Networks Analysis and Mining

Sorbonne Université
Master d'Informatique spécialité Réseaux

Examination

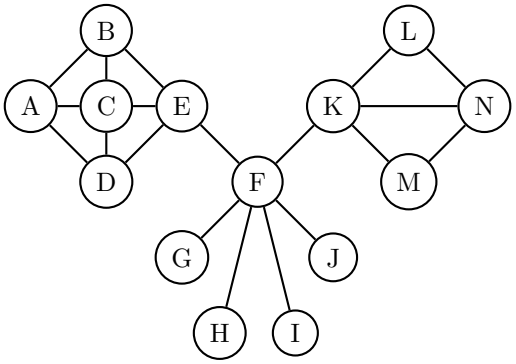
February 4^{th} 2020

Maximilien Danisch and Lionel Tabourier

The exam is 2 hours long. All documents are allowed. Please answer on a separate paper. The grading scale is only indicative. Good luck!

Exercice 1 — *Measures on a small graph (6pts)*

Consider the following graph:



Q1. Draw the degree distribution of the graph.

Q2. Give a possible Breadth First Search ordering (or a BFS tree if you prefer) of the nodes starting from node E.

Q3. What is the diameter of the graph? Justify briefly.

Q4. Compute the clustering coefficient of nodes F, K and A.

Q5. Compute the core value of each node.

Exercice 2 — *Basic description of dynamical contact data (6pts)*

Consider the following undirected interaction data in the format
`node1 node2 time`
Interactions are supposed to be instantaneous. It is not mandatory but recommended to draw the dynamical network before starting.

```
c e 1
c d 2
e f 2
c d 3
e f 3
c e 4
d e 5
c d 6
e f 6
a c 7
d e 8
a c 9
e f 9
b c 10
b c 11
```

Q1. We remind that for all pairs of interacting nodes, an inter-contact time is the time interval between two successive contacts.

What is the distribution of all inter-contact times? (You can plot it if you prefer.)

Q2. For this question, give the links and the moment when they occur.

a) Give a foremost path from node `c` to node `f` starting after time t=0.

b) Give a fastest path from node `c` to node `f` starting after time t=3.

Q3. Give the set of reachable nodes $\mathcal{R}_d(a)$ from node `a` considering the dynamics.

Q4. We remind that the (static) aggregated graph over a period T is the graph obtained when you consider all interactions taking place during T as a link in the graph.

a) Draw the aggregated graph corresponding to the whole period $[0, 12]$.

b) Give the set of reachable nodes $\mathcal{R}_s(a)$ from node `a` in the static aggregated graph, compare it to $\mathcal{R}_d(a)$.

c) In general (not only in this dataset), for a node `x`, what do you think the relation is between $|\mathcal{R}_s(x)|$ and $|\mathcal{R}_d(x)|$? (No justification is necessary.)

Exercice 3 — *Extreme cases (6pts)*

For any given $n \in \mathbb{N}$, let G_n be the graph composed of $2n + 1$ nodes and $3n$ edges such that:

- a unique node n_0 is connected to all other nodes in the network;
- all other nodes have degree 2.

Q1. Draw the case G_4.

The clustering coefficient of a node is defined as the number of triangles the node belongs to divided by its number of pairs of neighbors (it is not defined for a node with less than 2 neighbors). The clustering coefficient of a graph is defined as the average clustering coefficient of all its nodes.

Q2. Compute the clustering coefficient of n_0. What is the clustering coefficient of any other node? Deduce the clustering coefficient of G_4.

The transitivity ratio is defined as three times the number of triangles divided by the number of V-edges (a V-edge is a pair of edges that share an end node).

Q3. Compute the transitivity ratio of G_4.

Q4. How do these coefficients evolve when n goes to ∞? Justify briefly.

Q5. Provide an interpretation of these coefficients in terms of probability (one sentence for each coefficient is sufficient).

Exercice 4 — *Understanding a document (5pts)*

The following pictures are extracted from the article: *Human Wayfinding in Information Networks*, by R. West and J. Leskovec, *WWW 2012*.

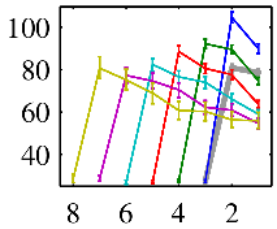
The authors have made a simplified replica of *Wikipedia*. They have organized an online game on this website and report its results. The principle of the game is as follows: given a source page and a target page, players have to find the shortest possible path from the source page to the target page by only following the hyperlinks which are present in the text of the Wikipedia page.

The purpose of their study is to investigate how efficient humans are at finding a way in this network and understand what kind of strategies they are using in that purpose.

Their first observation is that when a chain from the source to the destination is completed, it is usually short: the median is typically 3 clicks from the starting page and it is rarely larger than 10.

They focus on the paths completed by the players in between 3 and 8 clicks. They plot characteristics of the pages depending on their distance to the target of the path. Nodes are grouped depending on the length of the path from the source to the target they are in: the yellow curve corresponds to 8 clicks paths, violet is 7, light blue is 6, red is 5, green is 4, dark blue is 3.

Q1. First observe the following picture. It is the average out-degree of a page (Y-axis) as a function to its distance to the target (X-axis):



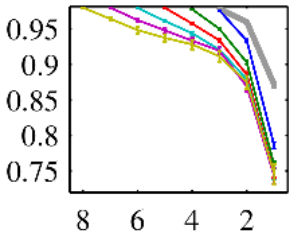
a) Read: how do you read the point of coordinates (4,85) on the red curve?

b) Analyze: for all paths length, where do we find the highest out-degree nodes?

c) Interpret: what kind of pages do you think most players are looking for at their first click?

Q2. We remind you that TF-IDF distance is a measure of how different the contents of 2 documents is.

Now observe the following picture, it is the average TF-IDF distance of the current page compared to the target page (Y-axis) as a function to its distance to the target (X-axis):



a) Let us focus on length 8 paths (leftmost, yellow curve), describe qualitatively the curve.

b) What does the curve tells you about the content of the pages on the path compared to the target?

Q3. What conclusion can you draw about the typical strategy a human is following in order to find target page? Which experiment that we have discussed in the course does it remind you of?

Exercice 5 — *Listing triangles (7pts)*

Consider the following algorithm:

Algorithm 1 Listing triangles

Let η be a total ordering on the nodes of the input graph G
$\vec{G} \leftarrow$ directed version of G, where $v \to u$ if $\eta(v) < \eta(u)$
function TR($\vec{G}$)
  for each node u of $\vec{G}$ **do**
    for each node v in $\Delta_u^{?_1}$ **do** ▷ replace "$?_1$", "$?_2$", "$?_3$" by either "+" or "-"
      $W \leftarrow \Delta_u^{?_2} \cap \Delta_v^{?_3}$ ▷ Δ_u^+: out-neighbors of u in $\vec{G}$
      for each node $w \in W$ **do** ▷ Δ_u^-: in-neighbors of u in $\vec{G}$
        output triangle $\{u, v, w\}$

Q1. In Algorithm 1, each one of the symbols "?" can be replaced by either "+" or "-".
a) Consider the case $?_1 = -$, $?_2 = -$, $?_3 = -$. Is the algorithm correct (meaning the algorithm outputs every triangle of G once and only once)? Justify your answer.
b) Consider the case $?_1 = +$, $?_2 = -$, $?_3 = +$. Is the algorithm correct (meaning the algorithm outputs every triangle of G once and only once)? Justify your answer.

For a node u, we define d_u the degree of node u, $d_u^+ = |\Delta_u^+|$ and $d_u^- = |\Delta_u^-|$.

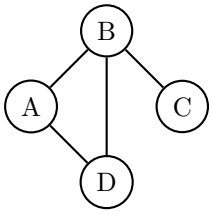
Q2. Assume all "?" are "+", explain the implementation details of the algorithm so that it runs in $O\big(m+n\big)$ memory and $O\big(m + \sum_{uv\in E}(d_u^+ + d_v^+)\big)$ time.

Q3. Prove that if η is the core ordering, then $O\big(m + \sum_{uv\in E}(d_u^+ + d_v^+)\big) \subset O(\delta \cdot m)$, where δ is the core value of the graph.

Q4. Prove that $\sum_{uv\in E}(d_u^+ + d_v^+) = \sum_{u\in V} d_u \cdot d_u^+$. Deduce which node ordering η minimizes $\sum_{uv\in E} d_u^+ + d_v^+$?

Exercice 6 — *The friendship paradox (bonus)*

Consider this simple friendship network with 4 nodes:



Its average degree is 2, so if δ_i is the degree of person x_i,

$$\frac{1}{4}\sum_{i=1}^{4}\delta_i = 2$$

Now, we consider the person x_i and measure the average degree of his or her friends, it is $\overline{x_i}$, and we do the same for all persons. We compute the average of all $\overline{x_i}$, it is $\frac{13}{6}$, in other words:

$$\frac{1}{4}\sum_{i=1}^{4}\overline{x_i} = \frac{13}{6} \simeq 2.17$$

The fact that $\frac{1}{n}\sum_{i=1}^{n}\overline{x_i} \geqslant \frac{1}{n}\sum_{i=1}^{n}\delta_i$ is true for any network. It is described as the friendship paradox: *most people have less friends than their friends have.*

Can you give an explanation of this fact?