Data Mining
Homework Assignment #11

Dmytro Fishman, Anna Leontjeva and Jaak Vilo

May 9, 2014

The goal for this homework is to get you started with basics of social network analysis.
There are many different network analyzing tools. Choose the one you prefer. Some of the well-known tools and packages are: NetworkX and igraph for Python, JUNG for Java, igraph for R, Gephi for the visualization with some built-in calculations and even NodeXL free template for Excel.
Most of the exercises require you to check presentation slides for the definitions.

Task 1

In this task, we will use the techniques of Social Network analysis to study a virus spread. Imagine the following situation: terrified biologists came to the Institute of Computer Science seeking for a help. Their email network was infected by the virus that was created by the student that received 'B' for his Master's Thesis and got offended. Your goal is to help poor biologists to estimate the worst-case scenario of this virus spread.
Biologists observed that if virus infects a node, it always infects all its immediate neighbors, if they are not already infected (100% of infection rate). Also, we know that virus travels only along the edge direction (e.g. if virus infects node A, which only has an incoming edge from node B, node B will not be infected).
Biologists provided you with their directed anonymous email network (email_virus.txt) that you can access on the course web-page.
Load the data. To get the first insights about biologists' network, calculate the list of the following statistics:

- number of nodes in the network
- number of edges in the network
- number of nodes with a self-loop
- number of mutual connections or *reciprocated* edges, i.e if there is a directed edge from node a to node b, there is also an edge from b to a.

- number of nodes with zero indegree (those that have only outgoing edges)
- number of nodes with zero outdegree (those that have only ingoing edges)
- degree distribution of the given network
- optionally calculate whatever measure you deem appropriate for better understanding

What intuition you can gather from these numbers?

Task 2

Next, biologists ask us to estimate the vulnerability of their network. In order to measure it, you have to calculate average virus spread assuming that the initial infected node is chosen uniformly at random. What is the probability that virus will affect at least 30% of the network nodes ('large scale' epidemics emerges).
For that, consider the following simplified version of the graph.

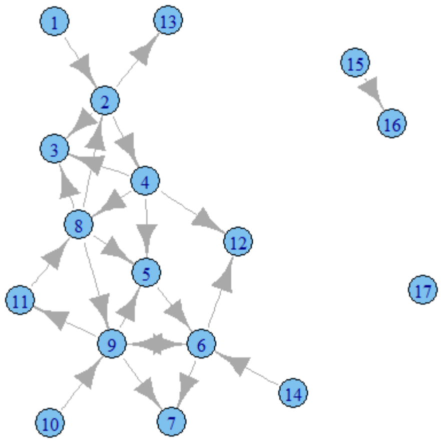

Figure: simplified version of the graph

a. According to the virus transmission rule, calculate number of infected nodes in three different cases:

    (a) the seed of the infection is node 1
    (b) the seed of the infection is node 8
    (c) the seed of the infection is node 12

b. Here we introduce the notion of Bow Tie structure, which is a recent concept that grasps the essence of both biological networks and the representation of links in the Internet. It divides a graph into 4 basic components (and two secondary that we omit for this task):

   - strongly connected component (SCC), which is the core of the graph, where all nodes can reach one other along directed links
   - ”IN“ component that consists of nodes that can reach the SCC, but cannot be reached from it.
   - ”OUT“ component that consists of nodes that are accessible from the SCC, but do not link back to it
   - disconnected components are those nodes that are not connected to neither IN, OUT or SCC components.

   For more information on described above graph structure read article by Andrei Broder et al.: `http://snap.stanford.edu/class/cs224w-readings/broder00bowtie.pdf`
   In this part, your task is to calculate the proportion of nodes that belong to SCC, IN, OUT and disconnected components.

c. Based on estimated proportions in the part ’b’, calculate the probability of emerging a ’large scale’ epidemics (at least 30% of the network nodes infected) given that the initial infected node is chosen uniformly at random from the network nodes.

Task 3

In this task we will explore generative mathematical models that try to simulate real-world network according to some characteristics. We will generate three models and compare it with the real-world graph of an academic collaboration network. For this:

- generate random graph according to Erdos-Renyi model with 5242 nodes and 10484 edges

  ```
  #Hint for R: use package igraph and function
  erdos.renyi.game(n = 5242, 10484, type = c("gnm"))
  #Hint for python library NetworkX:
  gnm_random_graph(5242, 10484)
  ```

- generate a graph with small world model (Watts and Strogatz model) so that it has 5242 nodes and 10484 edges.

```
#In R use function
watts.strogatz.game(1, 5242, 2, 0.5)
#Hint for python NetworkX:
watts_strogatz_graph(5242, 2, 0.5)
```

- generate preferential attachment model (Barabasi model)

```
#Hint for R:
barabasi.game(5242, power = 1, m = 2)
#Hint for NetworkX in python
barabasi_albert_graph(5242, 2)
```

- load the real collaboration graph undirected_real_world_graph.txt from the course webpage

Plot the degree distribution of all three networks both on original and log-log scale. Highly recommended to place all distributions on one plot for comparison. Describe the differences and the shape of the distributions.

Task 4

The local clustering coefficient for a node v is defined as:

$$C_i = \frac{2\|e_i\|}{k_i(k_i - 1)}$$

where k_i is the degree of v_i and e_i is the number of edges between the neighbors of v_i. Your task is to find *average clustering coefficient* that is defined as

$$C = \frac{1}{\|V\|} \sum_{i \epsilon V} C_i$$

Note that if a node has 0 or 1 neighbor, we will ignore it. Calculate average clustering coefficient for all four networks in the previous task. Compare them and describe the difference.
In case of using built-in functions, one point will be given. If you write a script that calculates average clustering coefficient (instead of built-in function), you will earn additional point.

Task 5

In this task you have to come up with your own model for real-world networks generation. Come up with the algorithm and write down the pseudo code for your model. Implement the model and generate the graph of approximately

the same size as in previous tasks. Compare degree distribution and clustering coefficient of your graph and other networks.
As an example, you may introduce a notion of nodes dying after some time. Note that without implementation only half of the point will be given.

Task 6

In this bonus task you will help biologists to fight with the virus epidemic. Using the data email_virus.txt, calculate the same components as in task 2b and calculate the probability of emerging a 'large scale' epidemics as in task 2c. Hint: you may want to use BFS algorithm to estimate IN and OUT components.

In this homework we were inspired and adapted some of the materials from Juri Leskovec stanford course "Social and Information Network Analysis“.