Problem Set 2

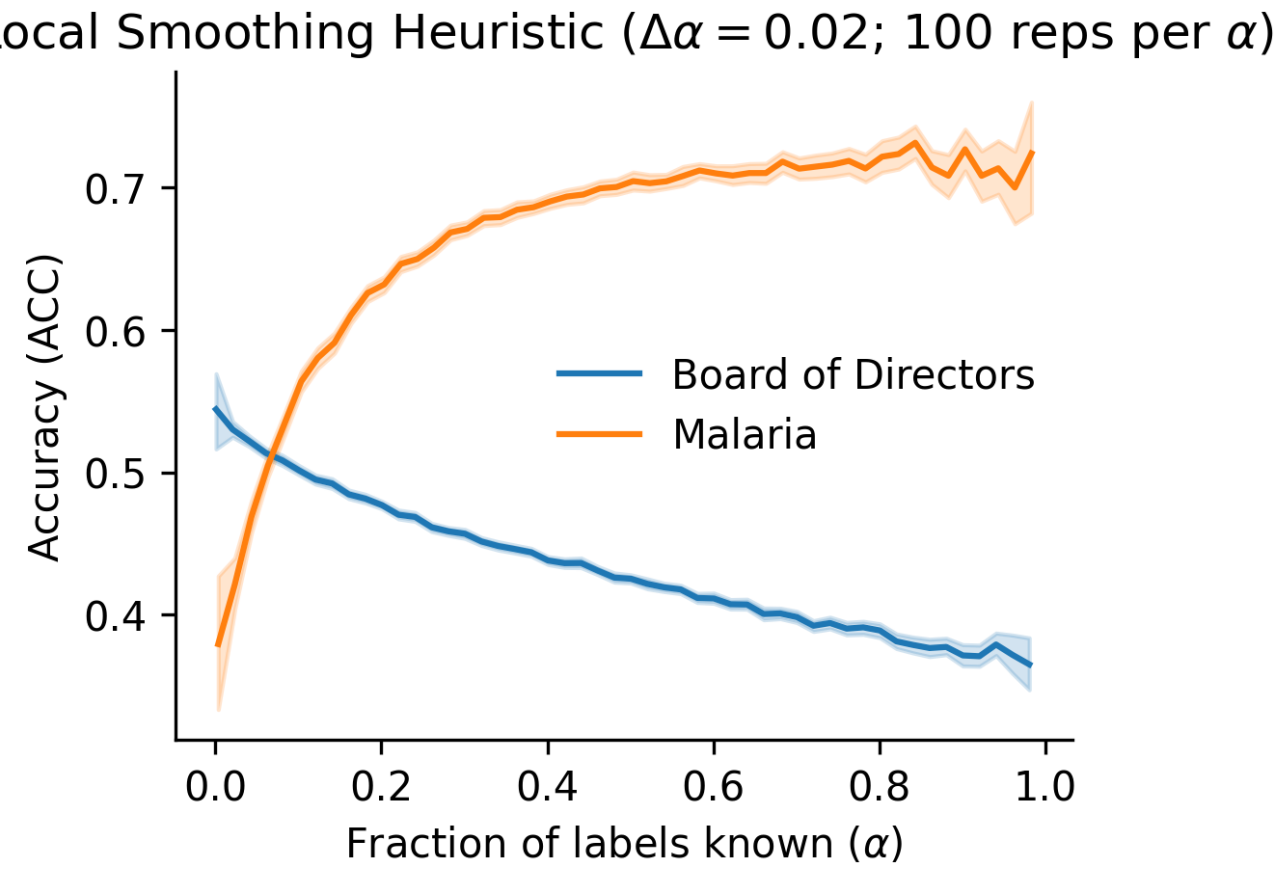
CSCI 5352

Ben Braun

2.28.2025

Problem 1

Part A



The accuracy curves for these two networks are very different. The board of directors network starts off at an accuracy barely better than chance (0.5) and steadily declines as more neighbor labels are added until leveling off at ~0.38 by the time all neighbor labels are added. Since we're predicting a simple binary attribute here, we would expect this performance to be much better if the social network exhibited assortative mixing. Since the accuracy decreases, I would guess that the network is actually disassortative, with edges between different genders being more common than edges between the same genders. It could also be due to the difference in frequency between males and females.

There are 908 males but only 513 females, so the predictions will generally be biased toward male, and this effect will scale linearly with the number of neighbors. At high values of α, the prediction accuracy is roughly proportional to the class imbalance.

In the Malaria network, there is a pattern of initial increase in accuracy as information is added which levels off when about half of the labels are known. The local smoothing heuristic can only take us so far in this case, as the highest accuracy achieved was only about 0.7 despite most nodes being known. There is class imbalance in both of these networks that make it difficult to get high accuracy, since local smoothing is often biased toward larger classes and makes assumptions that don't seem true in either of these networks. Mainly, it assumes a large amount of assortativity, which seems more true in the Malaria network than the board of directors, but is not a strong property of either network.

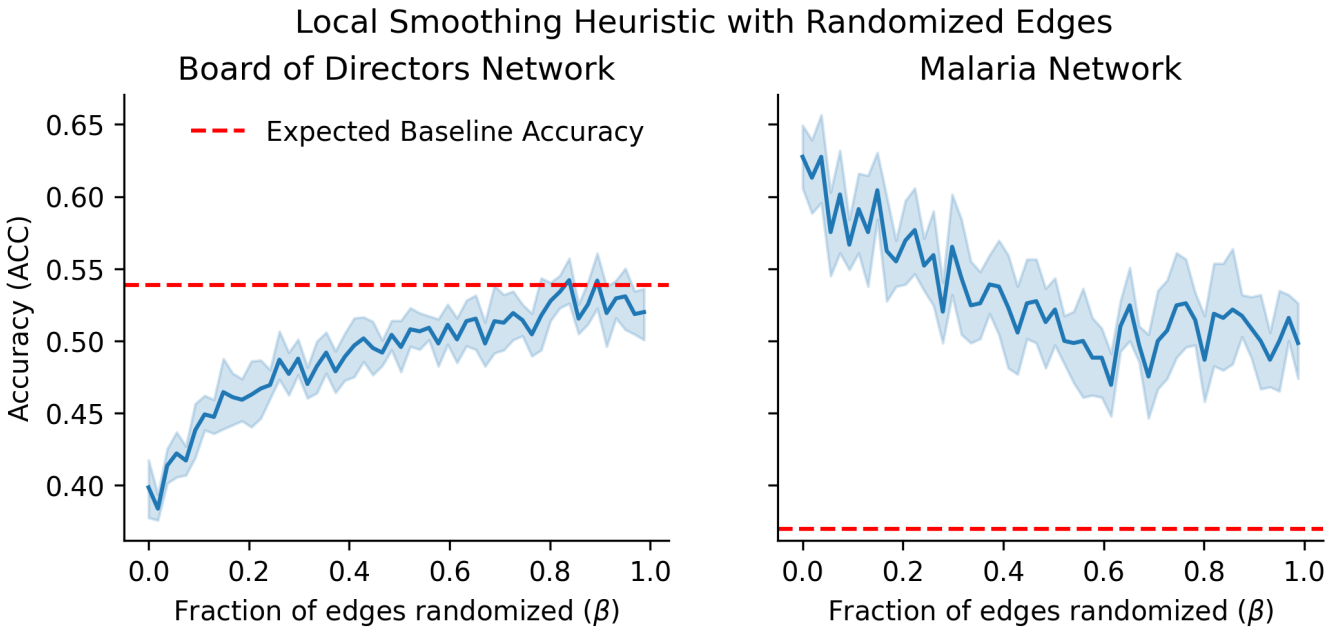
The probability of guessing the true value of the missing node is proportional to the number of nodes in the known distribution with that value.

If we have k unique labels and label x_i appears n_i times in the network, then the probability of guessing x_i is $p_i = \frac{n_i}{N}$, where N is the total number of nodes in the network. The probability that the guess is correct is also $p_i = \frac{n_i}{N}$, so we multiply those events to find the probability of guessing a missing label correctly. The expected accuracy is the sum of these probabilities over all labels:

$$\mathbb{E}[\mathrm{ACC}] = \sum_{i=1}^{k} p_i^2 = \sum_{i=1}^{k} \frac{n_i^2}{N^2} = \frac{1}{N^2} \sum_{i=1}^{k} n_i^2$$

The baseline expected accuracy for the board of directors network is ~0.54, and for the malaria network it is ~0.37. These are roughly equal to the accuracy of the local smoothing heuristic when α is small. As α increases, the accuracy begins to deviate from this baseline since more nodes have neighbors with known labels.

Part B

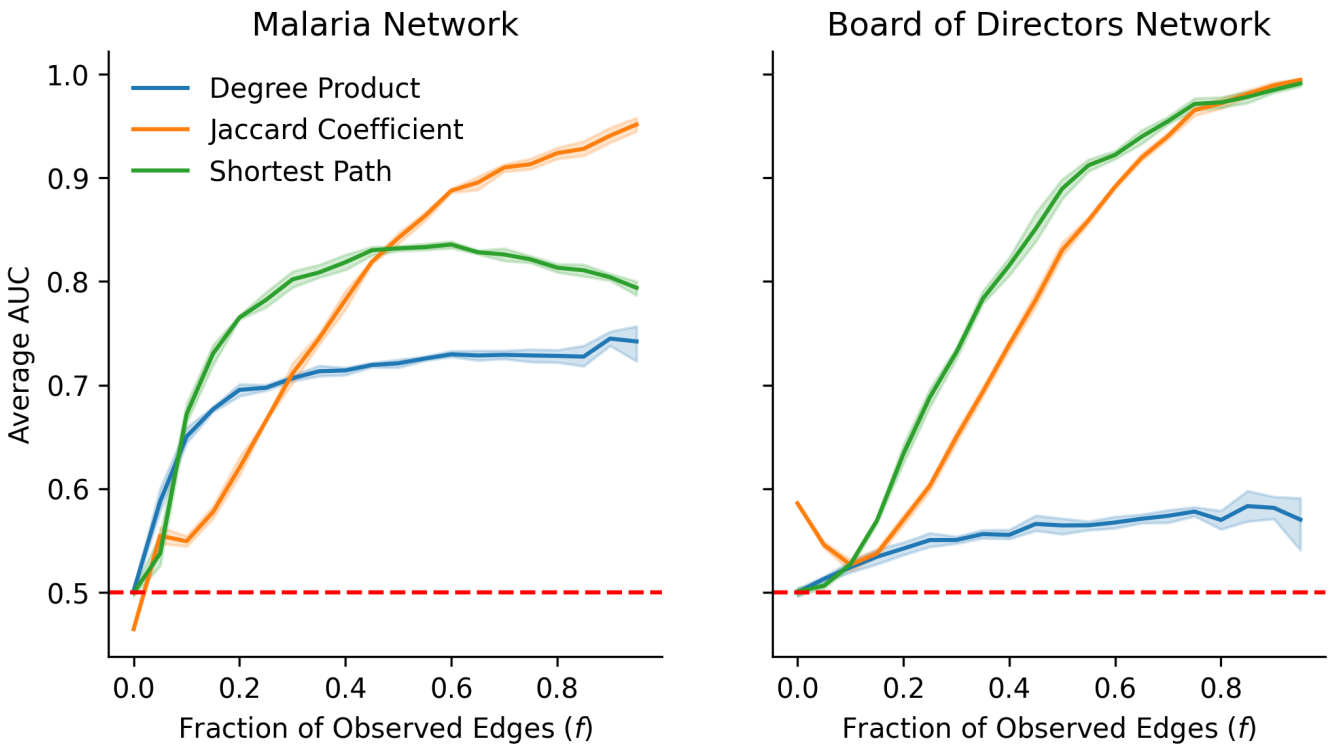


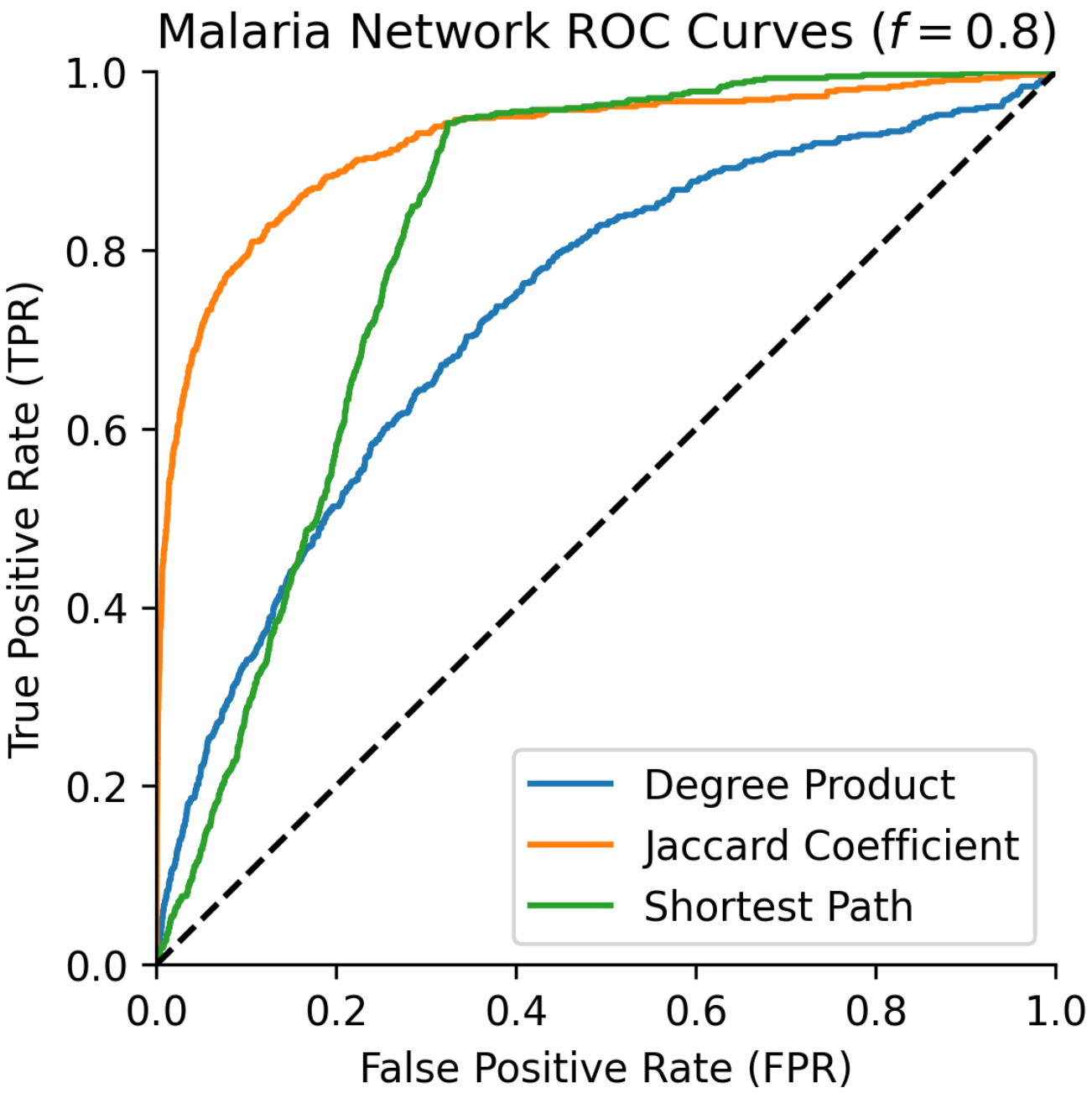
We find that in the board of directors graph, the accuracy is actually higher at baseline than with the local smoothing heuristic since it's a disassortative graph, so we see an increase in accuracy toward the baseline. For this graph, the hypothesis holds.

However, in the case of the malaria graph, we never quite reach baseline accuracy. Instead, the degradation of accuracy stops around 0.5 and stays there. I suspect that this may be due to the degree structure of the graph. If there is a tendency for high-degree or low-degree nodes to be in certain categories, this could skew the accuracy of the local smoothing heuristic. For instance, if the most common category also has the highest mean degree, then the local smoothing heuristic will be biased toward that category, resulting in higher-than-baseline accuracy despite graph randomization (since double-edge swaps preserve node degree). This illustrates the diminished usefulness of ACC when label frequencies are imbalanced.

Problem 2

Part A





In the Malaria network, the degree product takes an early lead in AUC at low f values. However, it significantly underperforms as f gets large. This is likely because the degree product is relatively effective at "easy" predictions, where the most high-degree nodes are likely connected to high-degree nodes as well. It becomes much less effective as the degree products go down since the nodes become lower-degree and the statistical trend the predictor relies on diminishes.

Since the Malaria network has the best performance with the Jaccard coefficient, that network likely forms tight clusters/communities. The board of directors network similarly has good performance with Jaccard, but also with the shortest path algorithm.

Overall, the ROC curves show that the Jaccard coefficient is an excellent predictor of node labels especially at high values. The degree product is moderately good at predicting labels but frequently makes incorrect predictions even at values that should provide confident predictions. Finally, the shortest path algorithm performs decently well for shorter paths until it plateaus at very short paths, where the information provided by local connectivity becomes more misleading than useful.