ISSN: 2321-7782 (Online)
Volume 4, Issue 2, February 2016
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Maximizing Influential spread in Network: A survey

Prabhat Mishra[1]
Computer Science
BMSCE,
Bangalore - India

Kaushal Kishore[2]
Computer Science
BMSCE,
Bangalore - India

Shivam Gupta[3]
Computer Science
BMSCE,
Bangalore - India

Monika Poddar[4]
Computer Science
BMSCE,
Bangalore - India

Asha G R[5]
Assistant Professor, Dept. of CSE
BMSCE,
Bangalore - India

Abstract: Social Media has always been a platform for marketing products and spreading innovations but this task may take a lot of budget which could have been invested in development of product. To market efficiently analysis of social media which can be thought of as graph needs to be done. This problem is universally identified as Influential maximization or viral marketing. In this paper we have analyzed various papers and their contribution towards this problem and further look community detection which when integrated with cascade models significantly reduces the number of computation.

Keywords: Diffusion model, Cascade model,. Viral Marketing, Community detection, Influence Maximization.

I. INTRODUCTION

This mining the social graph to get the useful information has always been an interest of researchers and this topic of finding most influence node in a graph has been researched in many journals. The literature reviewed is as follows: In the first step, a graph of any random user is taken out using API and application of social networking media, secondly, we surveyed the algorithms used to analyze and visualize the graph, and third step is, the utility and challenges of problem were identified

The amount of people in Social media like Facebook, Twitter and LinkedIn is just not to maintain and keep connection, this connection can be exploited in many areas of data mining to identify patterns and gaining some information out of them. One such area is marketing using Social media and in more formal term this also called "Viral Marketing "that uses word-of-mouth effect to increase profit. There are millions of users who are using Social network for sharing interests, posting information, commenting on various events and discussing various issues. The development and wide adoption of social networks among various users can be used to launch new product or compelling users to accept an innovation with a very significant amount of budget.To accomplish it in a cost efficient way we can target influencers in the social network, investing part of budget in getting them to adopt the product, for example by giving them product for free or at a discounted price. The belief is that these influencers will be able to motivate other people in the network to adopt the idea or product, generating a potentially large cascade in the social network. The social network analysis area can be divided into multiple fields of research. Various such fields, research in which has found widespread presence in recent years are Viral Marketing and Community identification or detection.Other approaches try to use heuristics completely treating both of them as independent areas and majorly work only on underlying dynamics of social networks.In this paper we have tried to exploit and understand these two areas and the way they connect with each other in real world social networks

Consider for example a start-up community which is growing today at an increasing rate. Everyone in one of two is available with an idea and a product but they don't have enough budget to develop as well as market their product this is where the technology steps in, this project will help them to ease their marketing and will make them able to target right communities to promote their product. This paper starts with survey followed by advantages of analysis then proposed method and design and concludes with the future work and improvisation

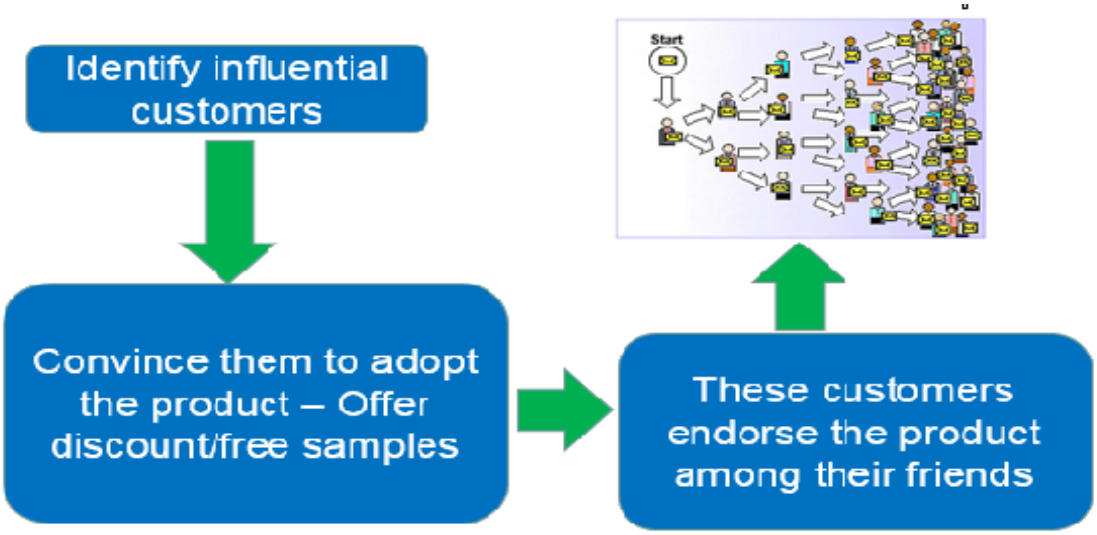


Fig. 1 : Pictorial representation of influential maximization problem

II. SURVEY

Influence maximization as a technique for viral marketing was first proposed by Domingos and Richardson [2001], with an algorithm based on probabilistic approach and Kempe et al. [2003]were the first to formulate influence maximization as a discrete stochastic optimization problem

Influence maximization: Given a social graph G =(V,E), a stochastic diffusion model on G, and a budget k, find a seed set A subset of V with $|A| \leq k$, such that the influence spread of A, $\sigma(A)$ under a diffusion model is maximized.

The problem was initially started long ago(Schelling and Granovetter) but the major landmark was established by [1].In [1]there are two diffusion models Independent cascade model[1] and Linear Threshold model[1].Independent cascade model is a diffusion model based on discrete steps in which user is activated only once. If user gets activated it is called a successful attempt and if he doesn't gets activated it is an unsuccessful attempt thus the model is a simulation of a coin flip. The event is successful with a probability with which two persons are attached (influence probability between them).Linear Threshold method is based on threshold value of the user. In this model each node is attached with other node with a weight and if the incoming weights to the user exceeds the threshold value of the node the node gets activated. The formula representing this is

$$\sum_{w \text{ neighbour of } v} bv, w \leq 1 \quad [1]$$

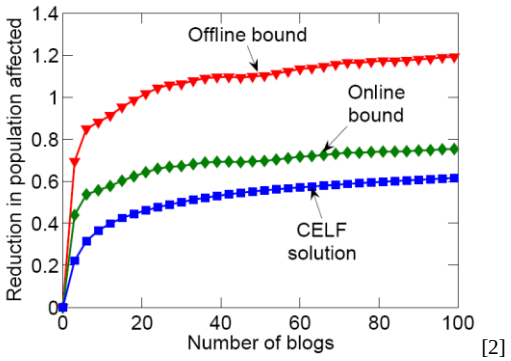
$$\sum_{w \text{ neighbour of } v} bv, w \leq \theta_v \quad [1]$$

Jure Leskovec et,al[2] states a different view to the problem he visualizes the problem as water distribution network where the problem is to place sensors to detect contamination so as to maximize their efficiency and cover the maximum area which is much same as influential maximization problem. It tells which blogs to study to detect cascades as effectively as possible.There are two parts of the problem Reward and Cost(location dependent).We have to select a subset of nodes A such that it maximizes the expected reward.If B is the budget of sensors.

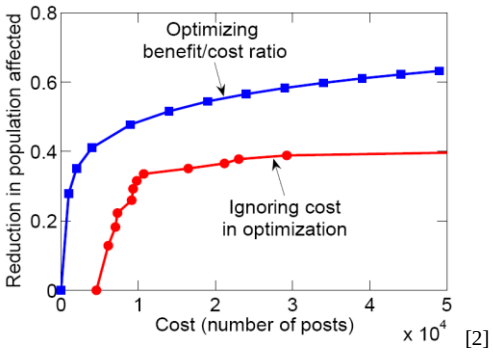
$$\max_{\mathcal{A}\subseteq\mathcal{V}} R(\mathcal{A}) \equiv \sum_i P(i)R_i(T(i,\mathcal{A}))$$

(reward for detecting contamination i) [2]

The solution provided for this is CELF(Cost Effective Lazy detection) which is a two pass greedy algorithm and is much faster than the standard greedy. CELF is near optimal.CELF achieves *½(1-1/e)* factor approximation



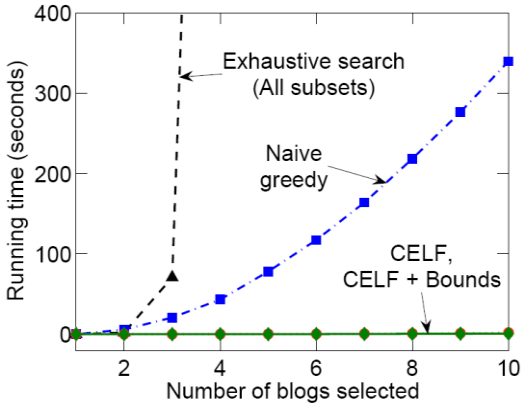
[2]



[2]

Fig. 2: **(a)** Bound of CELF **(b)** Cost of CELF

The CELF algorithm is 700 times faster than the greedy algorithm



[2]

Fig. 3: A comparison of naïve greedy and CELF

In [3] Wei chen et,al have worked on improving the efficiency of greedy based algorithm for solving influential maximization problem they have moved from general Monte carlo simulations to more reformed simulations and apart from this they have suggested new heuristics. In the paper they described three methods NewGreedyIC[3] , NewGreedyWC[3] , DegreeDiscountIC[3].NewGreedyIC reduces the size of graph performs linear scan on it and generates next best candidates for seed set and is improvement over simulations whereas DegreeDiscountIC is heuristic based in which suppose if u and v are two connected nodes (neighbour of each other)in graph and u is in seed set then degree of v will be discounted by one in the iteration of the algorithm before selecting v in the seed set. They both are improvement over CELF algorithm [2].The comparison of algorithm is given in the following table

Algorithms	Time complexity
Algorithm 1: GeneralGreedy	$O(knRm)$
Algorithm 2: NewGreedyIC	$O(kRm)$
Algorithm 3: NewGreedyWC	$O(kRTm)$
Algorithm 4: DegreeDiscountIC	$O(k \log n + m)$

[3]

Table 1: Comparison of Time complexity of different algorithms from [3]

Research that combines community detection algorithms with social (viral) marketing has been done by [4][5]. They have used a method called H_Clustering[4] for detecting communities in the social network

In [6] Amit Goyal et,al have improved the optimization of algorithm by suggesting an extension to the model proposed by [2] called CELF++ which not only eliminated Monte-Carlo simulations but also finds node in quadratic number of steps. It improves the efficiency by 35-55%.In CELF++ each user node has three information i.e. marginal gain, previous best marginal gain and a flag value. The flag value indicates the number of iteration the node was last updated because of which computation of marginal gain is not done in the next iteration. The data structure used in this algorithm was heap. The below table shows the comparison between CELF and CELF++

Dataset	Running time (min)			Avg. # node lookups		
	CELF	CELF++	Gain	CELF	CELF++	Gain
Hept WC	245	159	35%	18.7	13.4	28.3%
Hept IC	5269	2439	53.7%	190.5	101.5	46.7%
Phy WC	1241.6	667.7	46.2%	18.6	15.2	18.3%

[6]

Table 2: Comparison between CELF and CELF++ with Number of seeds=100.[6]

Till now we have understood that the traditional method of evaluating influential maximization is as follows

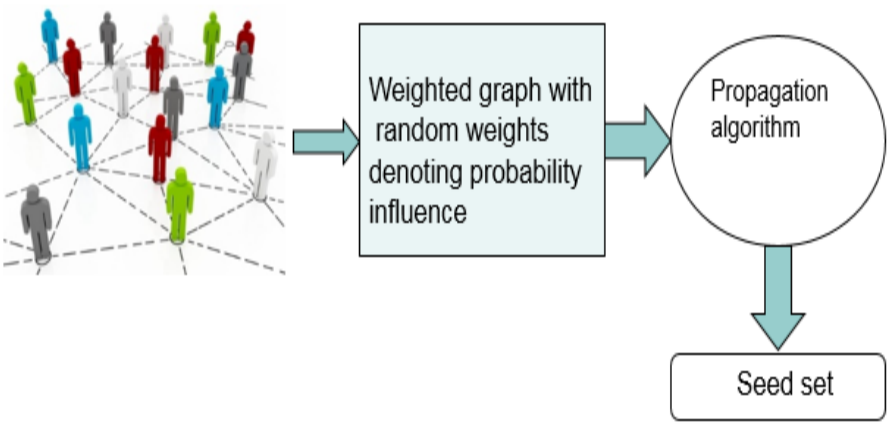


Fig. 4 : Traditional method of finding seed set in Influential Maximization problem.

Mark Granovetter [7] to model behaviour in a collective way, which aims at treating binary decisions problems, such as spreading of innovations, spreading rumours, voting and so on. He used the threshold model to explain the residential segregation, spiral of silence and spreading of rumours. According to Granovetters model, the "threshold" is "the fraction of people who make decision before a certain number of people before an actor does so. Another researcher ,T.Quian[8] have taken into account the effect of person in the network based on SNP value of the individual and used probabilistic approach to identify influential nodes. Currently three models have been in there like epidemic model[9,15],threshold model[6,10,11,12] and cascade model[1,3,12] to solve influential maximization problem. The flowchart of the model is shown in fig. 5.

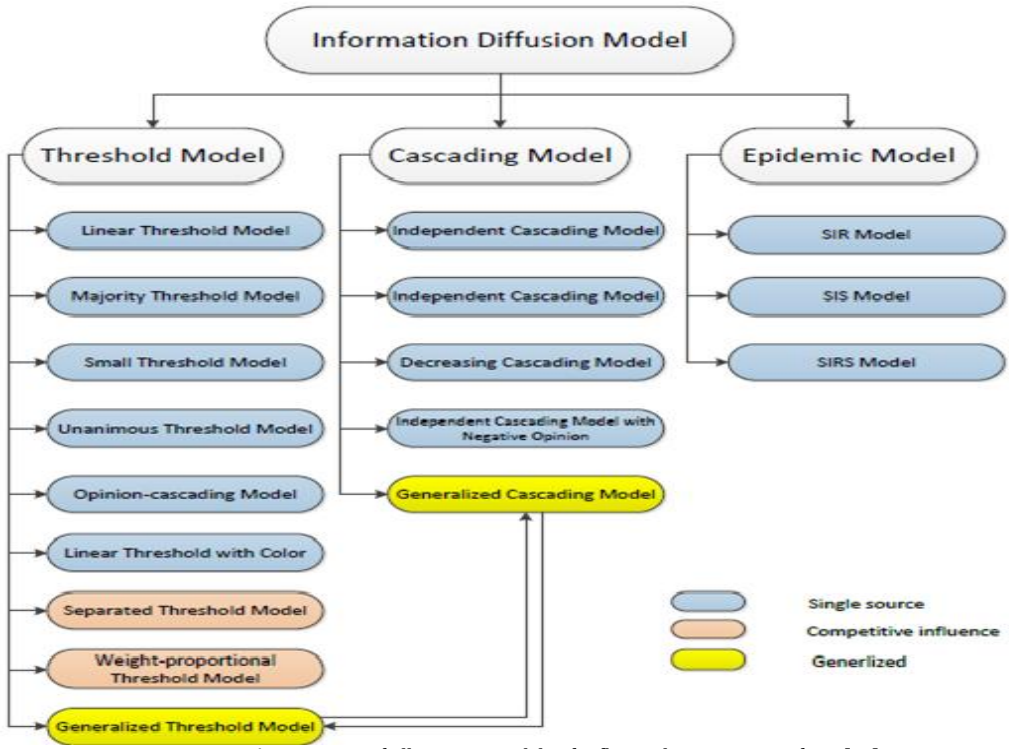


Fig. 5: An overview of all existing models of influential maximization from [17]

In [14] a data based perspective was proposed by Amit Goyal et,al which completely diverts the traditional method of influential maximization in which the probability between edges is simulated using Monte-Carlo.In this probabilities are learned which are not based on assumptions but using the real world data and credit based model these are learned and hence are less to error prone and accurate probabilities are computed and hence the need for Monte-Carlo simulations are eliminated.

Y.C chen et,al[16] have introduced a very innovative concept of integrating community detection algorithm with influential maximization so that irrelevant nodes can be removed and influential spread can be done at local community level thus the computations can be reduced.In[16] the paper CIM algorithm is suggested which has three phases.The first phase is coomunity detection then candidate generation and afterwards seed selection.This combination of data mining and community detection greatly enhances the efficiency of solving problem.Further the number of parameters to compute clustering are also reduced this freduces the number of computations to calculate seed set.Many algorithms like LDAG,SimPath,MIA has beed proposed in this context. One of these SimPath can be used to calculate similarity measure between two nodes of the graph.This similarity measure can be described as follows

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

The stopping condition of the above is based on modularity gain which is defined as

$$\text{Q(C)} = \sum_{i=0}^{p} \left[\frac{IS_i}{TS} - \left(\frac{DS_i}{TS} \right)^2 \right]$$

This work has further been improved by [17] by proposing CGA community based greedy algorithm in which they find top-K influential nodes within the community.In this the top influential nodes S is mined using extended Independent cascade model such that R(S) is maximised.The first step is construction of CDR(call detailed record) then community detection is used(MSN) and any dynamic greedy algorithm is applied on the obtained set.

We further look to extend the work done by [16,18] and propose a model by using well defined heuristics and community detection to further improvise the influential spread.

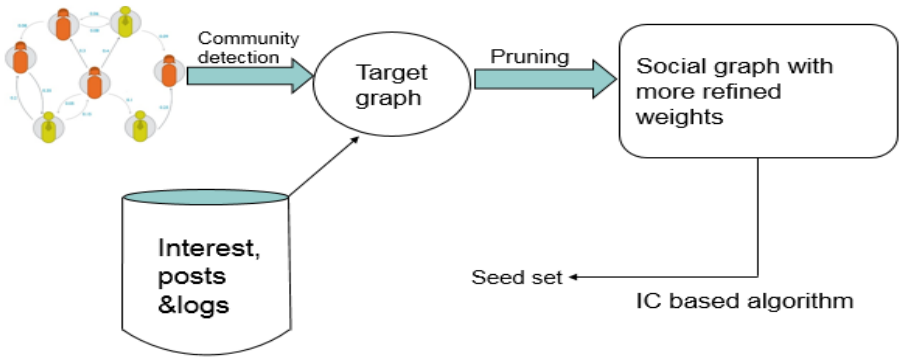
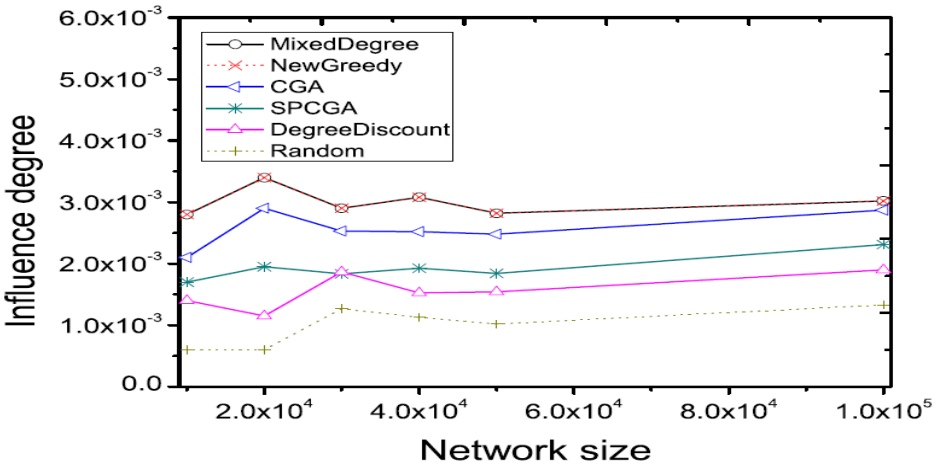
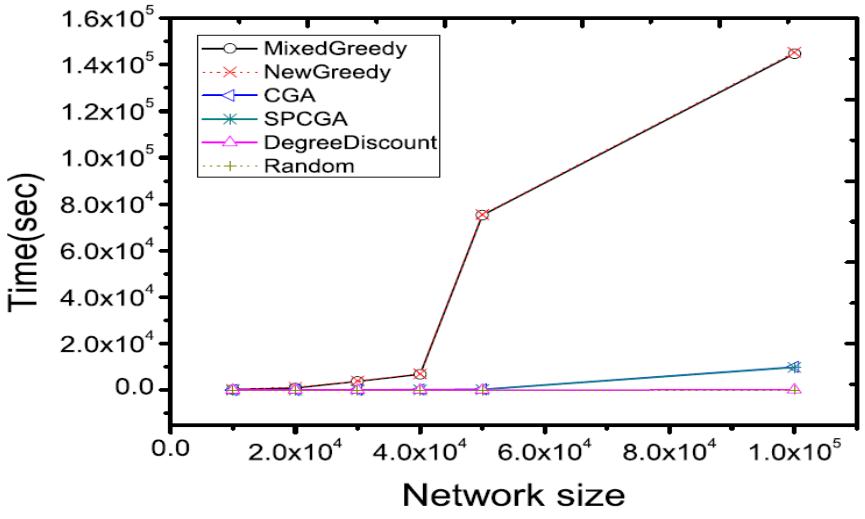


Fig. 6: Integration of community detection and heuristic approach in influential maximization



(a)



(b)

Fig. 7 : Influence degree and time vs network size (a) and (b) [18]

Another model which has been introduced in this context is CVAP (Cascade with Varying Activation Probability Model)[19] where they use online social network(OSN) and calculate the variation in activation probability between two users A and B to propagate the influence. In the paper they consider three phases of activation. One A is the first one to activate B, two.

B has been activated by many friends and third if A has activated him at last and concluded that activation probability increases first and then decreases with certain friends already attempted activation. Following figure demonstrates CVAP with real dataset (Renren).

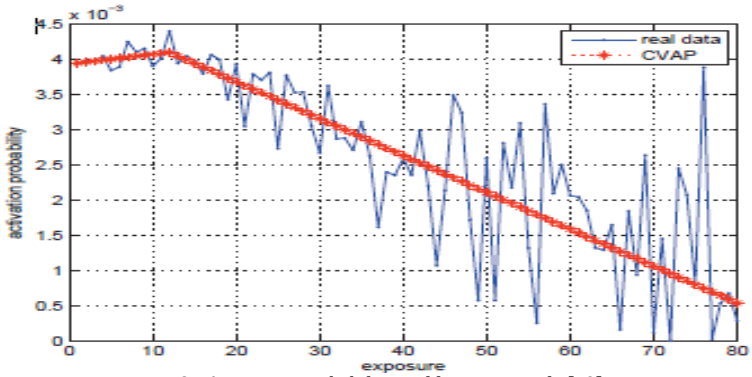


Fig. 8 : Activation probability and learning results[19]

III. Conclusion

We have made rigorous analysis of papers and did survey to find most appropriate heuristic to identify influencers in the network and integrated them to come up with a model.We further would like to improvise and work on it to make it more efficient so that viral marketing can done in reduced time complexity

Acknowledgement

The work reported in this paper is supported by the college through the TECHNICAL EDUCATION QUALITY IMPROVEMENT PROGRAMME [TEQIP-II] of the MHRD, Government of India.

References

1. D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network, 2003.
2. J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos,J. VanBriesen, and N. Glance. Cost-effective Outbreak detection in networks.2007.
3. W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. 2009.
4. M. Richardson and P. Domingos.Mining Knowledge-Sharing Sites for Viral Marketing 2002.
5. S. M. Flip Korn and S. Muthukrishnan. Influence sets based on reverse nearest neighbour queries, 2000.
6. A. Goyal, W. Lu, and L. V. S. Lakshmanan. Celf++: optimizing the greedy algorithm for influence Maximization in social networks. In 20th international conference companion on World wide web, USA,, 2011.
7. M. S. Granovetter. Threshold models of collective behavior. The American Journal of Sociology, 1978
8. T.Quian, Influence maximization through identifying seed nodes from implicit social networks-2010
9. M. Kermack. Contributions to the mathematical theory of epidemics. In Royal Society of Edinburgh Section A. Mathematics, volume 115, 1972.
10. C. Ning. On the approximability of influence in social networks. In Proceedings of the nineteenth annual ACM-SIAM symposium on discrete algorithms, 2008.
11. F. S. Roberts. Graph-theoretical problems arising from defending against bioterrorism and controlling the Spread of fires. In DIMACS/DIMATIA/Renyi Combinatorial Challenges Conference, 2006.
12. A. Borodin, Y. Filmus, and J. Oren. Threshold models for competitive influence in social networks. In Proceedings of the 6th international conference on Internet and network economics, WINE'10, 2010.
13. W. Chen, A. Collins, R. Cummings, T. Ke, Z. Liu, D. Rincon, X. Sun, Y.Wang, W. Wei, and Y. Yuan. Influence maximization in social networks when negative opinions may emerge and propagate. In In Proceedings of the 11th SIAM International Conference on Data Mining, 2011.
14. AmitGoyal, Francesco Bronchi, Laks V. S. Lakshmanan, A da.ta based approach to Social Influence Maximization.
15. J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In In SDM, 2007.
16. 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.
17. Recent Advances in Information Diffusion and Influence Maximization of Complex Social Networks.
18. Community-based Greedy Algorithm for Mining Top-K Influential Nodes in Mobile Social Networks, Yu Wang, Gao Cong, Guojie Song, Kunqing Xie, 2010.
19. Cascade with Varying Activation Probability Model for Influence Maximization in Social Networks, Zhiyi Lu, Yi Long, Victor O.K. Li,2015.

AUTHOR(S) PROFILE

Asha G R, received the B.E degree in Computer science & Engineering from Golden Valley Institute Of Engineering and M.E degree in Computer science & Engineering from BMS College Of Engineering, Bangalore. From 1999-Present she is Assistant professor in in Computer science & Engineering department BMS College Of Engineering, Bangalore.. Her research interest is in the field of Computer Networks.