Network Analytics: Twitter pre-class assignment

Due Friday, July 29, 2016

Submit all your answers as a single PDF on NYU Classes.

1 Download data

For this assignment you will use Twitter data, provided in the assignment in NYU Classes. When you unzip data.zip, you will have access to the following data files:

1. graph complete.txt provides the edges of the graph in the form from → to. Each line is an edge where each node is separated by a space.
2. graph subset.txt: a subset of the complete network. This file contains roughly 1% of the total number of edges (randomly selected). Each line is an edge where each node is separated by a space.
3. ids to usernames.csv: maps the integer ids given in the two data files to the actual Twitter usernames of the users in our dataset. There are two comma separated fields in this file: the integer id and the string username.

This data was collected in July 2014. We used four major Twitter accounts (Snapchat, Dropbox and two others) as seed nodes, collecting their most recent 4000 followers. We did the same for each of these followers. Then, for each of these accounts, we collected all the accounts they follow. Your data set includes 59,974 Twitter users (nodes) and 73,277 follower relationships (edges).

2 Network structure visualization

Install the igraph package for **R** (suggested) or **Python**. To do so, in R type:[1]

```
# Download and install the package
install.packages("igraph")
```

Plot the network by using the information in the file graph subset.txt. Note that this is not the complete network, but only a subset of its edges. By visualizing the graph, you get an idea of the structure of the network you will be working on. In addition to plotting, comment on anything interesting you observe.

You may find it useful to treat this data file as being in ncol *format in* igraph. *In addition, playing with the size, color, and layout of objects may make it easier to visualize. Using* lay out=layout.kamada.kawai *for this graph generally gives good results.*

[1] See http://igraph.org/r/ or http://igraph.org/python/ for more information on igraph for R and Python.

3 Data analysis

For the rest of the assignment, use the complete graph contained in the file `graph complete.txt` and the username file `ids to usernames.csv`. It will be in in your best interest to using a programming language such as R or Python.

1. Plot the distribution of the of number of followers of each user in our dataset (x-axis number of followers, y-axis number of nodes). Please note, that for each edge, user *a* is said to be a follower of user *b* if there is some edge a → b. The following steps will outline one way to approach this problem.

   (a) Start by counting the number of followers of each user. You may use the `table` command in R or a `dict` in Python to compute the number of followers of each user. This is the same as the in-degree of each node in the graph.

   (b) You can then apply the same process you just used so that you can count the number of users (nodes) that have a particular number of followers (in-degree). This is the in-degree distribution.

   (c) Once you are done, you can use the default plotting environment in R, `ggplot`[2] in R, or `maplotlib`[3] in Python to plot the distribution. Note that you can avoid step (b) if you use the `geom density()` function in `ggplot` or the `hist()` method in `matplotlib`. However, you may approach this anyway you wish.

2. Transform the x-axis of the previous graph to logscale, to get a better understanding of the distribution. Note here that you have some users that have 0 followers. This means that using the log of the x-axis will fail since $log(0)$ will not be valid. Due to this, you should replace 0 with 0.1. Comment on what you observe.

3. Compute the average number of followers and the standard deviation. Comment on the result.

4. Report the Twitter usernames of the top 10 users with the most followers. If there are multiple ties, you may print any users with that particular number of followers.

[2] http://ggplot2.org/
[3] http://matplotlib.org/