Heidelberg University
Institute of Computer Science
Database Systems Research Group

Lecture: Complex Network Analysis

Prof. Dr. Michael Gertz

Assignment 2

Graph Properties and Random Graphs

`https://github.com/nilskre/CNA_assignments`

Team Member: Patrick Guenther, 3660886,
Applied Computer Science
rh269@stud.uni-heidelberg.de

Team Member: Felix Hausberger, 3661293,
Applied Computer Science
eb260@stud.uni-heidelberg.de

Team Member: Nils Krehl, 3664130,
Applied Computer Science
pu268@stud.uni-heidelberg.de

1 Problem 2-1 Erdos-Renyi Network

Consider an Erdos-Renyi network with N = 80 nodes, connected to each other with probability p = 0.05.

1. What is (i) the expected number of links in the graph and (ii) the expected degree of a node?

   (i) The expected number of links in the graph:

   $$\langle L \rangle = p\frac{N(N-1)}{2} = 0.05\frac{80 \cdot (80-1)}{2} = 158$$

   The expected number of links in the graph is 158.

   (ii) The expected degree of a node:

   $$\langle k \rangle = p(N-1) = 0.05 \cdot (80-1) = 3.95$$

   The expected degree of a node in the network is 3.95.

2. In which regime is the network?

   $\langle k \rangle$ is 3.95, thus greater than 1 and not in the subcritical regime.

   $\langle k \rangle < ln(N)$ since $3.95 < 4.38$, thus it is not in the connected regime.

   This means that the network is in the **supercritical regime**.

3. What is the probability to find exactly $L = 200$ links in the graph?

   $$p_L = \binom{\frac{N(N-1)}{2}}{L} p^L (1-p)^{(N(N-1)/2)-L}$$

   $$p_{200} = \binom{\frac{80(80-1)}{2}}{200} 0.05^{200}(1-0.05)^{(80(80-1)/2)-200} \approx 1.26e^{-4}$$

   The probability to find exactly 200 links in the graph is arround $1.26e^{-4}$.

4. What is the probability that a node i in the graph has degree $k_i = 5$ (using the binomial distribution)?

   $$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

Node	Degree
1	3
2	2
3	2
4	1
5	5
6	1
7	1
8	3
9	1
10	1
11	1
12	1

Table 1: Degrees of each node

$$p_5 = \binom{80-1}{5} 0.05^5 (1-0.05)^{80-1-5} \approx 0.158$$

The probability of a node i having a degree of 5 is 0.158.

5. Use maximum likelihood estimation to estimate the model parameters (N, p) for the shown graph.

   From Table 1 follows that the average degree of the network is $\langle k \rangle \approx 1.83$.

$$\langle k \rangle = p(N-1)$$

$$p = \frac{\langle k \rangle}{N-1} = \frac{1.83}{12-1} \approx 0.166$$

The model parameters are $N = 12$ and $p = 0.166$.

Alternative: We also tried an alternative approach which should be 'real' maximum likelihood estimation. This uses the number of edges, as well as the number of vertices and does not rely on vertex degrees.

$p(G)$ is maximized for

$$p = \frac{m \cdot 2}{n(n-1)}$$

where m is the number of edges and n is the number of vertices in the graph.

$$p = \frac{11 \cdot 2}{12(12-1)} \approx 0.166$$

We multiply m with 2, since indegree = outdegree on this undirected graph.

Following this method, p would be 0.166.

2 Problem 2-2 Three-Dimensional Lattice Network

Consider an undirected network $G = (V, E)$ with $N = \ell^3$ nodes corresponding to points on a regular three-dimensional lattice $\{1, ..., \ell\} \times \{1, ..., \ell\} \times \{1, ..., \ell\}$. Two nodes x and y are connected if and only if $d(x, y) = 1$, where $d(x, y)$ denotes the Euclidean distance. A visualization of such a graph is shown in the following image:

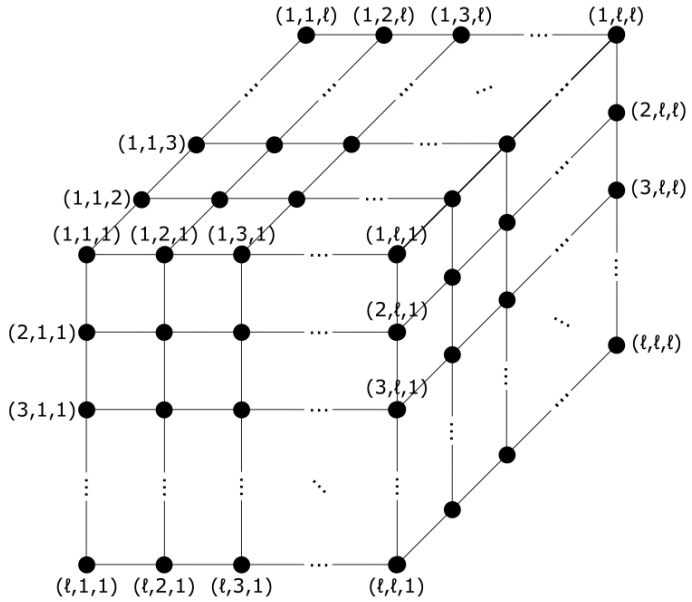


Figure 1: Three-Dimensional Lattice Network

1. What is the diameter d_{max} of the graph?

   The diameter is the longest shortest path in the graph. One example for a longest shortest path is the distance from the left upper edge in the front $(1, 1, 1)$ to the right lower edge in the back (ℓ, ℓ, ℓ). The diameter is therefore $d_{max} = 3(\ell - 1)$ (-1, because the corner nodes are not counted twice).

2. Provide an expression (in terms of N and/or ℓ) for the probability p_i that a randomly chosen node has degree i. You may assume that $\ell \geq 3$. What are the consequences for $N \to \infty$?

   Due to the fact that there are four different types of nodes (inner nodes, corner nodes, edge nodes and outer nodes) with differing degree three different expressions are provided in the following.

   First the distribution of the different node types is clarified in table 2:

Node type	edges	nodes
inner node	6	$(\ell - 2)^3$
corner node	3	8
edge node	4	$12(\ell - 2)$
outer node	5	$6\ell^2 - 12(\ell - 2) - 8$

Table 2: Generic number of nodes for each degree

   Secondly the expressions for the probability distribution are derived based on the following formula with $N = \ell^3$ and $p = \frac{\#nodes}{\ell^3}$:

$$p_k = \binom{N-1}{k} * p^k * (1-p)^{N-1-k}$$

   For inner nodes with $i = 6$:

$$p_6 = \binom{\ell^3 - 1}{6} * (\frac{(\ell-2)^2}{\ell^3})^6 * (1 - \frac{(\ell-2)^2}{\ell^3})^{\ell^3-1-6}$$

   For corner nodes with $i = 3$:

$$p_3 = \binom{\ell^3 - 1}{3} * (\frac{8}{\ell^3})^3 * (1 - \frac{8}{\ell^3})^{\ell^3-1-3}$$

   For edge nodes with $i = 4$:

$$p_4 = \binom{\ell^3 - 1}{4} * (\frac{12(\ell-2)}{\ell^3})^4 * (1 - \frac{12(\ell-2)}{\ell^3})^{\ell^3-1-4}$$

For outer nodes with $i = 5$:

$$p_5 = \binom{\ell^3 - 1}{5} * (\frac{6\ell^2 - 12(\ell - 2) - 8}{\ell^3})^5 * (1 - \frac{6\ell^2 - 12(\ell - 2) - 8}{\ell^3})^{\ell^3 - 1 - 5}$$

For $N \to \infty$ also $\ell \to \infty$, that is why the probability for an inner node increases and the probability for a corner, edge or outer node decreases. This means the average degree converges to 6 (probability p_k for $k = 6$ converges to 1).

3. What is (i) the clustering coefficient of a node i and (ii) the average clustering coefficient of this network?

   (i) the clustering coefficient of a node i

   The clustering coefficient is defined as:

   $$C_i = \frac{2 * L_i}{k_i * (k_i - 1)}$$

   with L_i being the number of edges between the neighbors of the node and k_i being the degree from the chosen node.

   The clustering coefficient is 0 for all types of nodes, because there are no edges (diagonal edges between the nodes) between the neighbors of the node ($L_i = 0$).

   (ii) the average clustering coefficient of this network

   The average clustering coefficient is defined as:

   $$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i$$

   As described in (i) the clustering coefficient is zero for all nodes. Consequently the average clustering coefficient is also zero.

4. Now assume we have a different three-dimensional lattice graph with nodes corresponding to points $\{1, ..., \ell\} \times \{1, ..., \ell\} \times \{1, ..., \ell\}$, where two nodes x and y are connected if and only if $d(x, y) \leq \sqrt{3}$. How does this change the average clustering coefficient in the limit $N \to \infty$?

   Now diagonal edges between the nodes are present. That is why the clustering coefficient is not zero any more.

For $N \to \infty$: it is sufficient to only look at the inner nodes. Each inner node is connected to all nodes of its surrounding $3 * 3$ cube. We denote l as the number of nodes in the intersection of two $3 * 3$ cube neighborhoods and n as the number of nodes of the same type in a neighborhood.

$$L_i = \frac{l_{inner} * n_{inner} + l_{edge} * n_{edge} + l_{corner} * n_{corner}}{2} = \frac{16 * 6 + 10 * 12 + 6 * 8}{2} = 132$$

$$k_i = 3 * 3 * 3 - 1 = 26$$

$$C_i = \frac{2 * L_i}{k_i * (k_i - 1)} = \frac{2 * 132}{26 * (26 - 1)} = 0,41$$

$$\langle C \rangle \to 0,41$$

3 Problem 2-3 Degree distributions

Select two networks, a small network with $N \leq 250$ nodes from the Koblenz Network Collection KONECT and a larger network with $N \geq 2500$ nodes from the Stanford Large Network Dataset Collection SNAP. For each of the networks, provide the following plots.

1. Show the cumulative degree distribution of each network in a separate plot. Specifically, plot the degree $k_x \in \{1, ..., N\}$ on the x-axis.

   On the y-axis, plot $P(k \geq k_x)$ that denotes the probability of a randomly chosen node in the network having a degree of k_x or higher.

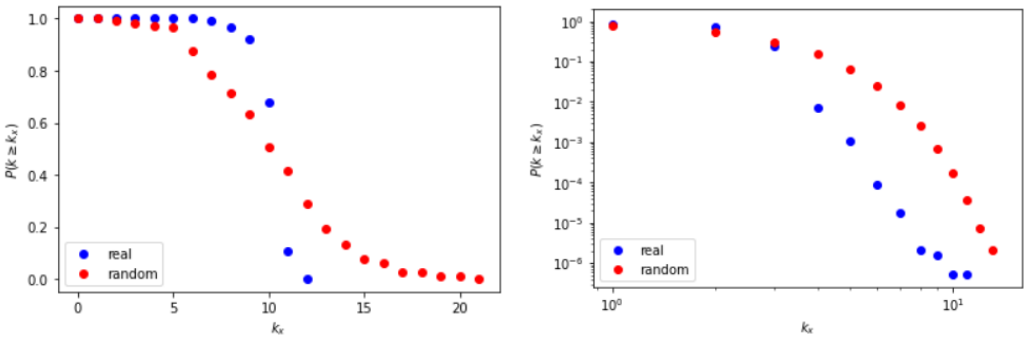
   Select appropriate scales for your plots!

2. To each plot, add the cumulative degree distribution of a random graph with a corresponding number of nodes and edges (approximately!).

   You are free to chose between the $G(N, L)$ and the $G(N, p)$ model for random graphs.

   Average the degree counts over at least 10 samples from the random graph model.

Include your plots and a brief discussion of your findings in the PDF with your solution to this assignment. Also hand in your source code along with specific instructions on how to compile and run your program in a separate archive file.



(a) KONECT American Football network (b) SNAP roadNet-CA (log log scale)

Figure 2: Complementary cumulative distribution functions of the degree distributions

When averaging the degree distribution for both randomly created networks over 10 instances and comparing their complementary cumulative distribution functions (CCDF) to those of the real networks, one can identify interesting results. In both cases the probability for high degree nodes is overestimated by the random networks. This can be explained by the hard organisatorial and physical thresholds that the real networks impose. In case of the american football network, each team normally has the same amount of matches during a season except the case that additional relegation matches are scheduled. Also road networks normally have a physical threshold as not too many roads can intersect each other. The random networks and their underlying binomial distribution aim for softer boundaries in the degree distribution and therefore lead to overestimation.

In case of the american football network, the random model even overestimates the amount of teams that have very few games in one season. As this is also rarely the case in reality and teams tend to have the same amount of matches, the CCDF of the real network remains on a plateau at the beginning while the CCDF of the random network already decreases in value. On the other hand, both real and random CCDF on the road networks seem to match at the beginning.

This leads to the conclusion that both random networks cannot be identified as the generative model behind the real networks.