Name:	Mark Gameng
IIT ID Number:	A20419026

CS 579: Online Social Network Analysis

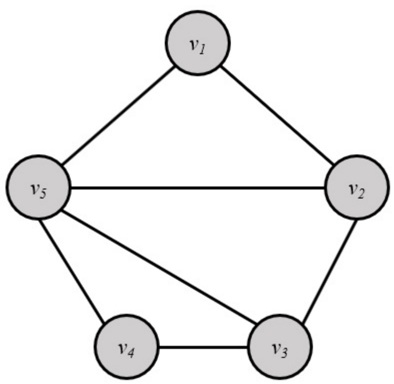
Homework II - Network Measures, Network Models and Clustering

Prof. Kai Shu
Due at 2022 Feb. 21, 11:59 PM

This is an ***individual*** homework assignment. Please submit a digital copy of this homework to **Blackboard**. For your solutions, even when not explicitly asked you are supposed to concisely justify your answers.

1. **[Network Measures]** Based on the following network answer the questions,

   (a) Fill the adjacency matrix.



	v_1	v_2	v_3	v_4	v_5
v_1	0	1	0	0	1
v_2	1	0	1	0	1
v_3	0	1	0	1	1
v_4	0	0	1	0	1
v_5	1	1	1	1	0

   (b) Calculate the "Degree Centrality" (normalized by the maximum degree) values and "Katz Centrality" values with $\alpha = 0.3$ and $\beta = 0.2$, and rank the nodes based on Katz Centrality (you can use Matlab or other mathematical software to calculate the eigenvalues).

d		**Degree Centrality**	**Katz Centrality**	**Ranks (Katz)**
2	v_1	2/4 = 1/2	1.33	3
3	v_2	3/4	1.72	2
3	v_3	3/4	1.72	2
2	v_4	2/4 = 1/2	1.33	3
4	v_5	4/4 = 1	2.03	1

maximum degree = n - 1 = 5 - 1 = 4

degree centrality = d/4

katz = B(I - aA^T)^-1 . 1

eigenvalues, using wolframalpha
= (2.94, -1.62, -1.47, 0.62, -0.46)

Name:	Mark Gameng
IIT ID Number:	A20419026

(c) Is the above alpha value a good choice for Katz centrality? Why?

I would say, yes. The above alpha value of 0.3 is a good choice for Katz centrality because it is recommended to choose alpha values below 1 / max(eigenvalues). So, in the case above, the max eigenvalue was 2.94. Thus 1 / 2.94 = 0.34. And the above alpha value was 0.3. 0.3 < 0.34, so the alpha value was a good choice for Katz centrality.

(d) Discuss what would happen if we set $\alpha = 0$?

If we set alpha to 0, then essentially the eigenvector centrality is removed and all the nodes katz centrality would just be beta, or the bias. This is because the formula for katz is just the eigenvector centrality + bias. For example, using part b, the katz centrality for each node would just all be 0.2.

(e) Calculate the global clustering coefficient of the graph.

C = ((# triangles) * 3) / (# connected triples of nodes)
= (3 * 3) / (3 * 3 + others) = 9 / (9 + 5)
= 9 / 14
= 0.64

open or others =
123, 153, 154
254
~~321, 351~~
~~452~~, 432

(f) Compute the similarity between nodes v_2 and v_5 using cosine similarity.

intersect = $\cap$

|{v1, v5, v3} intersect {v1, v2, v3, v4}| / sqrt(|{v1, v5, v3}| * |{v1, v2, v3, v4}|)
= |{v1, v3}| / sqrt(3 * 4)
= 2 / sqrt(12)
= 0.58

Name:	Mark Gameng
IIT ID Number:	A20419026

2. **[Network Models]**

(a) Why are random graphs incapable of modeling real-world graphs?

Real world networks have a bunch of properties such as degree distribution, clustering coefficient, and average path length. Degree distribution and power law distribution are common in real-world graphs, where small instances are very common while large instances are very rare. Like, twitter users with less than 10 followers and people with more than 1 million followers, or even just the wealth distribution. Another real-world graph can be friendships, which are highly transitive and thus results in high clustering coefficient. Studies has also been made that showcase that most people are connected to one another of at most 6 people, meaning 6 degrees of separation. That means that real world networks tend to have a short average path length.

The thing with random graphs is that the edges between nodes are formed randomly or following a certain distribution. However, real-world graphs are not at all random because human beings are rarely truly random. We tend to follow patterns, choose friends based on our own personality, buy popular or underrated stuff, etc. The more complex a random graph becomes, the closer it gets to a real-world graph but it still isn't good enough yet. Random graphs tend to not have a power-law degree distribution, and underestimate the clustering coefficient, but does well modeling the average path lengths. The closest we ever got was good degree distribution, average path length, but bad clustering coefficient.

(b) Show that in a regular lattice for small-world model, local clustering coefficient for any node is $\frac{3(c-2)}{4(c-1)}$, where c is the average degree.
Hint: See problem 5 in the textbook.

Regular lattice of degree c, nodes are connected to their previous c/2 and following c/2 neighbors.
In a regular lattice, the number of connections between neighbors is (3/8)c(c-2)
Denominator for undirected graph = di choose 2 = di(di - 1) / 2 = c(c - 1) / 2

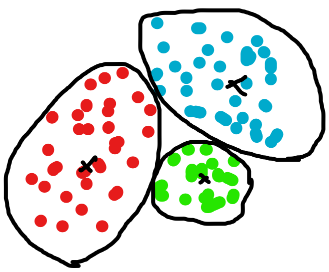
Thus, plugging in to the formula for local clustering coefficient for any node with c as the average degree
C = ((3/8)c(c-2)) / (c(c-1)/2)
= (2(3/8)c(c-2)) / (c(c-1))
= ((6/8)c(c-2)) / (c(c-1))
= ((6/8)(c-2)) / (c-1)
= (6(c-2)) / (8(c-1))
= (3(c-2)) / (4(c-1))

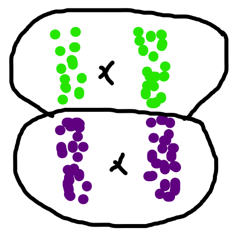
Name:	Mark Gameng
IIT ID Number:	A20419026

3. **[Unsupervised Learning]**

(a) What is the usual shape of clusters generated by k-means? Why is this the case? Justify your answer by referencing the algorithm.

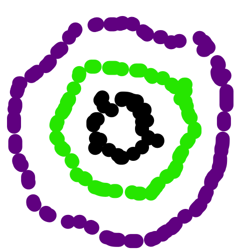
The usual shape of clusters generated by k-means are spherical. This is due to how the algorithm goes about grouping points together. Basically, it assigns a point to a cluster so that the distance between that point and the cluster's centroid is at the minimum. Due to the algorithm using a centroid to form clusters, the clusters tend to form spherical shapes. Now, this is great when the "correct" clusters are actually spherical like a circle or sphere, if three dimensional. However, for other shapes or forms, it doesn't do well.





wrong

(b) Give (or draw) an example of the case where k-means is unable to correctly classify data instances due to the pattern of these instances.
Hint: use your answer from part (a).



Correct Classification

K-means classification (wrong)

Good Luck