Prediction of Virality in Online Social Networks Using Core Periphery Structure

Social Network Analysis

Final Research Report

Summer Research Fellow
Indian Academy of Sciences, Bangalore

Student :
Vikramank Singh
Computer Engineering,
VES Institute of Technology

Guide :
Dr.Sudarshan Iyengar
Ph.D, Indian Institute of Science
Associate Professor, IIT Ropar

Prediction of Virality in Online Social Networks Using Core Periphery Strucuture

September 12, 2014

1 Introduction

Today information dissemination is very fast and easy. Presence of social networking sites and mash-ups with ever increasing user base has made information sharing fast and easy. Multimedia content such as news, video or image can spread in no time. Any such content that spreads across the social network is called a meme. One such example is the 'Ganganam Style' song uploaded on Youtube in August 2012. Within January 2013, the video managed to get 1 billion views. This is the perfect example for a piece of information attaining virality of epic proportion. Most of the times the spread of the information depends on initial days of its lifetime. A piece of information becoming viral in it's initial days is the most interesting. Our work aims at predicting the virality of a meme in its initial days of spread.

Virality prediction has found a lot of applications in the real world problems such as predicting the spread of epidemics [2]. By looking at the tweets posted by people on twitter and counting the number of tweets consisting of keywords pertaining to the disease influenza such as cold, fever, cough etc, the probability of influenza being an epidemic is predicted with a high efficiency. Given a (social) network, where the nodes are people and the edges of the graph represent the connections or the relationships between the people, we aim at predicting whether a meme will go viral in this network or not based upon the spreading pattern of the meme on this network in its initial days.
Another application is viral marketing[13]. When a company launches a new product or a brand, it might be enough for the company to identify highly influential people in an online community and advertise the product/brand to them. Stock market indicators[4], predicting the results of election[6], crime detection[3] are a few other areas where the applications can be found.

2 Previous Work

An enormous amount of study has been done in this direction before. First such study is based upon the content of the meme[7]. It says whether a meme will become viral or not depends on its content. (explanation). Another study is based upon the highly influential people in a group[8]. Essentially, it says that,

if a meme is getting shared between the highly influential people in an online community, its probability of going viral increases. It also devises methods to figure out these highly influential people in an online community. There are studies based upon the emotions of the people that a meme invokes. All these studies were quite rudimentary and based upon the content of the meme or the nature of the people spreading a meme. A seminal work has been done in this direction by L Weng, F Menczer, YY Ahn[9]. They say that taking the network structure and the way in which the meme is spreading over this network, it can be very efficiently predicted whether a meme is viral or not. Any online social networking site exhibits the property of being segregated into communities. If a meme is spread in multiple communities after few days of its launch, its probability of going viral increases. We are trying to take this approach further by concentrating on a different meso scale structure that is observed in such sites known as a core periphery structure.

3 Important definitions

3.1 Scale free Network

A network is known as a scale free network if its degree distribution follows a power law .
The proportion of nodes having degree k,

$$P(k) \propto k^{-\gamma}$$

where $2 < \gamma < 3$

An example of such a degree distribution is illustrated in figure 1.

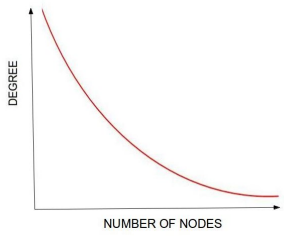

Figure 1: Power law degree distribution

Power law degree distribution implies that there are a large number of nodes in the graph with low degree and vice versa. So, there are hubs in the network albeit the number of hubs is very low.

3.2 Preferential Attachment Model

Preferential attachment model [10]is a standard algorithm for generating scale free networks given by Barabasi and Albert. The essence of the algorithm is that the scale free networks are evolved over time. Such a network starts its formation from a small number of nodes. As soon as a new node comes in the network, it tries to link with the existing nodes having greater degree.

A scale free network is shown in the figure 2.

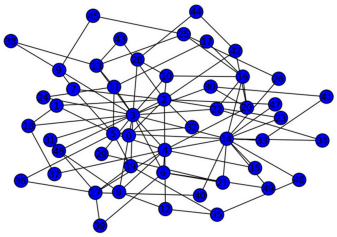

Figure 2: Example of a small scale free graph

The degree distribution for this graph is shown in figure 3.

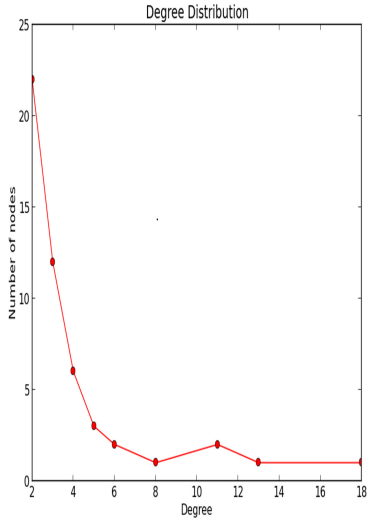

Figure 3: Degree distribution for the graph

3.3 Core Periphery Structure

Almost every scale free network tends to have a core periphery structure.[12] This means that whole of the graph can be divided in 2 kinds of nodes- core and periphery. Core is an ensemble of nodes that are very well connected amongst themselves as well as the periphery nodes. On the other hand, the periphery nodes are not so well connected amongst themselves.

An example of core periphery structure is shown in figure 4.

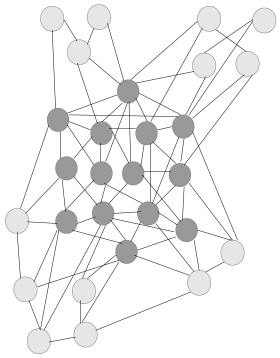

Figure 4: Example of a core periphery structure. Core nodes are denoted by dark grey nodes and peripheral nodes by light grey nodes

4 Our work

We wish to delve deeper in the process of spreading of a meme and come up with something non trivial. It is conceivable that leakage of a meme to the highly influential nodes in a network will have profound positive impact on its virality. One way of looking at an online community is that it is divided into communities. We focus at another facet of looking at such an online community known as the core periphery structure. In any social networking site, there is an ensemble of nodes known as the core. The core nodes generally form an elite group and have the power to convince the rest of the network to adopt an idea. Hence, the concept of using core periphery structure in the detection of virality is plausible. The use of the core periphery structure is the groundwork of our study. We unify the concept of predicting virality with community structure along with doing the same with core periphery structure.
In the paper – community structure and virality prediction, there are 4 baseline models used- random sampling model, simple cascade model, social reinforcement model and homophily model. It is shown that a viral meme spreads like the simple cascade model (simple contagion) as compared to the non viral memes that spread like social reinforcement or homophily model. We wish to show that there is something more going on inside the network except the community structure. It is shown that the social networking sites are evolved over time which can be described very well with the help of preferential attachment model that says that a node is likely to make connection with the existing nodes having a high degree thus supporting the idea of "rich gets richer". The networks formed in such a way are called scale free networks and their degree distribution follows a power law.

Now, research over the past years has shown that the existence of the core periphery structure in a scale free network is ubiquitous. Core nodes are the nodes which are well connected to each other as well as the periphery nodes, whereas periphery nodes are not so well connected to each other. We hypothesise that a meme gets viral if it leaks inside the core of a social networking site. However, leaking into the core is very tough for a meme, since core nodes do not get easily influenced by the periphery nodes albeit periphery nodes are

easily influenced by the core. We also hypothesize that a meme can leak inside multiple communities if it leaks inside the core. ie. , for a meme to infect multiple communities, it is mandatory for it to get spread to the core nodes. eg- the core of the twitter are the famous news sites and celebrities. Though a network can consist of more than one core, but in our work, we are using the assumption that the network consists of solely one core.
Our work consists of 3 parts :-

- Coming up with a synthetic network similar to a real world online social network.
- Developing a synthetic spreading model for the spread of meme on this network.
- Comparing our approach with real world networks.

4.1 Generation of the synthetic model

We are using our own synthetic network for the simulation purposes, and to better understand the process of spread of the meme. Our synthetic network consists the same features as the real world social networking sites given below :-

1. Scale free network
2. Divided into clusters/communities
3. Have a core periphery structure

We have made a scale free network according to the standard preferential attachment model. For also introducing communities in the network, we used the algorithm described in the paper "Cascades and breakdown in a scale free network" [14]where each new node make more links to its own community and less links to the nodes in the other community, both in a scale free fashion.

The procedure is shown in the figure 5:-
Since, it is quite feasible for a core periphery structure to coexist in a scale free network, we do not separately construct a core periphery structure, rather detect it using the random walker algorithm described in the paper "Profiling core periphery network structure by random walkers".
We detect the core of a scale free network with communities with the help of this model. Detection of the core is illustrated in the figure 6.

4.2 Synthetic spreading model

We divide the probability of a node infecting its neighbours in several categories in accordance with the type of the sender and the receiver nodes. Since, the core nodes are the highly influential nodes in any network, the probability of a core node infecting its neighbour (be it a core node or a periphery node) is the highest. Whereas, if it is a periphery node who is trying to infect its neighbour, then the probability of infection depends on many factors. If the neighbour is a core node, then the probability of infection is the least, as core nodes are the

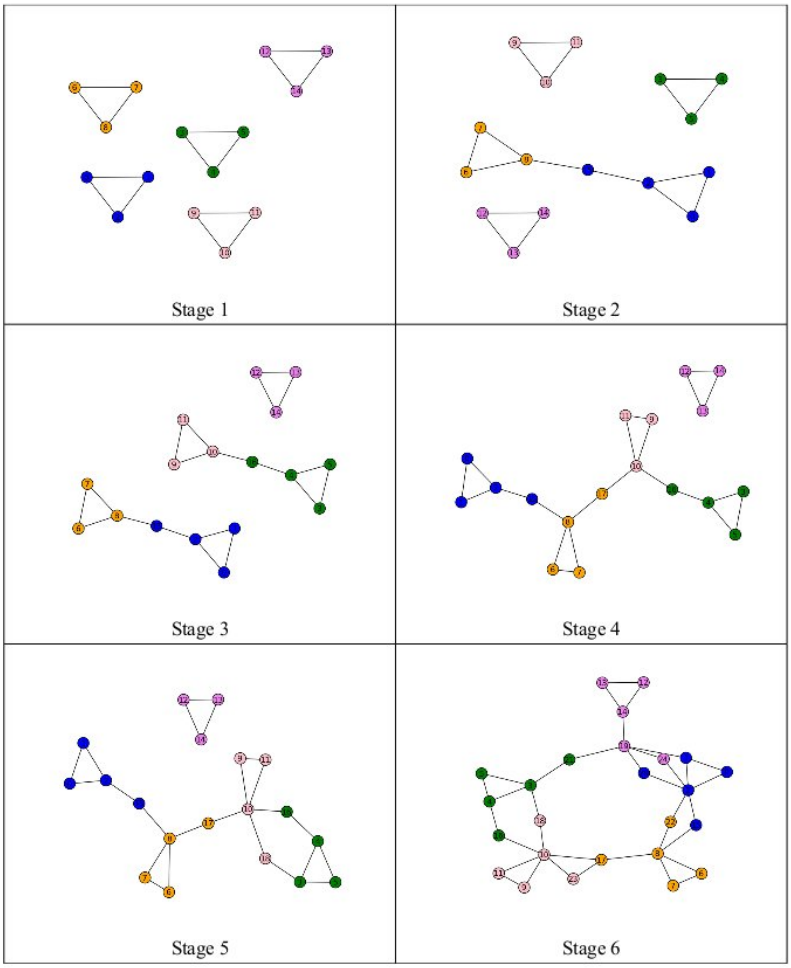

Figure 5: Construction of scale free networks with communities

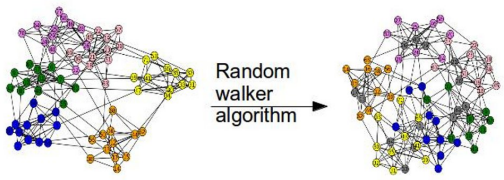

Figure 6: Detecting core using random walker algorithm

most important nodes in a network and they do not accept any information easily. If the neighbour is a periphery node, then the probability of infection depends upon the communities of both the nodes. If both the nodes are of the same community, then the probability of infection is higher as compared to the

case when both the nodes belong to two different communities.
Based on these observations, we can divide the probability of infection in various levels:

Sender node	Receiver node	Probability of infection
Core	Core	Highest
Core	Periphery	Highest
Periphery	Periphery(same community)	Medium
Periphery	Periphery(other community)	low
Periphery	Core	Very low

So, according to these probabilities, the probability of a meme leaking into the core is very tough, but as soon as a meme leaks inside the core, it gets disseminated into the whole network, as the core nodes have a very high potential to transmit the meme into the complete network.
A viral meme has a capacity to leak outside the boundaries of its community and enter into the core, subsequently reaching to every other community in the network and getting viral.

4.3 Comparison of the synthetic model with the real world networks

We have compared our synthetic network as well as the spreading model with the real world citation network as elucidated below:-

5 Experimental results

5.1 Synthetic network

We have generated synthetic scale free networks with communities having 1 million nodes and then applied our synthetic spreading model.

We have identified some of the factors that help in getting a meme viral which are described as follows:-

1. Fraction of initially infected nodes
2. Fraction of core nodes infected
3. Fraction of communities infected

According to these values, we can figure out whether a meme is going to be viral or not using the technique of machine learning. We consider a meme to go viral if it infects more than 90% of the total population.

1. Number of nodes in its own community that are linked to the core nodes
2. Shortest distance to the nearest core node
3. Its coreness value calculated according to the algorithm of the random walker's paper

4. Number of edges this node has to the other communities

5. Its own degree

6 Conclusion

Now, we are working with the citation network and the collaboration network formed in the field of high energy physics theory. In a collaboration network, the authors are the nodes and there is a link from one author to the another if both of them have co-authored the same paper. In this way, collaboration network are formed. In citation networks, research papers are the nodes and there is a directed edge between two nodes if one paper has cited the other. In this way, citation networks are formed. We obtained the citation network and the collaboration network for High Energy Physics theory from Stanford Large DataSet Collection. The citation network has 27770 nodes and 352807 number of edges. The collaboration network has 9877 nodes and 51971 number of edges. By applying the random walker algorithm for core detection, we can figure out the core authors in the collaboration network which are probably the authors who have collaborated with a large number of other authors and they tend to be the center of the collaboration network. Further, each author can be assigned a degree of coreness by the random walker algorithm. So, in the citation network, by looking at the authors of a paper, we probably can tell how much the impact of the paper will be which further can be proven by looking at the number of citations of the paper in the citation network.

References

[1] Aron Culotta , *Towards detecting influenza epidemics by analyzing Twitter messages, Proceedings of the First Workshop on Social Media Analytics Pages 115-122, 2010 .*

[2] Aron Culotta , Detecting influenza outbreaks by analyzing Twitter messages, Arxiv.org, July 2010.

[3] Matthew S. Gerber, *Predicting crime using Twitter and kernel density estimation, Decision Support Systems, May 2014.*

[4] Xue Zhang1,Hauke Fuehres,Peter A. Gloor, Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear", Procedia - Social and Behavioral Sciences, 2011.

[5] J. Berger, K. Milkman, *What Makes Online Content Viral?, 2012.*

[6] Gayo-Avello, Daniel, A Balanced Survey on Election Prediction using Twitter Data, 2012.

[7] J. Berger, K. Milkman, *Social Transmission, Emotion, and the Virality of Online Content, 2010.*

[8] M. Cha, K. Gummadi, Measuring User Influence in Twitter : The Million Follower Fallacy, 2010.

[9] L Weng, F Menczer, YY Ahn , *Virality prediction and community structure in social networks, 2013.*

[10] DIO Hein, DWIM Schwind, W König , Scale-free networks, 2006.

[11] D Easley, J Kleinberg, *Networks, crowds, and markets, 2010.*

[12] M Barthélemy, A Barrat , Velocity and hierarchical spread of epidemic outbreaks in scale-free networks, 2004.

[13] M Kitsak, LK Gallos, S Havlin, F Liljeros, *Identification of influential spreaders in complex networks, 2010.*

[14] J. Wu, Z.Gao, H. Sun , Cascade and breakdown in scale-free networks with community structure, 2006.

[15] SP Borgatti, MG Everett, *Models of core/periphery structures, 2010.*