ANALYSIS OF HUAWEI SOCIAL NETWORK DATA

Prathyusha Aravapalli, Aruna Venkatasubramaniam

IDS 564 – Social Media and Network Analysis, Fall 2018

1. Introduction

Huawei is the leading company that provides information and communication technology services. The brand is headquartered in China has its business operations spread out all over the world. At the time of its establishment, Huawei focused more on manufacturing phone switches and later it expanded its business to building telecommunication network and communication services. Huawei has launched a successful digital marketing campaign with Lionel Messi and Scarlett Johansson. The number of PR campaigns has increased 300 percent in Western Europe in 2017. All these are visible on social media platforms like Twitter, Facebook and Instagram. Huawei is maintaining its company profile on all these platforms. In this project, we tried to analyze how it is leveraging its network and come up with few inferences that can create significant business value to Huawei.

2. Data Description

We have sourced the Huawei social media data from Kaggle.com. The data is in 3 csv files, one for each platform, and is in the form of an adjacency matrix. We have done some preliminary exploration to understand the data and to identify the size of the network that it encompasses.

- Facebook Communication Network is Directed and labeled having 1000 nodes and 100306 edges.
- Twitter Communication Network is Directed and labeled having 1000 Nodes and 500630 edges.
- Instagram Communication Network is Directed and labeled having 1000 Nodes and 9866 edges.

The edges of this communication network are the interactions between people who use social media on daily basis and tweet or post regarding Huawei. The dataset has been cleaned and pre-processed already. Each of this dataset has the same 1000 nodes and the connections between them.

3. Project Objective and Goals

Huawei is mainly doing its business in the telecommunication area. Though it is a successful company, it is very difficult for to target new customers and more importantly retain the customers who are already using this brand. In a market that is saturated with smartphones its difficult for a new company to seize its customers. This project has two main goals:

- We try to understand how this company is being perceived by its customers and employees and what kind of information we can derive from analyzing the network formed by it.
- Understanding the layout and connections between the nodes, would help us generate insights into the company's marketing strategy and come up with suggestions, if any.

4. Data Summary Statistics

We focused on summary statistics in order to get an idea about the network. To help us understand the nature of network we have looked at statistics like the maximum degree connections, the average path that needs to be traversed for information spread, the diameter of the network etc. These have been summarized in the figure below.

Summary Statistics

Facebook	Twitter	Instagram
Number of Edges:100,306	Number of Edges:500,630	Number of Edges:9,866
Number of Users:1000	Number of Users:1000	Number of Users:1000
Clustering coefficient:0.1002897	Clustering coefficient:0.5012008	Clustering coefficient:0.008399037
Diameter:3	Diameter:2	Diameter:5
Avg. path:1.899632	Avg. path:1.498869	Avg. path:3.273137
Edge Density:0.1004064	Edge Density:0.5011311	Edge Density:0.009875876
Assortative Degree:-0.007703903	Assortative Degree:-0.004438593	Assortative Degree:0.005353542
Max Degree:254	Max Degree:1090	Max Degree:40
Min connections: 64	Min connections: 445	Min connections: 2
Max connections: 127	Max connections: 545	Max connections: 20
Most connected: Engkos Kosasih, Ernie	Most connected : Dililah	Most connected: Alexis, Ishku Ishku, Alveena
Least connected: Aleisha	Least connected: Saleem Hassan	Least connected: Denno, Rohan, Fahad Rehman

Figure 1 Summary Statistics

Interpretations

- Even though the number of nodes across all platforms are same, the connections/edges between the nodes make a network dense or sparse. Considering the number of edges, Twitter is very dense network, the Facebook has next highest edges. This gives us an idea of how Huawei is being talked about or exposed to in different platforms. A company can have huge follower base in a platform, but if no one are actually exchanging any information about it, it would mean that it is not having a good exposure to the public.
- Next metric is clustering coefficient, which is high in Twitter network. This means that the nodes that form clusters in Twitter are very high. It could signify that Twitter is tightly knit community. The clustering coefficient of Instagram is very less signifying the sparsity and low connectedness of the network.
- Diameter is the shortest distance between any two nodes that are very far. When we see the diameter of Twitter, it is less compared to other two platforms. This reinforces that the Twitter is tightly knitted community so the distance between the far nodes is just 2. This also signifies how fast information can spread in twitter compared to others. On the other hand, the shortest distance between far nodes in Instagram is highest reiterating the fact that it is very sparsely connected.
- Another interesting observation we made when looking at the statistics is, the nodes which are highly connected are not the same across 3 platforms. Also, the nodes least connected are not the same either. This shows that different nodes have different preference of media platform that they interact the most in.

5. Network Graphs

We have plotted the network graphs for all 3 platforms. For this we have opted to use the Fruchterman-Reingold layout as, this layout gives the most optimized layout by arranging the nodes in such a way that, the edges cross over the least allowing us to look at the network and understand how it is laid out.

The layout for Facebook has been given in figure 2. The yellow nodes represent the highly connected nodes and the blue circle represents the least connected nodes. Intuitively, we assume that most connected node would be in the middle, in the thick of network and the least connected would be in the fringes. This is what we see in the layout too. We see similar layouts for Twitter and Instagram as well, shown in figures 3 and 4. All the layouts are done in Fruchterman-Reingold. We can see at a glance that Twitter is the densest network of all implying that most of the information about the company is passed in this platform.

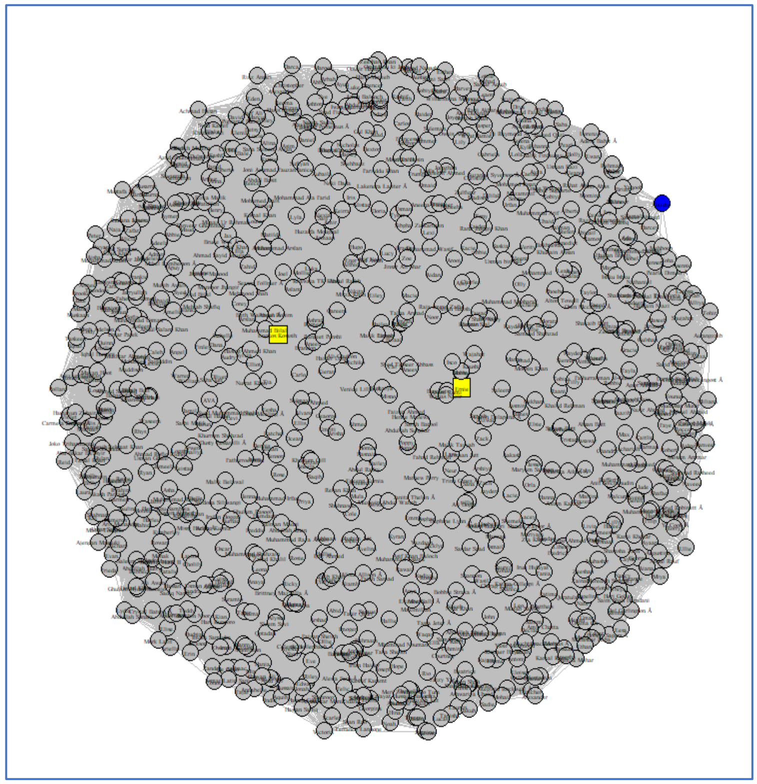


Figure 2 Fruchterman-Reingold Layout for Facebook

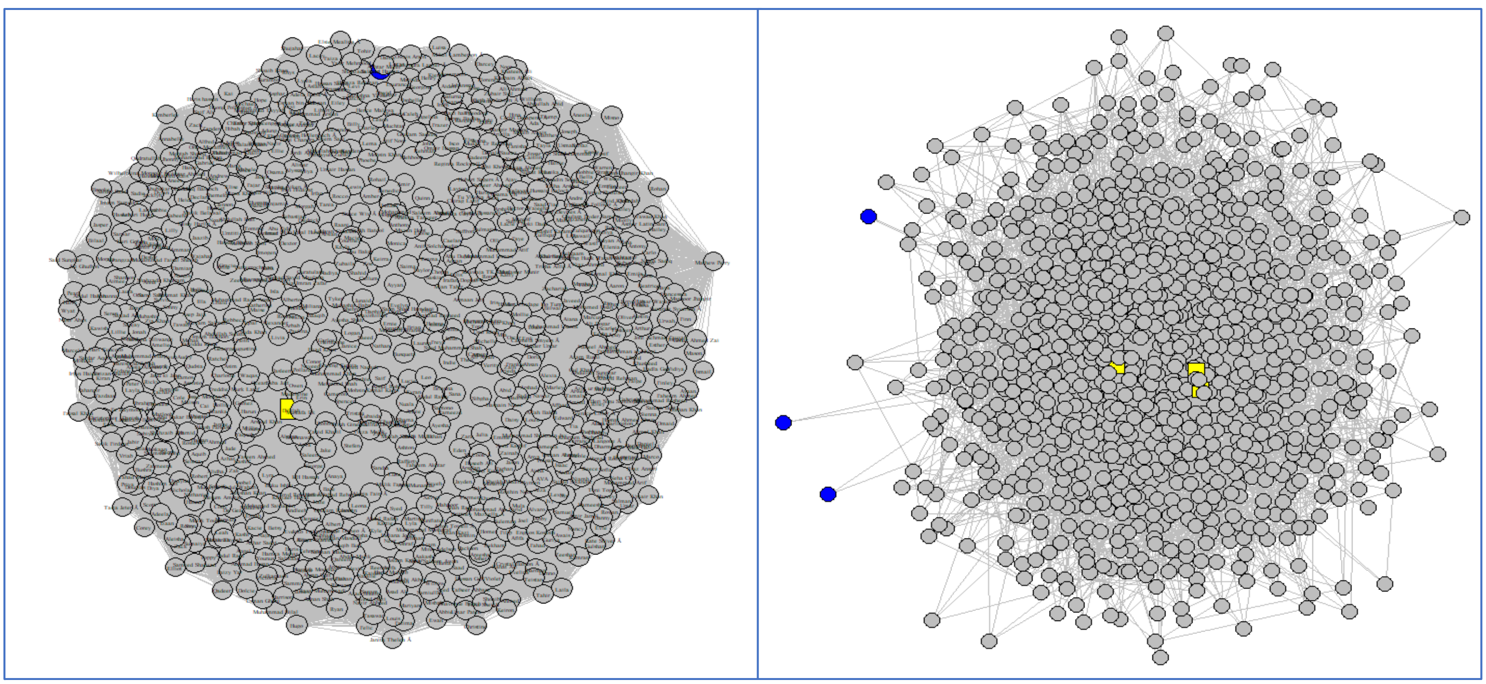


Figure 3 Layout for Twitter and Instagram

6. Community Detection

Now that we have the entire network layout, we can see that it is too dense for us to make any sense of it. So, we break down the graph to look at the communities formed within the network to try and understand how they interact with each other. We have used walktrap community detection algorithm to identify the communities within the network. Since this is an information network, walktrap algorithm made the most sense as we did not have to ignore the directions of information flow and also, walktrap algorithm detects communities based on how the nodes are laid out on random short walks between 2 pairs of nodes. The communities for the three platforms are shown in figures 4 and 5.

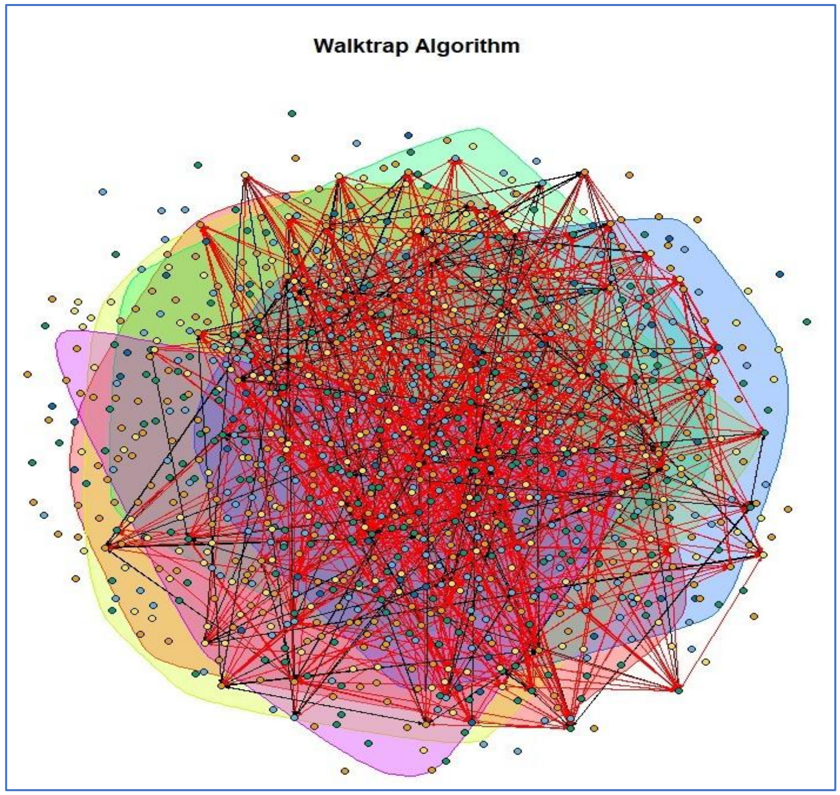


Figure 4 Community Detection for Facebook

For Facebook the distribution of nodes in various communities has been given in the table below. Except for communities 1,2 and 6, the rest of the communities seem to have an equal number of nodes among them. The red arrows show the information flow within communities and the black arrows show information flow between communities. We see more of red than black arrows, signifying good amount of information flow between communities. Similarly, we have run the walktrap algorithm for twitter and Instagram data as well and the community structure is as shown in figure 5.

For Twitter, we see a lot of black arrows signifying a lot of within community information flow than between communities. On the other hand, the Instagram communities show that there is hardly any communication, within communities.

	1	2	3	4	5	6
Facebook	225	281	153	147	149	45

The distribution of nodes in different communities for twitter has been given below:

	1	2	3
Twitter	477	315	208

Twitter just forms 3 communities showing the interconnectedness of the network with more than 70% of the nodes falling within first two communities. Instagram on the other hand forms 54 communities which are very sparsely divided between them with highest being 102 nodes in community 5 and least being 1 node per community for 53 and 54. This can also be seen in the right-side picture in figure 5.

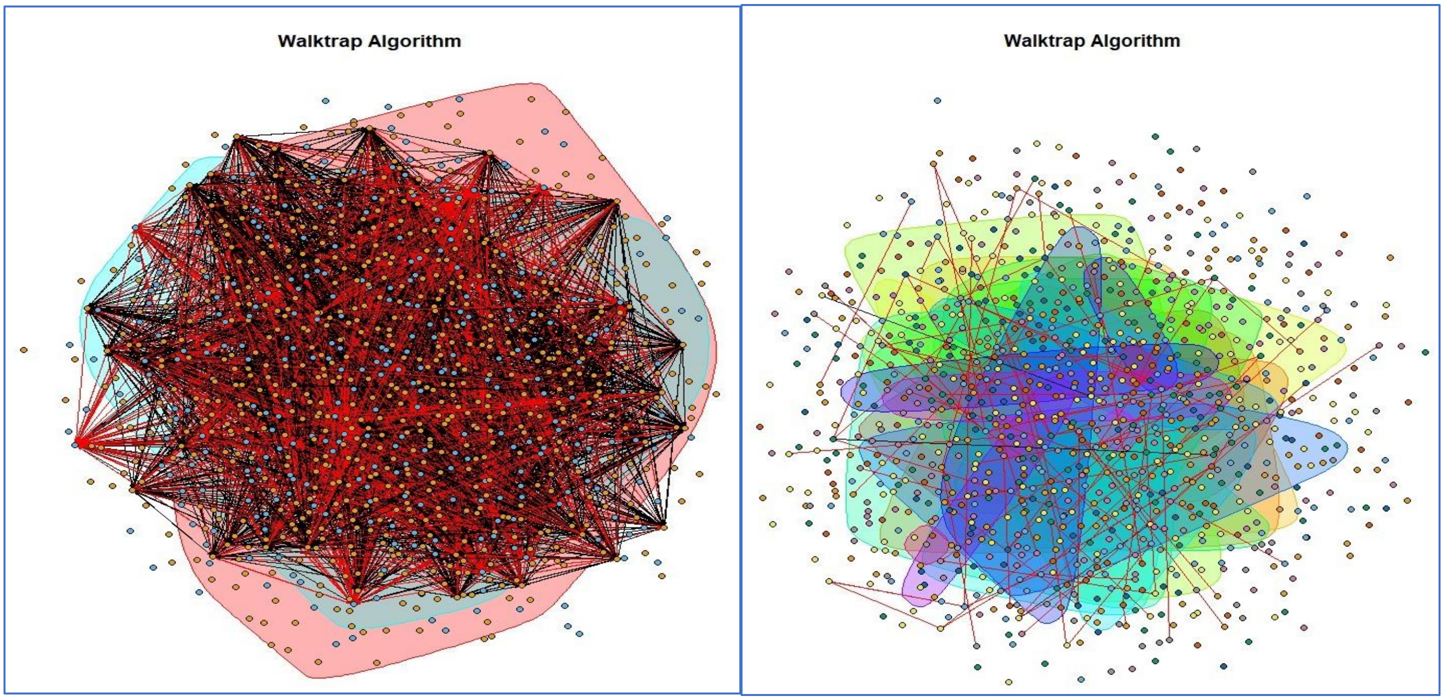


Figure 5 Community Structure for Twitter and Instagram

7. Subgraphs

The subgraphs are generally developed from the main network, by separating the nodes which fall under one community. When we tried the same for these datasets, we found that the network was still too dense to make any inferences. So, we induced subgraphs by isolating the top 10% of the nodes which had the most connections.

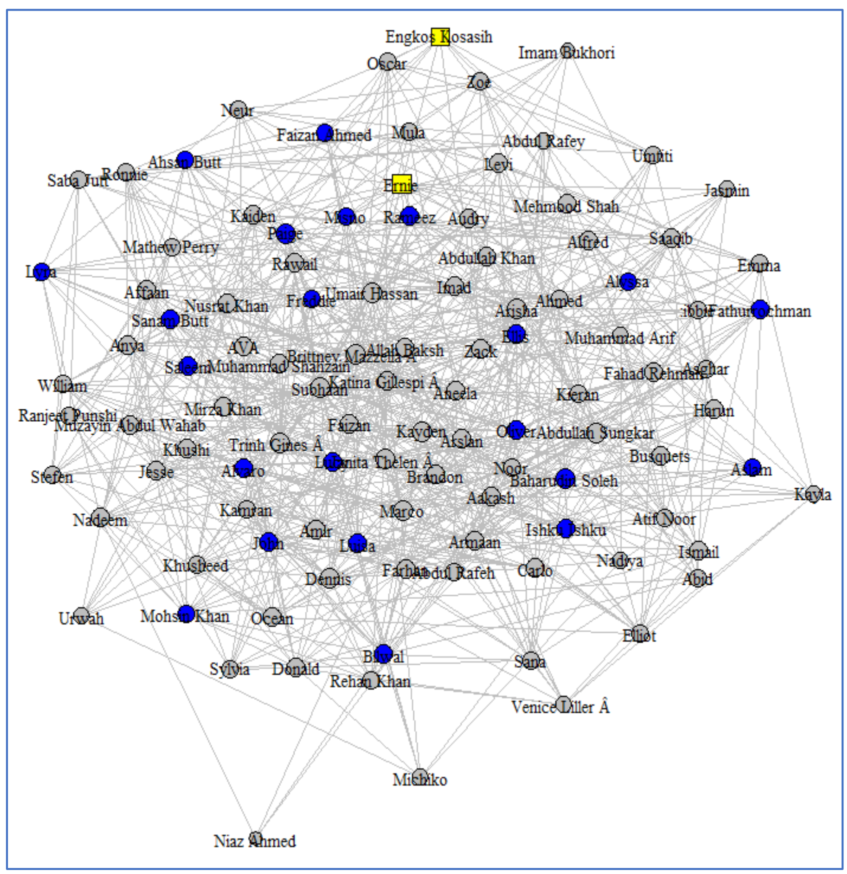


Figure 6 Subgraph for Facebook

Communities 1 and 2 together make up 50% of the number of nodes in Facebooks. Analyzing just community 1 or 2 did not help us much, as they were still dense. We have plotted the subgraph using top 10% of most connected nodes and represented the top connected and least connected nodes with yellow and blue respectively. The graph is shown in the figure 6 below. For Facebook the top 10% of connections implies, nodes with degree/connection between 112 and 124. The assumption behind considering this subgraph is that, we think the most connected part will be the densest part and would form a community. As we intuited before that the most connected nodes will be in the thick of the network, we can see that it does not hold true here. The yellow squares are more in the fringe of the network.

Similar graphs have been plotted for Twitter and Instagram and can be seen in figure 7.

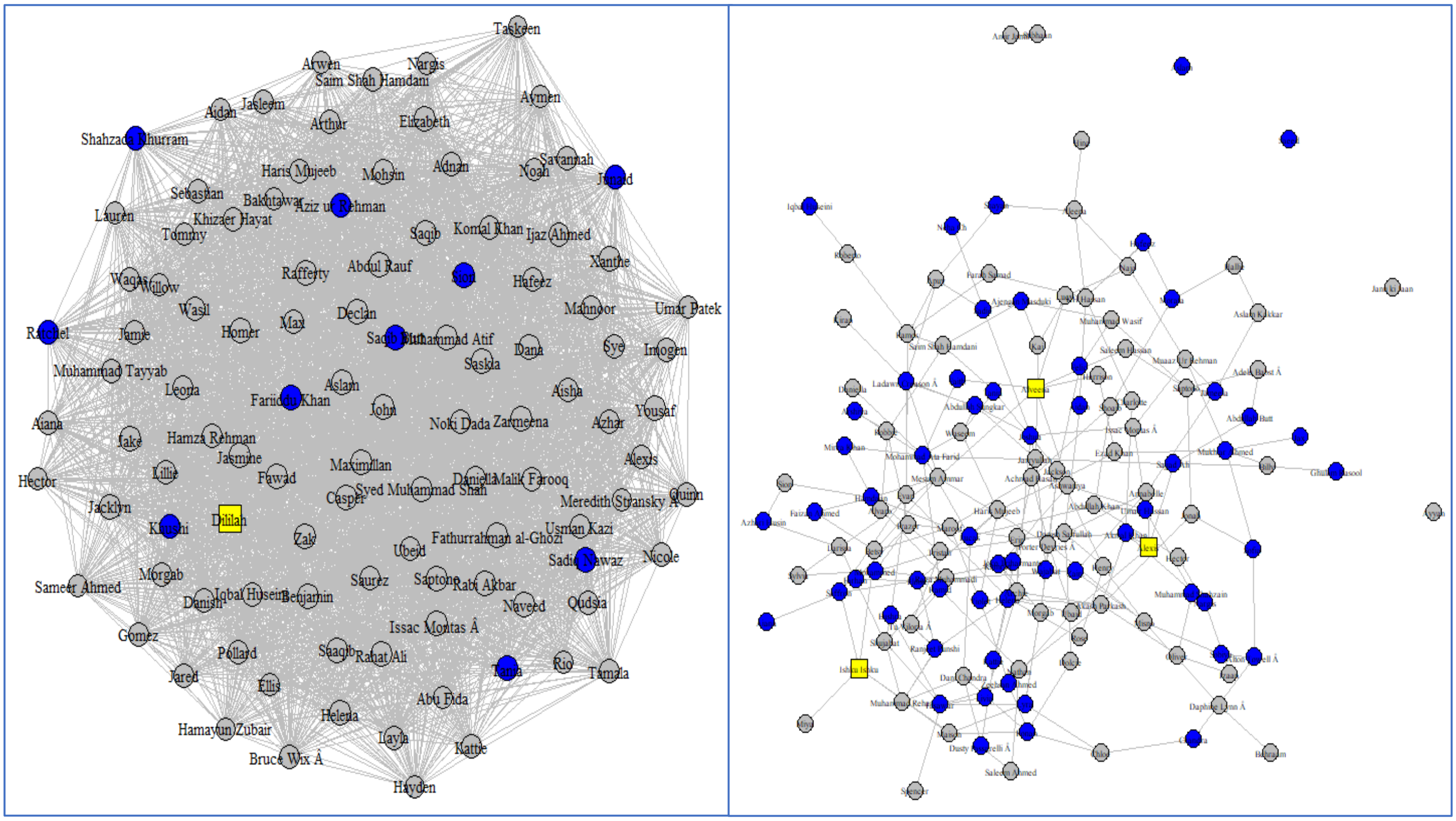


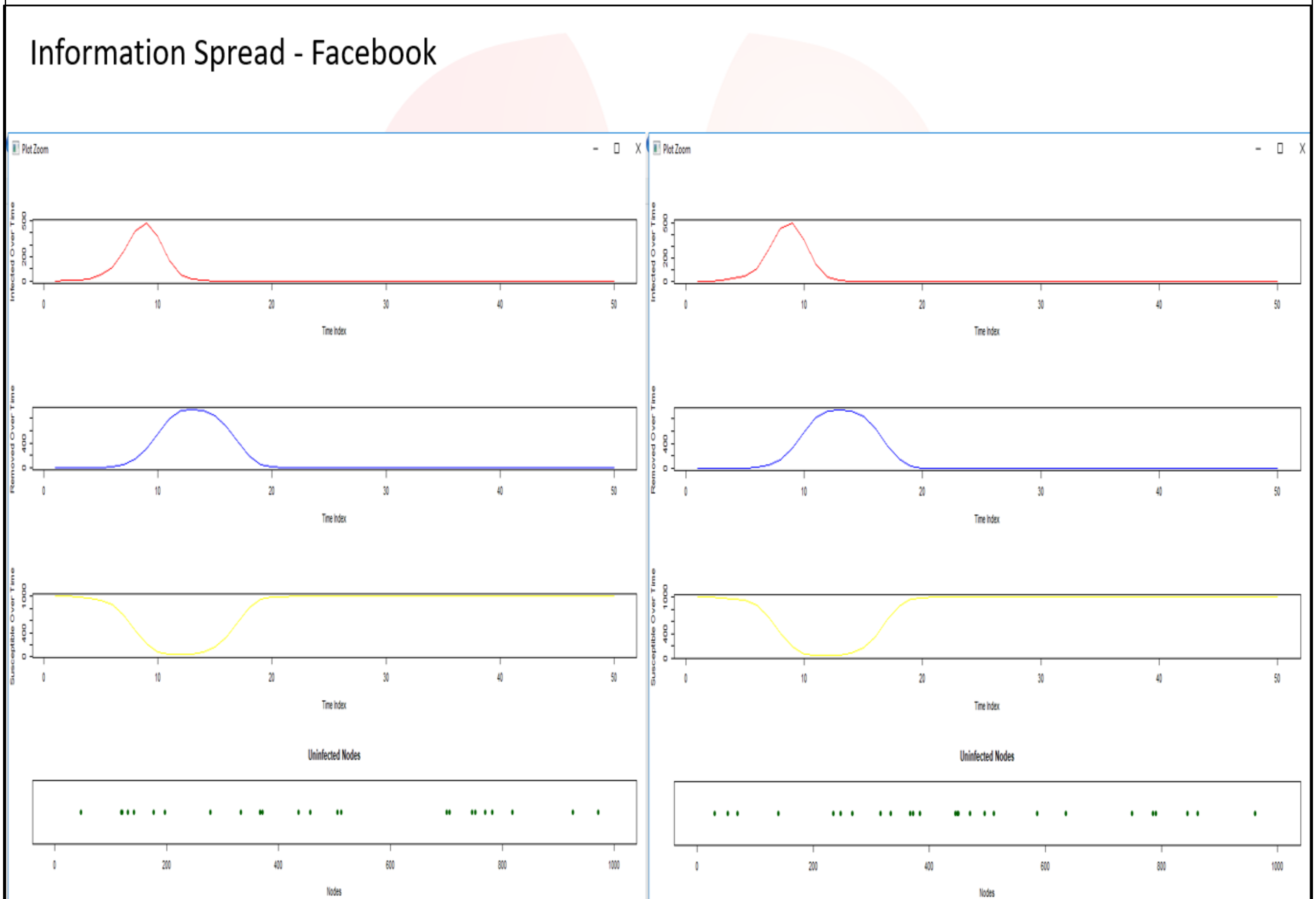
Figure 7 Subgraphs for Twitter(left) and Instagram (right)

8. Information Spread (Social Contagion)

The idea of spreading disease among people in a biological sense is called Social Contagion. For the purposes of our report, we shall call it information spread as this concept can be applied to all areas that deals with people and spread of information among people. The study of how an epidemic spread through a network could give significant idea of how information spreads across the network.

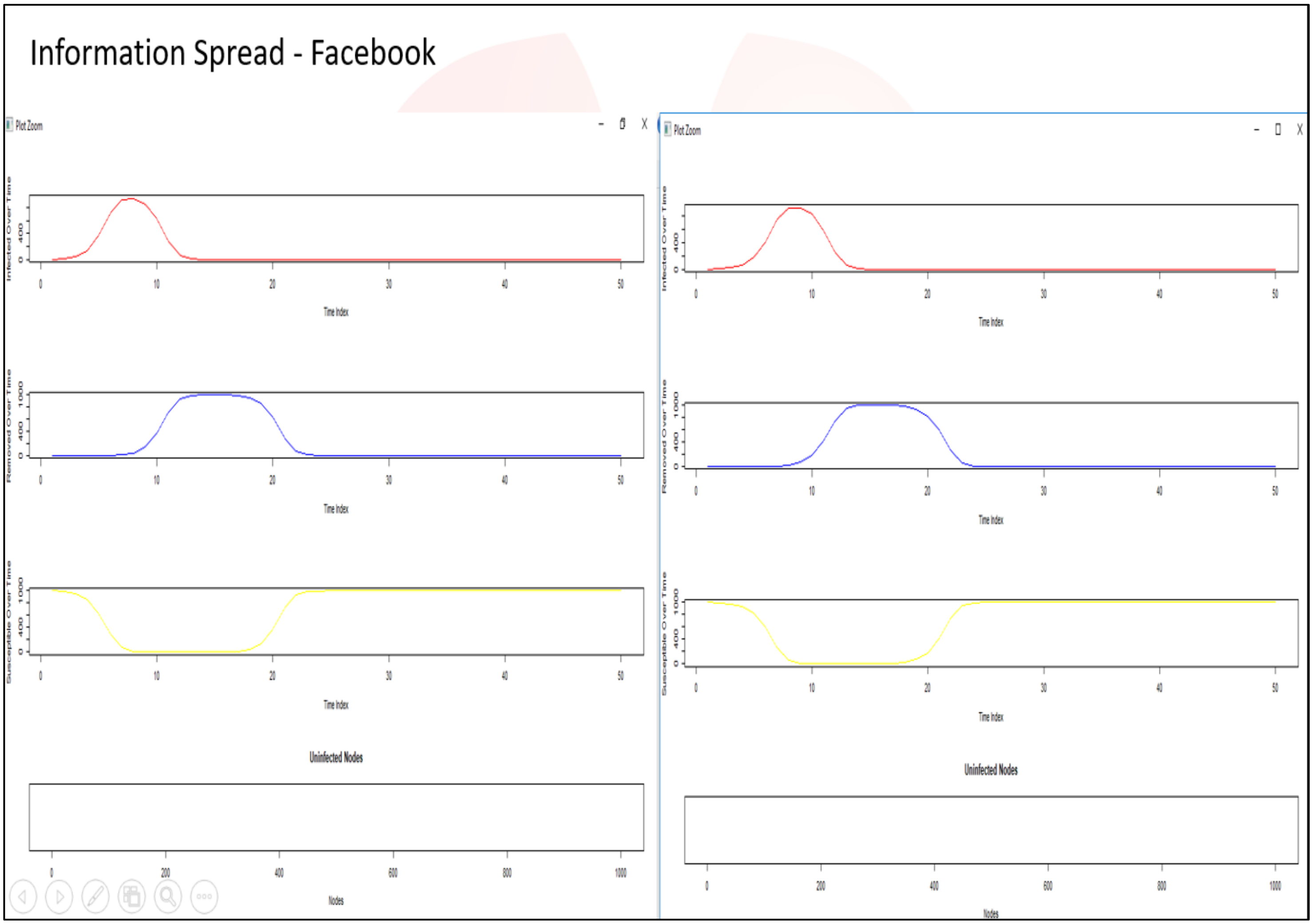
In a network, we identify those people who are very significant in diffusing the information so that, the company gets an idea on whom to target. We visualized how the information spreads in our network using SIR epidemic model. Here 'S' stands for susceptible, 'I' stands for infected and 'R' for recovered. In our case susceptible are those nodes that might join in the information spread, infected nodes are those nodes that that are in network diffusing information and recovered are the nodes that were once in the network but not anymore. The information circulates mainly in the center of the network compared to the peripheral region. This is mainly due to the community which is tightly knit at center compared to periphery (again, intuition). So, in order to increase the brands social presence, we need to identify influential targets which are capable of spreading the information throughout the network. Once our target nodes are identified, we can determine properties to make the brand popular.

In the first plot below, we observe two things. The first row in plot shows how the information is spreading. The curve is very narrow which signifies that the information spread could be less. We cannot for say for sure that it is low, as we need to compare it with other models. The second observation is that the number of people who are not participating in the information spread. This is shown in the last row of the plot.

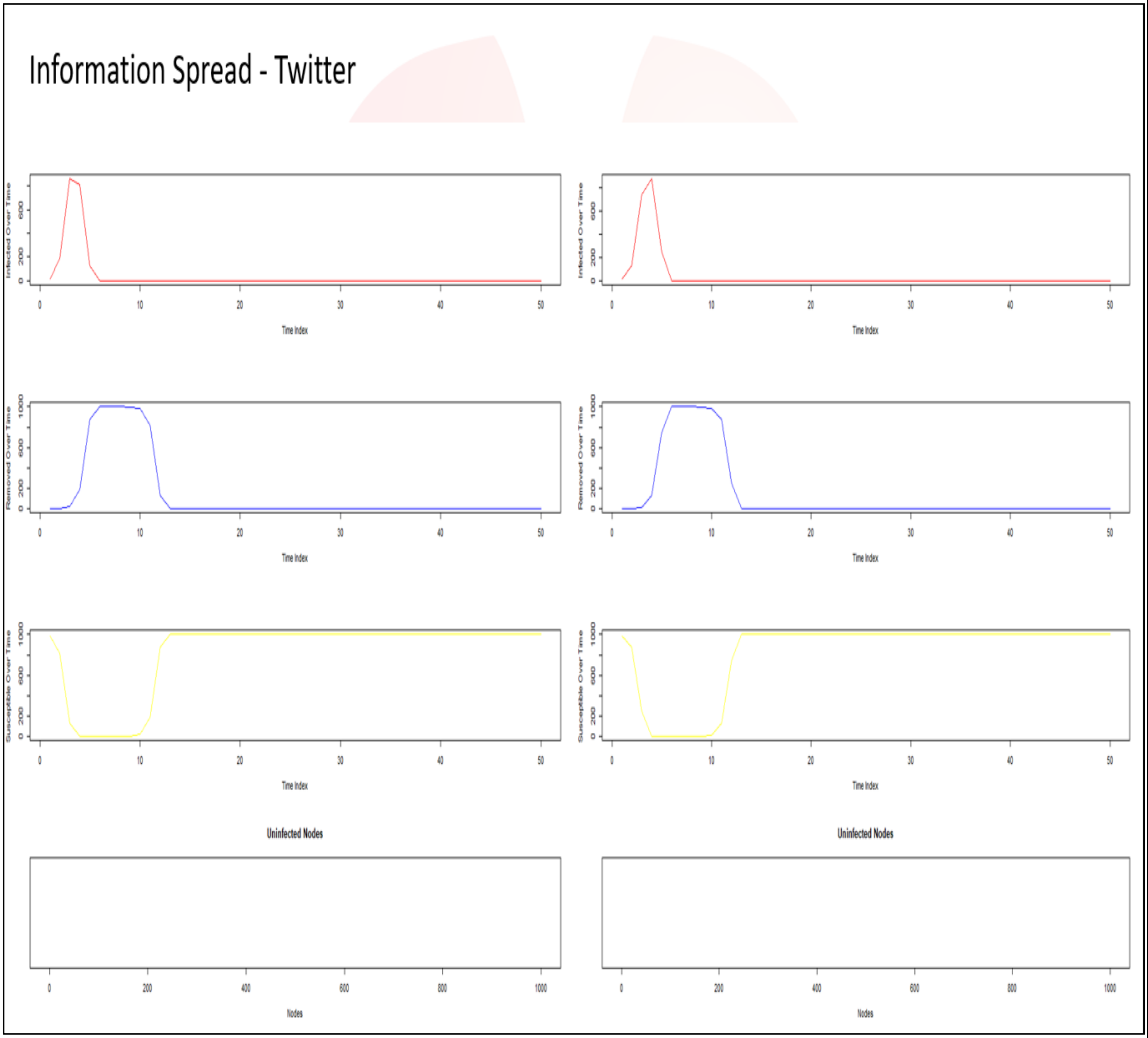


There are two parameters in the SIR model that are very important for tuning the model. The value for "R" and "S" determines how information spreads in network. "R" is removed after, so we give particular number of iterations as a value to this parameter. So, if a node stays in removed state even after R iterations then the node is moved to removed state. Similarly, if a node stays in susceptible state even after S iterations then the node is moved to susceptible state.

In our next plot, we changed the number of iterations of R and S to see how it changes the amount of information spread. We see that the amount of information that is spread through the nodes has increased compared to previous plot. The number of nodes that are not participating in the network is almost nil. Which implies that all nodes are participating in diffusing information.

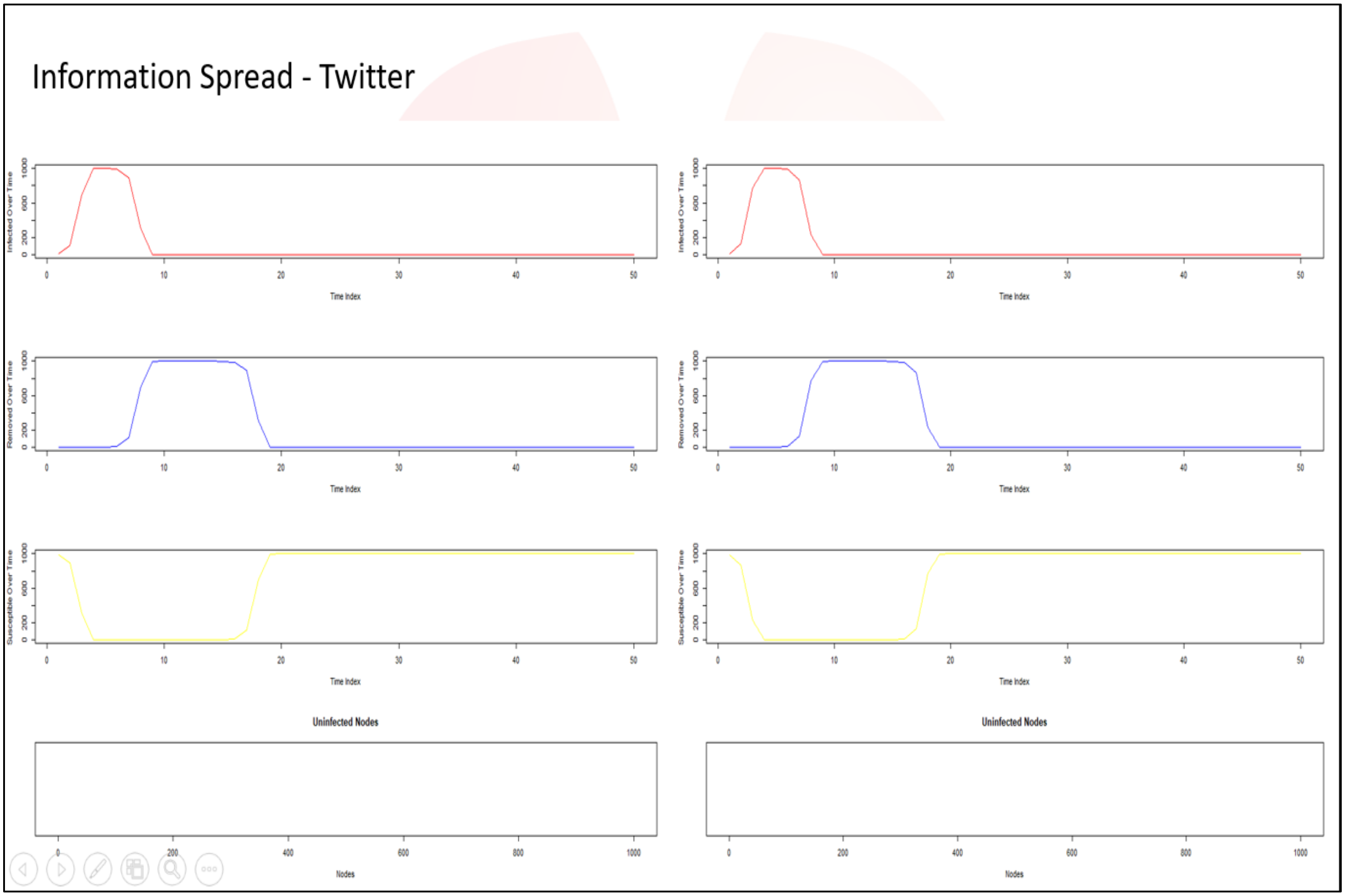


Similar analysis is done for Twitter and Instagram network.

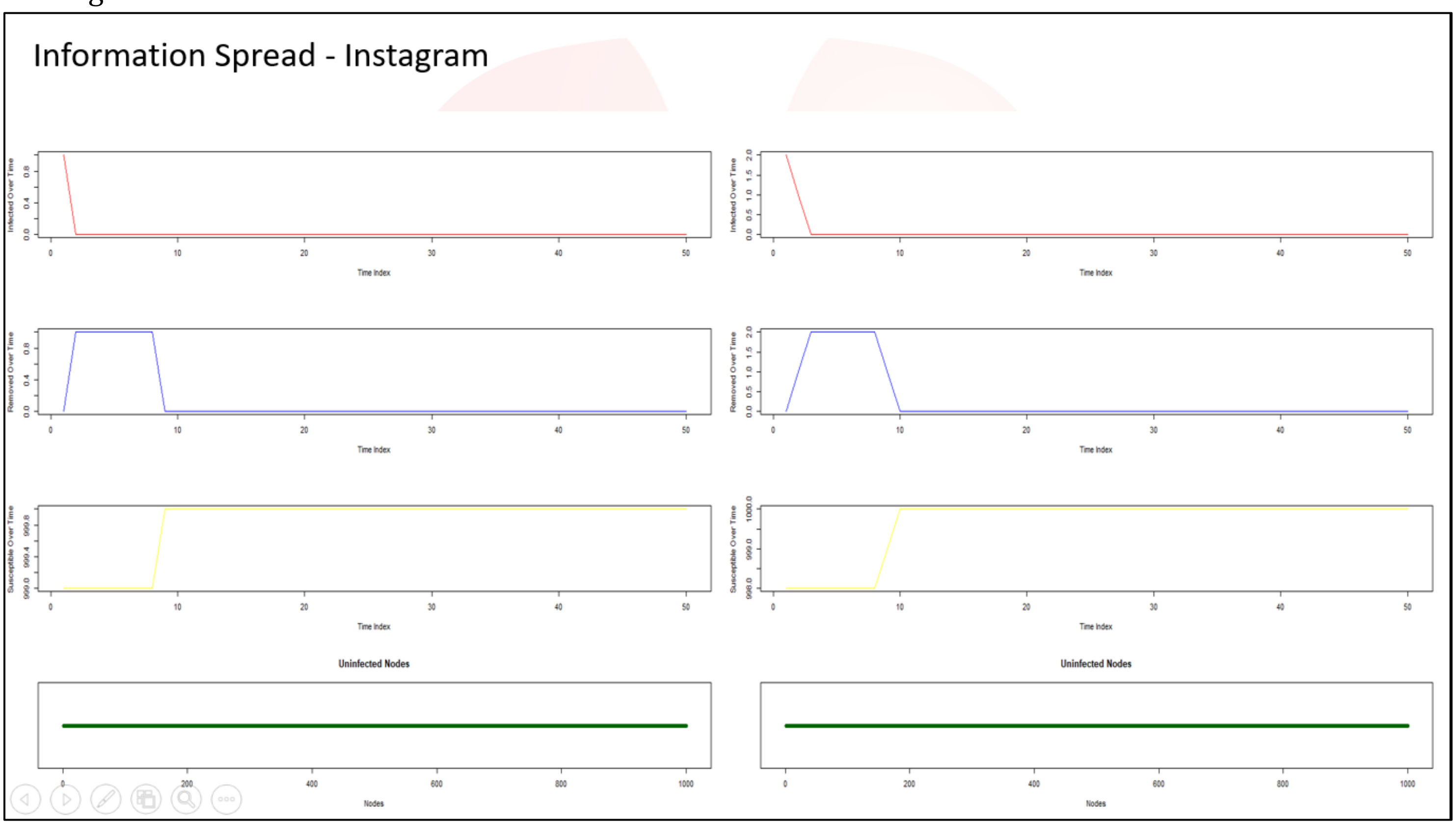


In this plot the amount of information spread (first row of the graph) is narrower meaning information spread is little less. But in the last row, there are no nodes. This signifies that all of the nodes are participating in the information spread. This is similar to what we would have hypothesized that as Twitter network is tightly knit, all nodes participate in the information spread. But again, changing the parameter R and S iteration values, we get the plot shown below:

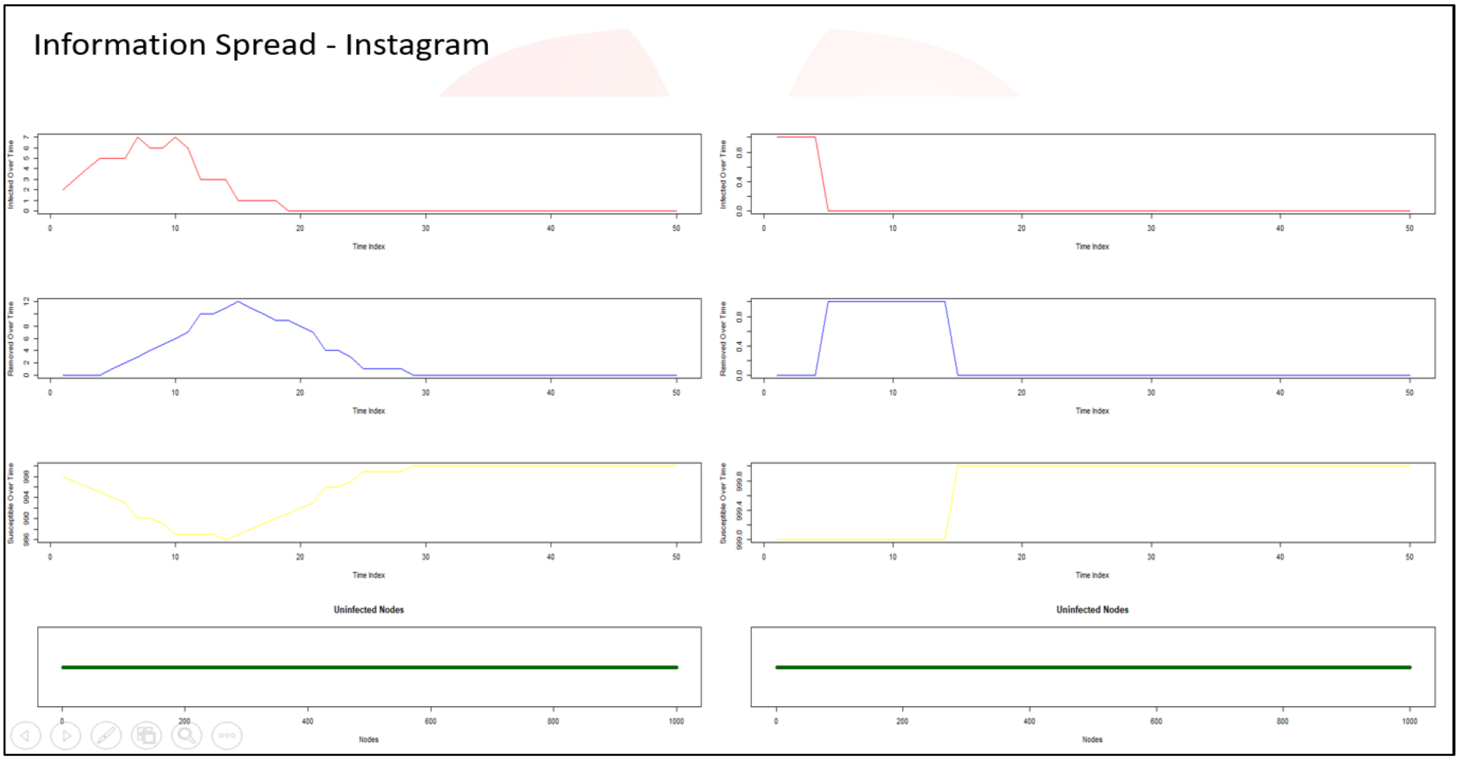
In the following plot, the curve is wider compared to previous plot. But there is no change in the last row.


Information Spread - Twitter
Infected Over Time
Time Index
Removed Over Time
Time Index
Susceptible Over Time
Time Index
Uninfected Nodes
Nodes

Instagram:


Information Spread - Instagram
Infected Over Time
Time Index
Removed Over Time
Time Index
Susceptible Over Time
Time Index
Uninfected Nodes
Nodes

Here there is not much information spread. The last row shows all the nodes. As the network is sparsely connected, there are lot of nodes that are not participating in the network to spread of information. We then verified by change the parameters in SIR model, the plot is shown as below:



There is still no change in the number of nodes not participating in the information spread. And the information dies very soon in Instagram network as per first row representation.

Interpretation

In the above plots we found out that the information dies after a finite amount of time. This could represent the fact that a new brand of product has come into the market and eventually popularity of Huawei decreased with time. When the value of R is equal to number of total iterations as SIR model then the information will never die. But achieving this is quite impossible because the information will die at some point in time and will be overtaken by the new information that will again be circulated in the network. Also, when the information is received first, the reaction to hearing to that information first time will be quite different when it is passed around and is again received by the first individual. Also, after running the function for varying iteration numbers, we found that by tuning the parameters in SIR model the Twitter network is most efficient means to spread information in the network.

9. Conclusions and Recommendations

Connecting all observations and exercises we carried out so far, we want to conclude that, even though we intuitively think the way the network is connected and the way the nodes interact with each other is evident, it is not always the case. Like we saw from the subgraphs, the most connected nodes are not always in the thick of things, and from information spread we see that, even though lesser connected nodes are in the fringes, it does not mean they are not important to the spread of information. They are as important to a network as the most connected nodes as they are key in recycling the information after it has passed a certain time.

The recommendation that we provide to the company would be that,

- Huawei should keep pumping new information about its products and services, or in absence of new information, it should keep tweeting or posting about the company to keep the quality of information high and to make sure the information does not die out and users lose interest.
- It should concentrate on the way the most connected people/users and talking about Huawei among all three platforms.
- It should increase its presence in Instagram as well, as recent trends suggest that the millennials are using Instagram more than Facebook.
- Since different users are more active/more connected in different platforms, the chances that the information would reach more people would be high if they make sure that similar content is posted across all platforms.

10. References

- Data Source: https://www.kaggle.com/andrewlucci/huawei-social-network-data/home
- Definitions – Class slides and Wikipedia.