Linking the Network Centrality Measures Closeness and Degree

Tim S. Evans*, Bingsheng Chen

Centre for Complexity Science, and Theoretical Physics Group,
Imperial College London, London, SW7 2AZ, U.K.

Abstract

Measuring the importance of nodes in a network with a centrality measure is an core task in any network application. There many measures available and it is speculated that many encode similar information. We give an explicit non-linear relationship between two of the most popular measures of node centrality: degree and closeness. Based on a shortest-path tree approximation, we give an analytic derivation that shows the inverse of closeness is linearly dependent on the logarithm of degree. We show that our hypothesis works well for a range of networks produced from stochastic network models and for networks derived from 130 real-world data sets. We connect our results with previous results for other network distance scales such as average distance. Our results imply that measuring closeness is broadly redundant unless our relationship is used to remove the dependence on degree from closeness. The success of our relationship suggests that most networks can be approximated by shortest-path spanning trees which are all statistically similar two or more steps away from their root nodes.

---

*Corresponding author. `T.Evans@imperial.ac.uk`

Introduction

Network science has proved to be an exceptionally useful tool, especially with the large scale data sets now available in every discipline. Many systems are driven by pairwise interactions at the microscopic level and the network formalism is ideal as it represents these bilateral relationships as edges between nodes. The real power of network analysis is that it gives us tools to look at the emergent behaviour such system even when driven by interaction's on meso- and macro-scopic scales [1].

A good example is the key task of finding the most important nodes in the system. In network science this is done by using a centrality measure which assigns nodes a centrality value with larger values indicating greater importance. Centrality has been developed over the last seventy years[2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15] and is a core part of many introductory texts on Network Science[16, 17, 19, 18] The simplest centrality measure is degree, the number of neighbours each node has, probably the first node property examined in any study. However, the degree is insensitive to the wider network structure which is the primary goal of network analysis.

In order to probe the broader network structure, many centrality measures are based on the distance between nodes defined as the length of the shortest path between nodes. For instance, our focus is on one of the oldest and most widely used centrality measures, closeness [2], the inverse of the average distance from a node to all others. Closeness is an important indicator in many different contexts. It has been used to investigate the role of different academic disciplines when academics choose their journals and bibliographies [12]. Closeness centrality measures the impact of an author on a field and their social capital [10]. When used to select potential leads in customer data, closeness led to a significant gain in the success rate [9]. In an air transport network, it has been shown that the closeness of a city is highly correlated with socio-economic indicators such as gross regional domestic product [13]. Closeness has been applied to biological networks [6] and closeness measures were able to identify more than 50% of the global regulators within the top 2% of the ranked genes [8]. Essential genes have been found to have higher closeness than nonessential genes in protein-interaction networks [7]. In a metabolic network the closeness of nodes can identify the most important metabolites [20]. Centrality measures such as closeness are an important tool to analyse many types of data.

However, there are a vast number of centrality indices available, as visualised nicely by Schoch [21, 22], suggesting that there is a lot of redundancy. Many different centrality measures encode similar information as seen in the strong correlations between centrality indices [23, 24, 25, 26, 27, 28, 29, 30, 19, 31, 32, 33]. In particular Pearson correlation coefficients are invariably used which are most sensitive to linear correlations between centrality measures. That is we measure $\rho_{cd} = \langle (c_u - \bar{c})(d_u - \bar{d}) \rangle / (\sigma_c \sigma_d)$ where c_u and d_u are two centrality measures for node u, the average is over all nodes with the means $(\bar{c}, \bar{d})$ and standard deviations (σ_c, σ_d) found using the same ensemble average. There seems to be no clear consensus from these studies other than there are often strong relationships between centrality measures but these vary from network to network. For instance, in the introduction of Schoch et al. [30] it is suggested that "Reported results, however, are often inconsistent with regard to the similarity of centrality indices". Of particular interest here is the conjecture made in the Introduction section of Valente et al. [27] where it is stated that "We expect that measures of degree and closeness centrality will be more highly correlated with each other than with other measures, because they are both based on direct ties." Later, in the discussion of results by Valente et al. [27], the authors conclude that "The amount of correlation between degree, betweenness, closeness, and eigenvector indicates that these measures are distinct, yet conceptually related" and the closeness-degree pair is only the third most correlated pair of centrality indices in their study.

In this paper we will focus on the relationship between closeness centrality and degree.

Our result is that closeness centrality and degree have a non-linear relationship, namely that the inverse of closeness ('farness' [34], a normalised version of 'status' [35, 5, 6]) is linearly dependent on the logarithm of the degree. This explains why linear correlation measures often link degree and closeness centrality but at the same time no general pattern has been seen before. Equally our results suggest that studies based on linear correlation measures may well miss important other features in the landscape of centrality measures.

This non-linear relationship means that measuring closeness gives the same broad information as closeness making the computationally expensive calculation of the latter redundant to a first approximation. To extract useful information from closeness the dependence on degree must first be removed and only then will closeness reveal important and distinct information on individual nodes.

Our starting point is that the shortest paths from any one node to all other nodes can be arranged as a spanning tree. We then conjecture that the branches of this tree are statistically similar, implying that the closeness of a node can only depend on the number of such branches, i.e. the degree of the node. If we also assume that the number of nodes in each branch grows exponentially, we find that the inverse of closeness is linearly dependent on the logarithm of degree centrality. The success of this non-linear relationship in a wide range of real-world networks suggests that both these conjectures are broadly true, an insight that can be exploited in a much wide range of network studies.

Results and Discussion

Theory

General Definitions

For simplicity, we will assume throughout this paper that we are analysing a simple graph $\mathcal{G}$ with just one component. We will denote the degree of each node v as k_v. A path in a network of length ℓ is a sequence of $(\ell + 1)$ distinct nodes such that each consecutive pair of nodes in the path is connected by an edge. We will define the distance between two nodes u and v in a network to be the length of a shortest path between two nodes, denoted here as d_{uv}.

The closeness c_v [2, 16, 17, 19, 18] of a vertex v is then defined to be the inverse of the average distance from v to every other vertex in the graph, so

$$\frac{1}{c_v} = \frac{1}{(N-1)} \sum_{u \in \mathcal{V} \backslash v} d_{uv} \tag{1}$$

where $\mathcal{V}$ is the set of nodes and $N = |\mathcal{V}|$ is the number of nodes. Clearly the closer a vertex v is to other vertices in the network, the larger the closeness. Thus closeness mimics the properties of points in a geometric shape where those points closest to the geometric centre will have highest closeness.

Trees [16, 17, 19, 18] are connected networks with no loops so the number of edges is always one less than the number of nodes. Here we use a spanning tree [36] which is a connected subgraph of the original graph $\mathcal{G}$ containing all the original vertices $\mathcal{V}$ but a subset of $(N-1)$ edges that are just sufficient to keep every node connected to all others. We are also going to work with rooted trees $\mathcal{T}(r)$ in which we have singled out one special node, the root r of the tree.

Estimate of Closeness

We start from the idea that some of the statistical properties of real-world networks may be captured by spanning trees [36]. Here we are interested in closeness which uses the lengths of

shortest paths between nodes so the most useful trees for this work are the shortest-path trees, $\mathcal{T}(r)$, that contain one shortest path from a root node r to each remaining node in the network. As our networks are unweighted, the shortest-path trees always exist and are easily defined as part of a breadth-first search algorithm, see the Supplementary Note 2 on the Shortest-Path Tree Algorithm for more details. Every node can act as a root node so there is at least one shortest-path tree, $\mathcal{T}(r)$, for every node r. These trees are not unique as there can be many shortest paths between a pair of nodes.

Our picture for these shortest-path trees is shown in Figure 1. We start with the observation that close to the root node the structure of these shortest-path trees will vary and in particular, the number of nearest neighbours k_r of the root vertex r will vary. However, as we move further away from the root node, the number of nodes $n_r(\ell)$ at some distance ℓ from root node r grows exponentially with each step in most networks. This is the origin of the small-world effect seen in many networks, where the distance between nodes is typically much smaller than is found in similar size networks that are constrained by Euclidean geometry, such as a regular square grid of streets or a random geometric graph. Regardless of the local context of a root node, the shortest-path trees quickly access a similar set of nodes in the main bulk of the network provided there is no large scale inhomogeneity in the network. Thus we conjecture that the structure and statistical properties of these trees away from the root node are similar for all possible root nodes. The contribution to closeness of each node in the bulk is bigger as they are further from the root and more numerous. So we expect that the largest contributions to closeness always come from the same bulk regions where we can expect statistical similarity.
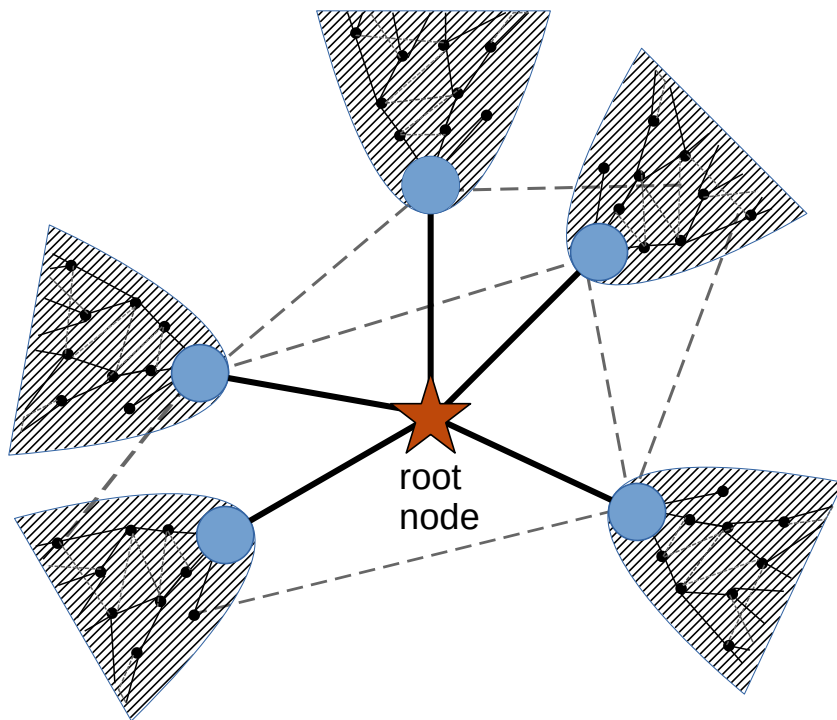

Figure 1: **An illustration of the shortest-path tree approximation.** Every node r, here the red star, is considered to be the root node of a shortest-path tree $\mathcal{T}(r)$. This has k_r nearest neighbours, as indicated by the five solid black lines. Each of the neighbouring nodes, here the blue circles, is treated as the root of a branch of the shortest-path tree. These branches are treated as being statistically identical with a branching number $(1 + \bar{z})$ as indicated here through the use of the same shaded shape rooted on each neighbouring node. The grey dashed lines represent some of the many edges in the graph $\mathcal{G}$ that are not included in the shortest-path tree.

The most important difference when comparing different root nodes therefore comes from

the local structure of each root node. In particular, the initial value for the exponential growth in the number of nodes at distance ℓ from the root will have a large effect. This will depend on the local structure and the simplest effect comes from the number of immediate neighbours the root node has, i.e. the degree of the root node k_r. The simplest approximation for the growth of these shortest-path trees is therefore $n_r(\ell) = k_r \bar{z}^\ell$, where $\bar{z}$ is some measure of the rate of exponential growth of the shortest-path tree. Note that our assumption of statistical similarity suggests that the branching factor of these trees is, on average, the same so we use a single parameter $\bar{z}$ to represent the exponential growth from any root node r.

Our crude approximation is clearly going wrong when we look at large distances from the root as eventually any real network will run out of vertices so $n(\ell) = 0$ for large ℓ. One can model the end of the exponential growth in different ways but we will use the simplest. Namely, we will define a sharp upper cutoff L_r and assume that $n_r(\ell) = 0$ for $\ell > L_r$ which gives

$$N = 1 + \sum_{\ell=1}^{L_r} k_r\, \bar{z}^{\ell-1} = 1 + k_r \frac{(\bar{z}^{L_r} - 1)}{(\bar{z} - 1)}\,. \tag{2}$$

We can invert this expression to express this cutoff L_r in terms of parameters N and k_r, giving us $L_r = L(N, k_r)$ where for large N

$$L(N, k) \approx \frac{\ln\left(N(\bar{z} - 1)/k\right)}{\ln(\bar{z})}\,. \tag{3}$$

Individual distances are integers but it is clear from the form in Eq. (3) that we need L to be a real number, in some sense L is an average over the actual distances from the root to the leaves (nodes with degree one) of the tree. This also tells us that the distance scale associated with a node depends on the logarithm of that node's degree. Since the inverse closeness is a sum of distances, this distance scale L controls the result and that explains why $\ln(k)$ appears in our expression for inverse closeness. As an aside, the $\ln(N)$ dependence of this distance scale L in Eq. (3) reflects the small-world effect seen in most networks.

We now have that $n_r(\ell) = k_r \bar{z}^{\ell-1}$ for $\ell < L(N, k_r)$ and zero for larger ℓ which depends on local parameters k_r, the degree of each root node, and two global parameters, the total number of nodes N and some measure of the growth rate of the shortest-path trees $\bar{z}$. We can now rewrite the closeness c_r Eq. (1) of a vertex r in terms of $n_r(\ell)$ as $1/c_r = (N-1)^{-1} \sum_{\ell=1}^{L_r} \ell n_\ell$ to find that

$$\frac{1}{c_r} = \frac{1}{(N-1)} \sum_{\ell=1}^{L_r} \ell k_r \bar{z}^{\ell-1} = \frac{k}{(N-1)} \left(\frac{(L_r+1)\bar{z}^{L_r}}{\bar{z}-1} - \frac{(\bar{z}^{L_r+1}-1)}{(\bar{z}-1)^2} \right) \tag{4}$$

By using Eq. (2) and Eq. (3) we can eliminate L_r to find that

$$\frac{1}{c_r} = -\frac{1}{\ln(\bar{z})} \ln(k_r) + \beta\,. \tag{5}$$

Our calculation shows the parameter β is also independent of the root vertex r chosen but it is a function of the global network parameters N and $\bar{z}$ so that $\beta = \beta(\bar{z}, N)$ where

$$\beta(\bar{z}, N) = \left(\frac{1}{(\bar{z}-1)} + \frac{\ln(\bar{z}-1)}{\ln(\bar{z})} \right) + \frac{1}{\ln(\bar{z})} \ln(N)\,. \tag{6}$$

Our prediction is that the inverse of closeness c_r of any node r should show a linear dependence on the logarithm of the degree k_r of that node with a slope that is the inverse of the log of the growth parameter $\bar{z}$.

In our analysis we will not assume the parameter β is given by Eq. (6). By adding one additional fitted parameter we lose a little predictive power since many parameters are needed to characterise a network. This leaves us with a conjecture based on the number of nodes N and degree of each nodes k_r which are usually known. Then in principle we have two unknown global parameter values which we find from a linear fit to our data for c_r and k_r giving $\bar{z}^{\text{(fit)}}$ and $\beta^{\text{(fit)}}$.

Numerical Results for Theoretical Models

We looked at the relationship between closeness and degree using simple networks produced from three different theoretical models [17, 18, 19]: the Erdős-Réyni (ER) model [37] the Barabási-Albert model with pure preferential attachment [38] and the configuration model [39] network starting from a network generated with the same Barabási-Albert model. In the first and third model, the edges are completely randomised so there are no vertex-vertex correlations. The last two models both have fat-tailed degree distributions. Our networks built from artificial models were created using standard methods in the `networkx` package [40].

For a single network, we get several nodes with the same degree and we use this variation to find a mean and standard error in the mean shown. The fit is done using Eq. (5) with two free parameters $\bar{z}^{\text{(fit)}}$ and $\beta^{\text{(fit)}}$ and the goodness of fit measures in Table 1 show this is a good fit, confirmed visually by the plots in Figure 2a. Note that a good linear fit of status $((N-1)/c)$ to the logarithm of degree for one example of each of the Erdős-Réyni and Barabási-Albert models was found in Fig. 1 of Wuchty & Stadler [6].

Roughly speaking we find that mean inverse closeness values for any one degree are typically within 2% of the prediction made from the best fit as we can see in Results are shown in Figure 2b. The small deviations seen in Figure 2c, especially for the Config-BA model, are for higher degree values where the data is sparse and uncertainties are large. This means no firm conclusions can be drawn from Figure 2c about the presence of higher order corrections to our form Eq. (5).

Network type	N	$1/\ln(\bar{z}^{\text{(fit)}})$	$\beta^{\text{(fit)}}$	$\bar{z}^{\text{(fit)}}$	$\beta(\bar{z}^{\text{(fit)}}, N)$	$\rho(c,k)$	χ^2_{r}
	1000	0.46 ± 0.01	4.29 ± 0.01	8.87 ± 0.20	4.23 ± 0.03	0.94	1.02
ER	2000	0.42 ± 0.01	4.52 ± 0.01	10.64 ± 0.18	4.28 ± 0.03	0.93	1.02
	4000	0.43 ± 0.01	4.82 ± 0.01	9.99 ± 0.12	4.67 ± 0.02	0.93	1.03
	1000	0.30 ± 0.01	3.59 ± 0.02	28.03 ± 3.11	3.09 ± 0.13	0.75	1.16
BA	2000	0.32 ± 0.01	3.86 ± 0.01	22.76 ± 2.22	3.46 ± 0.07	0.70	1.29
	4000	0.31 ± 0.01	4.03 ± 0.01	25.17 ± 2.61	3.60 ± 0.08	0.65	1.16
	1000	0.35 ± 0.01	3.76 ± 0.02	17.41 ± 1.42	3.46 ± 0.14	0.75	1.19
Config-BA	2000	0.36 ± 0.01	4.01 ± 0.02	16.08 ± 1.24	3.78 ± 0.15	0.70	1.28
	4000	0.35 ± 0.01	4.19 ± 0.01	17.41 ± 1.41	3.94 ± 0.08	0.66	1.19

Table 1: **Results for artificial networks.** Table of results for one example of a simple graph with average degree 10.0 produced using one of three artificial models with the same average degree $\langle k \rangle = 10.0$ but with a different number of nodes, N. Each 'ER' network is a standard Erdős-Réyni network, a 'BA' network is produced using pure preferential attachment in the Barabási-Albert model, and the 'Config-BA' network is a configuration model version of a Barabási-Albert model network. The results for $1/\ln(\bar{z}^{\text{(fit)}})$ and $\beta^{\text{(fit)}}$ come from linear fits of inverse closeness, $1/c_v$ to the logarithm of degree, $\ln(k_v)$ for each vertex v, i.e. $1/c_v = (\ln(\bar{z}))^{-1} \ln(k_v) + \beta$ Eq. (5). The value of β derived from $\bar{z}^{\text{(fit)}}$ and N using Eq. (6) is also shown for comparison. The fits are very good as indicated by the column for the reduced chi-square χ^2_{r}. The Pearson correlation measure between closeness and degree $\rho(c,k)$ is given for comparison with earlier work.

We now turn to look at the actual values obtained from these fits of data on closeness and
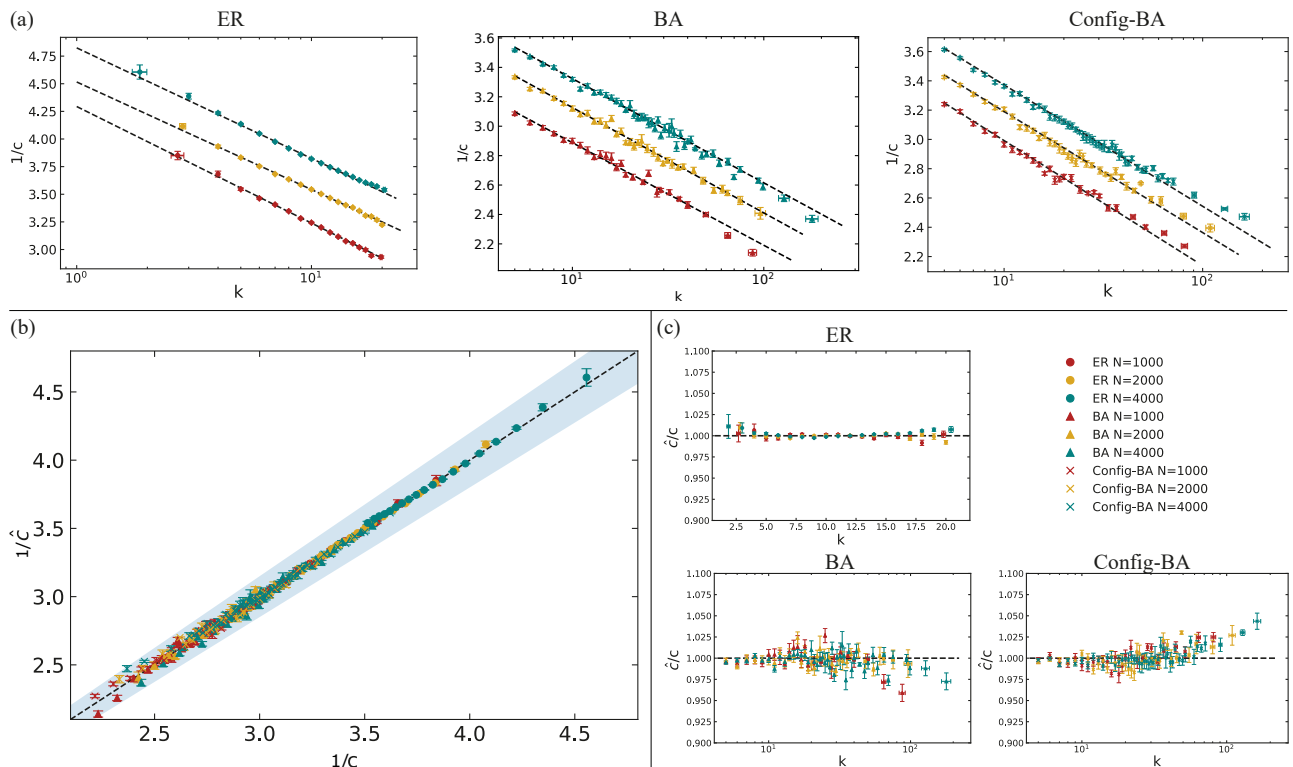
Figure 2: **Closeness and degree for artificial networks.** In panel (a), each plot shows results for networks formed from one artificial model: the Erdős-Réyni (ER) model, the Barabási-Albert (BA) model, and the configuration model network starting from a Barabási-Albert model (Config-BA). The dashed lines shows the best linear fit of $1/c$ to $\ln(k)$ using Eq. (5). The same data from all nine artificial networks is shown in the scatter plot in panel (b) with data $1/\hat{c}$ against predicted value c obtained from the best fit Eq. (5) and the shaded region corresponds to a 5% deviation from the theoretical prediction. Panel (c) shows the fractional error, the fitted value of closeness divided by data value. The results are for three different sized networks: $N = 1000$ (red points) $N = 2000$ (blue points) and $N = 4000$ (yellow points) where N is the number of nodes. All networks have average degree 10.0 and 100 realisations were taken for each case. The values of closeness for each value of degree are binned, the mean is shown as the data point with error bars the standard error of the mean. The results show that the non-linear correlation of closeness and degree predicted in Eq. (5) works most of the time within a 2% variation. There are some hints of small but systematic at higher degree value but the data is sparse and less reliable here.

degree from the artificial networks to Eq. (5). As Table 1 shows there is a small amount of variation in value of $\bar{z}^{(\mathrm{fit})}$, the fit for the shortest-path tree growth factor, with the size of the network. What is of more interest are the differences in values between these three types of artificial networks. All these networks had an average degree of about 10.0 and an infinite tree with constant degree 10 (a Bethe lattice) would have a growth factor $\bar{z} = 9$, one less than the average degree. So the best fit values for the growth factor $\bar{z}$ in the Erdős-Réyni networks are a little higher than this while the Barabási-Albert network and its randomised version are a lot bigger.

Another possible reference value for the shortest-path tree growth factor $\bar{z}$ is the average degree of a neighbour in a random graph with the same degree distribution which is $\langle k \rangle_{\mathrm{nn}} = \langle k^2 \rangle / \langle k \rangle$. This is the relevant value for diffusive processes on a random graph. For our finite Erdős-Réyni networks we have that $\langle k \rangle_{\mathrm{nn}} \approx \langle k \rangle$ so again the growth factor found to give the best fit, $\bar{z}^{(\mathrm{fit})}$, in actual Erdős-Réyni networks is still a bit higher than this estimate. For the Barabási-Albert networks and their randomised versions, the $\langle k \rangle_{\mathrm{nn}}$ is around twenty-two to

twenty-five for the networks in Table 1. This value is much closer but still not in complete agreement. This suggests our shortest-path trees are sampling nodes in a different ways from diffusion but still with a bias to higher degree nodes.

Since spanning trees have many fewer edges than the original graphs, it is perhaps somewhat surprising that we find that the growth factors comparable with any measures of the average degree in the original network. So the high values of $\bar{z}$ are telling us that the shortest-path trees are sampling the nodes of their networks with a large bias towards high degree nodes in the parts of the tree close to the root node and that is why we need such a high growth rate $\bar{z}^{\text{(fit)}}$ when we fit our data for closeness. That way when we prune the edges to produce a tree we will still have high degrees in the tree close to the root node. The corollary is that the outer parts of shortest-path trees are dominated by leaves (degree one nodes) and other low degree nodes, and these also correspond to low degree nodes in the original network.

It is also clear that node correlations play an important role as these are present in the Barabási-Albert model but absent in the randomised version. The large difference in $\bar{z}$ values for these two cases show such node correlations are important and yet, the non-linear relationship Eq. (6) still holds well in these artificial networks, with or without these correlations.

The β parameter in Eq. (5) is harder to interpret but Table 1 shows a comparison between the two values of β. The first is $\beta^{\text{(fit)}}$ derived from a two-parameter fit of the data to Eq. (5). The second value is $\beta(\bar{z}^{\text{(fit)}}, N)$ the value predicted using Eq. (6) where we use the $\bar{z}$ value obtained from the same two parameter fit and the number of nodes N. What we can see is that the values derived using Eq. (6), $\beta(\bar{z}^{\text{(fit)}}, N)$, are consistently poorer than the values $\beta^{\text{(fit)}}$ derived from a two-parameter fit. It highlights that the details of our theoretical form, such as the precise formula for β, here Eq. (6), can be improved. However, our simple calculation has captured the important features of the problem so that the form Eq. (5) does work in these theoretical models provided we treat both $\bar{z}$ and β in Eq. (5) as free parameters to be determined.

Numerical Results for Real Data

For networks representing real-world data, we used data which is open access and easily obtained [41, 42, 43, 44]. We aimed for a wide range of networks both in terms of size and in terms of the type of interaction encoded in these real-world networks.

The first set of eighteen data sets we refer to as the Konect-SNAP networks. These were derived from real-world data and were chosen to reflect five broad categories of network: social networks (`social-...`), communication networks (`commun-...`), citation networks (`citation-...`), co-author networks (`coauth-...`), and hyperlink networks (`hyperlink-...`). A more detailed description of these Konect-SNAP networks is given in Supplementary Note 3 on data sets.

Summary statistics are given in Table 2. The reduced chi-square χ_{r}^2 measure is between 1.05 and 1.61 for ten, more than half, of our examples and another four networks have values between 2.09 and 2.86. Given the wide range of both size and nature of these networks and the simplicity of our theoretical derivation, this level of agreement may not have been expected. We also give the Pearson correlation measure between closeness and degree, $\rho(c, k)$, and this is generally high as has been noted before [23, 24, 25, 26, 27, 28, 29, 30, 31, 32]. The success of our non-linear relationship between closeness and degree is not incompatible with high $\rho(c, k)$ values.

The data for each network is shown in more detail in Figure 3. Again, we can see that within the error bars the average closeness at each degree generally follows the form we predict within 5% when the best fit parameters are used. Further, the uncertainties estimated for these data points suggest that the vast majority of average closeness values are statistically consistent

Network	N	$1/\ln(\bar{z}^{\text{(fit)}})$	$\beta^{\text{(fit)}}$	$\bar{z}^{\text{(fit)}}$	$\beta(\bar{z}^{\text{(fit)}}, N)$	$\rho(c,k)$	χ_{r}^2
social-karate-club	34	0.460 ± 0.066	2.997 ± 0.095	8.81 ± 2.76	2.69 ± 0.25	0.77	1.23
social-jazz	198	0.367 ± 0.015	3.349 ± 0.048	15.28 ± 1.72	2.98 ± 0.08	0.86	13.15
social-hamster	1788	0.353 ± 0.009	4.129 ± 0.020	17.05 ± 1.22	3.68 ± 0.07	0.68	1.20
social-oz	217	0.403 ± 0.010	3.492 ± 0.038	11.96 ± 1.02	3.22 ± 0.08	0.89	2.99
social-highschool	70	0.561 ± 0.039	3.734 ± 0.079	5.95 ± 0.74	3.48 ± 0.17	0.87	1.25
social-health	2539	0.537 ± 0.008	5.605 ± 0.016	6.43 ± 0.17	5.31 ± 0.06	0.75	1.07
commun-email	1133	0.394 ± 0.007	4.309 ± 0.014	12.64 ± 0.54	3.83 ± 0.05	0.84	1.06
commun-UC-message	1893	0.264 ± 0.003	3.526 ± 0.008	43.92 ± 2.16	3.01 ± 0.03	0.72	2.28
commun-EU(core)-email	986	0.259 ± 0.004	3.324 ± 0.012	47.63 ± 2.67	2.80 ± 0.03	0.84	2.48
commun-DNC-email	1833	0.222 ± 0.010	3.499 ± 0.012	91.16 ± 18.84	2.67 ± 0.08	0.41	1.38
commun-DIGG-reply	29652	0.388 ± 0.002	5.078 ± 0.003	13.12 ± 0.18	5.05 ± 0.02	0.61	1.60
citation-DBLP	12494	0.361 ± 0.003	4.856 ± 0.004	15.98 ± 0.35	4.45 ± 0.03	0.54	1.17
citation-Cora	23166	0.503 ± 0.004	6.639 ± 0.008	7.31 ± 0.13	6.14 ± 0.04	0.48	1.10
coauthor-astro-ph	14845	0.441 ± 0.004	5.735 ± 0.010	9.67 ± 0.21	5.30 ± 0.04	0.61	1.07
coauthor-netsci	379	0.382 ± 0.080	6.553 ± 0.119	13.74 ± 7.51	3.31 ± 0.48	0.35	1.27
coauthor-pajek	6927	0.259 ± 0.002	3.894 ± 0.002	47.45 ± 1.33	3.30 ± 0.02	0.64	12.56
hyperlink-polblog	1222	0.240 ± 0.004	3.316 ± 0.011	64.84 ± 4.44	2.72 ± 0.03	0.72	2.15
hyperlink-blogs	1222	0.239 ± 0.004	3.316 ± 0.011	65.07 ± 4.48	2.71 ± 0.03	0.72	2.18

Table 2: **Results for the largest components the Konect-SNAP networks.** These networks are derived from a variety of derived from real-world data, see the description of Konect-SNAP networks in Supplementary Note 3 on data sets for more details. The results for $1/\ln(\bar{z}^{\text{(fit)}})$ and $\beta^{\text{(fit)}}$ come from linear fits of inverse closeness, $1/c_v$ to the logarithm of degree, $\ln(k_v)$ for each vertex v, i.e. $1/c_v = (\ln(\bar{z}))^{-1} \ln(k_v) + \beta$ Eq. (5). The value of β derived from $\bar{z}^{\text{(fit)}}$ and N using Eq. (6) is also shown as $\beta(\bar{z}^{\text{(fit)}}, N)$ for comparison. The fits are very good as reduced chi-square χ_{r}^2 values show. The Pearson correlation measure between closeness and degree $\rho(c,k)$ is given for comparison with earlier work.

with the predicted value for that degree, something already captured by the reduced chi-square values in Table 2.

Finally we looked at our relationship between degree and closeness in 112 networks taken from the Netzschleuder archive of 276 network data sets [44]. Our only selection criterion was that we could automatically download a network and that it could be analysed successfully by our standard code without further work. This excluded several examples in this archive such as those with multiple networks (e.g. amazon_copurchases), or some which were too large for our code (e.g. academia_edu). Supplementary Note 3 contains further information on these Netzschleuder networks. As a result, our sample contains many for which we would not expect much success: some are very small, some are very dense, some have add additional known structure e.g. a bipartite network. At the same time there are many for which we would expect to be successful, typically anything sparse and large. We represented each as a simple graph and analysed the largest connected component which had up to 40,000 nodes and an average degree of less than 300.

The results are shown in Figure 4 with additional information and a table of results on the Netzschleuder networks provided in Supplementary Note 4. We found 50 networks had an excellent fit with a reduced chi-square χ_r^2 close to 1.0 and always less than 2.0. Another 13 networks gave a reasonable fit $2.0 \leq \chi_r^2 < 3.0$ and 9 networks had $3.0 \leq \chi_r^2 < 4.0$. For the remaining networks, our code did not find a χ_r^2 for 13 of these as there was at most one node for each degree value and our code could not estimate the uncertainty in the measurement of closeness. Overall, of the 99 networks where we had a χ_r^2 result, our degree-closeness relationship was very successful ($\chi_r^2 < 2.0$) in these arbitrary networks 50% of the time.

However, from Figure 4 we can see that the density of a network, the number of edges divided by the number of node pairs, has an impact on the success rate. Our sample has
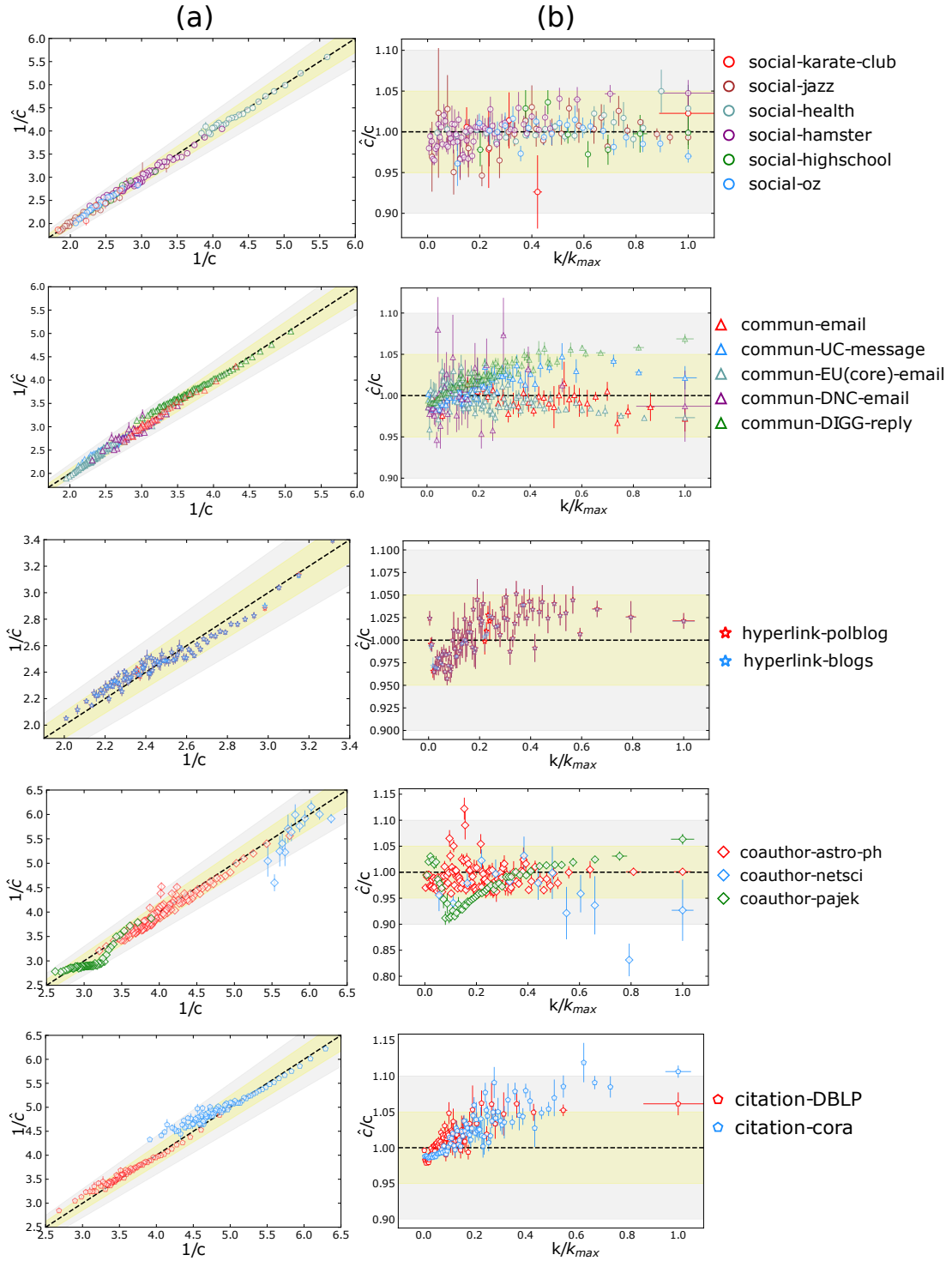
Figure 3: **Closeness and degree for Konect-SNAP networks.** Results for the largest components of eighteen Konect-SNAP networks derived from real-world data, see Table 2 for the statistics of each dataset. The yellow shaded region corresponds to 5% deviation and grey region corresponds to a 10% deviation. Plots in column (a) show the inverse of the predicted result $1/\hat{c}$ from the best fit against the inverse of the mean measured value $1/c = 1/\langle c \rangle_k$ averaged over nodes with the same degree k. Both axes are essentially $\ln(k)$. Column (b) shows the ratio of best fit value $\hat{c}$ over measured value c as a function of degree k rescaled by the largest degree in each network $k_{\max}$. If the prediction matched data perfectly, points will lie on the dashed lines. The error bars represent from standard error of mean of the inverse closeness. For majority of points, we can see our prediction Eq. (5) captures the relation between closeness and degree, usually with within a 5% margin.
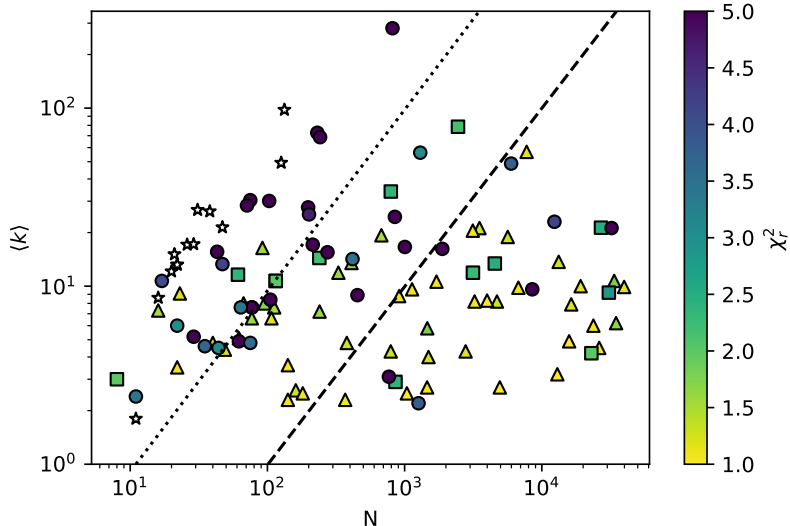
Figure 4: **Closeness and degree for Netzschleuder real-world networks.** Results for the largest components of an additional 112 networks taken from the Netzschleuder archive [44]. Each point represents one network, plotted using the number of nodes N and the average degree $\langle k \rangle$ as coordinates. The symbol colour indicates reduced chi-square (χ_r^2) value with yellow being the best results (closest to 1.0). The symbol shape also indicates χ_r^2 values with triangles representing best results where $\chi_r^2 < 2.0$, squares are for $2.0 \leq \chi_r^2 < 3.0$ and circles represent poor performance with $3.0 \leq \chi_r^2$. The stars (white fill) are networks where our χ_r^2 calculation failed due to a lack of an error estimate on each point which occurs when we have one node per k value. To the right (left) of the dashed (dotted) line, networks have a density $\langle k \rangle/(N-1)$ of less than 0.01 (more than 0.1).

included several very dense networks where there was a lot of additional information on the edges so the existence of an edge was much less important. Indeed from Table 3 we see that for Netzschleuder networks with densities below 0.04 we have around 70% fitting with a reduced chi-square of less than 2.0, rising to roughly 80% for $\chi_r^2 < 3.0$.

Upper Density	Number Networks	% Networks $\chi_r^2 < 2.0$	% Networks $\chi_r^2 < 3.0$
0.01	32	66%	84%
0.02	44	70%	84%
0.03	48	69%	81%
0.04	55	67%	80%

Table 3: **The success rate for sparse Netzschleuder networks.** These results are for sparse networks taken from the 112 Netzschleuder networks analysed in this paper. Each row gives the statistics for all networks whose largest connected component has a density less than or equal to the density given. The percentage of these networks whose reduced chi-square is less than two or less than three is given in the last two columns.

Using Closeness

Our numerical results confirm our analytical work that the inverse of closeness depends linearly on the logarithm of degree Eq. (5) for most networks. This is a correlation, true on average but not an exact result for every node. It has long been known that nodes with larger degrees

tend to have smaller closeness which leads to significant correlation measures [23, 24, 25, 26, 27, 28, 29, 30, 19, 31, 32, 33] but this is often discussed in terms of the Pearson correlation coefficient which is most sensitive to linear correlations. A non-linear relationship between degree and closeness is not discussed in previous studies and our relationship puts this intuitive understanding of a degree-closeness relationship on a firm footing.

Our work suggests that in the majority of networks, to a first approximation, closeness of individual nodes captures little more information on average than is contained in the degree. There is no point spending time calculating closeness if you only want a rough measure of centrality, you may as well just use degree. However, closeness measurements can give useful information on a network if used correctly.

First, closeness and degree measurements yield our fitted parameters, $\beta^{(\text{fit})}$ and $\bar{z}^{(\text{fit})}$. These characterise each network and can be used to compare different types of networks and even networks of different sizes. In particular, $\bar{z}^{(\text{fit})}$ characterises the exponential growth that is a feature of most complex networks and which is behind many well-known phenomena as we shall note later in this discussion.

There are a number of ways to determine $\bar{z}$ and β from our relation Eq. (5). We have found the most effective approach is the simpler method where we determined $\bar{z}$ and β from a linear fit of our data to Eq. (5). There are alternatives but these throw light on various possible approximations rather than being of practical use, see the sections on determining $\bar{z}$ and β in Supplementary Note 1 and results for higher-order polynomial fits in Supplementary Note 4.

For networks where degree and closeness are linked by our relationship Eq. (5), our work shows that useful information from closeness centrality for individual nodes can only come by comparing values for individual nodes against the expected value derived using our relationship Eq. (5) using the logarithm of degree. The only useful information in closeness values is the deviation from their expected value. In these situations, we could start by examining the degree centrality of every node. This would be the primary measure of centrality. We then fit our closeness values using Eq. (5) to produce an expected value of closeness $c_v^{(\text{fit})}$ for each node. Finally, we use this fit to find nodes which are noticeably more (or less) central than expected. One way to do this is to look at the normalised closeness

$$c_v^{(\text{norm})} = \frac{c_v}{c_v^{(\text{fit})}} \, . \tag{7}$$

Our normalised closeness measure $c_v^{(\text{norm})}$ will highlight the outliers which would then be of most interest. See Supplementary Note 4 to see some examples of the fluctuations in closeness values around our predicted value.

One could also compare closeness to degree by running the configuration model, which keeps node degree constant, and measuring the closeness of each node in such a null-model network. However our method, requiring a simple linear fit to data already acquired, will be much faster than running the configuration model.

General Network Insights

The success of our conjecture also suggests that most networks satisfy two key assumptions built into our derivation.

First we assumed that the number of nodes a distance ℓ from any node grows exponentially. Such exponential growth is common in networks as it is the mechanism behind the concepts of the "six degrees of separation" and the "small world" effect [45] often reported in networks. More formally, this is linked to length scales in networks models with N vertices which grow as $O(\ln(N))$. This is to be contrasted with a network controlled by the geometry of a d-dimensional Euclidean space (such as Random Geometric Graphs) the number of nodes at distance ℓ from

a node grows as a power law ℓ^{d-1} and length scales in such networks grow as $O(N^{1/d})$. We will consider the length scales in more detail in the Network Length Scales subsection below.

This non-Euclidean behaviour of most networks highlights one situation where closeness is a useful centrality measure independent of and uncorrelated with degree, so a situation where our relationship Eq. (5) will fail. That is for graphs embedded in Euclidean space. This is somewhat ironic as this type of planar network, such as a network representing the connections between intersections of streets in a city, is the prototypical example used to motivate the idea that the network measure closeness is related to our intuition about the concept of centrality in a network. Indeed, Bavelas [2] only used planar graph examples to develop closeness so a link between closeness and degree in most networks was never an issue in the original motivation for the closeness measure.

The second assumption that our work supports is that the branches of the shortest-path trees are statistically similar as illustrated in Figure 1. The success of our analysis suggests this assumption works well whenever we are looking at measurements that depend on the bulk of the network. This simple approximation may therefore help analyse other network measurements, and we consider some of these in the Network Length Scales subsection below.

Closeness and Real-World Networks

Given the simple assumptions and approximations used in deriving our relationship Eq. (5), it is perhaps surprising to find that this closeness-degree relation works well for so many networks based on real world data. So how often is our relationship a success? Also, can we understand when our closeness-degree relationship may succeed and when it is likely to fail.

For the first set of eighteen Konect-SNAP networks shown in Figure 3 and Table 2, we had eleven (61%) with a reduced chi-square of less than 2.0 and all but two (89%) have a reduced chi-square with less than 3.0. The success rate for the 112 Netzschleuder networks as measured by reduced chi-square is lower. Of the ninety-nine Netzschleuder networks where we had a reduced chi-square measurement, fifty (51%) had a reduced chi-square of less than 2.0 while sixty-three (64%) had a a reduced chi-square of below 3.0. The lower success rate for the Netzschleuder networks can be understood as we are sampling without bias a varied collection of network data sets and, in many cases, we have information on the networks which suggests we might never have expected success.

The simplest issue is network density, the fraction of possible edges that are actually present in a network. If the density is very high, the distances between nodes will be low, perhaps just one or two. This leads to very small variations in the distances encoded in the closeness of each node making it harder to distinguish meaningful patterns in closeness measurements so our relationship is more likely to struggle with high density networks. This trend is clear in Figure 4. Of the two Konect-SNAP networks with poor chi-square, one, the `social-jazz` network, has a high density of 0.13. Conversely, we find that the success rate is much higher in our Netzschleuder networks for networks with density below 0.04: roughly two-thirds of these low density networks fit our degree-closeness relationship well with $\chi_r^2 < 2.0$ and this rises to 80% if we accept anything with $\chi_r^2 < 3.0$, see Table 3.

More generally, a failure of our relationship is a powerful tool to highlight where there is additional structure in the network. Our theoretical analysis assumes a generic network where every node sees the same exponential growth in the number of nodes at distance ℓ away as discussed in the General Network Insights subsection. If this generic structure is not present, we have no reason to expect our relationship to work well.

To see this structural issue, consider some of the notable exceptions to our low density criterion. Looking at Figure 4, the low density networks with N between 700 and 2000 and $\langle k \rangle$ between 2.0 and 3.5, you can see two good networks with good fits (triangles) and three

with poor fits (two circles and a square). One circle comes from the `crime` network with $N = 1263$, $\langle k \rangle = 2.2$ for which we find a reduced chi-square of $\chi_r^2 = 3.8$. The description of the `crime` network mentions that the data was obtained using "snowball sampling from five initial homicides" so we are looking at a biassed sample of a much larger network. We suspect this sampling produces a different structure from a typical complex network leading to a failure of our relationship in this case. The other circle is the `plant_pol_kato` network with $N = 768$, $\langle k \rangle = 3.1$ and $\chi_r^2 = 34.1$ which is a "bipartite network of plants and pollinators" from a forest. Our analysis does not allow for a bipartite structure and in this case the difference between the two types of nodes is large with 91 plant species nodes but 715 nodes representing species of insects. The square represents the `unicodelang` network with $N = 858$, $\langle k \rangle = 2.9$ and $\chi_r^2 = 2.4$. This is still a good fit but it is also a "bipartite network of languages and the countries" so again, a better result might be achieved if we adapted our approach for two-mode networks.

In some cases, the meta-data we have for a network may already tell us about large-scale structure, such as a bipartite network, that will invalidate our relationship Eq. (5). In simple cases, it may be possible to adjust our derivation to find a more appropriate relation that matches the known structure. For instance, for bipartite networks we could represent our shortest-path trees using two growth rate parameters, $\bar{z}_a$ and $\bar{z}_b$, for odd and even distances from the root node.

In other cases failure may indicate the presence of structure that is not already known. For example, the network may have strong inhomogeneities such as high degree nodes clustering together in a dense core. That would invalidate our assumption that all the branches of the shortest-path trees look statistically similar. We can see this type of problem in discussions of the average path length in random graph models [46, 47]. There it was noted that the typical $\ln(N)$ dependence of length scales in complex networks is not seen when these models have a degree distribution of the form $p(k) \sim k^{-\gamma}$ with $2 < \gamma < 3$. In these cases this failure is linked to the network taking on a rather different structure, perhaps an "octopus" [46], a star-like structure with one dense core connected to many legs. This type of network invalidates our assumptions and we would not expect to see our closeness-degree relationship.

Network Length Scales

We can compare our analytical approach to that of other theoretical work on length scales in networks. The vast majority of the analytical results on distance in networks is for well defined simple theoretical models, typically the Erdős-Réyni graph [46, 48, 49, 17], the Barabási-Albert model [50, 51], or scale-free random graphs [52] (our BA-Configuration model). The focus in the literature tends to be on characteristic length scales for a graph, such as the average path length $\langle \ell \rangle$ or the diameter, and not on length scales associated with each vertex, such as closeness that we consider.

What is distinctive about our approach is that we focus on generic network properties that appear to hold in many networks. So our approach can be used on a much wider range of networks and in particular work for real-world networks not just for one simple model. The downside is that we do not have the mathematical rigour of those working with simple models. So, let us consider how our approach can be related to other length scale calculations in the literature.

We have already noted that the concepts of the "six degrees of separation" and the "small world" effect [45] are linked to a network models where the length scale grows as $O(\ln(N))$, much slower that for $O(N^{1/d})$ behaviour expected for networks embedded in d-dimensional Euclidean space. One common way to study this more precisely is to consider the average distance between nodes $\langle \ell \rangle$, where each node pair contributes equally to the average. This is what we obtain if we take half the average of the inverse closeness Eq. (1) over all vertices,

$\langle \ell \rangle = (2N)^{-1} \sum_r (c_r)^{-1}$. From our results for closeness, we can see that the N dependence in $\langle \ell \rangle$ comes from the $\ln(N)$ term in the expression for β of Eq. (6), see the section on average shortest path length in Supplementary Note 1 for more details. So our result is consistent with the behaviour found more rigorously for the average path length in random graph models [46, 47]. The success of our method in terms of this average path length result shows that small world behaviour can be linked to the simple network features built into our method, those discussed in the General Network Insights subsection.

Our theoretical cutoff $L(N, k)$ of Eq. (3) gives another length scale for each node and again this has the same small world $\ln(N)$ behaviour. One might conjecture that this cutoff length scale $L(N, k)$ could be linked to the eccentricity length scale. The eccentricity [16, 6] e_r of a node r is the largest distance from r to any other node, $e_r = \max\{d_{rv} | v \in \mathcal{V}\}$. In that case we are making a new conjecture that eccentricity should also be linearly dependent of $\ln(k)$ with a gradient of $-1/\ln(\bar{z}^{(\mathrm{fit})})$ to match our theoretical cutoff $L(N, k)$. As seen in Fig. 1 of Wuchty & Stadler [6] and confirmed in our own analysis, eccentricity does depend linearly on $\ln(k)$ but the gradient does not seem to match the prediction from our theoretical cutoff $L(N, k)$. Roughly speaking our cutoff $L(N, k)$ represents a typical average large scale, not an extremal value of a distribution, so we should never expect a close link with eccentricity. For more details see the section on eccentricity and $L(N, k)$ in Supplementary Note 1.

As the largest eccentricity is the diameter of a graph, we can see the $O(\ln(N))$ behaviour of our $L(N, k)$ expression as matching this behaviour seen analytically in the diameter in simple models. However, just as for eccentricity and $L(N, k)$, we don't expect to be able to get a precise handle on diameter in our approach.

We also see strong similarities between our approach and that used in several papers[52, 48] where the network is reduced to a set of rings of nodes, each ring containing all the nodes at the same distance ℓ from a root node. For instance, the mean first-passage time of random walkers on a network to a given vertex v is just the inverse of closeness where now closeness is defined in terms of a new distance function d_{uv} in Eq. (1), where d_{uv} is the average first-passage time for a random walker to move from vertex u to vertex v (this is also known as Markov centrality [53] or Random walk closeness centrality). Mean first-passage time has been observed to be proportional to degree [48] and applying our simple approximations to the "ring" method of Baronchelli & Loreto [48] quickly reveals this feature, see the section on the ring calculations of first passage times in Supplementary Note 1 for more details.

Overall, our approach can give insights in the the behaviour of many network lengths scales, sometimes only very roughly, sometimes with more precision. It can never match the precision of the analytic calculations done for the simplest models but our approach can be used in a much broader range of networks.

Improvements

We have already noted one simple improvement when working with bipartite networks. That is to use two growth rate parameters, $\bar{z}_a$ and $\bar{z}_b$ for odd and even distances from the root node in the shortest-path tree of bipartite networks.

Other extensions are suggested by probing our numerical results in more detail. For our smaller set of eighteen Konect-SNAP networks, beyond the two poorly fitting cases, our relationship Eq. (5) is very successful for most individual nodes within a 5% margin, a success which may not be expected given the simple analytical derivation. However we can see some clear if small trends in the deviations in Figure 3. We suggest these trends highlight the limitations of our analytical approach but it is possible to improve our theoretical methods.

At the simplest level, we could replace the sharp cutoff used for $n_\ell(r)$ where $n_\ell(r) = 0$ for $\ell > L$. There are examples of these distributions for some simple models in Baronchelli

& Loreto [48]. A better cutoff may well lead to better predictions for β allowing one to fit a function with one independent parameter rather than two that we used by keeping β as a fitted parameter. However, while fitting one rather than two parameters may be theoretically satisfying, it does not seem much of a gain for analysis of real-world networks.

Another option might be to calculate a different network parameter, namely the second degree $k_r^{(2)} = n_{\ell=2}(r)$ [54] for each node r. By finding the number of nodes two steps away from every node, we can make a better approximation for $n_\ell(r)$, that is $n_0(r) = 1$, $n_1(r) = k_r$, and $n_\ell(r) = k_r^{(2)} \bar{z}^{\ell-2}$ for $2 \leq \ell \leq L_r$ and $n_\ell(r) = 0$ for $\ell > L_r$. This approach cannot be worse than the method used here as the latter is included as a special case where the second degree $k_r^{(2)} = \bar{z} k_r$ for all nodes r. To leading order we get the same type of result, namely that $1/c_r = (\bar{z})^{-1} \ln(k_r^{(2)}) + \beta$ since the degree k_r now only contributes a small number of terms to closeness. So in this approach using second degree we need to measure a different set of N parameters, the second degree of each node. Finding second degree is slower numerically than degree but both scale in the same way with increasing network size. The success of our simpler method here points to the idea that second degree and degree may often be correlated so it is likely that using second degree may only enhance results in a few cases.

More serious changes will be needed to the calculation if other effects neglected here, such as community structure or degree assortativity, are to be included.

Distance and Logarithm of Degree

The logarithm of degree $\ln(k)$ has been found to play an important role in network analysis before. A large fraction of papers on networks will show degree distributions where the horizontal axis is the dependent variable $\ln(k)$ and not simply the degree k. A more specific example comes from Zhou et al. [55] where the ratio of the degrees of nodes at the two ends of each edge (largest value in the numerator) is used to assign a 'distance' $\eta(u,v)$ to each edge (u,v). This is equivalent to defining $\lambda(u,v) = \ln(\eta(u,v)) = |\ln(k_u) - \ln(k_v)|$. In fact one can quickly see that while both η and λ are semi-distances on the set of edges in the formal mathematical sense, only λ is also a semi-metric and so λ is in some sense the more natural 'distance' measure in a qualitative sense. Our work suggests that an alternative view is to replace the logarithm of degree by the inverse of closeness. Since $(c_u)^{-1}$ is the actual average of the shortest-path distances from u to all other nodes, we can immediately see it is natural to work with inverse closeness when considering distances. For instance Zhou et al. [55] we could look at a different edge measure $\tilde{\lambda}(u,v) = |(c_u)^{-1} - (c_v)^{-1}|$. While the inverse closeness is a more natural distance, the degree is much easier to calculate in practice. Our work allows researchers to move between these two pictures.

Closeness and Gromov Centrality

It has been noted [57] that the inverse of closeness is related to another centrality measure based on the average of the Gromov product. This centrality measure captures the extent to which the triangle inequality is not saturated between three nodes in a network, and so this is deeply connected to the geometry of a network. Our result leads to a natural prediction for this Gromov Centrality G_r^D of node r defined on the scale of the network's diameter D [57]. This Gromov Centrality is defined on other network length scales, G_v^ℓ, and Babul et al. [57] suggest there are useful generalisation of closeness. Our approximations will prove just as effective for such generalisations of closeness.

Conclusions

We have derived a non-linear relationship between degree and closeness Eq. (5), two of the most important centrality measures in Network Science. This was achieved by assuming that every node can be seen as the root of a shortest-path spanning tree where for all nodes these trees are statistically similar two or more steps away from the root. We have shown that this degree-closeness relationship works on a large number of artificial and real-world networks, particularly low density networks where there is no special macroscopic structure.

Our results mean that measuring closeness is of little use unless our relationship is used to remove the dependence on degree from closeness. Further, most networks can be seen as shortest-path spanning trees which are statistically similar two or more steps away from their root nodes. Equally, if our relationship fails, it is an indication that our basic assumptions about network structure are wrong and so there are strong inhomogeneities and macroscopic structure in the data.

Acknowledgements

Data Availability

All the data used in this paper is publicly available from several repositories. Our copies were downloaded from one of the following: KONECT [42], SNAP [43], and the Netzschleuder [44] repositories. More detailed information on some of the networks used along with additional references are also given in Supplementary Note 3 on data sets. The networks used in this study along with tables of many of the results are available in [56].

Author Contributions

TSE performed the analytical research and BC performed the numerical research. TSE and BC designed the research, analysed data and wrote the paper.

Competing Interests

The authors declare no competing interests.

References

[1] Brandes, U., Robins, G., McCranie, A. & Wasserman, S. What is network science? *Network Science* **1**, 1–15 (2013). DOI http://dx.doi.org/10.1017/nws.2013.2.

[2] Bavelas, A. Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America* **22**, 725–730 (1950).

[3] Sabidussi, G. The centrality index of a graph. *Psychometrika* **31**, 581–603 (1966).

[4] Freeman, L. C. Centrality in social networks conceptual clarification. *Social networks* **1**, 215–239 (1978).

[5] Hage, P. & Harary, F. Eccentricity and centrality in networks. *Social Networks* **17**, 57–63 (1995).

[6] Wuchty, S. & Stadler, P. F. Centers of complex networks. *Journal of Theoretical Biology* **223**, 45–53 (2003).

[7] Hahn, M. W. & Kern, A. D. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution* **22**, 803–806 (2004).

[8] Koschützki, D. & Schreiber, F. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regulation and Systems Biology* **2**, GRSB.S702 (2008).

[9] Kiss, C. & Bichler, M. Identification of influencers — measuring influence in customer networks. *Decision Support Systems* **46**, 233–253 (2008).

[10] Yan, E. & Ding, Y. Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology* **60**, 2107–2118 (2009).

[11] Landherr, A., Friedl, B. & Heidemann, J. A critical review of centrality measures in social networks. *Business & Information Systems Engineering* **2**, 371–385 (2010).

[12] Ni, C., Sugimoto, C. & Jiang, J. Degree, closeness, and betweenness: Application of group centrality measurements to explore macro-disciplinary evolution diachronically. In *Proceedings of ISSI*, 1–13 (2011).

[13] Wang, J., Mo, H., Wang, F. & Jin, F. Exploring the network structure and nodal centrality of China's air transport network: A complex network approach. *Journal of Transport Geography* **19**, 712–721 (2011).

[14] Brandes, U. & Hildenbrand, J. Smallest graphs with distinct singleton centers. *Network Science* **2**, 416–418 (2014). DOI `http://dx.doi.org/10.1017/nws.2014.25`.

[15] Das, K., Samanta, S. & Pal, M. Study on centrality measures in social networks: a survey. *Social Network Analysis and Mining* **8** (2018).

[16] Wasserman, S. & Faust, K. *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)* (Cambridge University Press, 1994).

[17] Newman, M. *Networks: an introduction* (Oxford University Press, 2010).

[18] Coscia, M. *The Atlas for the Aspiring Network Scientist* (Michele Coscia, 2021). `arXiv:2101.00863`.

[19] Latora, V., Nicosia, V. & Russo, G. *Complex Networks: Principles, Methods and Applications* (Cambridge University Press, 2017).

[20] Ma, H.-W. & Zeng, A.-P. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* **19**, 1423–1430 (2003).

[21] Schoch, D. *A Positional Approach for Network Centrality*. Ph.D. thesis, Universität Konstanz (2015).

[22] Schoch, D. Periodic table of network centrality (2016). URL http://schochastics.net/sna/periodic.html.

[23] Bolland, J. M. Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks. *Social networks* **10**, 233–253 (1988).

[24] Rothenberg, R. B. *et al.* Choosing a centrality measure: epidemiologic correlates in the colorado springs study of social networks. *Social Networks* **17**, 273–297 (1995).

[25] Faust, K. Centrality in affiliation networks. *Social networks* **19**, 157–191 (1997).

[26] Lee, C.-Y. Correlations among centrality measures in complex networks. *arXiv preprint physics/0605220* (2006).

[27] Valente, T. W., Coronges, K., Lakon, C. & Costenbader, E. How correlated are network centrality measures? *Connections (Toronto, Ont.)* **28**, 16 (2008). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875682/.

[28] Batool, K. & Niazi, M. A. Towards a methodology for validation of centrality measures in complex networks. *PloS one* **9**, e90283 (2014).

[29] Lozares, C., López-Roldán, P., Bolibar, M. & Muntanyola, D. The structure of global centrality measures. *International Journal of Social Research Methodology* **18**, 209–226 (2015).

[30] Schoch, D., Valente, T. W. & Brandes, U. Correlations among centrality indices and a class of uniquely ranked graphs. *Social Networks* **50**, 46–54 (2017).

[31] Oldham, S. *et al.* Consistency and differences between centrality measures across distinct classes of networks. *PLOS ONE* **14**, e0220061 (2019).

[32] Bringmann, L. F. *et al.* What do centrality measures measure in psychological networks? *Journal of Abnormal Psychology* **128**, 892–903 (2019).

[33] Arnaudon, A., Peach, R. L. & Barahona, M. Scale-dependent measure of network centrality from diffusion dynamics. *Physical Review Research* **2**, 033104 (2020).

[34] Opsahl, T., Agneessens, F. & Skvoretz, J. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks* **32**, 245–251 (2010).

[35] Harary, F. Status and contrastatus. *Sociometry* **22**, 23 (1959).

[36] Šubelj, L. Algorithms for spanning trees of unweighted networks. Tech. Rep., University of Ljubljana (2021). `arXiv:2205.06628`.

[37] Erdős, P. & Réyni, A. On random graphs. i. *Publicationes Mathematicae* **6**, 290–297 (1959).

[38] Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 173 (1999).

[39] Molloy, M. & Reed, B. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms* **6**, 161–180 (1995).

[40] Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T. & Millman, J. (eds.) *Proceedings of the 7th Python in Science Conference (SciPy2008)*, 11–15 (2008).

[41] Batagelj, V. Pajek datasets. `http://vlado.fmf.uni-lj.si/pub/networks/data/` (2017).

[42] Kunegis, J. The KONECT project. `http://konect.cc/`.

[43] Leskovec, J. & Krevl, A. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data` (2014).

[44] Peixoto, T. P. The Netzschleuder network catalogue and repository. URL `https://networks.skewed.de/`.

[45] Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998). DOI `http://dx.doi.org/10.1038/30918`.

[46] Chung, F. & Lu, L. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences* **99**, 15879–15882 (2002).

[47] Dorogovtsev, S. N., Mendes, J. F. F. & Samukhin, A. N. Metric structure of random networks. *Nuclear Physics B* **653**, 307–338 (2003).

[48] Baronchelli, A. & Loreto, V. Ring structures and mean first passage time in networks. *Physical Review E* **73**, 026103 (2006).

[49] Blondel, V. D., Guillaume, J.-L., Hendrickx, J. M. & Jungers, R. M. Distance distribution in random graphs and application to network exploration. *Physical Review E* **76**, 066101 (2007).

[50] Bollobás, B. & Riordan, O. The diameter of a scale-free random graph. *Combinatorica* **24**, 5–34 (2004).

[51] Evans, T., Calmon, L. & Vasiliauskaite, V. The longest path in the Price model. *Scientific Reports* **10**, 10503 (2020). DOI `https://doi.org/10.1038/s41598-020-67421-8`.

[52] Cohen, R. & Havlin, S. Scale-free networks are ultrasmall. *Physical Review Letters* **90**, 058701 (2003).

[53] White, S. & Smyth, P. Algorithms for estimating relative importance in networks. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '03* (ACM Press).

[54] Falkenberg, M. *et al.* Identifying time dependence in network growth. *Physical Review Research* **2**, 023352 (2020). DOI `https://doi.org/10.1103/PhysRevResearch.2.023352`.

[55] Zhou, B., Meng, X. & Stanley, H. E. Power-law distribution of degree-degree distance: A better representation of the scale-free property of complex networks. *Proceedings of the National Academy of Sciences* **117**, 14812–14818 (2020).

[56] Evans, T. S. & Chen, B. Linking the network centrality measures closeness and degree: Additional data. figshare. DOI `https://doi.org/10.6084/m9.figshare.19216812`.

[57] Babul, S'A., Devriendt, K. & Lambiotte, R. Gromov Centrality: A Multi-Scale Measure of Network Centrality Using Triangle Inequality Excess. Tech. Rep., Oxford University (2022). `arXiv:2205.04974`.

Supplementary Information for Linking the Network Centrality Measures Closeness and Degree

Tim S. Evans, Bingsheng Chen

Centre for Complexity Science, and Theoretical Physics Group, Imperial College London, SW7 2AZ, U.K.
Version 3. 11th May 2022

Contents

Supplementary Note 1: Calculations

In this Supplementary Note, we will give greater detail on the calculations referred to in the main text. As we have used the Zachary karate club network [1] as an example to illustrate some key principles, e.g. in Supplementary Figure 1.3, we start by illustrating the correlation between degree and closeness in this network in Supplementary Figure 1.1.
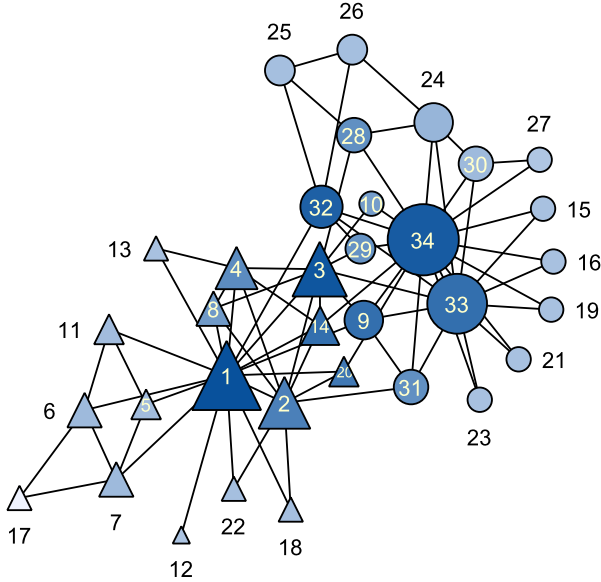
Supplementary Figure 1.1: The Zachary karate club network [1]. The nodes are coloured so that the smaller the closeness, the darker the colour while the size of the node is proportional to the degree. Node labels correspond to those used in [1]. The largest nodes tend to be darkest, the smallest nodes lightest indicating that there is significant correlation between degree and closeness. The node shape indicates the two communities found by Zachary [1] using the Ford-Fulkerson binary community algorithm.

1.1 Estimate of Closeness

In this section we will look at our calculation of closeness in more detail. This enables us to highlight in more detail the approximations and assumptions made, and to show the type of terms we are neglecting in our expansions.

We will work with simple graphs with one component, so every vertex is connected by a path to every other vertex. We will analyse this in terms of TREES (for example see section 4.2.10 of Wasserman & Faust [2], section 6.7 of Newman [3] or section 1.4 of Latora et al. [4]), subgraphs with no closed loops and one component so the number of edges in a tree is one less than the number of vertices in the tree. Our trees $\mathcal{T}(r)$ are ROOTED TREES in that they are defined by starting at a special node, the root node r for that tree. An example is shown in Supplementary Figure 1.2.

We are interested in the closeness c_r of a vertex r which is defined to be[1]

$$\frac{1}{c_r} = \frac{1}{(N-1)} \sum_{v \in \mathcal{V} \setminus r} d_{vr} \tag{1.1}$$

where d_{uv} is the length of the shortest path between any pair of vertices u and v and the sum is over all vertices except r. To study the closeness c_r of a vertex r we defined the SHORTEST-PATH TREES which are rooted on node r. These trees contain one shortest path from a root node r

---

[1]Care is needed as at least one high impact paper [5] defines closeness to be the sum of distances to all other nodes, $\sum_v d_{vr} = (N-1)/c_r$, something known elsewhere as "status" [6, 7, 8] and which when normalised is proportional to what we call the "farness" $f_r = 1/c_r$ of vertex r. The definition of closeness in [5] is the inverse of the expression used here (1.1) and in all other papers we have found. The paper [5] derives it's definition from two other high impact papers: [9] (eq.2, pp.583) and [10] (pp.225). These last two papers also use what we call farness but those authors call their measure "point centrality" so avoiding confusion.
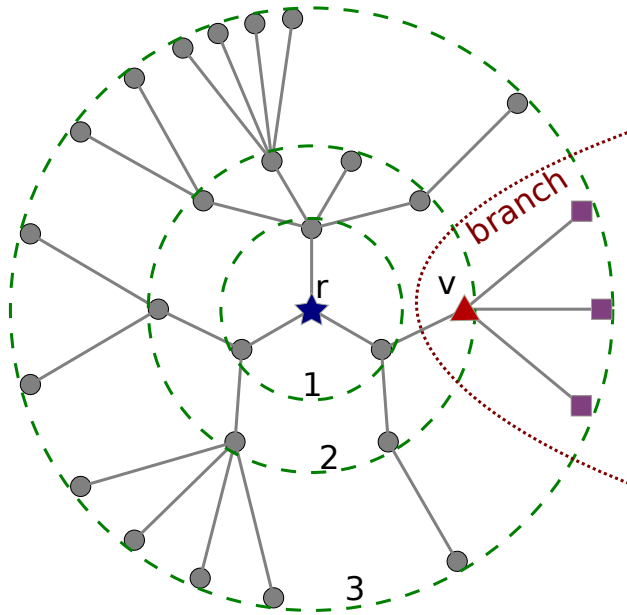
Supplementary Figure 1.2: An example of a rooted tree $\mathcal{T}(r)$ defined in terms of a root node r, the blue star at the centre. All nodes at the same network distance from the root node, are placed at the same distance from the root node in this visualisation, as indicated by the green dashed circles. The red triangle, node v, is the root of a BRANCH $\mathcal{T}(v,r)$, a smaller tree containing all the nodes u which lie on a path from the root through v (which is also a shortest path from r to u in the full graph $\mathcal{G}$). These nodes u in the branch are therefore further from the root than v, $d_{ur} \geq d_{vr}$. The branch $\mathcal{T}(v,r)$ illustrated here includes v (red triangle), the nodes indicated with purple squares and the edges between these nodes. The degree of a node is the number of neighbours so here node v has degree 4.

to every other node in the full network $\mathcal{G}$. At least one shortest-path tree exists for every root node r and these trees can be defined using a breadth-first search algorithm. The shortest-path trees are not unique, one root node can have several different shortest-path trees, but any one of these trees will be sufficient for our purposes. The key property of these trees is that the distance from every node in the tree to the root node is the same in both the shortest-path tree $\mathcal{T}(r)$ and in the full graph $\mathcal{G}$. We give more details on how such trees may be found and a proof of the properties we have stated in Supplementary Note 2. We show an example of a shortest-path tree in Supplementary Figure 1.3.
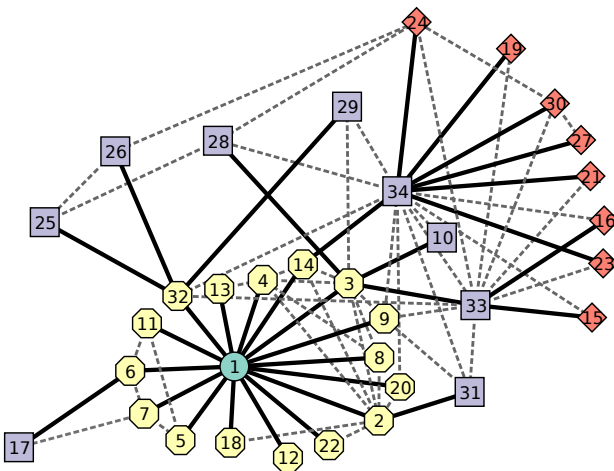
Supplementary Figure 1.3: The Zachary Karate club network [1]. The nodes of the same colour and shape have shortest paths to node 1 of the same length. The thick black lines are edges which are part of one possible shortest-path tree $\mathcal{T}(1)$ with node index 1 as the root node of the tree. The dashed grey lines indicate edges which are not in the tree. These shortest-path trees are not unique as can be seen here since we can include edge $(7, 17)$ in $\mathcal{T}(1)$ instead of the edge $(6, 17)$ used here. Node labels correspond to those used in [1].

The picture we have is as follows. Suppose we start at node r of degree k_r and we look a few steps always from the root at all the nodes $\{v_\ell\}$ which are distance ℓ from the root node. Each of these nodes v_ℓ will be the root of a branch[2] $\mathcal{T}(v_\ell, r)$ of our tree $\mathcal{T}(r)$.

---

[2]This branch $\mathcal{T}(v_\ell, r)$ containing all the nodes u which are further from the root v_ℓ (i.e. $d_{ur} \geq \ell$) and these nodes u lie on a path from the root node r passing through v_ℓ (so the branch $\mathcal{T}(v_\ell, r)$ includes v_ℓ itself) which is also a shortest path from r to u in the both the full graph g and in the tree $\mathcal{T}(r)$. The edges in the branch $\mathcal{T}(v_\ell, r)$ are all the edges from the full tree $\mathcal{T}(r)$ which run between nodes in the branch. So all nodes in the branch $\mathcal{T}(v_\ell, r)$ will be at least distance ℓ from the root node and the branch itself is also a tree, a subgraph of

One key assumption we make is that *all* these branches have similar statistical properties because the vast majority of such branches are in the 'bulk' of the network. To start with, we will assume that in terms of our measurements $\mathcal{T}(v_\ell, r) \equiv \mathcal{T}(v'_\ell, r)$ for any two nodes v_ℓ and v'_ℓ distance ℓ from our root. For the same reason, in most graphs we might expect these subgraphs to be similar whatever the root node was[3] so $\mathcal{T}(v_\ell, r) \equiv \mathcal{T}_\ell$.

In particular, we can look at the number of nodes at distance ℓ from the root r which we denote as $n(\ell; r)$ ask how this grows with distance ℓ. If the average number of child nodes (neighbours which are one step further out) is $\bar{z}(v_\ell, r)$, then our assumption that branches are similar statistically means we are assuming that this $\bar{z}$ only depends on the distance from the root node so $\bar{z}(v_\ell, r) = \bar{z}_\ell$. Hence, we estimate that the number of nodes ℓ steps away from our root vertex r as

$$n_\ell(r) \approx \bar{z}_\ell \, n_{\ell-1}(r), \qquad n_\ell(r) \approx \prod_{\ell'=1}^{\ell} \bar{z}_{\ell'}, \quad \text{for} \quad \ell \geq 1, \quad n_0(r) = 1. \tag{1.2}$$

It is immediately possible to improve on this approximation. For a start, we usually know, or can easily find, the degree k_v of each node v so we will assume that we know the values $n_1(r) = k_r$ since neighbours are the nodes at distance one from the root. So at this stage we have a model with $n_0(r) = 1$, the N local values $n_1(r) = k_r$, and then global (independent of root vertex r chosen) parameters $\bar{z}_\ell$ for $\bar{z} \geq 2$. However, our goal is to find a simple statistical relationship between closeness and degree, so we will make a further approximation. If the graph looks the statistically similar once we are looking at nodes in the bulk, then assuming that $\bar{z}_\ell \approx \bar{z}$ independent of ℓ for $\ell \geq 2$ is consistent with our picture. This then leaves us with

$$n_\ell(r) \approx \bar{z}^{\ell-1} k_r, \quad \text{for} \quad \ell \geq 1. \tag{1.3}$$

The exponential growth in the number of nodes distance ℓ from *any* root node, as encoded in (1.3) is our second key assumption. This will not be true for networks embedded on a plane or other Euclidean spaces as there we expect $n(\ell)$ to measure the surface area of a shape of radius ℓ which would follow a power-law $n_\ell \sim \ell^{D-1}$ for a D-dimensional Euclidean space.

Clearly, to get a network of significant size we need $\bar{z}_\ell \geq 1$. However, the total number of nodes N is given by

$$N = \sum_{\ell=0}^{\infty} n_\ell(r) \tag{1.4}$$

so a model with constant $\bar{z}_\ell = \bar{z} \geq 1$ gives an infinite graph which is of little use for the finite graphs found in real data sets. So we know that in practice $\bar{z}_\ell < 1$ for larger ℓ in a graph of a finite number of nodes N. This contradicts our need for $\bar{z}_\ell > 1$ for small ℓ in order to get a network of any size.

The crudest solution, and the one we will follow here, is to assume that $\bar{z}_\ell = \bar{z}$ for $2 \leq \ell \leq L_r$ where L_r is some long distance cutoff which may depend on the root vertex r considered with $\bar{z}_\ell = 0$ for larger ℓ. So we work with the following model

$$n_\ell(r) = \begin{cases} 1 & \text{if} \quad \ell = 0, \\ \bar{z}^{\ell-1} k_r & \text{if} \quad 1 \leq \ell \leq L_r, \\ 0 & \text{if} \quad L_r < \ell. \end{cases} \tag{1.5}$$

---

the original tree $\mathcal{T}(v_\ell, r) \subset \mathcal{T}(r) \subset \mathcal{G}$.

[3]To be more precise, if our measurement M is a map from a branch to some real numbers, then we require that $M(\mathcal{T}(v_\ell, r)) \approx M(\mathcal{T}(v'_\ell, r')) \approx M_\ell$ so that the average value of these measurements for any given ℓ, say M_ℓ, is a fair representation of the individual measurements.

To improve clarity of the expressions, we will now drop the explicit dependence on the root vertex r chosen and write $c \equiv c_r$, $k \equiv k_r$, and $L \equiv L_r$.

We will determine the distance cutoff L by imposing (1.4) which given our model for $\bar{z}_\ell$ now becomes

$$N = 1 + \sum_{\ell=1}^{L} \bar{z}^{\ell-1} k = 1 + k \frac{(\bar{z}^L - 1)}{(\bar{z} - 1)} \, . \tag{1.6}$$

Inverting this we see that the distance cutoff L we need is given by, for large N,

$$L(N, k) \approx \frac{\ln\left(N(\bar{z} - 1)/k\right)}{\ln(\bar{z})} \, . \tag{1.7}$$

Even in this simplest approximation, it is clear that the distance cutoff L depends on our choice of root vertex through the degree of the root node.

In principle L in (1.5) and (1.6) is an integer but it is clear from the form in (1.7) that we need L to be a real number, in some sense an average over the actual distances from the root to the leaves (nodes with degree one) of the tree. So the real number valued L given by (1.7) sets the scale of the distance beyond which the terms in these sums become negligible. We also note as an aside that L depends on the size of the network through a $\ln(N)$ factor, not as a power $N^{1/D}$, and this is the classic "small-world" effect seen in many network length scales such as diameter and average distance, see Supplementary Section 1.3 for further discussion.

We will assume our network is connected so a shortest path exists between all node pairs in the network. We can now rewrite the closeness c_r (1.1) of a root vertex r using (1.5) and (1.7) to gives us

$$\frac{1}{c_r} = \frac{1}{\Omega} \sum_{\ell=1}^{L} \ell n_\ell \, , \quad \Omega = \sum_{\ell=1}^{L} n_\ell \, . \tag{1.8}$$

where for simplicity we will drop the explicit dependence on the root vertex r and write $c \equiv c_r$, $k \equiv k_r$, $L \equiv L_r$, and $\Omega \equiv \Omega_r$.

For the normalisation Ω we have that $\Omega = (N - 1)$ and using (1.6) we can express this in terms of our other variables

$$\Omega = \sum_{\ell=1}^{L} k \bar{z}^{\ell-1} = k \frac{(\bar{z}^L - 1)}{(\bar{z} - 1)} = N - 1 \, . \tag{1.9}$$

Note this gives us a link between L, N and $\bar{z}$. We will eventually use (1.9) to eliminate L as we assume N is known. However the expressions are simpler in terms of L so here we will use (1.9) to eliminate N and Ω.

Using (1.5) and (1.7) gives us that

$$\frac{1}{c} = \frac{1}{\Omega} \sum_{\ell=1}^{L} k\ell \bar{z}^{\ell-1} = \frac{k}{\Omega} \frac{d}{d\bar{z}} \sum_{\ell=0}^{L} \bar{z}^{\ell} = \frac{k}{\Omega} \frac{d}{d\bar{z}} \left(\frac{\bar{z}^{L+1} - 1}{\bar{z} - 1} \right) \tag{1.10}$$

$$= \frac{k}{\Omega} \left(\frac{(L+1)\bar{z}^L}{\bar{z} - 1} - \frac{(\bar{z}^{L+1} - 1)}{(\bar{z} - 1)^2} \right) \tag{1.11}$$

Using (1.9) to eliminate Ω (i.e. N) in terms of L and $\bar{z}$, we have that

$$
\begin{align}
\frac{1}{c} &= \frac{(\bar{z}-1)}{(\bar{z}^L-1)}\left[\frac{(L+1)\bar{z}^L}{\bar{z}-1} - \frac{(\bar{z}^{L+1}-1)}{(\bar{z}-1)^2}\right] \tag{1.12}\\
&= \frac{L\bar{z}^L}{(\bar{z}^L-1)} + \frac{\bar{z}^L}{(\bar{z}^L-1)} - \frac{(\bar{z}^{L+1}-1)}{(\bar{z}^L-1)}\frac{1}{(\bar{z}-1)} \tag{1.13}\\
&= L\left(1 - \frac{1}{(\bar{z}^L-1)}\right) + \frac{1}{(\bar{z}-1)}\frac{1}{(\bar{z}^L-1)}\left(\bar{z}^L(\bar{z}-1) - (\bar{z}^{L+1}-1)\right) \tag{1.14}\\
&= L\left(1 - \frac{1}{(\bar{z}^L-1)}\right) + \frac{1}{(\bar{z}-1)} \tag{1.15}
\end{align}
$$

Now we can use (1.7) in (1.15) to produce a prediction of the relationship between the closeness of a node and its degree, also showing how closeness should vary with the size of the network and we find that

$$
\frac{1}{c} \approx \left(\frac{1}{(\bar{z}-1)} + \frac{\ln(\bar{z}-1)}{\ln(\bar{z})}\right) + \frac{1}{\ln(\bar{z})}\ln(N) - \frac{1}{\ln(\bar{z})}\ln(k) + O\left(\frac{\ln(N)}{N}\right). \tag{1.16}
$$

We now restore the dependence on the root vertex in our notation to emphasises which quantities depend on this choice, and which are fixed network values. The prediction is that the inverse of closeness c_v of any node v should show a linear dependence on the logarithm of the degree k_v of that node with a slope that is the inverse of the log of the branching ratio parameter, that is

$$
\frac{1}{c_r} = -\frac{1}{\ln(\bar{z})}\ln(k_r) + \beta. \tag{1.17}
$$

Our calculation suggests that the parameter β is a function of other known parameters but that it is also independent of the vertex v chosen, so that

$$
\beta = \beta(\bar{z}, N) = \left(\frac{1}{(\bar{z}-1)} + \frac{\ln(\bar{z}-1)}{\ln(\bar{z})}\right) + \frac{1}{\ln(\bar{z})}\ln(N). \tag{1.18}
$$

In our analysis we will assume that the number of nodes N and degrees of the nodes k_v are known as such information is often available. Then in principle we have one unknown global parameter, $\bar{z}$, which are fixed whatever vertex v we consider.

However, our calculation is fairly crude. The key assumptions are the statistical similarity of the branches and the exponential growth in the number of nodes at distance ℓ from any root. These are the idea which lead to the form (1.17). The details of the implementation, such as the precise form for the cutoff[4] of our sums, here (1.5) summarised by our single parameter L, will alter the detail form of β but not the broad dependence of closeness c_r on degree k_r and the number of nodes N. For that reason, we can regard β along with $\bar{z}$ as two global parameters (i.e. the same for all root nodes) to be determined.

1.2 Determining $\bar{z}$ and β

There are a number of ways of looking this relation between closeness and degree (1.17) when determining $\bar{z}$ and β.

---

[4]A final note is that there are some formal issues here. The calculation went through a parameter L which was initially an integer yet later it became a real valued parameter. We are of course making a particular analytic continuation of the results of sums of integers such as (1.1) and (1.9). Technically these analytic continuations are not even unique without an additional criterion but the 'natural' forms given here define the continuation chosen.

First we could set these parameters based on (1.16). That means we would find shortest-path trees for all vertices v, find their average degree $\bar{z}^{(\text{num})}$, and use this to set the value of $\bar{z}$. Logically we would then choose $\beta^{(\text{num})}$ using N and $\bar{z} = \bar{z}^{(\text{num})}$ in (1.18).

Another approach would be to use the result for the average degree of a neighbour in a random graph [11] (also see section 13.3 [3], section 5.4 [4] and the discussion in Supplementary Section 1.3) which leads to the suggestion that

$$\bar{z}^{(\text{rnd})} = \frac{\langle k^2 \rangle}{\langle k \rangle} - 1 \,. \tag{1.19}$$

Again we can then use (1.18) with N and $\bar{z} = \bar{z}^{(\text{rnd})}$ to suggest a value for $\beta^{(\text{rnd})}$. Here the expectation values are averages over the degree distribution in the full graph $\mathcal{G}$. Random graphs do become very similar to trees close to their percolation transition and this approach for $\bar{z}^{(\text{num})}$ and $\beta^{(\text{num})}$ ought to work well in that region. However, a typical shortest-path tree has many fewer edges than the full graph and many of those edges are involved in short loops so in practice it is not clear that these averages on the original graph averages are going to be of much relevance to our shortest-paths. Nevertheless, (1.19) provides us with a useful reference point.

Given the very simple minded approximations, neither of the previous approaches is likely to be very effective for most cases. The driving force behind the form (1.17) is the idea that the number of nodes at distance ℓ from any one chosen node rises exponentially, something found in most networks. The precise link between the parameters $\bar{z}$ and β and properties of the network is not going to be as simple or universal as the the simple derivation given here. So the most effective approach may be to treat $\bar{z}$ and β as two *independent* parameters. That is we ignore (1.18) and just do a linear fit of inverse closeness values $1/c_v$ to the logarithm of degree k_v using data from as many vertices v as we can to give $\bar{z}^{(\text{fit})}$ and $\beta^{(\text{fit})}$ (1.17). We lose little predictive power in using the data to fix these two model parameters rather than one. We can then turn this around. The fitted values $\bar{z}^{(\text{fit})}$ and $\beta^{(\text{fit})}$ give us two new global network measurements. Looking at differences between the fitted values and the alternative values suggested above can give us insights into the complexity of our network.

Indeed we can take this a step further and define new network vertex measures $\bar{z}_v$ and β_v for each vertex v by taking the closeness and degree values for that vertex and inverting (1.17) and (1.18). The β_v parameter is hard to interpret but the $\bar{z}_v$ tells us what sort of shortest-path tree that vertex sees, independent of its degree. Our assumptions state that such a value $\bar{z}_v$ will be roughly constant but individual variations could give insights into the network structure.

There is another way we can look at the effective branching ratio parameter $\bar{z}$ and that is to actually measure it in actual shortest-path trees. Calculating the average degree of the nodes in a finite tree does not tell us much as this is close to one by definition. What we really want is to look away from the outer edges of the tree, away from the degree one leaf nodes, to look at the degree of nodes in the central part of the tree as it grows in size moving away from the root node. One way might be to look at the modified average degree where we average over all nodes that have degree larger than one (so excluding all leaf nodes) and we also exclude the root node. However, for our artificial networks in our examples, where the average degree was 10.0, we find values which are much lower (typically between 2.2–2.3 for the Erdős-Réyni networks, 4.5–6.6 for the Barabási-Albert networks and 4.0–4.1 for the Configuration Barabási-Albert networks) than the degree of the original network and our $\bar{z}^{(\text{fit})}$ values. What is happening is that even without out the leaf nodes, close to the edge the branches are often made up of low degree node and these pull the average measured down. The growth of the tree as we move away from the root is driven by the presence of large degree nodes closer to the root node in the tree. By definition there are many fewer nodes close to the root node so these large degree nodes do not have much effect on the average degree measured in non-leaf nodes of the tree.

1.3 Average Shortest Path, Closeness and Random Graphs

Finding a typical length scale of a network was one of the earliest challenges in Network Science. The length d_{uv} of the shortest path between two nodes u and v provides the natural measure of distance, as it satisfies both the mathematical criteria for a distance function and our own intuition about the importance of short paths to minimise the costs and losses in communication in real networks.

Characteristic length scales are important in any system and for networks the average distance between all node pairs $\langle \ell \rangle$ is such a length scale where

$$\langle \ell \rangle = \frac{1}{N(N-1)} \sum_{u \in \mathcal{V}} \sum_{v \in \mathcal{V} \setminus u} d_{uv}. \tag{1.20}$$

This is the mathematical quantity that represents Milgram's "six-degrees of separation" [12, 13] and investigated on large scales in modern data sets, for instance [14, 15, 16]. It has been the focus of great interest in some of the earliest theoretical papers such as [17, 18, 19].

The average distance of a network $\langle \ell \rangle$ has a simple relationship to the closeness considered in this paper as average distance $\langle \ell \rangle$ is simply the half the average over all nodes of the inverse closeness

$$\langle \ell \rangle = \frac{1}{N} \sum_{r \in \mathcal{V}} (c_r)^{-1}. \tag{1.21}$$

If we insert our expression (1.17) we have that

$$\langle \ell \rangle = -\frac{1}{N} \frac{1}{\ln(\bar{z})} \sum_{r \in \mathcal{V}} \ln(k_r) + \beta \tag{1.22}$$

giving us

$$\langle \ell \rangle = +\frac{1}{\ln(\bar{z})} \ln(N) - \frac{1}{\ln(\bar{z})} \langle \ln(k) \rangle + \frac{1}{(\bar{z}-1)} + \frac{\ln(\bar{z}-1)}{\ln(\bar{z})} \tag{1.23}$$

where we have used β from (1.18). This gives for large N and fixed degree distribution that

$$\lim_{N \to \infty} \langle \ell \rangle = \frac{1}{\ln(\bar{z})} \ln(N). \tag{1.24}$$

This is the same dependence of average path length $\langle \ell \rangle$ on N found for large random graphs [18, 19]. That is for an ensemble of networks defined by configuration model, so no node-node correlations, where the degree distribution has a finite second moment in the infinite graph limit, the leading term in N is found to be [18, 19]

$$\lim_{N \to \infty} \langle \ell \rangle = \frac{\ln(N)}{\ln(\bar{z}^{(\mathrm{rnd})})} \tag{1.25}$$

where $\bar{z}^{(\mathrm{rnd})}$ is defined in (1.19).

We can, however, repeat our calculation for closeness with rather more certainty for random graphs, that is those defined by the configuration model. The necessary results are all contained in equation (13.73) of [3] or equation (5.41) of [4] (and in several other places) which gives that for random graphs we have

$$\frac{n(\ell)}{n(\ell-1)} = \bar{z}^{(\mathrm{rnd})}. \tag{1.26}$$

It is interesting to note that this result is used to look at average distance, e.g. equation (13.73) of [3] and (5.43) of [4], and so gives an *average* closeness value of (1.21). This contrasts with our focus on the average result for an individual node of a given degree. Indeed, as far as we can tell, the literature in this area has always focussed on average closeness, on the average path length[5].

In terms of our calculation for closeness, we can see that our form for $n(\ell)$ in (1.5), an assumption in most cases, is also found in the results for *ensemble averages* given by the random graph model for large graphs provided $\bar{z} = \bar{z}^{(\mathrm{rnd})}$. From here we can follow the rest of our calculation with its additional assumptions, as given in Supplementary Section 1.1, to see that our formula for closeness of a random graph will have the same form but now there is a very specific prediction for $\bar{z}$, namely that for a random graph $\bar{z} = \bar{z}^{(\mathrm{rnd})} = (\langle k^2 \rangle / \langle k \rangle) - 1$ of (1.19).

While the exact result (1.19) for $\bar{z}$ in a large random graph is of interest, we do not expect the precise value to be informative for real networks. As Newman says about random graph models and simple network models in general (p.448 [3])

> they can give you a feel for the types of effect one might expect to see, or the general directions of changes in quantities. But they don't usually give quantitative predictions for the behaviour of real networks.

So the value of random graph models here is that they give an example where the assumed form for $n(\ell)$ of (1.5) is to be expected (on average) giving some further justification for its use in general. In fact our work on the Erdős-Réyni model and the randomised Barabási-Albert model, both examples of random graph models, show there are still clear differences between the value of $\bar{z}$ found from fits even in these finite size cases and the value $\bar{z}^{(\mathrm{rnd})}$ suggested by the random graph work for infinite graphs. However the overall form, the prediction of dependence of closeness on the logarithm of degree $\ln(k)$, does work well in these models. The lack of a perfect match for random graph models is presumably because other approximations we have made come into play for finite graphs suggesting improvements could be made in future work on closeness in random graph models.

The random graph example also highlights another issue, that of networks with power-law degree distributions with exponents between two and three, $p(k) \sim k^{-\gamma}$ for $2 < \gamma < 3$. In network models with this distribution, the average degree $\langle k \rangle$ is well defined for infinite graphs but the second moment $\langle k^2 \rangle$ diverges as we take $N \to \infty$. The result on p15879 of [18] for such random graphs is that the average path length $\langle \ell \rangle$ is simply $\lim_{N \to \infty} \langle \ell \rangle_{\mathrm{CL}} = O(\ln(N)/\ln(\ln(N)))$ for some constant A. This is because random graphs with such power-law distributions are arranged like "an 'octopus' with a dense subgraph having small diameter as the core" (from p.15881 of Chung & Lu [18] but also see discussion in Bollobás [20]). This result for the average path length points to the need for a different theoretical approach to closeness in these cases. So the failure of our closeness-degree relationship in a few networks constructed from real-world data maybe due to the nature of their large scale structure which will require a different theoretical approach.

A plot comparing the predicted average path $\langle l \rangle$ from fitted value $\beta^{(\mathrm{fit})}$ and theoretical value β for the eighteen Konect-SNAP real-world datasets is shown in Supplementary Figure 1.4.

---

[5]This is also a normalised version of the WIENER INDEX W of a graph which is the sum of the distances between all node pairs, so $W = [N(N-1)/2]\langle \ell \rangle$.
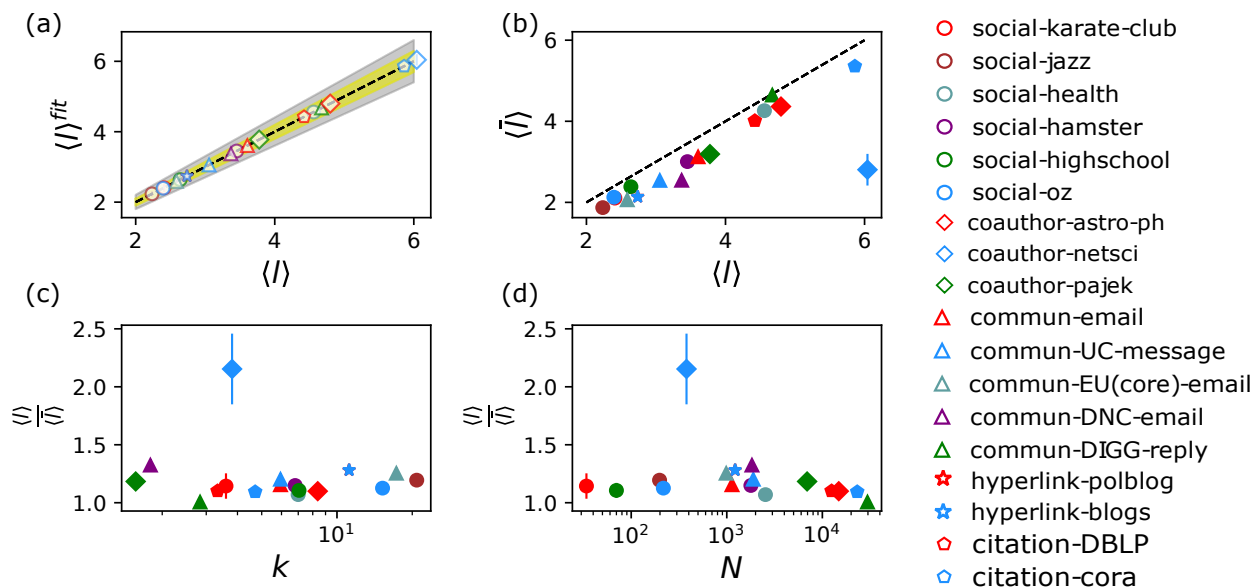
Supplementary Figure 1.4: A plot comparing the predicted average path $\langle l \rangle$ from fitted value $\beta^{\text{(fit)}}$ and theoretical value β for the eighteen Konect-SNAP real-world datasets. The empty fill markers correspond to predictions based on the fitted parameter $\beta^{\text{(fit)}}$ using (1.22) and solid fill markers correspond to predictions based on (1.23). In (a), we used fitted value $\beta^{\text{(fit)}}$ to compute average path length using (1.22). Our model can predict average shortest path in all 19 datasets within 5% deviation (shade area correspond to deviation between theoretical value and real value: 5% for yellow region and 10% for grey). From (b), we notice the result from the theoretical calculation using (1.23) are consistently lower than measured value from the network. This may suggest some other constant appears, however, the average shortest path in the large N limit will converge to $\ln N / \ln \bar{z}$. To reveal the large limit behaviours predicted by (1.24), we investigate the relation between average shortest path $\langle l \rangle$ in (c) and k, and $\langle l \rangle$ and $\ln N$ in (d).

1.4 Eccentricity and $L(N,k)$

So far we have focussed our attention on closeness. The inverse of closeness, farness $f_r = 1/c_r$, is the average distance from a given node r to all other nodes. So farness is a node-dependent length scale which we can contrast against the average path length which sets a network length scale. However, we have another node-dependent length scale present in our calculation, $L(N,k)$ of (1.7). This $L(N,k)$ is the cutoff we use in our sum so in some sense it represents the *typical* largest distance from any node r of degree k to any other node. One can think of $L(N,k)$ as being some type of *average* of the distance to end every branch in the shortest-path tree.

The actual largest distance from a node r to any other node is known as the ECCENTRICITY e_r of node r [2, 8]

$$e_r = \max\{d_{rv}|v \in \mathcal{V}\}. \tag{1.27}$$

Clearly as both $L(N,k)$ and e_r are dealing with largest distances there might be some merit in comparing the two length scales and thinking they could scale in similar ways or at least be strongly correlated. Certainly we would expect $L(N,k) \leq e_r$. However the cutoff parameter $L(N,k)$ is an average over a set of maximum values while the eccentricity e_r is simply a maximum value of a set so we expect far more fluctuations in eccentricity values even if there was a strong correlation with $L(N,k)$.

We have done a quick comparison for the theoretical models we have used with results given in Supplementary Table 1.1. What we see is that there is a good linear fit of eccentricity to $\ln(k)$, $e_r = a + b\ln(k_r)$, so this aspect of the prediction is good. Note that a good linear fit of eccentricity to the logarithm of degree was shown in Figure 1 of [8] for one example of each of the Erdős-Réyni and Barabási-Albert models. So our results confirm this observation.

Network type	N	$1/\ln(\bar{z}^{(\text{fit})})$	χ_r^2	gradient b	intercept a
ER	1000	0.46	1.12	-0.64	6.17
	2000	0.42	1.09	-0.19	5.48
	4000	0.43	1.07	-0.70	6.90
BA	1000	0.30	1.03	-0.20	4.50
	2000	0.32	1.05	-0.42	5.28
	4000	0.31	1.04	-0.43	5.67
Config-BA	1000	0.35	1.04	-0.22	4.55
	2000	0.36	1.05	-0.43	5.32
	4000	0.35	1.05	-0.43	5.68

Supplementary Table 1.1: Table of results for showing how the linear fit of eccentricity to the logarithm of degree behaves for simple graphs with average degree 10.0 produced using one of three artificial models with the same average degree $\langle k \rangle = 10.0$ but with a different number of nodes, N. Each 'ER' network is a standard Erdős-Réyni network, a 'BA' network is produced using pure preferential attachment in the Barabási-Albert model, and the 'Config-BA' network is a configuration model version of a Barabási-Albert model network. The results for $1/\ln(\bar{z}^{(\text{fit})})$ (the uncertainty is 0.01) come from linear fits of inverse closeness, taken from a table in the main text. We have fitted $e_r = a + b\ln(k_r)$ to the same models. The values of the intercept a and gradient b given are the mean over fits to each of 100 realisations (standard deviation is roughly 0.01). The reduced chi-square χ_r^2 is for the eccentricity linear fit.

However, the results in Supplementary Table 1.1 shown no clear link between the gradient b of this eccentricity fit to the $-1/\ln(\bar{z}^{(\text{fit})})$ predicted if $e_r \propto L(N, k_r)$. As eccentricity is an extremal value unlike the average used to find closeness, it could be we need to look at much

larger networks to see any link between eccentricity and our simple cutoff scale $L(N, k_r)$. As there is no clear signal here, we have not pursued this further.

1.5 Ring Calculations of First-Passage Times

Random walks play an important role in both physical processes and as a theoretical and computational tool to probe systems. The use of random walks on networks is widespread as can be seen in any textbook such as [2, 3, 4, 21]. This section is prompted by the study of first-passage times on a network using a "ring" approximation as suggested in [22]. We aim to show how the "ring" approximation in [22] is closely related to the shortest-path tree approximation used here. These are types of mean-field approximations, see section 3.2.5 of [23] for an extensive review of these and alternative approaches as well as for known results.

The ring approximation represents a network as a sequence of rings, where the ring labelled ℓ contains all $n_r(\ell)$ nodes which lie at a distance ℓ from a given root node r. Then we represent the probability of finding a random walker at a node at distance ℓ from the root node r at time t as $w_\ell(t)$. The master equation is then

$$w_i(t+1) = \begin{cases} p_{01}w_1(t) & \text{if} \quad i = 0 \\ \sum_{j=1}^{e(r)} B_{ij}(r)w_j(t) \end{cases} \tag{1.28}$$

where $e(r)$ is the eccentricity of node r, the largest distance from r to any other vertex. Here random walkers which arrive at the root node r at time t are removed at time $(t+1)$. So the mean first-passage time $\tau(r)$ for a node r is

$$\tau(r) = \sum_{t=1}^{\infty} t\, w_0(t) \tag{1.29}$$

The parameter p_{01} in (1.28) is the probability of moving a neighbour of the root node to the root node, from ring one to ring zero. From the master equation (1.28), it is clear that to find the mean first-passage time $\tau(r)$ it is sufficient to know $w_1(t)$, the probability of finding a random walker at ring one at any time,

$$\tau(r) = \sum_{t=0}^{\infty} t\, p_{01}\, w_1(t)\,. \tag{1.30}$$

This means we need to analyse the matrix $B_{ij}(r)$ which is the probability of a walker moving from ring j to ring i when we are considering random walkers moving towards root node r. This matrix $\mathsf{B}(r)$ is a constant tridiagonal stochastic matrix. Assuming a strongly connected network, standard properties of such matrices tell us that the largest eigenvalue λ_1 is therefore constrained to be $1 \geq \lambda_1 \geq 1 - B_{01}(r)$, and it is linked to the only right-eigenvector $\mathbf{v}^{(1)}$ whose entries can be chosen to be all positive. We can the write the average time for a random walker to first reach the root node in terms of some initial distribution of random walkers $w_i(t=0)$ as

$$\tau(r; \mathbf{w}(t=0)) = \sum_{t=1}^{\infty} t \sum_{i=1}^{e(r)} p_{01}[(\mathsf{B}(r))^{t-1}]_{1i} w_i(t=0) \tag{1.31}$$

If the starting point of the random walk is chosen with equal probability, then in terms of the number of nodes in each ring we have that $w_i(t=0) = n_r(i)/(N-1)$. The mean first-passage time $\tau(r)$ for a root node r is then

$$\tau(r) = \sum_{t=1}^{\infty} t \sum_{\ell=1}^{e(r)} B_{01}(r)[(\mathsf{B}(r))^{t-1}]_{1\ell} \frac{n_r(\ell)}{(N-1)} \tag{1.32}$$

This represents the most general form of the mean first-passage time within the mean-field ring approximation. It is hard to make progress as we still need to know $(3e(r) - 2)$ entries for each of the N $\mathsf{B}(r)$ matrices. In most cases, results are obtained by working with a specific simple model, such as the Erdős-Réyni model considered in [22], so the matrices $\mathsf{B}(r)$ are known expression in terms of a few parameters.

This is where we can make contact with the very simple approximations made when considering closeness using the shortest-path tree. While simpler than many approaches in the literature, our approximations work well for closeness in many real networks so it is interesting to see what they give here for first-passage times. One part of our approximation is simply that $n_r(\ell) = k_r \bar{z}^{\ell-1}$ and with that we find

$$\tau(r) \;=\; \frac{k_r p_{01}}{(N-1)} \sum_{t=1}^{\infty} t \sum_{\ell=1}^{L(N,k)} [\mathsf{B}^{t-1}]_{1\ell} \bar{z}^{\ell-1} \,. \tag{1.33}$$

Note we have to use our cutoff $L(N, k)$ rather than the eccentricity $e(r)$. Also note that inherent in our approach is that away from the root node, the network looks the same statistically. So another important aspect of our approximation is that the matrix B is independent of the root node, but this is something that will also occur in most models found in the literature.

The key point is that the main dependence on the root node is through the degree k_r used to start the exponential growth in nodes encoded in our expression $n_r(\ell) = k_r \bar{z}^{\ell-1}$. There is some logarithmic dependence coming through the cutoff $L(N, k)$ but we assume this will be a smaller correction in any practical finite size network. Thus one simple conclusion is that on average, the mean first-passage time $\tau(r)$ will be proportional to degree. This is something which was seen in [22] for a real network and in the Erdős-Réyni model but not explained there.

Assuming the dominant contribution is from the largest eigenvalue λ_1, with normalised left- and right-eigenvectors $\mathbf{u}^{(1)}$ and $\mathbf{v}^{(1)}$ respectively, we have that

$$\tau(r) \;=\; \sum_{t=1}^{\infty} t \sum_{\ell=1}^{L(N,k_r)} p_{01} [\mathsf{B}^{t-1}]_{1\ell} \frac{k_r \bar{z}^{\ell-1}}{(N-1)} \tag{1.34}$$

$$\;=\; \frac{k_r p_{01}}{(N-1)} c_1 \sum_{t=1}^{\infty} t \sum_{\ell=1}^{L(N,k_r)} (\lambda_1)^{t-1} v_\ell^{(1)} \tag{1.35}$$

where c_1 gives the overlap of the initial condition with the left-eigenvector $\mathbf{u}^{(1)}$, namely

$$c_1 \;=\; \sum_{\ell=1}^{L(N,k_r)} u_\ell^{(1)} \bar{z}^{\ell-1} \,. \tag{1.36}$$

This gives us

$$\tau(r) \;=\; k_r \frac{c_1 p_{01}}{(N-1)} \frac{1}{(1-\lambda_1)^2} \sum_{\ell=1}^{L(N,k_r)} v_\ell^{(1)} \,. \tag{1.37}$$

If we have a simple model for the entries in the matrix B, it is relatively easy to find the leading eigenvalue and its eigenvectors of this tridiagonal matrix so a more sophisticated result is possible. However the main point for this paper is that the dominant dependence of the mean first-passage time $\tau(r)$ on the choice of root vertex r is through the factor of the degree k_r and this factor appeared simply through our assumption that the number of nodes grew exponentially as $n_r(\ell) = k_r \bar{z}^{\ell-1}$.

Supplementary Note 2: Shortest-Path Tree Algorithm

Our algorithm to find shortest-path trees is a slight variation on well known depth-first search algorithms used to find shortest paths between nodes. We have included it because our focus is rather different from those standard algorithms and because we will use our algorithm to deduce a key property of shortest-path trees. That is the property that every node in a connected network is the root node of at least one shortest-path tree. We can always find a shortest path from each node to any other node (in the same connected component). However, if we take a set of shortest paths, one from a root node to each other node, the union of such a set of paths does not in general give a tree. The success of the algorithm below shows that shortest-path trees can always be constructed.

This shortest-path tree algorithm generates an example of a shortest-path tree starting from a given root node v. It is simply a breadth-first search algorithm [24] (Dijstra's algorithm for unweighted networks) where we record which edge was used to reach each node for the first time in the breadth first search as these edges form a shortest-path tree. In practice, there are various ways to optimise this implementation, for example see section 10.3 [3] and [24], but this version serves as a simple example which we will then use to highlight some properties of shortest-path trees that we use in our work.

1. Label all nodes v with distance to the root node as -1, `distance[v]=-1`.

2. Label all nodes v with inner neighbour -1, `inner_neighbour[v]=-1`.

3. Start from root node r, set `current_distance=0`, `distance[r]=current_distance`.

4. Create a set `next_set` containing just node r.

5. Increment current distance, `current_distance+=1`.

6. Copy the contents of `next_set` into `current_set`.

7. Remove the contents of `next_set` so it is now an empty set.

8. Loop through all nodes u in `current_set`. For each u do the following.

   (a) Add a neighbouring node v to `next_set` if the distance from v to the root has not been set, i.e. add if `distance[v]=-1` and set `inner_neighbour[v]=u`.

   (b) On the other hand if `distance[v]=current_distance` then you may choose to change to the new node u using `inner_neighbour[v]=u`. This node v has already been found and is in a shortest-path tree.
   Note that at this point we could change the tree defined by using this new neighbour, the current u, instead of the existing node v already found. This might be done with a random number, say 50% of the time.

9. Once the loop in 8 has finished, if `next_set` is not empty, then loop back to 5.

10. The edges in the tree (v, u) are given by `u=inner_neighbour[v]` where u is one step closer to the root than v. Note the one exception is the root node which has no inner neighbour and `inner_neighbour[r]=-1`.

This algorithm also gives us a proof that in a single component simple graph, there always exists at least one such shortest-path tree $\mathcal{T}(r)$ for every node r. The proof can be expressed as follows where we set $N(u) \equiv$ `inner_neighbour[u]`.

1. Every node in the graph is visited by this algorithm as we are assuming a single component. So $N(u)$ is always defined for every node except for the root node.

2. The edge set of the tree $\mathcal{T}(r)$ is $\mathcal{E}_r = \{(u, N(u)) \mid u \in \mathcal{V} \setminus r\}$.

3. This edge set $\mathcal{E}_v$ contains all the vertices in $\mathcal{V}$, the vertex set of the original graph, so $\mathcal{V}$ is also the vertex set of the tree.

4. Our tree is then the subgraph $\mathcal{T}(v) = \{\mathcal{V}, \mathcal{E}_v\}$ of the original graph.

5. For all non-root nodes u, all nodes in $N(u)$ are exactly one step closer to the root node than u.

6. For any node u, the sequence $\{u_i\}$, where $u_0 = r$ is the root node, $u_i = N(u_{i-1})$ for $i = 1, 2, \ldots, \ell$, and $v_\ell = u$, always exists.
   Note that u_i in the sequence is i distance from the root node from 5.
   That this path exists then follows from 5, since $N(u)$ is always defined so we can always start from v_ℓ and iterate down the sequence. The iteration terminates at $u_0 = r$, the root node, as the node u_1 will always be one step away from the root so we must have $N(u_1) = r$.
   It then follows that this defines a path between any given node u and the root r.

7. This path $\{u_i\}$ must be a shortest path because the edge from u_{i-1} to u_i would always be visited in the algorithm before any edge between u_i and nodes further away from the root node than u_{i-1} as this is what the breadth first search guarantees. Each edge is visited when we are studying the neighbours of a node in step 8 of the algorithm.

8. The edge set $\mathcal{E}_v$ therefore contains paths from every vertex to the root. Hence, all vertices in are connected in the $\mathcal{T}(v)$ subgraph so this is a single component subgraph.

9. The edge set $\mathcal{E}_v$ has one less edge than the the total number of vertices in the graph which is a necessary and sufficient condition for a single component graph to be a tree.

10. Thus this algorithm defines a spanning tree $\mathcal{T}(v) = \{\mathcal{V}, \mathcal{E}_v\}$ that contains a shortest path from every vertex to the root node. That is it is a shortest-path spanning tree.

Supplementary Note 3: Data Sets

We used a variety of networks for which data is openly available. These were analysed in two groups. The first group of eighteen networks we refer to as the "Konect-SNAP networks" were analysed in greater detail and we provide further information on each data set here. The second set of one hundred and twelve networks we describe as our "Netzschleuder" networks. We provide have provided independent copies of the data used in [25].

3.1 Description of Konect-SNAP networks

For these networks, all but one can be found on KONECT [26, 27] with the `commun-EU(core)-email` coming from the Stanford Large Network Dataset Collection [28]. However, many of these networks can be found on other repositories of network data. Our aim was to find networks of different sizes representing contrasting types of interaction which we break down into five broad categories: social networks (`social-...`), communication networks (`commun-...`), citation networks (`citation-...`), co-author networks (`coauth-...`), and hyperlink networks (`hyperlink-...`). These networks have been used in many contexts in other publications but we will only give a brief summary of each one.

In each case we created a simple graph, ignoring edge directions and weights, node types, time stamps, and any other such information. We took the largest connected component (LCC) of the graph and performed our analysis on this. Some basic statistics on each graph is given in (3.1) and then more detailed information on each data set follows.

Network Name	Number of nodes	Number of edges	Mean distance
`social-karate-club`	34	78	2.44
`social-jazz`	198	2742	2.21
`social-hamster`	1858	12534	3.39
`social-oz`	217	2672	2.33
`social-highschool`	70	366	2.66
`social-health`	2539	12969	4.52
`commun-email`	1133	5451	3.65
`commun-UC-message`	1899	59835	3.07
`commun-EU(core)-email`	1005	25571	2.59
`commun-DNC-email`	2029	39264	3.37
`commun-DIGG-reply`	30398	87627	4.68
`citation-DBLP-cite`	12590	49759	4.37
`citation-Cora`	23166	91500	5.74
`coauthor-astro-ph`	16046	121251	5.10
`coauthor-netscience`	1461	2742	6.28
`coauthor-pajek`	6927	11850	3.79
`hyperlink-polblog`	1224	33430	2.75
`hyperlink-blogs`	1224	19025	2.72

Supplementary Table 3.1: Summary statistics for the original data sets used in this paper, for the whole graph not just the largest connected component. The mean distance is the average length of the shortest paths between all pairs of connected nodes.

Social networks

Social networks capture the social interactions between actors, such as friends, colleagues, clients and students. We used five data sets, the size of networks ranged from 34 to 2539 nodes. On average, we find the mean shortest distance are quite small compare other type of networks (apart from `social-health` dataset).

The `social-karate-club` is the well-known and much-used Zachary karate club dataset. The original data was collected from the members of a university karate club by Wayne Zachary in 1977 [1] and each edge represents some type of social interaction between two members of the club.

The `social-highschool` network represents friendships between boys in a small highschool in Illinois, USA. Each boy was asked once in the fall of 1957 and the spring of 1958. This dataset aggregates the results from both dates. A node represents a boy and an edge between two boys shows that at least one boy chose the other as a friend. The original network [29] is directed, weighted and allows multiple edges.

The `social-hamster` network comes from the Koblenz Network Collection (KONECT) [27] where it is described as the "Hamsterster households network dataset" but no further information is provided.

The `social-jazz` network is the collaboration network between Jazz musicians. Each node is a Jazz musician and an edge denotes that two musicians have played together in a band [30].

The `social-oz` network is a network recording the friendships between 217 residents living at a residence hall located on the Australian National University campus [31]. A node represents a person and edge represent the friendship between them.

The `social-health` network is a network created from a survey of students in 1994/1995 [32]. Each student was asked to list their five best female and five best male friends. A node represents a student and an edge between two students shows at least one chose the other as a friend.

Communication networks

Communication networks describe the individual messages exchanged between people. Communication networks are often directed and typically contain multiple edges each with distinct time stamps so we are neglecting a lot of information when working with simple graphs representations.

The `commun-email` network is the based on emails sent between members of the University Rovira i Virgili in Tarragona in the south of Catalonia in Spain [33]. Nodes are users and each edge represents that at least one email was sent between two users.

The `commun-DNC-email` network is built from the emails from the Democratic National Committee, the formal governing body for the United States Democratic Party. A dump of emails of the Democratic National Committee was leaked in 2016. Nodes in the network correspond to persons in the dataset. An edge in the dataset denotes that at least one email has been sent between the two linked nodes.

The `commun-UC-message` network represents messages sent between the users of an online community of students from the University of California, Irvine [34]. An edge connects two users if they exchanged at least one message.

The `commun-EU(core)-email` is a network representing email sent between members of a large European research institution [35]. An edge represents an email sent between members of the institution (nodes). This data was downloaded from sourced from the Stanford Large Network Dataset Collection [28].

The `commun-DIGG-reply` data [36] gives a network of users of the social news website Digg. Each node is a user of the site two users are connected by an edge is one of those users replied

to another user at any point.

Citation networks

Citation networks represent documents as nodes in the network, with two nodes linked if one document cites another. These are direct acyclic graphs in principle but here we use a simple graph representation.

The `citation-DBLP-cite` is the citation network built from the DBLP database of computer science publications [37].

The `citation-Cora` network uses another database of computer science papers, CORA [38, 39]. Our simple network is constructed as for the DBLP network.

Co-authorship network

Co-authorship networks are networks connecting authors who have written articles together. Co-authorship networks are normally weighted but we ignore that here.

The `coauthor-astro-ph` network is the co-authorship network from the astrophysics section (`astro-ph`) of arXiv preprint archive constructed in [40]. Nodes are authors and an edge denotes a collaboration on at least one paper.

The `coauthor-netscience` network is a network of co-authors in the area of network science [41]. Nodes represent authors and edges denote collaborations.

The `coauthor-pajek` is the co-authorship graph around Paul Erdős [42] which can be used to is used to define the "Erdős number".

Hyperlink networks

In hyperlink networks the nodes are pages or documents. These are linked by an edge if there is at least one hyperlink between these two documents in either direction as here we ignore the direction inherent to hyperlinks.

We use two examples from hyperlinks between blogs about politics during the U.S. Presidential Election of 2004 [43], `hyperlinks-blogs` and `hyperlink-polblog`.

3.2   Description of Netzschleuder networks

We have used an additional 112 networks taken from the Netzschleuder repository [44] though again, many of these are available elsewhere. We made use of simple interface to these network data sets provided by the `graph-tool` package [45] and some examples of our code, along with copies of the datasets, doing this can be found in [25]. Our sole selection criteria was that the network would download and run without additional work on our existing code. For this reason some excessively large data sets or those with unusual data structures were not included in our analysis. This means that some of the 112 Netzschleuder networks that we analysed are not good candidates for our relationship, e.g. they have bipartite structure. We did not exclude such cases to ensure that we have as unbiased a sample of network data sets as possible. For the 112 networks that we looked at, some basic properties along with additional results are given in Supplementary Note 4.4 below. As there are too many datasets to discuss in detail here, please refer to the Netzschleuder repository [44] for further information on individual examples.

Supplementary Note 4: Additional Results

These results are not central to the work in the main paper but they are mentioned briefly in the Discussion section. In this section we show these results in more detail.

4.1 Dependence of fit on N

For the artificial models we have looked at how the parameters $\bar{z}^{(\mathrm{fit})}$ and $\beta^{(\mathrm{fit})}$, found by fitting to (1.17), depend on the number of nodes N. For $\beta^{(\mathrm{fit})}$ we compare the fitted value against the value $\beta = \beta(\bar{z}^{(\mathrm{fit})}, N)$ predicted from (1.18) so are factoring out the expected $\ln(N)$ contribution to β. Results are shown in Supplementary Figure 4.1. In all cases there is no strong variation and while there are signs of some systematic variation, it is at a small scale and no clear pattern emerges.
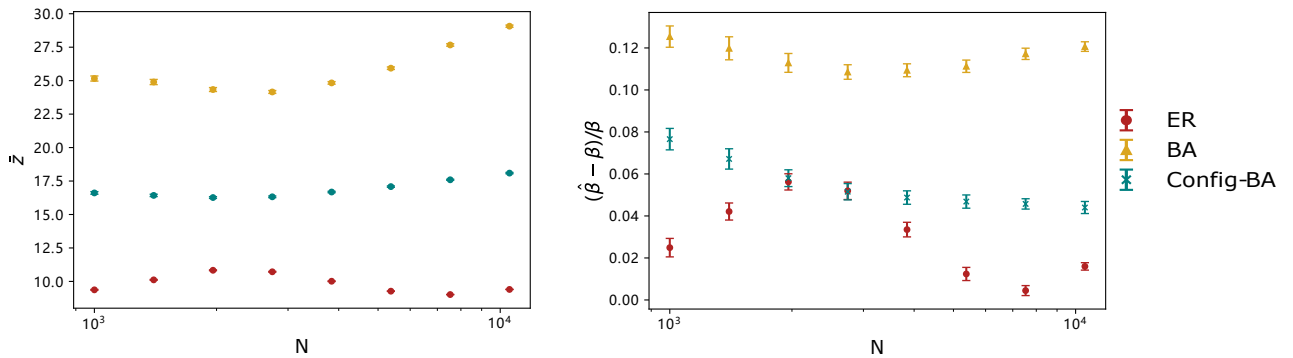
Supplementary Figure 4.1: Plots showing the dependence of the best fit values $\bar{z}^{(\mathrm{fit})}$ and $\beta^{(\mathrm{fit})}$ on the number of nodes N. These are shown for networks from three artificial models: the Erdős-Réyni model (ER, red data), the Barabási-Albert model (BA, yellow data), and the randomised Barabási-Albert model (Config-BA, blue data), all for average degree 10.0. Data points are the mean values with error bars showing the standard error of the mean estimated from 100 realisations. On the left we can see the best fit value of the growth factor $\bar{z}^{(\mathrm{fit})}$ has some non-linear dependence on system size N. On the right we compare the fitted value $\beta^{(\mathrm{fit})} = \hat{\beta}$ to the value $\beta = \beta(\bar{z}^{(\mathrm{fit})}, N)$ predicted from (1.18) using the fitted value $\bar{z}^{(\mathrm{fit})}$. It is clear that the predicted value $\beta = \beta(\bar{z}^{(\mathrm{fit})}, N)$ from (1.18) is 5% to 15% below the best fit value though no strong trends are visible on this small range of N values.

4.2 Higher-order polynomial fit

An interesting feature of our work is the failure of four of our eighteen real-world networks to give us a good fit: `social-jazz`, `commun-UC-message`, `coauthor-astro-ph` and `coauthor-pajek`. This is clear from their values of reduced chi-square which are all greater than ten. Several of the plots, particularly those showing the fractional error, also show issues with these data sets but also some clear trends in some other data sets even if this is within statistical fluctuation for each individual point. In the case of the `social-jazz` network we could dismiss this as this is such a small network, though we note that our relationship has worked well for several smaller networks. However the other three networks with high χ_{r}^2 one to fifteen thousand nodes and these show our relationship (1.17) is not the last word. Even when the chi-square measure looks good, the plots of fractional error show a convincing trend that is not captured by our relationship. Again we stress that these deviation are not that large, no worse than 5% in most cases. Nevertheless this points to the need to go beyond our simple derivation.

One way to get get a better fit is to try to fit a higher order polynomial in $\ln(k)$ to the inverse closeness values, that is

$$1/c = \sum_{i=0}^{m} p_i (\ln k)^i \tag{4.1}$$

where m is the maximum number of the parameters in the polynomial fit, p_i is the coefficient for i-th power of $\ln k$. Working with $m > 1$ is not motivated by any theoretical consideration but we work with it here because it is easy to implement. We have not found this to be particularly effective except in one case, `commun-DIGG-reply`, where we already had a good fit but could see clear trends in the fractional deviation plots. Results for some examples are shown below. The data on first order ($m = 1$) corresponds to what was used in the main text.
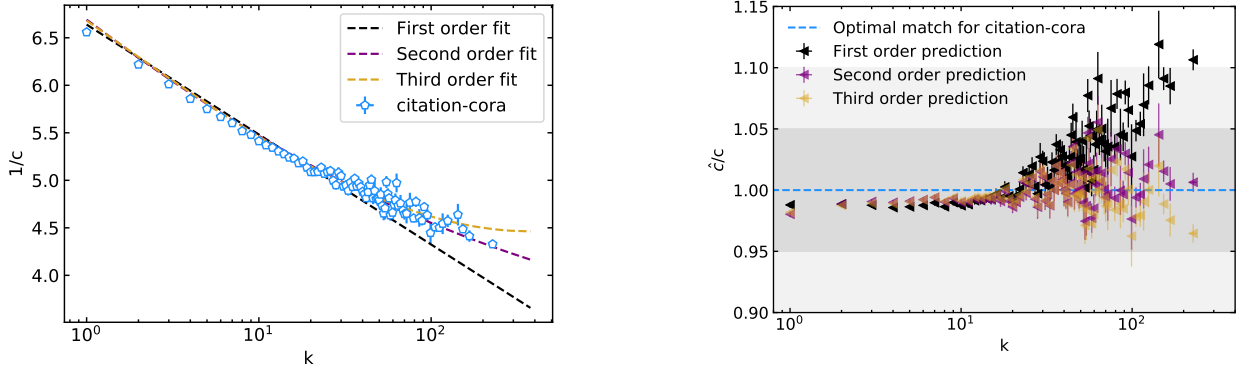
Supplementary Figure 4.2: The effect of fitting inverse closeness to higher order polynomials in the logarithm of degree $\ln(k)$ (4.1) for the CORA citation data `citation-Cora`. On the left we show the data points (means with standard error of mean for error bars) against the dashed lines for different polynomial fits. On the right we show the fitted value $\hat{c}$ divided by the data c against degree k. The shaded bands mark 5% and 10% deviations.
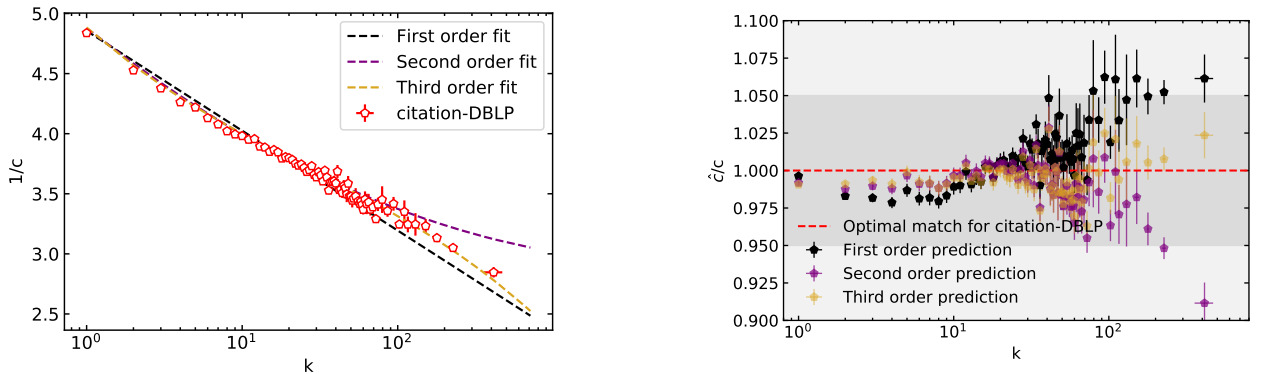
Supplementary Figure 4.3: The effect of fitting inverse closeness to higher order polynomials in the logarithm of degree $\ln(k)$ (4.1) for the DBLP citation data `citation-DBLP-cite`. On the left we show the data points (means with standard error of mean for error bars) against the dashed lines for different polynomial fits. On the right we show the fitted value $\hat{c}$ divided by the data c against degree k. The shaded bands mark 5% and 10% deviations.
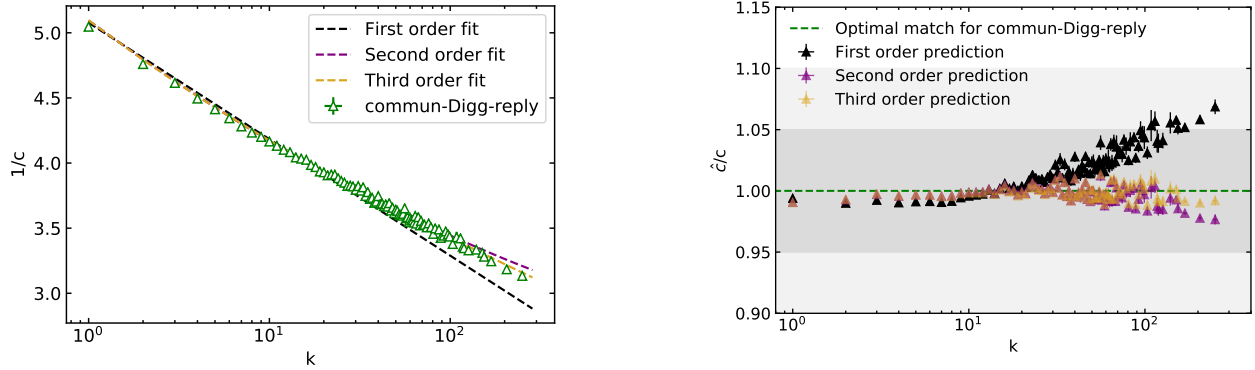
Supplementary Figure 4.4: The effect of fitting inverse closeness to higher order polynomials in the logarithm of degree $\ln(k)$ (4.1) for the DIGG communication network `commun-DIGG-reply`. On the left we show the data points (means with standard error of mean for error bars) against the dashed lines for different polynomial fits. On the right we show the fitted value $\hat{c}$ divided by the data c against degree k. The shaded bands mark 5% and 10% deviations.
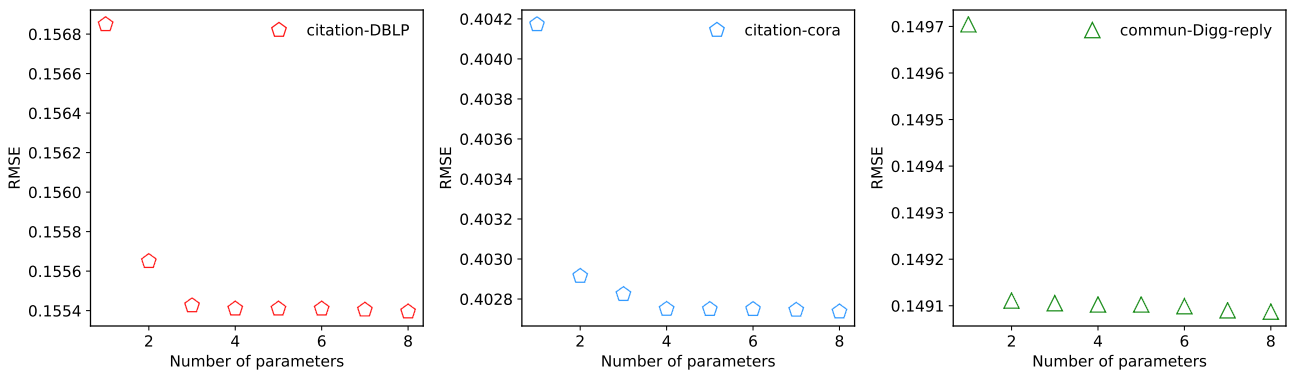
Supplementary Figure 4.5: Higher order correction revealed by comparing root mean squared error (RMSE). For those datasets suggesting a dependency of higher order relation between $\ln(k)$ and $1/c$, we use the higher order polynomial fit of (4.1) but the improvement is not significant.

4.3   Fluctuations in Closeness

We can take a closer look at some of the fluctuations in closeness around the predicted value for the six social networks in Supplementary Figure 4.6 and the five communication networks Supplementary Figure 4.7. The distribution of the number of nodes for a given range of fractional error in their closeness value, the closeness measured compared to the predicted value from the fit, shows variations between data sets but generally confirms that most individual nodes have a closeness that is reasonably similar to the prediction.
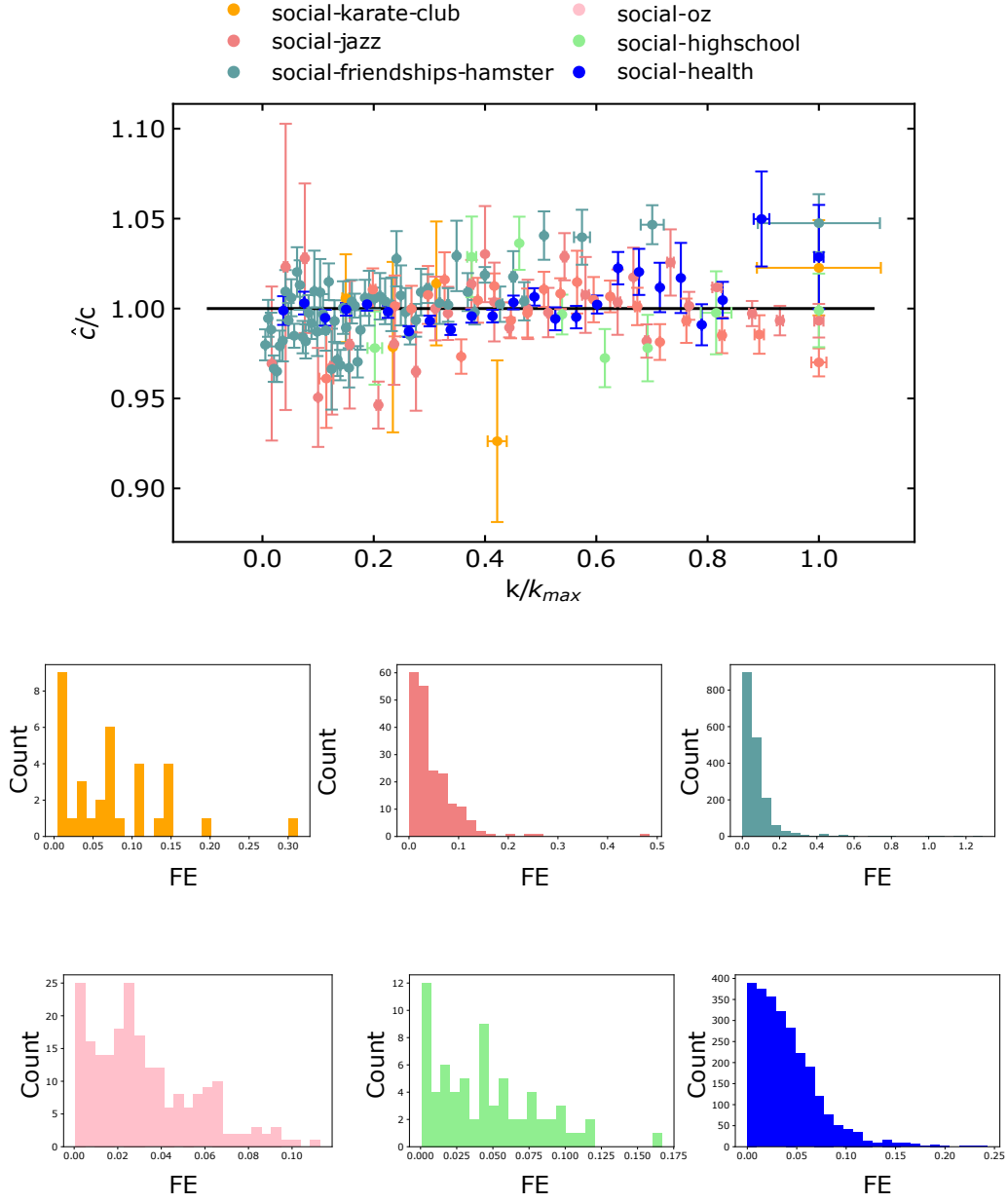
Supplementary Figure 4.6: Results for six friendship networks derived from real-world data, see Supplementary Table 3.1 for the statistics of each dataset. In the top plot, the horizontal axis is the degree divided by the largest degree in each data set. The vertical axis is the predicted value $\hat{c}$ for degree k divided by the equivalent measured value $c = \langle c \rangle_k$ averaged over nodes with the same degree k. The predicted value comes from (1.17) using values $\bar{z}^{(\text{fit})}$ and $\beta^{(\text{fit})}$ obtained by fitting the data to (1.17). The error bars show the standard error of the mean. The histograms show the number of data points in each data set with a specified absolute value of the fractional error where FE $= (|c^{-1} - \hat{c}^{-1}|)/\hat{c}^{-1}$. We can see that even for a small network, such as the the Karate club data set, our conjecture (1.17) is successful.
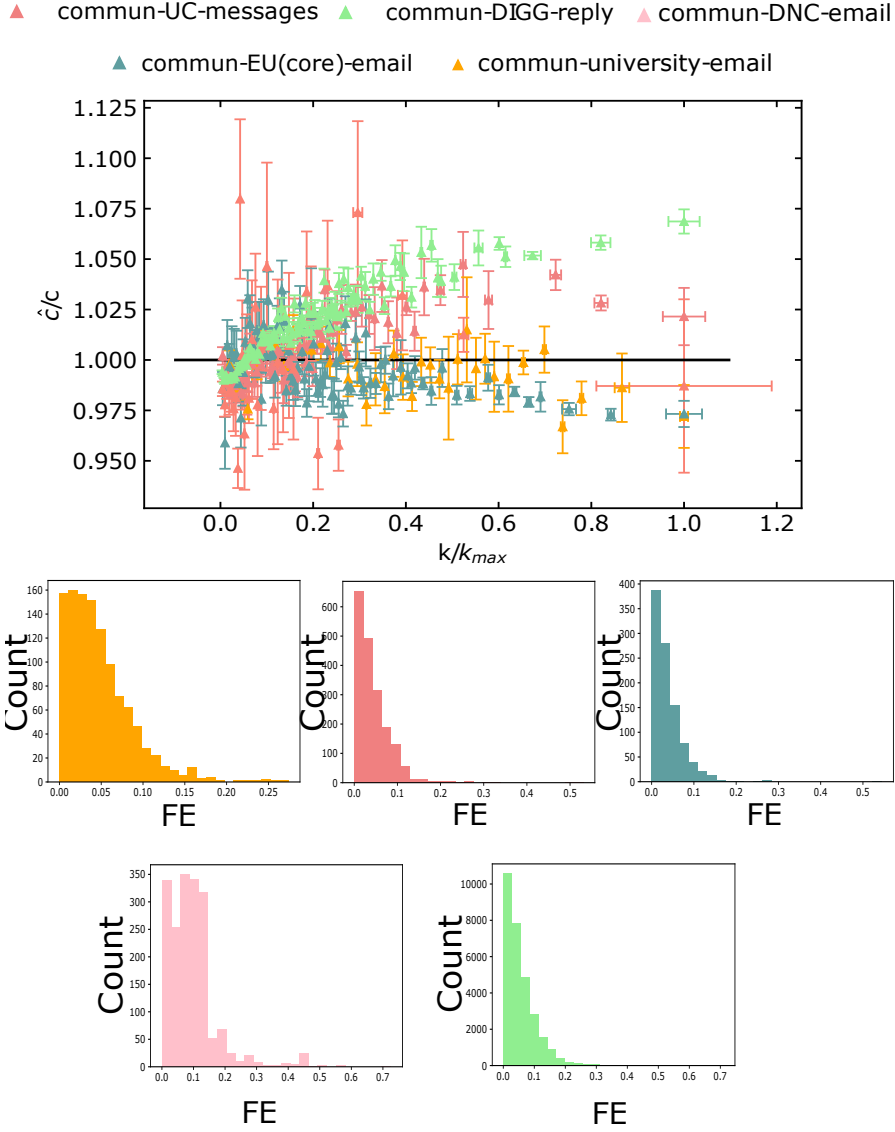
Supplementary Figure 4.7: Results for five communication networks derived from real-world data, see Supplementary Table 3.1 for the statistics of each dataset. In the top plot, the horizontal axis is the degree divided by the largest degree in each data set. The vertical axis is the predicted value $\hat{c}$ for degree k divided by the equivalent measured value $c = \langle c \rangle_k$ averaged over nodes with the same degree k. The predicted value comes from (1.17) using values $\bar{z}^{(\text{fit})}$ and $\beta^{(\text{fit})}$ obtained by fitting the data to (1.17). The error bars show the standard error of the mean. The histograms show the number of data points in each data set with a specified absolute value of the fractional error where $\text{FE} = (|c^{-1} - \hat{c}^{-1}|/\hat{c}^{-1})$. The poor fit of `commun-DIGG-reply` reflected in the bad χ_r^2 value is clearly visible here.

4.4 Additional Results for the Netzschleuder Networks

Further information on 112 networks taken from the Netzschleuder repository [44]. These were the networks which would download and run automatically using our code without any additional work. The information is provided here is available as a spreadsheet file provided `extra_result220208tidy.xlsx` in [25]. We have split the results into three tables: in Supplementary Table 4.1 are the 49 networks with χ_r^2 of 2.0 or more, in Supplementary Table 4.2 are the 49 networks with an excellent fit to our degree-closeness relationship (1.17), and finally in Supplementary Table 4.3 are the 12 networks where the topology in the simple graph representation used in our analysis means their results are of little relevance here.

The columns of tables 4.1, 4.2 and 4.3 are labelled as follows:

ID	internal identification number.
Name	name of the Network Netzschleuder repository [44].
N	Number of nodes
$\langle k \rangle$	Average degree
ρ	Density of network, $\rho = \langle k \rangle / (N-1)$
Rk	Rank of network by χ_r^2 value
χ_r^2	Reduced chi-square value for fit of closeness to (1.17)
$\bar{z}^{(\mathrm{fit})}$	Growth parameter value found by a fit to (1.17)
$\bar{z}_{\mathrm{err}}^{(\mathrm{fit})}$	Error in $\bar{z}$
$\beta^{(\mathrm{fit})}$	Value of β parameter found by a fit to (1.17)
$\beta_{\mathrm{err}}^{(\mathrm{fit})}$	Error in $\hat{\beta}$
β	Value of beta obtained using $\bar{z}$ value in formula (1.18)
β_{err}	Error in β
ρ^{P}	Pearson correlation measure
ρ^{IP}	Improved Pearson correlation measure
$\langle \ell \rangle$	Average path length in network
$\langle \ell \rangle^{\mathrm{FP}}$	avgPathLengthFitPredicted
$\langle \ell \rangle_{\mathrm{err}}^{\mathrm{FP}}$	Error in $\langle \ell \rangle^{\mathrm{FP}}$
$\langle \ell \rangle^{\mathrm{zP}}$	avgPathLengthzbarPredicted
$\langle \ell \rangle_{\mathrm{err}}^{\mathrm{zP}}$	Error in $\langle \ell \rangle^{\mathrm{zP}}$

ID	Name	N	$\langle k \rangle$	ρ density	Rk	χ^2_r	$\bar{z}^{(\text{fit})}$	$\bar{z}^{(\text{fit})}_{\text{err}}$	$\beta^{(\text{fit})}$	$\beta^{(\text{fit})}_{\text{err}}$	β	β_{err}	ρ^{P}	ρ^{IP}	$\langle \ell \rangle$	$\langle \ell \rangle^{\text{FP}}$	$\langle \ell \rangle^{\text{FP}}_{\text{err}}$	$\langle \ell \rangle^{\text{zP}}$	$\langle \ell \rangle^{\text{zP}}_{\text{err}}$
47	un_migrations	231	72.7	0.32	1	3187.1	37	4	2.81	0.03	2.53	0.05	1.0	-0.9	1.7	1.68	0.05	1.412	0.014
77	fediverse	816	280.9	0.34	2	2941.2	1170	150	2.370	0.014	1.95	0.02	1.0	-0.9	1.7	1.671	0.019	1.251	0.005
67	collins_yeast	1004	16.6	0.0166	3	162.2	8	0.9	6.60	0.06	4.39	0.18	0.7	-0.5	5.5	5.55	0.09	3.34	0.13
72	dnc	849	24.5	0.029	4	60.1	1200	500	3.01	0.02	1.95	0.05	0.7	-0.5	2.8	2.76	0.03	1.69	0.04
29	sp_hospital	75	30.4	0.41	5	37.3	15.9	1.3	2.78	0.04	2.61	0.05	1.0	-1	1.6	1.60	0.05	1.423	0.015
58	plant_pol_kato	768	3.1	0.0040	6	34.1	29	5	4.040	0.015	2.99	0.11	0.5	-0.6	3.9	3.903	0.017	2.85	0.10
50	contact	274	15.5	0.057	7	28.0	88	6	2.756	0.007	2.26	0.02	1.0	-1	2.4	2.424	0.009	1.929	0.014
27	law_firm	71	28.4	0.41	8	20.7	16.1	1.3	2.78	0.03	2.58	0.05	1.0	-1	1.6	1.60	0.05	1.395	0.013
32	polbooks	105	8.4	0.081	9	19.5	12	4	3.86	0.11	2.9	0.3	0.6	-0.6	3.1	3.08	0.15	2.14	0.16
14	montreal	29	5.2	0.186	10	14.1	8.3	0.9	2.77	0.04	2.67	0.09	0.9	-1	2.2	2.15	0.05	2.06	0.06
42	jazz_collab	198	27.7	0.141	11	13.1	15.3	1.7	3.35	0.05	2.98	0.08	0.9	-0.9	2.2	2.24	0.07	1.87	0.04
24	terrorists_911	62	4.9	0.080	12	12.8	8.6	2	3.55	0.08	3.0	0.2	0.8	-0.8	2.9	2.95	0.10	2.39	0.16
89	anybeat	8518	9.6	0.00113	13	9.8	6600	2000	2.976	0.006	2.03	0.04	0.2	-0.3	2.9	2.856	0.007	1.91	0.03
80	jung	77	7.6	0.100	14	9.6	8.2	1.6	3.28	0.08	3.1	0.2	0.8	-0.8	2.5	2.46	0.11	2.32	0.13
82	jdk	77	7.6	0.100	15	9.6	8.2	1.6	3.28	0.08	3.1	0.2	0.8	-0.8	2.5	2.46	0.11	2.32	0.13
49	sp_primary_school	242	68.7	0.29	16	8.2	24.3	0.7	3.032	0.013	2.75	0.02	1.0	-1	1.7	1.732	0.018	1.450	0.005
109	epinions_trust	32223	21.2	0.00066	17	7.7	31.5	0.4	4.198	0.003	4.03	0.01	0.5	-0.8	3.7	3.676	0.004	3.510	0.011
37	foodweb_baywet	103	30.1	0.30	18	7.6	22	1.6	2.80	0.03	2.53	0.04	1.0	-1	1.7	1.72	0.04	1.459	0.013
56	celegans_metabolic	453	8.9	0.0197	19	7.0	20	4	3.25	0.04	3.06	0.14	0.6	-0.6	2.7	2.66	0.06	2.48	0.10
16	windsurfers	43	15.6	0.37	20	6.0	10.5	1.2	2.80	0.06	2.67	0.09	1.0	-1	1.7	1.67	0.08	1.53	0.03
46	residence_hall	214	17.1	0.080	21	5.1	12.2	1.1	3.47	0.04	3.20	0.08	0.9	-0.9	2.4	2.38	0.06	2.11	0.04
71	plant_pol_robertson	1882	16.2	0.0086	22	5.0	68	5	3.718	0.009	2.80	0.03	0.7	-0.8	3.3	3.306	0.011	2.39	0.02
43	cintestinalis	201	25.3	0.127	23	4.7	15.1	1.1	3.27	0.03	3.00	0.05	0.9	-0.9	2.2	2.15	0.04	1.89	0.03
18	macaque_neural	47	13.3	0.29	24	4.2	9.7	1.2	2.93	0.06	2.76	0.10	1.0	-0.9	1.9	1.85	0.08	1.68	0.04
92	google	12354	23	0.00186	25	4.2	1E+20	7E+20	2.173	0.006	1.20	0.02	0.1	-0.1	2.1	2.118	0.008	1.148	0.015
3	kangaroo	17	10.7	0.67	26	4.1	12.3	2	2.24	0.06	2.18	0.08	1.0	-1	1.4	1.36	0.08	1.30	0.02
63	crime	1263	2.2	0.00174	27	3.8	3.1	0.4	13.41	0.08	7.4	0.7	0.3	-0.2	13	12.96	0.10	7.0	0.7
81	reactome	5973	48.8	0.0082	28	3.8	14.6	0.7	5.27	0.02	4.29	0.06	0.6	-0.6	4.2	4.21	0.03	3.23	0.04
13	dutch_criticism	35	4.6	0.135	29	3.7	3.9	0.5	3.66	0.10	3.8	0.3	0.9	-0.9	2.7	2.71	0.13	2.81	0.16
28	blumenau_drug	75	4.8	0.065	30	3.6	10.6	1.7	3.32	0.04	2.89	0.13	0.8	-0.9	2.8	2.81	0.06	2.38	0.10
55	eu_airlines	417	14.2	0.034	31	3.5	39	6	3.25	0.02	2.66	0.07	0.8	-0.8	2.8	2.76	0.03	2.17	0.05
0	sa_companies	11	2.4	0.24	32	3.4	2.27	0.23	3.56	0.14	4.0	0.3	0.9	-0.9	2.6	2.62	0.18	3.1	0.2
25	train_terrorists	64	7.6	0.121	33	3.4	10.5	2.6	3.40	0.08	2.8	0.2	0.9	-0.8	2.7	2.69	0.11	2.12	0.12
17	elite	44	4.5	0.105	34	3.1	5	0.8	3.59	0.09	3.5	0.2	0.9	-0.8	2.8	2.76	0.12	2.62	0.16
84	elec	1300	56.2	0.043	35	3.1	76	4	3.087	0.011	2.67	0.02	0.8	-0.9	2.2	2.243	0.015	1.823	0.011
6	november17	22	6	0.29	36	3.0	9.7	2.1	2.66	0.07	2.43	0.14	0.9	-0.9	2	1.98	0.10	1.75	0.08
107	facebook_wall	30793	9.2	0.00030	37	2.9	6.52	0.05	6.187	0.005	6.61	0.03	0.7	-0.8	5.3	5.288	0.006	5.71	0.02
110	slashdot_zoo	26997	21.3	0.00079	38	2.5	38.1	0.5	4.019	0.003	3.82	0.01	0.6	-0.9	3.4	3.446	0.003	3.249	0.009
104	inploid	4542	13.4	0.0030	39	2.5	400	70	2.843	0.010	2.41	0.04	0.2	-0.5	2.5	2.544	0.013	2.11	0.03
64	polblogs	793	34	0.043	40	2.4	123	12	3.036	0.013	2.39	0.03	0.8	-0.9	2.4	2.424	0.018	1.781	0.016
21	cs_department	61	11.6	0.193	41	2.4	9.2	1.1	3.10	0.06	2.92	0.10	0.9	-0.9	2.1	2.06	0.08	1.88	0.05
59	unicodelang	858	2.9	0.0034	42	2.4	13.6	2.2	4.41	0.02	3.64	0.17	0.4	-0.5	4.2	4.21	0.03	3.44	0.16
75	openflights	3147	11.9	0.0038	43	2.3	20	1.3	4.411	0.014	3.72	0.06	0.6	-0.6	3.9	3.924	0.018	3.24	0.05
87	wiki_rfa	2449	78.6	0.032	44	2.2	102	5	3.151	0.009	2.70	0.02	0.7	-0.9	2.3	2.292	0.013	1.837	0.009
51	celegansneural	239	14.4	0.061	45	2.2	12	1.2	3.48	0.04	3.26	0.09	0.8	-0.9	2.5	2.49	0.06	2.27	0.05
73	interactome_figeys	8	3	0.43	46	2.0	5.2	2	2.30	0.16	2.4	0.3	0.9	-0.9	1.7	1.7	0.2	1.75	0.18
95	internet_as	22963	4.2	0.00018	47	2.0	24.3	1.1	4.061	0.004	4.18	0.04	0.2	-0.4	3.8	3.842	0.005	3.96	0.04
35	football	115	10.7	0.094	48	2.0	26	28	3.2	0.2	2.5	0.5	0.3	-0.3	2.5	2.5	0.3	1.8	0.3
36	football_tsevans	115	10.7	0.094	49	2.0	26	28	3.2	0.2	2.5	0.5	0.3	-0.3	2.5	2.5	0.3	1.8	0.3

Supplementary Table 4.1: Further information on 49 of the 112 networks taken from the Netzschleuder repository [44]. These have $\chi^2_r \geq 2.0$ when fitting the closeness-degree data to the relationship (1.17). Taken from the file extra_result220208tidy.xlsx provided in [25].

ID	Name	N	$\langle k\rangle$	ρ density	Rk	χ_r^2	$\bar{z}^{(\mathrm{fit})}$	$\bar{z}^{(\mathrm{fit})}_{\mathrm{err}}$	$\beta^{(\mathrm{fit})}$	$\beta^{(\mathrm{fit})}_{\mathrm{err}}$	β	β_{err}	ρ^{P}	ρ^{IP}	$\langle\ell\rangle$	$\langle\ell\rangle^{\mathrm{FP}}$	$\langle\ell\rangle^{\mathrm{FP}}_{\mathrm{err}}$	$\langle\ell\rangle^{z\mathrm{P}}$	$\langle\ell\rangle^{z\mathrm{P}}_{\mathrm{err}}$
103	email_enron	33696	10.7	0.00032	50	1.8	132	7	4.297	0.004	3.14	0.02	0.4	-0.4	4	4.025	0.005	2.87	0.02
65	new_zealand_collab	1463	5.8	0.0040	51	1.8	169	26	2.920	0.008	2.43	0.04	0.5	-0.7	2.7	2.749	0.009	2.25	0.04
102	topology	34761	6.2	0.00018	52	1.7	39	1.3	3.989	0.003	3.87	0.03	0.2	-0.5	3.8	3.768	0.004	3.65	0.03
48	physician_trust	95	8	0.085	53	1.7	10.9	1.4	3.10	0.05	2.96	0.11	0.9	-0.9	2.5	2.46	0.06	2.13	0.06
30	lesmis	77	6.6	0.087	54	1.6	23	8	3.10	0.06	2.41	0.16	0.8	-0.7	2.6	2.64	0.08	1.95	0.11
31	sp_office	92	16.4	0.180	55	1.6	17.5	2	2.90	0.04	2.62	0.07	0.9	-0.9	2	1.96	0.05	1.68	0.03
53	facebook_friends	329	11.9	0.036	56	1.6	17	4	4.29	0.07	3.07	0.18	0.6	-0.5	3.6	3.58	0.09	2.36	0.12
88	dblp_cite	240	7.2	0.030	57	1.6	11.5	2.4	3.78	0.06	3.3	0.2	0.6	-0.6	3.1	3.08	0.09	2.61	0.14
86	sp_infectious	410	13.5	0.033	58	1.6	7.8	0.9	4.78	0.06	4.00	0.16	0.7	-0.7	3.6	3.63	0.09	2.86	0.10
34	adjnoun	112	7.6	0.068	59	1.5	11.7	1.5	3.22	0.04	2.97	0.10	0.8	-0.9	2.5	2.54	0.05	2.28	0.07
1	new_guinea_tribes	16	7.3	0.49	60	1.5	6.6	0.6	2.57	0.05	2.56	0.08	1.0	-1	1.5	1.54	0.07	1.53	0.03
57	wiki_science	677	19.3	0.029	61	1.5	9.9	0.5	4.50	0.03	3.91	0.07	0.8	-0.9	3.4	3.43	0.04	2.84	0.04
15	ceo_club	40	4.8	0.123	62	1.4	14	8	2.99	0.13	2.5	0.4	0.7	-0.6	2.4	2.45	0.18	1.9	0.2
44	interactome_pdz	161	2.6	0.0163	63	1.3	5.1	1.1	5.70	0.08	4.2	0.4	0.5	-0.5	5.3	5.33	0.10	3.9	0.4
85	chess	5643	18.9	0.0033	64	1.3	12.67	0.27	4.541	0.009	4.46	0.03	0.8	-0.8	3.6	3.603	0.012	3.52	0.02
62	faa_routes	792	4.3	0.0054	65	1.3	4.13	0.29	6.07	0.05	5.8	0.2	0.6	-0.6	5.2	5.21	0.06	5.0	0.2
111	prosper	3513	21.2	0.0060	66	1.3	60.7	2.7	3.510	0.007	3.00	0.02	0.7	-0.8	2.9	2.858	0.010	2.348	0.015
79	bitcoin_trust	4709	8.2	0.00174	67	1.3	35.2	1.9	3.846	0.007	3.40	0.04	0.5	-0.7	3.5	3.512	0.009	3.06	0.03
66	netscience	379	4.8	0.0127	68	1.3	14	8	6.55	0.12	3.3	0.5	0.3	-0.2	6	6.04	0.16	2.8	0.4
26	highschool	67	8	0.121	69	1.3	5	0.6	3.83	0.09	3.7	0.2	0.8	-0.9	2.6	2.60	0.13	2.48	0.11
91	foldoc	13274	13.7	0.00103	70	1.3	8.12	0.16	5.044	0.011	5.61	0.04	0.5	-0.7	3.9	3.872	0.015	4.44	0.03
68	interactome_stelzl	1493	4	0.0027	71	1.3	6.2	0.4	5.23	0.02	5.11	0.14	0.5	-0.6	4.8	4.78	0.03	4.66	0.12
20	contiguous_usa	49	4.4	0.092	72	1.2	1.85	0.15	6.4	0.3	7.2	0.7	0.7	-0.7	4.2	4.2	0.4	5.0	0.4
8	zebras	23	9.1	0.41	73	1.2	6.4	0.9	2.95	0.09	2.78	0.14	0.9	-0.9	1.9	1.86	0.12	1.70	0.06
74	interactome_vidal	2783	4.3	0.00155	74	1.2	7.9	0.4	5.299	0.016	4.92	0.10	0.5	-0.6	4.8	4.84	0.02	4.46	0.09
33	game_thrones	107	6.6	0.062	75	1.2	8.9	1.3	3.58	0.05	3.21	0.15	0.8	-0.8	2.9	2.90	0.07	2.54	0.11
96	word_assoc	7751	56.8	0.0073	76	1.2	111.5	3.4	3.361	0.005	2.91	0.01	0.8	-0.9	2.5	2.550	0.007	2.096	0.007
105	pgp_strong	39796	9.9	0.00025	77	1.2	12.45	0.32	6.086	0.008	5.25	0.04	0.4	-0.4	5.5	5.498	0.010	4.67	0.04
22	dolphins	62	5.1	0.084	78	1.2	4.9	0.9	4.24	0.12	3.7	0.3	0.7	-0.7	3.4	3.36	0.15	2.8	0.2
94	marvel_universe	19182	10	0.00052	79	1.2	22.2	0.7	4.957	0.006	4.21	0.03	0.3	-0.6	4.5	4.450	0.008	3.71	0.03
108	slashdot_threads	16377	7.9	0.00048	80	1.2	24.7	0.5	4.388	0.003	4.05	0.02	0.6	-0.8	4	3.971	0.004	3.637	0.017
76	bitcoin_alpha	3235	8.2	0.0025	81	1.1	27.4	1.6	3.878	0.009	3.47	0.04	0.6	-0.7	3.5	3.514	0.011	3.10	0.04
101	linux	913	8.8	0.0096	82	1.1	30	4	3.41	0.02	3.02	0.08	0.6	-0.7	2.9	2.93	0.03	2.55	0.06
5	moreno_taro	22	3.5	0.167	83	1.1	5.2	2.6	3.2	0.2	3.0	0.6	0.6	-0.6	2.5	2.5	0.3	2.2	0.4
83	advogato	3140	20.5	0.0065	84	1.1	42.7	1.8	3.564	0.008	3.16	0.03	0.6	-0.8	2.9	2.941	0.011	2.540	0.018
98	movielens_100k	23761	6	0.00025	85	1.1	17.24	0.31	5.095	0.003	4.58	0.02	0.5	-0.7	4.8	4.778	0.004	4.26	0.02
69	bible_nouns	1707	10.6	0.0062	86	1.1	13.2	0.9	4.09	0.02	3.93	0.08	0.6	-0.7	3.4	3.38	0.03	3.22	0.06
38	revolution	141	2.3	0.0164	87	1.1	33	16	3.30	0.03	2.4	0.2	0.5	-0.5	3.2	3.23	0.03	2.4	0.2
100	nematode_mammal	26197	4.5	0.00017	88	1.1	7.8	0.22	7.047	0.009	6.03	0.07	0.4	-0.4	6.7	6.662	0.010	5.65	0.07
54	london_transport	369	2.3	0.0063	89	1.1	1.251	0.027	17.2	0.4	24.2	1.9	0.5	-0.5	13.7	13.7	0.5	20.7	1.6
99	digg_reply	6746	9.8	0.00145	90	1.1	17.3	0.4	4.367	0.005	4.13	0.02	0.8	-0.9	3.7	3.714	0.007	3.480	0.019
52	marvel_partnerships	181	2.5	0.0139	91	1.1	2.6	0.4	8.65	0.18	6.6	0.9	0.4	-0.4	7.9	7.9	0.2	5.9	0.8
41	student_cooperation	141	3.6	0.026	92	1.1	1.57	0.11	9.2	0.4	11.4	1.4	0.4	-0.5	6.5	6.5	0.6	8.7	1.0
70	interactome_yeast	1458	2.7	0.00185	93	1.1	4.6	0.4	7.22	0.04	5.9	0.3	0.4	-0.4	6.8	6.81	0.05	5.5	0.3
60	uni_email	1133	9.6	0.0085	94	1.1	12.6	0.5	4.309	0.014	3.83	0.05	0.8	-0.9	3.6	3.606	0.018	3.12	0.04
93	escorts	15810	4.9	0.00031	95	1.0	9.05	0.25	6.190	0.008	5.46	0.06	0.1	-0.5	5.8	5.785	0.009	5.05	0.05
90	chicago_road	12978	3.2	0.00025	96	1.0	5.7	2.1	42.58	0.14	6.5	1.2	0.1	0	42	42.0	0.2	5.9	1.1
97	cora	3991	8.3	0.0021	97	1.0	5.65	0.21	6.71	0.03	5.89	0.11	0.5	-0.6	5.7	5.65	0.03	4.84	0.08
78	power	4941	2.7	0.00055	98	1.0	2.69	0.17	19.80	0.06	9.7	0.5	0.2	-0.2	19	18.99	0.08	8.9	0.5
61	euroroad	1039	2.5	0.0024	99	1.0	1.299	0.031	21.5	0.3	25.3	1.9	0.4	-0.3	18.4	18.4	0.4	22.1	1.7
106	paris_transportation	11	1.8	0.180	100	0.0	12.915	6E-15	1.9	0	1.99	0	1.0	-1	1.8	1.80	0	1.91	0

Supplementary Table 4.2: Further information on 51 of the 112 networks taken from the Netzschleuder repository [44]. These networks fit the closeness-degree fitting to the relationship (1.17) extremely well with a reduced χ^2 between 1.0 and 2.0. Taken from the file extra_result220208tidy.xlsx provided in [25].

ID	Name	N	$\langle k \rangle$	ρ density	Rk	χ_r^2	$\bar{z}^{(\mathrm{fit})}$	$\bar{z}^{(\mathrm{fit})}_{\mathrm{err}}$	$\beta^{(\mathrm{fit})}$	$\beta^{(\mathrm{fit})}_{\mathrm{err}}$	β	β_{err}	ρ^{P}	ρ^{IP}	$\langle \ell \rangle$	$\langle \ell \rangle^{\mathrm{FP}}$	$\langle \ell \rangle^{\mathrm{FP}}_{\mathrm{err}}$	$\langle \ell \rangle^{z\mathrm{P}}$	$\langle \ell \rangle^{z\mathrm{P}}_{\mathrm{err}}$
2	rhesus_monkey	16	8.6	0.57	101	0.0	8.7	0.9	2.39	0.05	2.35	0.07	1.0	-1	1.4	1.43	0.06	1.39	0.02
9	cattle	20	12.1	0.64	102	0.0	5.44	0.31	2.82	0.05	2.87	0.06	1.0	-1	1.4	1.36	0.07	1.419	0.015
4	high_tech_company	21	15.1	0.76	103	0.0	4.24	0.16	3.11	0.05	3.23	0.06	1.0	-1	1.2	1.24	0.07	1.362	0.009
10	moreno_sheep	22	13.2	0.63	104	0.0	11.4	1.5	2.41	0.06	2.33	0.08	1.0	-1	1.4	1.37	0.08	1.297	0.019
40	foodweb_little_rock	22	13.2	0.63	105	0.0	7.4	0.6	2.64	0.05	2.63	0.07	1.0	-1	1.4	1.37	0.08	1.359	0.017
7	bison	26	17.1	0.68	106	0.0	6.5	0.5	2.81	0.06	2.84	0.07	1.0	-1	1.3	1.32	0.08	1.347	0.014
11	7th_graders	29	17.2	0.61	107	0.0	7.2	0.6	2.80	0.06	2.80	0.07	1.0	-1	1.4	1.38	0.08	1.381	0.017
12	college_freshmen	31	26.8	0.89	108	0.0	3.6	0.08	3.67	0.05	3.81	0.05	1.0	-1	1.1	1.11	0.06	1.253	0.003
23	macaques	38	26.3	0.71	109	0.0	4.67	0.17	3.40	0.05	3.48	0.06	1.0	-1	1.3	1.29	0.07	1.368	0.008
19	sp_kenyan_households	47	21.4	0.47	110	0.0	10.5	0.9	2.81	0.05	2.70	0.06	1.0	-1	1.5	1.53	0.07	1.426	0.018
39	email_company	126	49.3	0.39	111	0.0	24.6	3.2	2.77	0.05	2.54	0.06	1.0	-0.9	1.6	1.61	0.07	1.371	0.017
45	fao_trade	133	97.8	0.74	112	0.0	5.45	0.18	3.93	0.05	3.99	0.06	1.0	-1	1.3	1.26	0.08	1.315	0.007

Supplementary Table 4.3: Further information on 12 of the 112 networks taken from the Netzschleuder repository [44]. Our automated code was able to download and run on these networks but in the form used here these give no insight into the closeness-degree relationship (1.17). This can be seen by the 0.0 for the reduced χ_r^2 and the average path length is less than 2.0. Typically the information in the topology of the simple graph representation of the data used in our code is trivial and all the interest in these examples lies in weights on the edges. This table is provided only for completeness. Taken from the file extra_result220208tidy.xlsx provided in [25].

Supplementary References

[1] Zachary, W. Information-flow model for conflict and fission in small-groups. *Journal Of Anthropological Research* **33**, 452—473 (1977).

[2] Wasserman, S. & Faust, K. *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)* (Cambridge University Press, 1994).

[3] Newman, M. *Networks: an introduction* (Oxford University Press, 2010).

[4] Latora, V., Nicosia, V. & Russo, G. *Complex Networks: Principles, Methods and Applications* (Cambridge University Press, 2017).

[5] Borgatti, S. P. & Everett, M. G. A graph-theoretic perspective on centrality. *Social Networks* **28**, 466–484 (2006).

[6] Harary, F. Status and contrastatus. *Sociometry* **22**, 23 (1959).

[7] Hage, P. & Harary, F. Eccentricity and centrality in networks. *Social Networks* **17**, 57–63 (1995).

[8] Wuchty, S. & Stadler, P. F. Centers of complex networks. *Journal of Theoretical Biology* **223**, 45–53 (2003).

[9] Sabidussi, G. The centrality index of a graph. *Psychometrika* **31**, 581–603 (1966).

[10] Freeman, L. C. Centrality in social networks conceptual clarification. *Social networks* **1**, 215–239 (1978).

[11] Molloy, M. & Reed, B. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms* **6**, 161–180 (1995).

[12] Milgram, S. The small world problem. *Psychology Today* (1967).

[13] Travers, J. & Milgram, S. An experimental study of the small world problem. *Sociometry* **32**, 425 (1969).

[14] Leskovec, J. & Horvitz, E. Planetary-scale views on an instant-messaging network. Tech. Rep., Microsoft Research (2007).

[15] Backstrom, L., Boldi, P., Rosa, M., Ugander, J. & Vigna, S. Four degrees of separation. In *Proceedings of the 3rd Annual ACM Web Science Conference on - WebSci '12* (ACM Press, 2012).

[16] Boldi, P. & Vigna, S. Four degrees of separation, really. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (IEEE, 2012).

[17] Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998). DOI http://dx.doi.org/10.1038/30918.

[18] Chung, F. & Lu, L. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences* **99**, 15879–15882 (2002).

[19] Dorogovtsev, S. N., Mendes, J. F. F. & Samukhin, A. N. Metric structure of random networks. *Nuclear Physics B* **653**, 307–338 (2003).

[20] Bollobás, B. Mathematical results on scale-free random graphs. In *Handbook of Graphs and Networks*, 1–37 (Wiley, 2003).

[21] Coscia, M. *The Atlas for the Aspiring Network Scientist* (Michele Coscia, 2021). arXiv:2101.00863.

[22] Baronchelli, A. & Loreto, V. Ring structures and mean first passage time in networks. *Physical Review E* **73**, 026103 (2006).

[23] Masuda, N., Porter, M. A. & Lambiotte, R. Random walks and diffusion on networks. *Physics Reports* (2017).

[24] Šubelj, L. Algorithms for spanning trees of unweighted networks. Tech. Rep., University of Ljubljana (2021). arXiv:2205.06628.

[25] Evans, T. S. & Chen, B. Linking the network centrality measures closeness and degree: Additional data. figshare. DOI https://doi.org/10.6084/m9.figshare.19216812.

[26] Kunegis, J. KONECT – The Koblenz Network Collection. In *Proc. Int. Conf. on World Wide Web Companion*, 1343–1350 (2013).

[27] Kunegis, J. The KONECT project. http://konect.cc/.

[28] Leskovec, J. & Krevl, A. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data (2014).

[29] Coleman, J. S. *Introduction to Mathematical Sociology* (London Free Press Glencoe, 1964).

[30] Gleiser, P. M. & Danon, L. Community Strucure in Jazz. *Advances in Complex Systems* **06**, 565–573 (2003). DOI https://doi.org/10.1142%2Fs0219525903001067.

[31] Freeman, L. C., Webster, C. M. & Kirke, D. M. Exploring social structure using dynamic three-dimensional color images. *Social Networks* **20**, 109–118 (1998).

[32] Moody, J. Peer influence groups: Identifying dense clusters in large networks. *Soc. Netw.* **23**, 261–283 (2001).

[33] Guimerà, R., Danon, L., Díaz-Guilera, A., Giralt, F. & Arenas, A. Self-similar community structure in a network of human interactions. *Physical Review E* **68**, 065103 (2003).

[34] Opsahl, T. & Panzarasa, P. Clustering in weighted networks. *Social Networks* **31**, 155–163 (2009).

[35] Leskovec, J., Kleinberg, J. & Faloutsos, C. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data* **1**, 2 (2007).

[36] Choudhury, M. D., Sundaram, H., John, A. & Seligmann, D. D. Social synchrony: Predicting mimicry of user actions in online social media. In *Proc. Int. Conf. on Comput. Science and Engineering*, 151–158 (2009).

[37] Ley, M. The DBLP computer science bibliography: Evolution, research issues, perspectives. In *Proc. Int. Symposium on String Process. and Inf. Retr.*, 1–10 (2002).

[38] McCallum, A. K., Nigam, K., Rennie, J. & Seymore, K. Automating the construction of internet portals with machine learning. *Information Retrieval* **3**, 127–163 (2000).

[39] Šubelj, L. & Bajec, M. Model of complex networks based on citation dynamics. In *Proc. of the WWW Workshop on Large Scale Network Analysis*, 527–530 (2013).

[40] Newman, M. E. J. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 404–409 (2001).

[41] Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **74** (2006).

[42] Batagelj, V. Pajek datasets. http://vlado.fmf.uni-lj.si/pub/networks/data/ (2017).

[43] Adamic, L. A. & Glance, N. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, 36–43 (ACM, 2005).

[44] Peixoto, T. P. The Netzschleuder network catalogue and repository. URL https://networks.skewed.de/.

[45] Peixoto, T. P. The graph-tool python library (2014). DOI https://doi.org/10.6084/m9.figshare.1164194.v14