**International Journal of Engineering Research and Development**
*e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com*
*Volume 10, Issue 6 (June 2014), PP.01-04*

# Data-Based Influence Maximization using Community Detection

Abhishek Singh
*Department of Computer Engineering Indian Institute of Technology Banaras Hindu University*

**Abstract:-** Influence maximization is the problem of finding a set of users in a social network, such that by targeting this set, one maximizes the expected spread of influence in the network .In this work, I am proposing a new algorithm drawing inspiration from the works done in the same field by [9] and [10]. The main motive behind this work is to emulate the relationship between the fields of Community Detection and Viral Marketing whose existence in the real world is intuitive. To do this I have taken inspiration from the work done by [10]. But as far as the actual diffusion process goes I am proposing using the Credit Distribution Model described by [9], which has been proven to be faster than any other existing models instead of the Heat Diffusion Model that has been used originally.

**Keywords:-** Influence Maximization, Community Detection, Viral Marketing

## I. INTRODUCTION

In today's era with the spread and penetration of the internet in our day to day lives.there has been an increased interest in analysing the social behaviour of individuals on online social networks. With people becoming more and more active on the web, these social platforms provide an almost alternative lifestyle. The field of social network analysis looks to explore possibilities in this alternate society and has with time grown out of the technical domain into business, economics, politics and many such fields.

The field of social network analysis can be broken down into multiple areas of research. Two such areas, research in which has found prevalence in recent times are Community detection and Viral Marketing. Lots of work has been done in these fields and huge amount of literature can be found on both of them. In this paper we have tried to understand and exploit these two fields and the way they connect with each other in real world networks.

Community detection as the name suggests is about detecting groups in a network of actors, actors being individuals who form the network, based on parameters that define the connection between these actors. These parameters may be as direct as friendship or kinship to more complex as shared interests, connections etc. These communities are important as they can be viewed as a platform for sharing knowledge, data, emotions, sentiments etc.

Viral Marketing may have found its initial use in business analytics but with growing research it has outgrown the field of computer research and found its application in myriads of domains. Viral marketing is an advertisement technique where one identifies a subset of 'actors' of the social network so as to obtain a "word of mouth" effect in promoting the product.

As we go about tackling these problems we find that they are connected at a very basic level by a special set of actors who are uniquely positioned in the network. These actors are known as influencers, as be it the survival of community or spreading of an innovation/idea/product in a network, *there is always a need for certain 'actors' who have influence over the rest of the community.* This can be seen in viral marketing as companies trying to identify an initial seed set of individuals to launch their producttp with the aim of achieving as much spread of word as possible. In terms of community detection, a certain set of individuals that share interests can be seen influencing each other. For example, an individual who shares interests in terms of his movie preferences with another individual would be compelled to watch a new movie if he/she sees positive feedback from the other. When this happens, the community detection algorithms increases the parameter used to represent the common interests among such individuals which results in detection of community. So, although both community detection and viral marketing are different areas, the underlying problem in both the cases is the same. This is the problem of maximizing the spread of influence in a network.

Formally, the problem of Influence maximization involves finding a seed set of users in an online social network to adopt an innovation and spread the information, so that the influence of innovation or product in the network is maximized. Influence maximization is a problem applied not only to tasks related to social networks but can be used for different other applications.'

Finding these 'few initial users' in these large networks is the major challenge of the problem and for which, huge amount of data is needed to be processed. Not only the processing of huge amount of data is required, timeliness of the processes are also important. For this, the time complexity of the process should be small. The most popular approaches in this area are greedy algorithm and/or optimizations to the greedy algorithms. It was observed that the individuals are connected to many others based on different interests. So a product/idea/event, which is to be 'spread' in the network, belongs to a particular community of the individual only. So, instead of targeting the entire network to find 'influential seeds', one can find the community first, where the probability of spreading is high and subsequently finding the 'seed'/'seeds'.

Using the community detection approach is also advantageous to us in scenarios where the locality of influence needs to be controlled. For example, a global company launching a new product would want the publicity to be maximized as well as evenly distributed across the globe. In this paper we try to combine the work done on Influence Maximization by [], with community detection and put forward an efficient algorithm for Influence Maximization in real world social networks. The aim of this algorithm is to try and optimize the number of nodes that are processed in the reverse k neighbour step for determining SNP value for nodes by emulating the relationship between community detection and viral marketing.

The rest of the paper is organized as follows: section 2 discusses about the related works, section 3 describes our method, section 4 provides experimental results and sections 5 concludes and discuss the future aspect of the work.

## II. RELATED WORK

The concept of spreading of an idea, or an innovation or influence for that matter was first studied in the field of economy, giving birth to viral marketing. Several models have been proposed to simulate this process. There are two models in particular that have gained widespread acceptance. These are the Linear Threshold Model and the Independent Cascade Model. The Linear Threshold model states that every node in a network has some threshold which is needed to be achieved after which it can become active. The Independent Cascade model on the other hand gives a probabilistic methodology to this. It proposes that any active node in a network would get a single chance to activate an inactive node, which it can do with certain probability. The problem of Influence Maximization was first studied by Domingos and Richardson[1][3]. Although there attempts at solving this problem were probabilistic. Since Domingos and Richardson studied the problem of Influence Maximization, the other researchers [5][7] have proposed greedy approaches rather than the probabilistic approach suggested by the formers. In 2003, Kempeet.al[3], proposed a greedy approach to solve what they viewed as a discrete optimization problem. After this several modifications and improvements have been proposed over this original approach.

AmitGoyal et al, proposed a credit distribution based Influence Maximization model, which worked by distributing credits to users for activating other users based on an action log of activities performed by all users. Another work combining viral marketing and community detections has been done by [10]. They have proposed a clustering method called H_Clustering for community detection. In our approach, we have taken inspiration from this clustering approach in identifying communities.

## III. PROPOSED METHOD

In this work, I propose to combine the work done by [9] and [10], in order to give a new algorithm for the Influence Maximization problem, that keeps into consideration the history of actions that have been taken by the users in determining their influence over each other. Also, it uses the concept of community detection and its relationship with the field of Viral Marketing. I propose that instead of the HDM model which has been used to simulate propagation of influence, the credit distribution model given by [9], would prove to be a far better and accurate model for the same. The proposed algorithm starts by assigning credits to users based on their activities as shown in the action log. Here's an example as to how this would work. Flixster(www.flixster.com) is one of the main players in the mobile and social movie rating business. Here, an action is a user rating a movie. In other words, if user v rates “The King's Speech”, and later on v's friend u does the same, it would be considered that the action of rating “The King's Speech” has propagated from v to u. Although unlike the actual algorithm proposed by [9], which is not a propagation algorithm, I propose to use this scanning of action log to determine probabilistic influence between any two users. Once we have these influence values we can apply the Independent Cascade Model on the network, with probability values that are actually significant. This approach is clearly more practical and hence more accurate than assigning random probability values to each of these edges.

Once we have a graph of users with connection between them showing probabilistic influence, I propose to use the work done by [10], on community detection based Influence Maximization. The community detection step starts by assigning similarity values between two nodes. Similarity for any given pair of nodes is defined as

$$Sim(u,v) = \frac{|adj(u) \cap adj(v)|}{\sqrt{|adj(u)| \times |adj(v)|}}$$

The stopping criterion for this algorithm is the modularity gain between iterations. It is defined mathematically as

$$Q(C) = \sum_{i=1}^{p} \left[ \frac{IS_i}{TS} - \left( \frac{DS_i}{TS} \right)^2 \right]$$

A graphical representation of the community detection process has been shown below.

Once this community detection process is over I propose to use the IC model for determining the Influence spread and determining seed set accordingly.

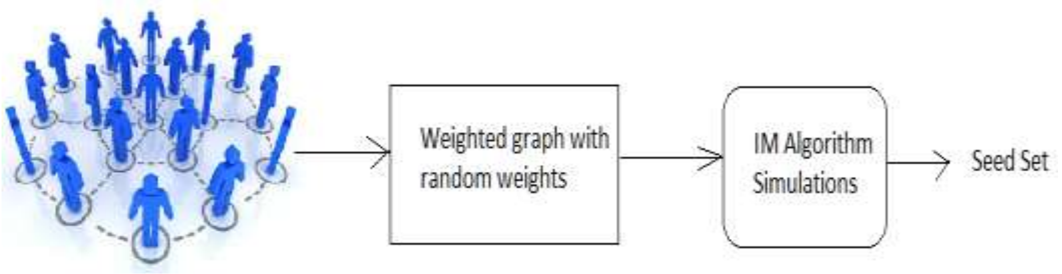

**Fig 1.Genereal Influence Maximization Process**

Vs.

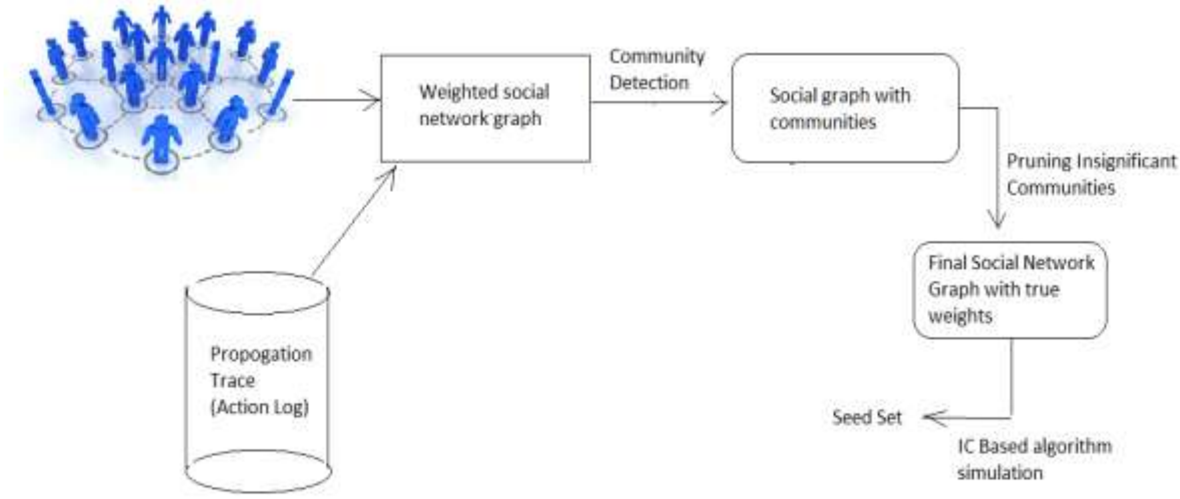
Weighted social network graph
Community Detection
Social graph with communities
Pruning Insignificant Communities
Final Social Network Graph with true weights
Propogation Trace (Action Log)
Seed Set
IC Based algorithm simulation

**Fig 2. The proposed algorithm**

The figures above show a comparison between general Influence Maximization algorithm vs the algorithm proposed in this paper. The algorithm simulations in the proposed algorithm are based on the Independent Cascade Model.

## IV. CONCLUSION

In this work I have proposed a new algorithm for Influence Maximization in Social Networks drawing inspiration from the work done in the same field by [9] and [10]. The aim of this work is to,

1. Redefine the way we take weights between users for the IC model. I have proposed that instead of random assignment of weights, we learn from the actions taken by users and use these action logs to determine mutual influence between these users.

2. Emulate the relationship between the fields of Viral Marketing and Community Detection in order to improve the time complexity of the algorithm.

To achieve this goal, I have drawn inspiration from the work done on CIM by [10]. Although instead of using the HDM, I propose to use any algorithm based on the IC model. As this would be more accurate.

## REFERENCES

[1]. P. Domingos and M. Richardson. Mining the network value of customers 2001

[2]. M. Richardson and P. Domingos.Mining Knowledge-Sharing Sites for Viral Marketing 2002.

[3]. D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network, 2003

[4]. S. M. Flip Korn and S. Muthukrishnan. Influence sets based on reverse nearest neighbour queries, 2000

[5]. W. Chen, Y. Wang, and S. Yang.Efficient influence maximization in social networks. 2009.

[6]. J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos,J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks.

[7]. W. Chen, Y. Wang, and C. Wang.Scalable Influence Maximization for Prevalent Viral Marketing in Large-Scale Social Networks, 2010.

[8]. Influence Maximization through Identifying Seed Nodesfrom Implicit Social Networks

[9]. AmitGoyal, Francesco Bronchi, Laks V. S. Lakshmanan, A data based approach to Social Influence Maximization.

[10]. Y. C. Chen, W. Y. Zhu, W. C. Peng, W. C. Lee, S. Y. Lee. CIM: Community based Influence Maximization in Social Networks, ACM 2010.