Written Report – 6.419x Module 3

Name: ajland

Problem 1

Part (c). (2 points) How does the time complexity of your solution involving matrix multiplication in part (a) compare to your friend's algorithm? (Maximum 100 words)

Solution: Generally, for an (n x k) X (k x m) matrix multiplication, the cost is $2nmk$ [1], where the multiplication involves three nested for-loops. Therefore, if $A \in \mathbb{R}^{nxn}$, A^TA and AA^T is of order $\mathcal{O}(n^3)$. The two approaches are asymptotically similar, but matrix multiplication is readily parallelized with modern libraries.

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? (Maximum 200 words)

Solution: Co-citation counts how many papers cited both papers in question while bibliographic coupling counts how many citations two papers have in common. In other words, a co-citation network counts the in-degree from common source nodes while bibliographic coupling counts the out-degree from source nodes to common terminal nodes. In this way, the co-citation count for two papers can only grow from papers published afterwards while the bibliographic coupling can be non-zero as soon as a couple is formed. Furthermore, the co-citation network is potentially changing as time continues but the bibliographic coupling will be static. It seems to me that bibliographic coupling is a better measure for similarity between papers because it is static and it also does not depend on the seminal nature of the works.

Problem 2

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? (Maximum 200 words)

Solution: As stated, the number of nodes in Figure 1 grows quickly until phase 3, then levels out. The data collected to build the network came from 11 wiretaps that lasted two months each; thus, the first three phases correspond to the first 6 months

of the investigation and the building of the network, which explains the growth in these phases. There was possibly a delay in the growth of the network after phase 3 because the first police seizure happened in phase 4. Indeed, it was during phase 4 that the number of edges dropped, meaning that relationships were dropped or there was an in and out flux of people. Neither of these affect my conclusions about which nodes had the highest mean centrality measures through the 11 phases.

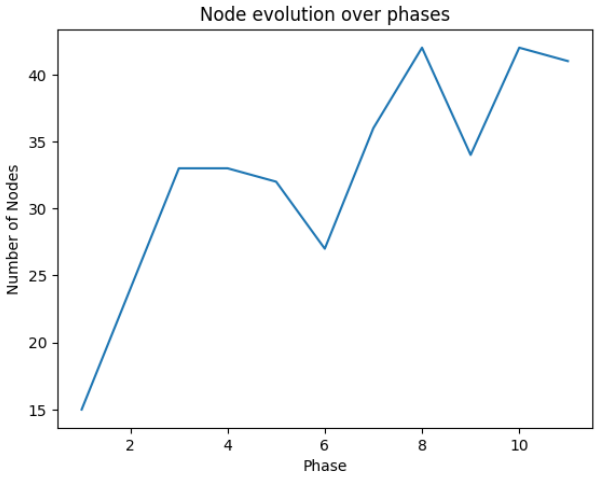

Figure 1: Evolution of number of nodes through the 11 phases

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. (Maximum 400 words)

Solution: The goal of a centrality measure is to, "capture the importance of a node's position in the network" according to the lecture notes. The degree centrality yields the degree of each node, meaning that nodes with higher degrees have more connections. In the context of criminal networks, the highest degree centrality nodes will have the most connections to other criminals and therefore likely more influence and access to information. Degree centrality gives no indication, however, of which nodes high degree nodes are connected to (i.e. important or non-important nodes). This means a high degree centrality node may not be as important as it seems. Eigenvector centrality answers this drawback by ranking

people on whether they are connected to many nodes and/or nodes of high importance; in this way, high eigenvector centrality individuals know powerful and/or connected people. Betweenness centrality measures, "the extent to which a node lies on [shortest] paths between other nodes" [2]. In the criminal network, high betweenness nodes are nodes that are most required to traverse across to reach any two other nodes. In this way, information will likely require traversing across these people which gives them influence and removing these people would most break the network apart. The centrality measure that is most relevant is one that will identify Daniel Serero as the mastermind of the network. In this situation, I will say that the degree centrality is best as it assigns the highest proportion of mean centrality to this node. Specifically, n1 had 44.7% of the mean betweenness centrality mass, 12.7% of the mean eigenvector centrality mass, and 21.4% of the mean degree centrality mass through all 11 stages.

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. (Maximum 200 words)

Solution: I would define important nodes to be ones that, if removed, would most break up the network since this would cripple the trafficking network, potentially making it unworkable. This would correspond to removing nodes with high betweenness centrality. Using this metric and the results from Part (b) Question 5, the top three players are n1, n12 and n3 based on mean betweenness through time. Every other player could be considered peripheral depending on the betweenness threshold; some even have zero betweenness.

Part (f) Question 2. (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. (Maximum 300 words)

Solution: The major event that took place in Phase 4 during the actual investigation was the first drug seizure by the police. Visually, the graph becomes less clustered as several edges are removed. Indeed, the average clustering coefficient decreases from 0.32 to 0.25 and, as stated in a previous answer, the number of edges decreases. Phase 4 betweenness restructures from [n1, n89, n3] to [n1, n12, n31] top three nodes, while both degree and eigenvector centrality restructure from

[n1, n3, n83] to [n1, n12, n3] top nodes. This means that the seizure likely resulted in a re-ranking of the nodes, resulting in different nodes/people being more or less important. Visually, the seizure appeared to lessen the number of relations between people, perhaps because people decided to lay low over the following two month phase or for other reasons.

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? (Maximum 300 words)

Solution: The evolution of the network is visualized in Figure 2.

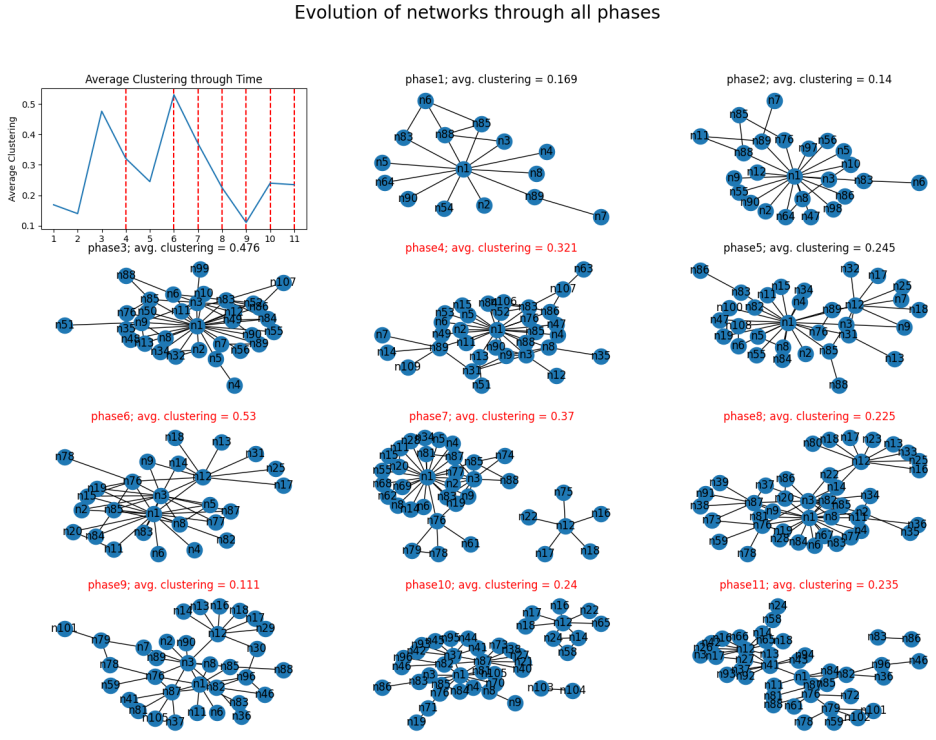

Figure 2: Network evolution through all 11 phases. Phases with seizures have a red title. Clustering coefficients are given for each phase and shown in the first plot with red dashed vertical lines indicating seizure phases.

Interestingly, Phases 7 and 10 see the formation of isolated clusters after cer-

tain connections are broken; for example, the connection between n3 $\rightarrow$ n12 (two key nodes) breaks from 9 to 10 and n30 drops out, leaving the n12 centric cluster isolated.

Furthermore, the average clustering tends to increase after non-seizure phases and decrease after seizure phases. For example, there are no halts in seizures starting in phase 6 until the end, and from 6 to 9 we see a continuous decrease in average clustering, 10 to 11 sees a minor decrease while 9 to 10 sees a decent up-tick. Perhaps the latter two phases were a result of the network gaining resilience under the prolonged stress. It is also possible that 10 saw a large uptick as there was a large amount of marijuana being imported ($18.7 Million worth), which required a decent amount of coordination among 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. (Maximum 100 words)

Solution: Taking the mean betwennness score and sorting it in descending order, the first node not in the 23 listed is n41 with a score of 0.0504. Reviewing Figure 2, we see that n41 actually occupies a critical position in phase 11 in that it shares the sole edge with n1 connecting the two largest clusters. If "importance" is defined to be those nodes that if removed would most break up the network, then n41 is certainly in that list. We see n14 and n22 playing a similar but less critical role in phase 8.

Part (i). (2 points) What are the advantages of looking at the directed version vs. undirected version of the criminal network? (Maximum 250 words)

Solution: In a directed graph, we could examine more directly the flow of information, whether the information is flowing from or to given nodes, via the in and out degrees. This could be used the define importance since one could, say, remove highly informed nodes or remove highly informative nodes. Similarly, the left eigenvector centrality (LEC) will have nodes derive their importance from other important nodes pointing to it while right eigenvector centrality (REC) is determined by importance of nodes it points to. Using the LEC may reveal nodes that leaders direct (i.e. lieutenants) while REC could reveal nodes that report to leaders (i.e. informants). In summary, a directed network could allow for better hierarchical analysis of the criminal group.

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)? Optional: Also comment on what the hub and authority score can tell you about the actors you identified in Part (e) (Maximum 400 words)

Solution: Using the stated function I computed the hub and authority score for each node through all 11 phases then plotted the time histories for n1 and n3; the result is Figure 3.

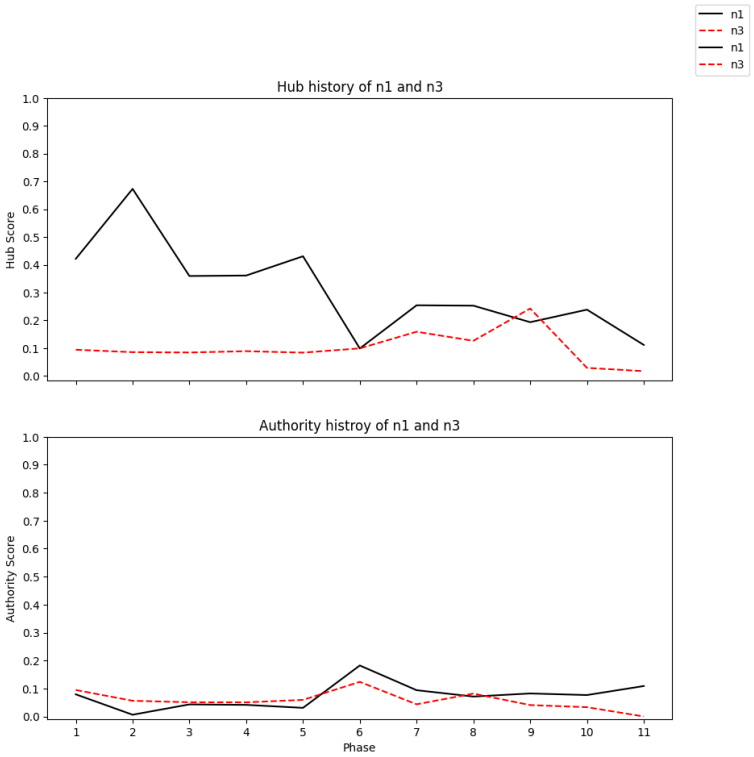

Figure 3: Time histories of hub and authority score for n1 and n2

Both n1 and n3 are relatively similar in their authority score throughout. For the Hub score history, n1 starts high, but becomes similar to n3 by phase 6 and beyond, though it is still typically larger. Other than this, n1 and n3 not appear to

correlate. The results do not necessarily compare to my results in part (g), but it can be observed that n1 does become proportionately less of a hub as more nodes are added.

Project

Question: Using the Facebook ego network [3], are low degree nodes and high degree nodes assortative (i.e. low friend count people are friends with people who also have low friend counts and vice-versa) and how do they compare to one another?

Methodology:

1. Identify which nodes are in the first and last quartiles according to number of degrees and put them in their own lists.
2. Extract two separate edge lists using the nodes in the low/high degree node lists.
3. Construct two separate graphs using the previous node lists. These graphs consist of all nodes 0 and 1 hops from originals.
4. Identify all nodes that are 2 hops from the originals and add these to their respective graphs.
5. Analyze the degree distribution of all 1-hop nodes for each graph (these are the nodes that the originals are connected to) to answer questions about degree assortativity.
6. Visualize assortativity accross degree levels in one plot

A graph of the entire Facebook egonet is shown in Figure 4.

Following step one, the first quartile of node degrees range from 1-11 while the last quartile of nodes have degrees ranging from 57-1045. The degree assortativity is a number in the range [-1, 1] and describes the tendency of nodes to share edges with similar degree nodes. By following steps 1-4 we can construct the the graphs consisting of all nodes 2 hops or less from the original. The motivation for this is that we want to see what kind of nodes are connected to the originals; therefore, we need to the graph to include nodes up to 2 hops away. These graphs are shown in Figure 5

The assortativity computation was carried out using a networkx function on the entire graph while specifying the subset of low/high nodes; for example,

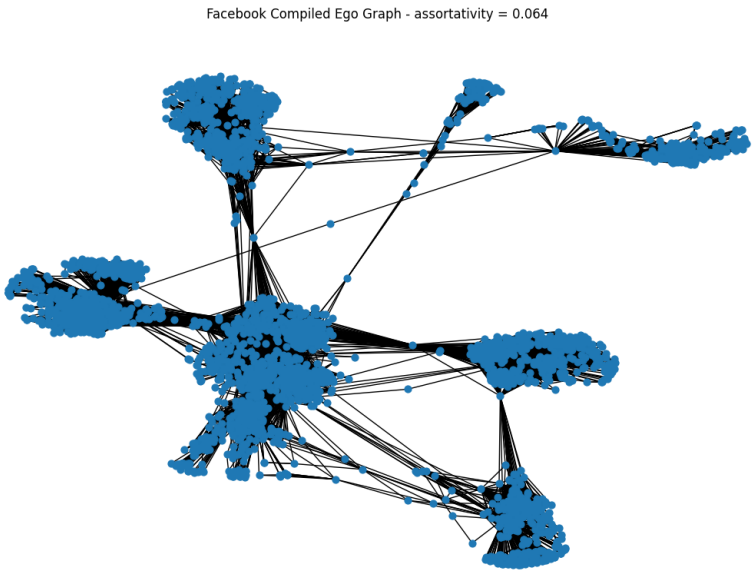

Figure 4: Visualization of entire Facebook egonet with nearly neutral assortativity (4039 nodes and 88234 edges)

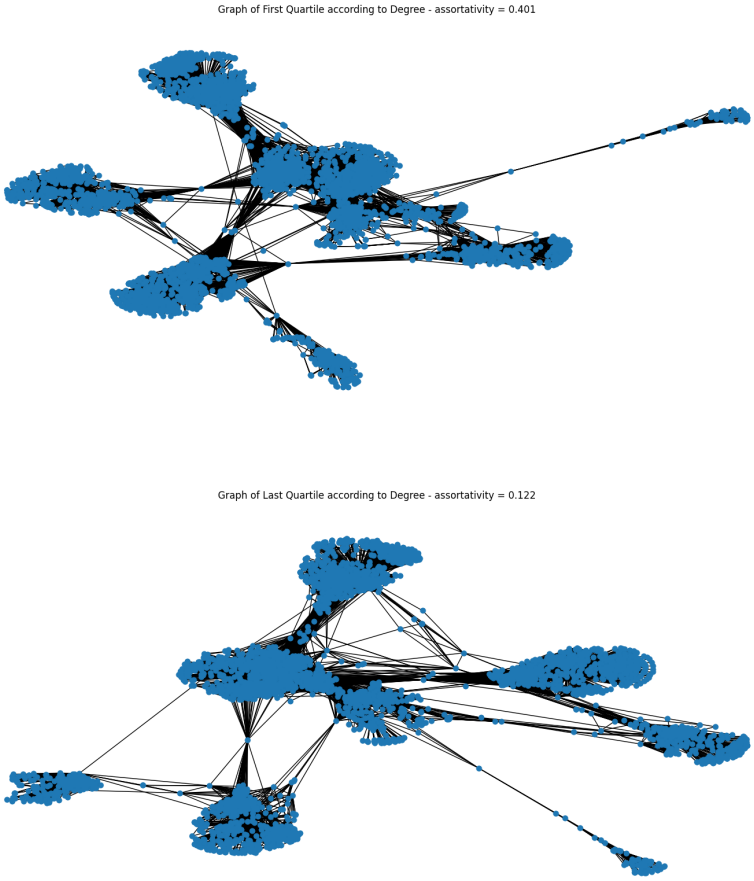

Figure 5: Top graph includes low degree nodes and all nodes up to 2 hops away (4039 nodes and 44572 edges). Bottom graph is similar but with high degree nodes (4039 nodes and 87895 edges)

```
print('Low nodes assortativity: ',
nx.degree_pearson_correlation_coefficient(G,
                                          nodes=low_nodes))
```

It is worth noting now that the assortativity coefficient for node degree is the pearson correlation coefficient between pairs of linked nodes. In this way, a positive number indicates a correlation between nodes of same degree and negative number indicates a correlation between nodes of different degrees. From these results, the sub-graphs do have a positive correlation and the low-node subgraph has an even larger positive correlation, though it may be hard to tell if these are significant. One may be able to formulate a statistical test to help answer this question, but the problem is that the correlation is computed over a range of different levels of degrees. That is, since quartile one, for example, covers nodes of degrees 1-11, one cannot say that the degree of the nodes they are connected to are iid; though it is possible that, conditioned on degree, the degree of nodes they are connected to are iid. Let's examine the degree distribution of all 1-hop nodes to get another look into the assortativity.

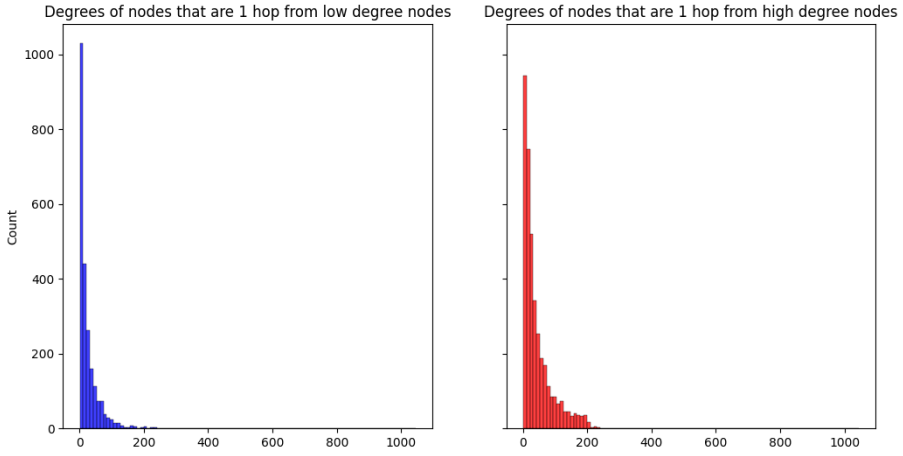

Figure 6: Degree distribution performed only on the nodes that are 1-hop away from the nodes of interest. Left count: 2319; Right count = 3895

It can be seen from these plots that both node subsets follow something like a power law and that the low degree 1-hop nodes decay quicker. Since there are many more lower degree nodes than higher degree nodes, it might make sense that the low-degree nodes exhibit higher assortativity than high-degree nodes. In other words, connections with low-degree nodes are more likely because there are more of them, regardless of initial degree. In fact, it can be seen from Figure 7 that nodes

of all degrees tend to connect to nodes of 200 degrees or less and that from 1-200 degrees there appears to be a definite positive correlation/assortative mixing.

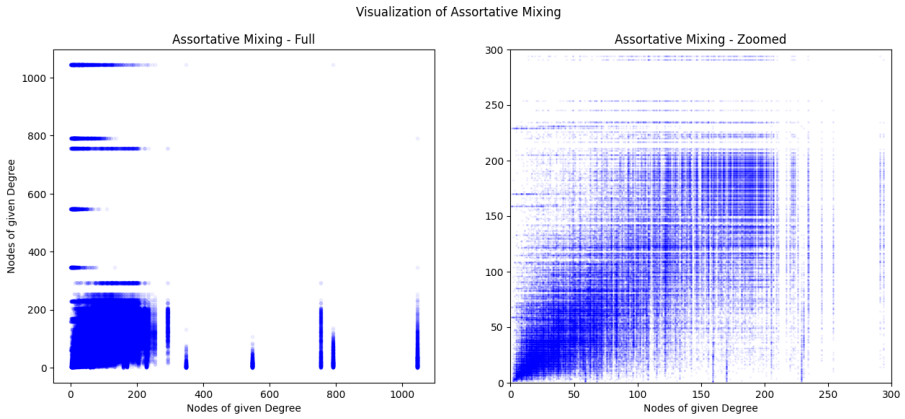

Figure 7: Data points represent connections between nodes of different degrees. Densely populated regions indicate frequent connections. Left: The full xlim and ylim ranges; Right: Zoomed view of first graph to xlim and ylim of 300 degrees.

Looking more closely at the right image of Figure 7, the densely populated region appears to fan out in a cone pattern until it drops off/vanishes around 200 degrees. This means that the assortative mixing starts high for very low degree nodes then decreases until nodes of 200 degrees until it suddenly goes away. In fact, this would explain why the first quartile had a higher corrleation than the final quartile in Figure 5. The plot also verifies the idea that the degree distribution of the nodes attached to the originator is conditional on the original's degree; in general, the distribution shifts rightward and upward as the degree increases and appears to become more dispersed.

To answer the original question in plain English, both low and high friend count people exhibit some tendency to be friends with other low and high friend count individuals, respectively. Lower friend count people tend to show this assortative property more than high friend count people at an increasing rate from people with 1 friend up to people with 200 friends, but after 200 friends there is no longer any assortativity, meaning that ultra high friend count people are friends with almost entirely people who have 200 friends or less. The best visualization for these results is captured in Figure 7, which can be reproduced using the following code [4]:

```
xdata = []
ydata = []
```

```
for i, j in G.edges():
    xdata.append(G.degree(i)); ydata.append(G.degree[j])
    xdata.append(G.degree(j)); ydata.append(G.degree[i])
fig_assort, ax_assort = plt.subplots(1, 2, figsize=(15,6))
ax_assort[0].plot(xdata, ydata, 'bo', alpha = 0.05, ms=3)
ax_assort[1].plot(xdata, ydata, 'bo', alpha = 0.05, ms=1)
...
```

References

[1] M. E. M. et. al., "Linear algebra: Foundations to frontiers - notes to laff with," http://www.ulaff.net/downloads.html], note = [Online; accessed 10-Jul-2024], p. 153, 2020.

[2] C. Uhler, "Criminal networks module," http://www.ulaff.net/downloads.html] ,pp.2,5,Online; accessed 11-Jul-2024].

[3] J. McAuley and J. Leskovec, "Learning to discover social circles in ego networks," *NIPS*, 2012.

[4] H. Sayama, "17.6: Assortativity," https://math.libretexts.org/Bookshelves/Scientific_Computing_Simulations_and_Modeling/Introduction_to_the_Modeling_and_Analysis_of_Complex_Systems_(Sayama)/17%3A_Dynamical_Networks_II__Analysis_of_Network_Topologies/17.06%3A_Assortativity#:~:text=Assortativity%20(positive%20assortativity)%20The%20tendency,dissimilar%20properties%20within%20a%20network., [Online; accessed 11-Jul-2024].