Homework 2

Jackson Paull
University of Texas at Austin
jackson.paull@utexas.edu

November 2023

Problem 1

For my mitigation strategy, I chose to label all nodes as either ”essential” or ”non-essential” with probability $p = 0.25$, which is expected to mirror real world data since roughly 25% of the population was essential during the Covid-19 pandemic. All non-essential persons then isolated from all but 10 of their neighbors, chosen randomly. Since all nodes isolate independently, in practice it is expected for non-essential nodes to have a degree between 4-10.

My simulated results are shown in Figure 1.

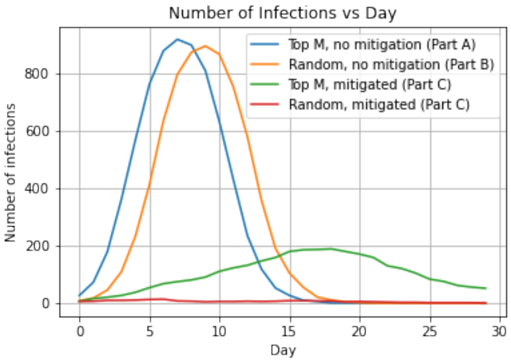


Figure 1: Simulation Results

Problem 2

My assigned Ego-Net was the Ego-Net centered around node 'U546'. Global network analysis metrics form Table 1 shows a small network with some scale free properties. The largest supporting evidence for this is the degree distribution, which follows a power law distribution.

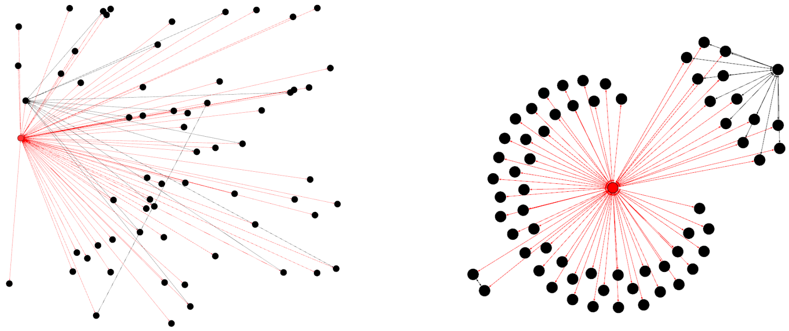
Circle of Truth Analysis

There are only two circle of truths with a non-trivial amount of nodes, those being circle 6 and circle 8. These circles do share similar characteristics, as the entropy of their features is low compared to the log of the number of nodes. For reference, circle 6 contains 28 nodes, and therefore an expected entropy of roughly 5 bits for a population with no homophily. Most of the entries in Table 2 show entropy in the range of 2-3 bits, which is significantly lower than 5, showing significant evidence of homophily.

There is a lot of overlap between circle 6 and circle 8, with circle 8 almost being a subset of circle 6. Beyond this specific pair, there is no overlap between circles with the sole exception that the root node, U546, is the only overlap between several circle pairs.

Num Nodes	Num Edges	Avg Node Degree	Avg Clustering Coeff	Avg Path Length	Num Communities
64	154	2.4	0.24	1.91	3

Table 1: Global Network Analysis



(a) Random Initialization

(b) Force Atlas Visualization

Figure 2: Ego-Net for U-546

	clustering coefficient	Homo College	Homo Employer	Homo Location
circle 1	0.0	0.69	1.38	0.69
circle 2	0.0	0.69	2.08	0.69
circle 3	0.14	0.69	2.78	1.91
circle 4	0.0	0.69	2.12	0.95
circle 5	NA	NA	NA	NA
circle 6	0.0	2.51	3.57	2.38
circle 7	NA	NA	NA	NA
circle 8	0.84	2.04	2.96	2.16

Table 2: Homophily Analysis

Problem 3

We see modularity peak around 20 communities, and the addition of more communities serves only to slightly decrease modularity. When preforming modularity based detection with resolution=0.1, we observe more communities, but some of them seem arbitrary. Notably, there is a set of densely connected nodes near the bottom right of the graph which is grouped into two separate communties with a resolution of 0.1, and a single community with a resolution of 0.7. Intuitively, the fewer communities seem to be a better representation of the network. This aligns with our experimental modularity results, as the fewer communities provide greater modularity.

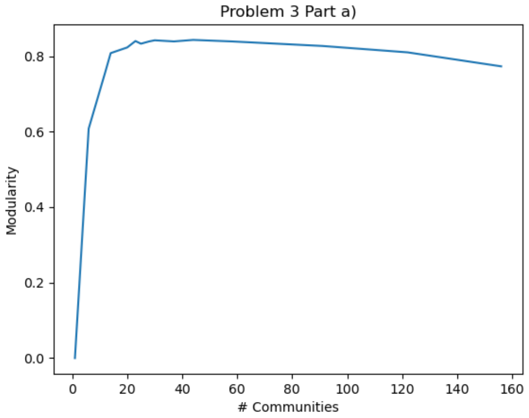
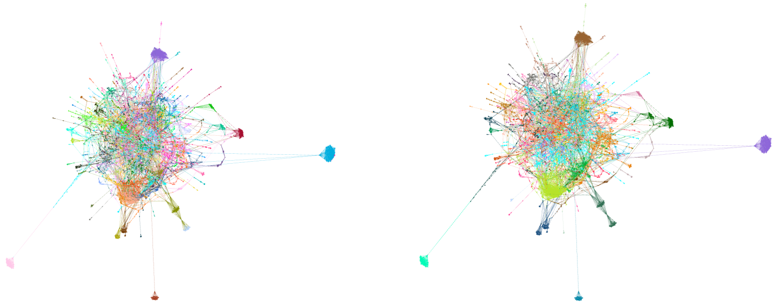


Figure 3: Modularity vs Number of Communities



(a) 0.1 Resolution

(b) 0.7 Resolution

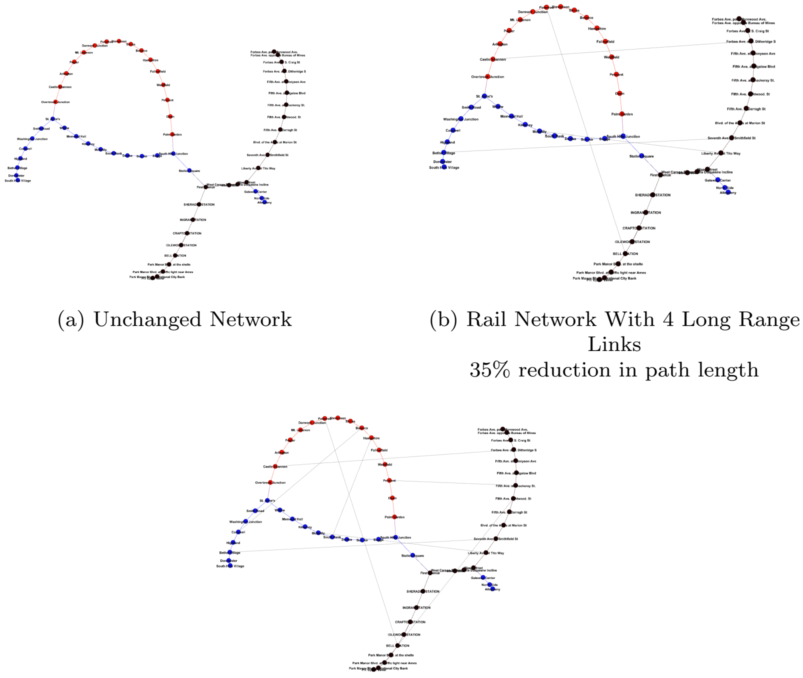
Figure 4: ArXiv Network

Problem 4

The rail network does not exhibit much of a small-world effect due to its incredibly low clustering coefficient evidenced in table 3. By inserting just 4 long range links, I reduce the average path length by 35% to 7.59. By inserting 4 more long range links, I further reduce the average path length to 6.02, which is a 49% reduction from the original. By introducing these lengths, the small-world effect is amplified as the reduction in path length greatly outweighs a reduction in clustering coefficient. Intuitively this makes sense as these long range links greatly contribute to the traversability of the network despite not creating any triangles.

Avg Node Degree	Avg Clustering Coeff	Avg Path Length	Num Triangles
2.03	0.03	11.75	1

Table 3: Global Network Analysis



(a) Unchanged Network

(b) Rail Network With 4 Long Range Links
35% reduction in path length

(c) Rail Network With 8 Long Range Links
49% reduction in path length.

Figure 5: Rail Network

Problem 5

For preforming *node2vec_random_walk* I used parameter values of $p = q = 1.6$. To tune the parameters I would evaluate the importance of communities to my application, and increase p and decrease q proportional to community importance.

As shown in Figure 6, these raw embeddings do not clearly separate the pubmed classes. To improve interpretability of features, I would first pass the features through a GCN model and use the features at the end of the model, but before the classifier head, as what would be presented as interpretable. I would apply dimensionality reduction to allow for visualization and attempt to map regions of the visualized feature space to qualitative features of the classes they represent. This would also showcase distance between different classes, as two regions which are closer, or even overlap, would have similar qualitative class features, whereas one that can be separated more easily would have more unique features.

When training a two layer GCNConv model, I achieve 79% accuracy.

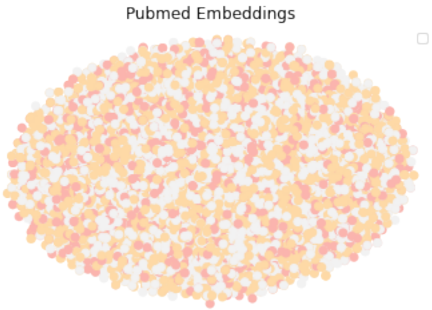


Figure 6: Node2Vec Embeddings