A Project report submitted

on

Searching for superspreaders of information in real-world social media

Sen Pei, Lev Muchnik,José S. Andrade,Zhiming Zheng & Hernán A. Makse
(Published - July 3,2014)

By

Priya Anubhuti Tiru (17BCS021)
Riya Gupta (17BCS026)
Department of Computer Science and Engineering

Under the guidance of

Dr. Yayati Gupta
Assistant Professor
Department of Computer Science and Engineering

PROBLEM STATEMENT AND SCOPE

Information in today's world is one of the most important assets. It can be used as a tool to become rich, as a secret weapon in wars, as a shield against outbreak of epidemics, and the use of information is everywhere. The most astonishing feature of information is that it creeps in out of nowhere and we are often unaware of how a piece of information from us, reaches to others and from others reaches us. People think there has to be edges, here connections between the nodes, here people for a piece of information to travel among them, but fascinating as it looks there are cases where information have reached from a source node to nodes which had no connecting path reaching them. Therefore, the concept of information diffusion network underlying the network of connections among the nodes becomes vital.

The influence of a node in the spreading of information can be seen as the power this node has to impact a number of people in good or bad ways. If we want our information to reach a huge number of people, where we want these people to be from the set of target people to whom we want this information to reach, we would want a node with high influence to support us in spreading information. Hence, finding such highly influential nodes from a network that is huge and so dense, becomes a difficult task.

Seeing the present scenario of COVID-19 where it seems to have the ability to cross all the boundaries and spread rapidly; finding the most influential spreader has been the most pertaining task in order to prevent the spread of this pandemic and monitor the outbreak. However, with a densely populated country like India, it becomes quite challenging. This in many ways seems similar to finding the influential super spreader in the real world social network where the network is dense and the spread of information is quite incomplete.

This work, however, focuses on the spread of information or rather diffusion of information in the real world social networks such as a blogging site-LiveJournal, social networking sites-

Facebook, Twitter and many more and find the most influential spreader in the network. We gave an overview of methods currently modeled to solve the problem statement.

IMPORTANCE AND PREVIOUS TECHNIQUES

Several methods have been employed in searching the super-spreader nodes in a network that is voluminous, highly interconnected and also directed as it is in real networks. From real networks, however it was very difficult to get a complete and accurate information diffusion network. Among the various different methods aimed to find the influential spreader, the most well known includes degree, Page rank, Betweenness Centrality and K-core(also called K-shell). However drawbacks of some previous studies of predictors have been done by modelling the spread of information rather than using the real spreading dynamics. Particular models for the same include Random walks for Page Rank, susceptible-infectious-recovered (SIR) and susceptible-infectious susceptible (SIS). There were contradictory predictions concluded by simulations of different models. In the simulation of the SIR and SIS model, K-core outperforms the other measures(degree and centrality).

In the paper, searching for superspreaders of information in real-world social media, the researchers collected a real-work network with full information diffusion network with a large number of nodes interconnected with directed edges, the dataset represented public blog posts published at LiveJournal.com (LJ). In this research, importance given to the 'information' was more than to 'information-carriers'. This paper also addresses that the search for influential spreaders is done by following the real spreading dynamics.

There were basically four algorithms used to solve the problem which includes degree, which stated that the nodes with highest degree or the people with more number of connections may help in easy and faster spreading of information. Next algorithm used was Random walk for Page Rank which was basically used to rank the large array of data and accordingly the one with more rank(containing more hyper links) would be the one responsible for spreading information. It was also addressed that the K-core was used to identify the location of the person in the network. The one with high K-shell value was prone to be present in the core of the network and the one with lower K-shell value was prone to be present at the periphery. The K-core was found

to be more reliable, outperforming the Page Rank as it did not only predict the average influence of nodes but also helped in identifying the topmost super spreader more accurately.

Further with the incompleteness of the datasets such as that of the twitter dataset, K-core was modified on the local network information-K_{sum} and K_{2sum} which outperformed Kin and Pagerank and therefore it was later concluded that the Ksum can be used to search for influential super spreader.

CONTRIBUTIONS

While the paper motivated us in many different aspects, one thing which fascinated us to contribute to this was the study of information spreading the real world social network.

We at first collected the wikipedia vote network from Stanford Large Network Data Collection. The dataset consisted of 7115 nodes and 103689 edges which were directed. Then we followed our work with the cleaning of the data set. Thereafter we identified the K-core nodes of the network and found the influence by implementing the SIR model into the network. For the comparison of algorithms, we implemented the Random Walk for Pagerank to identify the super spreader using rankings. Later, we found the average influence of the K-core and it's logarithmic values and plotted the comparison results. We also compared the variation of the influence of the nodes within fixed measure intervals.

For the comparison again, we collected another dataset of social networking site - Twitter from Stanford Large Network Data Collection. This dataset had 81306 nodes and 1768149 edges which were directed too. Moreover, this dataset was an information diffusion network. Out of 1768149 edges, 1048564 were able to be loaded in our system. We made a network using the edges and obtained the node IDs, after which we found out the k-core (ks), k-degree (k-in) and pagerank (pr) values for each node following which we found out their influence over the network.

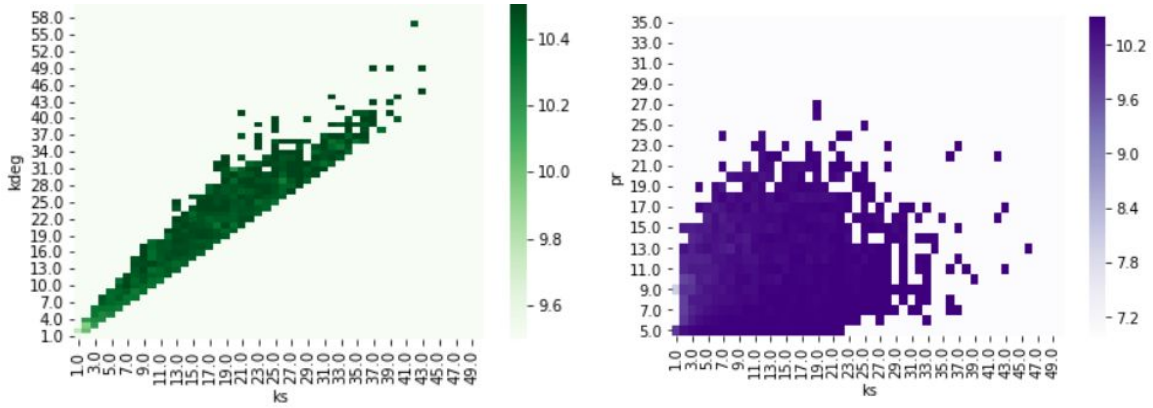
We then found out the average influences for combinations of ks and k-in over the range of ks and k-in and plotted the same on a heatmap. The average influence, M(ks,k-in) for a pair of ks

and k-in was a summation of the influence value of the nodes that had this pair of ks and k-in averaged by the number of such nodes after which we obtained their logarithmic values. Similar work was done for ks and pr combinations. We then divided the range of ks and divided nodes into five bins and found out the standard deviations among the influence of nodes in a bin, in a similar manner we calculated for kin and pr and these values were plotted using cluster bar plots.

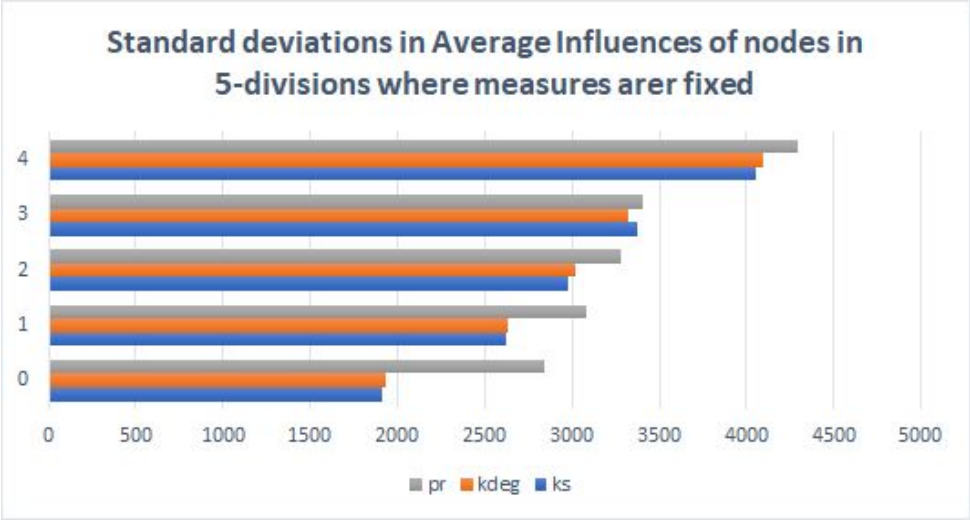
For each node, ksum i.e., sum of k-in of neighbor nodes was obtained and then k2sum values, sum of k-in of neighbors of neighbors. Average influences were evaluated for combinations of k-in and ksum values, k-in and k2sum values and pr and ksum values.

ANALYSIS AND RESULTS

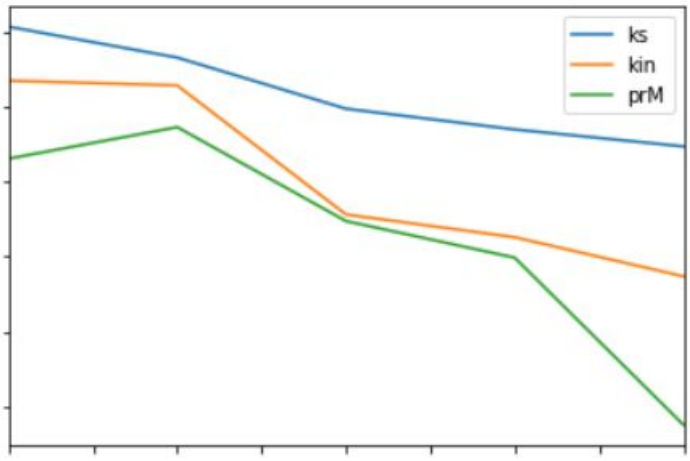
Since the twitter dataset was dense and due to the limitations of the excel sheet, we found out that the original network consisted of 81306 nodes among which only 40651 nodes with 1048564 edges were read. It was found that in the network(number of nodes and edges) read, the maximum k-shell value was 104 with 124 nodes in the innermost shell. The average in-degree(K_{in}) of the network was 42.357.



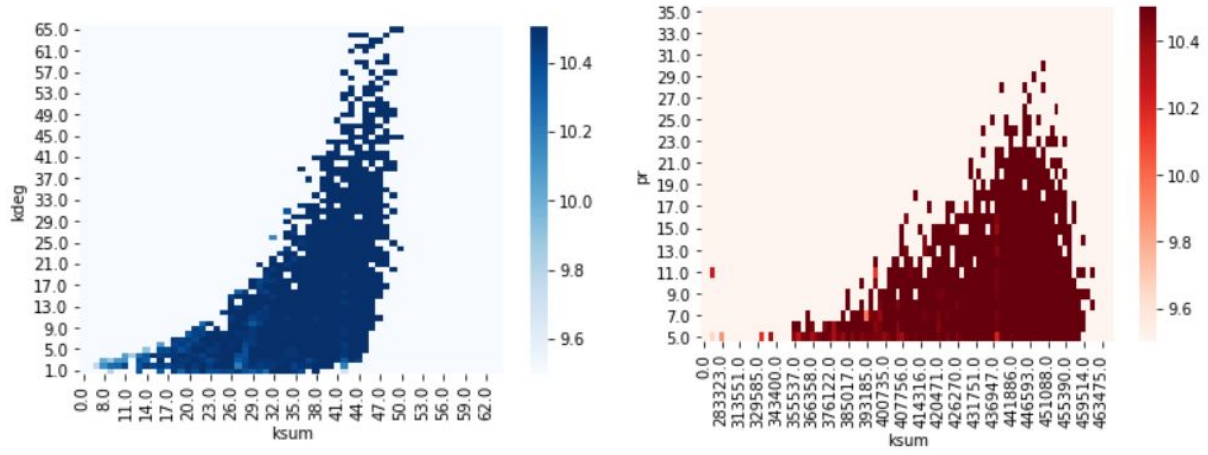
The above plots describe the comparison between k-core, ks with in-degree, k-in and then with pagerank, pr. Logarithmic values of average influences for combinations of ks and k-in/pr are plotted. For plotting the graphs values of pr and k-in have been normalized. The higher values of ks usually results in a greater influence over the network. The plot from k-in vs ks is more linear than the graph between pr and ks, suggesting that influence values from pr with ks is more random as compared to values of k-in.



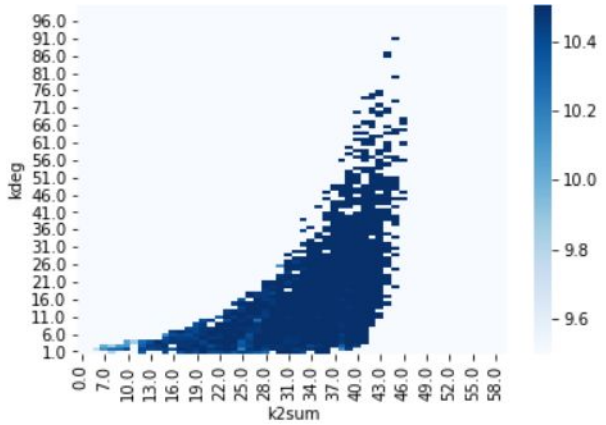
In the above plot, we have ks in blue color, k-in (kdeg) in orange and pr in grey. The ranges of ks, k-in and pr were divided into five bins and the standard deviation among the influence values keeping the measure fixed were obtained. From the graph we can see although there is an increase in the randomness of the values of influence as we get towards the higher values of the respective measures, the ks values perform better than the other two and pagerank performance is the worst.



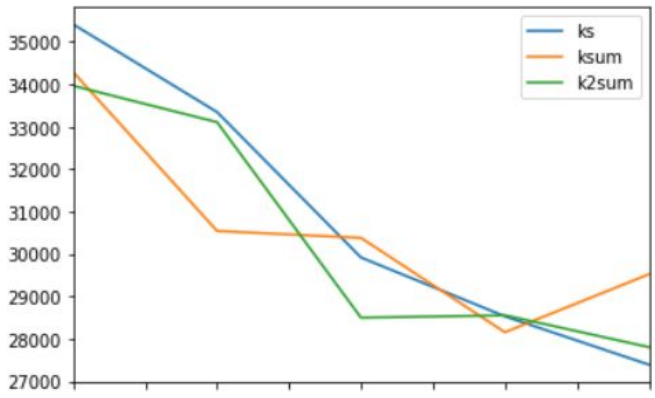
In the above graph, y-axis is for the average influence of top ranking nodes and x-axis for top 10% fraction of the range of measure (ks/k-in/pr), followed by top 20%, 30%, 40% and 50%. The plot clearly depicts that the influence of top most k-shell nodes are relatively higher than the other two measures.



The above plots describe that ksum was able to predict the average influence of nodes with a combination of ks and kin/pr values, more reliably than k-in (in-degree) and pagerank (pr).



The above graph is plotted for average influence of combinations of k-in and k2sum, the graph seems to be a slightly improved version of the graph for k-in and k2sum.

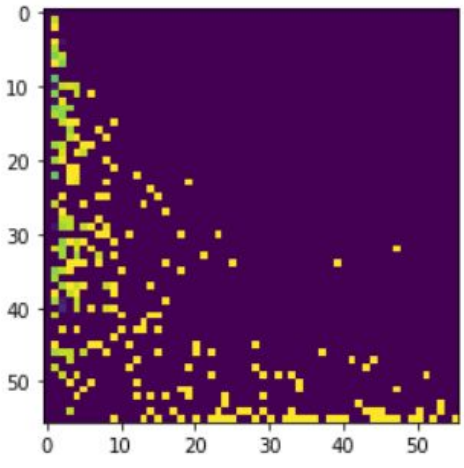


In the above graph, y-axis is for the average influence of top ranking nodes and x-axis for top 10% fraction of the range of measure (ks/ksum/k2sum), followed by top 20%, 30%, 40% and

50%. Randomly, 50 nodes were picked from each top fraction. The three measures are almost similar. The variation in the peaks can be justified by the fact that nodes were chosen randomly.

We also had deployed an artificial information simulation using the SIR model on a dataset called Wikipedia Vote Network to obtain a diffusion network. We found out the k-shell number, kin values, page rank values, ksum, k2sum and influence of each node. Information of each node in the network was arranged in ascending order of the k-shell value. We found, the largest value of k-shell was 56, the average kin value, <kin> was 14.573, number of nodes, N in the network was 7115, number of edges, Ne was 103687 and number of nodes involved in diffusion, Nd was only 2243.

The plotted logarithmic values of average influence for combinations of ks and kin in the graph looked meaningless, hence implying that employing artificial information diffusion on networks was futile. But even by this method we found out that most of the nodes with top-most values of ksum and also in k2sum were present in the innermost k-core of the network.



In the graph, y-axis is ks and x-axis is kin.

CHALLENGES ENCOUNTERED

The biggest challenge faced by us was searching for an appropriate dataset for carrying out methods suggested by the paper. We needed a real network whose diffusion graph was static and obtainable. We had first considered the facebook dataset from https://snap.stanford.edu/data/ but all the nodes in this dataset were in the same k-core, we considered the LJ dataset, the one mentioned in the paper and the dataset was voluminous beyond the capability of our system to

load and process it. We chose the twitter dataset in which we had to compromise with the deletion of 10,000 network edges.

The influence of most the nodes in the twitter dataset were almost the same, 36433 so we had to decrease the depth limit of breadth-first search, for finding the individual influence of nodes, to 10. Moreover, traversing through the entire network of more than 40000 nodes took a long computation time.

FUTURE WORK

We would want to work on Twitter posts, by extracting out tweets on trending topics using keywords in the Twitter API and then collecting the diffusion data by following the retweet ID and tags, thereby generating an information diffusion graph of our own. We shall then validate and analyze our work by employing the methods used earlier on this network.

MEMBERS CONTRIBUTION

Most parts of the work were combined efforts of the team members Priya and Riya, where contribution of Priya was more towards studying the network and working with K-core, in-degree, pagerank and individual influence of the nodes and Riya contributed in finding out the average influences of the combinations of measures and evaluation of ksum and k2sum values for the nodes.