Exam 2023/2024: Social and Graph Data Management

Pablo Mollá Chárlez

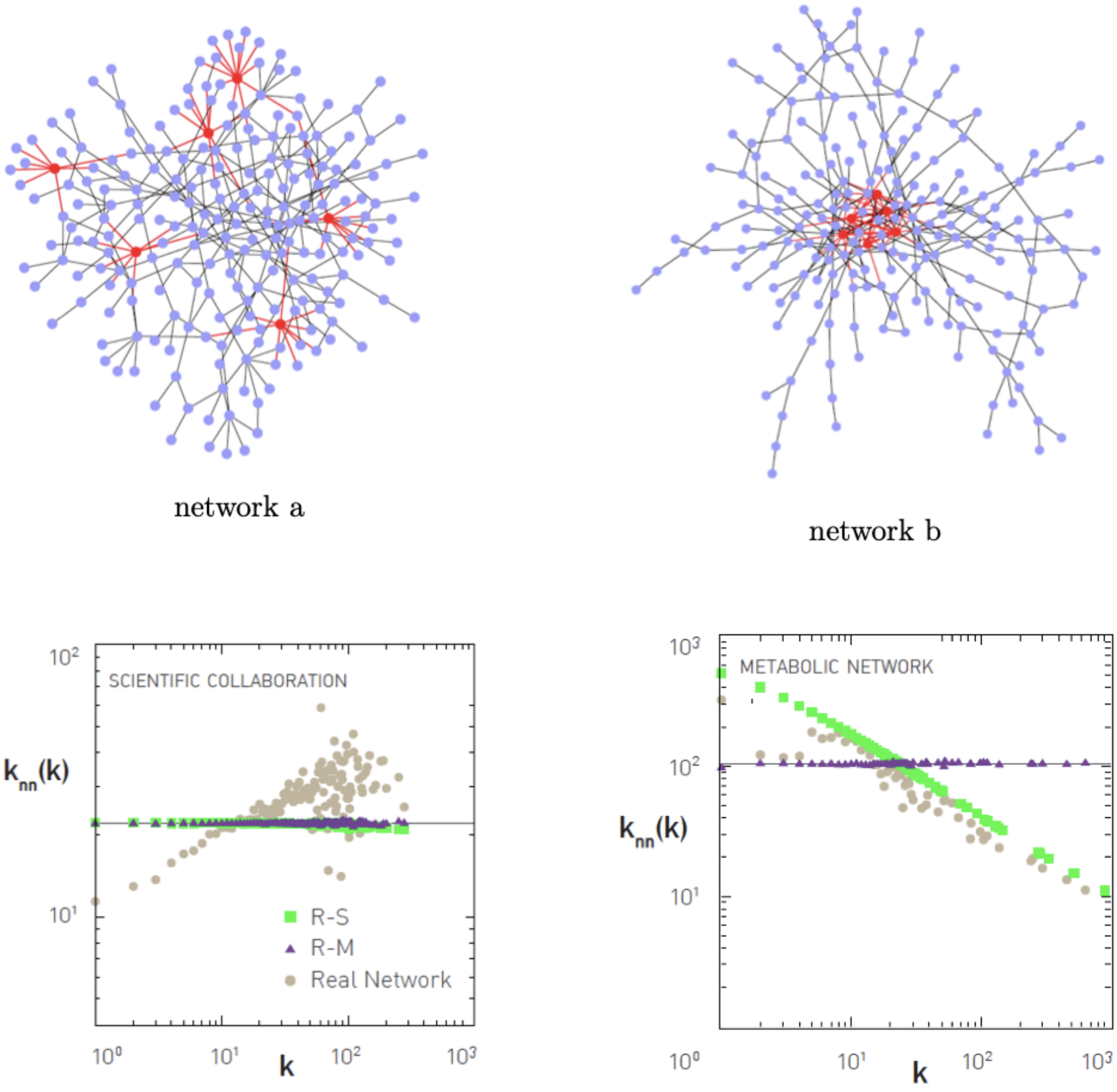
Contents

1 Exercise 1: Communities (10 points)

- **Question 1**. Give an intuition why it is extremely unlikely that there are 2 giant components in a randomly wired network.

- **Question 2**. Explain the hypotheses used for community detection in social networks.

- **Question 3**. For each of the following 4 networks, indicate if they are assortative, disassortative or neutral, and justify ($k_{nn}(k)$ is the degree correlation function, see on the last page a reminder for the definition). For networks c and d (scientific collaboration and metabolic network), you should discuss the assortativity of the "real network" but also what additional information is provided by the green square plots (network perturbed randomly while preserving the degree distribution and without multilinks).



network a

network b

network c

network d

Figure 1: Networks

1.1 Answers

- **Question 1**. Give an intuition why it is extremely unlikely that there are 2 giant components in a randomly wired network.

  In a randomly wired network, having **two giant components is extremely unlikely** because, as the network grows, nodes are added and connected randomly, making it highly probable that small components will merge into one large component rather than forming multiple large ones, which happens in the supercritical regime ($p > \frac{1}{N}$).

  When a network grows and reaches a certain size, even a few random links are enough to connect smaller parts into one big component. Since connections are made randomly, it's very unlikely for two large, separate groups to form without eventually joining together.

- **Question 2**. Explain the hypotheses used for community detection in social networks.

  - **Hypothesis 1: Topological Encoding of Community Structure**

    Hypothesis 1 posits that a graph's community structure is uniquely encoded in its topology. This means that the arrangement and connectivity of nodes inherently contain the information necessary to discern distinct communities without additional metadata.

  - **Hypothesis 2: Locally Dense Subgraph**

    Hypothesis 2 defines a community as a locally dense connected subgraph within a network. In other words, a community is a subset of nodes that exhibit a higher density of connections among themselves compared to their connections with the rest of the network.

  - **Hypothesis 3: Absence of Community Structure in Random Networks**

    Hypothesis 3 asserts that random networks lack a community structure. Unlike structured networks where communities emerge due to specific interaction patterns, random networks distribute connections uniformly, resulting in no discernible clusters.

  - **Hypothesis 4: Optimal Community Structure via Maximum Modularity**

    Hypothesis 4 suggests that the partition with maximum modularity corresponds to the optimal community structure. Modularity measures the strength of division of a network into communities, and the hypothesis claims that the highest modularity partition best represents the inherent community organization.

- **Question 3**. For each of the following 4 networks, indicate if they are assortative, disassortative or neutral, and justify ($k_{nn}(k)$ is the degree correlation function, see on the last page a reminder for the definition). For networks c and d (scientific collaboration and metabolic network), you should discuss the assortativity of the "real network" but also what additional information is provided by the green square plots (network perturbed randomly while preserving the degree distribution and without multilinks).

  Let's classify each network.

  - **Network A:**
    - **Assortativity:** Disassortative
    - **Justification:** High-degree nodes are primarily connected to low-degree nodes, as seen by the red hubs linking to peripheral nodes.
  - **Network B:**
    - **Assortativity:** Assortative
    - **Justification:** High-degree nodes (red) tend to connect with other high-degree nodes, forming clusters.

- **Network C (Scientific Collaboration):**
  - **Assortativity:** Assortative
  - **Justification:** The real network exhibits an upward trend in $k_{nn}(k)$, indicating that high-degree nodes are more likely to connect with other high-degree nodes.
  - **Additional Information from R-M and R-S:** Both R-M and R-S show no assortative or disassortative trends ($k_{nn}(k)$ remains flat). This suggests that the assortative mixing in the real network is not due to random chance but is a structural property of the network.
- **Network D (Metabolic Network):**
  - **Assortativity:** Disassortative
  - **Justification:** The real network shows a decreasing $k_{nn}(k)$, meaning high-degree nodes are more likely connected to low-degree nodes.
  - **Additional Information from R-M and R-S:** Both R-M and R-S maintain a neutral trend ($k_{nn}(k)$ flat), indicating that the disassortative behavior in the real network is a structural feature, not a result of degree distribution alone.

Reminders

- An increasing $k_{nn}(k)$ indicates assortativity.
- A decreasing $k_{nn}(k)$ indicates disassortativity.
- A flat $k_{nn}(k)$ suggests a neutral network.

In the two graphs, the acronyms **R-M** and **R-S** stand for:

- **R-M (Random Multigraph):** A randomly generated multigraph preserving the degree distribution of the real network. Multilinks (multiple edges between two nodes) and self-loops are allowed.
- **R-S (Random Simple Graph):** A random simple graph generated while preserving the degree distribution of the real network. Multilinks and self-loops are disallowed.

  These graphs are used as references to compare the degree correlation properties of the real network against purely random versions with the same degree distribution.

2 Exercise 2: Graph Measures (10 points)

- **Question 1** Represent the graph as an adjacency matrix.
- **Question 2** Write down the degree distribution of G, and the average degree $\langle k \rangle$.
- **Question 3** Compute the diameter d_{max} of G, and show a path of length d_{max} in G.
- **Question 4** Assume that the graph was computed using an Erdös–Rényi random network model with parameter p. What is the value of p? Explain how you found it.
- **Question 5** If we assume that links represented with thick edges are strong links, and the other links are weak, does the graph G satisfy the strong triadic closure property ?
- **Question 6** Give at least 2 measures for which social networks generally behave differently from the Erdös–Rényi model. Explain succinctly how they differ.
- **Question 7** Recall the phase transitions observed for connected components when a random graph is generated using Erdös–Rényi model.

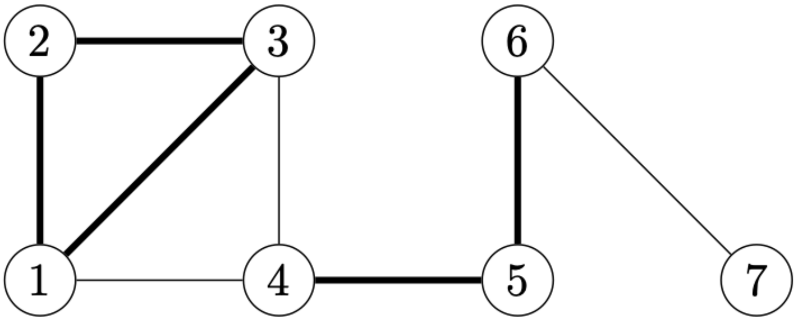


Figure 2: Graph G

2.1 Answers

- **Question 1** Represent the graph as an adjacency matrix.

  The graph has 7 nodes, and the adjacency matrix A is an 7×7 matrix where $A[i][j] = 1$ if there is an edge between nodes i and j, otherwise $A[i][j] = 0$, then :

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

- **Question 2** Write down the degree distribution of G, and the average degree $\langle k \rangle$.

  The degree distribution can be extracted by counting how many $1's$ appear per row in the adjacency matrix A:

$$k_1 = 3; \quad k_2 = 2; \quad k_3 = 3; \quad k_4 = 3; \quad k_5 = 2; \quad k_6 = 2; \quad k_7 = 1$$

Therefore, the degree distribution of G is:

$$p_0 = 0; \quad p_1 = \tfrac{1}{7}; \quad p_2 = \tfrac{3}{7}; \quad p_3 = \tfrac{3}{7}$$

Then, the average degree is:

$$\langle k \rangle = 0 \times p_0 + 1 \times p_1 + 2 \times p_2 + 3 \times p_3 = \frac{16}{7}$$

- **Question 3** Compute the diameter d_{max} of G, and show a path of length d_{max} in G.

  The d_{max} is the longest shortest path between any two nodes in the graph, therefore the $d_{max} = 5$. One path with such length could be for instance the path that connects node 2 and node 7, which is:

  $$2 \underbrace{\longrightarrow}_{1} 3 \underbrace{\longrightarrow}_{1} 4 \underbrace{\longrightarrow}_{1} 5 \underbrace{\longrightarrow}_{1} 6 \underbrace{\longrightarrow}_{1} 7$$

  Another path with maximum distance could be:

  $$2 \underbrace{\longrightarrow}_{1} 1 \underbrace{\longrightarrow}_{1} 4 \underbrace{\longrightarrow}_{1} 5 \underbrace{\longrightarrow}_{1} 6 \underbrace{\longrightarrow}_{1} 7$$

- **Question 4** Assume that the graph was computed using an Erdös–Rényi random network model with parameter p. What is the value of p? Explain how you found it.

  The theory studied during the lectures tells us that the average degree $\langle k \rangle$ in a random network with size N nodes, is described as:

  $$\langle k \rangle = p \cdot (N - 1) \longleftrightarrow p = \frac{\langle k \rangle}{N - 1} = \frac{\frac{16}{7}}{7 - 1} = \frac{16}{42} = \frac{8}{21}$$

- **Question 5** If we assume that links represented with thick edges are strong links, and the other links are weak, does the graph G satisfy the strong triadic closure property ?

  The Strong Triadic Closure Property (STCP) states that if a node has strong links to two other nodes, then there must be at least a (strong or weak) link between those two nodes. Let's analyze the graph step by step to check if it satisfies STCP.

  - **Node 1:** Node 1 has strong links to nodes 2 and 3. There is a strong link between nodes 2 and 3, so STCP is satisfied for Node 1.
  - **Node 2:** Node 2 has strong links to nodes 1 and 3. There is a strong link between nodes 1 and 3, so STCP is satisfied for Node 2.
  - **Node 3:** Node 3 has strong links to nodes 1 and 2 and a weak link to node 4. Nodes 1 and 2 are strongly connected, so STCP is satisfied for Node 3.
  - **Node 4:** Node 4 has a strong link to node 5 and weak links to nodes 1 and 3. Node 1 and node 3 are weakly connected to node 4, so there is no violation of STCP (since weak links don't enforce the requirement for a connection between neighbors).
  - **Node 5:** Node 5 has strong links to nodes 4 and 6. There is **no strong or weak link** between nodes 4 and 6, so STCP is **not satisfied** for Node 5.
  - **Node 6:** Node 6 has a strong link to node 5 and a weak link to node 7. No strong triads are formed here, so STCP is not violated for Node 6.
  - **Node 7:** Node 7 has a weak link to node 6. No strong triads are formed here, so STCP is not violated for Node 7.

In conclusion, the graph does not satisfies the Strong Triadic Closure Property because for node 5 which is strongly connected to nodes 4 and 6, there is no link (strong or weak) between both nodes 4 and 6.

- **Question 6** Give at least 2 measures for which social networks generally behave differently from the Erdös–Rényi model. Explain succinctly how they differ.

  Here are two measures where social networks (~ Power-Law model) behave differently from the Erdös–Rényi model:

  1. **Clustering Coefficient**
     - **Social Networks** ~ Power-Law model: Typically have **high clustering coefficients**, meaning that if two individuals have a common friend, there is a high likelihood that they are also friends (triadic closure is common).
     - **Random Networks** ~ Erdös–Rényi model: Edges are placed randomly, so the **clustering coefficient is much lower** and depends only on the edge probability p. The probability of a triangle forming is approximately p^3, which is often very small.

     Social networks exhibit significantly **more local structure** (tightly-knit groups) and **clustering** than random networks, reflecting real-world tendencies for communities and friendships to form.
  2. **Degree Distribution**
     - **Social Networks** ~ Power-Law model: Typically have a **heavy-tailed degree distribution**, meaning that most nodes have a small degree, but a **few nodes (hubs) have very high degree** (e.g., influencers in social networks).
     - **Random Networks** ~ Erdös–Rényi model: The degree distribution follows a binomial distribution (or Poisson in the large graph limit). **Degrees are more concentrated around the average degree**, and **hubs are unlikely**.

     Social networks are more heterogeneous, with a few highly connected nodes playing a central role, while the random networks are homogeneous, with nodes having roughly the same degree.

- **Question 7** Recall the phase transitions observed for connected components when a random graph is generated using Erdös–Rényi model.

  The behavior of random networks varies significantly depending on the edge probability p relative to $\frac{1}{N}$ (where N is the number of nodes) or the average degree $\langle k \rangle$ relative to 1. These variations define distinct regimes within random networks:

  1. $\boxed{\textbf{Subcritical Regime: } p < \frac{1}{N} \text{ or } \langle k \rangle < 1}$

     The network consists of numerous small, disconnected components. The size of the largest connected component N_G grows logarithmically with N, i.e., $N_G \in O(\ln N)$. The structure contains clusters which are predominantly tree-like, lacking cycles.
  2. $\boxed{\textbf{Critical Point: } p = \frac{1}{N} \textbf{ or } \langle k \rangle = 1}$

     This is the threshold where a phase transition occurs. A large component emerges with $N_G \sim N^{2/3}$, alongside many smaller tree-like clusters. The large cluster may contain loops, while smaller clusters remain mostly trees.
  3. $\boxed{\textbf{Supercritical Regime: } p > \frac{1}{N} \textbf{ or } \langle k \rangle > 1}$

     A single dominant connected component, known as the giant component, forms and $N_G \sim (p - p_c)N$, where $p_c \approx \frac{1}{N}$ is the critical probability. The giant component contains **numerous loops**, whereas **other smaller clusters are typically tree structures**.
  4. $\boxed{\textbf{Connected Regime: } p > \frac{\ln(N)}{N} \textbf{ or } \langle k \rangle \geq \; ln(N)}$

     The network becomes fully connected. N_G approaches N, meaning almost all nodes are part of a single connected component. The network is richly interconnected with **multiple loops**.

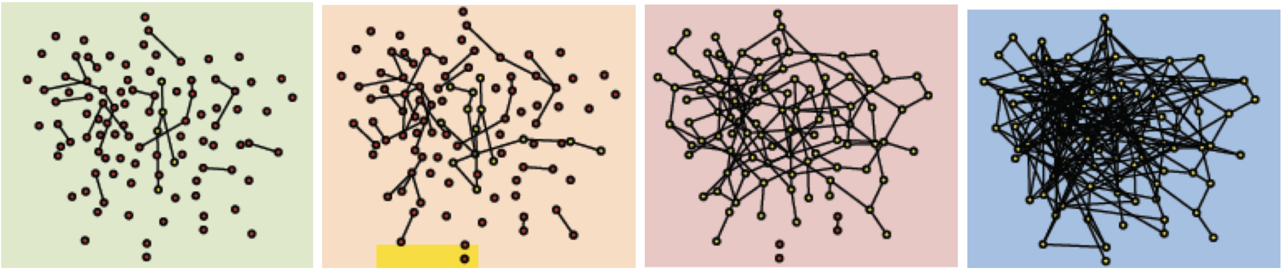


Figure 3: Subcritical, Critical, Supercritial and Connected Regime in Random Networks

3 Exercise 3: Power-Laws (10 points)

- **Question 1** Depending on the value γ, which moments exist for Power-Law distributions with exponent γ? Prove it (you may consider directly the continuous version of the distribution).
- **Question 2** Explain why networks following a Power-Law distributions are called scale-free (provide two explanations for the term).
- **Question 3**. Explain the differences between how robust the social networks are as compared to random network under targeted attacks that remove the optimal nodes to disconnect the network.

Reminder

Degree correlation function: $k_{nn}(k) = \sum_{k'} k' \cdot P(k', k)$ where $P(k', k)$ is the conditional probability that by following a link of a degree-k node, we reach a degree-k' node.

3.1 Answers

- **Question 1** Depending on the value γ, which moments exist for Power-Law distributions with exponent γ? Prove it (you may consider directly the continuous version of the distribution).

  The Power-Law distribution is given as $p(k) \sim k^{-\gamma}$. For the moments, we calculate:

$$\langle k^m \rangle \sim \int_{k_{\min}}^{\infty} k^{m-\gamma}\, dk \quad \text{diverges} \longleftrightarrow m - \gamma > -1$$

  Case Analysis Based on γ:

  1. $\boxed{\gamma \le 2:}$ We have that, $\gamma \le 2 \longleftrightarrow m - \gamma \ge m - 2$ and for $m = 1$, $m - \gamma \ge -1$ thus $\langle k \rangle$ **diverges** and for $m = 2$, we have that $m - \gamma \ge m - 2 = 0 > -1$, then $\langle k^2 \rangle$ **diverges** too.
  2. $\boxed{2 < \gamma < 3:}$ We have that, $2 < \gamma < 3 \longleftrightarrow m - 2 > m - \gamma > m - 3$ and for $m = 1$, $-1 > m - \gamma > -2$, therefore the integral converges and $\langle k \rangle$ **is finite**. For $m = 2$, $0 > m - \gamma > -1$, the integral diverges, thus, $\langle k^2 \rangle$ **diverges**.
  3. $\boxed{\gamma \ge 3:}$ We have that, $\gamma \ge 3 \longleftrightarrow m - \gamma \le m - 3$ and for $m = 1$, $m - \gamma \le -2$ and for $m = 2$, $m - \gamma \le -1$, both integrals converge. Thus, both $\langle k \rangle$ and $\langle k^2 \rangle$ **are finite**.

- **Question 2** Explain why networks following a Power-Law distributions are called scale-free (provide two explanations for the term).

  For a random network with a Poisson degree distribution $\sigma_k = \langle k \rangle^{\frac{1}{2}}$, which is always smaller than $\langle k \rangle$. Hence the network's nodes have degrees in the range $k = \langle k \rangle \pm \langle k \rangle^{\frac{1}{2}}$. In other words nodes in a random network have **comparable degrees** and the average degree $\langle k \rangle$ serves as the "**scale**" of a random network.

  For a network with a power-law degree distribution with $\gamma < 3$ the first moment is finite but the second moment is infinite. The divergence of $\langle k^2 \rangle$ (and of $\langle \sigma_k \rangle$) for large N indicates that the **fluctuations**

around the average can be arbitrary large. This means that when we randomly choose a node, we do not know what to expect: The selected node's degree could be tiny or arbitrarily large. Hence networks with $\gamma < 3$ **do not have a meaningful internal scale**, but are "scale-free"

- **Question 3**. Explain the differences between how robust the social networks are as compared to random network under targeted attacks that remove the optimal nodes to disconnect the network.

  Social networks, which often follow a Power-Law distribution, are more **robust to random attacks** but **highly vulnerable to targeted attacks**. This is because their structure is dominated by hubs (nodes with very high degrees). Random immunization (removing nodes at random) is less effective because it is unlikely to affect these critical hubs. However, selective immunization (targeting hubs) quickly disconnects the network, as hubs are crucial for maintaining its connectivity.

  In contrast, random networks have a degree distribution centered around the **average degree**. They are less dependent on hubs, so both random and targeted attacks are similarly effective, making them more evenly robust under different attack strategies.