Complex network analysis project

Rade Nježić

July 2020

Abstract

In this short project, we have analyzed the who-trust-whom online social network of a general consumer review site Epinions. It was a website where people could review products. Users were able to register for free and earn money according to how much their reviews were found useful. Due to attempts to game the system and as a possible fix, a trust system was put in place. Users were able to put people in their web of trust, where they would put people whose reviews were consistently found useful, and their "Blocklist", i.e. "authors whose reviews they find consistently offensive, inaccurate, or in general not valuable".

We have performed the structural analysis, interpreted and described structural balance properties, and provided an example of the adoption of an opinion.

Preprocessing

The dataset was downloaded from the SNAP[Les] website. The project was completed in python, using the library NetworkX. To produce plots and get other useful statistics, we have made a data frame containing the statistics for all types of degrees for every node.

Structural analysis

The network is given is a form of a directed weighted graph, where the edges of weight one denote the trusting relationships, while the ones with the weight -1 denote the distrusting ones. The values of the important measures are the following[1] :

Number of nodes	131827
Number of edges	841741
Density	4.84e-05
Avg. path length	4.1
Avg. clustering	0.0956
Transitivity	0.0742
Diameter	12

This is a sparse network with a high clustering coefficient, which is common for a social network. The average clustering coefficient is high because it is much bigger than it would be if we had a random network. Transitivity is to the average clustering coefficient. The diameter is small with respect to the network size. Before proceeding further we note that positive edges make around 85% of the total amount of edges.

Power law As we were working with the trust network, we were interested in the degree distribution of the nodes that have at least one indegree (that is, at least one person trusts or distrust that person). The indegree here is serving as a proxy for popularity. Our first assumption is that this distribution follows a power law (a small number of highly trusted nodes, bug number of lowly trusted ones). One way to see it to check if it exists is if the log-log plot of degree distribution is linear.

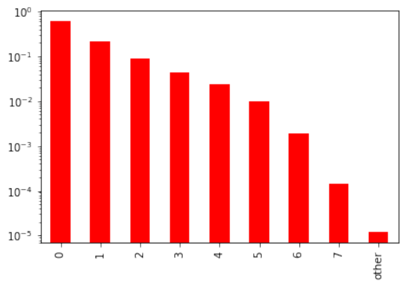

Figure 1: Log-log degree distribution

[1] The average path length and the diameter taken from the SNAP website (we were unable to compute with existing libraries due to disconnectedness).

We observe linearity, so our assumption of the power-law was true. The next thing we want to check is if the degree distribution is in a scale-free regime, that is

$$P(k) \sim k^{-\gamma},$$

where k is denoting degree size and $\gamma \in (2,3)$. We run the built-in function and get that power-law coefficient is equal to 1.70, so this network is not in a scale-free regime. Even if we are not in a scale-free regime, the degree distribution is heterogeneous, and mean carries little statistical significance. We repeated the same calculation for the density of the outdegrees, and the results were similar (power-law coefficient of 1.72). We can interpret that in a way that few users make most of the site's activity (or maybe just care about trusting or distrusting). We have also calculated the maximum degree values.

Total indegree	3478
Positive indegree	3338
Negative indegree	540
Total outdegree	2070
Positive outdegree	2070
Negative indegree	1562

Table 1: Maximum degree values

Small world phenomena The average shortest path of the biggest WCC is 4.1, which is less than both the transitional baseline of 6 and 11 (logarithm of the total amount of nodes in it). So the underlying directed graph of our network is a small-world model. This is common for the networks having hubs: for the scale-free regimes, it is proportional to $\log \log N$ (it would be 2.4 in this case, if we were in a scale-free regime).

Bow-Tie structure We can get more information about the structure of the network by examining the connected components. Maximum strong are weakly connected The second biggest connected component was equal to 15, and the second-largest weakly connected component was equal to 20. Therefore we can conclude that our graph has one big component where people follow each other amounting to around 30% of the total network and around 90% of the nodes are connected in one way or another. The number of edges in these components is 693737 and 833695, representing 0.825% and 0.991% of the total amount of edges. We made an analogy with the structure of the web, end checked if it had a bow-tie structure. We found out the sizes of the outer components and got the following results:

Component	Size
OUT component	37682
IN component	29732
Tendrils	10275

Table 2: Sizes of the components

We indeed have a bow-tie structure. It is comparable to the original web structure, except that the overall proportion of tendrils is much smaller. That is normal, considering we have a small-world network.

Reciprocity The measure that is of big importance to our networks in the reciprocity. Reciprocity can be connected with social selection (I trust people who trust me) . For our graph, the overall reciprocity is equal to 0.31. For the strongest connected component, the reciprocity was equal to 0.37.

Degree pair plots We have also explored the relationship between different types of degrees:

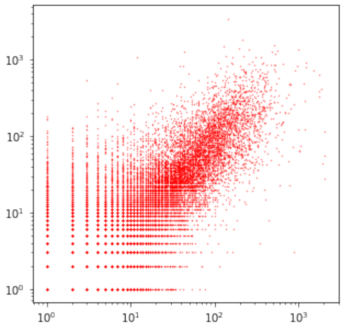

Figure 2: Outdegree vs indegree relationship

We see that on average, the nodes with then higher outdegree tend to have higher outdegree. How nodes with smaller indegrees and outdegrees, there is no noticeable correlation.

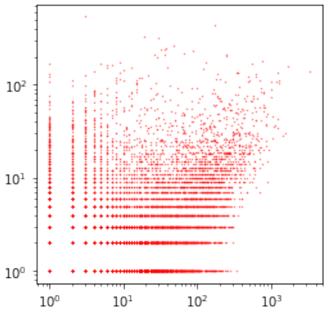

Figure 3: Positive vs Negative indegree

From this plot, we see that the most trusted authors don't tend to be the most distrusted ones. And that highly distrusted authors usually have around 100 negative ingoing edges.

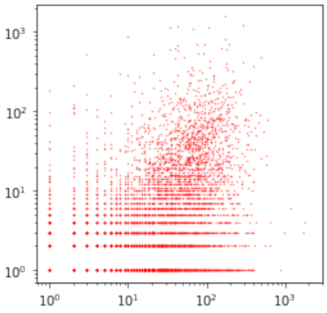

Figure 4: Positive vs Negative outdegree

From the plot, we see that a big majority of people prefer trusting to distrusting relationships. But we already know that from the edge weight distribution.

Degree Correlation We recall that assortativity means that the nodes tend to attach to the other nodes having similar power, while disassortativity means that nodes with higher degree attach to the ones having the lower degree.

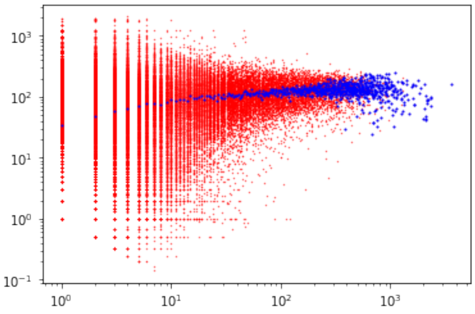

Figure 5: Outdegree assortativity

From the plot, we see that the nodes are assortative based on their outdegree.

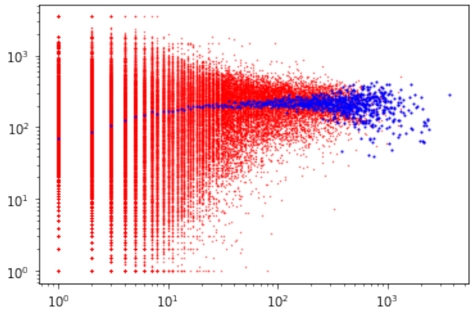

Figure 6: Indegree assortativity

However, the ingoing edges tend to be slightly disassortative. That is not a surprise here, as edges with the high ingoing degree will tend to connect to the edges with lower indegree, simply because they don't have too many edges that are similar to them. The values of degree assortativity coefficients are 0.0051, and -0.0064 respectively, so our assumptions hold.

Rich club effect We want to see how much interconnection there is between the most trusted edges. Due to the size and the power-law nature of our network, we decided not to use the percentages and observed only the top hundred nodes. We have investigated densities for both signs.

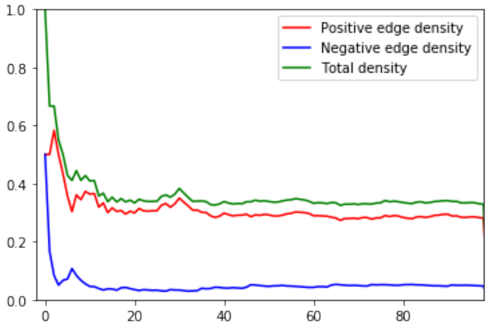

Figure 7: Rich club effect in popular vertices

From the plot, we see that we don't have a rich club effect. After the ten most popular edges, the density stays around 0.4. Since density is close to one half, we want to check is the connections tend to be one-sided or random pair of nodes follow each other. The reciprocity for this subgraph is 63%, so we conclude that popular edges usually aren't connected, but when they are, the following is reciprocal. One of the interpretations is that we may have that these popular nodes are from different groups, and therefore do not communicate in the network. From the plot, we have also seen that some of the most trusted members distrust other highly trusted members.

Negative rich club This may seem a bit uncommon, but we also wanted to check if there is some relationship between edges with the highest negative indegree. Imagine that in our social network we have a certain group of people with opinion deviating from consensus (think of conspiracy theorists or a group of internet trolls), that is generally disliked, but very popular among their group. By checking the rich-club effect between such nodes, we can check the existence of such a group.

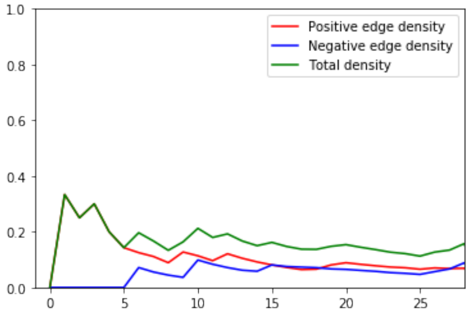

Figure 8: Rich club effect in distrusted vertices

As the top four most distrusted vertices were not connected between themselves, we removed them and plotted another graph. However, we see that there is not a high interconnection between distrusted nodes (there is some level of trust, but not significant). And we see that it turns fast into distrust. So, the distrusted members of this site tend to operate like isolated islands.

Homophily As this network is a trusted network, given without distinctive groups, there are no ways to put this data into meaningful groups that will tell us more about whether homophily is present or not.

Community structure Our graph didn't come with any previous information about the structure of the underlying communities in the graph. So, we used the Louvain algorithm to detect them. Its advantage over similar algorithms is that it is very fast, even for big networks. Our main goal was to find the most central communities, so we have observed the underlying undirected graph of the biggest SCC (the edge weights were ignored, as distrusted members are still the part of the community).

Louvain algorithm yielded a small number of big components and a huge number of very small communities. In total, eleven communities had at least 100 members, and they were of the following sizes: 10032, 10540, 453, 11587, 221, 143, 945, 311, 111, and 124.

So there are three big communities, each taking approximately 25% of the biggest SCC. Since this is a big network, there was a possibility that the algorithm will fail to detect the small networks (resolution limit). However, as we obtained the total modularity equal to 0.43, we are satisfied with the obtained results.

Structural balance

We are working with the signed network, so it is natural to search for the structural balance property. We know that a graph is balanced if it contains no cycle with an odd number of negative edges. Due to the size of our graph, that behaviour can't occur, but we can search for other interesting statistics. One of them is to examine weak structural balance. We are interested in the proportion of triangles that have one negative edge, as they will be unbalanced. Due to computational complexity, we were not able to produce the results, so we will present results from the original research on this dataset [J L10]. In our dataset, authors explored the balance of triads which form a triangle, ignoring the order of edges of the assumption. They obtained the following results:

Triad type	Total count	Percentage
+ + +	11.640.257	0.870
+ - -	947.855	0.071
+ + -	698.023	0.052
- - -	89.272	0.007

In network theory, we can assume that we have an approximately weak balance if the number of triangles with all the positive edges much greater then it would be if it was obtained randomly (random expectation was based on the random distribution). The results are given in the table below.

Triad type	Perc.	Random Exp.
+ + +	0.870	0.621
+ - -	0.071	0.055
+ + -	0.052	0.321
- - -	0.007	0.003

We will not state the advanced assumptions of the structural balance theory at this time, so we will just note that the all-positive triangles are overrepresented compared to a random chance and that the all-negative triangles are underrepresented compared to a random chance.

Balance for reciprocal triads Structural balance theory is usually viewed in the context of undirected graphs. We are looking for a way in which we can view the structural properties in the context of directed graphs. We have looked at the triads irrespectively of their order, but we did not focus on the case when the edges in the triads were reciprocated. However, reciprocated edges in triads are not a dominant mode of triads creation. The total proportion of triads having reciprocated edges was 3.1% per cent, and the following results were obtained:

	Triads	P(RSS)	P(++)	P(−)
Bal.	348.538	0.929	0.941	0.688
Unb.	74.860	0.788	0.834	0.676

In the table, P(RSS) denotes the probability that reciprocated edges have the same sign, P(++) and P(−) probabilities that positive and negative edges are reciprocated with the same sign. We see that it balanced triads, it is rare that trusting (positively labelled) relationships aren't reciprocated, and it holds in the majority of unbalanced graphs. While for the negative edges, reciprocation isn't so strong. This is indeed a sign of social selection.

Behaviour adoption

We are interested in how would the adoption of a certain behaviour look in a trusted network. For example, suppose that this is a network of doctors and that the new cure has appeared. In the beginning, some number of the highly trusted medicians will start thinking that this cure is good (we assume that by being more trusted, they will be to some degree more knowledgeable about medical advances). Also, some highly distrusted medicians will adopt the same opinion. But, the nodes in the network don't have the information about the popularity of other nodes. So, they will adopt a certain opinion only if a certain rate of people they trust will adopt it as well. Also, adoption of behaviour by an adopted person will negate the effects of a behaviour adoption of a trusted node. We assume that once an opinion is adopted, it won't be changed (otherwise the person could switch opinion because people that distrust would negate it). In that way, we enforced monotonic spread (the cure may work and people will not check the behaviour of people they distrust). To produce these results, we set the same random seed. We assume that initially, a set of 200 out of 300 most trusted people adopts an opinion, and at the same time 50 of 100 most distrusted people will adopt it (they may be as knowledgeable, but are distrusted for a variety of other reasons).

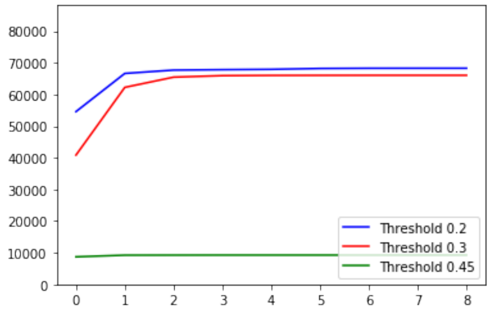

Figure 9: Adoption of an opinion.

From the plot, we see that for smaller thresholds, the opinion adoption will be fast, while for the bigger ones, it won't spread too much. However, we want to see more interesting dynamics. We produced them by setting the threshold at 0.4.

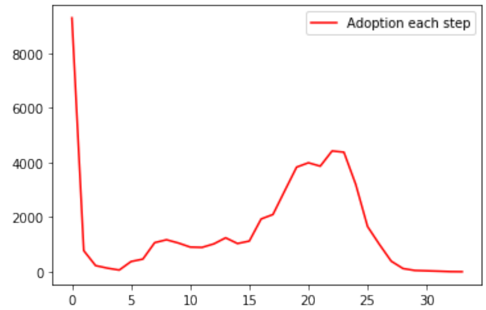

Figure 10: Number of adoptions for each step.

We see that at some point, people have almost stopped adopting opinion, but at some point, it started to go up. We know that opinion spreads quickly, due to the power-law structure, but slows down later. That happened when it reached a dense cluster, and after spending time on its border, it started expanding. On average, it took 17.12 time units to adopt an opinion! That is half of the duration of the whole process. Also, due to the power-law structure, we may attribute the big initial spread of opinion to the fact a sizable minority of nodes mostly follows only popular nodes, therefore making them more ready to adopt an opinion. Next, we look at the evolution graph.

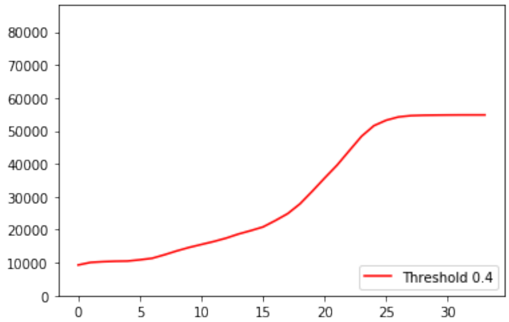

Figure 11: Evolution of behaviour adoption.

We are interested to see how the other structural properties enforce the adoption of an opinion. It seemed logical that the opinion won't spread in the OUT component, and that is true, since only 26% of its members have adopted a new opinion.

The next thing we checked was the adoption of the opinion in the found communities. We want to see how much has the opinion spread through them. We observed six smallest communities containing at least a hundred nodes.

Size	Adoption rate
453	0.26
311	0.35
221	0.66
143	0.39
124	0.38
111	0.14

Table 3: Adoption rate for small communities.

We see that inside these communities there are dense clusters that stop the adoption of an opinion. That further shows that this dataset contains a certain number of small heterogeneous communities. This is one of the reasons why we don't have the information cascades in this case. Some other reasons for that are:

- Some nodes have no outdegree, therefore they can't have a base of trusted users.
- Nodes being outside the biggest weakly connected component are isolated from the main events.
- Nodes that are way more distrusting than trusting (like in real life, some people only focus on negative aspects).

Conclusion

We have seen that the Epinions network exhibits typical behaviour of a social network: sparsity, power-law structure, and high clustering coefficient. We have analyzed the pair plots of different degree types and their correlations and gained valuable insights. We have shown there is social selection, and that we do not have a rich club effect. We have isolated 11 noticeable communities and checked the behaviour adoption in the smallest of them. As trust is based on the quality of the information received, we compared it's structure to the general web structure, and we have shown that they are similar. We have also checked the adoption behaviour in the OUT community and found that it is mostly isolated from outside influence.

Regarding structural balance, we have seen that this network is approximately balanced and that this balance gets even stronger in the triads having reciprocated edges.

Simulations of the opinion adoptions have shown how the network reacts to the adoption of opinion with respect to the different thresholds, and how it spreads through components and communities.

References

[J L10] J. Kleinberg J. Leskovec D. Huttenlocher. *Signed Networks in Social Medias.* 2010.

[Les] J. Leskovec. URL: `https://snap.stanford.edu/data/soc-Epinions1.html`.