Epidemic Models in Social Networks (Facebook)
CS 590 – Elements of Network Science

ChoungRyeol Lee, Shubham Agrawal, Ashwin Jiwane

1. Abstract

Much attention is given to understand information diffusion behavior on social network that deeply effects interpersonal interactions in the real world. In this project we have analyze epidemic models with respect to information diffusion on social networks, Facebook in our experiments. The analysis is done for different versions of the same Facebook network simulating different real world scenarios. We observe for viral information diffusion in social networks users with highest value in degree distribution should be targeted.

2. Introduction

Social networks, especially Facebook, have become an important part of human life. Facebook has more than 1 billion users. Users have hundreds of friends and acquaintances. Users read and follow their friends' posts, share the information, promote some information, spread breaking news etc. Xu[8] model SIS on Facebook data and conclude all online users are tend to get infected due to mass connections in social network. Thus, social networks have become a very important medium to spread information among population, may it be advertisers, viral videos, breaking news etc.

The structure and the dynamics of social networks have been investigated intensively. With the advance of Information Technology, Facebook has emerged as the most influential online social networking. Many researchers have recently focused on the influence of Facebook affecting real relationship between individuals as a particular type of social networks that have the scale-free property i.e. a small average shortest path between individuals and a high value of the clustering coefficient. Matt and Ken [1] tried to mix social networks and epidemic models. J. McAuley and J. Leskovec [2] studied how friendships can be studied to detect different friend circles of user. Woo and Chen [3] studied how information diffusion happens on web forums using SIR model. This study can be easily converted to social network information diffusion study with required changes.

In our study, we try to examine topological properties of partial sample data over Facebook in order to shed light on fundamental understanding about the structure of complete Facebook and investigate temporal aspects of information diffusion behavior as epidemic spreading over the users.

We study how the information is spread among population. We study this important topic from various perspectives. When the content of the information is not of much important (for e.g. a funny joke), users share information only from certain friends (mostly close friends with whom user is in good relationship in real life). We simulate this by calculating weightage of friendship between the user and a friend. When the content of the information is of high importance (for e.g. breaking news of Boston Attacks) users share information no matter how important the friendship is. We simulate this by giving very high weightage to friendship between the user and a friend. When the content of the information is of low importance (for e.g. personal information of the friend) users hesitate to share information no matter how important the friendship is. We simulate this by giving very low weightage to friendship between the user and a friend.

For all the versions of information spread mentioned above we study how this information diffusion is affected when the information is alive throughout the time period of the diffusion (i.e. once the user shares information the information remains active on user's homepage for others to share). We study this by simulating Susceptible Infected (SI)[4] epidemic model on Facebook network. We also study how information diffusion is affected when we make information die after certain time period of the diffusion (i.e. once the user shares information the information remains active only for certain time period on user's homepage for others to share). We study this by simulating Susceptible Infected Removed (SIR)[4] epidemic model on Facebook network.

For various possibilities in information diffusion we study on how to maximize information diffusion in minimum time. We consider users (topmost to top5) who have maximum values with different centrality measures as our starting point of diffusion.

We observe as compared to different centrality measures selecting users with highest degree values in real world scenarios performs better. We chose these top users such that no two of them are in the same clique.

3.Problem Definition

Our main problem statement to study is 'how to maximize the information dissemination in social networks?' i.e. which users should be chosen such that viral marketing is achieved with minimum cost and in minimum time period. We analyze different scenarios of information diffusion in a social network by giving different weightage to friendship.

We also observe how information diffusion happens for these scenarios when starting point of diffusion is chosen as users having maximum values in PageRank, Eigenvector and Degree. For real world network (when weightage for friendship is calculated from the real data in our experiment) we find which centrality measure and which users should be chosen such that information is spread to maximum number of users as fast as possible. We run numerous experiments – described in implementation section- to draw inference and conclude answers for our problem statement.

This problem is extremely important to study. It can be used for viral marketing, advertisers to spread ads, for public leaders to reach people, for news agency to spread news etc. It studies which users to target in order to spread information faster and wider.

4.Models/Algorithms:

We create different versions of Facebook network by giving different edge weights. We calculate PageRank, Eigenvector and Degree centrality measures for networks.

Giving weightage to edges:

Each edge in Facebook network is given different weightage. We give weights 1 and cosine similarity measure between two users. The cosine similarity measure determines how close two users are in terms of mutual friends. The formula to calculate cosine is

$$cos(X,Y) = \frac{\sum_{t\varepsilon T}(X_t)(Y_t)}{\sqrt{\sum_{t\varepsilon T}(X_t)^2} * \sqrt{\sum_{t\varepsilon T}(Y_t)^2}}$$

where T = number of users in graph, X_t=1 when X is a friend of t and X_t=0 is not a friend of t. For graph with weight as 1 we create different versions where an infected user can infect others with constant probability as 0.3, 0.2, 0.1 or 0.05.

Susceptible Infected Model:

For each network and for each centrality measure we consider different users as starting nodes in SI model. In each centrality measure we consider five different set of starting users as infected nodes. Set1={top user with highest centrality value}, Set2={top2 users with highest centrality value}, Set3={top3 users with highest centrality value}, similarly we get Set4 and Set5.

At each time step, there are infected users and susceptible users. Susceptible users are the neighboring users of the infected users. At each time step an infected user can infect its susceptible users with probability assigned as weight. We run this model and analyze the results.

Susceptible Infected Removed Model:

In SIR model, at each time step, there are infected users, susceptible users and removed users. Susceptible users are the neighboring users of the infected users. At each time step an infected user can infect its susceptible users with some probability. Removed users are the once who have remained infected for certain timeperiod and then removed from the network such that they can't infect others or get infected anymore.

SIR model simulates the real world social network. Once the information is shared the importance of the information/post degrades over time period. Hence, we have performed rigorous simulations of SIR model.

For each network and for each centrality measure we consider different users as starting nodes in SIR model. In each centrality measure we consider five different set of starting users as infected nodes. Set1={top user with highest centrality value}, Set2={top2 users with highest centrality value}, Set3={top3 users with highest centrality value} and similarly we get Set4 and Set5.

We run this model for different timepriod (1,2,3,4 and 5) where timeperiod is a time for which infected user remains infected before it gets transform into removed user.

We also run a SIR model simulation for timeperiod = 2, 3, 4 and 5 and probability=0.2, 0.3 and cosinie weight such that the effect of infected user decreases with timeperiod. Thus, at each time step the probability

that the infected user infects others decreases by 0.25. This is more relevant to real world Facebook scenario since value of information shared by users degrade over time and gets removed after certain timeperiod. We analyze the results obtained.

5.Implementation and Analysis

We collected raw Facebook data available at Stanford Network Analysis Platform (SNAP)[5]. The data available is a disconnected ego network of 10 nodes. An ego network of a node is a network in which a node is connected to all of its neighbors and connections between all its neighbors as shown in Fig1. We traced back from disconnected ego network to create a Facebook graph (a part of Facebook network) as shown in Fig2. We used python and igraph for all of our experiments.

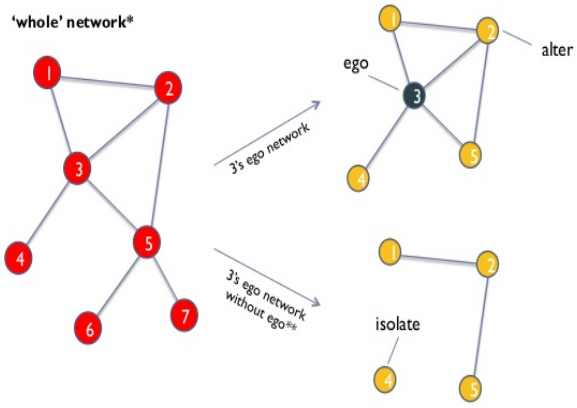


Figure 1

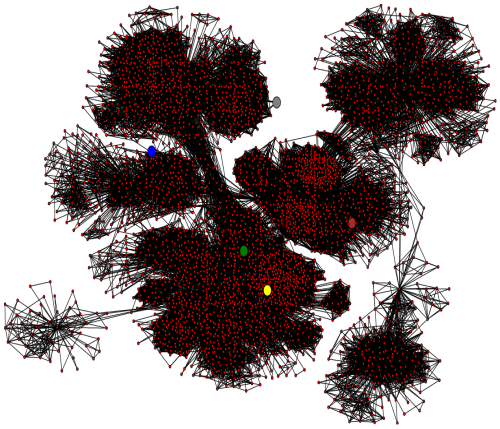


Figure 2

This Facebook network is a friend's graph. Nodes represent Facebook users and edge represents friendship between users. The Facebook graph has 3963 nodes and 88156 edges. Minimum degree is 2, maximum degree is 1034 and average degree is 22.245. Thus this graph contains users with different characteristics. Average path length is 3.776, thus following small world phenomena. Diameter of the graph is 8. Maximum size of clique is 57. The colored points in Fig2 are the users that are important with different centrality measures mentioned in Table1.

We assign weightage to each edge of the network by using cosine similarity measure described previously in the section4. Thus we have two basic version of the network, weighted (edge weights =cosine similarity) and unweighted (each edge has equal weightage).

Weight	Centrality Measures	Node
Weighted	Eigenvector	2160
	Pagerank	1641
	Closeness	100
	Betweenness	100
	Degree	100
Non-Weighted	Eigenvector	1868
	Pagerank	3381
	Closeness	100
	Betweenness	100
	Degree	100

Table 1

Centrality Measure	Facebook Interpretation
Pagerank	It is very likely to visit his profile in random surfing starting from anyone else's profile
Eigenvector	This person has 'influential' (or social) friends
Betweenness	This person is an important connection between different people
Closeness	This person uses minimum amount of 'mutual friends' link to connect to anyone else
Degree	This person has maximum number of friends

Table 2

Table1 shows different centrality measures and important users.

Table2 explains what these centrality measures represent in terms of Facebook.
This graph is not a scale-free network since it doesn't follow Power Law, it is easily observable from Fig3

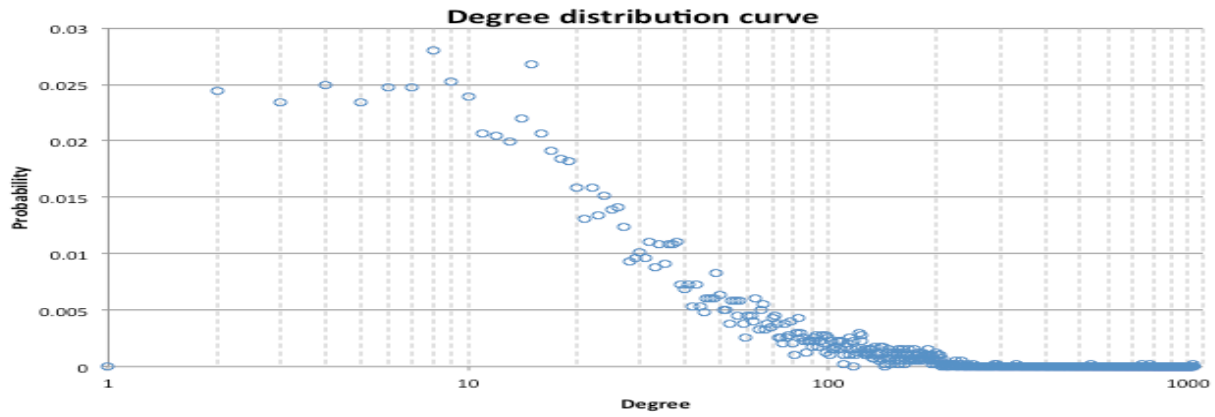


Figure 3

Susceptible Infected Model Simulation:

We ran SI model simulation considering different scenarios.

For unweighted graph, we considered 3 different starting users separately for each centrality measure (PageRank, Eigenvector and Degree) from Table1. For each starting user, we considered different constant probability (p=0.3, 0.2, 0.1 and 0.05) of infection between users. We observed time taken for the SI model to stop i.e. when the model infects no more users.

We also considered a model where instead of starting the SI model with just one user we started the infection with more than one user. For each centrality measure we calculated top5 users such that no two users from these top5 nodes are in the same clique in the network. We implemented a greedy algorithm to create cliques. We ran SI model simulation with these top2, top3, top4 and top5 users as starting infected users and the probability of infection as 0.1. We observed time taken for SI model to complete the infection.

For weighted graph, we considered 3 different starting users separately for each centrality measure (PageRank, Eigenvector and Degree) from Table1. For each starting user, the probability of infection between users is equal to the weight of the edge. We observed time taken for the SI model to stop i.e. when the model infects no more users.

Instead of starting the SI model with just one user we started the infection with more than one user. For each centrality measure we calculated top5 users such that no two users from these top5 nodes are in the same clique in the network. We ran SI model simulation with these top2, top3, top4 and top5 users as starting infected users and probability of infection as weight of the edge. We observed time taken for SI model to complete the infection.

SI model is highly probabilistic, hence we repeated all the simulations 5 times with different random files.

Susceptible Infected Removed Model Simulation:

We ran SIR model simulation considering different scenarios.

For unweighted graph, we considered 3 different starting users separately for each centrality measure (PageRank, Eigenvector and Degree) from Table1. For each starting user, we considered different constant probability (p=0.3, 0.2, 0.1, 0.05) of infection between users. For each probability, we considered different timeperiod (timeperiod=1, 2, 3, 4, 5) of infection i.e. the timeperiod for which the infected user remains infected before being removed.

We also considered the scenario when there is degradation in the infection probability i.e. as the time proceeds the probability of infection by the infected user degrades by 25% at every timeperiod. For this degradation scenario we considered p=0.2 and 0.3 and timeperiod = 2, 3, 4 and 5. We observed time taken for the SIR model to stop i.e. when the model infects no more users for continuous three timestamps.

Instead of starting the SIR model with just one user we start the infection with more than one user. For each centrality measure we calculated top5 users such that no two users from these top5 nodes are in the same clique in the network. We implemented a greedy algorithm to create cliques. We ran SIR model simulation with these top2, top3, top4 and top5 users as starting infected users and the probability of infection as 0.1 and timeperiod as 3. We also considered the scenario when there is degradation of 25% in the infection probability (constant p = 0.1) when starting with multiple users. We observed time taken for SIR model to complete the infection.

For weighted graph, we considered 3 different starting users separately for each centrality measure (PageRank, Eigenvector and Degree) from Table1. For each starting user, we considered weight between edges as probability of infection between users. For this probability, we considered different timeperiod (timeperiod=1, 2, 3, 4, 5) of infection i.e. the timeperiod for which the infected user remains infected before being removed.

We also considered the scenario when there is degradation in the infection probability i.e. as the time proceeds the probability of infection by the infected user degrades by 25% at every timeperiod. For this degradation scenario we considered timeperiod = 2, 3, 4 and 5. We observed time taken for the SIR model to stop i.e. when the model infects no more users for continuous three timestamps.

Instead of starting the SIR model with just one user we start the infection with more than one user. For each centrality measure we calculated top5 users such that no two users from these top5 nodes are in the same clique in the network. We implemented a greedy algorithm to create cliques. We ran SIR model simulation with these top2, top3, top4 and top5 users as starting infected users and the probability of infection as weight of edge and timeperiod as 3. We also considered the scenario when there is degradation of 25% in the infection probability (p = weight of the edge) when starting with multiple users. We observed time taken for SIR model to complete the infection.

SIR model is highly probabilistic, hence we repeated all the simulations 5 times with different random files.

In total, it took more than 90 hours to run all simulations and we created more than 600 csv files. We have plotted different graphs from all these csv files. Important graphs are explained in next section.

6. Results and Discussion:

We present and discuss our results in this section. For most of our graphs we have probability as 0.3, 0.2, 0.1, 0.05 and w. It means we have plotted those graphs for unweighted network when probability of infection is 0.3, 0.2, 0.1 and 0.05 and for weighted network when weight is w. In our graphs, Degree stand for observations made when model was simulated starting with best user/users corresponding to centrality measure Degree. Similarly Page represents PageRank and Eigen represents Eigenvector.

Observations of SI Model:

In, Fig4 each graph represents number of timesteps required for SI model to stop when starting with single topmost user with different centrality measures. In Fig4a, 4b and 4c, we show a box graph with median and standard deviation of timesteps required for various probabilities, since we run our simulations with 5 different random files to reduce the effect of randomness in our observations.

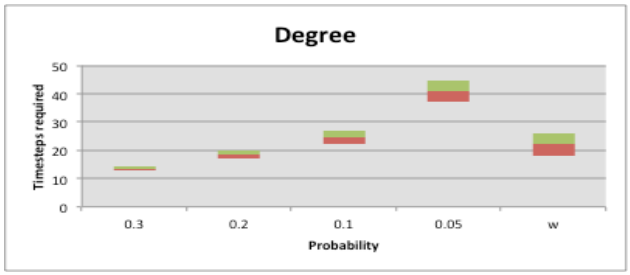


Figure 4a

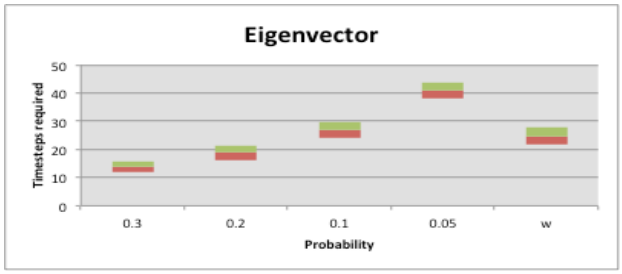


Figure 4b

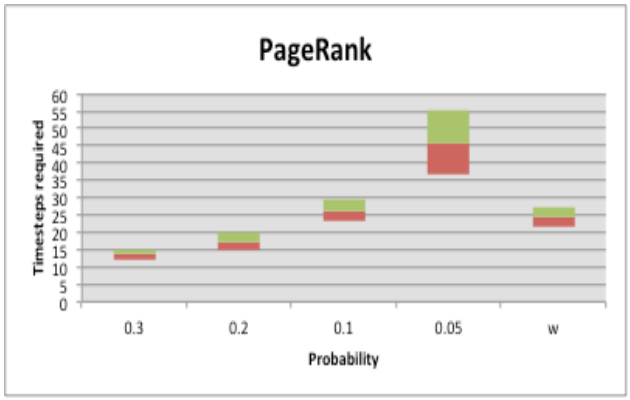
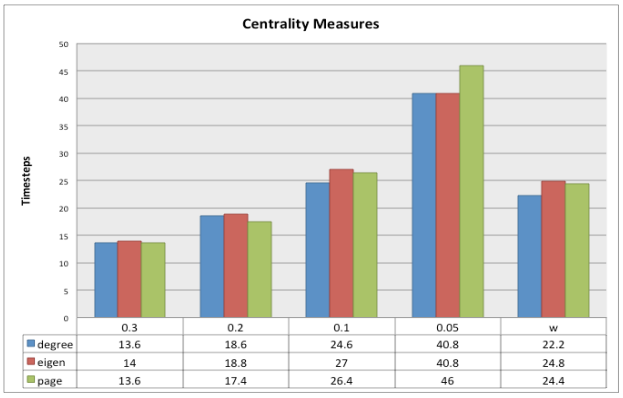


Figure 4c



	0.3	0.2	0.1	0.05	w
degree	13.6	18.6	24.6	40.8	22.2
eigen	14	18.8	27	40.8	24.8
page	13.6	17.4	26.4	46	24.4

Figure 4d

Fig4d shows a comparison between all of the centrality measures with different probabilities. From results, it's obvious degree measure stands out best in performance for information diffusion in minimum time. In Fig5 we plot average (since we run simulations for 5 different random files) number of users getting affected at each timestep for Degree centrality measure. We also have similar plots for PageRank and Eigenvector, due to lack of space we can't include them in report, they are available on the github link provided in Apendix. Degree plots stood best among the three, rate of infection was fastest with degree. Initially the number of users getting infected increases exponentially till almost all the users are infected.

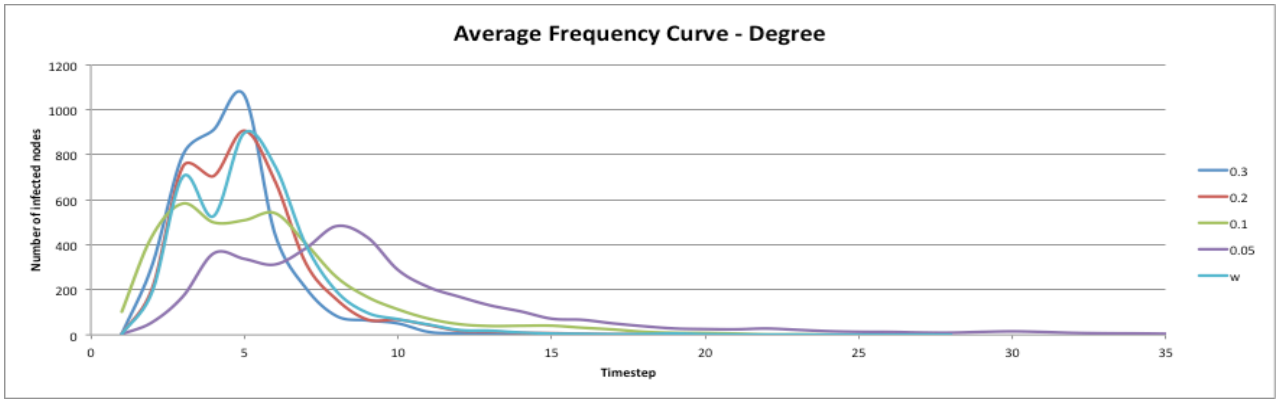


Figure 5

In Fig6a and Fig6b, we plot timesteps required to finish SI cascading for different centrality measures on unweighted graph (p=0.1) and weighted graph when number of starting nodes are varied. On the x-axis we have number of starting nodes. The results obtained shows that as number of starting nodes increases the time period is increasing which is not expected. This is due to many reasons. It's a very high probabilistic model, depending on random file we used for simulation, the limitation of Facebook network data, selection of starting infected users. We also observed more users are infected when number of starting nodes is increased hence it takes longer time to stop. Fig6 shows on an average starting users=3 performs better than others in both variations of Facebook network. Degree centrality measures outperform other measures.

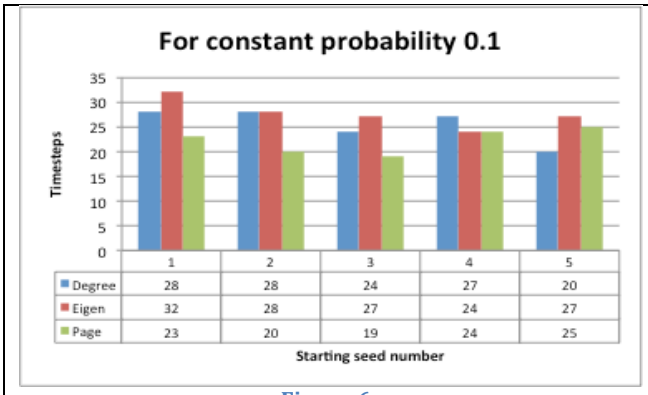


Figure 6a

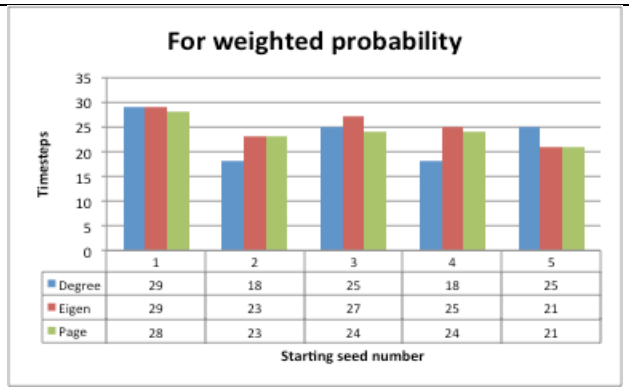


Figure 6b

SIR Model without degradation:

In this section, we discuss results from SIR model when probability of infection between users remains constant over increasing timesteps starting from the infected time. We plot radar graph for analyzing SIR model. In Fig7 'timestep n' represents total timesteps SIR model to stop when starting number of infected users are n. 'nodes n' represent percentage of total users(nodes),normalized on the scale of 20, who got infected in the model when starting number of infected users are n. We plot this for different probabilities (0.3,0.2,0.1,0.05 and w) of infection.

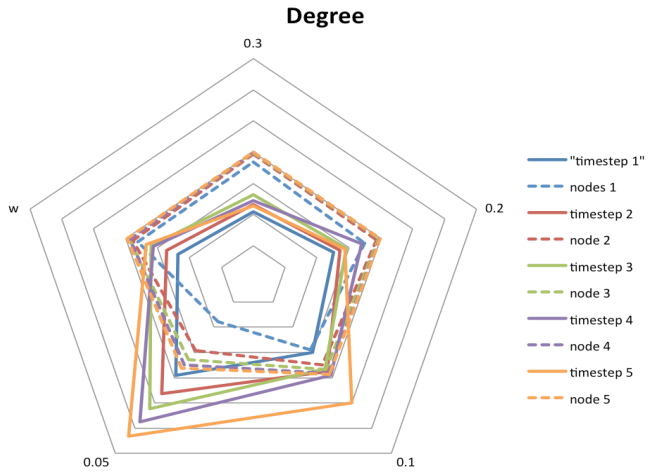


Figure 7

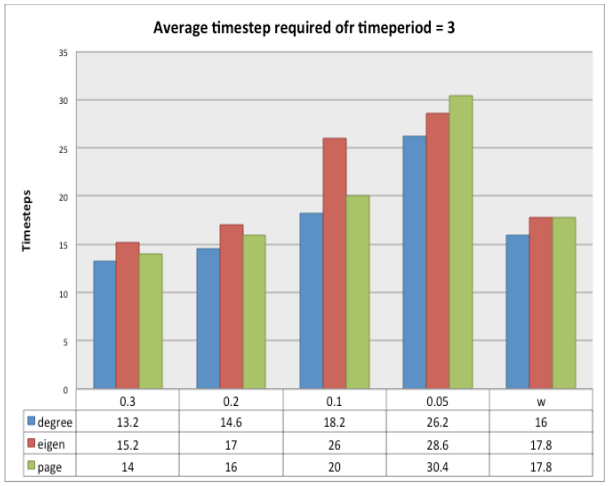


Figure 8

Fig7 plots when starting users are taken from degree centrality measure. We have similar graphs for other centrality measures. Fig7 shows result obtained with p=0.2 is very much similar to that of p=w. Thus concluding in real life scenario (w) on average information is spread by at least 20% of friends. For p=0.3 time taken by all variations of starting nodes is almost same and number of users infected is same. This is due to very high probability of infection. For p=0.05, due to very low probability time taken is highest for all variations and the number of users infected varies increasing with timeperiod of infection. Comparing p=0.1 and p=0.2, we observe number of users getting infected for variation in timeperiod is same. However, the time taken to cascade is higher when p=0.1. Thus, for viral infection p=0.2 stands best when a user with highest degree is chosen as a starting node of infection. As mentioned earlier readings of p=0.2 are similar to p=w (real data scenario). Thus, infecting node with highest value with degree distribution will spread information fastest and to the maximum users.

Fig8 shows average time taken for SIR model to spread the infection when timeperiod=3. This figure is a small part of data that is shown in Fig7. It can be observed that degree distribution performs best for viral information diffusion with all probabilities.

In Fig9a and 9b we plot timesteps required for SIR cascading for different centrality measures on unweighted graph (p=0.1) and weighted graph when number of starting nodes are varied. On the x-axis we have number of starting nodes. The results obtained shows that as number of starting nodes increases the time period is decreasing as expected. We also observed that number of users getting infected increases as we increase number of starting infected users. Degree centrality measures outperform other measures.

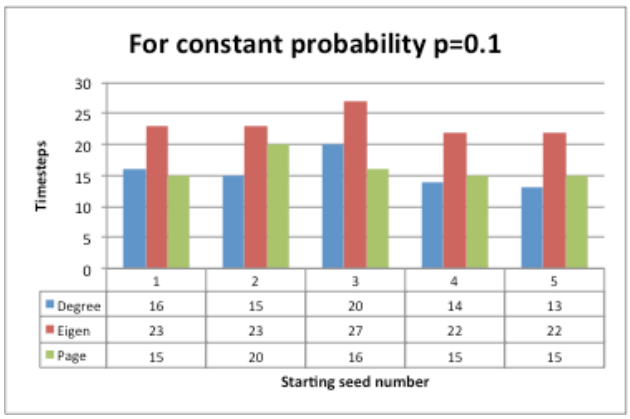


Figure 9a

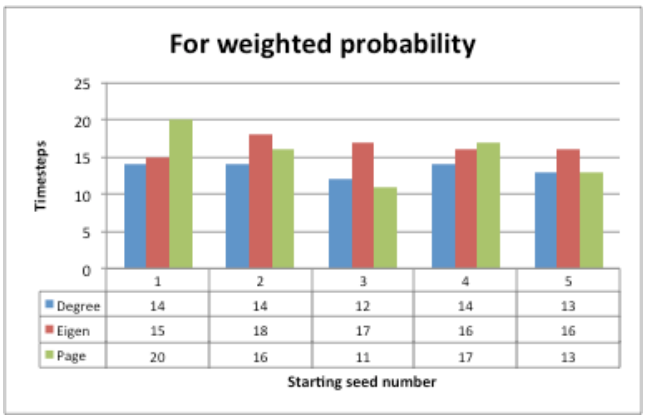


Figure 9b

SIR Model with 25% degradation:

In this section, we discuss results from SIR model when probability of infection between users degrades by 25% over increasing timesteps starting from the infected time. We plot radar graph for analyzing SIR model. In Fig10 'timestep n' represents total timesteps SIR model to stop when starting number of infected users are n. 'nodes n' represent percentage of total users(nodes),normalized on the scale of 20, who got infected in the model when starting number of infected users are n. We plot this for different probabilities (0.3,0.2 and w) of infection.

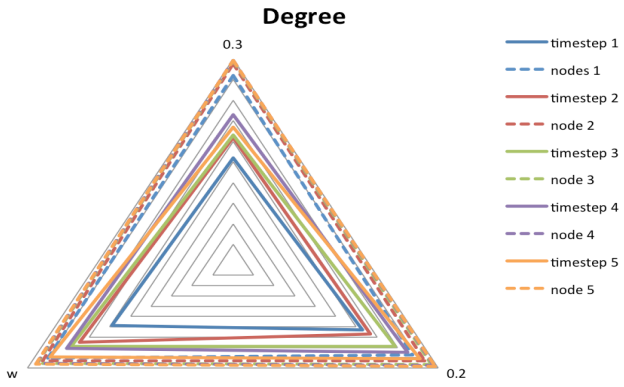


Figure 10

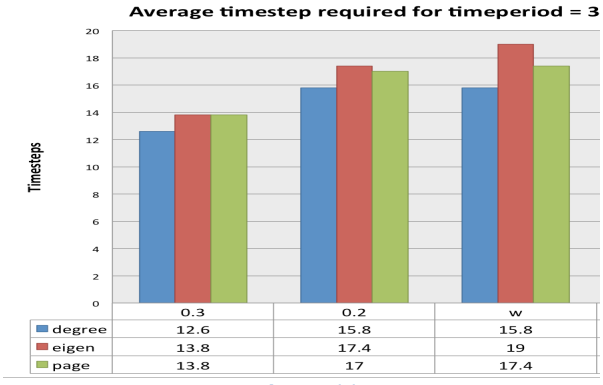


Figure 11

Fig10 plots when starting users are taken from degree centrality measure. We have similar graphs for other centrality measures. Fig10 shows result obtained with p=0.2 is very much similar to that of p=w. Thus concluding in real life scenario (w) on average information is spread by at least 20% of friends.

For p=0.3, number of users infected is same but time taken by all variations of starting nodes is not same. This is in contradiction with what we observed when probability of infection is not degrading in Fig7. Comparing p=0.2 and p=0.3, we observe number of users getting infected for variation in timeperiod is same. However, the time taken to cascade is higher when p=0.2. Thus, for viral infection p=0.3 stands best when a user with highest degree is chosen as a starting node of infection. However, p=0.3 meaning 30% of user's friends will share the information is not a real world scenario. As mentioned earlier, readings of p=0.2 are similar to p=w

(real data scenario). When we compare p=0.2 in Fig7 and Fig10, we observe that degrading probability follows the trend of increasing number of timeperiods infects more users in less time. Thus, infecting node with highest value with degree distribution will spread information fastest and to the maximum users.

Fig11 shows average time taken for SIR model with degrading factor of 25% to spread the infection when timeperiod=3. This figure is a small part of data that is shown in Fig10. It can be observed that degree distribution performs best for viral information diffusion with all probabilities.

In Fig12a and 12b we plot timesteps required for SIR cascading for different centrality measures on unweighted graph (p=0.1) and weighted graph when number of starting nodes are varied. On the x-axis we have number of starting nodes. The results obtained shows that as number of starting nodes increases the time period is not decreasing as expected. This is mainly due to the degrading factor. The infection effect of more number of users is decreased if we select more starting infected users. We also observed that number of users getting infected remains the same as we increase number of starting infected users. This is again due to the degrading factor. Degree centrality measures outperform other measures when we only consider top3 users.

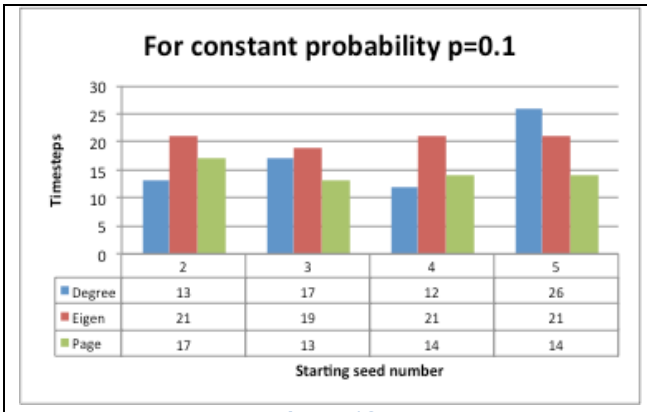


Figure 12a

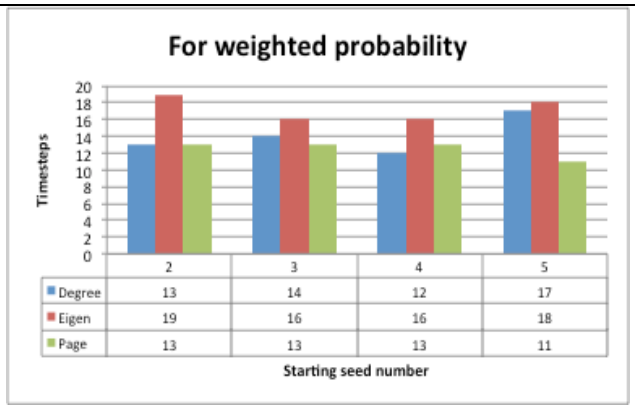


Figure 12b

7.Related Work:

Many studies have been done on studying information diffusion on social networks. Matt and Ken [1] tried to mix social networks and epidemic models. Woo and Chen [3] studied how information diffusion happens on web forums using SIR model. J.Yang and J. Leskovec[6] tried to model information diffusion in implicit networks developing a Linear Influence Model. Agarwal etl[7] studied how information diffusion in social networks affect user's social interests. Xu[8] model SIS on Facebook data and conclude all online users are tend to get infected due to mass connections in social network. Lada et al [9] studies weak ties are as important as strong ties for spreading novel information in social networks. There are many such numerous studies that try to relate information diffusion with various aspects of user's life. We couldn't find any published work that tries to solve the problem we address in our study with similar approach.

8.Conclusion:

We performed rigorous study of SI and SIR epidemic models for information diffusion on Facebook network. We simulated different real world social network scenarios. Degree centrality measure outperforms other centrality measures in almost all scenarios. Considering the real world scenario when information degrades over time, one should induce information to any top3 users with high degree centrality measure. We observe that p=0.2 for unweighted network performs very similar to real world Facebook network, thus concluding if information is induced in users who perform best at least 20% of users' friends would spread the information; thus making information reachable to almost every users of the network.

9.Bibliography:

[1]. Matt J Keeling and Ken T.D Eames. *Networks and epidemic models.* J R Soc Interface. 2005 September 22; 2(4): 295–307

[2]. J. McAuley and J. Leskovec. *Learning to Discover Social Circles in Ego Networks.* NIPS, 2012.

[3]. Jiyoung Woo, Hsinchun Chen: *An event-driven SIR model for topic diffusion in web forums*. ISI 2012: 108-113
[4]. Kermack, W. O.; McKendrick, A. G. (1927). *A Contribution to the Mathematical Theory of Epidemics*. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 115 (772): 700.
[5]. *http://snap.stanford.edu/data/egonets-Facebook.html*
[6]. J.Yang and J. Leskovec. *Information Diffusion in Implicit Networks.* ICDM 2010.
[7]. Divyakant Agrawal, Ceren Budak and Amr El Abbadi. *Information Diffusion In Social Networks: Observing and Influencing Societal Interests.* VLDB 2011
[8]. Bo Xu. *Information diffusion through online social networks.* ICEMMS 2010
[9]. Eytan Bakshy, Itamar Rosenn, Cameron Marlow and Lada Adamic. *The Role of Social Networks in Information Diffusion.* ACM WWW 2012.

10.Appendix:
All graphs, data, code is available on github. https://github.com/ajiwane/FNA