Networks Analysis and Mining

Sorbonne Université
Master d'Informatique spécialité Réseaux

Examination corrected

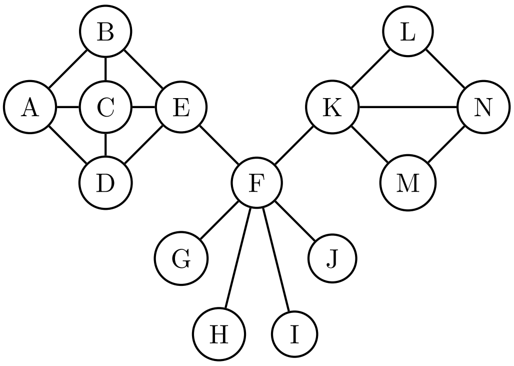
February 4th 2020

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. The degree distribution of the graph is as follows:
1:4 (G,H,I,J), 2:2 (L,M), 3:4 (A,B,D,N), 4:3 (C,K,E), 5:0, 6:1 (F)

Q2. Possible BFS from E ordering:

$$[E, B, C, D, F, A, G, H, I, J, K, L, M, N]$$

Q3. The diameter of the graph is 5, it is the maximum distance between two nodes, it is the case here of nodes N and A (one shortest path is $N - K - F - E - D - A$).

Q4. Clustering coefficients:
$cc(F) = 0$, $cc(A) = \frac{2}{3}$, $cc(K) = \frac{1}{3}$

Q5. Core values:
1: $\{G, H, I, J\}$, 2: $\{F, K, L, M, N\}$, 3: $\{A, B, C, D, E\}$

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.

The list of inter-contact times is:
$c - e : \{3\}$, $c - d : \{1, 3\}$, $e - f : \{1, 3, 3\}$, $d - e : \{3\}$, $a - c : \{2\}$, $b - c : \{1\}$
Hence the distribution:
1:3, 2:1, 3:5

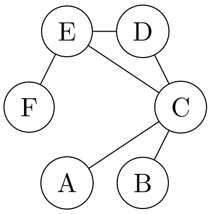
Q2. For this question, give the links and the moment when they occur.

a) A foremost path from node `c` to node `f` starting after time t=0:
$(c, e, 1); (e, f, 2)$

b) A fastest path from node `c` to node `f` starting after time t=3:
$(c, e, 4); (e, f, 6)$

Q3. The set of reachable nodes $\mathcal{R}_d(a)$ from node `a` considering the dynamics:
$\mathcal{R}_d(a) = \{b, c\}$

Q4. a)



b) The set of reachable nodes $\mathcal{R}_s(a)$ from node `a` in the static aggregated graph is:
$\mathcal{R}_s(a) = \{b, c, d, e, f\}$ and $\mathcal{R}_d(a) \subset \mathcal{R}_s(a)$.

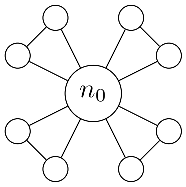
c) We always have $|\mathcal{R}_s(x)| \geqslant |\mathcal{R}_d(x)|$

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. The graph G_4 is:



Q2. In the general case, n_0 belongs to n triangles, and it is connected to $2n$ nodes, thus it has $\frac{2n(2n-1)}{2}$ pairs of neighbors. So $cc(n_0) = \frac{n}{n.(2n-1)} = \frac{1}{2n-1}$.

In the case where $n = 4$, $cc(n_0) = \frac{1}{7}$.

Any other node x only has two neighbors which are always connected so $cc(x) = 1$.

In the case of G_4, there is one node with $cc = \frac{1}{7}$ and 8 nodes with $cc = 1$, so $\overline{cc} \simeq 0.90$.

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. G_4 contains $n = 4$ triangles, $n.(2n-1) = 28$ forks having n_0 as central node, and 2 other forks per triangle, then 36 forks. So its transitive ratio is $\frac{3.4}{36} = \frac{1}{3}$.

Q4. In the general case, $tr(G_n) = \frac{3n}{n(2n-1)+2n} = \frac{3}{2n+1}$.

Following Q2, $\overline{cc}(G_n) = \frac{1}{2n+1}\left(\frac{1}{2n+1} + 2n\right)$

We observe that $tr(G_n) \to 0$, while $\overline{cc}(G_n) \to 1$ when $n \to \infty$.

Q5. The clustering coefficient of x is the probability that 2 random neighbors of x are connected, and thus the graph clustering coefficient is the average probability that when selecting a node, 2 random neighbors of this node are connected. The transitive ratio is the probability that a randomly selected fork is actually a triangle.

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.*

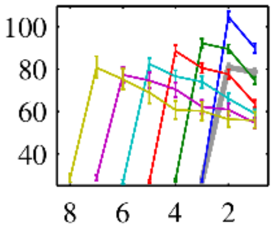
The authors have made a simplified replica of *Wikipedia.* They have organized an online game on this website and report its results. The principle of the game is as follows: given a source page and a target page, players have to find the shortest possible path from the source page to the target page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



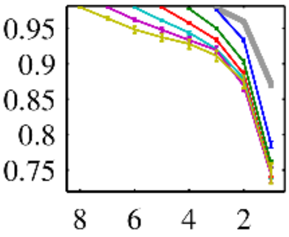
a) The point of coordinates (4,85) on the red curve means that the out-degree of nodes located at distance 4 to the target on paths of length 5 have an average out-degree of 85.

b) At fixed path length, the highest out-degree nodes are the second node visited, that is to say the page visited after the first click.

c) Pages with high out-degree are certainly generalist pages on Wikipedia, so players are looking for general topic pages 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) Considering the yellow curve, the TF-IDF is slowly decreasing from distance 8 to the target to distance 3 to the target, then it decreases abruptly during the last two steps.

b) So the content of the page remains quite different compared to the target as long as the page is more distant than 2 clicks to the target, but the last and second-to-last pages visited have content much more similar to the target.

Q3. So the typical strategy is to find first a general topic page which may be quite different from the target page, hoping that this would widen the possibilities of paths to the target. Then during a few steps, the player is looking for a specific path that could bring him or her semantically closer to the target. When a possible path is detected, the player gets very fast to the target by moving from pages to pages which are semantically closer.

This strategy is quite similar to what people were collectively doing in Milgram experiment: first people try to find a hub, someone who has a large range of acquaintances that gives a large range of paths available, then they search for a direct path to the target using more precise semantic information.

Exercice 5 — *Listing triangles (7pts)*

Consider the following algorithm:

Algorithm 1 Listing triangles

```
Let $\eta$ be a total ordering on the nodes of the input graph $G$
$\vec{G} \leftarrow$ directed version of $G$, where $v \rightarrow u$ 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 "-"
            $W \leftarrow \Delta_u^{?_2} \cap \Delta_v^{?_3}$          ▷ $\Delta_u^+$: out-neighbors of $u$ in $\vec{G}$
            for each node $w \in W$ do          ▷ $\Delta_u^-$: in-neighbors of $u$ in $\vec{G}$
                output triangle $\{u, v, w\}$
```

Q1. Case a) $?_1 = -$, $?_2 = -$, $?_3 = -$.

As seen in class, this case gives a correct algorithm: for each directed edge (s, t) in the DAG, it computes the intersections of the in-neighbors of nodes s and t. Each triangle is thus counted once and only once.

Case b) $?_1 = +$, $?_2 = -$, $?_3 = +$.

This case does not give a correct algorithm as it outputs no triangles: for each directed edge (s, t) it computes the intersection of the in-neighbors of node s with the out-neighbors of node t and the intersection is empty (otherwise there is a cycle in the underlying DAG).

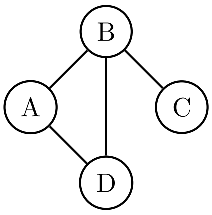
Q2. It suffices to store the out-neighbors of each node in an adjacency array data structure such that each array containing the out-neighbors is sorted (using bucket sort for efficiency). This requires $O(m + n)$ memory. The intersection of the out-neighbors of nodes u and v can then be computed in $O(d_u^+ + d_v^+)$ time, leading to the stated running time.

Q3. As seen in class, if the core ordering is used, then for each node u we have $d_u^+ \leqslant \delta$. This implies that, for each edge (u, v) we have $d_u^+ + d_v^+ \leqslant 2\delta$ and thus the stated complexity is obtained.

Q4. Given a node w, d_w^+ appears exactly d_w times in the sum, we thus obtain the equality. The ordering should thus minimize the right hand side of the equality and thus the largest the degree of a node u, the smallest d_u^+: this is achieved using the degree ordering.

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.*

The explanation is the following: the neighbors of a node are not random nodes of the graph, consequently there is a degree bias in the nodes appearing in the computation of the average of all $\overline{x_i}$. For example, a node with degree 0 has no neighbor, thus it will never appear in the computation of the average of all $\overline{x_i}$, a node with degree 1 appears once, a node with degree two appears twice, etc. The number of times a node appears in the computation of $\frac{1}{n}\sum_{i=1}^{n}\overline{x_i}$ is proportional to its degree.