Ukrainian Catholic University

Faculty of Applied Sciences

Data Science Master Programme

Reddit submissions graph analysis

Network science final project

Authors:
Andrew Kurochkin

26 June 2019

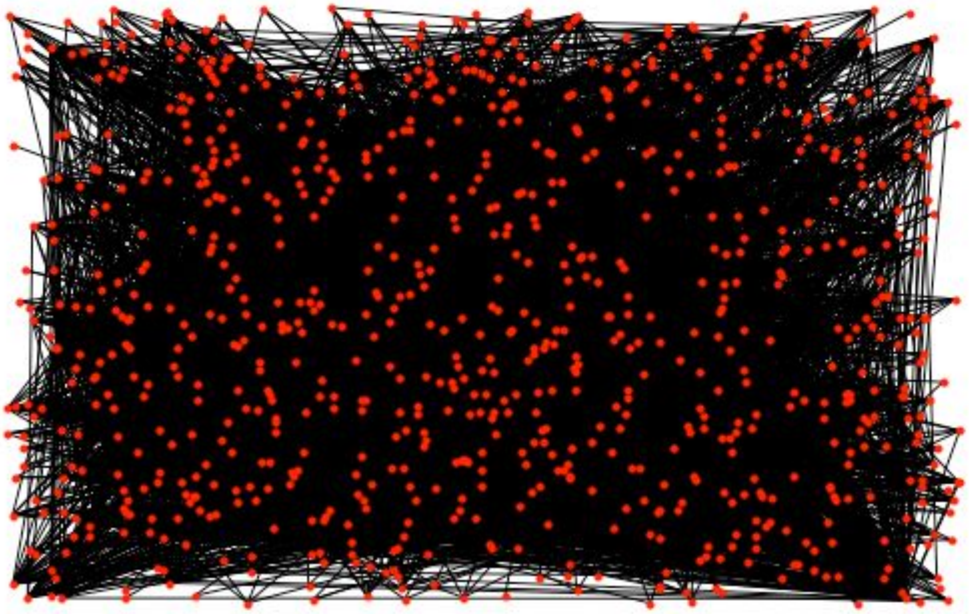
1 Introduction

In this project Reddit content aggregation site was analyzed from network point of view.

Graph is based on a collection of 132,308 reddit.com submissions dataset. Each submission is of an image, which has been submitted to reddit multiple times. For each submission, were collected features such as the number of ratings (positive/negative), the submission title, and the number of comments it received. [1]

In my definition of the graph a **node** is a single subreddit (user-created board) where the image had been submitted, a **link** is a fact of submission in both subreddits.

Total **number of nodes** is 868, total **number of links is** 5321.



Picture 1. Reddit submissions graph (random layout).

2 Reference network

Due to the fact that the nature of the **Erdos-Renyi graph** almost perfectly shows the nature of the random network, Erdos-Renyi (ER) graph was chosen for comparison. Key idea of this work is to compare different characteristics between Reddit submissions (RS) graph and ER to extract inferences about first one.

For the purity of the experiment setting with the same number of nodes and links was chosen as the best scenario. Selected ER graph has a **total number of nodes** equal 868 and a **total number of links** equal to 5321 as our main exploring graph.

3 Global network characteristics

All main network properties are described in the Table 1 below.

Characteristic name	Reddit graph	Erdos-Renyi graph
Total number of nodes	868	868
Number of links	5321	5321
Density of links	0.01	0.01
Average node degree	12.26	12.26
Min node degree	1	3
Max node degree	743	25
Average shortest path length	2.06	2.95
Diameter	4	5
Number of connected components (CC)	1	1
Size of the largest CC	868	868
Average clustering coefficient	0.89	0.01
Global transitivity	0.07	0.01
Assortativity	-0.41	-0.01
Average degree of the nearest neighbour	374.91	13.24
Average betweenness centrality	145.79	209.09

Table 1. Global network characteristics.

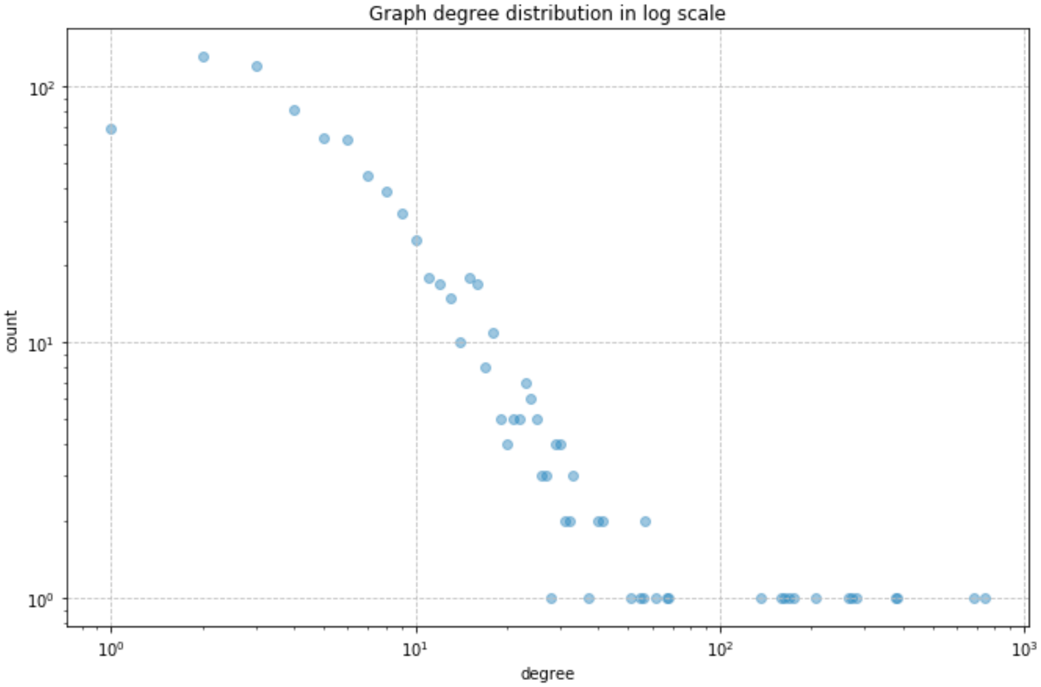
Analysis of the present table was essenced as next:

- Max node degree in the RS graph is much bigger than for ER, it shows us that our network has at least one significant “central” node.
- Shorter average shortest path length in the RS represents the nature of content aggregation communities nature, so in this system, actors are closer to each one.
- Average clustering coefficient helps to figure out the fact that subreddits can be clusterized better than a random graph.
- Another fascinating insight is that RS graph is less assortative than random, it can be explained, as more prominent communities have a lower probability to have connections between themselves.

- We think that the average degree of the nearest neighbor is much higher for the RS because it has hubs (nodes with high degree).
- The fact that ER graph has a bigger average value for betweenness centrality can be illuminated. We think that reason is in the nature of Reddit communities, and they have a high level of diversity in the interests and content.

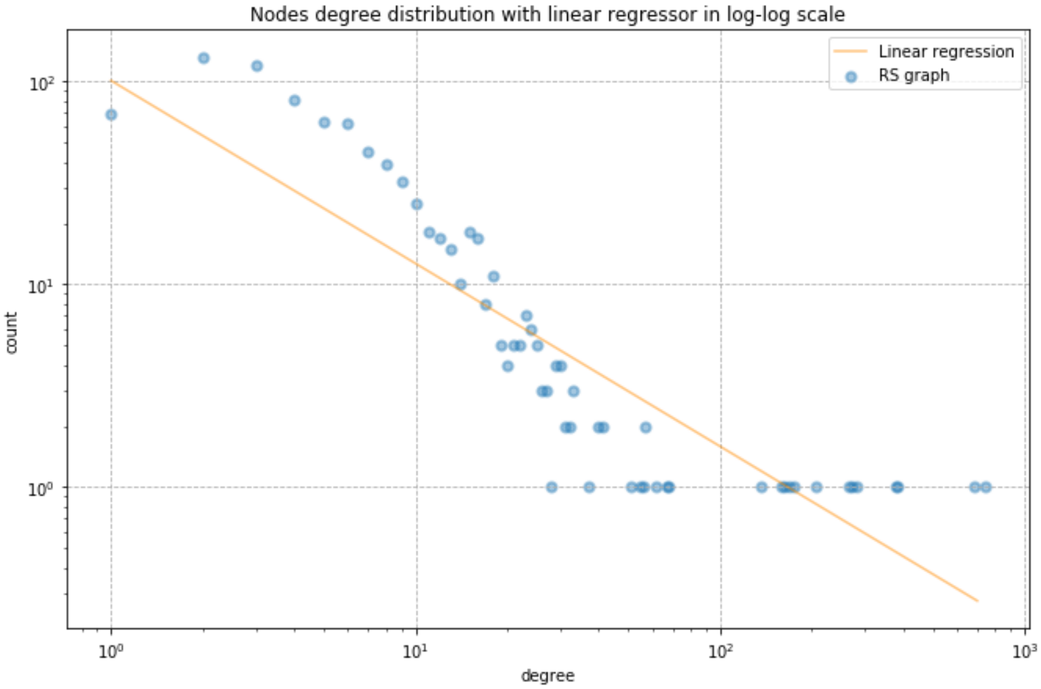
4 Distributions

4.1 Distribution of node degrees

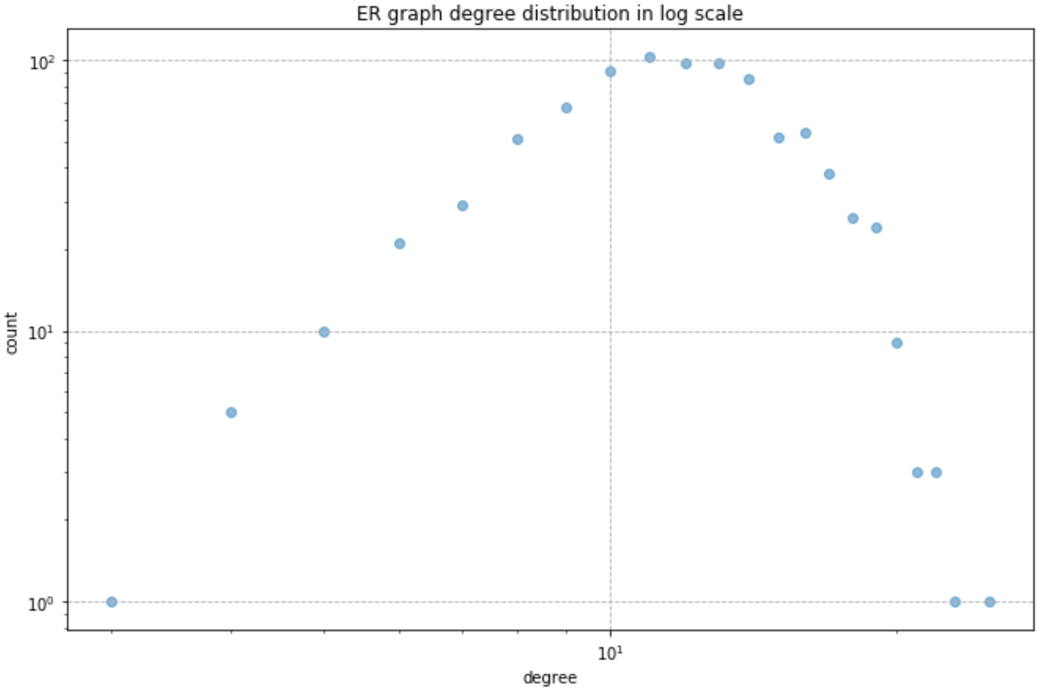


Picture 2. Node degrees distribution for RS graph.

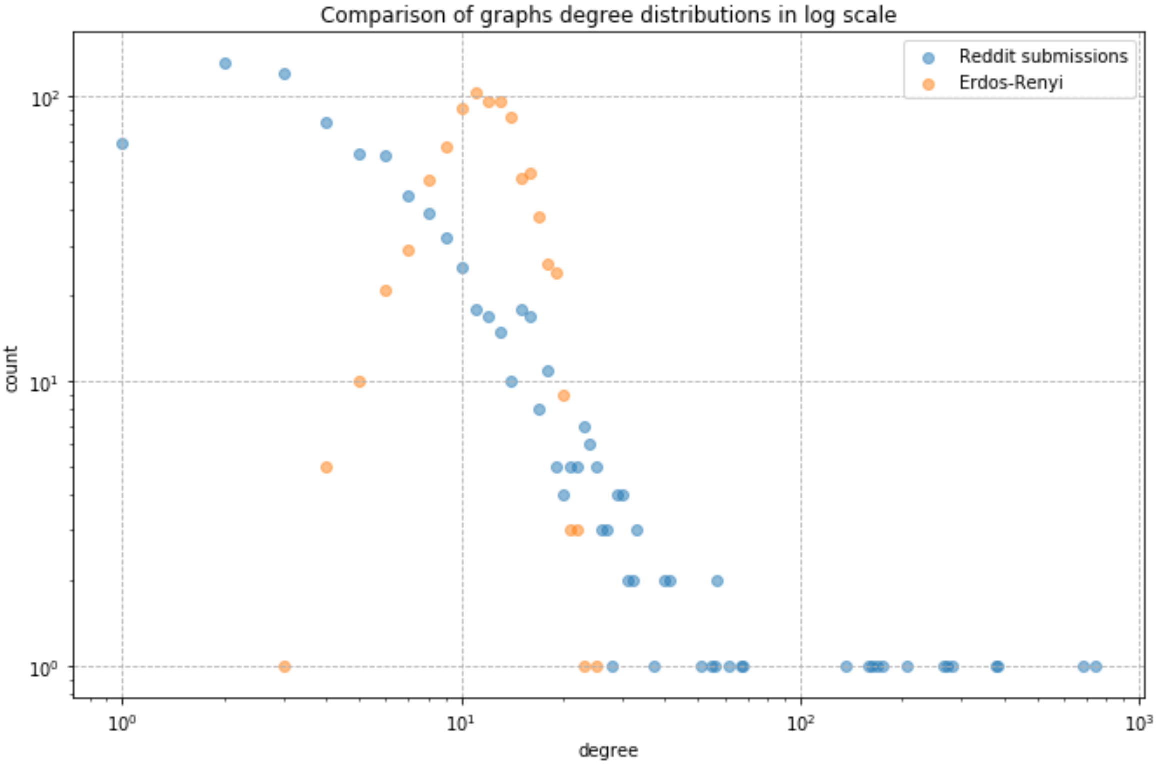
Exploratory data analysis showed that the main part of the observation can be fitted with power-law function, but right tail will add the noises to the accuracy of the approximation function. Exponential function is not applicable for our degrees distribution.



Picture 3. Node degrees distribution fitted with power law, $\lambda =- 0.9$.



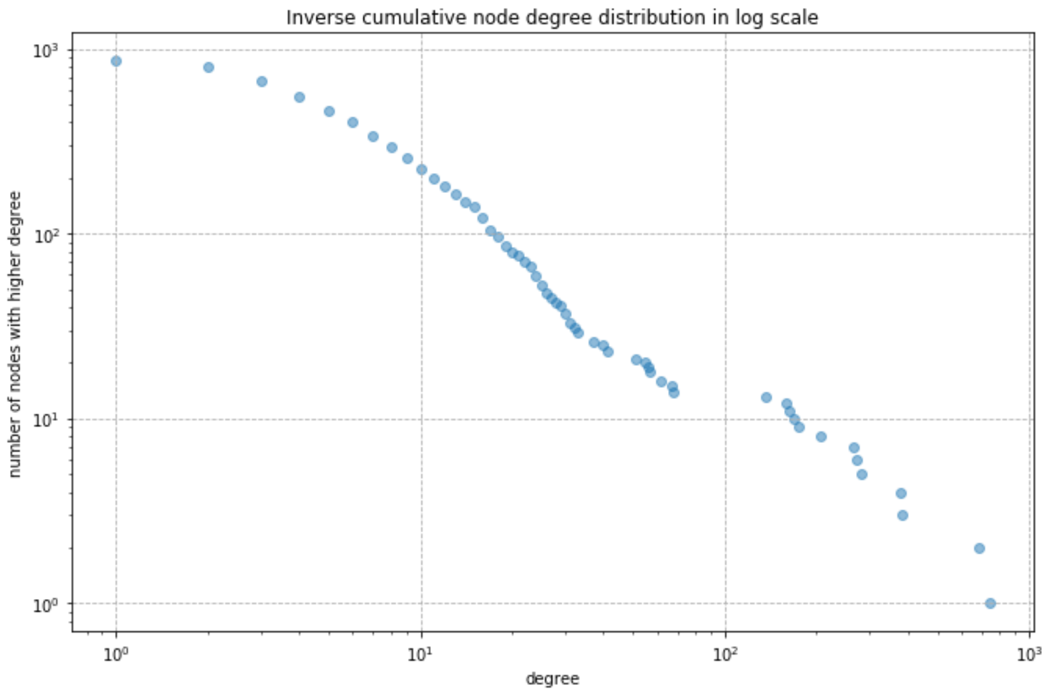
Picture 4. Node degrees distribution for ER graph.



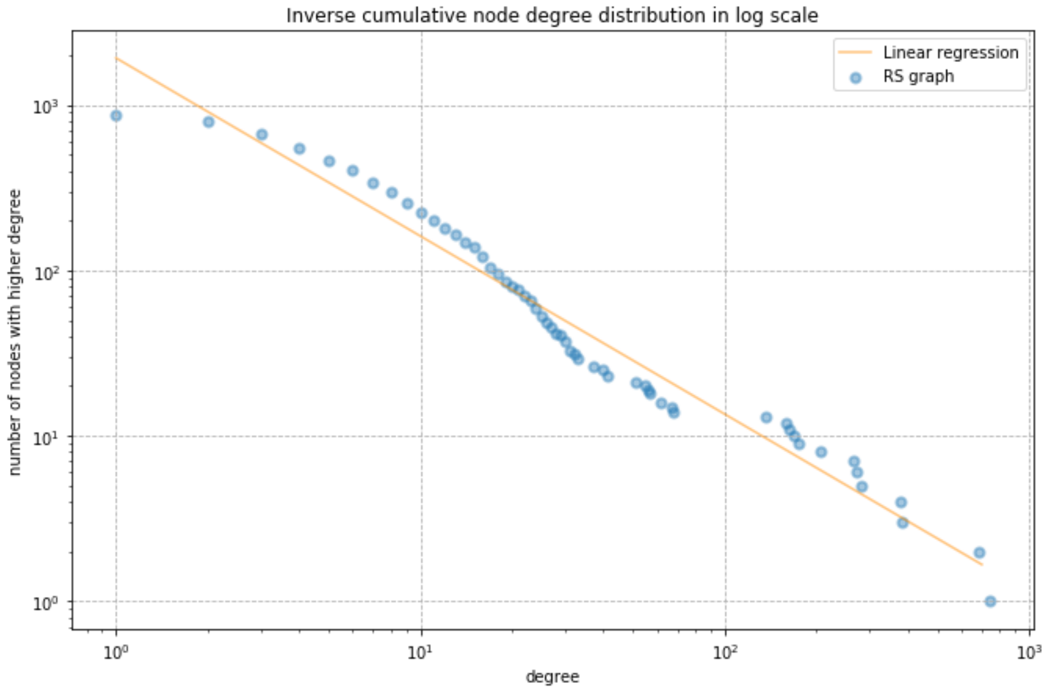
Picture 5. Comparison of degree distributions for graphs.

Reddit submissions graph shows a picture that is very typical for a scale-free networks, while Erdos–Renyi graph has different nature.

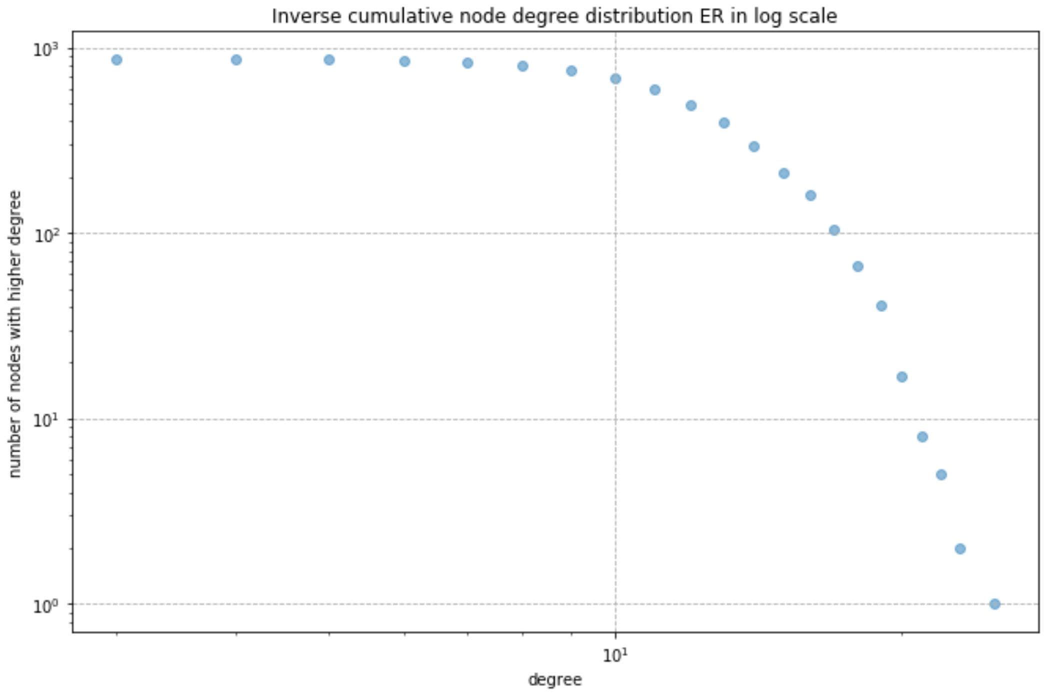
4.2 Inverse cumulative node degree distribution



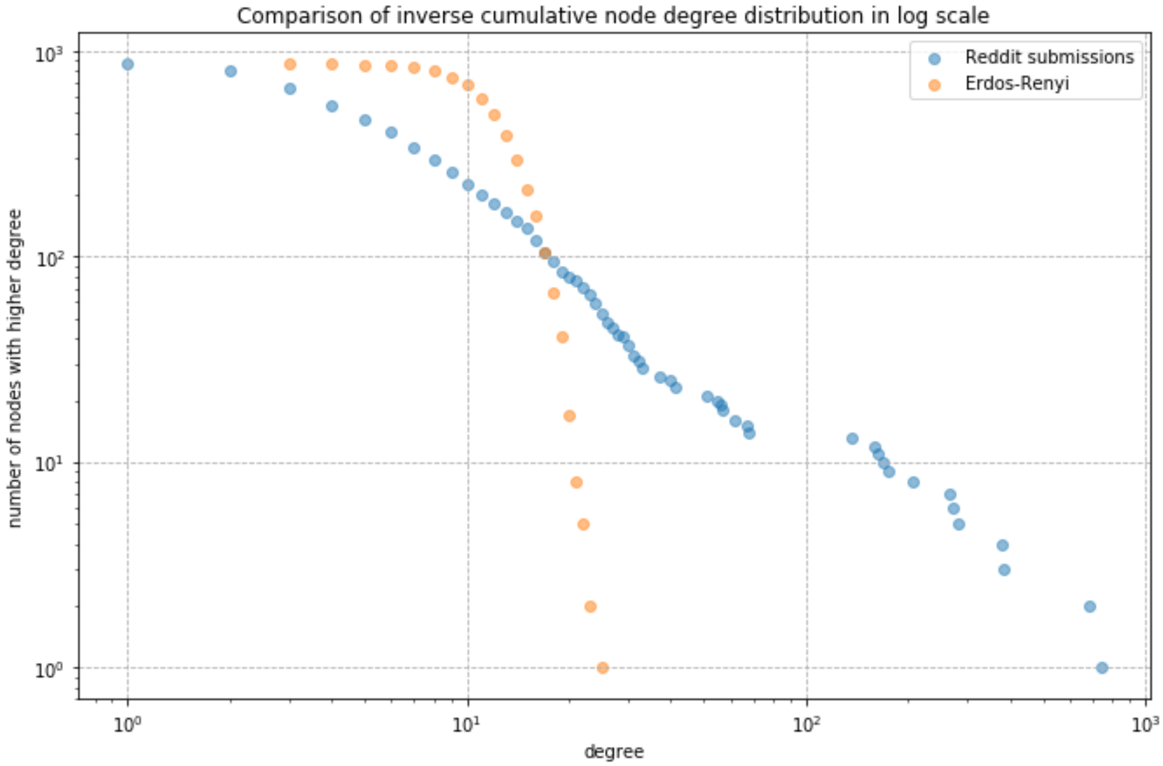
Picture 6. Inverse cumulative node degree distribution.



Picture 7. Inverse cumulative node degree distribution fitted with power law, $\lambda = -1.07$.



Picture 8. Inverse cumulative node degree distribution for ER graph.

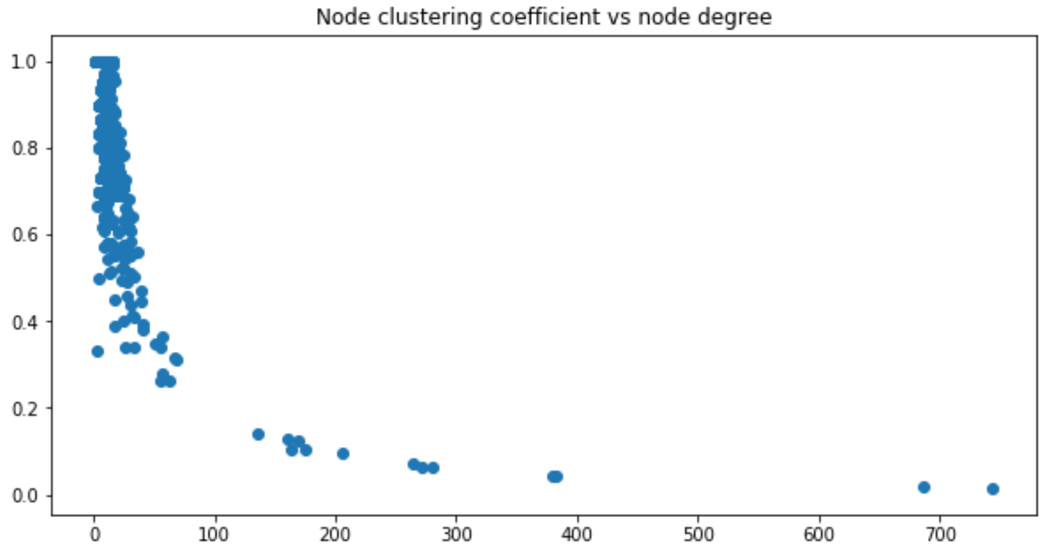


Picture 9. Comparison of inverse cumulative degree distributions for graphs.

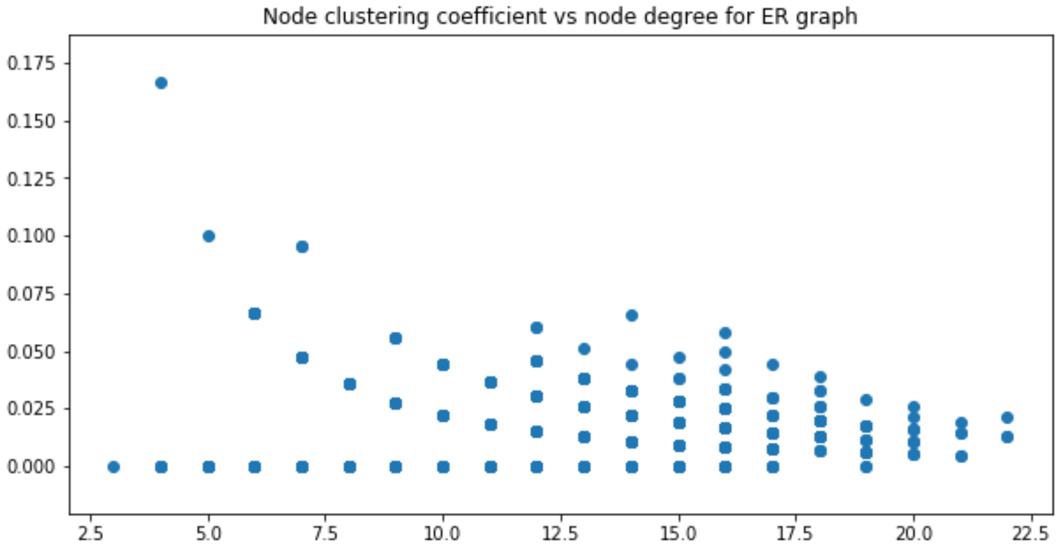
The degree distribution captures only a small amount of information about a network, however, this information gives important clues into the structure of a network. In our case, we can conclude that the network has power law distribution and it has long right tail.

5 Correlations

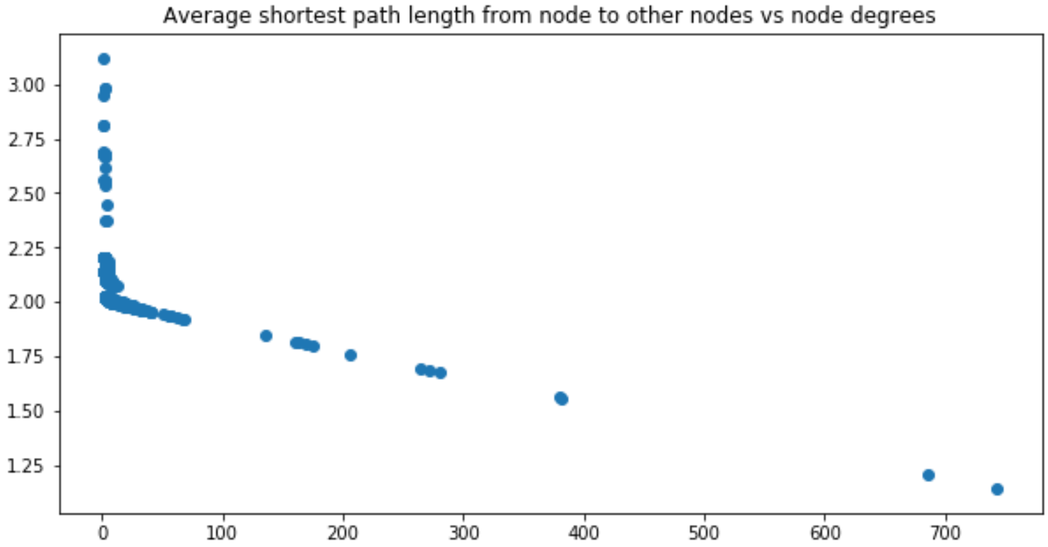
5.1 Clustering coefficient of a node vs node degree



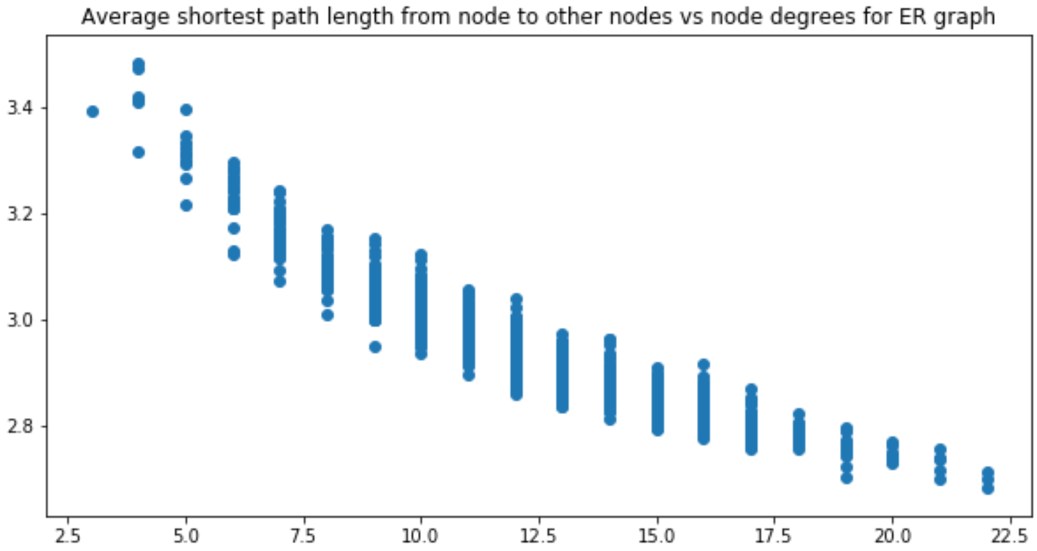
Picture 10. Node clustering coefficient vs node degree.



Picture 11. Node clustering coefficient vs node degree for ER.



Picture 12. Average shortest path length from node to other nodes vs node degrees.



Picture 13. Average shortest path length from node to other nodes vs node degrees for ER.

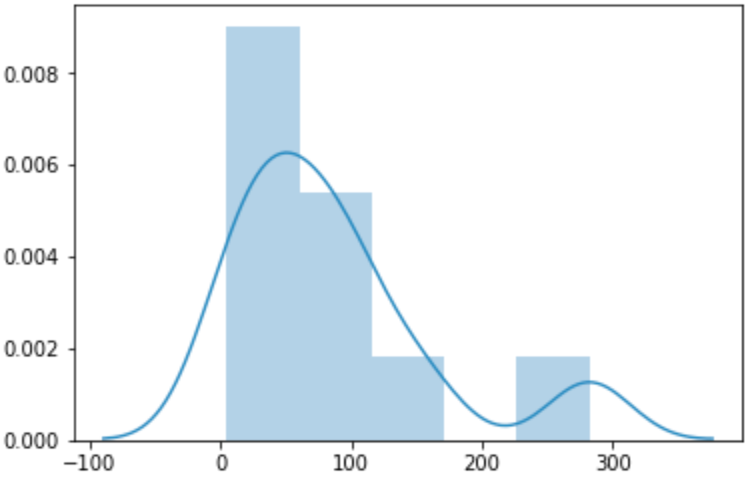
Correlation analyses helped to become more confident in the statement that the clustering coefficient is higher for a significant amount of nodes in the RS graph, and the average shortest path length is less for it. Both these things tell us about social-content nature of the structure that extremely differentiates from the random network.

6 Community structure

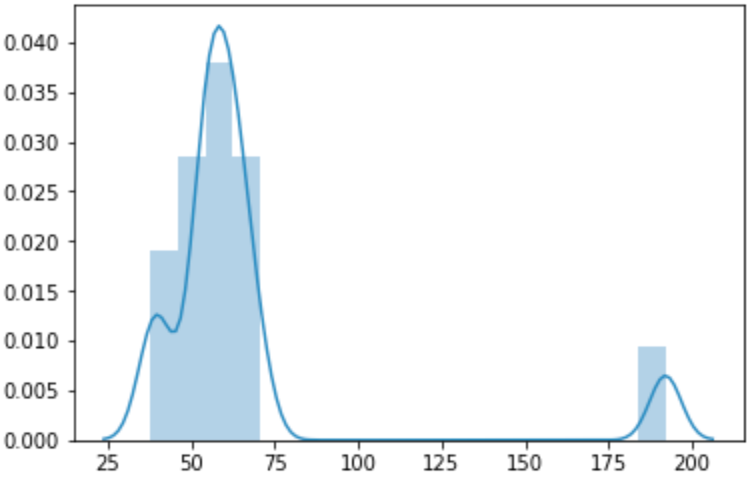
Characteristic name	Reddit graph	Erdos-Renyi graph
Modularity of the optimal partition	0.24	0.25
Number of clusters in optimal partition	10	13

Table 2. Community metrics characteristics.

6.1 Distribution of cluster sizes



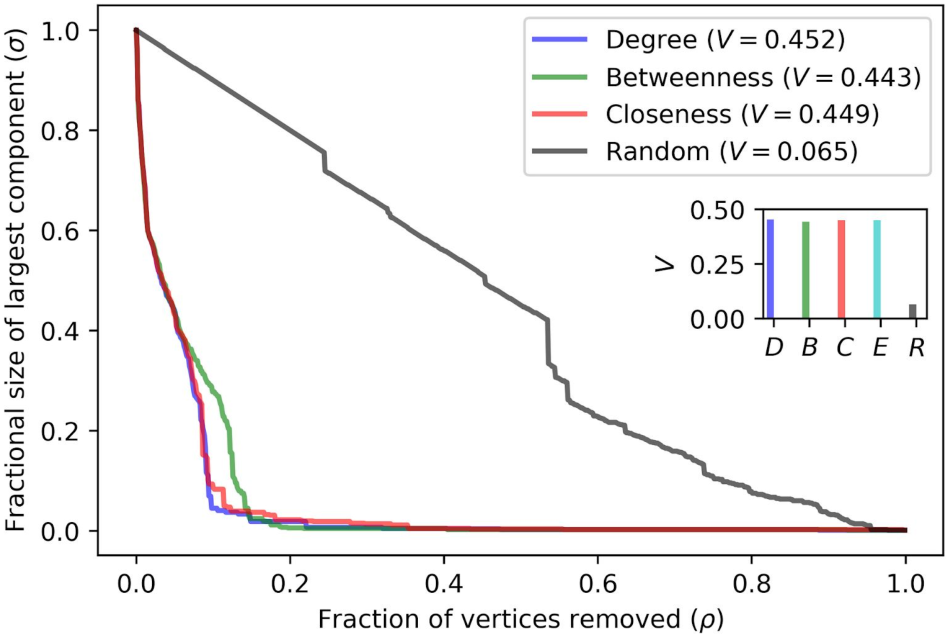
Picture 14. Distribution of cluster size for RS graph.



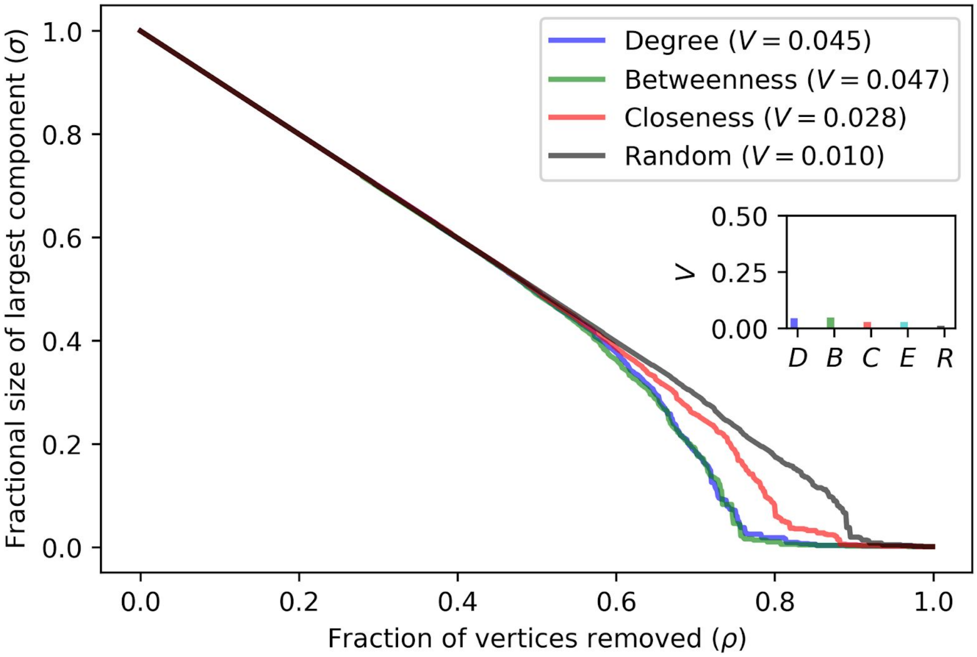
Picture 15. Distribution of cluster sizes for ER graph.

This part of analyzes helped us to extract information that social site submissions tend to have clusters with a higher number of the elements inside than the random graph with the same number of nodes and edges.

7 Network Stability



Picture 16. Quantitative network robustness analysis for Reddit submissions graph.

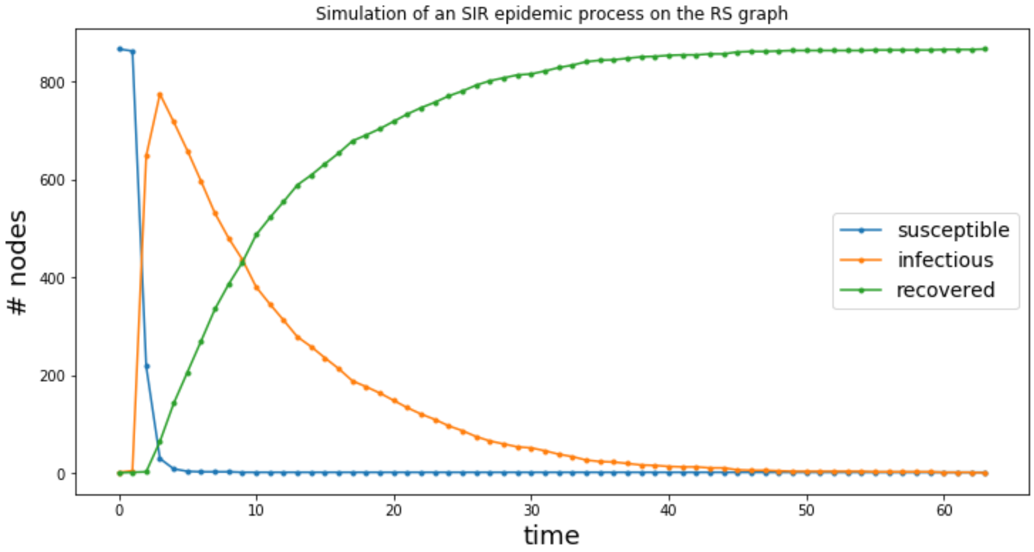


Picture 17. Quantitative network robustness analysis for ER graph.

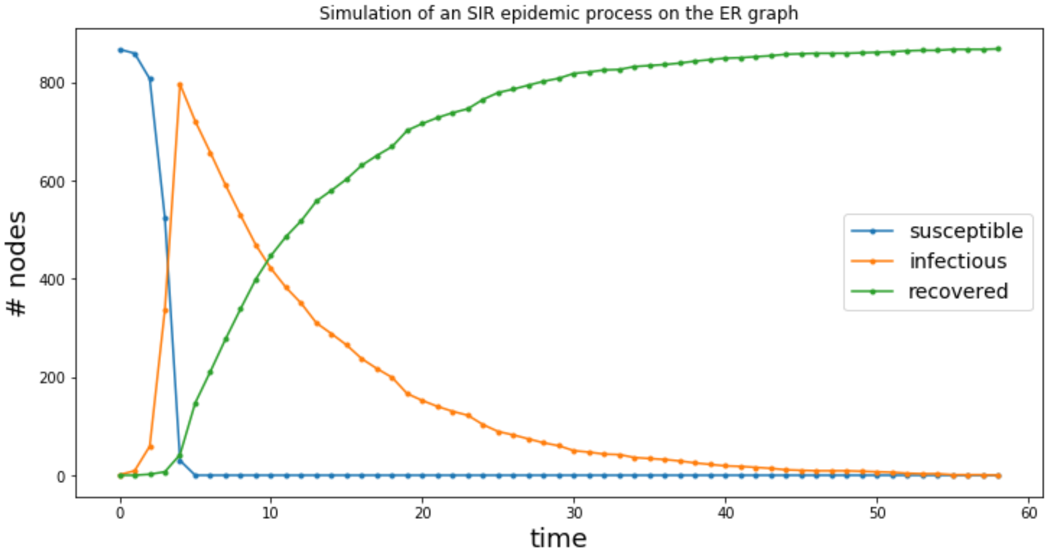
Quantitative and qualitative network robustness analysis under different multiple failure scenarios showed that it’s enough to remove ~20% nodes with a higher value of the importance metric such as node degree, betweenness centrality or closeness, to crucially damage the network structure, we believe this causes from the nature of the scale-free graph. On the other hand, the random graph showed some critical behavior only after ~60% of the nodes had been removed.

8 Spreading Processes

8.1 Fraction of infected nodes vs time without vaccination



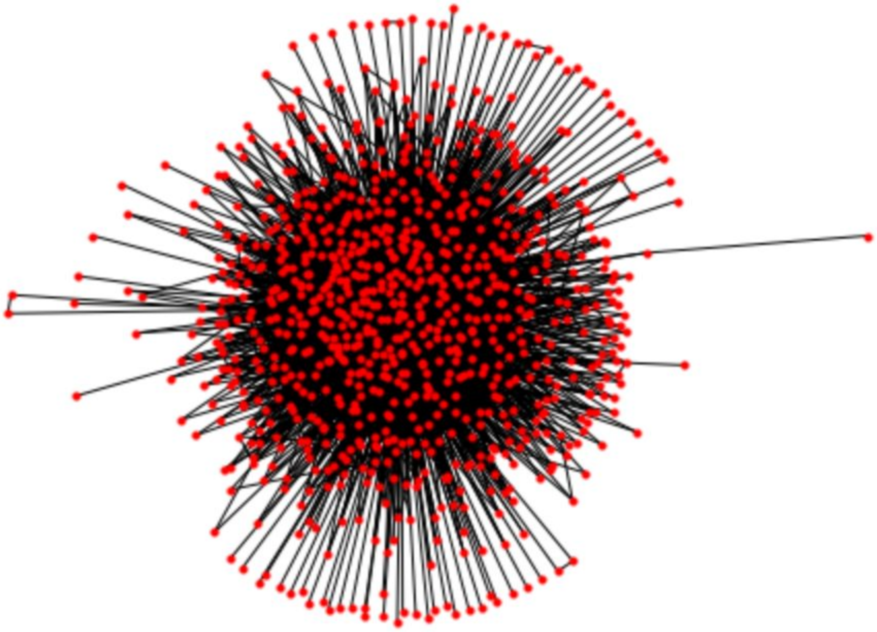
Picture 18. Simulation of an SIR epidemic process on the RS graph.



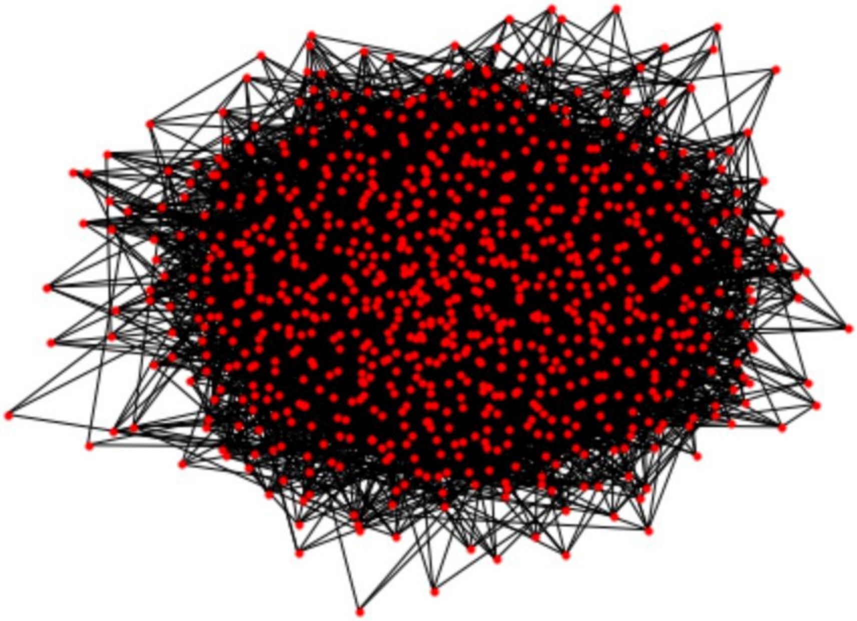
Picture 19. Simulation of an SIR epidemic process on the ER graph.

Simulations were run with parameter beta equal to 0.6 and mu equal to 0.1. After a few iterations of running simulations (as it is a stochastic process), we observed that difference between RS and ER graph isn’t significant, so we can not extract information about unique network patterns with this approach.

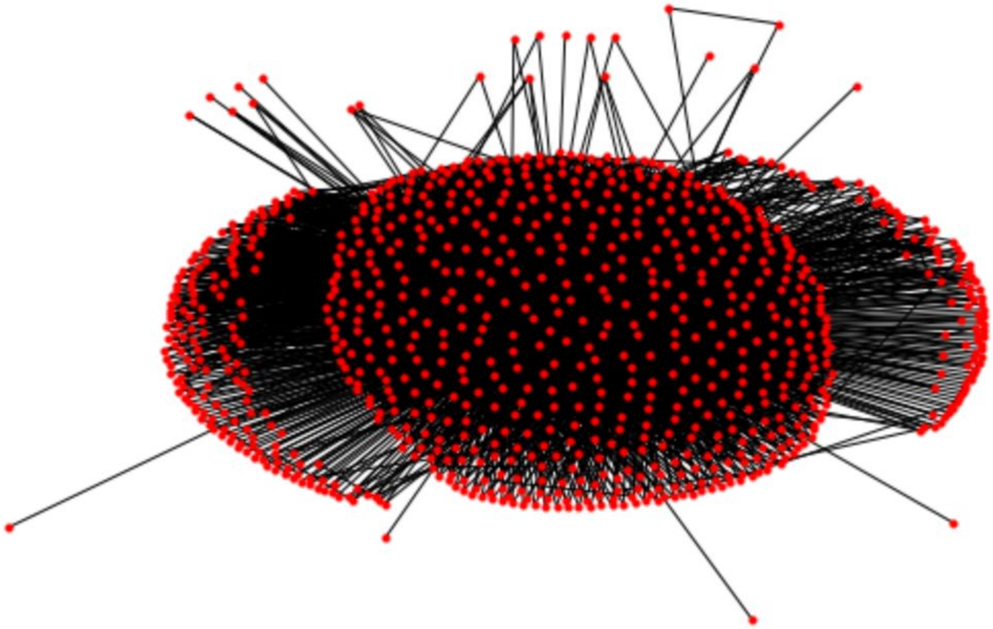
9 Visualization



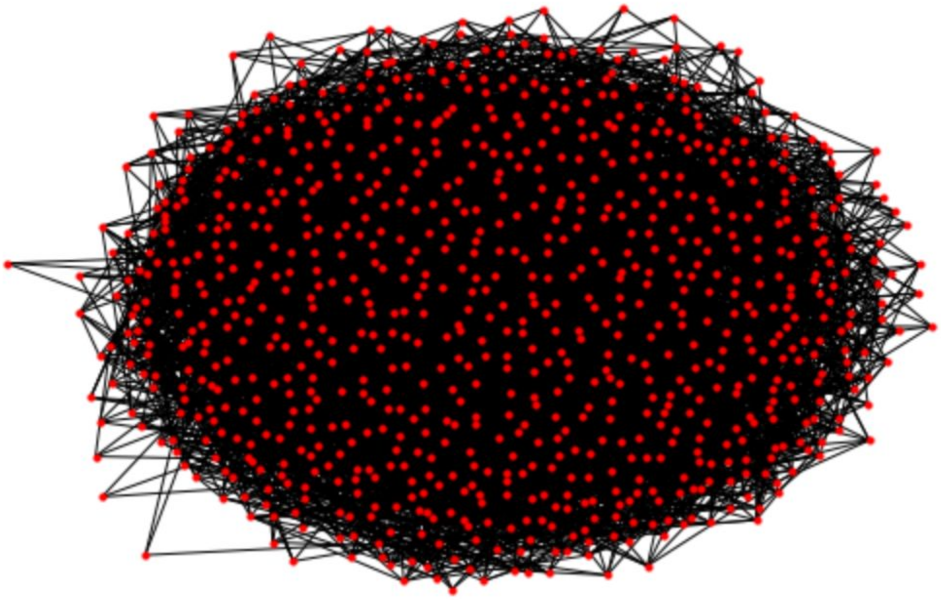
Picture 20. RS graph plotted with spring layout.



Picture 21. ER graph plotted with spring layout.



Picture 22. RS graph plotted with Kamada & Kawai layout.



Picture 23. ER graph plotted with Kamada & Kawai layout.

During the visual comparison of graphs, the difference between random and Reddit graph was significant. Kamada & Kawai is an algorithm for drawing general undirected graphs. The graph represents a dynamical system that tries to reach a minimum-energy state. On the Picture 22 result of the algorithm work can be observed.

10 Conclusion

The proposed approach of representation for Reddit submissions as graph helps to find a bunch of inferences. Graph-specific nature was highly remarkable during this research. Reddit is the union of thousands of different communities, even though they can be clustered in a few different groups, length of the information spreading is pretty small inside this system. RS graph shown properties of the scale-free graph, that was expectable. Important founding is that this network has strongly negative assortativity, as there are hubs - places of the content concentration, that don’t have an inheritance to have a close relationship between them.

References

1. Web data: Reddit submissions, http://snap.stanford.edu/data/web-Reddit.html.