RESYS 2019 TP1

Vincent Cabeli vincent.cabeli@curie.fr - Hervé Isambert

September 2019

1 Graph theory : measures and network topology

1.1 Definitions and measures

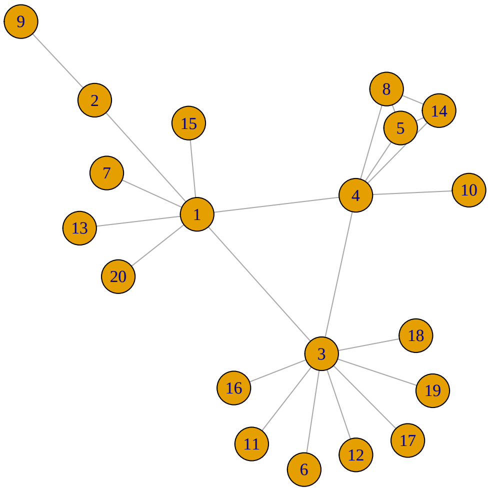


Figure 1: A simple graph

Answer the following questions related to the graph in Figure 1 :

1. Find two different ways to write down this network's topology.
2. What are the degrees of the nodes 1, 2, 5 ? Specify in, out and total degree if applicable.
3. How many connected components are there ? How many cliques (fully connected subgraphs) of more than 3 nodes ?
4. What is the betweenness centrality of the node 5 ? Which has a higher centrality : 3 or 4 ? Which nodes have non-zero betweenness ?
5. What are the clustering coefficients of the nodes 2, 4, 5 ? Reminder : the clustering coefficient can be computed as the number of triangles among the neighbours of a node over the number of possible subgraphs with 2 edges over all its neighbours.

1.2 Algorithms

1.2.1 Finding the shortest paths

The betweenness centrality measures requires to find the shortest path(s) between two nodes. Propose an algorithm that, given a network and a pair of nodes, returns the shortest path between them in the most efficient way you can think of.

Bonus question : which data structures are the most efficient for your solution ?

1.2.2 Directed acyclic graph generation

Suppose you want to direct the edges of the graph in Figure 1 in such a way that you do not create cycles. How would you do it ?

Bonus question : how many acyclic orientations are there for a given graph ?

1.2.3 Generating random networks

A common way to estimate the statistical significance of the results of an analysis is to generate a large number of randomized datasets and see how often they can be replicated "by chance". How would you go about generating random networks ? Try to find at least two ways that will result in different network connectivity. With your methods, can you find the average degree ?

1.3 Guess the network

In this section, you have three randomly generated networks with the same number of nodes ($|V| = 1000$) and edges ($|E| = 5000$), but with very different topologies. Figure 2 gives you the degree distribution in normal and log-log scale of each network. Figure 3 represents, **not in the same order**, examples of smaller networks with the same density generated the same way as those in Figure 2.

First, try to match the degree distribution of Figure 2 to each of the smaller examples of 3. Then, define them formally using the different types of networks that you have seen in class.

After guessing their topologies, can you order them :

1. from lowest to highest average path length ?
2. from lowest to highest average clustering coefficient ?

1.4 Using igraph

At this point we will switch to R and use the `igraph` package to answer all of the earlier questions programmatically.

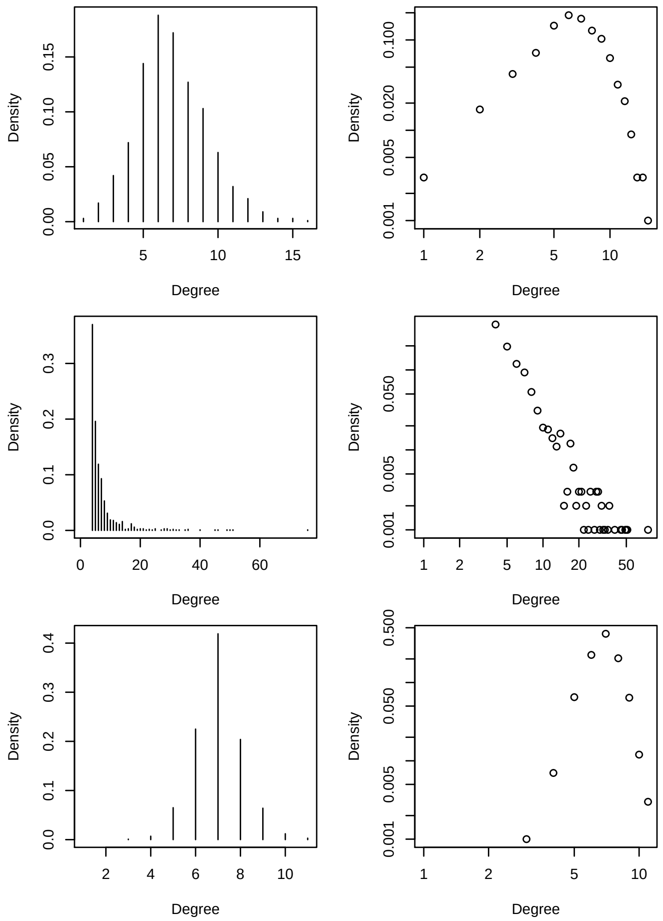


Figure 2: Degree distribution of 3 networks

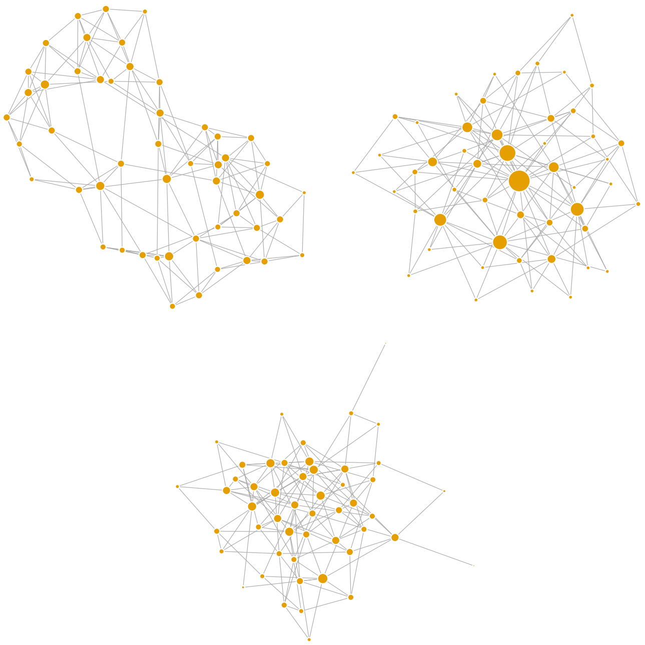
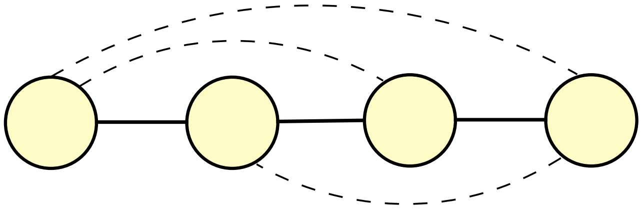


Figure 3: Network examples

2 Network inference from observation studies using ARACNE

Up to now, we were working on a graph that was given *as is*, but in most cases you will first have to infer the network from available data. One of the most popular approach used in bioinformatics is ARACNE which is based on information theory and the Data Processing Inequality (DPI).



ARACNE removes the weakest of the 3 links

The idea behind ARACNE is very simple and is motivated by looking at gene regulatory networks. Imagine that two genes $g1$ and $g3$ interact only through a third gene $g2$, (i.e., the interaction network is $g1 \leftrightarrow g2 \leftrightarrow g3$ and no alternative path exists between $g1$ and $g3$). Then the DPI states that $I(g1;g3) \leq min(I(g1;g2), I(g2;g3))$. If you assume that your network is a tree, then the lowest of the three MI between 3 nodes can only come from an indirect relation, and you can remove it from the graph (see example above).

ARACNE starts from a complete graph and examines each gene triplet for which all three MIs are greater than 0 and removes the edge with the smallest value. Each triplet is analyzed irrespectively of whether its edges have been marked for removal by prior DPI applications to different triplets. Thus the network reconstructed by the algorithm is independent of the order in which the triplets are examined.

Below is the Mutual information matrix measured between 6 genes. To simplify the problem, consider that anything < 0.25 is non significant.

	x1	x2	x3	x4	x5	x6
x1		0.12	0.21	0.11	0.08	0.09
x2	0.12		0.46	0.23	0.16	0.18
x3	0.21	0.46		0.47	0.29	0.35
x4	0.11	0.23	0.47		0.64	0.88
x5	0.08	0.16	0.29	0.64		1.00
x6	0.09	0.18	0.35	0.88	1.00	

Apply the ARACNE algorithm to reconstruct the graph from the MI matrix.