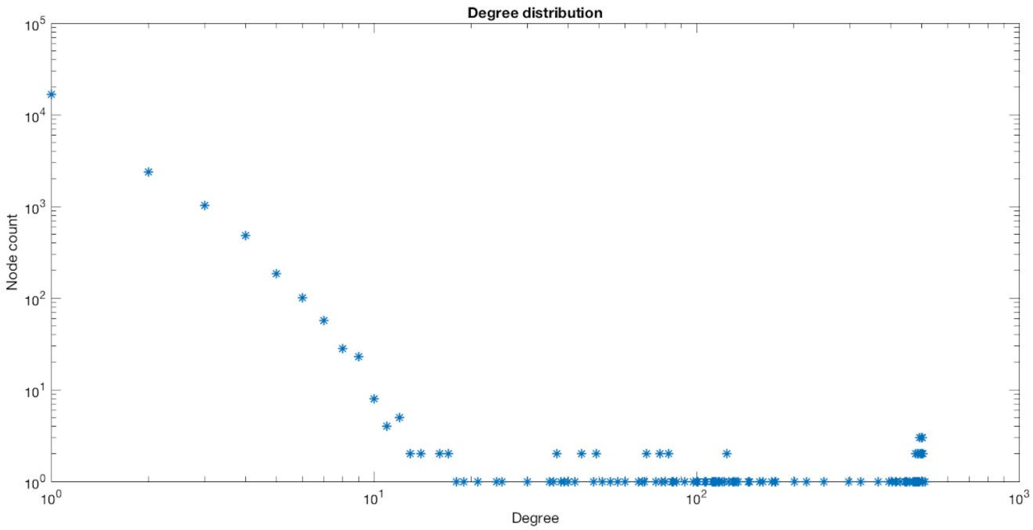
CS235 - Data Mining Techniques - Assignment

Instructor: Vagelis Papalexakis, University of California Riverside

Description

Introduction: In this assignment we are going to work with a real-world Twitter graph (adapted from [1,2] in order to be symmetric, thus *undirected*) in which we have injected a number of "suspicious" dense communities. As we saw in class, abnormally dense blocks in graphs are frequently associated with big communities or fraudulent users [3,4], therefore, our job in this assignment is to identify 1) how many injected blocks are there, and 2) which nodes are included in those blocks.

Along with the description, you are given a graph.txt with the edge-list of the graph. In addition to the actual graph, below we provide the degree distribution of the original clean graph, to use as a frame of reference. We are not including the clean graph as part of the assignment, because a simple diff between the two graphs may be able to reveal a lot about the questions we need to answer, and in reality we will only have the "unclean" graph. For all intents and purposes, however, you may use the clean graph degree distribution as the ideal scenario.



Question 1: First encounter with the data *[10%]*

1. Write code to read the graph into a matrix *[2.5%]*
2. Plot the spy-plot *[2.5%]*
3. Do you see any dense block(s) standing out? Annotate the spy plot by circling this block / these blocks. *[5%]*

Question 2: Degree distribution *[35%]*

The degree distribution of a graph is a very important feature of that graph. It has been shown that real-world graphs have very skewed degree distributions [5], as also evidenced by the distribution of the clean graph that is given to you above. Significant deviations from that "ideal" pattern may signify suspicious behavior. In this question, you will use the degree distribution of the graph in order to identify potentially abnormal blocks in the graph:

1. Write code to compute the degree per node. For this you may *not* use existing libraries that implement this functionality [10%]
2. Plot the degree distribution in log-log scale [10%]
3. Do you observe any differences between this degree distribution and the clean one? [5%]
4. Based on the degree distribution [*explain your process!!!*]
    a. How many abnormal blocks are there? [5%]
    b. Which nodes are included in each block? [5%]

Question 3: Singular value decomposition *[45%]*

The Singular Value Decomposition (SVD) is a very powerful tool that decomposes a matrix into a sum of elementary rank-one matrices. In our case, dense subgraphs in our graph typically manifest as dense blocks that are picked up by the SVD, since blocks are very well modeled as rank-one matrices (i.e., an outer product of two vectors). In this question, you will use the SVD to analyze the graph and identify any potential abnormal blocks. A word of warning: the matrix is very big but very sparse! Make sure you represent it as a "sparse" matrix and accordingly use SVD for sparse matrices (e.g., svds in Matlab). If you don't, you may experience your computer freezing (with all those zeros that are being represented as double floating point numbers) and that's not fun!

1. Literature survey
    a. Read and summarize the Eigenspokes [4] paper, in three short paragraphs: 1) what is the main problem, 2) what is the main innovation/idea, 3) what are the results. [**Important**: *use your own words!*] [10%]
2. Analysis:
    a. Find the number of singular values/vectors (aka rank) for which 90% of the original data is reconstructed correctly, in terms of squared (euclidean) error. [10%]
    b. Plot the left singular vectors corresponding to the top 5 singular values on a single plot [5%]
    c. Can you identify how many abnormal blocks there are from the above plot? [10%]
    d. For each of the top 5 left singular vectors: do the spy-plot of the induced subgraph corresponding to the top-100 nodes per singular vector vector. In order to determine the top nodes per vector, pick the nodes with the highest absolute value. [Note: Since the graph is symmetric, you may use only the left singular values to extract those nodes] [10%]

References:

[1] Twitter (icwsm) network dataset -- KONECT, April 2017 http://konect.uni-koblenz.de/networks/munmun_twitter_social

[2] Munmun De Choudhury, Yu-Ru Lin, Hari Sundaram, K. Selçuk Candan, Lexing Xie, and Aisling Kelliher. How does the data sampling strategy impact the discovery of information diffusion in social media? In ICWSM, pages 34--41, 2010.

[3] Shah, Neil, Alex Beutel, Brian Gallagher, and Christos Faloutsos. "Spotting suspicious link behavior with fbox: An adversarial perspective." In *2014 IEEE International Conference on Data Mining*, pp. 959-964. IEEE, 2014.

[4] Prakash, B. Aditya, Ashwin Sridharan, Mukund Seshadri, Sridhar Machiraju, and Christos Faloutsos. "Eigenspokes: Surprising patterns and scalable community chipping in large graphs." In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 435-448. Springer, Berlin, Heidelberg, 2010.

[5] Faloutsos, Michalis, Petros Faloutsos, and Christos Faloutsos. "On power-law relationships of the internet topology." In ACM SIGCOMM computer communication review, vol. 29, no. 4, pp. 251-262. ACM, 1999.

PLEASE READ BELOW - IMPORTANT INFORMATION

Programming language: For this assignment, you may use Matlab or Python. You may use existing packages for the SVD but you may not use graph-handling packages.

Deliverables: 1) A PDF report with all the plots and answers required, and 2) a .zip archive with all your code (please DON'T include the graph.txt, we already have it! :-)).

Academic Integrity: Each assignment should be done individually. You may discuss general approaches with other students in the class, and ask questions to the TA, but you must only submit work that is yours . If you receive help by any external sources (other than the TA and the instructor), you must properly credit those sources, and if the help is significant, the appropriate grade reduction will be applied. If you fail to do so, the instructor and the TA are obligated to take the appropriate actions outlined at http://conduct.ucr.edu/policies/academicintegrity.html. Please read carefully the UCR academic integrity policies included in the link.