MU5IN075 – Networks Analysis and Mining

Sorbonne University
Master of Computer Science – Networks specialty

Final Exam: correction

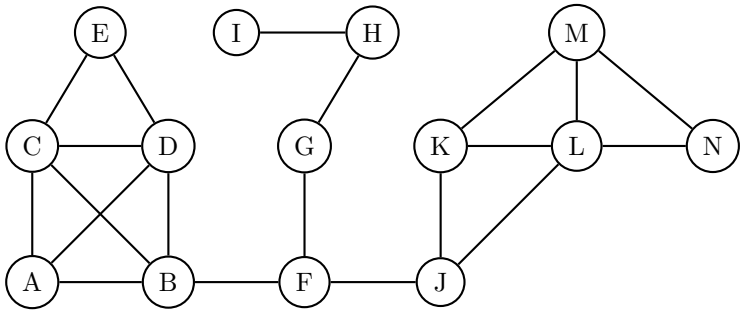
February 8 2022

Esteban Bautista Ruiz and Lionel Tabourier

The exam is 2hrs long. All documents are authorized. Points are only indicative. Good luck!

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

Consider the following graph:



Q1. degree 1: {I}
degree 2: {H, G, N, E}
degree 3: {M, K, F, A, J}
degree 4: {B, C, D, L}
Hence the points of the CDD: (1,1), (2,5); (3,10); (4,14)

Q2. cc(K)=2/3 ; cc(F)=0 ; cc(C)=4/6

Q3. One possibility (between brackets: father of the node in the tree)
root: G
distance 1: H(G), F(G)
distance 2: I(H), B(F), J(F)
distance 3: A(B), C(B), D(B), K(J), L(J)
distance 4: E(C), M(K), N(L)

Q4. With the tree above distances from G:
1: 2 , 2: 3, 3: 5, 4: 3.

Q5. Closeness centrality is defined as $C_c(G) = \frac{1}{\sum_{i \neq G} d(i,G)}$
So $C_c(G) = \frac{1}{2+3.2+5.3+3.4} = \frac{1}{35}$

Exercice 2 — *Understanding the course (5pts)*

In this exercise, you are asked to answer these questions and **justify your answer** with one or two simple sentences.

Q1. The experiment tends to indicate that Mr Jacobs is probably located on many shortest paths from the sources to the target, he would probably have a high betweenness centrality in the network.

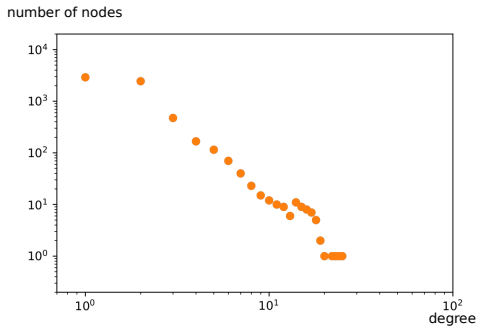
Q2. For instance, the Erdős-Rényi model, which exhibit a typical average distance between nodes in $log(n)$. As a random model with fixed density exhibits small typical distances, it is not surprising that a real network with the same density also has this property.

Q3. The existence of hubs in the network: if a hub becomes infected, it can spread the infection to all its neighbors, reigniting the spreading process. Another good answer was that the infection goes through a weak tie to reach a new community.

Q4. No, a collaborative filtering system tends to recommend content based only on other evaluations. It is true for content-based recommenders though.

Q5. We have measured the degree distribution of a subpart of the Internet (at the IP level) by repeating traceroute measurements from a source S to 3000 different destinations. By putting together all the paths obtained using traceroutes, we obtain the *observed network*. The underlying real IP-network is simply called the *real network*.

We measure that the degree distribution of the *observed network* is as follows:

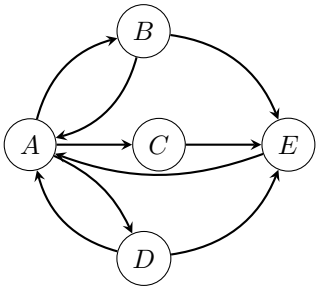


Do you think that the following statements are true? If not, explain why.

- There are about 3000 nodes of degree 1 in the *observed network*: true.
- The degree distribution of the *real network* follows a power law: false, as seen in the course, we can observe a heterogeneous degree distribution from both a homogeneous and a heterogeneous real distribution.
- The maximum degree of the *real network* is equal or lower than 25: false, the maximum degree in the real network cannot be lower than the one in the observed network.
- The maximum degree of the *real network* is equal or larger than 25: true.

Exercice 3 — *Comparison between PageRank and HITS (5pts)*

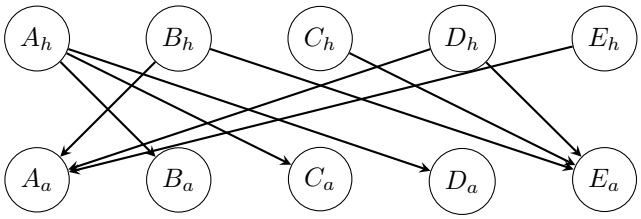
We consider the following directed graph G_d:



Q1. If $PR(A) = x$ then the algorithm in its stationary state imposes $PR(B) = PR(C) = PR(D) = \frac{x}{3}$ and $PR(E) = PR(B)/2 + PR(C) + PR(D)/2 = \frac{2x}{3}$.
The normalization condition $PR(E)+PR(B)+PR(C)+PR(D)+PR(A) = 1$ imposes $x = \frac{3}{8}$.
So finally, $PR(A) = \frac{3}{8}$,$PR(B) = PR(C) = PR(D) = \frac{1}{8}$ and $PR(E) = \frac{2}{8} = \frac{1}{4}$.

Q2. If the edge (C, E) didn't exist, C would be a dead-end in the graph, consequently the PageRank at this node disappears. The renormalization phase of the algorithm allows to avoid this issue.

Q3. Bipartite graph obtained from the transformation of the initial graph G_d:



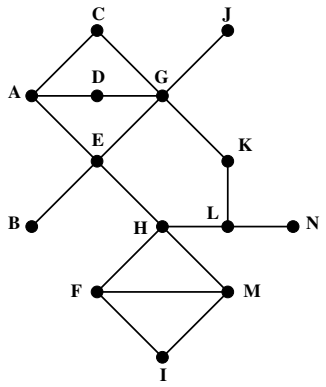
Q4.
- First iteration of HITS:
  Authority scores. $A_a : 3$, $B_a : 1$, $C_a : 1$, $D_a : 1$, $E_a : 3$.
  Hub scores: $A_h : 3$, $B_h : 6$, $C_h : 3$, $D_h : 6$, $E_h : 3$.
  Normalized authority scores: $A_a : \frac{1}{3}$, $B_a : \frac{1}{9}$, $C_a : \frac{1}{9}$, $D_a : \frac{1}{9}$, $E_a : \frac{1}{3}$.
  Normalized hub scores: $A_h : \frac{1}{7}$, $B_h : \frac{2}{7}$, $C_h : \frac{1}{7}$, $D_h : \frac{2}{7}$, $E_h : \frac{1}{7}$.

- Second iteration of HITS:
  Authority scores. $A_a : \frac{5}{7}$, $B_a : \frac{1}{7}$, $C_a : \frac{1}{7}$, $D_a : \frac{1}{7}$, $E_a : \frac{5}{7}$.
  Hub scores: $A_h : \frac{3}{7}$, $B_h : \frac{10}{7}$, $C_h : \frac{5}{7}$, $D_h : \frac{10}{7}$, $E_h : \frac{5}{7}$.
  Normalized authority scores: $A_a : \frac{5}{13}$, $B_a : \frac{1}{13}$, $C_a : \frac{1}{13}$, $D_a : \frac{1}{13}$, $E_a : \frac{5}{13}$.
  Normalized hub scores: $A_h : \frac{3}{33}$, $B_h : \frac{10}{33}$, $C_h : \frac{5}{33}$, $D_h : \frac{10}{33}$, $E_h : \frac{5}{33}$.

Q5. We can suggest that HITS will make A and E the largest authority scores after a large number of iterations and that it will make B and D the largest hub scores.

Exercice 4 — *Link prediction using Borda method (8pts)*

The network below represents a social network on year A. We aim at predicting links appearing during year $A+1$, knowing the network on year A.



As in the course, we denote the set of neighbors of node i: $\mathcal{N}(i)$.

To simplify the calculations, we consider from now on that the only edges that can appear on year $A+1$ are chosen among the 20 following candidate pairs of nodes:

(A,B), (A,F), (A,G), (A,H), (A,J), (B,F), (C,D), (C,E), (C,J), (D,E),
(E,F), (E,K), (G,H), (G,L), (H,K), (I,H), (J,K), (L,E), (L,M), (M,N).

Moreover, each node has a feature s which represents the academic level of the person (from 1 to 9):

node	A	B	C	D	E	F	G	H	I	J	K	L	M	N
s	6	2	2	5	4	2	8	4	1	9	8	8	3	4

We summarize the answers in the following table: for each candidate pair, we compute its PA, CN, sim.

pair	(A,B)	(A,F)	(A,G)	(A,H)	(A,J)	(B,F)	(C,D)	(C,E)	(C,J)	(D,E)
PA	3	9	**15**	**12**	3	3	4	8	2	8
CN	1	0	**3**	0	0	0	**2**	**2**	1	**2**
sim	5	5	7	7	6	**9**	6	7	2	**8**
Borda	0	0	**9**	3	0	**5**	4	4	0	**7**

pair	(E,F)	(E,K)	(G,H)	(G,L)	(H,K)	(I,H)	(J,K)	(L,E)	(L,M)	(M,N)
PA	**12**	8	**20**	**15**	8	8	2	**12**	9	3
CN	1	1	1	1	1	**2**	1	1	1	0
sim	7	5	5	**9**	5	6	**8**	5	4	**8**
Borda	2	0	**5**	**9**	0	4	3	2	0	3

Q1. The top five pairs for CN are
1. (A,G)
2. (C,D), (C,E), (D,E), (I,H)

Q2. The top five pairs for PA are
1. (G,H)
2. (G,L), (A,G)
3. (A,H), (E,L), (E,F) - or any combination of 2 out of these 3 pairs

Q3. The top five pairs for sim are
1. (B,F), (G,L)
3. (D,E), (J,K), (M,N)

We observe experimentally that 5 links indeed appear on year $A+1$, that is the links:

(A,G), (B,F), (G,L), (J,C), (J,K).

Q4. The number of true positive predictions are:
for CN: 1 (pair (A,G))
for PA: 2 (pairs (G,L),(A,G))
for sim: 3 (pairs (B,F), (G,L), (J,K))
Thus, the number of false positive are 4 for CN, 3 for PA and 2 for sim.

Q5. Precision is the number of true positive divided by the total number of predictions, so:
$\frac{1}{5}$ for CN, $\frac{1}{5}$ for PA, $\frac{3}{5}$ for CN.
And the recall is the number of true positive divided by the total number of links to predict, which is also 5, so the recall values are identical.

Q6. We report the Borda score for each pair of nodes in the table above. The top five pairs for Borda are
1. (A,G), (G,L)
3. (D,E)
4. (B,F), (G,H)

Q7. The number of true positive predictions for Borda ranking is 3 (pairs (B,F), (A,G), (G,L)) and 2 false positive (pairs (G,H) and (D,E))

Q8. As for sim, precision and recall are both $\frac{3}{5}$.

Exercice 5 — *Community detection (6pts)*

Reminders on the notion of conductance

Consider an undirected graph $G(V, E)$. For an arbitrary pair of vertices $u, v \in V$, we encode the existence of an edge between u and v as follows:

$$a_{uv} = \begin{cases} 1 & u \text{ is connected to } v \\ 0 & \text{otherwise} \end{cases}$$

Let $S \subseteq V$ be a subset of vertices of G. The cut of S, denoted $cut(S, S^c)$, is the number of edges between the set S and its complement S^c. Therefore, if S represents the nodes of a community, the cut of S is the number of edges that connect the community S to the other communities of the graph. The cut is more precisely defined as:

$$cut(S, S^c) = \sum_{u \in S} \sum_{v \in S^c} a_{uv}.$$

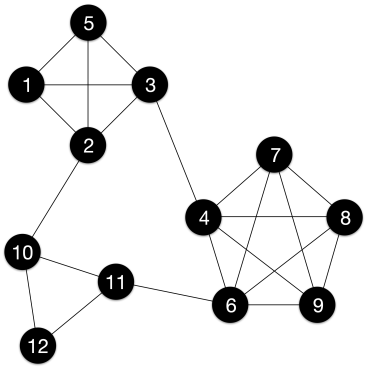
The volume of S, denoted $vol(S)$, is the sum of degrees of the vertices of S. Therefore, if S represents a community, the volume of S indicates the number of edges that have an endpoint in the community S. It is defined as:

$$vol(S) = \sum_{u \in S} d(u).$$

The conductance of the set S, denoted h_S, is a metric that allows us to evaluate if a chosen set S can be considered as a good community or not. It is defined as follows:

$$h_S = \frac{cut(S, S^c)}{min(vol(S), vol(S^c))}.$$

We consider now the following graph G:



Q1. (open question) Explain the reasons why h_S can be used to evaluate if S is a good community or not. Your answer must address the following three points:

(a) the relationship between h_S and the definition of communities

(b) the range of values that h_S can take

(c) how h_S must be used to evaluate communities.

Q2. For the graph G, what is the conductance of the following groups:

(a) $S = \{1, 2, 3, 5\}$

(b) $S = \{4, 6, 7, 8, 9\}$

(c) $S = \{10, 11, 12\}$

(d) $S = \{7, 8, 9\}$

(e) $S = \{2, 12, 7\}$

Q3. Write the expression of the optimization problem (based on conductance) that must be solved in order to partition all the graph at once into several disjoint communities.

Q4. We aim to identify the set S^* that has the smallest conductance. What is the time complexity of solving this problem exactly?

Q5. (open question) Modularity is another metric to verify if a set S is a good community or not. Explain in which way modularity differs from the conductance. It is not necessary to write equations but your answer must cover the following three points:

(a) how modularity verifies if S satisfies the community criterium

(b) the range of values of the modularity metric

(c) the optimization problem based on modularity.