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 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:

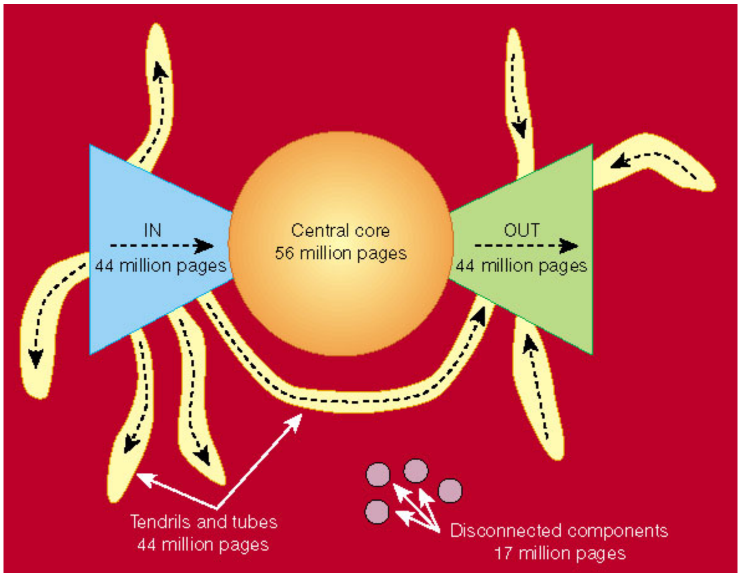
- number of nodes in the network
- number of edges in the network
- number of nodes with a self-loop
- number of mutual connections or *reciprocated* edges, i.e if there is a directed edge from node a to node b, there is also an edge from b to a.

- 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 step is to estimate the vulnerability of biologists' networks. We will approach this problem by using bow-tie structure. You can read about it here:
`http://www.nature.com/nature/journal/v405/n6783/full/405113a0.html`
or more thoroughly here: `http://snap.stanford.edu/class/cs224w-readings/broder00bowtie.pdf`.
Let us take a closer look at the following diagram:



Ignore tendrils and tubes in this task. For the virus spread we want to find four components of the network:

- proportion of nodes that belong to the SCC (Strongly Connected Component) (*CORE* of the network)
- proportion of nodes that belong to IN component (nodes that belong to the weak connected component and have zero indegree)

- proportion of nodes that belong to OUT component (nodes that belong to the weak connected component and have zero outdegree)
- proportion of disconnected components (do not belong to the weak component)

Task 3

Task 4

write your own function that generates erdos-renyi random graph with input parameter p, where p is the probability of edge creation. generate the graph with 50 nodes and at least five different p values. Plot the result.

Task 5

Community detection

Task 6