Written Report – 6.419x Module 3

Name: (Felipe Mehsen Tufaile)

▪ Problem 1: Suggesting similar papers

Part (c): *(2 points) How does the time complexity of your solution involving matrix multiplication in part (a) compare to your friend's algorithm? As above, for a brief introduction to the big-O notation, refer to the optional problem 1.7 in Module 1.*

Solution:

In part (a) we identify the time complexity of our friend's algorithm to be O(n^3). That is, the algorithm's running time grows cubically with the number of nodes (n). However, our approach, which uses matrix multiplication, rely on how efficient NumPy is for multiplying matrices. According to NumPy's documentation, NumPy utilizes Basic Linear Algebra Subprograms (BLAS) to provide efficient low-level implementation of standard linear algebra algorithms. This library, in turn, implements the Strassen's algorithm, which gives a time complexity of O($n^{2.807}$), resulting in an approach that is $n^{0.193}$ times faster. To illustrate how fast it is, if we had a graph with 10^4 nodes, the matrix multiplication approach would be almost 6 times faster than our friend's approach.

Part (d): *(3 points) Bibliographic coupling and cocitation can both be taken as an indicator that papers deal with related material. However, they can in practice give noticeably different results. Why? Which measure is more appropriate as an indicator for similarity between papers?*

Solution:

Considering that co-citation indicates two papers that are citated by the same third paper and bibliographic coupling corresponds to the number of common citations between two papers, it is more likely that two **papers that have high bibliographic coupling are more alike** than two papers that have high co-citation. The explanation for this statement is that as the number of common references increase between two papers (high bibliographic coupling) the probability that these two papers are conducting similar studies increases, even more so when the subject of study of the common references is quite specific. On the other hand, two papers that are cited by a set of papers might be important papers for the community, but not necessarily similar. In order to state that two papers with high co-citation are similar we would have to guarantee that the set of papers that cite these two papers are also very similar.

▪ Problem 2: Investigating a time-varying criminal network

Part (c): *(2 points) Observe the plot you made in Part (a) Question 1. The number of nodes increases sharply over the first few phases then levels out. Comment on what you think may be causing this effect. Based on your answer, should you adjust your conclusions in Part (b) Question 5.*

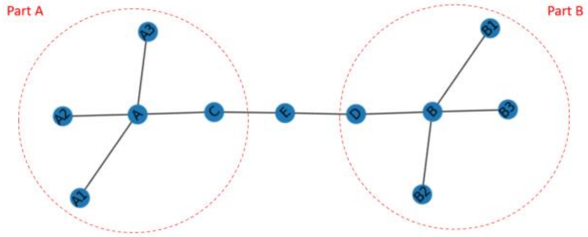
Solution:

The investigation lasted two years (1994-1996), in which 11 wiretap warrants were obtained, constituting an operation with 11 phases. Drugs seizure started on phase 4 and went on until phase 11, with exception of phase 5. In this sense, it is likely that before the first drug seizure, police officers were rapidly discovering new suspects through wiretapped conversations. However, after the first drug seizure, criminals possibly became more cautious and began to communicate more discreetly. To support this argument, we see that some nodes vanish and reaper in some of the 11 phases, suggesting that these perpetrators are still engaged in the illegal act, but it has become more difficult to detect them. In this context, we should disregard the first phases of the operation (phases 1, 2 and 3) in the calculation of the mean centralities in question 5 (b), since not all agents of the criminal operation would have been on the police list in these phases. This would bias the centrality calculation of some agents to lower values since we are setting the centrality o zero to a given agent that is missing in each phase.

Part (d): *(5 points) In the context of criminal networks, what would each of these metrics (including degree, betweenness, and eigenvector centrality) teach you about the importance of an actor's role in the traffic? In your own words, could you explain the limitations of degree centrality? In your opinion, which one would be most relevant to identify who is running the illegal activities of the group? Please justify.*

Solution:

The degree centrality will only capture importance up to one-hop neighbors of a node and may not be representative of the importance of the entire criminal network. To illustrate this problem, let's us consider the following network.



In this theoretical network, node A and B have the higher degree centrality among all nodes. However, node E is fundamental to hold the two portions of the graph together and this fact is not captured by degree centrality.

Eigenvector centrality will assign a relatively high importance value to a node that not only has many connections, but is also connected to other important nodes in the network. This centrality metric does a better job at capturing the importance of node E, since it takes into consideration that node E is also connected to other important nodes. However, the metric still doesn't capture the structural importance of node E to the network.

Betweenness centrality, on the other hand, will count the number of shortest paths that pass through a node, which allows the identification of nodes that acts as bridge or mediator in the flow of information between nodes and removing it could best break the network apart. In this sense, betweenness centrality will give higher centrality score to node E in the previous graph.

In the context of the criminal network, degree centrality and eigenvector centrality would indicate how well-connected a given criminal is to other criminals in the network, while betweenness centrality would allow to measure the degree at which perpetrators are high in the criminal organization hierarchy and, therefore, act by giving orders to other members in the organization. **In this regard, betweenness centrality would be the best centrality metric to identify who is running the illegal activities in the group** and removing the criminals with the highest betweenness centrality would best break (disturb) the criminal network.

Part (e): *(3 points) In real life, the police need to effectively use all the information they have gathered, to identify who is responsible for running the illegal activities of the group. Armed with a qualitative understanding of the centrality metrics from Part (d) and the quantitative analysis from part Part (b) Question 5, integrate and interpret the information you have to identify which players were most central (or important) to the operation.*

Solution:

As mentioned in the previous question, betweenness centrality would be the best metric to identify who is running the illegal activities in the group since it measures the degree at which perpetrators are high in the criminal organization hierarchy and, therefore, act by giving orders to other members in the organization.

Looking at the results of question 5 (b), we see that Daniel Serero (n1) and Ernesto Morales (n12) have the highest betweenness centrality, which is aligned with the goal of identifying key people in the criminal network, since we know that Serero is the mastermind of the entire network and Morales is intermediary between the Colombians and Serero in the cocaine import.

The third “player” with the highest betweenness centrality is Pierre Perlini (n3), who is responsible for executing Serero’s instructions. This result is also aligned with the goal at hand since the act of intermediating Serero’s instructions to other members in the network gives a central role to Pierre Perlini in the network.

Part (f): *(3 points) The change in the network from Phase X to X+1 coincides with a major event that took place during the actual investigation. Identify the event and explain how the change in centrality rankings and visual patterns, observed in the network plots above, relates to said event.*

Solution:

Phase X and phase X+1 correspond to phase 4 and phase 5, respectively. From the problem statement, we know that after the first seizure, happening in Phase 4, traffickers reoriented to cocaine import from Colombia, which is the major event mentioned in this problem. Once the cocaine import starts in phase 5, Ernesto Morales (n12), who until then wasn’t a central figure in the criminal network, becomes more relevant as he is put in charge of the cocaine operation, providing instructions to several subordinates. Looking at the betweenness centrality rank, we note that Morales (n12) becomes the second most central figure in the criminal network in phase 5, whereas in phase 4 his betweenness centrality metric was equal to zero. Therefore, betweenness centrality succeeds at capturing the structural change in the criminal network.

Part (g): *(4 points) While centrality helps explain the evolution of every player's role individually, we need to explore the global trends and incidents in the story in order to understand the behavior of the criminal enterprise. Describe the coarse pattern(s) you observe as the network evolves through the phases. Does the network evolution reflect the background story?*

Solution:

We observe some key changes in the criminal network as each phase develops, especially after the first drug seizure, in Phase 4. As mentioned in the previous questions, the criminal network expands its illegal operation to include cocaine import after phase 4, which results in a structural change in the criminal network: Morales (n12) becomes a new key player, and the network seems to form two cores that represents the original marijuana operation and the new cocaine operation. Additionally, after phase 4, some criminals seem to “come and go” in the network in future phases, which suggests that criminals are being more discreet in their communications.

It is also noticeable that as the phases develop, Daniel Serero begins to communicate more with non-traffickers (Players 83-110), investors, accountants, and transportation managers than with traffickers. This suggests the criminal network becomes more sophisticated as the phases develops with a more well-defined hierarchy, where players on the top of the hierarchy have less contact with traffickers, reducing their exposure.

Finally, we see that from time to time the cocaine import core is shown disconnected from the main core. Possibly, this is a result of the effort of the perpetrators to make it difficult for the police to investigate the criminal network. Nevertheless, it does show that the cocaine import become more mature and less dependent from the main core of the network.

Part (h): *(2 points) Are there other actors that play an important role but are not on the list of investigation (i.e., actors who are not among the 23 listed above)? List them, and explain why they are important.*

Solution:

Yes. If we calculate the betweenness centrality for all nodes in phase 11, we observe that node 41 is ranked as the most important node. This node appears as central (connected to many other nodes) in the cocaine import operation and seems to act as an intermediator between Serero (n1) and Morales (n2). Additionally, we note that removing this node we would best break the network separating the cocaine import core from the main core of the criminal network.

Part (i): *(2 points) What are the advantages of looking at the directed version vs. undirected version of the criminal network?*

Solution:

If we were to study the directed version of the graph, we would be able to learn the direction that information flows in the network. That is, we would be able to identify who is given instruction and who is receiving instructions. This would allow us to better understand the role of the “players” in this criminal network. In this sense, nodes with high left-eigenvector centrality would indicate important nodes that receive (and probably execute) instructions from many other (important) nodes in the networks whereas nodes with high right-eigenvector centrality would indicate important nodes that give instructions to many other (important) nodes in the network.

Part (j): *(4 points) Recall the definition of hubs and authorities. Compute the hub and authority score of each actor, and for each phase. (Remember to load the adjacency data again this time using create_using = nx.DiGraph().)?*

With networkx you can use the nx.algorithms.link_analysis.hits function, set max_iter=1000000 for best results.

Using this, what relevant observations can you make on how the relationship between n1 and n3 evolves over the phases. Can you make comparisons to your results in Part (g)?

Solution:

In the context of a social network, the hub score indicates the degree at which an individual of the network is highly connected to others (have a wide reach), whereas the authority score indicates the degree at which an individual of the network is a trusted source of information.

By calculating the hub and authority scores for all nodes in the network in all phases, we notice that Serero (n1) decreases its hub score (in average) at the same time as it increases its authority score (in average). This would indicate that as the phases develop, Serero have less contact with lower-level members in the hierarchy (terminal nodes of the network), keeping only communication with key members (key nodes). However, as instructions to many members is the organization indirectly come from him, his authority increases.

In contrast, Pierre Perlini (n3), principal lieutenant of Serero, slowly increases its hub score (in average) suggesting that he becomes more exposed (more connected) to the lower-level members of the network, which makes sense since he is responsible to execute Serero's instructions. On the other hand, the authority score of Perlini seems to slightly decrease (in average) as the phases develop, which highlights his intermediary role in the network between Serero and lower-level players.

These observations are aligned with what was described in part (g), which suggested that the criminal network became more sophisticated as the phases developed, with a better structured hierarchy where key members on the top of the hierarchy have less contact with traffickers (mostly terminal nodes), reducing their exposure, but still remaining as the source of information (instructions).

- **4. Open Project - Network chosen: Twitter**

(12 points) To what extent does the power-law distribution hold in the Twitter network? How does degree distribution, clustering, and centrality metrics can be used to select a candidate model among the four network models discussed (Erdos-Renyi, configuration, preferential attachment, and small-world)?

Solution:

The following project was conducted using the Twitter network available on [3]. Although the reference mentions that the Twitter graph is directed, this project focused on the analysis of the undirected version of the graph. The reason for this choice is because a similar analysis for the suitability of the power-law distribution in the Twitter network is available in [4], which can be used to validate the findings in this project.

In order to accomplish the goal of this project, let us first plot the node degree distribution of the undirected Twitter graph (**Figure 1**). Looking at the plot, it is possible to note that the distribution is right skewed, which is a key characteristic of a power law distribution.

Furthermore, if we calculate the ratio between the maximum degree size and the minimum degree size, we will find a ratio of 3383, which is considerably larger if compared to a centered distribution like a normal distribution (ration would be 1). Additionally, is not difficult to show that if any scaling factor is used to scale up or down the distribution, its shape would not change. For one thing, the ration between the maximum degree size and the minimum degree size would still be the same. With these observations, it makes sense to further investigate if the distribution of the degree size of the Twitter graph follows a power-law distribution.

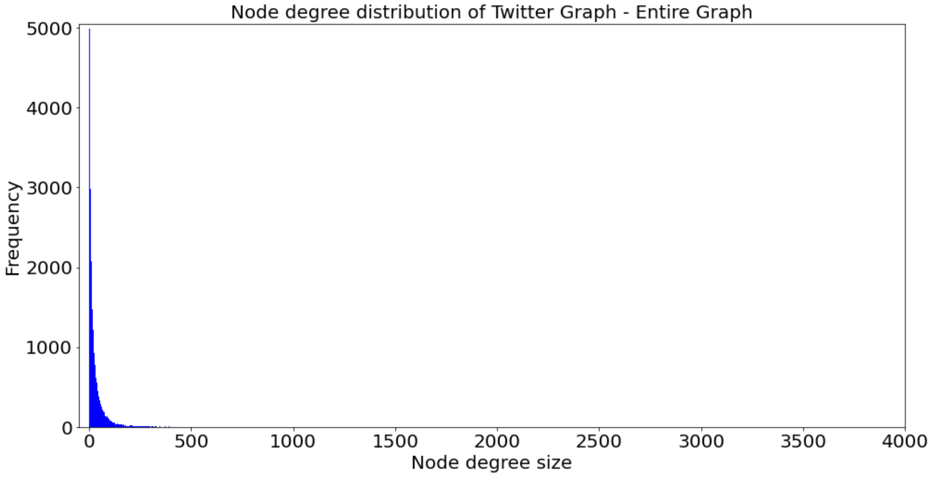


Figure 1 – Node degree distribution of the Twitter graph using all nodes and edges (entire graph).

The discrete probability of a power-law distribution, indexed by the degree value k can be expressed as follow:

$$p(k) = \frac{(\alpha - 1)}{k_{min}} \left(\frac{k}{k_{min}}\right)^{-\alpha}$$

Where,

- α is the power in the power law distribution;
- k_{min} is the minimum degree for which the discrete probability law applies;

If we take the logarithm on both sides of the equation, we have:

$$\log(p(k)) = log\left(\frac{(\alpha - 1)}{k_{min}}\right) - \alpha * log\left(\frac{k}{k_{min}}\right)$$

If we adjust a linear equation expression to the above equation, we have:

$$y = b + a * x$$

Where,

- $y = \log(p(k))$;
- $b = log\left(\frac{(\alpha-1)}{k_{min}}\right)$;
- $a = -\alpha$;
- $x = log\left(\frac{k}{k_{min}}\right)$;

Therefore, if we plotted the degree distribution from **Figure 1** in a log – log graph and the distribution followed a power-law distribution, then the curve in the log – log figure should approximate a straight line. For this purpose, **Figure 2** shows the log – log plot of the degree distribution of the Twitter network and the regression line that best fit the datapoints ($k_{min} = 1$ and alpha = 1.83).

It is not difficult to note that the curve in **Figure 2** doesn't approximate a straight line, which suggest that the degree distribution of the nodes in the Twitter graph does not follow a power-law distribution (at least not for $k_{min} = 1$). However, it is arguable that the tail of the distribution would follow a power-law distribution.

In this sense, if k_{min} was set to 25 and another regression curve was fit to the data (**Figure 3**), we would notice a much better fit. This suggests that the tail degree distribution could indeed be described by a power-law distribution (alpha = 1.96). A similar behavior was highlighted in for [4] and, according to the reference, one of the reasons would be Twitter's police that would interfere on how the graph (connections) grows.

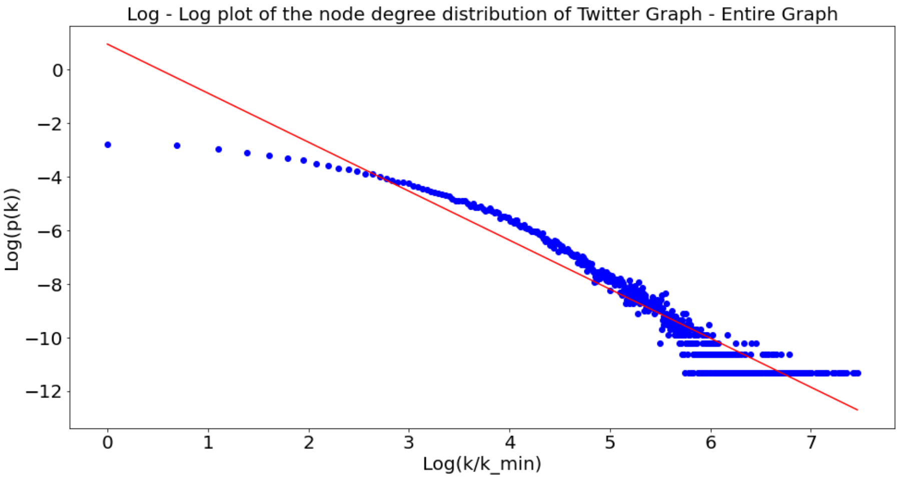


Figure 2 – Log – log plot of the node degree distribution of the Twitter graph (entire graph) and regression line (red) created to approximate the power-law distribution behavior considering $k_{min} = 1$.

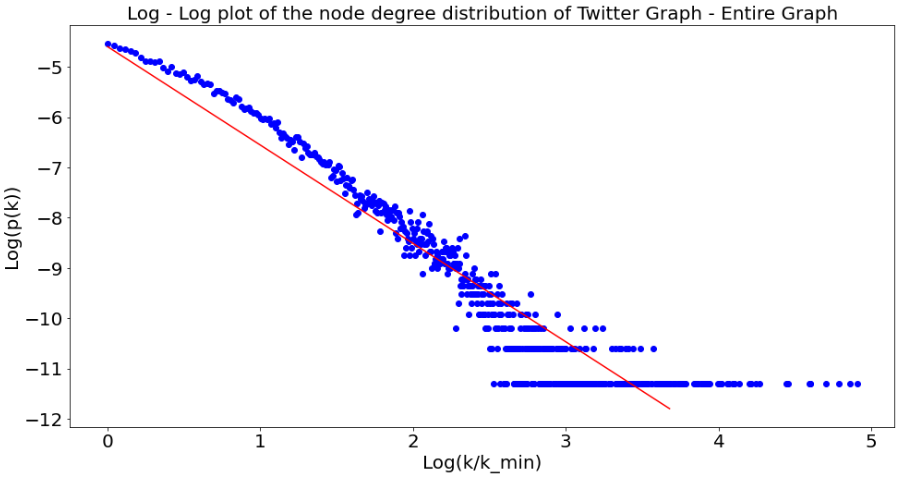


Figure 3 – Log – log plot of the node degree distribution of the Twitter graph (entire graph) and regression line (red) created to approximate the power-law distribution behavior considering $k_{min} = 25$.

If we now compare the node degree distribution of the Twitter network with the node degree distribution of other models, namely Erdos-Renyi, Preferential Attachment and Small World, we note that the node distribution of the Twitter network is very atypical and does not approximate any of the mentioned distribution (see **Figure 4**). Therefore, none of the network models seen in the course is suitable to model the Twitter network.

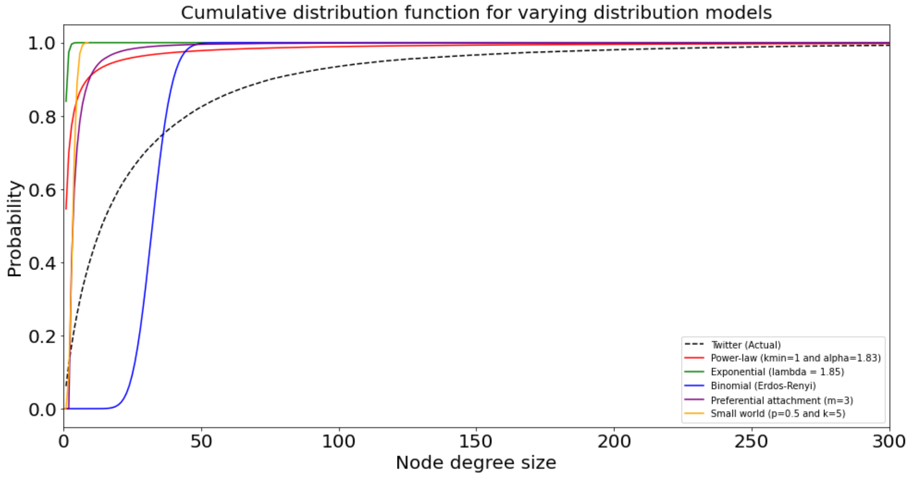


Figure 4 – Cumulative distribution function plot for varying network models.

Figure 4 was created with the aid of the NetworkX library, which has functions that allow the simulation of Preferential Attachment and Small World network models. Erdos-Renyi and Exponential network models were simulated using the corresponding distribution functions.

As mentioned, **Figure 4** also shows the cumulative distribution curve of the power-law distribution and exponential distribution. We note that the node degree distribution of the preferential attachment model is the distribution that best approximates the power-law distribution as expected. Additionally, we also note that the node degree distribution of the small world model does not approximate a power-law distribution as also expected from theory.

Reference

[1] Linear Algebra, NumPy, https://numpy.org/doc/stable/reference/routines.linalg.html#module-numpy.linalg, visited on 23 Oct 2023.

[2] William, P, Flannery, B, Teukolsky, S, Vetterling, W, 2007, *Numerical Recipes: The Art of Scientific Computing*, 3rd Edition, Cambridge University Press. p. 108. ISBN 978-0-521-88068-8.

[3] Leskovec, J, Social Circles: Twitter, Stanford University, https://snap.stanford.edu/data/ego-Twitter.html, visited on 29 Oct 2023.

[4] Trolliet, T, Giroire, F, Pérennes, S, *A Random Growth Model with any Real or Theoretical Degree Distribution*, https://arxiv.org/pdf/2008.03831.pdf, visited on 29 Oct 2023.