Imperial College London

Approximating Networks with Shortest Path Trees

Harry Bennett & William Leach
Supervisor: Dr Tim Evans
Group: Theoretical Physics

Introduction

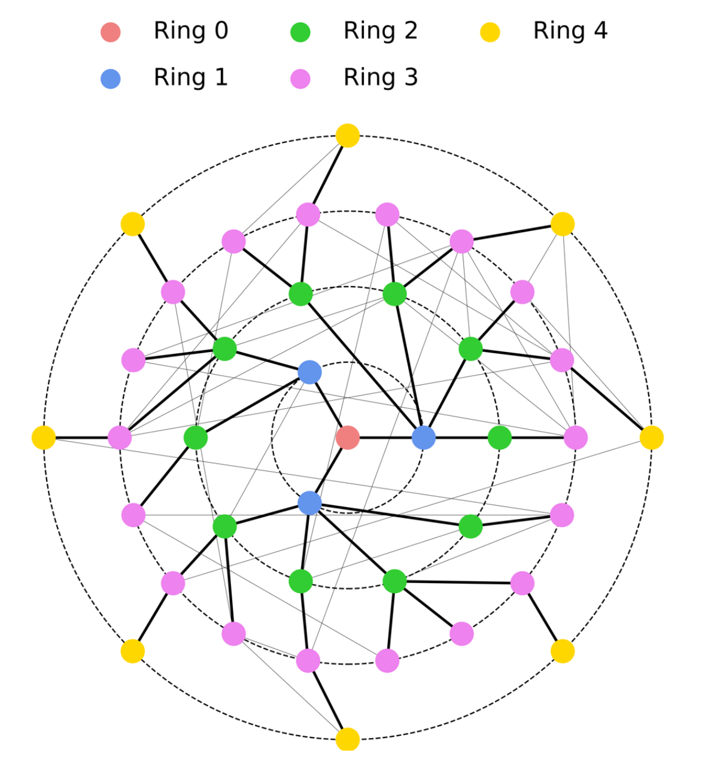

Figure 1: Ring Based Shortest Path Tree Approximation

Networks are a structured collection of objects called *nodes* (vertices), connected by *edges* (links). It is often important to assess the importance of nodes in a network – and centrality metrics are a core way of doing this. There are many types of centrality. Our research focusses on two types:

- **Degree** – The number of edges it has, denoted k.
- **Closeness** – The average distance of node from all others in a network [1], defined for node v as:

$$\frac{1}{c_v} = \frac{1}{N-1}\sum_{u\in\mathcal{V}/v} d_{uv}$$

d_{uv} → distance between nodes u and v
$\mathcal{V}$ → set of all nodes

Research Question: Can we form a relation between degree and closeness by approximating networks as shortest path trees? When does it work Well? What do the failures tell us? We follow on from the results of Evans and Chen [2]

Theory

Here we repeat the derivation done in [2]. It starts with noting the properties of degree and closeness are captured by rooted shortest path trees. These trees contain all nodes but only the edges necessary for shortest paths between them.

① Exponential Growth
- Trees grow exponentially out as series of rings
- Allow number of nodes on l^{th} ring to be $n_r(l) = k_r\bar{z}^l$
- $\bar{z}$ is growth parameter

② Statistical Similarity
- Take 'branches' of tree to start at each of k nearest neighbors'
- Assume branches, when averaged across all root nodes, are statistically similar

This growth can not continue forever, hench model as a sharp cut off L_r, meaning $n_r(l) = 0$ for $l > L_r$. Constraining by the number of nodes N we get:

$$N = 1 + \sum_{l=1}^{L_r} k_r\bar{z}^{l-1} = 1 + k_r\frac{(\bar{z}^{L_r} - 1)}{(\bar{z} - 1)} \quad \longrightarrow \quad L(N,k) \approx \frac{\ln(N(\bar{z}-1)/k)}{\ln(\bar{z})}$$

Moving to the definition of closeness, and reformulating in terms of rings and exponential growth returns a relation for closeness:

$$\sum_{u\in\mathcal{V}/r} d_{ur} = \sum_{l=1}^{L_r} l\cdot n(l) \rightarrow \frac{1}{c_r} = \frac{1}{N-1}\sum_{l=1}^{L_r} l\, k_r\bar{z}^{l-1} = \frac{k}{(N-1)}\left(\frac{(L_r+1)\bar{z}^{L_r}}{\bar{z}-1} - \frac{(\bar{z}^{L_r+1}-1)}{(\bar{z}-1)^2}\right)$$

Eliminating L and rearranging gives us an equation for inverse closeness of:

$$\boxed{\frac{1}{c_r} = -\frac{1}{\ln(\bar{z})}\ln(k_r) + \beta} \quad \longrightarrow \quad \beta = \left(\frac{1}{\bar{z}-1} - \frac{\ln(\bar{z}-1)}{\ln(\bar{z})}\right) + \frac{1}{\ln(\bar{z})}\ln(N)$$

Thus, from a simple approximation and ansatz we expect inverse closeness to be linear in the logarithm of degree.

Results

Method: We fit each network to the relation derived. We use reduced chi-square, χ_r^2 to grade the goodness of fit. $\chi_r^2 = 1$ implies a perfect fit, and $\chi_r^2 \gg 1$ a poor fit.

Artificial Networks: Generated Erdős–Rényi (ER) network [3] – where the only parameters are the number of nodes N, and the probability of an edge forming between nodes. We see The relation performs well here. With χ_r^2 just above 1.

Real Networks: Also processed 126 real world networks with up to 400,000 nodes from [4]. Fig 3. Shows the cumulative distribution of χ_r^2. We see relatively good fitting with extreme failures.

Density: This is the proportion of possible edges that are present in each network. We find that for real networks this is a large factor in how well the relation holds with more dense networks performing worse.

Clustering: The clustering coefficient for a network gives information on the level to which nodes form groups within the network [5]. We find that restricting clustering < 0.3 (so remove highly grouped networks) gives a large improvement in how well the networks fit.

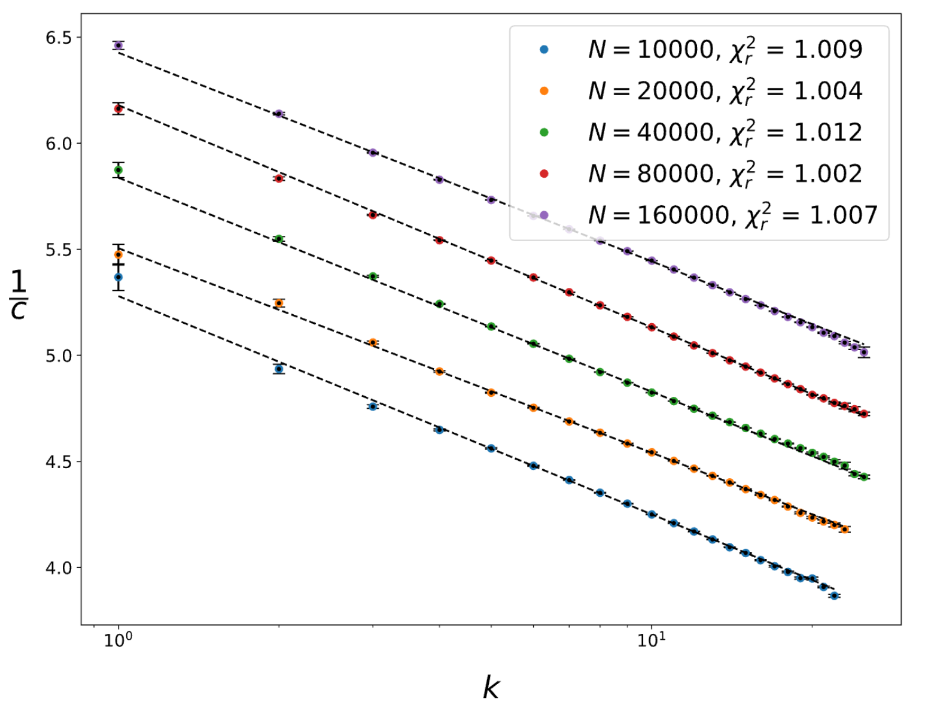

Figure 2: Inverse closeness against degree for an Erdős–Rényi network with average degree of 10. Several Different size networks are shown, and the error bars are the error in the mean. The χ_r^2 is also in the legend.

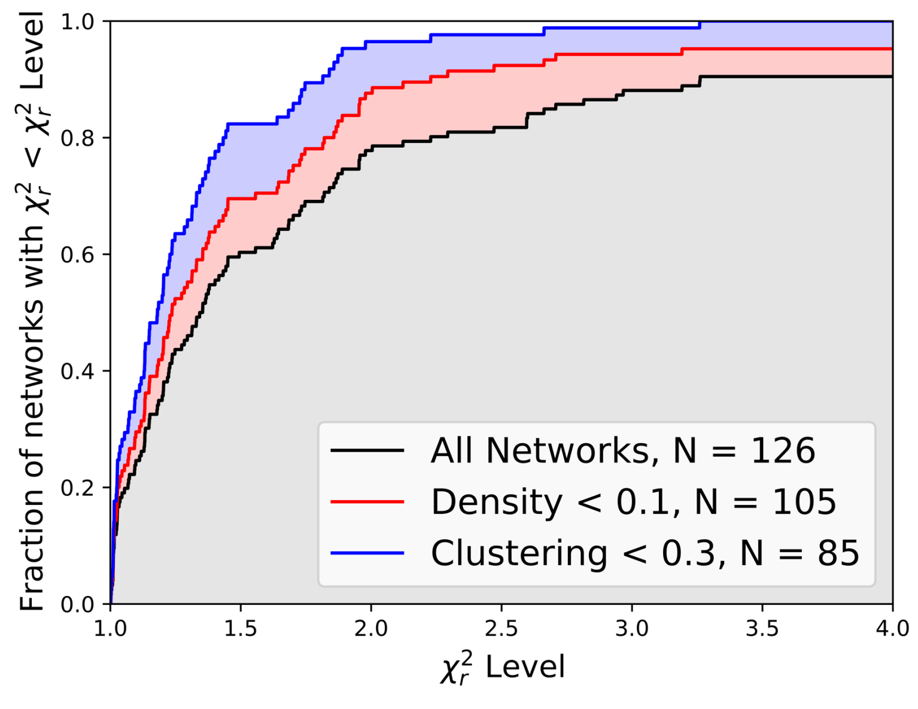

Figure 3: Results of processing real networks and associated goodness of fit. The fraction of networks that have a better (lower) χ_r^2 than a given level is displayed.

RANSAC

Hypothesis: Outlier points correspond to further network structure.

Test: Applied RANSAC algorithm which assumes some of data follows regression, the rest are outliers [6].

Result: Partitioning network in k, c^{-1} variables allows for weakly connected group structure to be found.

Impact: Potentially motivate using expected vs numerical results to investigate structure.

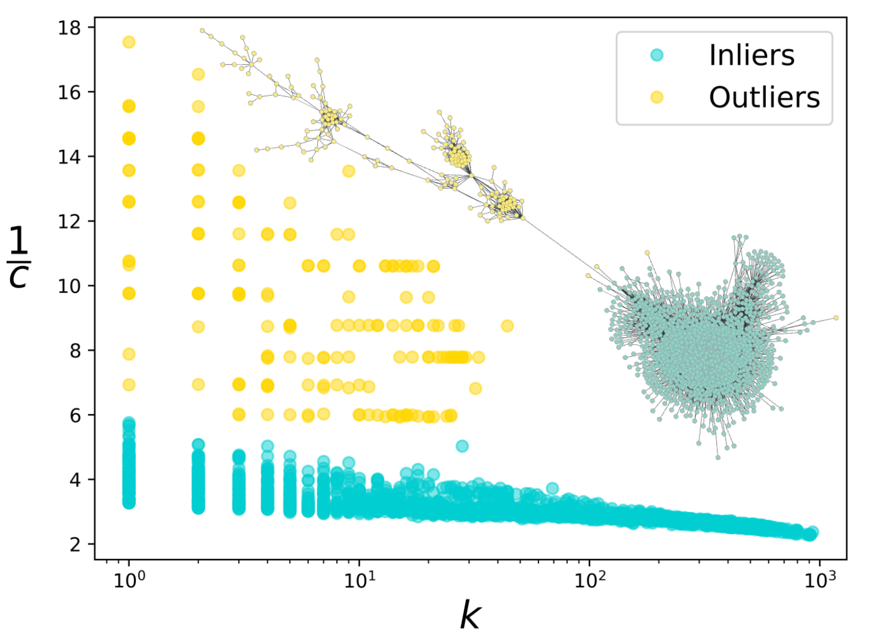

Figure 4: Graph of C. elegans Interactomes - Microarray. This is a protein interaction network. The data allows us to infer network structure.

Room for Improvement

Assumed Exponential growth and cut-off is imperfect. Testing by calculating this growth explicitly for an ER network, and averaging over all nodes, we see this clearly. This is plotted in Fig. 5. alongside the result for $\bar{z}$ and L_r found from fitting the same network for inverse closeness against degree. In future research this may provide motivation for improved ansatz - although the complexity of the relation would increase.

It is worth noting the contribution to closeness from higher rings goes roughly as $^1/_l$ - so these further nodes have less impact.

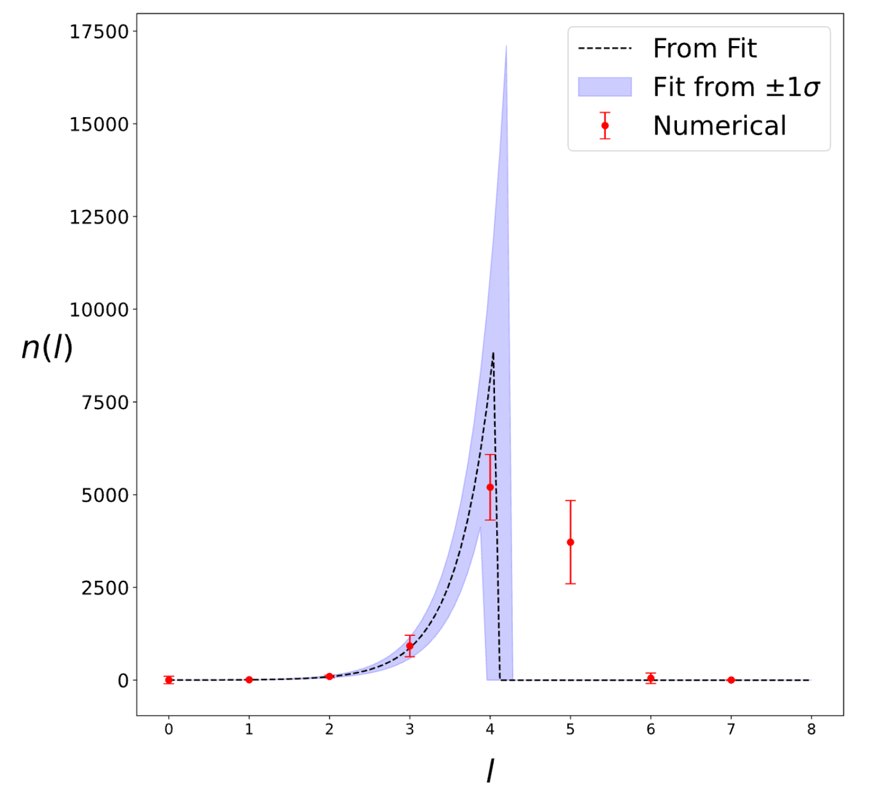

Figure 5: Numerically finding number of nodes at distance. The dashed line is formed using the growth and cut off parameters evaluated by fitting closeness and degree. N = 10000. $\langle k\rangle = 10$

Extending to Bipartite Networks

Bipartite networks consist of two distinct groups of nodes, where edges are only present between the two types. For example, customers and products. Bipartite networks are thus a special subset of the regular networks we have considered so far – and we may hope to learn more about the structure of these networks by extending the theory to them.

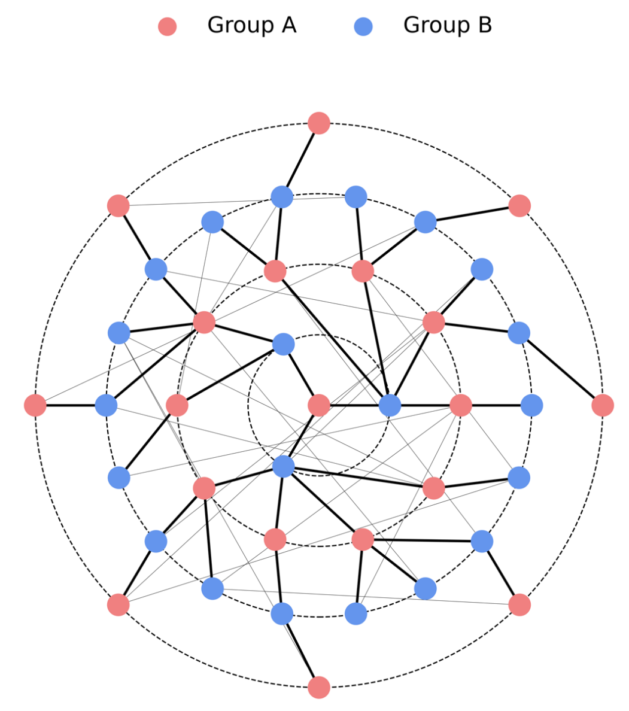

Figure 6: Bipartite Network

We label the two groups a and b. Each group has its own growth parameter, $\bar{\bar{z}}_a$ and $\bar{z}_b$. Hence for a node in group a

$$N = 1 + \sum_{l=1}^{L_r/2} k_r(1+\bar{z}_b)(\bar{z}_a\bar{z}_b)^{l-1}$$

Compare to old theory

$$L_r \rightarrow \frac{L_r}{2},\quad \bar{z} \rightarrow \bar{z}_a\bar{z}_b,\quad k_r \rightarrow k_r(1+\bar{z}_b)$$

Hence, we may write: $$\boxed{\frac{1}{c_r} = -\frac{2}{\ln(\bar{z}_a\bar{z}_b)}\ln(k_r(1+\bar{z}_b)) + \gamma}$$ [For type A nodes]

For nodes of type b, $\bar{\bar{z}}_a \rightarrow \bar{z}_b$. We expect two lines per network – with a common gradient. We may now repeat our results process – analyzing for both real and artificial networks. We find the artificial work well, but for many real networks – while there are two distinct lines – the gradient is not always shared.

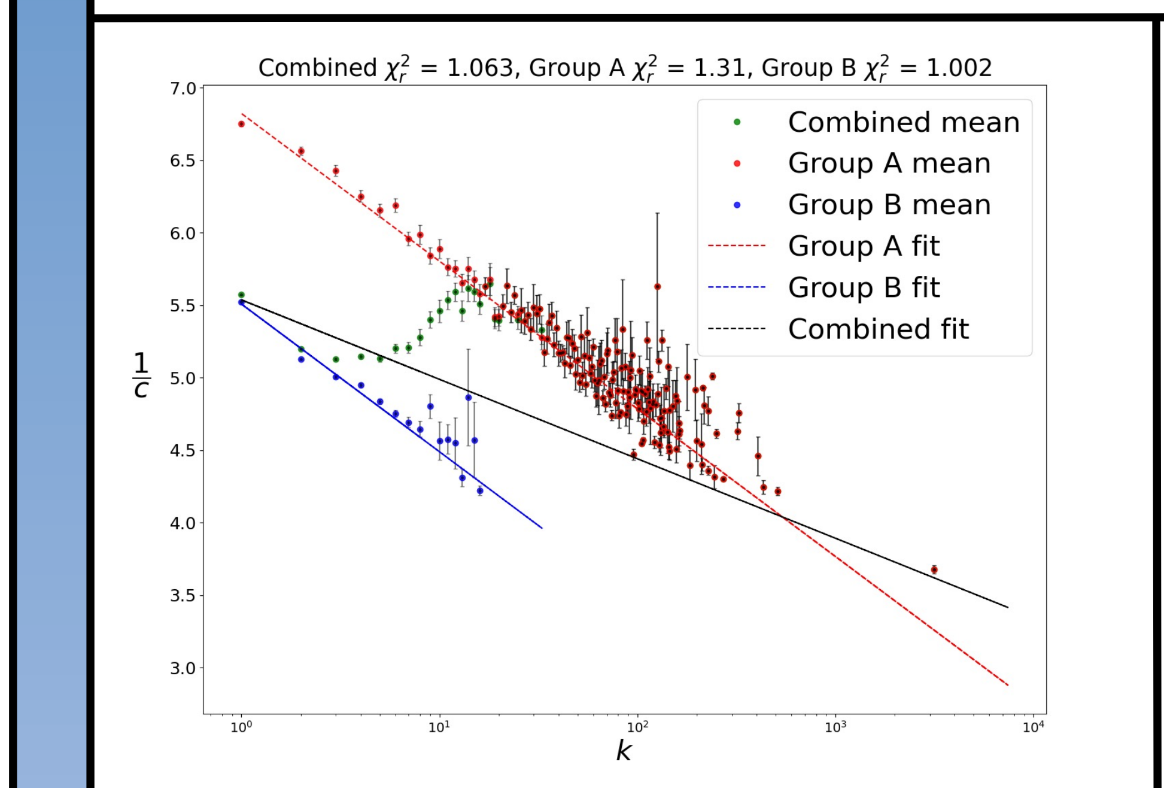

Figure 6: Inverse closeness vs degree for real world network. The red and blue data is mean from the partition by group, and the green is the combined average. *Network: DBpedia – Record Label.*

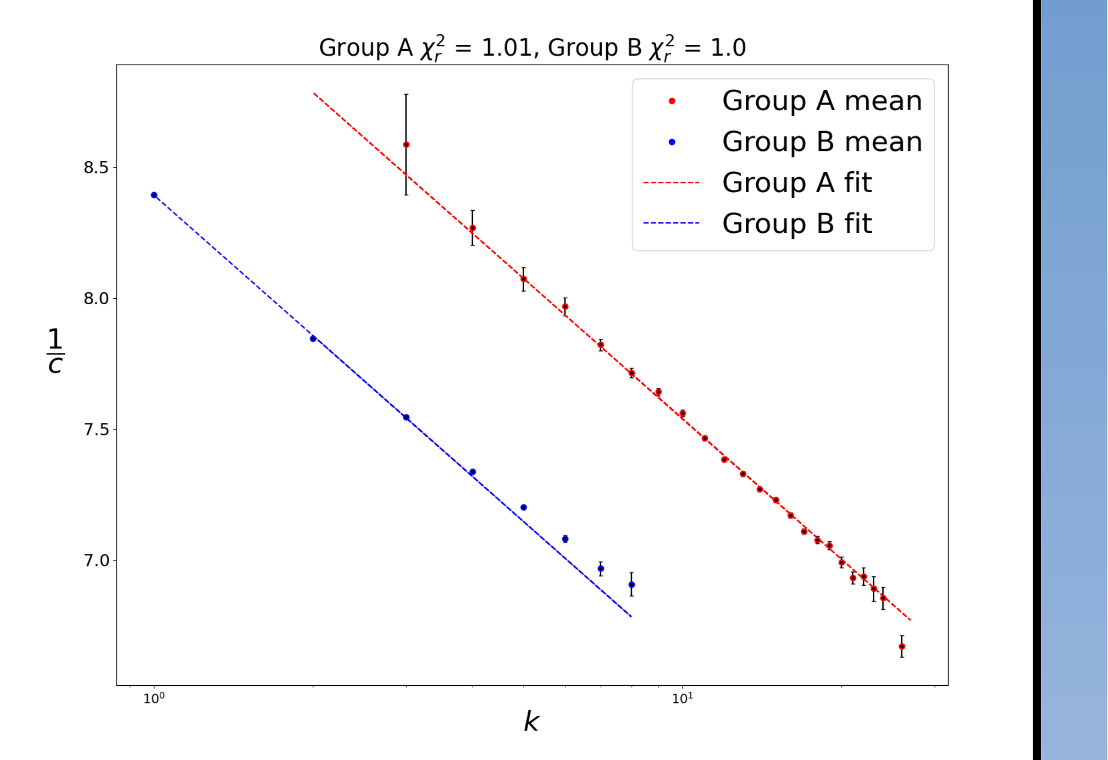

Figure 7: Inverse closeness vs degree for ER-Bipartite network. The blue and red data is the mean from the partition by group. Group A: N = 50000, Group B: N= 5000, p = 0.00025.

Conclusion & Outlook

1. Artificial and Real networks support new theory – but often real networks have extra structure that leads to a worse fit.
2. Can gain insights about networks through looking at 'failures'.
3. Theory for bipartite networks is successful – but we don't always see common gradient between the two lines.

Outlook and Further Research

- When is bipartite approximation better than unipartite?
- Investigate impacts of clustering through analysis of artificial model.
- Investigate when Second Degree and Higher order is an improvement.
- Improve growth and cut-off ansatz – do results improve?

References

[1] Coscia, M. The Atlas for the Aspiring Network Scientist (Michele Coscia, 2021)
[2] Evans, T.S., Chen, B. Linking the network centrality measures closeness and degree. Commun Phys 5, 172 (2022). https://doi.org/10.1038/s42005-022-00949-5
[3] Erdős, P. & Réyni, A. On random graphs. i. Publ. Mathematicae 6, 290–297 (1959)
[4] Tiago P. Peixoto, "The Netzschleuder network catalogue and repository", https://networks.skewed.de (2020)
[5] Barabási, Albert-László. Network Science. Cambridge: Cambridge University Press, 2016
[6] Martin A. Fischler and Robert C. Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24, 6 (June 1981), 381–395. https://doi.org/10.1145/358669.358692