LEBANESE AMERICAN UNIVERSITY

Influence Maximization in Social Networks:

A Multi-objective Approach

By

Ghinwa Bou Matar

A thesis

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science

School of Arts and Sciences

May 2019

THESIS APPROVAL FORM

Student Name: Ghinwa Bou Matar I.D. #: 201506323

Thesis Title: Influence Maximization in Social Networks: A Multi-objective Approach

Program: M.S. in Computer Science

Department: Computer Science and Mathematics

School: Arts and Science

The undersigned certify that they have examined the final electronic copy of this thesis and approved it in Partial Fulfillment of the requirements for the degree of:

Masters of Science in the major of Computer Science

Thesis Advisor's Name: [Redacted]

Signature: [Redacted] Date: 14 / 05 / 2019
Day Month Year

Committee Member's Name: Dr. May Hamdan

Signature: [Redacted] Date: 14 / 05 / 2019
Day Month Year

Committee Member's Name: Dr. Azzam Mourad

Signature: [Redacted] Date: 14 / 05 / 2019
Day Month Year

THESIS COPYRIGHT RELEASE FORM

LEBANESE AMERICAN UNIVERSITY NON-EXCLUSIVE DISTRIBUTION LICENSE

By signing and submitting this license, you (the author(s) or copyright owner) grants the Lebanese American University (LAU) the non-exclusive right to reproduce, translate (as defined below), and/or distribute your submission (including the abstract) worldwide in print and electronic formats and in any medium, including but not limited to audio or video. You agree that LAU may, without changing the content, translate the submission to any medium or format for the purpose of preservation. You also agree that LAU may keep more than one copy of this submission for purposes of security, backup and preservation. You represent that the submission is your original work, and that you have the right to grant the rights contained in this license. You also represent that your submission does not, to the best of your knowledge, infringe upon anyone's copyright. If the submission contains material for which you do not hold copyright, you represent that you have obtained the unrestricted permission of the copyright owner to grant LAU the rights required by this license, and that such third-party owned material is clearly identified and acknowledged within the text or content of the submission. IF THE SUBMISSION IS BASED UPON WORK THAT HAS BEEN SPONSORED OR SUPPORTED BY AN AGENCY OR ORGANIZATION OTHER THAN LAU, YOU REPRESENT THAT YOU HAVE FULFILLED ANY RIGHT OF REVIEW OR OTHER OBLIGATIONS REQUIRED BY SUCH CONTRACT OR AGREEMENT. LAU will clearly identify your name(s) as the author(s) or owner(s) of the submission, and will not make any alteration, other than as allowed by this license, to your submission.

Name: Ghina Ben Matar

Signature:

Date: 22 / 04 / 2019
Day Month Year

PLAGIARISM POLICY COMPLIANCE STATEMENT

I certify that:

1. I have read and understood LAU's Plagiarism Policy.
2. I understand that failure to comply with this Policy can lead to academic and disciplinary actions against me.
3. This work is substantially my own, and to the extent that any part of this work is not my own I have indicated that by acknowledging its sources.

Name: Ghin

Signature:

Date: 22 / 04 / 2019

Day Month Year

ACKNOWLEDGMENT

This thesis was only possible due to the support of many people in my personal and professional life.

I cannot thank enough my advisor, Dr. Faisal Abu-khizam, for his major support throughout my entire journey at LAU. His professionalism and knowledge guided me through all the confusions, and his continuous encouragement, support and guidance made this work possible. I would also like to take this opportunity to thank my committee members, Dr. Azzam Mourad and Dr. May Hamdan for offering generous help and support.

At last, I would like to thank my family and friends for encouraging me to succeed and for bearing with me throughout this research process and during all the years of graduate studies.

Influence Maximization in Social Networks:

A Multi-objective Approach

Ghinwa Bou Matar

ABSTRACT

The main challenge in viral marketing, that is powered by social networks, is to minimize the seed set that will initiate the diffusion process and maximize the total influence at its termination. The aim of this thesis is to study influence propagation models under the influence maximization problem and to investigate the effectiveness of a new model that is based on a multi-objective approach. We propose a *Depth-Based Diminishing Influence model (DBDM)* that is based on adding nodes to the seed set by considering influenced in-neighbors and how far these in-neighbors are from the initial activated set. As an enhancement to our approach, we used a clustering mechanism to help increase the influence spread. Several experiments were conducted to compare between our approach and previous work. As a result, the selection of the seed set under the DBDM model boosted the influence spread substantially compared to previously proposed models.

Keywords: Influence Propagation, Diffusion Models, Influence Maximization, Social Networks, Viral Marketing

TABLE OF CONTENTS

Chapter	Page
I Introduction	1
II Preliminaries	4
2.1 Types of Graphs	5
2.2 Influence Propagation	7
2.3 The Fixed-Distance Influence Propagation Problem	10
2.4 The Influence Maximization Problem	10
III Related Work	12
IV Proposed Model	15
4.1 The Multiple Path Asynchronous Threshold Model	15
4.2 The Depth-Based Diminishing Influence Model	17
V Influence Maximization Algorithm	20
5.1 Breadth First Search(BFS)	20
5.2 Influence Maximization Algorithm	21
VI Experimental Analysis	24
6.1 Model Comparison	24
6.2 Comparison on Different Type of Graphs	27
6.3 Comparison Between Different Series	31
6.4 Clustering	32
VII Conclusion and Future Work	35

LIST OF FIGURES

Figure	Page
1 A dense Facebook Graph	6
2 Sparse graph representing Facebook messages	6
3 Bitcoin Hub Graph	7
4 Sample Graph	17
5 Influence spread achieved under MAT and DBDM models on PGP Data-set	25
6 Influence spread achieved under MAT and DBDM models on Wiki- vote Data-set	26
7 Influence spread achieved on various data-sets	27
8 Visualization of the Data-sets	28
9 "Brightkite" Data-set	29
10 Visualization of "Brightkite" Data-set	30
11 Performance Comparison on Different Series: a Socfb-Brandeis99 dataset b fb-messages dataset c ia-radoslaw-email dataset d socfb-swarthmore42 dataset	31
12 The Effect of Clustering	34

LIST OF TABLES

Table		Page
1	Symbols used in this thesis	5
1	Social Network Data-sets	25
2	Various Types of Social Network Data-sets	28

Chapter One

Introduction

Online social platforms have been intruding our lives tremendously in the past years. Nowadays, Facebook, Twitter, and Instagram are the main source of information and form the basic means of communication in a real-world setting. Facebook official news proclaimed that the number of daily active users reached 1.49 billion users for September 2018. In June 2017, Facebook celebrated 2.01 billion monthly users [3], this number reached 2.27 billion users by September 30, 2018 [1]. This huge increase in the number of people engaged in social network platforms, in less than 1.5 years, shows that these platforms are widely spreading all over the globe. Due to this growth, many companies that base their work on viral marketing, like media co-operations, strive to reach their target audience through those platforms.

The immense number of users prompted some businesses to shift to the online world. Information can spread faster when using a highly connected network that is capable of reaching and affecting every person in one way or another. Facebook researchers found out that each user present on the Facebook network is connected to every other person by an average of three and a half other people [2], thus by using the share option that is provided by Facebook every news/campaign can reach almost the entire population of the Facebook graph.

With the growing use of online social networks as a medium for information propagation, they became a reliable and powerful sources of “information communication.”

This fact led to the exploitation of such power to spread false news or impose specific campaigns to overtake others. For example, during the 2016 US presidential elections many argued around the possibility that Donald Trump would not have won the elections if it wasn't for the influence of the false news that was spread all over the internet. Statistics showed that 62% of the US adults use social media to get news and many people believe what they read [9]. In the case of the presidential elections, most of the stories that were spread at that time were in Donald Trump's favor.

In a viral marketing campaign, the main target is to maximize the number of influenced people in a specific period of time. Thus, a company must spread advertisements to eventually reach as many people as possible. This approach raises the important question of how to select a set of individuals who are able, when adopting the campaign, to influence the majority of the targeted network. Influence propagation studies the technical way in how information overtakes the network. Independent cascade model and linear threshold model are the most widely used to illustrate the diffusion process. Those models are based on selecting a small seed set that will eventually propagate the news to the majority of the network.

In this work, we introduce a new model that considers the real world setup on how information diffuses through the network. In our *Depth-Based Diminishing Influence model*, we took into consideration that as far as you go from the initial seed set that carries the campaign, the influence over the individuals will decrease accordingly. After the seed selection process, we will use weighted edges to decide whether the individual will be influenced through the diffusion process or not. Those weights will represent the intercommunication factor between two individuals on the network. We measure the effectiveness of our model by producing a set of influenced nodes at the termination of our algorithm.

After extensive experiments that were performed on our model in comparison with the MAT model, which was recently introduced by Wang et al. in [33]. The experimental results show that our *Depth-Based Diminishing Influence model* along with our maximization algorithm performs better than the proposed MAT model. After that,

We introduced a clustering approach. This approach will help in further scattering our selected seed set by applying our algorithm on the obtained sub-graphs from the clustering technique. The results showed that the influence spread increased with a significant decrease in the selected seed set.

The rest of the thesis is organized as follows. Chapter II presents the preliminaries needed throughout this thesis. In Chapter III, we discuss the relevant previously known adopted approaches. Chapter IV presents the diffusion models and the influence maximization techniques. Chapter IV presents our proposed *Depth-Based Diminishing Influence model*. In Chapter V we introduce an influence maximization algorithm under our proposed model. We describe our implementation and experimental results in chapter VI and we conclude in chapter VII with a summary and potential future work.

Chapter Two

Preliminaries

We model a social network as a graph $G = (V, E)$ that could be directed or undirected depending on the nature of the social network. The sets V and E consist of the vertices/nodes of the graph and its edges, respectively. In this context, the nodes represent the individuals that are engaged in the network and the edges represent the connections/links between them. Our aim in this thesis is to find a set of seed nodes that can influence the majority of the network. These nodes must have high influence propagation ratio for our target to be achieved successfully. We will elaborate on how to obtain the influence of the node using our proposed influence function in the chapter five.

We consider the variables d and α , where d is the depth that we can reach in our network and α is the maximum amount of seed nodes that we can afford. To select the seed nodes, we will use a decreasing sequence function that calculates the influence on the initial node and its neighbors until we reach the depth d .

The need-degree of a vertex v , denoted by $Need(v)$ is the number of in-neighbors that are still needed for a vertex to be influenced. In the context of PIDS, it was assumed that half of the neighbors of a node must adopt a campaign in order for the node to adopt it, or to be influenced. Thus each vertex would initially have a need value that is equal to half its number of neighbors. In this thesis we do not use the same notion but rather adopt a different model that assumes a vertex has a certain effect on its neighbors that is modeled as a weight function. The below table provides a summary

Symbols	Description
G	Influence Graph
V	Set of Users of G
E	Set of Edges of G
d	Depth
N_d	The neighbors at distance d
α	Size of the Seed Set
S	Set of Seed Nodes
$I(S)$	Set of Influenced Nodes
e_{uv}	Influence of u on v
$degree(v)$	Number of neighbors of v

Table 1: Symbols used in this thesis

of the terminology used in the rest of this text.

2.1 Types of Graphs

We will be testing our proposed algorithm on different types of graphs, classified mainly based on degree structure.

- Dense Graph: where the number of edges is close to (or a fraction of) the number of edges in a complete graph (or clique).

Figure 1: A dense Facebook Graph

- Sparse Graph: where the number of edges is bounded above by a small constant multiple of the number of vertices.

Figure 2: Sparse graph representing Facebook messages

- Hub Graph: where a few nodes have very high degree while most vertices have a much lower degree.

Figure 3: Bitcoin Hub Graph

2.2 Influence Propagation

Influence propagation is the formal notion of the act of spreading information from a specific set of source nodes in a portion of a social network graph. Influence propagation does not aim to spread influence to the entire network like what the classical Dominating Set problem aims to achieve. Its main target is to be able to reach the largest portion possible of the network with the least cost. The main elements of the diffusion process that will occur according to the rules of a pre-specified diffusion model are the piece of information that is tending to propagate throughout the network, and the network that consists of nodes and edges and the pre-selected seed set that will initiate the diffusion process.

Every node in the network has its unique influence capability. The influence of a node represents the effectiveness of this node in the influence spread process. This notion of influence capability of a node plays the major role in the seed selection process in the sense that the higher the influence capability of a node, the higher the chance that this

node will be selected in the seed set.

Mathematical models have been proposed to describe the notion of influence propagation. Those models came to set some ground rules for the diffusion process to be able to illustrate mathematically what is actually happening in the real world. The rules focus on how one node would influence another based on certain aspects that are defined in the model itself. Most of the models that are presented in this area follow two basic rules [11]:

- Information diffusion starts from a set of seed nodes;
- information can be accessible through direct neighbors only.

Let us take the case where a song artist is promoting his/her new song. This artist will share a snippet of the upcoming song on twitter for example, and this snippet will circulate among the fans. The artist in this case is considered as a source node that activated most of his/her fans. An active node is said to be the node that adopted the piece of information or in this case the node that shared/listened to the new song and an inactive node is the one that did not yet adopt the piece of information or didn't listen to the new song.

We can categorize the models into two categories: progressive and non progressive models. A progressive model is the model that allows its nodes to switch from being inactive to being active but does not allow them to switch in the other direction. Non progressive models are the models that allows its nodes to switch in both directions.

We should not fail to mention that we have two types of influence when it comes to the nature of the node that is spreading influence. We have the direct and indirect influence. The direct influence is when the influence is transmitted directly from the seed set to neighboring nodes. Indirect influence is when an influenced node spreads influence to non-active neighbors. In our proposed method we will be considering both the indirect and the direct influence since we believe that the influence spread is highly affected by influenced neighbors since they are the main means of interaction between the seed nodes and the rest of the network.

The major models that have been adopted by most researchers were the Independent Cascade Model and the Linear Threshold Model.

2.2.1 The Independent Cascade Model

In the Independent cascade model (IC), proposed by Goldenberg et al. in [18], the diffusion process starts from the seed set S and each node u activates its neighbors with influence probability P_{uv} that is exerted on each edge to determine the transmission possibility. If a node v has more than two active neighbors, they will effect v independently. If the activation process took place successfully, and assuming the two neighbors of v have become active at time t , node v will be active at time stamp $t + 1$. Node v cannot be switched back to become inactive. The diffusion process will terminate when there are no more nodes that can be activated. The main drawback of this model was that each node u has a single chance to activate its out-going neighbor. Note that the IC model is said to be progressive and stochastic which implies that the end result might change if the initial seed set changed.

2.2.2 The Linear Threshold Model

In the Linear Threshold Model (LT) that was introduced by Granovetter in [20], every node u is associated with a threshold $\theta \in [0, 1]$ that will indicate the capacity of a node to be influenced by a campaign. Each edge uv has a weight w_{uv} that reflects on the intercommunication between two nodes. During the activation process, the total influence weight will be calculated upon all incoming edges and compare the result to the assigned threshold of the node. If the total influence is greater than this threshold, the node will become active and will be able to activate other inactive neighbors.

$$\sum_{u \in N_{in}(v)} w_{uv} \geq \theta(v)$$

The LT model encompasses the fact that a node needs to be triggered by multiple influencers in order to be activated successfully. This notion was not reflected in the

IC model, that is the reason why the LT model is more suitable for viral marketing than the IC model.

2.3 The Fixed-Distance Influence Propagation Problem

If edge-weights are uniform and the IC model is used, influence depends merely on distance. In such cases, influence maximization becomes simply the following partial version of the classical Firehouse problem (seemingly defined for the first time in [25]).

The *Fixed-Distance Influence Propagation* problem can be formally defined as a partial version of the classical *Firehouse* problem, which is also known as *Distance- k Dominating Set* and some other names. We provide the following parameterized version of the problem.

Fixed-Distance Influence Propagation (DFIP)

Given: Graph $G = (V, E)$; parameters k, d and t .

Question: Does V have a subset D of cardinality at most k such that $|N_d(D)| \geq t$?

The problem is obviously *NP*-Complete by reduction from Partial Dominating Set (simply by subdividing all edges $d - 1$ times). The set D will be referred to as the seed set in this paper. The same problem has been given many names in the realm of social networks when authors of some published papers use a fixed-distance of allowed propagation in their influence maximization problems/models. In fact, it is easy to see that Influence Maximization, under the Independent Cascade model is nothing but DFIP. Below we describe the common notion of influence maximization.

2.4 The Influence Maximization Problem

Given a social network graph $G = (V, E)$, a diffusion model, and an empty seed set S . The main objective of Influence Maximization problem is to select the minimum number of nodes that will initiate the diffusion process and be able to cover the majority

of the network. We can measure influence by the total number of influenced nodes at the termination stage of the diffusion process.

Kempe et al. [26] and Borgs et al. [10] both worked on different influence maximization approaches under the IC model. Kempe's approach was based on building the seed set using multiple rounds and in each round they will add one node to the seed set. This node will be chosen based on the larger marginal increase of the seed set. The authors used Monte Carlo simulations to estimate this marginal increase and select the seed set nodes accordingly. The greedy approach introduced by Kempe et al. runs for a long period of time even when the number of nodes in the social network graph is relatively small.

As an enhancement over the greedy method proposed by Kempe in 2003, Borgs proposed reverse influence sampling method to overcome some limitations. The main objective of this method is to find a reverse reachable set (RR) of a certain node v which contains all the direct paths from any node to node v . For a node v to be activated by a seed set S , the set has to overlap the reachability set of node v . If a node u appears in many RR sets then this node must have high influence on the network since it can be reached from several other nodes and can reach several nodes as well.

Chapter Three

Related Work

Influence Maximization was shown to be an NP – *hard* optimization problem in [26] for the Independent Cascade model (although this is already known by interpreting it as a variant of the Dominating Set problem as we pointed out before). As we shall see from previous work, some of which is described below, the topic has been of major interest mostly due to the fact that viral marketing campaigns have been heavily relying on social network platforms lately. We will discuss various approaches in this section.

Granovetter et al. were the first to propose the linear threshold model (LT) in 1978 [21]. The model assumes that every node v is associated with a threshold θ_v that belongs to the interval $[0,1]$ and $\sum_{u \in V} w_{uv} \leq 1$. The basic idea of this model is that if a node is activated in previous steps it will remain active and an inactive node becomes active if $\sum_{u \in N_{in}(v)} w_{uv} \geq \theta_v$ where $N_{in}(v)$ is the set of v 's active in-neighbors. In [15], the authors showed that computing the influence in a directed acyclic graph (DAG) requires not more than linear time. Using the linear threshold model, and the DAG theory, they proposed the first scalable influence maximization algorithm called local directed acyclic graph algorithm (LDAG). [24] extended the work of [15] and proposed the competitive local directed acyclic graph that uses the competitive linear threshold model and the notion of blocking the influence propagation by targeting the nodes that might initiate a propagation on their own.

Kemp et al. in [26] proved that the influence maximization problem is NP-hard both

for the independent cascade and the linear threshold models. They raised the question of how to convince a subset of individuals on a social network to adopt a campaign, idea or even a product to boost their existence by reaching as many people as they can. Using submodularity, Kemp et al. managed to prove that a natural greedy algorithm is 63% optimal for several classes of models, but this approach needs days to execute even for a few thousand nodes, which is prohibitive for a social network setup considering the huge number of nodes that is constantly increasing.

Kimura and Saito in [27] introduced the shortest path heuristic targeting an approximate solution to the influence maximization problem. The algorithm computes the shortest path between a specific inactive node and the initial activated set. Then they generalized the shortest path model (SPM) into SP1M. The main difference is that node v is allowed to get activated not only at $t = d(S, v)$ but also at $t = d(S, v) + 1$ where $d(S, v)$ is the topological distance from S to v .

Chen et. al in [13] extended the work of [27] from a shortest path notion to consider the maximum influence path (MIP) notion. As we know, every edge is considered to be associated with an influence propagation probability value. A maximum influence path is the path with the maximum propagation probability. The algorithm employs an influence threshold and computes a maximum influence path between every pair of nodes. MIP will ignore all those below the previously specified threshold.

In [23] the authors proposed a multi objective optimization approach under influence maximization. This approach was based on creating a function that determines the influence spread probability and they used genetic algorithms for the seed selection process. In [19] the authors used a discrete particle swarm optimization for the sake of reaching more exploration and increase the influence spread.

A degree discount heuristic was adopted by the authors of [14] to see how the degree of a specific node might affect its influence. The authors proposed two degree discount heuristics. The single discount heuristic which is based on minimizing the degree of the selected seed node by 1 and the degree discount IC which was made to fit the

IC model with propagation probability. When a node v is added to the seed set, the number of active nodes will increase by:

$$1 + (d_v - 2t_v - t_v p(d_v - t_v) + o(t_v)) * p$$

where:

d_v : is the degree of v

t_v : number of v 's neighbors that are already in the seed set

In conclusion, the larger the number of seed neighbors of v are, the more discount its degree will get.

In [11] the authors conducted a study that encapsulates all the work that has been done on information diffusion, influence evaluation, influence maximization and information source detection. The four categories are essential to the misinformation prevention problem. In information diffusion we have two rules to be considered:

1. The initialization of information diffusion starts from several nodes;
2. Information can be accessed only through neighbors.

The information diffusion models present in the literature can be categorized into progressive models and non progressive models. The difference is that in a progressive model it is prohibited to let an active node be switched to inactive while in the non progressive models this rule is not valid.

Based on the information diffusion models, influence evaluation can be handled. In this context, influence evaluation is the study of how much an individual can be influential to others. Kempe in [26] measured the influence of the initial node set A by the number of active neighbors of A at the end of the diffusion process which will help reaching influence maximization. The last step of this process is the information source detection which allows us to find the source of the rumor and diminish information circulation.

Chapter Four

Proposed Model

In this chapter, we describe our *Depth-Based Diminishing Influence model (DBDM)* and discuss the difference between our proposed model and previously introduced models. We will be mainly comparing with the MAT model, introduced recently (in 2018) by Wang et al. in [33].

4.1 The Multiple Path Asynchronous Threshold Model

The reachability-based influence diffusion model was introduced by Wang and Street in [31, 32]. It is mainly a depth-limit based model: the information visibility over the network is studied, regardless of the influence, using a pre-specified depth.

In [33], the authors extended the above approach to include the “activation” notion. They measure the total influence of a node using a *multiple-path asynchronous threshold model* (MAT). This model takes into consideration the word-of-mouth (WOM) communication, where they consider that any influencer that is initially in the seed set and any influenced node can be a carrier of information and can potentially influence other inactive neighbors.

This MAT model encompasses the following rules:

- No loops are allowed. They consider that the message should not circulate and

get back to the origin source.

- A node can be visited multiple times from different paths. Multiple influencers can trigger the same node.
- The three-degree-of-influence was introduced in [16]. The MAT model adopted this approach by stopping the diffusion at distance three from the influencer.
- Asynchronous diffusion from an influencer parent node to its neighbors.
- Influence decay over time.
- Intercommunication strength is illustrated by the weights exerted on edges.

The authors introduced the normalized influence weight as follows:

$$\hat{w}_{uv} = \frac{w_{uv}}{\sum_{k \in N_v^{in}} w_{kv}}$$

The directed edge from u to v is assigned a value, $\hat{w}_{uv}$, that is calculated using the above equation where N_v^{in} denotes the set of nodes of v 's in-neighbors.

Assuming a maximum distance d is pre-specified for the propagation of influence, the depth factor is denoted by:

$$\alpha = d^{-2}$$

The authors also introduced the following influence function that quantifies influence over each and every seed node. X_i denotes the influence spread of the influencer i , Y denotes the diffusion path from i , Y is the diffusion path from an influencer to a node j and $P\{Y = y\}$ is the probability of influence reaching j along a path y .

$$E[X_i] = \sum_y E[X_i|Y = y]P\{Y = y\}$$

4.2 The Depth-Based Diminishing Influence Model

This model focuses on the reachability and information visibility of a marketing campaign using a depth-limit approach. Our main target is to measure the total influence after the diffusion process, taking into consideration that the influence diminishes as we go deeper in our path to the desired node. To illustrate this approach, let us consider the directed graph below:

Figure 4: Sample Graph

Let us consider that V_3 was selected into the seed set S . As we know, the seed set is the place where the diffusion process starts and assume V_7 is where the termination of the diffusion occurs in this case. We consider that the influence of V_3 on V_0 is direct, a direct influence is the most effective influence. Practically speaking, if I adopted a campaign on Facebook, I will have a direct influence on my direct neighbors/friends and if one of my neighbors was influenced then I will have indirect influence on his neighbors and so on. The basic idea here is that if I am interested in something as a person, I will be following multiple pages that address that topic and I will be sharing articles, photos and comments that are related to this idea or product. Following this real world setup, a person (vertex) would have a substantial influence on his/her friends (neighbors) since they are always viewing posts that (s)he have shared and most probably being attracted by one of them. This notion of a real life scenario formed the base of our diminishing influence model.

We adopt the following general formula for measuring the influence of a node in a graph/network. We shall refer to it as the *diminishing influence function*:

$$Influence(v) = \sum a_i w(N_i)$$

where a_i is a decreasing sequence that is used to quantify the influence as we go deeper in the network. The weight function (w in the above formula) assigned to a vertex in our model is assumed to be its *local influence* which we decided to calculate it as the number of neighbors (degree) when no other information is given. The weight of a $N_i(v)$ is nothing but the sum of weights of the vertices that are at distance i from v .

The above formula has been proposed by Abu-Khzam [6]. A variant that was suitable for computing positive influence dominating sets was also used in [4].

Our work follows the rules of the independent cascade model that were explained earlier. We added the notion of the fixed depth and we used our influence function to select the seed set that will initiate the diffusion process.

The following ground rules are assumed by our model:

- A node can be visited from multiple seed set nodes and the effect of the influence will be independent for each and every seed node in the used set.
- We exert a depth d , that is the maximum depth to be reached in our diffusion process depending on a threshold θ determined depending on the maximum depth of each graph and the connectivity of the graph.
- The intercommunication e_{uv} strength between nodes will be computed uniformly at random.
- We measure the influence by outputting the A set (activated set) which is the total number of influenced nodes at the end of the diffusion process
- We captured the indirect influence but we do not consider that an in-active neighbor can pass influence and activate another in-active neighbor.

It is important to mention that in a real world social network graph, we are not able to quantify the exact intercommunication degree between two nodes. For simplicity, we estimated this probability randomly.

Chapter Five

Influence Maximization Algorithm

5.1 Breadth First Search(BFS)

Breadth First Search is well known graph traversal algorithm that allows us to measure the distance from a node to other nodes in a graph when edges have a uniform weight. It is used by our algorithm to estimate influence propagation, among other things. We used an adjacency list representation (to list the direct neighbors of each node). The influence of a node will be calculated using the influence function described earlier.

If we take figure 4 as an example. Starting at v_3 , the algorithm will recognize v_0 and v_5 as a direct neighbors. Afterwards, the algorithm will continue by considering the neighbors of the first node which is in this case v_0 . The direct neighbor of v_0 is v_1 which will be considered as a second level neighbor. This process will continue until we reach v_7 . After that we will go back to v_5 to list its neighborhood, and so on.

The calculation of the influence function will consider the various levels of neighborhood of each vertex so we can accurately quantify the influence that will be used in the influence maximization algorithm presented in the next section.

5.2 Influence Maximization Algorithm

Let us consider the election setup, as an example. Several political parties are engaged in a specific community and want the majority of the community to support them. The main and most efficient way to reach those people is by breaking through their social network accounts. For simplicity, we can assume that campaign G is targeting a specific community and trying to spread pledges to win the election.

First of all, we need to convert this problem to an optimization problem. If we want to consider a social network graph, it can be represented as a directed graph $G(V, E)$. Our aim is to find a set of seed nodes that can influence the majority of the network. These nodes must have high influence propagation ratio for our target to be achieved successfully.

A node "i" can influence a node "j", if "i" is a direct neighbor of "j". In a real time setup, while scrolling through our timelines and viewing the same post more than once such as all of our friends support the topic, then "i" will more likely be convinced by it. So we need to select a seed set that will initiate the marketing campaign and make sure that it is highly connected to the network with high influence ability. We will use a decreasing sequence approach to calculate the influence of the nodes and carefully select our seed set to handle the diffusion process later on.

To calculate the influence degree of each node, we use a Breadth First Search (BFS) approach to traverse the whole network and attain the connections of each and every node. During the BFS process we calculate the influence of each node using our proposed influence function mentioned in our proposed model. At the end of the BFS algorithm, we will retrieve the influence array representing the influence degree of each node. After that we will select the top k nodes with the highest influence, where k is a pre-specified threshold (such as the number of people we can afford paying to start a marketing campaign). Along with selecting the seed set we will be calculating the the edge probability of the diffusion, which is generated randomly and denoted by e_{uv} ($e_{uv} \in [0, 1]$.)

After selecting the seed set S , we use BFS again for the diffusion process. This BFS will take as an input the Seed set S along with all the edge-weights that were computed earlier and the maximum depth d that we can afford. The output of the diffusion algorithm will be the set of nodes that were successfully influenced by our selected seed set.

A pseudocode of our influence maximization algorithm is given below.

Algorithm 1 InfluenceMaximization (G, d, K)

Result: Set A of activated nodes

Initialize $\theta \leftarrow 0$

Initialize $I[\] \leftarrow 0$

Initialize EdgeWeight[] $\leftarrow \phi$

foreach $v \in G$ **do**

$I[v] \leftarrow \text{Calculate_Influence}(G, v)$

end

foreach edge $e \in G$ **do**

$w(e) \leftarrow \text{Calculate_Edgeweight}(G, e)$

if $w(e) < \theta$ **then**

 delete edge e

end

end

$S \leftarrow$ set of nodes having top K influence values in $I[\]$

foreach $v \in S$ **do**

$A \leftarrow A \cup \text{BFS_Activate}(G, v, d)$

end

Return A

Algorithm 2 BFS_Activate(G, v, d)

Result: Set A of activated nodes

Initialize $A \leftarrow \emptyset$ $Q \leftarrow \text{Queue}(v)$ // v is initially inserted into queue Q **while** Q is not empty **do** $u \leftarrow \text{dequeue}(Q)$ **if** $d[u] \leq d$ **then** $A \leftarrow A \cup \{u\}$ **foreach** x in $\text{Neighbors}(u) \setminus S$ **do** **if** $x \in Q$ **then** enqueue(x) $d[x] = d[x] + 1$ **end** **end** **end****end**Return A

Chapter Six

Experimental Analysis

We implemented our Influence Maximization algorithm and ran multiple tests using different types of graphs to assess the efficiency and (most importantly) effectiveness of our algorithm. The algorithm was implemented using *JAVA* and the tests were performed on an Intel(R) Core(TM) i7-7600U CPU @ 2.80 GHz machine. We compared our results with the results presented by [33] on the MAT model.

We obtained our social network sample graphs from the network data repository [30] and Stanford Large Network Data-set Collection [29]

6.1 Model Comparison

First, we describe briefly the characteristics of the two data-sets that we used to compare the performance of our model with that of the MAT model. The data-sets that were used in the comparison are the PGP data-set and the wiki-vote data-set. The reason we chose those two graphs is that they are different in nature and in density. In the next section we will be testing our approach on different type of graphs.

The below table shows the major differences between the two network graphs.

Data-sets	PGP	Wiki-vote
Directed	No	Yes
Weighted	No	No
Nodes	10,680	7115
Edges	24316	103,689
Connected components	1	24
Average component size	10,680	296.5
Largest component size	10,680	7,066

Table 1: Social Network Data-sets

Figure 5: Influence spread achieved under MAT and DBDM models on PGP Data-set

The PGP subgraph has relatively high number of nodes and large component size. This makes it a highly connected graph. Since the graph is undirected, we considered that every edge is bidirectional in our code. In other words, if node u is connected to node v in the data-set, we will assume that u is connected to v and v is connected to u in our implementation process.

It is clear that we were able to duplicate the influence spread using our purposed algorithm. For simplicity, we considered that we cannot exceed the depth of 15% of the maximum depth that is found in the graph naturally. In a real world set up, the influence propagates through the direct neighbors in a higher intensity than that of the

second-level neighbor (i.e., node at distance 2) and so on. That is why our diminishing influence propagation function, considering this fact to select the seed set that will initiate the process, succeeded to achieve a high influence rate with a high degree of convergence. The influence spread that is achieved by 10 seed set nodes is not much lower than the influence spread that is achieved by 100 seed set nodes. This high convergence ensures the effectiveness of the selection process of the seed set. We were able to maximize the propagation of influence while minimizing the seed set and the depth threshold, which is our main target using the proposed model.

Figure 6: Influence spread achieved under MAT and DBDM models on Wiki-vote Data-set

Wiki-vote is a directed graph with a smaller size of connected components, allowing the diffusion to achieve its convergence with the smallest seed set possible. As we can see in the above graph, the total number of influence of our proposed model did not exceed the maximum influence that was achieved in the MAT model. However, we were able to cover about 2000 nodes by selecting a seed set of size 10, while using the MAT model this number was achieved by selecting 50 nodes into the seed set. This is a substantial improvement over the most recent model for influence propagation.

This result proved that our model is effective along all the set objectives, in the sense

of (i) minimizing the seed set, (ii) the depth degree, and (iii) maximizing the influence spread. Our multi-objective model achieved its target regardless of the spread of influence. Despite the fact that MAT0.1 reached a higher influence as the number of seed nodes increased, the DBDM was able to reach a close influence way before MAT0.1. Recall that our main objective was to decrease the total cost in terms of number of seeds, influence spread, and depth.

6.2 Comparison on Different Type of Graphs

In this section, we presents a number of experiments on various types of graphs, and we analyze the effectiveness of our algorithm accordingly. Those graphs are chosen to be of different densities so we can test the performance of our algorithm based on the connectivity of a graph, which is the most important aspect in determining the propagation efficiency/effectiveness. The below graph illustrates the obtained results.

Figure 7: Influence spread achieved on various data-sets

Data-sets	Nodes	Edges	Density	Maximum Degree
fb-messages	1.3k	6.5k	0.00805625	112
Brightkite	56.7k	212.9	0.000132294	1.1k
Socfb-Brandeis99	3.9k	137.6	0.0181122	2k
ia-radoslaw-email	167	82.9k	5.97980	9.1k
socfb-swarthmore42	1.7k	61.1k	0.04439	577

Table 2: Various Types of Social Network Data-sets

Figure 8: Visualization of the Data-sets

As illustrated in the above graph, we were able to reach high influence spread compared to the total number of nodes that are found in each of the studied graphs. The convergence is still relatively high if we compare statistical values of each graph over seed set of size 10 and seed set of size 100.

Following the obtained results, we notice that we have similar convergence rate that is common between the four presented graphs. We used the same depth rate, which is 15% of the maximum depth available in the graph, yet we were able to reach a near-maximum coverage over the "ia-radoslaw-email" which is the most dense graph in our collection. Considering the "Socfb-Brandeis99", which is relatively dense also, we were able to cover 75% of the entire graph.

To illustrate the effect of the density of the graph on our implemented methodology, we consider the test results that were performed on the "Brightkite" graph. This graph had the lowest density among all the used graphs and has a lot of hubs, which makes it much harder for us to be able to cover a large area of the network considering a relatively small seed set. Below are the results performed on this graph along with corresponding visualization.

Figure 9: "Brightkite" Data-set

Figure 10: Visualization of "Brightkite" Data-set

The influence spread that was recorded on the semi-hub structure graph was relatively small if we consider the total number of nodes that belong to this graph. On the other hand, a size 10 set was able to reach around 2400 nodes.

A seed set of size 1000 was able to reach up to 7000 nodes. Considering the nature of the graph, our test measures changed a bit. Since we are dealing with a deep network but a not highly connected one, we considered that we can reach up to 30% of the maximum valid depth. Those measures were taken into consideration since we have low connectivity and high degree graph.

6.3 Comparison Between Different Series

The diminishing influence function was tested over multiple decreasing sequences to record the influence spread variation with respect to the rate of decrease of the used sequence (i.e., *sequence decreasing factor*). In the influence function mentioned earlier we replaced the a_i factor with the below series.

- Series A: $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \frac{1}{8}, \dots$
- Series B: $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \dots$
- Series C: $1, \frac{1}{2^2}, \frac{1}{3^2}, \frac{1}{4^2}, \frac{1}{5^2}, \dots$

The first term is always set to “1” to indicate the almost complete influence of a node on its direct neighbors. As we move away from a certain node, our model assumes its influence decreases according to a certain rate represented by the value of the fraction that is associated with the distance/level of the nodes that are not direct neighbors.

We demonstrated the effect of the above series on the influence spread achieved in the end of the diffusion process using the below graph.

Figure 11: Performance Comparison on Different Series: **a** Socfb-Brandeis99 dataset **b** fb-messages dataset **c** ia-radoslaw-email dataset **d** socfb-swarthmore42 dataset

We ran multiple tests over each series and took the average value of the resulting influence spread to be able to represent the most accurate values of the actual spread that might occur in a real time setup. The results showed that adopting series *C* recorded the highest influence spread on all considered graphs. As we can see the rate of decrease of this series is the fastest one among the tested instances.

This fact helped in selecting the most effective seed sets as far as the influence spread in concerned. The success of such a fast-decreasing sequence (i.e., its sharpness) reveals that assigning smaller factors to the influence on farther nodes can single out those nodes that are essential for the spread of influence. Because of the fast convergence of the used series (the series tail tends to zero as the distance grows), the value of our influence functions essentially affect by the first few neighborhood-levels (mainly the first three levels, as we observed).

We noticed that the sharpness of the sequence affected the data-sets with respectively low densities. If we considered graphs *a, b* and *d*, the influence spread increased significantly in comparison with graph *c* which represents the influence spread on the "ia-radoslaw-email" data-set which is the densest graph used in our collection..

In conclusion, the rate of decrease of the adopted series affects mostly sparse graphs while if we consider a dense graph the effect does not seem to vary with the rate of decrease, which is technically speaking due to the large value of the first term of each series (due to large degrees). In fact, the first term of each series is the same and it has a large value compared to the remaining ones when the graph is very dense. This makes the values of the influence functions close to each other.

6.4 Clustering

Social networks are known to be very large, so they tend to have large diameters. If our influence function assigns high values to a seed set that belongs to a certain local community, or neighborhood, we would have much less propagation effect than having a well-scattered seed set. For this purpose we decided to enhance our model by

applying graph clustering prior to seed selection.

Clustering consists of partitioning a graph into a number of subgraphs such that nodes of the same subgraph have a high connectivity (in the subgraph) compared to their relation to the rest of the network. In our work, the use of graph clustering helped us more than we expected. In addition to maximizing the influence spread, it helped us minimize the size of our selected seed set. An important related notion used in literature is *modularity*. Modularity is used to discover the structure of the network and partition it to non-overlapping communities without taking into consideration the size of each component [17, 22, 12].

We used a third party application provided by "github" named "Gephi". This application divided our network into communities depending on the connectivity of the provided networks. The program took the graph file as an input and provided us with multiple subgraphs (the clusters). We tested our approach on the subgraphs by providing each corresponding subnetwork as a separate component to our algorithm. We recorded the results to see if the clustering approach increased our influence spread.

Multiple experiments were performed on different social network graphs that were used earlier in our experiments. Most of the results obtained proved that the clustering substantially increased the influence spread and decreased the seed set. The results showed that we were able to decrease almost 60% of the seed set while obtaining the same influence spread. We used the "fb-messages" data-set as an example to illustrate (below) the results obtained after applying graph clustering.

The employed clustering software partitioned the network into 10 separate communities and we fed the resulting sub-networks to our algorithm. We took into consideration the size of each obtained subgraph when deciding on the best possible seed-set size per cluster, and we performed our method as described earlier.

The obtained results are shown in Figure 12, which shows clearly that scattering the seed set among different clusters can significantly improve the influence spread, as expected. Applying this notion on the "fb-messages" data-set, we saw that the maximum

Figure 12: The Effect of Clustering

result obtained by the DBDM approach using 100 seed set nodes was easily obtained by DBDM + clustering using 30 seed nodes only.

Chapter Seven

Conclusion and Future Work

Our proposed multi-objective approach proved to be efficient and effective compared to previously studied models. It is mainly based on minimizing the selected seed set and the depth achieved to maximize the influence spread. It would be interesting to study this problem from a parameterized complexity standpoint. In fact, the use of multiple parameters enhances the model and can yield faster algorithmic solutions, as observed in [5] in the context of multi-parameterized cluster editing.

We compared our results to the results obtained under the MAT model, which was recently presented by Wang et al. in [33]. We noticed that our algorithm works perfectly fine under dense and highly connected graphs, but the approach seems to be less effective, relatively, when it comes to hub structured graphs. Furthermore, we proposed a clustering-based approach to scatter as much as possible the seed set and thus enhance the reachability of a larger set of influenced nodes. Our experiments show that our clustering-based method does indeed enhance the results obtained by merely applying the DBDM method. This particular clustering-based method should be the method of choice on very large networks.

Social network optimization problems are of a high interest nowadays, and there are other important flavors of influence propagation that can be tackled by applying our method. For example, the online version of the problem could be very interesting since the whole network may not be known upfront. The dynamic nature of social

networks could also be modeled using the notion of *parameterized dynamic problems* that are recently studied in [7, 8] and [28].

Another interesting variant of this work is to consider its use for the so-called *misinformation prevention* where two competing campaigns are considered and the objective of each is to limit the spread of the other. We believe our proposed model can be applied effectively in this context. Finally, the influence maximization model considered in this thesis can be enhanced further, in future work, if a better approximation is used for quantifying the real intercommunication rate between individuals in a network.

Bibliography

- [1] Facebook newsroom. <https://newsroom.fb.com/company-info/>, 2018.
- [2] facebook research. <https://research.fb.com/three-and-a-half-degrees-of-separation/>, 2018.
- [3] The wordstream blog. <https://www.wordstream.com/blog/ws/2017/11/07/facebook-statistics>, 2018.
- [4] F. Abu-khzam and K. Lamaa. Efficient heuristic algorithms for positive-influence dominating set in social networks. pages 610–615, 04 2018.
- [5] F. N. Abu-Khzam. On the complexity of multi-parameterized cluster editing. *J. Discrete Algorithms*, 45:26–34, 2017.
- [6] F. N. Abu-Khzam, 2018. Personal Communication.
- [7] F. N. Abu-Khzam, J. Egan, M. R. Fellows, F. A. Rosamond, and P. Shaw. On the parameterized complexity of dynamic problems with connectivity constraints. In *Combinatorial Optimization and Applications - 8th International Conference, COCOA 2014, Wailea, Maui, HI, USA, December 19-21, 2014, Proceedings*, pages 625–636, 2014.
- [8] F. N. Abu-Khzam, J. Egan, M. R. Fellows, F. A. Rosamond, and P. Shaw. On the parameterized complexity of dynamic problems. *Theor. Comput. Sci.*, 607:426–434, 2015.
- [9] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31:211–236, 05 2017.

- [10] C. Borgs, M. Brautbar, J. T. Chayes, and B. Lucier. Influence maximization in social networks: Towards an optimal algorithmic solution. *CoRR*, abs/1212.0884, 2012.
- [11] B. Chang, T. Xu, Q. Liu, and E.-H. Chen. Study on information diffusion analysis in social networks and its applications. *International Journal of Automation and Computing*, 15(4):377–401, Aug 2018.
- [12] J. Chen, O. R. Zaïane, and R. Goebel. *Detecting Communities in Social Networks using Max-Min Modularity*, pages 978–989.
- [13] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD*, 2010.
- [14] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 199–208, New York, NY, USA, 2009. ACM.
- [15] W. Chen, Y. Yuan, and L. Zhang. Scalable influence maximization in social networks under the linear threshold model. In *2010 IEEE International Conference on Data Mining*, pages 88–97, Dec 2010.
- [16] N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379, 2007. PMID: 17652652.
- [17] M. E.J. Newman. Newman mej.. modularity and community structure in networks. *proc natl acad sci usa* 103: 8577-8582. *Proceedings of the National Academy of Sciences of the United States of America*, 103:8577–82, 07 2006.
- [18] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, Aug 2001.

- [19] M. Gong, J. Yan, B. Shen, M. Lijia, and Q. Cai. Influence maximization in social networks based on discrete particle swarm optimization. *Information Sciences*, 367, 07 2016.
- [20] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.
- [21] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.
- [22] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral. Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E*, 70:025101, Aug 2004.
- [23] J.-b. Guo, F.-z. Chen, and M.-q. Li. A multi-objective optimization approach for influence maximization in social networks. In G. Q. Huang, C.-F. Chien, and R. Dou, editors, *Proceeding of the 24th International Conference on Industrial Engineering and Engineering Management 2018*, pages 706–715, Singapore, 2019. Springer Singapore.
- [24] X. He, G. Song, W. Chen, and Q. Jiang. Influence blocking maximization in social networks under the competitive linear threshold model technical report. *CoRR*, abs/1110.4723, 2011.
- [25] J. Hopcroft and J. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley Series in Computer Science and Information Processing. Addison-Wesley, 1979.
- [26] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 137–146, New York, NY, USA, 2003. ACM.
- [27] M. Kimura and K. Saito. Tractable models for information diffusion in social networks. In J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, editors, *Knowledge*

Discovery in Databases: PKDD 2006, pages 259–271, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

- [28] R. Krithika, A. Sahu, and P. Tale. Dynamic parameterized problems. *Algorithmica*, 80(9):2637–2655, 2018.
- [29] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. <http://snap.stanford.edu/data>, June 2014.
- [30] R. A. Rossi and N. K. Ahmed. The network data repository with interactive graph analytics and visualization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- [31] W. Wang and W. N. Street. A novel algorithm for community detection and influence ranking in social networks. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 555–560, Aug 2014.
- [32] W. Wang and W. N. Street. Modeling influence diffusion to uncover influence centrality and community structure in social networks. *Social Network Analysis and Mining*, 5(1):15, May 2015.
- [33] W. Wang and W. N. Street. Modeling and maximizing influence diffusion in social networks for viral marketing. *Applied Network Science*, 3(1):6, Apr 2018.