Complex Networks

Teachers: M. Emmerich, D. Garlaschelli, F. den Hollander.
Written examination: 11 January 2019, 14:00-17:00.

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 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 the homework assignments and 70% for the exam.

Success!

---

1. a. **[3 points]** List the four main classes of complex networks.
   b. **[4 points]** Give one example in each class and provide a brief description of each example.
2. Consider the graph $G(V, E)$ with vertex set $V = \{1, 2, 3, 4\}$ and edge set $E = \{12, 23, 34, 41, 13, 24\}$.
   a. **[3 points]** Compute the typical distance.
   b. **[4 points]** Compute the number of wedges and the number of triangles.
   c. **[2 points]** Compute the global clustering coefficient.
3. Consider the Erdős-Rényi random graph with n vertices and with retention probability λ/n, where $\lambda \in (0, \infty)$. Start an exploration of the graph from a given vertex $*$.
   a. **[4 points]** What branching process stochastically dominates the number of vertices explored at the k-th generation of the exploration process for $0 \leq k \leq n$?
   b. **[5 points]** For which values of k is this branching process a good approximation of the actual number of vertices? Why?
   c. **[4 points]** Explain what the 'depletion effect' in the exploration process is.
4. Consider a real-world binary undirected graph $\mathbf{G}^*$ with n vertices and degree sequence $\{k_i^*\}_{i=1}^n \equiv \{k_i(\mathbf{G}^*)\}_{i=1}^n$, where $k_i^* = k_0$ for all i.
   a. **[2 points]** Describe the procedure how to construct the maximum-

entropy ensemble of binary undirected graphs with n vertices and expected degree sequence $\{\langle k_i \rangle\}_{i=1}^n = \{k_i^*\}_{i=1}^n$.

b. **[2 points]** Write the resulting probability $P(\mathbf{G})$ of any graph $\mathbf{G}$ in such ensemble, as a function of only the degree sequence $\{k_i(\mathbf{G})\}_{i=1}^n$ of $\mathbf{G}$ (note that $\mathbf{G}$ is *any* graph, not necessarily $\mathbf{G}^*$).

c. **[3 points]** Write the equation(s) required to fix the value of the parameters of $P(\mathbf{G})$ and solve these equations explicitly, writing the parameters as a function of k_0.

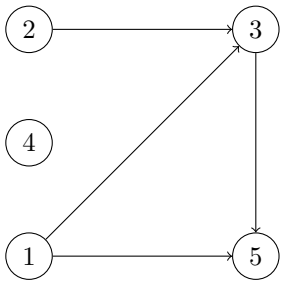
d. **[3 points]** Discuss how this ensemble relates to the Configuration Model and to the Erdős-Rényi random graph model.

e. **[3 points]** Now consider two graphs $\mathbf{G}_A$ and $\mathbf{G}_B$ belonging to this ensemble. Calculate the ratio $P(\mathbf{G}_A)/P(\mathbf{G}_B)$ of the probabilities of generating the two graphs, only as a function of the numbers $L(\mathbf{G}_A)$ and $L(\mathbf{G}_B)$ of undirected links of $\mathbf{G}_A$ and $\mathbf{G}_B$ respectively.

f. **[3 points]** Discuss how one may choose $\mathbf{G}_A$ and $\mathbf{G}_B$ in order to maximize $P(\mathbf{G}_A)/P(\mathbf{G}_B)$, based on the value of k_0.

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

a. [**4 points**] Describe the graph in the picture as an adjacency matrix and as an igraph graph formula ($\hat{=}$ adjacency list). Use the syntax of igraph (R or Python).



b. [**5 points**] A sparse graph is a graph where the largest degree is bounded by a constant, say k. What is the time complexity of computing the degree of all nodes of a sparse graph when the graph is stored (1) as an edge list, (2) as an adjacency matrix? Provide upper and lower time complexity bounds in the Big O, respectively, Big Ω notation.

6. Hypergraphs are networks in which one edge can consist of more than two nodes. Such edges are called *hyperedges*. Let us consider now triplet-graphs, where hyperedges consists of triplets, that consist of exactly three different nodes. Assume the nodes in the edges are ordered but there is no repetition of a node in one edge.

a. **[3 points]** What is the maximum number of triplet-graphs with m

hyperedges?

b. **[6 points]** Discuss an efficient way to (uniformly) randomly generate a random triplet-graph with m random hyperedges for a given set of n nodes (denoted by the indexes 1, ..., n). What is the time complexity and space complexity of your algorithm in terms of m and n.

7. a. **[4 points]** Describe the algorithm that generates the invasion percolation cluster on $\mathbb{Z}^d$, $d \geq 2$.

   b. **[4 points]** Give three characteristic properties of the invasion percolation cluster.

   c. **[6 points]** For each of the three properties, provide a heuristic explanation.

8. Given a generic binary undirected graph, let k_i denote the degree of vertex i and k_i^{nn} the arithmetic average of the degrees of the vertices connected to vertex i. Imagine that you produce a scatter plot where each vertex i is represented as a point with coordinates (k_i, k_i^{nn}) in the plane.

   a. **[4 points]** Describe what the scatter plot looks like in typical realizations of the Erdős-Rényi random graph model with n nodes and connection probability p. Explain your answer.

   b. **[4 points]** Describe what the scatter plot looks like in typical realizations of the canonical configuration model, as a function of the input degree sequence. Explain your answer.

   c. **[5 points]** Describe what the scatter plot looks like in different real-world networks, and what can be concluded from it about the assortativity of these networks.

9. Consider a star-graph with n nodes $V = \{v_1, \ldots, v_n\}$, a central node v_1 and $n-1$ edges $E = \{(v_1, v_2), (v_1, v_3), \ldots, (v_1, v_n)\}$. Consider the SI model of epidemiology. At time $t = 0$ the central node v_1 gets infected, while all the other nodes are in the susceptible state.

   a. **[4 points]** Describe the **generator matrix** of the process. Note that, due to symmetry, the number of infected nodes can be used to represent the state of the graph.

   b. **[3 points]** What is the average time of a Continuous Time Markov Chain process on the graph until it infects every node in the star network. The initial state at time t_0 is the state where the first node just got infected. Assume $n > 1$ and a contagiousness (infection rate) λ.

   c. **[3 points]** What is the closeness centrality of the central node (v_1) in a star graph? What is the closeness centrality of a peripheral node (e.g., v_2) in terms of n?

SOLUTIONS

1a. The four main classes of complex networks are: (1) Social Networks, (2) Technological Networks, (3) Economic Networks, (4) Biological Networks.

1b. Examples in each class are: (1) WWW, Facebook, Twitter, WhatsApp, (2) Internet, power grids, traffic, transportation, (3) trade, interbank, interfirm, input/output, (4) metabolic, neural, protein interaction. Brief descriptions can be found in Sections 1.2-1.5 of the lecture notes.

2a. All pairs of vertices are connected by an edge. Hence the typical distance is $H_G = 1$.

2b. Each of the 4 vertices has 3 wedges and 1 triangle. Hence the total number of wedges is 12 and the total number of triangles is 4.

2c. The global clustering coefficient is $C_G = \frac{3!\times 4}{2!\times 12} = \frac{24}{24} = 1$. Indeed, every wedge is part of a triangle.

3a. The number of neighbours 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 neighbours of $*$, because the complete graph looks the same from every vertex. Hence the average number of neighbours of neighbours of $*$ is equal to the size of a branching process with offspring distribution BINOMIAL$(n-1, \lambda/n)$ after 2 generations. Here we over-count because the exploration process may choose vertices that were chosen before.

3b. In the limit as $n \to \infty$, as long as $k = o(n)$ the over-counting is negligible because the exploration rarely creates a loop.

3c. The 'depletion effect' is the fact that the exploration removes vertices from the set of vertices that are yet to be explored, and is the cause of the discrepancy noted in 3b.

4a. The construction of the maximum-entropy ensemble of graphs with n nodes and expected degree sequence $\{\langle k_i \rangle\}_{i=1}^n = \{k_i^*\}_{i=1}^n$ proceeds through the definition of the Hamiltonian $H(\mathbf{G}, \vec{\theta}) = \sum_{i=1}^n \theta_i k_i(\mathbf{G}) = \sum_{i<j}(\theta_i + \theta_j)g_{ij}$, the calculation of the partition function $Z(\vec{\theta}) = \sum_{\mathbf{G}} e^{-H(\mathbf{G},\vec{\theta})}$, and of the probability $P(\mathbf{G}|\vec{\theta}^*) = e^{-H(\mathbf{G},\vec{\theta}^*)}/Z(\vec{\theta}^*) = \prod_{i<j}(p^*)_{ij}^{g_{ij}}(1-p_{ij}^*)^{1-g_{ij}}$, where $p_{ij}^* = e^{-\theta_i^*-\theta_j^*}/(1+e^{-\theta_i^*-\theta_j^*})$ and each θ_i^* maximizes the likelihood $P(\mathbf{G}^*|\vec{\theta})$, i.e., is such that $k_i^* = \sum_{j\neq i} p_{ij}^*\ \forall i$.

4b. Writing $x_i^* = e^{-\theta_i^*}$, we have $P(\mathbf{G}|\vec{x}^*) = \prod_i (x_i^*)^{k_i(\mathbf{G})} / \prod_{i<j}(1+x_i^* x_j^*)$.

4c. The parameters should realize the maximum-likelihood condition, which in this case ensures that the expected degree sequence equals the empirical

one, i.e., $k_i^* = \sum_{j \neq i} p_{ij}^* \ \forall i$, where $p_{ij}^* = x_i^* x_j^*/(1 + x_i^* x_j^*)$. Since in this case $k_i^* = k_0$ for all i, we have $k_i^* = (n-1)x_0^2/(1+x_0^2)$ and, inverting, $x_0 = \sqrt{k_0/(n-1-k_0)}$.

4d. This ensemble is the canonical version of the configuration model, also called the Park-Newman model, where each degree has the same expected value k_0. Since in this case $x_i^* = x_0$ for all i, the connection probability is a constant $p_{ij}^* = x_0^2/(1+x_0^2)$. Therefore this ensemble also coincides with the Erdős-Rényi random graph model with n nodes and connection probability $p = x_0^2/(1+x_0^2)$.

4e. Using $P(\mathbf{G}|\vec{x}^*) = \prod_i (x_i^*)^{k_i(\mathbf{G})} / \prod_{i<j}(1 + x_i^* x_j^*)$ and $x_i^* = x_0$ (see above), we get $P(\mathbf{G}_A)/P(\mathbf{G}_B) = \prod_i (x_i^*)^{k_i(\mathbf{G}_A)} / \prod_i (x_i^*)^{k_i(\mathbf{G}_B)} = x_0^{L(\mathbf{G}_A)}/x_0^{L(\mathbf{G}_B)} = x_0^{L(\mathbf{G}_A)-L(\mathbf{G}_B)}$.

4f. If $x_0 > 1$ (i.e., $k_0 > (n-1)/2$), then $P(\mathbf{G}_A)/P(\mathbf{G}_B)$ is maximized if $L(\mathbf{G}_A) - L(\mathbf{G}_B)$ is maximized, i.e., if $L(\mathbf{G}_A) = n(n-1)/2$ ($\mathbf{G}_A$ is the complete graph) and $L(\mathbf{G}_B) = 0$ ($\mathbf{G}_B$ is the empty graph). This gives the maximum value $P(\mathbf{G}_A)/P(\mathbf{G}_B) = x_0^{n(n-1)/2}$. If $x_0 < 1$ (i.e., $k_0 < (n-1)/2$), then $P(\mathbf{G}_A)/P(\mathbf{G}_B)$ is maximized if $L(\mathbf{G}_A) - L(\mathbf{G}_B)$ is minimized, i.e., if $L(\mathbf{G}_A) = 0$ ($\mathbf{G}_A$ is the empty graph) and $L(\mathbf{G}_B) = n(n-1)/2$ ($\mathbf{G}_B$ is the complete graph). This gives the maximum value $P(\mathbf{G}_A)/P(\mathbf{G}_B) = x_0^{-n(n-1)/2}$. If $x_0 = 1$ (i.e., $k_0 = (n-1)/2$), then $P(\mathbf{G}_A)/P(\mathbf{G}_B) = 1$ for each pair of graphs, because all graphs are equiprobable in this case.

5a.
```
g<-graph.adjacency(
   matrix(
   c(0,0,1,0,1,
     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-+3:5,2-+3.3-+5, 4,5)
```

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

6a. The maximum number is $n(n-1)(n-2) = n^3 - 3n^2 + 2n$.

6b. The following algorithm produces m random edges.

1. Make a list of all $|V| \times (|V|-1) \times (|V|-2)$ possible hyperedges.
2. Shuffle the list by applying the Fisher-Yates method.
3. Output the first m edges.

The algorithm can be made more efficient by applying only the first k iterations of the Fisher Yates shuffle. The time complexity is $O(|V|^3)$, because in the worst case all hyperedges have to be generated. The Fisher Yates shuffle scales linearly with the length of the list.

7a. Draw edges between all neighbouring vertices of $\mathbb{Z}^d$. Assign i.i.d. weights to these edges, drawn from $(0,1)$ according tothe uniform distribution. At time 0, invade the origin. At time 1, look at the edges touching the origin, pick the one with the smallest weight, and invade the vertex at the other end. At time $n \geq 2$, look at all the edges touching the vertices that have been invaded up to then, pick the one with the smallest weight, and invade the vertex at the other end (possibly this vertex was invaded before). Continue. The invasion percolation cluster IPC is the random set of vertices that are invaded eventually.

7b. Three characteristic properties of IPC are (with probability 1): (1) IPC $\subsetneq \mathbb{Z}^d$; (2) $\lim_{N\to\infty} |\mathrm{IPC} \cap [-N,N]^d|/|[-N,N]^d| = 0$; (3)$\limsup_{n\to\infty} W_k = p_c$, where W_k is the weight of the edge that is invaded at time k and p_c is the critical threshold for ordinary percolation on $\mathbb{Z}^d$.

7c. (1) and (2): The invasion algorithm is so greedy that it shoots out to infinity and skips most vertices. (3): After a finite time, the invasion gets stuck in an infinite cluster that is barely supercritical and contains only weights that are $\leq p_c$.

8a. In typical realizations of the Erdős-Rényi model, the degrees of vertices are distributed around the expected value $\langle k \rangle = p(n-1) \approx pn$ according to a binomial distribution (which, in the sparse case, approaches a Poisson distribution) and the degrees of neighbouring nodes are not correlated among themselves. This means that a typical scatter plot is a 'ball' of points randomly scattered around the point (pn, pn).

8b. In typical realizations of the canonical configuration model with given degree sequence, the k_i coordinates of the scatter plot are fixed by the chosen degree sequence, while each of the k_i^{nn} coordinates will be randomly scattered around its expected value $\langle k_i^{nn} \rangle = \sum_{j\neq i} p_{ij} k_j / k_i$, where $p_{ij} = x_i x_j/(1 + x_i x_j)$ and each x_i is such that $k_i = \sum_{j\neq i} p_{ij}$. If the input degree sequence is a constant vector, then the canonical configuration model reduces to the Erdős-Rényi model discussed in point 8a above. As the degree sequence becomes more and more heterogeneous, the scatter plot acquires a typically decreasing trend. If the degree distribution has a decreasing power-law tail, then the tail of the scatter plot is a decreasing power law as well.

8c. In different real-world networks, the empirical scatter plot is often decreasing (disassortativity) and sometimes increasing (assortativity). However, what matters in order to assess whether there is a genuine tendency towards (dis)assortativity is not the empirical trend itself, but rather its comparison with the one obtained in the configuration model with the same input degree sequence as the empirical one. Assortativity/disassortativity is then signalled *locally* by some points of the empirical scatter plot being significantly above/below the points of the corresponding vertices in the scatter plot obtained in the configuration model. Both behaviours are observed in real-world networks. In some cases (e.g. interbank networks),

an almost complete consistency with the configuration model is observed, which means neither assortativity nor disassortativity.

9a. Let n denote the number of nodes $s, 1 \leq s \geq n-1$ denote the number of infected nodes. Then, the transition rate $q_{s,s+1}$ from a system with s infected nodes to a system with $s+1$ infected nodes is $\lambda(n-s)$. The transition rate q_{ss} is equal to $-n+s$. All other rates are zero.

9b. $E(t) = \sum_{s=1}^{n-1} \frac{1}{(n-s)\lambda}$.

9c. The closeness of the central node is 1. Of all other nodes the closeness centrality is $\frac{(1+2(n-1))}{n} = 2 - \frac{1}{n}$.