Complex Networks

Teachers: M. Emmerich, D. Garlaschelli, F. den Hollander.
Written examination: Wednesday, 10 January 2018, 10:00–13:00.

Open book exam: the lecture notes may be consulted, but no other material.

Answer each question on a separate sheet. Put your name, student number and the number of the question you are answering on each and every sheet. Provide full explanations with each of the answers!

Each question is weighted by a number of points, as indicated. The total number of points is 100. The final grade will be calculated as a weighted average: 30% for homework assignments and 70% for this exam.

Success!

1. [**7 points**]
   Describe the static and the dynamic features of Internet. Explain what can be measured to support the presence of these features.

2. Recall the definition of a graph, a random graph and a sequence of random graphs.

   2a. [**3 points**] What is the difference between a simple graph and a multi-graph?

   2b. [**3 points**] Is the random graph generated according to the configuration model always simple?

   2c. [**3 points**] If the degrees of the configuration model are drawn independently from a given probability distribution f on $\mathbb{N}_0$, then is the resulting sequence of random graphs labelled by the number of vertices n always sparse as $n \to \infty$?

3. Consider the Erdős-Rényi random graph $\mathrm{ER}_n(\lambda/n)$ on n vertices with retention probability λ/n for $\lambda \in (0, \infty)$.

   3a. [**4 points**] Explain what is meant by the phrase "$(\mathrm{ER}_n(\lambda/n))_{n\in\mathbb{N}}$ has a percolation transition at $\lambda = 1$".

   3b. [**5 points**] Consider the exploration process $(N_d)_{d\in\mathbb{N}}$ from a given vertex $*$. Here, N_1 counts the number of neighbours of $*$, N_2 counts the number of neighbours of these N_1 neighbours, etc., where overcounting is allowed. Argue why this exploration process is a branching process, give the associated offspring distribution, and explain how this offspring distribution arises.

3c. [**4 points**] Compute the probability that $N_1 = 5$ when $\lambda = 2$ for arbitrary $n \in \mathbb{N}$.

4. Consider the Park-Newman model (for undirected random graphs) defined by the connection probability $p_{ij} = \frac{x_i x_j}{1+x_i x_j}$. Assume $x_i = x_0$ for all $i = 1, \dots, N$, where N is the total number of vertices. Calculate, as a function of x_0 and N only, the following quantities:

   4a. [**2 points**] the expected degree $\langle k_i \rangle$ of each node;

   4b. [**2 points**] the expected average nearest-neighbour degree $\langle k_i^{nn} \rangle$ of each node;

   4c. [**2 points**] the expected local clustering coefficient $\langle c_i \rangle$ of each node;

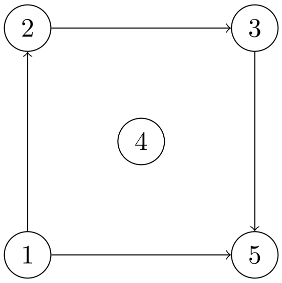
   4d. [**2 points**] the expected total number of undirected links $\langle L_u \rangle$;

   4e. [**2 points**] the expected link density;

   4f. [**4 points**] the entropy of the probability distribution $P(\mathbf{G})$ generating the possible graphs in the model.

5. Complex networks can be represented as adjacency matrix, adjacency list, and edge list.

   5a. [**4 points**] Describe the graph in the picture as an adjacency matrix and as an graph formula ($\hat{=}$ adjacency list) using the igraph notation.



   5b. [**5 points**] What is the time complexity of computing the degree of a node in these three data structures? Consider directed networks and unsorted edge lists and express the complexity in the 'Big O' notation in terms of $|V|$: number of nodes, and $|E|$: number of edges.

6. A bipartite graph is a graph where the node set $V = V_1 \cup V_2$ and $V_1 \cap V_2 = \emptyset$, and for all $(u, v) \in E$, $u \in V_1$ and $v \in V_2$ or $u \in V_2$ and $v \in V_1$.

   6a. [**4 points**] How many different undirected bipartite networks can be formed (in terms of $|V_1|$ and $|V_2|$?

   6b. [**5 points**] Describe an algorithm for producing a random undirected bipartite graph. *Hint:* Consider the idea of a Fisher Yates Shuffle. What is the time complexity of your algorithm?

7. 7a. [**4 points**] Consider ordinary percolation on $\mathbb{Z}^d$. Let $p_c(\mathbb{Z}^d)$ denote the critical threshold. Prove that $d \mapsto p_c(\mathbb{Z}^d)$ is non-increasing.

7b. [**4 points**] Prove that $p_c(\mathbb{Z}) = 1$. What is known about the value of $p_c(\mathbb{Z}^d)$ for $d \geq 2$?

7c. [**6 points**] The contact process has a phase transition that resembles the percolation transition. This is seen by comparing Figures 8.2 and 11.1 in the lecture notes. Can you think of a reason for this similarity? *Hint:* For the contact process on $\mathbb{Z}^d$, think of time as an extra dimension in order to compare it with the percolation process on $\mathbb{Z}^{d+1}$.

8. Consider two possible graphs $\mathbf{G}_1$ and $\mathbf{G}_2$ generated by the Erdős-Rényi random graph model on N vertices and with connection probability p. Let $P(\mathbf{G}_1)$ and $P(\mathbf{G}_2)$ denote the probabilities of generating the graphs $\mathbf{G}_1$ and $\mathbf{G}_2$, respectively. Let L_1 and L_2 denote the number of edges realized in graphs $\mathbf{G}_1$ and $\mathbf{G}_2$, respectively, and similarly let $\vec{k}_1$ and $\vec{k}_2$ denote the degree sequences realized in the two graphs.

8a. [**3 points**] If $\vec{k}_1 = \vec{k}_2$, then do we get $P(\mathbf{G}_1) = P(\mathbf{G}_2)$ in general? Why?

8b. [**3 points**] If $\vec{k}_1 \neq \vec{k}_2$ but $L_1 = L_2$, then do we get $P(\mathbf{G}_1) = P(\mathbf{G}_2)$ in general? Why?

8c. [**3 points**] If $L_1 \neq L_2$, then do we get $P(\mathbf{G}_1) = P(\mathbf{G}_2)$ in general? Why?

8d. [**3 points**] Write the probability $P(\mathbf{G}_1)$ of generating the graph $\mathbf{G}_1$ as a function of p, N and L_1.

8e. [**3 points**] Write the probability $P(\mathbf{G}_2)$ of generating the graph $\mathbf{G}_2$ as a function of p, N and $\vec{k}_2$.

9. A clique is a network with $E = \{(u, v) | u \in V, v \in V\}$. Consider the task of simulating an SI epidemic process, where in the initial state only one node is infected and λ is the infection rate for two neighboring nodes.

9a. [**3 points**]
Describe the state space for a clique of size three.

9b [**3 points**]
What is the time it takes on average for the virus to infect the second node? And the third node?

9c [**4 points**]
Given a clique with N nodes. How can one simulate the total time it takes for the virus to invade the full network?

SOLUTIONS

1. Internet is an example of a technological network. It is a physical network of computers, connected by cables transferring data. It is an undirected network: information can travel both ways along the cables.

   Internet evolves over time, which is why it is usually studied at a coarse-grained level that treats as vertices whole groups of computers, within which rearrangements may occur frequently due to local handling, but between which there are large-scale stable connections. These groups of computers are called Autonomous Systems.

   Features of Internet that can be measured are, for instance, the hop-count (= the number of routers traversed by an e-mail message between two uniformly chosen routers) and the AS-count (= the number of Autonomous Systems that are traversed by an e-mail data set). It turns out that both are small (typically not more than 7), and so Internet is small world. The degree distribution of Internet has a polynomial tail, with an exponent τ that lies between 2.15 and 2.20, and so Internet is scale free.

2a. A simple graph has no self-edges (= edges between a vertex and itself) and no multiple edges (two or more edges between the same pair of vertices). A graph that is not simple is called a multi-graph.

2b. No. The configuration model is generated by assigning a prescribed number of half-edges to each vertex and randomly pairing the half-edges to form edges. In the pairing it may happen that two half-edges attached to the same vertex get paired, or that two half-edges attached to different vertices get paired and afterwards this happens again. For large n, both events are unlikely, and under certain mild conditions on the degree sequence there is a strictly positive probability, bounded from below uniformly in n, that the configuration model generates a simple graph.

2c. If the degree sequence $D = (D_i)_{i=1}^n$ is drawn i.i.d. from f, then the empirical degree distribution $f_{\mathrm{CM}_n(D)} = \frac{1}{n}\sum_{i=1}^n \delta_{D_i}$ converges to f pointwise as $n \to \infty$, by the law of large numbers. Since both $f_{\mathrm{CM}_n(D)}$ and f are probability distributions, the convergence also holds in the supremum-norm. Hence the configuration model with i.i.d. degrees is sparse.

3a. As $n \to \infty$, for $\lambda < 1$ the largest cluster is of size $\Theta(\log n)$, while for $\lambda > 1$ the largest cluster is of size $\Theta(n)$. Thus, as λ crosses the critical threshold $\lambda = 1$, the growth of the largest cluster undergoes a transition from being slow (much less than n) to being fast (of order n). At $\lambda = 1$, the size of the largest cluster is $\Theta(n^{2/3})$.

3b. The number of neighbours N_1 of a given vertex $*$ has a distribution that is BINOMIAL$(n-1, \lambda/n)$: each of the $n-1$ neighbours of $*$ in the complete graph with n vertices is connected to $*$ with probability λ/n. The same is true for each of the N_1 neighbours of $*$, because the complete graph looks the same from every vertex. Hence N_2 is equal to the size of a branching process with offspring distribution BINOMIAL$(n-1, \lambda/n)$ after 2 generations. Here we allow for over-counting, but for n large this causes only a negligible error.

3c. Compute

$$\begin{aligned}\mathbb{P}(N_1 = 5) &= \text{BINOMIAL}(n-1, 2/n)(5) \\ &= \binom{n-1}{5}\left(\frac{2}{n}\right)^5\left(1-\frac{2}{n}\right)^{n-1-5}.\end{aligned}$$

4a. Note that in this particular case the Park-Newman model coincides with the Erdős-Rényi model with probability $p = \frac{x_0^2}{1+x_0^2}$. The expected degree of each node is

$$\langle k_i \rangle = \sum_{j\neq i} p_{ij} = (N-1)\frac{x_0^2}{1+x_0^2} \quad \forall\, i.$$

4b. The expected average nearest-neighbour degree of each node is

$$\langle k_i^{nn} \rangle = \frac{\sum_{j\neq i}\sum_{k\neq j} p_{ij}p_{jk}}{\sum_{j\neq i} p_{ij}} = \frac{(N-1)^2\left(\frac{x_0^2}{1+x_0^2}\right)^2}{(N-1)\frac{x_0^2}{1+x_0^2}} = (N-1)\frac{x_0^2}{1+x_0^2} \quad \forall\, i.$$

4c. The expected local clustering coefficient of each node is

$$\langle c_i \rangle = \frac{\sum_{j\neq i}\sum_{k\neq i,j} p_{ij}p_{jk}p_{ki}}{\sum_{k\neq i,j} p_{ij}p_{ki}} = \frac{(N-1)(N-2)\left(\frac{x_0^2}{1+x_0^2}\right)^3}{(N-1)(N-2)\left(\frac{x_0^2}{1+x_0^2}\right)^2} = \frac{x_0^2}{1+x_0^2} \quad \forall\, i.$$

4d. The expected total number of edges is

$$\langle L_u \rangle = \frac{x_0^2}{1+x_0^2}\frac{N(N-1)}{2}.$$

4e. The expected link density is

$$\langle c_u \rangle = \frac{2\langle L_u \rangle}{N(N-1)} = \frac{x_0^2}{1+x_0^2}.$$

4f. The entropy of the probability distribution $P(\mathbf{G})$ is

$$\begin{aligned}S &= -\sum_{\mathbf{G}} P(\mathbf{G}) \ln P(\mathbf{G}) \\ &= -\frac{N(N-1)}{2}\left[\frac{x_0^2}{1+x_0^2}\ln\frac{x_0^2}{1+x_0^2} + \frac{1}{1+x_0^2}\ln\frac{1}{1+x_0^2}\right]\end{aligned}$$

because it can be rewritten as a sum of the (identical) $N(N-1)/2$ entropies of each edge appearing.

5a.

```
g<-graph.adjacency(matrix(c(0,1,0,0,1,
                            0,0,1,0,0,
                            0,0,0,0,1,
                            0,0,0,0,0,
                            0,0,0,0,0),nrow=5, ncol =5));

g<-graph.formula(1-+2:5,2-+3.3-+5, 4)
```

5b. The time complexity of computing the degree of a node is $\Theta(|V|)$ for the adjacency matrix, $\Theta(2|V|) = \Theta(2|V|)$ for the adjacency list, and $\Theta(|E|)$ for the edge list.

6a. There are $|V_1| \cdot |V_2|$ possibilities to form undirected bipartite graphs.

6b.

1. Make an array of all $|V_1| \cdot |V_2|$ possible edges.
2. Shuffle the list by applying the Fisher Yates method.
3. Output the first k edges.

The algorithm can be made more efficient by applying only the first k iterations of the Fisher Yates shuffle. The time complexity is $O(|V_1| \cdot |V_2|)$.

7a. Note that $\mathbb{Z}^{d-1} \subset \mathbb{Z}^d$. If $p > p_c(\mathbb{Z}^{d-1})$, then with strictly positive probability 0 is connected to infinity in $\mathbb{Z}^{d-1}$. But then it is also connected to infinity in $\mathbb{Z}^d$, and so necessarily $p \geq p_c(\mathbb{Z}^d)$. Consequently, $p_c(\mathbb{Z}^{d-1}) \geq p_c(\mathbb{Z}^d)$.

7b. In $\mathbb{Z}$ the origin is connected to infinity if and only if all edges to the right of 0 are open or all edges to the left of 0 are open, or both. For every $p \in [0, 1)$ these events have probability $p^\infty = 0$, and so $p \leq p_c(\mathbb{Z})$. Consequently, $p_c(\mathbb{Z}) = 1$. It is know that $p_c(\mathbb{Z}^2) = \frac{1}{2}$. No closed form expression is known for $p_c(\mathbb{Z}^d)$ with $d \geq 3$. Numerical estimates are known, as well as expansions in powers of $\frac{1}{2d}$, with $2d$ the degree of the vertices in $\mathbb{Z}^d$.

7c. In the contact process on $\mathbb{Z}^d$, an infected vertex transmits its infection at rate λ along the edges of $\mathbb{Z}^d$. The probability that an infection is transmitted along a given edge after a time interval of length 1 equals $e^{-1}(1 - e^{-\lambda})$. Here, e^{-1} is the probability that the infection does not become healthy before time 1, and $e^{-\lambda}$ is the probability that it is not transmitted until time 1 while staying an infection. Roughly speaking, there is an epidemic if and only if infections survive along the $(d + 1)$-dimensional lattice $\mathbb{Z}^d \times \mathbb{N}_0$. This happens if and only if there is percolation in $\mathbb{Z}^d \times \mathbb{N}_0$ at $p = e^{-1}(1-e^{-\lambda})$. Thus, the critical threshold for percolation, respectively, epidemic are linked to each other, at least in the rough sense indicated above. Above these thresholds, the probability that 0 lies in an infinite cluster, respectively, in the space-time flow of the epidemic, increase with p, respectively, λ.

8a. Yes, if $\vec{k}_1 = \vec{k}_2$, then we get $P(\mathbf{G}_1) = P(\mathbf{G}_2)$ because $\vec{k}_1 = \vec{k}_2$ implies $L_1 = L_2$, and any two graphs with the same number of links (and the same number of nodes) are generated with equal probability in the Erdős-Rényi model.

8b. Yes, if $L_1 = L_2$, then we get $P(\mathbf{G}_1) = P(\mathbf{G}_2)$, again because any two graphs with the same number of links (and the same number of nodes) are generated with equal probability in the model.

8c. No, if $L_1 \neq L_2$, then in general we get $P(\mathbf{G}_1) \neq P(\mathbf{G}_2)$, because two graphs with different numbers of links are generated with different (in general) probability in the model.

8d. The probability of generating the graph $\mathbf{G}_1$ is

$$P(\mathbf{G}_1) = p^{L_1}(1-p)^{N(N-1)/2-L_1}.$$

8e. Since $L_2 = \sum_{i=1}^{N} k_i/2$, the probability of generating the graph $\mathbf{G}_2$ is

$$P(\mathbf{G}_2) = p^{L_2}(1-p)^{N(N-1)/2-L_2} = p^{\sum_{i=1}^{N} k_i/2}(1-p)^{N(N-1)/2-\sum_{i=1}^{N} k_i/2}.$$

9a. The state space has size 2^3, all possible settings of the three nodes to S and I.

9b. The first node infection takes time $1/(2\lambda)$, the second node infection also $1/(2\lambda)$.

9c. The total time can be simulated by the following program:

```
Time ← 0
For $i = 1$ to $N$
        Time ← Time + ExpDistribution($1/\left(i(N-i)\right)$)
Return Time
```

Remarks:
- This makes use of the fact that the probability that in a single differential time step more than one node gets infected is zero.
- Each susceptible node at time step i can get infected from i nodes (at rate $i\lambda$), and there are $n-i$ nodes that can get infected in the next time step.