Sumer Bangera, July 2020

Network Analysis of Zachary's Karate Club using Python (NetworkX)

Summary

Network A in this report refers to the "soc-karate" network which contains social ties among the members of a university karate club collected by Wayne Zachary in 1977. The tool used for analysis is Python along with the Networkx package.

Since the original network has an even number of nodes i.e. k = 34, a random sample of 17 nodes has been selected to create Network A_{sample}. Summary of the two networks is given below:

	Network A	Network A_{sample}
No. of nodes	34	17
No. of edges	78	22
Connected	Yes	Yes
Directed	No	No

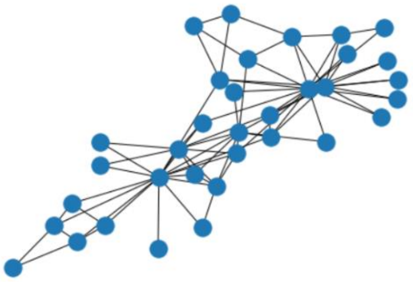


Fig. 1: Network A

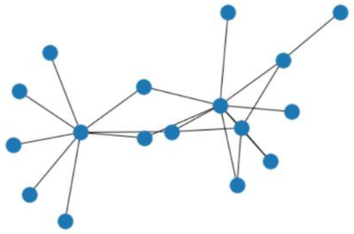


Fig. 2: Network A_{sample}

The 3 graph-level metrics used to compare Network A and Network A_{sample} are as follows:

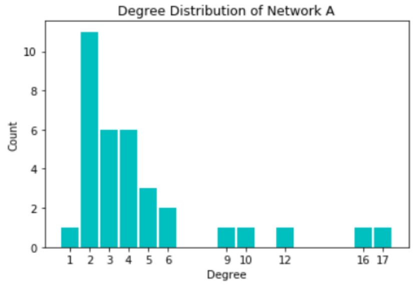
1. **Graph Density:** Graph density is the ratio of the actual number of edges in the network to all possible edges in the network measured on a scale of 0 to 1. It helps us understand how closely knit the network is.
   a. Network A: 0.139
   b. Network A_{sample}: 0.1618

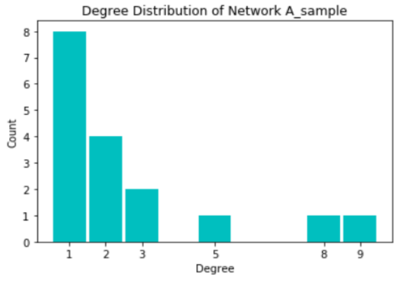
   Interpretation: Here, we observe that Network A is not very dense and in fact, the randomly sampled Network A_{sample} has higher graph density. This can be attributed to the increase in the number of possible edges due to a higher number of nodes in Network A even though the proportion of actual edges is higher as compared to actual edges in Network $A_{sample.}$

2. **Diameter:** Diameter is the longest shortest path in the network, i.e. it the shortest path between the two most distant nodes derived by calculating all possible shortest paths between all pairs of nodes in the network
   a. Diameter of Network A: 5
   b. Diameter of Network A_{sample}: 5

   Interpretation: The diameters of both the graphs are identical implying that the furthest apart nodes require a minimum of 5 links (edges) to be connected to each other.

3. **Degree Distribution**
   Degree distribution helps us understand the probability distribution of degrees of nodes across the network





Interpretation: Here we observe that the degree distributions of both the networks are skewed to the right. This indicates that many nodes in the network have a smaller number of connections (edges) with other nodes in the network, however, there exist few significantly connected nodes in both the networks.

In conclusion, these metrics help us delineate similarities and differences in the two networks. Apart from subtle differences in the graph densities and degree distributions, both networks behave in a similar manner. The randomly sampled network tends to exhibit the characteristics of the original network based on the aforementioned metrics. This may or may not hold true for other metrics not tested in this report.

Centrality measures and their alternative business applications

- **Degree:** The degree of a node is the number of nodes each node is connected (has links) to in the network
  Business case: Strategizing formations and player transfers for football (soccer) managers
  Nodes: Players on the football field during a match
  Edges: Passes between the players. If player A passes the ball to player B during the game, they are connected. The edge strength/thickness defines the number of passes between the two players.
  Application: Degree and edge strength would help the manager identify key players or player positions during a match. For example, a central midfielder would be pivotal in controlling the game, connecting passes from the defenders and providing them to the forwards to create goal-scoring chances. Studying his/her passing patterns, the manager can devise new formations or strategies. This study can also be useful for buying new players that fit the team's playing style.

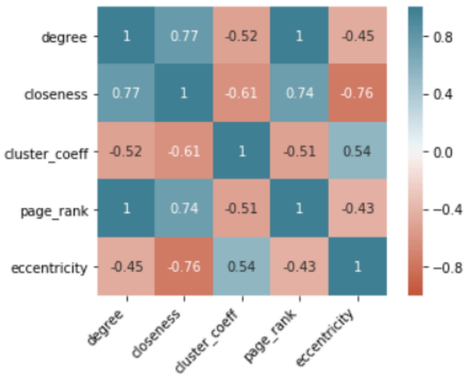
- **Closeness:** Closeness measures the distance of a node with respect to all other nodes in the network
  Business case: Disseminate vital information/guidelines to employees within an organization
  Nodes: Business teams within the organization
  Edge: Contact connection. If a team can communicate or contact another team then they are connected.
  Application: In case of a change in, say, accounting requirements the organization would need all concerned departments to comply with the new norms. Hence, understanding the most central team(s) can facilitate quick compliance and training as these teams would be most closely connected with other teams.

- **Clustering Coefficient:** Clustering coefficient of a node is the fraction of its neighbours that are connected to each other
  Business case: Designing marketing strategies by studying customer segmentation and predicting new customers
  Nodes: Customers (and potential customers)
  Edges: Profiling similarities i.e. two customers are connected if they share certain similar attributes. These attributes may differ based on the business e.g. location, purchase patterns, demographics, etc.
  Application: Visualizing such a network can help businesses understand various segmentations in their customer base and their behaviours. This can aid strategies such as preferential pricing, rewards or loyalty programs and even predicting the potential clusters/segmentation of new customers

- **Page Rank:** Page rank provides a ranking (importance) for each node in the network as a function of (incoming) links to the respective node in a directed network. For an undirected network, the algorithm converts it into a directed network by transforming each edge into two edges.
  Business case: Branding campaigns for new products through social media influencers
  Nodes: Individual social media accounts - Influencers (highly connected nodes) and their followers
  Edges: Connection between the accounts, i.e. if an account follows another account, there is a link between them. In a directed graph, there will be incoming links to the influencer node due to the higher number of followers while there may not be a reverse link.
  Application: Brands can use the influencer nodes to publicize their new products or promotions faster than trying to reach each follower node individually. Using the influencer nodes, the brands can swiftly access their cluster of users to increase their potential marketing reach

- **Eccentricity:** In a connected network, the eccentricity of a node is the maximum distance between the node and any other node in the network. All nodes in a disconnected network have infinite eccentricity.
  Business case: Choosing the main store to stock majority inventory for retail product businesses (eg. Nike)
  Nodes: Stores
  Edges: Paths between the stores. If there exists a path/route between two inventories (nodes) then they are connected
  Application: Every business would want to stock its excess inventory strategically which can aid the efficient transfer of the products. Although all stores tend to have their own inventory, there usually exist some central outlets for retails brands. For this, nodes with low eccentricity will have better direct connectivity with other nodes in the network and therefore can be crucial and efficient in supplying backup stocks when required.

Pearson's correlation among the 5 node-level metrics



Observation:

1. An evident observation is the strong (almost perfect) positive correlation between Degree and Page Rank implying that an increase in the Degree of a node will cause its Page Rank to increase. This is plausible as the Degree indicates the connectivity of a node and therefore, its importance in the network which is further translated into Page Rank of the node.
2. Degree and Closeness also indicate a strong positive correlation in Network A which implies that a node with a higher Degree is, in general, closely connected to all nodes in the network.
3. There exists a strong negative correlation between Closeness and Eccentricity as Closeness is concerned with the minimum distance between the nodes and Eccentricity calculates the maximum distance between them. Therefore, higher Closeness would translate to lower Eccentricity. This is evident even in the relationship between Degree and Eccentricity.
4. Clustering Coefficient is moderately correlated with the other four metrics of Network A. While it has an inverse relationship with Degree, Closeness and Page Ranks (which show strong positive correlations with each other), Clustering Coefficient has a positive correlation with Eccentricity. Therefore, as Eccentricity of the node increases, the Clustering Coefficient is also likely to increase.

Random graph creation:

Assumptions:

1. The random network created is connected and undirected
2. There are no loops in the network
3. The network is chosen uniformly at random from the set of all networks with 34 nodes and 78 edges

Summary of metrics for both networks are as follows:

Metric	**Network A**	**Network $_{random}$**
Graph Density	0.139	0.139
Diameter	5	4
Degree Distribution	Positively skewed with a long tail	Slight positive skew (almost Normal)

The graph densities are identical as the number of nodes and edges is the same. The diameter, however, is smaller for the random Network $_{random}$ which means it is more well connected in terms of shortest paths as compared to Network A. Moreover, the degree distribution of both networks differs significantly since Network A has a positive skew whereas Network $_{random}$ has an approximately normal curve with some positive skewness.

Summary of statistical tests (one way F-test) for significance (at 95%)

Metric	**p-value**	**Conclusion**
Degree	1.0	Significantly not different
Closeness	0.6532	
Clustering	6.3324e-10	Significantly different
Page Rank	1.0	Significantly not different
Eccentricity	0.1016	

Conclusion:

Network A and Network $_{\text{random}}$ show subtle differences in graph-level except for degree distribution where they differ significantly. In terms of node-level metrics, the two networks differ significantly only on the basis of Clustering. They do not differ significantly with respect to other terms. Therefore, we can conclude that Network A is sparse yet relatively more clustered.