MU5IN075 – Networks Analysis and Mining

Sorbonne University
Master of Computer Science – Networks specialty

Final Exam

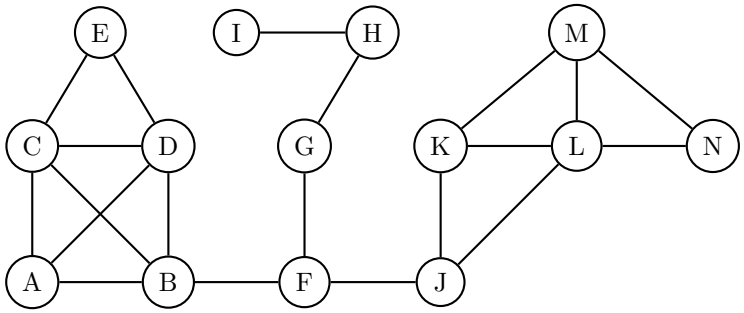
February 8 2022

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. Plot the cumulative degree distribution (CDD) of this graph.
Reminder: the CDD of a graph shows on the Y-axis the number of nodes which have a lower or equal degree to the value on the X-axis.

Q2. Compute the clustering coefficient of nodes K, F and C.

Q3. Implement a Breadth First Search (*BFS*) from node G. Give the result of the BFS in the form of a tree.

Q4. Deduce from Q3 the distribution of distances from source node G to all other nodes in the graph.

Q5. Deduce also from Q3 the closeness centrality of node G.

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. We remind that Milgram's small-world experiment consisted in asking several persons located in Nebraska to send a file to a target located in Michigan using exclusively the acquaintance network. In this experiment, we observe that a high fraction of the files (25%) that reached the target went through the hands of a man called Mr.Jacobs.

Supposing that we know exactly the acquaintance network, which measurement on this graph would account for the particular role of Mr.Jacobs?

Q2. We stated in the course that the average distance between two nodes in a real network is small in general and that this property is not surprising.

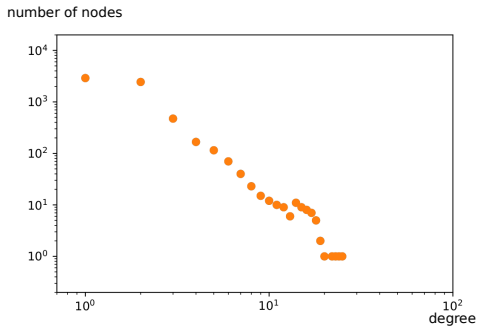
Which random network model allows to make this statement and why?

Q3. Which characteristic of the structure of real-world networks (which does not exist in an Erdős-Rényi network for instance), explains that an epidemic spreading which is about to disappear in a real network can suddenly restart?

Q4. Is it true that a recommender system based on collaborative filtering tends to propose always the same kind of content to a user?

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:

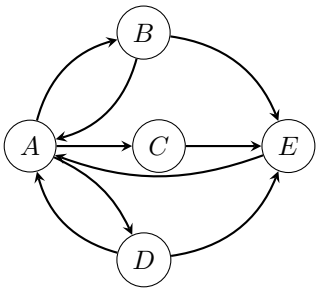


Do you think that the following statements are true? If not, explain why.

- There are about 3000 nodes of degree 1 in the *observed network*.
- The degree distribution of the *real network* follows a power law.
- The maximum degree of the *real network* is equal or lower than 25.
- The maximum degree of the *real network* is equal or larger than 25.

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

We consider the following directed graph G_d:



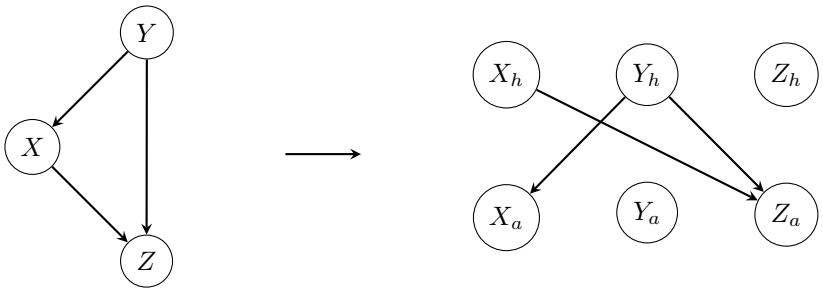
Q1. Compute the stationary state of the PageRank algorithm (of course 11 on Information Retrieval) on this graph.

Q2. What would happen if the edge (C, E) didn't exist? What phase of the PageRank algorithm is imperative to avoid this issue?

Now, we propose to compare the PageRank algorithm to the HITS algorithm. To do so, we start by transforming the initial graph into a bipartite graph in the following way:

- each node N is transformed into 2 nodes N_h and N_a (h stands for hub, a for authority)
- if there is an arc (Y, Z) in the initial graph, it is replaced by the arc (Y_h, Z_a).

In order to illustrate this, we give the transformation of the following graph:



Q3. Draw the graph obtained from the initial graph G_d.

Q4. Apply 2 iterations of the HITS algorithm on this graph. To help you, we remind below how HITS works.

Q5. Without iterating HITS further, can you make a hypothesis on the node(s) which will have the highest authority score after a large number of iterations? Same question for the node(s) which will have the highest hub score after a large number of iterations.

Reminder on HITS algorithm:

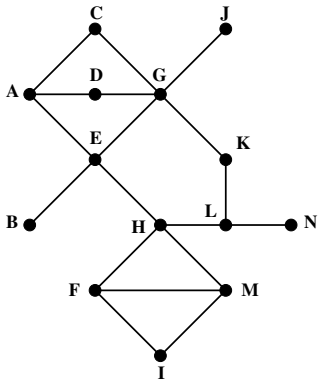
- *initialization: we initialize the hub scores of all hub nodes to 1 and the authority scores of all authority nodes to 0.*

then we iterate the following loop until it converges:

1. *authority update: for every authority node, its authority score is replaced by the sum of the hub scores of hub nodes pointing at it,*
2. *hub update: for every hub node, its hub score is replaced by the sum of the authority scores of the authority nodes at which it is pointing,*
3. *normalizing authorities: each authority score is replaced by itself divided by the sum of all authority scores,*
4. *normalizing hubs: each hub score is replaced by itself divided by the sum of all 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).

We want to predict the apparition of 5 new edges in the network using the ranking method.

Q1. We use the number of common neighbors $CN(i,j) = |\mathcal{N}(i) \cap \mathcal{N}(j)|$ to rank the pairs of nodes. Give the 5 candidate pairs of nodes that are most likely to be connected according to this ranking.

Q2. We use the preferential attachment index $PA(i,j) = |\mathcal{N}(i)|.|\mathcal{N}(j)|$ to rank the pairs of nodes. Give the 5 candidate pairs of nodes that are most likely to be connected according to this ranking.

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

Q3. We know that in this network, two individuals with a similar academic level have a higher probability to be connected. Using the following similarity measurement to rank the pairs of nodes,

$$sim(i,j) = 9 - |s(j) - s(i)|$$

give the 5 candidate pairs of nodes that are most likely to be connected according to this measure.

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. Give the number of true positive predictions and the number of false positive predictions using the ranking method for the three scores proposed (CN, PA and sim).

Q5. Deduce the precision and recall of the predictions for each of the three scores.

Now, we propose to improve our predictions by using Borda method, which principle is described here:

- *If a pair appears first in a ranking, it is given 5 points, if it is ranked second it is given 4 points and so on until the fifth rank (1 point).*
- *Note that if several pairs are ranked equally, they are given the same number of points. For instance, if 3 pairs are ranked second because they have a similar score, they all receive 4 points and the next one will be ranked fifth and will receive 1 point.*
- *A pair which does not appear in the top 5 does not receive any point for this ranking.*
- *We sum all the points obtained for each candidate pair over all rankings and then we create a new ranking based on this sum of points.*

Q6. Create the ranking obtained using Borda method from the three rankings proposed previously.

Q7. Compute the number of true positive and false positive if we predict that the top-5 pairs of this new ranking are indeed linked.

Q8. Deduce the precision and recall for this new prediction.

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:

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

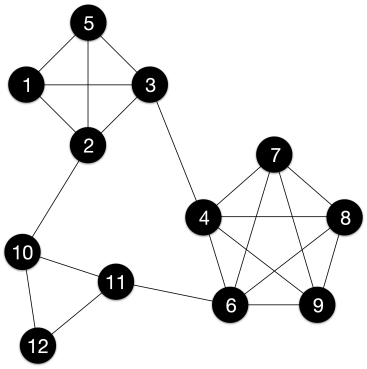
The volume of S, denoted $vol(S)$, is the sum of degrees of the vertices of S. Therefore, if S represents a community, the volume of S indicates the number of edges that have an endpoint in the community S. It is defined as:

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

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.