# Social Network Analysis

Shantanu Tyagi (201801015)* and Arkaprabha Banerjee (201801408)†
*Dhirubhai Ambani Institute of Information & Communication Technology, Gandhinagar, Gujarat 382007, India*
*SC-435,Introduction to Complex Networks*

In this paper we shall attempt to holistically analyze social networks by looking at trust based networks. It shall model out the key features of such networks and also engage in a qualitative as well as quantitative discussion of key opinion influencers in such networks.

## I. INTRODUCTION

A social network can be defined as a set of relationships or interactions among diverse social entities such as individuals, groups or even organizations. Social networks play a pivotal role in the modern age and lately there has been an exponential growth of social network data available on the web. Social networks play a paramount role in disseminating information and even opinions by tapping into user behaviour.

A significant number of consumer review as well as e-commerce websites allows its users to review or rate other products, sellers or even existing reviews in a bid to enable the end consumer to make an informed decision. Prominent market studies and reviews also indicate that a significant number of consumers are heavily influenced by such reviews stemming out of opinions of other users. By incorporating these features of interacting with other users from social networks, these websites end up forming various social structures which enable us to give a deeper understanding into human behaviour and psychology.

One such consumer review website is Epinions.com which was later on acquired by Shopping.com. Epinions allowed its users to review and rate products, services, sellers as well as existing reviews. Users expressed their trust or distrust towards other reviewers/ users via this method. This interplay of trust and distrust towards other users in this networks led to the formation of a **Web of Trust** which can be modelled in the form of a signed directed network[3]. A positive edge from one user to another signifies that the first user trusts the other while a negative edge suggests that the first user distrusts the latter user. Users who are trusted more often hold key influence while shaping opinions.

In this paper we shall investigate this network from 2 important perspectives :

1. Analyzing key network and structural metrics by comparing it with other similar networks.

2. Identifying key opinion influencers by use of various centrality measures and engaging in qualitative and quantitative arguments.

For the first segment we shall compare standard network metrics of the epinions dataset[4] along with the Wikipedia Vote Network[4] as well as the Bitcoin OTC trust weighted signed network[5, 6]. Furthermore, degree-relationships and rich club effect phenomena shall be modelled for the epinions dataset. For the second segment we shall engage in various standard as well as enhanced centrality measures in order to identify key opinion influencers by looking at a variety of parameters.

Analyzing such data can often give us very interesting insights into the dynamics of the network and even potential information about how large scale societies work as a whole

## II. DATASETS

Although the primary focus is the epinions dataset, however two additional datasets representing similar trust based networks have been used for a comparative analysis.[1]

- **Epinions Social Network** : This represents a who-trusts-whom network derived from the consumer-review site: epinions.com. This is a directed signed relationship where each node represents an user and a positive edge from one user to another represents a sign of trust from the former to the latter. On the contrary, a negative edge represents a sign of distrust from the former to the latter. This interplay of trust and distrust forms the web-of-trust. The data has been downloaded from the official SNAP website (`https://snap.stanford.edu/data/soc-sign-epinions.html`) . On account of the computational complexity involved, experiments have been performed on the maximal weakly connected component unless mentioned otherwise.

- **Wikipedia Vote Network** : This network represents the vote network formed in the official Wikipedia page whenever there is a public discussion on making a certain user an administrator giving them higher control over technical features and maintenance of the website. This an unweighted directed network where an edge from one user to another represents that the former votes in favour of the

*Electronic address: 201801015@daiict.ac.in
†Electronic address: 201801408@daiict.ac.in

latter for the aforementioned position. The data has been downloaded from the official SNAP website (https://snap.stanford.edu/data/wiki-Vote.html).

- **Bitcoin OTC trust weighted signed network** : This represents a who-trusts-whom network formed among users who trade using bitcoin on an online platform : Bitcoin OTC. In this network, users rate each others based on the trustworthiness of the other user, post a transaction. It has primarily been formed in order to identify risky and fraudlent users and transactions. This network is characterized by a weighted directed network with rating varying from -10 to +10 in steps of 1. A positive edge from one user to another represents a sign of trust from the former to the latter while a negative edge represents a sign of distrust from the former to the latter. The data has been downloaded from the official SNAP website (https://snap.stanford.edu/data/soc-sign-bitcoin-otc.html).

## III. NETWORK STATISTICS

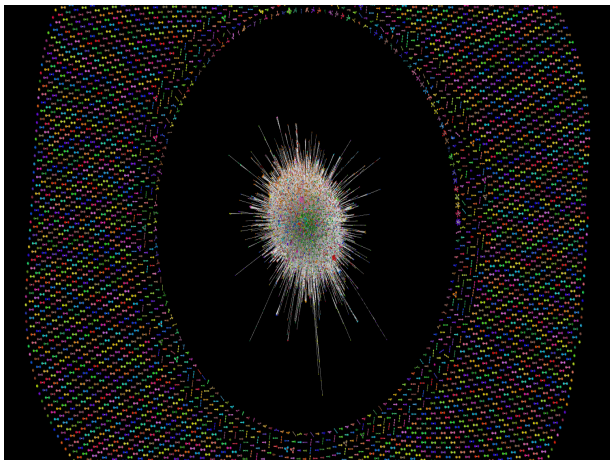


FIG. 1: Graph Visualization : Color coded according to community (made using graphia.app) For other visualizations click here

The epinions dataset is a relatively sparse network/graph. The degree distributions in log scale for in-degree and out-degree in the form of a histogram have been shown below.
A tentative fitting of the degree distribution with a power law distribution, for both cases reveal the power law exponent to be approximately 1.7. However, KS Test at lower significance levels (0.05) rejects the hypothesis that the aforementioned degree distribution is a Power-law distribution. At higher error tolerance levels one may obtain the hypothesis to be true. Since the power-law exponent is less than 2 thus we can't classify this network under the hood of scale-free regime. The average characteristic path length for the maximal weakly connected component is around 4.2.

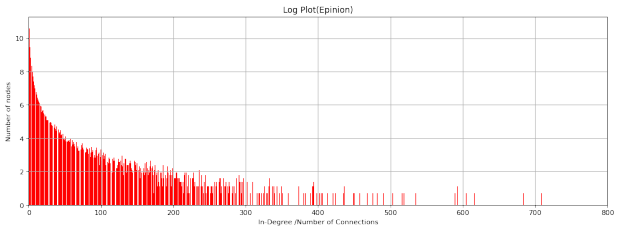



FIG. 2: Log plot for number of nodes at given in-degree/out-degree value

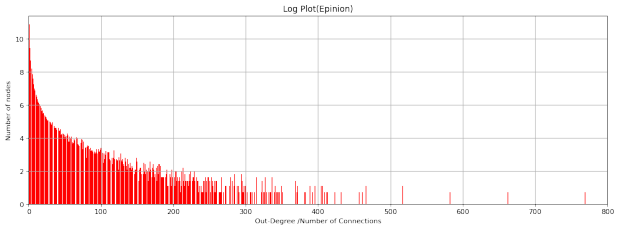



FIG. 3: Log plot for number of nodes at given in-degree/out-degree value

We find the size of the maximal weakly connected component to be 119130 nodes and the maximal strongly connected component to be 41441 nodes. The maximal weakly connected component consists of tentatively 99% of the net number of edges present in the entire graph. Thus the network consists of one major component along with multiple other small components. This is also clearly visible in the visualization of the entire network (Fig 1). This observation also affirms our assumption of working with the maximal weakly connected component for our experiments and generalizing it to the entire network.
Community detection algorithms showcase four major communities (of size 1000 or more) with the following sizes : 11587, 10540, 10032,4309 and 7 more communities of size 100 or more : 945, 453, 311, 221, 143, 124, 111. All of these communities belong to the maximal weakly connected component. Apart from this we also obtain multiple other small communities.This highlights the polarized structure of the network and can also been seen in Figure 1.

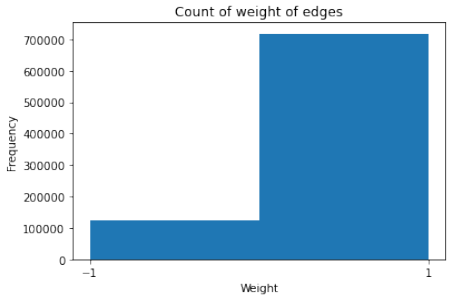



FIG. 4: Edge weight distribution

A quick look at the edge weight distribution reveals that a majority of the segment were trusting relationships. Since the website, didn't display distrust relationships publicly , thus people usually didn't

engage with other users having low trust scores. This effect possibly increased multi-fold over time in the network leading to some users being trusted too much.

A comparison of key network metrics for all the 3 graphs can be seen in I. The size of the Wikipedia and Bitcoin OTC network are comparable but vastly smaller than the epinions network. However, the Epinion dataset is very sparse as compared to other two which are comparatively dense. This is probably on account of the fact that a vast majority of people are not active users who rate each other but rather use the consumer review site to gain information regarding a product.

Clustering Coefficient usually represents the extent to which nodes in a network tend to cluster together. Social networks in general tend to have higher values of this metric on account of the formation of groups and tightly knit communities. In this case although the absolute value of this metric for epinion dataset is less however, after accounting for its size and comparing it to a random network of similar properties it has a relatively high average clustering coefficient. The comparatively high clustering coefficient and low path length is characteristic of a small-world network which is often seen in such social networks.The Bitcoin OTC network also has a high value. This is perhaps on account of the fact that people usually tend to have transactions with trusted users whom they know via someone. This is in turn leads to the formation of closely knit communities.

| **Metric** | **Epinion** | **Wikipedia** | **Bitcoin** |
|---|---|---|---|
| Nodes | 131828 | 7115 | 5881 |
| Edges | 841372 | 103689 | 35592 |
| Reciprocity for +ve edges | 0.347 | 0.0564 | 0.83911 |
| Reciprocity for -ve edges | 0.0379 | NA | 0.1706 |
| Density | 4.84145e-05 | 0.002048 | 0.00102 |
| Average Clustering Coefficient | 0.0956 | 0.081563 | 0.15106 |
| Maximum Out-degree | 2070 | 893 | 763 |
| Maximum In-degree | 3478 | 457 | 535 |
| Degree-Assortavity Coefficient (in,out) | 0.01754 | -0.07107 | -0.09253 |
| Degree-Assortavity Coefficient (out,in) | 0.07307 | -0.005002 | -0.09253 |
| Degree-Assortavity Coefficient (in,in) | 0.05284 | 0.02533 | -0.07953 |
| Degree-Assortavity Coefficient (out,out) | 0.050377 | -0.017617 | -0.1072 |
| Alpha Value | 1.704944 | 3.629870 | 2.2708 |

TABLE I: Network Metrics for all 3 cases

Reciprocity in directed graphs refers to the the tendency that when a certain node connects to another node, then the latter also connects to the previous one. For our experiments we have considered reciprocity for positive (trust) and negative (distrust) edges separately. One can observe that for all networks the reciprocity value for positive edges is significantly higher as compared to its negative counterpart. This translates to the fact that if you trust a certain user then he is also likely to trust you. This effect is less pronounced in Wikipedia vote network because mutual voting may not be semantically correct.

Assortavity coefficient represents the extent to which nodes are connected to each other based on a given property. In this case we shall be looking at their degree via 4 use cases[7] :

1. **Out-degree for source node and In-degree for Target node** :
2. **In-degree for source node and Out-degree for Target node**
3. **In-degree for source node and In-degree for Target node**
4. **Out-degree for source node and Out-degree for Target node**

In the above table I, Degree-Assortavity Coefficient (a,b) defines type of links for source node a and target node b. We find a positive correlation for epinion in all cases but a negative value in most cases for the others. The highest value for this coefficient in the epinion dataset is for source nodes with out-degree and target nodes with in-degree. Thus nodes with similar out-degree tend to connect with similar in-degree nodes. This effect is more pronounced for higher degree nodes as can be seen from the graphs plotted below. Thus, one can conclude that users who rate a lot of people also often tend to rate influential reviewers who have been reviewed a lot.

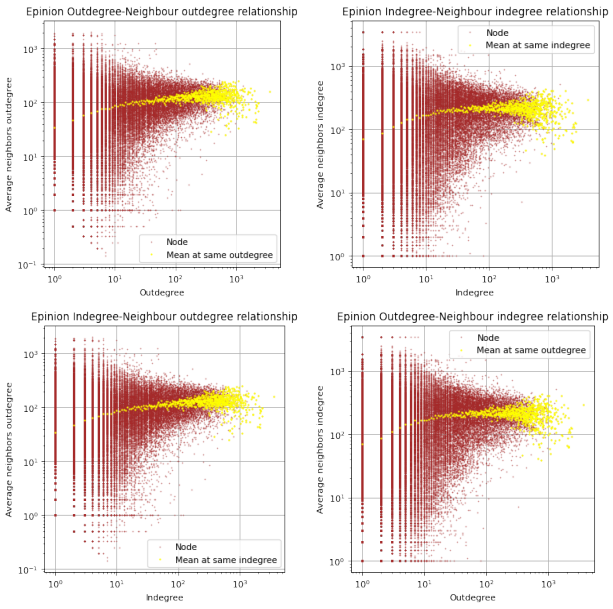



FIG. 5: Assortavity Analysis

The above plots represent the in-degree or out-degree on the x-axis and the average neighbour neighbor in-degree or out-degree on the y-axis (both of

them are on log-scale) and serves as a good medium for visualizing and understanding the assortavity coefficients. One can clearly see that for lower degree nodes for all cases there is a lot of variation regarding their neighbour's degree, but for higher degree the effect is slightly more clear and one can see the tendency of higher degree nodes to connect with each other albeit with some variation.

The power law exponent for Wikipedia and Bitcoin OTC are 3.69 and 2.27 respectively and both of them satisfy the KS-test with 0.05 significance level. Since the exponent for Bitcoin OTC is between 2 and 3 thus it also falls in the scale free regime.
The comparative analysis plots are primarily histograms with frequency on the y-axis and bin number on the x-axis for better visibility. The starting cutoff for all metrics is from 0.01 (bin 0) and goes up to 100 bins in a linear fashion.The enhanced in-degree centrality plot refers to degree centrality proposed in [2] by considering in-degree (instead of degree in undirected networks).

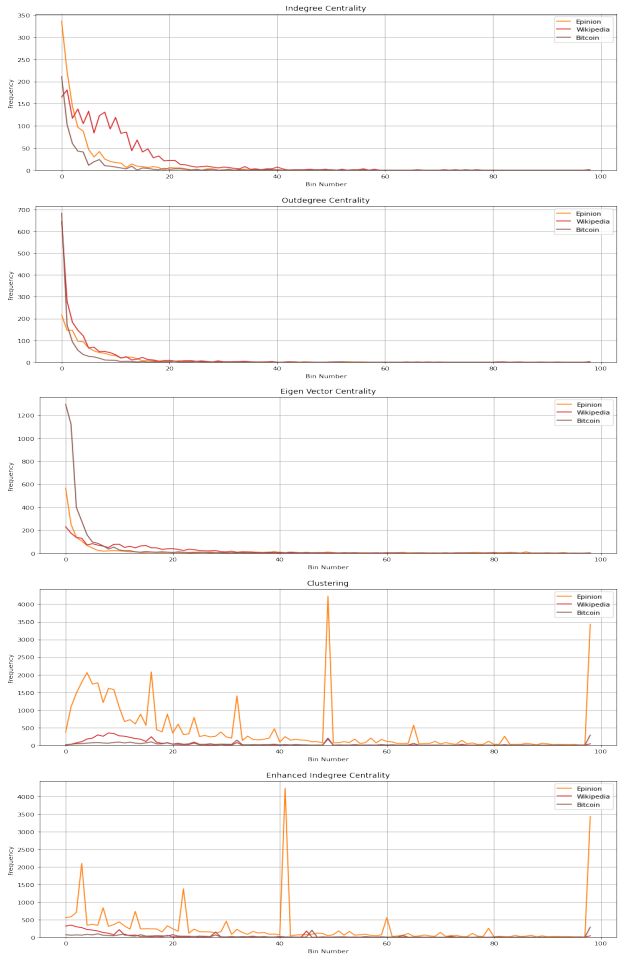



FIG. 6: Comparative Analysis

Although the in-degree centrality plot trends varies for the 3 networks however the trend for all the other metrics reveal almost identical trends among the 3 networks, with similar peaks in most cases. The plot for enhanced in-degree centrality also has peaks which almost coincide with those of clustering coefficient. This is perhaps on account of the fact that clustering coefficient is also used to calculate the enhanced degree centrality. This gives us an insight that most trust based social networks have similar centrality distributions even though their network size may vary vastly.

The Clustering coefficient plot also reveals that for the weakly connected component that there is significant clustering in some cases (around key opinion influencers) along with the majority of the nodes having values on the slightly lower end. This is perhaps on account of the sparse nature of the network.

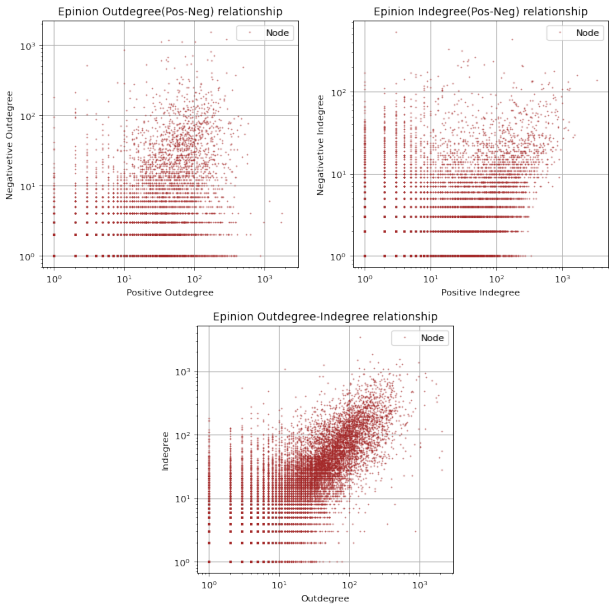



FIG. 7: Degree Analysis

The above plots showcase the degree relationships for a given node. From the second figure (+ve v/s -ve indegree) one can see that even for the most unreliable reviewers the number of negative reviews don't usually go above 100. But for the counterpart good reviewers the number of good reviews is in 1000s. Thus people tend to give more attention to good reviewers and not spend their effort on already distrusted reviewers. This is also visible from the first figure as the previously mentioned edge weight distribution. The overall in-degree v/s out-degree plot also reveals that people having higher in-degree also tend to have higher out-degree, i.e people having higher number of reviews also tend to give higher number of reviews to others. However, this is only a correlation and doesn't imply causality.

Rich club effect[12] measures the extent to which well-connected nodes also connect to each other. The above graph shows positive and negative rich club effect for the 100 most influential nodes in terms for the indegree count at those nodes. This indegree count represents trust or distrust depending on the positive or negative value respectively. The higher the sum of these two values, the more influential the node.

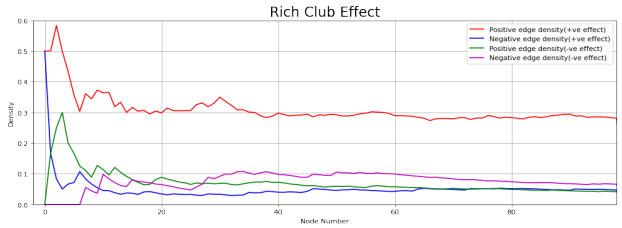



FIG. 8: Rich Club Effect

The density values plotted on the y axis represent the degree of interconnection between these most trusted and distrusted nodes. First we consider the top 100 most influential nodes in terms of their in-degree. We observe in the red curve(8), the density stays around 0.3 after the first few nodes which tells us that these influential nodes are not that well connected with each other and thus rich club effect is not seen for positive influencers. We see in the blue curve(8) that the most influential members also tend to distrust the other most influential members. Next we consider the the top 100 most distrusted nodes. These nodes have the highest negative indegree and can include people who are disliked by others maybe because they have a fake/spam/troll account with regards to our dataset. However these kind of people might be trusted among their controversial peers. We see in the green curve(8) that such a scenario does not arise due to the low density values indicated by the curve. Infact the trust is overpowered by distrust, as observed in the magenta curve(8), which depicts that the most distrusted nodes tend to distrust the other most distrusted nodes with a low reciprocity value of 0.39.

## IV. ANALYZING KEY INFLUENCERS

In order to quantify influence, centrality measures prove to be a very useful metric specially for social networks.However different centrality measures hold different contextual meanings and it's upto to the user to choose a valid centrality measure depending on the use case. For the purpose of our experiments the following centrality measures[8] have been employed on the epinions dataset such that they help us understand social trends within our limited computational complexity prowess:

1. In-degree Centrality : This refers to the direct number of nodes which point / connect to our node of interest and can be helpful in finding people who have been rated by a huge segment of people. Previous network statistical analysis also reveals that usually trust relationships are associated with higher in-degree centrality.

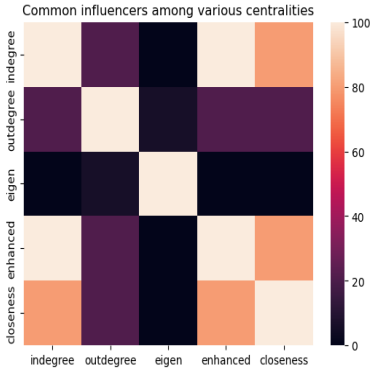
2. Out-degree Centrality : This refers to the direct number of nodes which our node of interest points / connects to. It can be helpful in identifying people who often rate/review other people.

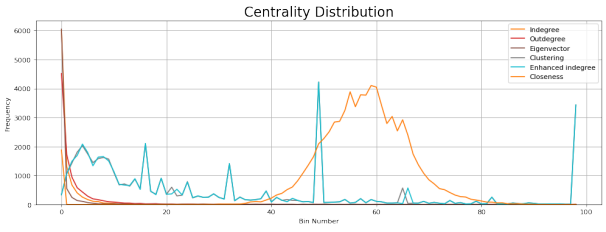
3. Eigen Vector Centrality : This refers to a node's influence based on its connectivity to other nodes in the network which themselves are also well connected. Thus it takes into account, the overall network structure rather than local node level trends as seen in degree-centrality measures.

4. Closeness Centrality : This refers to a measure where the inverse of the sum of all shortest paths to a given node is calculated. In this case incoming edges have been considered to calculate the shortest path. Thus nodes with high centrality value are much more accessible to other users. Under the assumption that users gain an idea of the reviewee before giving a rating this centrality value can ideally give us nodes which can help spread an idea or opinion very productively through the network. [9, 10].

5. Enhanced In-degree centrality : This refers to the centrality measure proposed in [2]. This particular method enhances the basic degree centrality formula by incorporating the clustering co-efficient of a particular node. In this case,only in-degree has been considered. This is being done on account of the fact that social networks in general associate with higher clustering coefficient values. Thus incorporating this measure will help in a more holistic exploration of influential nodes especially for social networks by giving us a threshold from the average value of enhanced degree centrality which can then be employed as a cutoff for standard in-degree-centrality.

All the above centrality experiments have been performed on the maximal weakly connected component and the ranked nodes can be obtained from this link. The heatmap shown below highlights the number of common elements between each centrality measure by considering the top 100 influencers. One can observe that enhanced in-degree centrality has a very high number of common influencers as that found out by in-degree centrality. This is on account of the fact that both methods use in-degree as the primary metric and the enhanced method is useful when we have defined threshold values[2] so that we can identify more active /influential nodes however their order may still remain the same. In-degree centrality also contains a large number of common elements with closeness centrality thus signifying that nodes that are reviewed a lot often tend to be much closer to other nodes in a network and potentially help spread their opinion in a much faster fashion. In-degree centrality and out-degree also have a fair number of common elements thus affirming our observation that nodes that have higher in-degree also tend to have higher out-degree. The lowest number of common elements is found between in-degree and eigen vector centrality. This can be explained by the fact that due to limited number of key influencers, although they may be interconnected (as seen by the assortavity coefficient), but the absolute value may still be small.




We find that except enhanced degree centrality and closeness distribution, all other distributions only have a high number of users with very low centrality values and then decreases very sharply. This is more or less consistent with the overall network structure. The trend for enhanced degree centrality is similar to the clustering coefficient rather than in-degree centrality in spite of contribution from both. This is on account of the comparatively larger value of clustering coefficient as opposed to the normalized value of in-degree used in its formula. Closeness centrality has a very high number of nodes with very low centrality as well as with moderate centrality. The high count of moderate centrality is perhaps on account of the fact that a large number of users are connected to each other via someone else, something which is also synonymous with small-world network.

## V. CONCLUSIONS

This paper has performed a comparative analysis of the epinion dataset with the wikipedia vote network and the Bitcoin OTC network, all of which are more or less trust based networks. Our experiments showcase that although they have similar centrality and clustering metrics but they do vary to a slight extent on their structural metrics. However, those differences are on account of the semantical meaning of the network. We find the epinion dataset to be a large sparse network with comparatively high values of reciprocity, clustering coefficients and degree assortavity coefficients and low characteristic path length. Although there is a tentative power law fitting for the network however it only happens with high error tolerance levels.The comparatively high clustering coefficient and low path length is characteristic of a small-world network which is often seen in such social networks. The degree assortavity coefficients indicate that people who often tend to rate others also rate key influencers thus leading to high credibility. Furthermore, key opinion influencers also tend to rate/review other users a lot. A closer look at key influencers via various centrality measures indicate that influencers via in-degree centrality , enhanced in-degree centrality and closeness centrality have a lot of common users followed by users from out-degree and eigen vector centrality . Apart from enhanced in-degree and closeness centrality all other centrality distributions follow a similar decaying distribution.


[1] Hashmi, A., Zaidi, F., Sallaberry, A. and Mehmood, T., 2012, August. Are all social networks structurally similar?. In 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 310-314). IEEE.

[2] A. Srinivas and R. L. Velusamy, "Identification of influential nodes from social networks based on Enhanced Degree Centrality Measure," 2015 IEEE International Advance Computing Conference (IACC), 2015, pp. 1179-1184, doi: 10.1109/IADCC.2015.7154889.

[3] Patil, A., Ghasemiesfeh, G., Ebrahimi, R. and Gao, J., 2013, September. Quantifying social influence in epinions. In 2013 International Conference on Social Computing (pp. 87-92). IEEE.

[4] Leskovec, Jure, Daniel Huttenlocher, and Jon Kleinberg. "Signed networks in social media." Proceedings of the SIGCHI conference on human factors in computing systems. 2010.

[5] Kumar, Srijan, et al. "Edge weight prediction in weighted signed networks." 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 2016.

[6] Kumar, Srijan, et al. "Rev2: Fraudulent user prediction in rating platforms." Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. 2018.

[7] Foster, Jacob G et al. "Edge direction and the structure of networks." Proceedings of the National Academy of Sciences of the United States of America vol. 107,24 (2010): 10815-20. doi:10.1073/pnas.0912671107

[8] Du, Donglei. "Social network analysis: Centrality measures." University of New Brunswick (2019).

[9] Landherr, A., Friedl, B. Heidemann, J. A Critical Review of Centrality Measures in Social Networks. Bus Inf Syst Eng 2, 371–385 (2010). https://doi.org/10.1007/s12599-010-0127-3

[10] Wasserman, Stanley, and Katherine Faust. Social Network Analysis: Methods and Applications. Cambridge University Press, 1994.

[11] http://olizardo.bol.ucla.edu/classes/soc-111/textbook/_book/

[12] Muscoloni, Alessandro, and Carlo Vittorio Cannistraci. "Rich-clubness test: how to determine whether a complex network has or doesn't have a rich-club?." arXiv preprint arXiv:1704.03526 (2017).