, S., Oltvai, Z. N., and Barabási, A.-L. (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics*, 35(2):176–179.

Yaveroglu, O. N., Malod-Dognin, N., Davis, D., Levnajic, Z., Janjic, V., Karapandza, R., Stojmirovic, A., and Przulj, N. (2014). Revealing the hidden language of complex networks. *Scientific Reports*, 4.

Zoraghi, R. and Reiner, N. E. (2013). Protein interaction networks as starting points to identify novel antimicrobial drug targets. *Current Opinion in Microbiology*, 16(5):566 – 572.