CIM: Community-Based Influence Maximization in Social Networks

Report submitted in fulfillment of the requirements
for the Exploratory Project of

Second Year B.Tech.

by

Praveen Kumawat,Pushpendra Kumar Vaishya,Shivam Tomar

Under the guidance of

Dr. Lakshmanan Kailasam

Department of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY (BHU) VARANASI
Varanasi 221005, India
May 2017

Dedicated to

My parents, teachers,.....

Declaration

I certify that

1. The work contained in this report is original and has been done by myself and the general supervision of my supervisor.

2. The work has not been submitted for any project.

3. Whenever I have used materials (data, theoretical analysis, results) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

4. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Place: IIT (BHU) Varanasi

Date: 03/05/2019

Pushpendra Kumar Vaishya, Praveen Kumawat, Shivam Tomar
B.Tech. or IDD Student
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Certificate

*This is to certify that the work contained in this report entitled "**CIM: Community-Based Influence Maximization in Social Networks**" being submitted by **Praveen Kumawat,Pushpendra Kumar Vaishya,Shivam Tomar** (**Roll No. [17075044, 17075045, 17075059]**), carried out in the Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, is a bona fide work of our supervision.*

Place: IIT (BHU) Varanasi
Date: 03/05/2019

Dr. Lakshmanan Kailasam
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Acknowledgments

We would like to express our sincere gratitude to Dr. Lakshmanan Kailasam for his guidance and constant super- vision as well as for providing necessary information regarding the project and also for his support throughout the duration of the project. We would like to express our gratitude towards colleagues in developing the project and people who have willingly helped us out with their abilities.Last but not the least, we would like to thank our parents for their due consideration to our busy schedule, providing us constant motivation and support along with their unconditional love.

Place: IIT (BHU) Varanasi

Date: 03/05/2019 **Pushpendra,Praveen,Shivam**

Abstract

Social graph has been in use to study and to get insights about social structures for several reasons. One example of it is Influence maximization problem, i.e., to efficiently select the number of seeds to maximize information spreads, which has various purposes. But Social network is usually incredibly vast and most of the state of art algorithm also not very efficient.

This project is an implementation of Community Influence Maximization(CIM) a feasible and efficient algorithm for an influence maximization problem. CIM algorithm utilizes many other techniques of network analysis like clustering, community detection and diffusion models which can determine an of set seeds efficiently in comparison to other algorithms.

Contents

List of Figures

List of Tables

List of Symbols

Symbol	Description
Q	Modularity
h_0	Intial heat
h_t	Heat at time
λ	Eigen Values
α	Diffusivity

Chapter 1

Introduction

1.1 Overview

Influence Maximization problem is to determine a set of nodes in a social graph that maximizes the spread of influence [**?**]. The idea is to use word-of-mouth phenomenon seen in the exchange of information, this means that information travel to other of people and influence them. Realistic social networks are vast so brute force or greedy algorithms are not very efficient as the search space is too big. Heuristic algorithm, called enhanced greedy algorithm (EGA) [Ma et al.2008], has been proposed to select influential nodes greedily, it still incurs excessive computation[]. CIM divides the problem into three phases which reduces the search space significatly, three phases are (i) community detection, (ii) candidate generation, and (iii) seed selection.

1.2 Motivation for choosing this project

Influence maximization problem is very apt in today's marketing and social networking media domains. A scenario when a company would plan for a marketing campaign via social-networking media, it would try to target a small number for trial and check how much other people would get influenced to buy a productcan be very common

in today's world. This motivated us to implement the CIM as it had practical usage in many domains.

1.3 Organisation of the Report

CIM framework comprises of three phases: (i) community detection, (ii) candidate generation, and (iii) seed tuning. So our report has a chapter for each phase in the following order.

1. Community Detection - discovers the community structure of the network.
2. Candidate Generation - uses the information of communities to narrow down the possible seed candidates
3. Seed tuning - finalizes the seed nodes from the candidate set.

Chapter 2

Community Detection

Designing an effective clustering algorithm for CIM is an immediate task we face. We believe that the notion and principles of community can capture human nature in social networks. Thus, our community detection algorithm intends to detect the most natural communities of a social network without relying on heuristics, for example, the number of partitions.

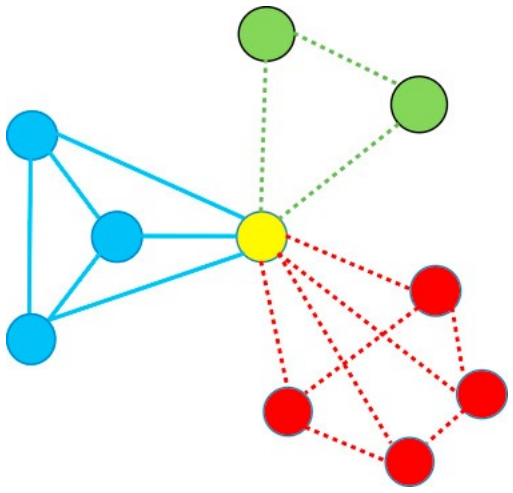

Figure 2.1 Communities in a social nework graph

APPROACH

In this approach for H Clustering, we use modularity, on a bottom-up approach to merge vertex with strong structure similarity into communities. Initially, we calculate structural similarity between a node and its neighboring nodes.

The similarity between two adjacent nodes u and v is defined as follows:

$$\text{Sim(u, v)} = |\text{adj(u)} \cap \text{adj(v)}| \ / \ |\text{adj(u)}| * |\text{adj(v)}|.$$

After the calculation of similarity scores, H clustering first treats each node as a community and then recursively merge two nodes if the similarity between these two nodes is the largest among their surrounding edges from each other. Next, we treat each newly created community as a node, and the process continues until a termination condition is reached which is decided by modularity gain. Borrowing an idea from the SHRINK algorithm [Huang et al. 2010], the definition of modularity gain is as follows.

Given a social network G = (V, E) and its clustering result C = c1, c2,..., cp, the modularity function is defined as:

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

where ISi is the summation of total similarity of nodes in cluster ci, DSi is the summation of similarity of nodes in cluster ci and other nodes in the network, and TS = is the summation of similarity between any two nodes in the network. For G, given two different clustering results C and C' , the modularity gain from C to C' is defined as $\triangle$ Q(C'$\rightarrow$ C)= Q(C') - Q(C), and if it is negative, then we terminate the process. Outliers are those homeless nodes whose neighbors are within only one community, and other homeless nodes are hubs that connect different communities.

Chapter 3

Candidate Generation

Since social networks are vast, the search space for selecting seeds with maximal influence spread is also huge. So, we need to reduce the number of candidate seeds effectively. The main issue faced in CIM is to narrow down the size of candidate seeds.

3.1 Approach 1

One of the easiest approach is to select centroid nodes of the k-largest communities in the social network as the k-influential seeds of that community.

However, this approach has some problems:-

(1) This approach implies that we should select more seeds in case of large communities.

(2) Some valuable information in the community structure is ignored.

The centroids of communities are natural candidates for seed selection, but the hubs which connect multiple communities should also be considered, as they can easily spread influences from one community to another.

3.2 Approach 2

In this approach instead of simply declaring some number of large communities as significant, we define significant communities as those who have large number of nodes than the average number of nodes a seed may influence in a given influence maximization task. By removing the insignificant communities, we can reduce the number of seed candidates.

1) Generally, nodes with a high degree usually present in large communities. So, our first strategy is to consider high degree nodes as the centroids of communities.

2) And our second strategy is to identify the nodes with a large summation of similarity scores as the centroids of communities

3) In this method, our strategy is to consider only the nodes with high degree and large score sum in significant communities and the hub nodes as candidates

4) We collect the top p% of high degree nodes and large score sum nodes in each significant community and all hub nodes connecting significant communities as the candidate set. (generally, we take p=10)

Chapter 4

Seed tuning using HDM

4.1 Overview

In this chapter, we will see how we modeled the information flow in the social network. The Linear Threshold and Independent Cascade Models are two of the most basic and widely studied diffusion models. But both models don't capture the temporal process of information. Heat diffusion model (HDM) is a realistic model that simulates social behavior in accordance with a physical phenomenon, heat flow (diffusion).

$$\frac{f_i(t+\Delta t) - f_i(t)}{\Delta t} = \alpha \Sigma_{j:(vj,vi)\in E} \left(f_j(t) - f_i(t)\right) = \alpha H f(t) \tag{4.1}$$

Heat equation

A Laplacian heat diffusion process calculates the heat distribution over a graph after at a specific time t:

$$h_i(t) = h_i(0) \exp(-\lambda_i t) \tag{4.2}$$

where $\mathbf{h}_0$ is the initial heat distribution, $\mathbf{h}_t$ is the heat distribution at time t and $\boldsymbol{\lambda}$ are the eigenvalues of the *Laplacian* of graph.

4.2 Seed tuning using HDM

CIM heuristically replaces initial seeds with the nodes remaining in the candidate set to test whether we can increase the influence spread. This process allows us to tune the influence spread under different parameter settings in HDM, including flow duration, activation threshold, and thermal conductivity.

In our implementation primary reason to tune the seeds is to prevent seed clustering or shortage of seed in communities i.e. a community shouldn't have too many or too less seed. Two evaluation metrics, left and seed load, are defined as the heuristics to determine the which node to remove from set of seed and which node to add in set of seeds.

Definition 8 (Left)- Given a social network G = (V, E) and a set of significant communities Cs = {c*1*, c*2*, . . . , cq}, suppose that we have selected a set of initial seeds, PS. After running HDM on G with PS, we obtain a set of active nodes I P S (G). We define a function,

$$left(c_i') = |\{u|u \in c_i' \text{ and } u \text{ is a non-activated node }\}| \tag{4.3}$$

Definition 8 (Seed Load). Given a social network G = (V, E) and a set of significant Cs = {c*1*, c*2*, . . . , cq},

$$\text{seed load}\,(c_i') = \frac{\text{size}\,(c_i')}{|\{u|u \in c_i' \cap PS\}|} \tag{4.4}$$

on seed$_{\text{l}}$oad (c_i') indicates whether too many seeds are selected from c i . When seed load(c i) is small, there are too many seeds in c i .

4.2.1 Swapping

The tuning takes r iterations to test the swapping heuristically.In each iteration (where$1 \leq \ell \leq r$) , CIM first selects the node with the maximal priority node from community C'_i in C_s which has ℓ -largest left (c'_i) as an add-node.Then CIM selects a delete node, delete-node, from PS, where the delete-node is the minimum priority node in C'_j which has minimum ℓ -largest left (c'_i). Finally, we test whether we should substitute an add-node for the delete-node. If the influence spread after swapping increases, we formally make the swap and continue the process.

for $\ell = 1$ to r **do** // initial seed tuning
  $T \leftarrow PS$; // *a temp set*
  $add\_node \leftarrow$ select maximum priority node $u \in candi\_set.c'_i$ where *left* (c_i) is top-ℓ maximum in C_s;
  $delete\_node \leftarrow$ select minimum priority node $v \in c'_j \cap PS$ where $seed\_load(c'_j)$ is minimum in Cs;
  replace the $delete\_node$ with add_node in T;
  $I_T\ (G) \leftarrow$ execute HDM on G with T;
  if $I_T(G) > IM$ **then**
    $PS \leftarrow T$; $IM \leftarrow |I_T(G)|$;
Output nodes in PS as seed nodes;

Chapter 5

Experimental Results

5.1 Social circles: Facebook

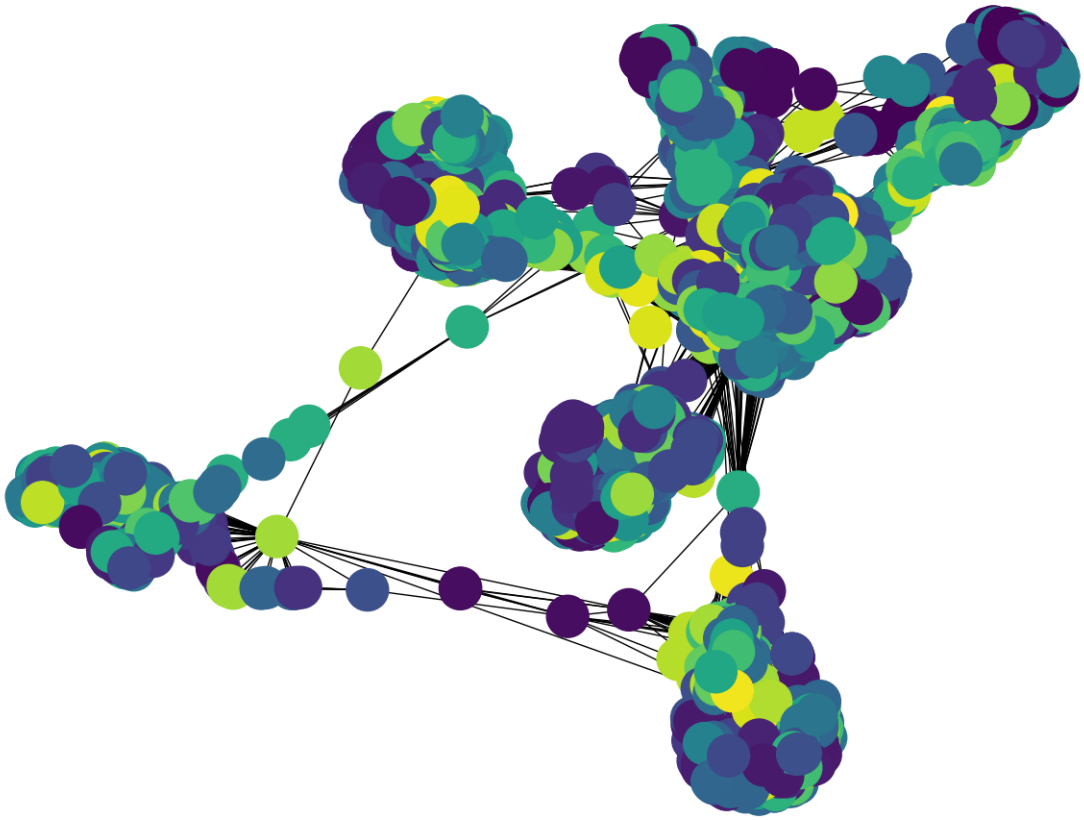

Figure 5.1 Communities in Social circles: Facebook

Nodes with different colour represent different communities. These communities are derived from hierarchical clustering using similarity score as modularity function and modularity gain as terminating condition.

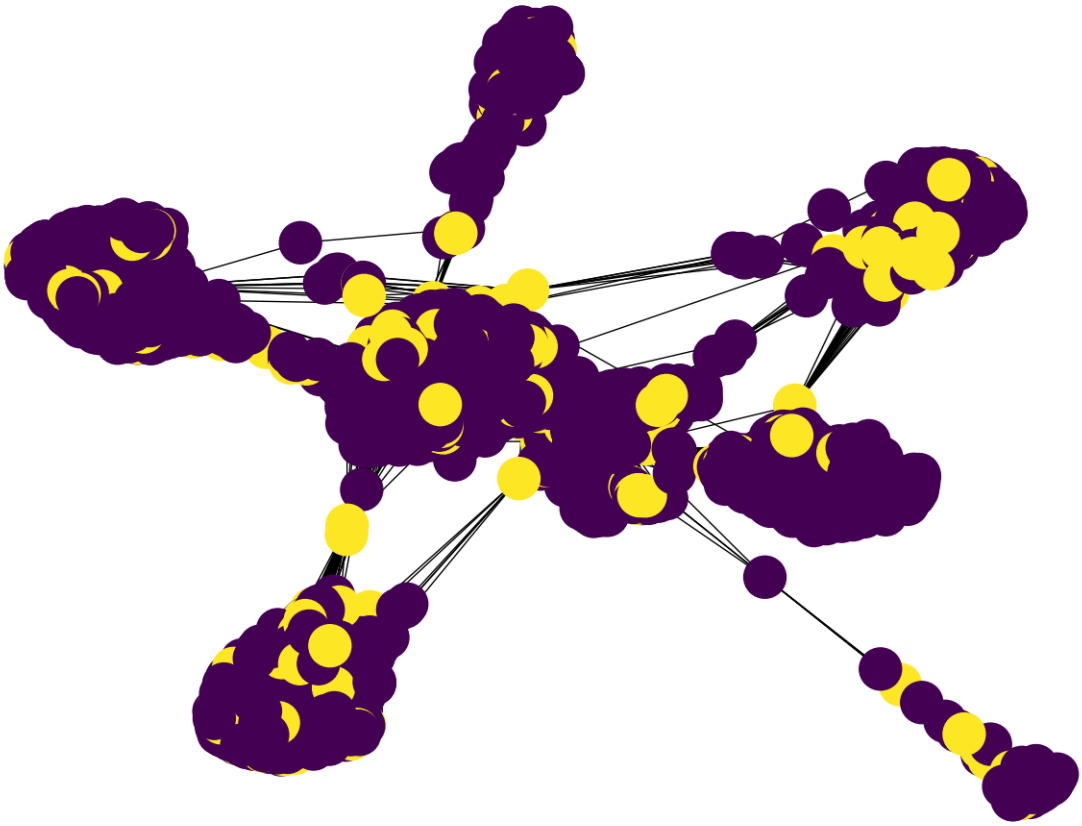

Figure 5.2 Final activated nodes

Yellow nodes are activated while purple are not activated. Here theta represents threshold value influence exerted on a node by all its active neighbors and alpha is 1 for simplicity.

total no. of nodes	time	θ	α	initial heat	activated nodes
4039	2 seconds	5	1	20	737

Table 5.1 Influence spread

Community detection	HDM	Total time
2.78	132	660

Table 5.2 Efficiency

5.2 Zachary's karate club

Figure 5.3 Communities in Zachary's karate club

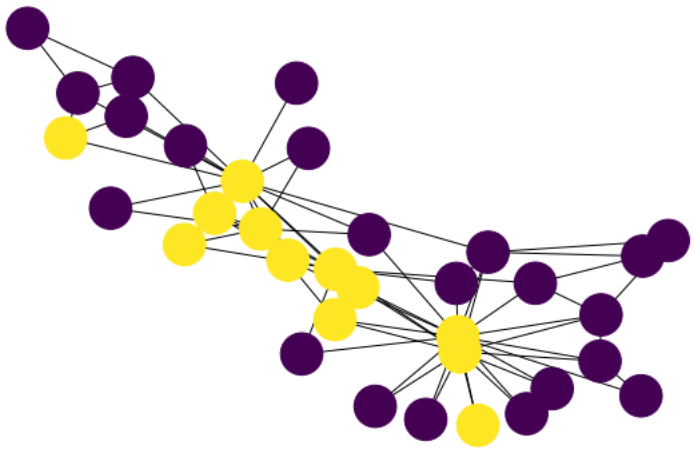

Figure 5.4 Final activated nodes

total no. of nodes	time	θ	α	initial heat	activated nodes
34	2 seconds	5	1	20	14

Table 5.3 Influence spread

Community detection	HDM	Total time
0.02	0.03	0.65

Table 5.4 Efficiency

Chapter 6

Conclusions and Discussion

In this project we have explored a new way to find seeds in influence maximization problem which was earlier considered a very computation heavy task. Social networks naturally cluster together, this property has been used to reduce the search space. More heuristics have been used to reduce the overhead in this algorithm, like a larger community will require more number of seeds. All the heuristics that are used are not taken for the cost of performance of algorithm which can be seen from the fact that influence spread are at par with other algorithms.

Future Directions

Tough this project was based on applying the CIM algorithm on social networks. Influence maximization problem can be interpreted differently in other networks/datasets. CIM algorithm can be applied on varied networks to get insights about the network. Some of the other interpertations are listed below.

- In a network of events we can identify most influential events that happened in history.
- In a network of political landscape we can identify the most influential political

influential persons.

- In a E-commerce network it can be used to identify target customers/markets.

Bibliography