Influence Maximization on Twitter: A Mechanism for Effective Marketing Campaign

Yan Mei*, Weiliang Zhao†, and Jian Yang‡
Department of Computing, Macquarie University
New South Wales 2109, Australia
*Email: *yan.mei@students.mq.edu.au, †weiliang.zhao@mq.edu.au, ‡jian.yang@mq.edu.au*

***Abstract*—When Influence Maximization (IM) is applied to social network to maximise the network coverage, it becomes an effective mechanism for marketing applications. In this paper, we focus on a specific influence maximization problem, i.e., selecting a set of seeds on twitter to maximise information propagation, which can be used for information reaching out in marketing campaigns. The proposed approach is taking into the consideration of social ties, user interactions, and information propagation on Twitter. The influence probability is calculated according to users' action history including tweet, favourite, mention/reply, and retweet. An information diffusion model is proposed with the capability to simulate the dynamic process of information spread on Twitter. A concise heuristic algorithm is developed for influence maximization accordingly. Experimental results and analysis are provided based on a real Twitter network including 3,292 users in Darwin city in Australia.**

***Keywords*-influence maximization; information diffusion; social networks; marketing campaigns**

I. Introduction

In recent years, with the rapid growth of Online Social Networks (OSNs) such as Twitter, Facebook, Google+, and LinkedIn, there has been a revolutionary change in the way people communicate with each other. Motivated by applications such as viral marketing [1], information diffusion in online social networks has received tremendous attention. Online social networks have become new channels for companies to carry out their marketing campaigns and brought in business opportunities for enterprises.

Suppose a company has released a new product or an online service. This company would like to promote or advertise its new product on Twitter. Since the company's Twitter account only has a small amount of followers, the company needs to select a set of users to help the propagation of the marketing information on Twitter. The company expects that these seed users will influence their followers, and then these followers will influence their own followers as well. As a result, a large number of users could receive this marketing information through the online word-of-mouth effect. Due to the constraints of the budget or relevant resources, it is necessary to find a set of influentials as seed users to maximize the expected coverage.

The problem of selecting seed users, referred to as *Influence Maximization*, was firstly proposed by Kempe et al. [2]. It has attracted a lot of interest in the research field of online social networks. Most of existing works on this topic focus on algorithms for the selection of seed nodes. In particular, when the influence maximization problem is studied in a specific social network, the following key questions have often been ignored.

Firstly, what kind of influence is it in regards to? Influence is a concept which could show in various ways in different contexts. For example, it might refer to passing a message successfully to others in a task of information diffusion. It might mean that audiences agree with the speaker's arguments in a campaign speech. It might imply that customers are persuaded to buy products in a marketing activity. A clear definition/description for influence in a specific context is crucial when studying the influence maximization problem.

Secondly, how are influence probabilities obtained or computed? The data of influence probabilities are essential in this problem. Most of the studies in this area assume these probabilities are given as input. Only some recent studies [3], [4] have shown how to learn influence probabilities from the historical data of user actions. It is necessary to identify which types of actions in a specific social network should be used in the calculation of influence probabilities.

Thirdly, how is a diffusion model defined? The diffusion model determines how influence propagates in the networks. A well-defined diffusion model should capture the major characteristics of information spread in a specific social network. It plays an important role in dealing with the influence maximization problem.

In this work, we specify an influence maximization problem which can cover a wide range of marketing campaign scenarios on Twitter. The main contributions are summarized as follows:

- An influence maximization approach has taken into consideration of social ties, user interactions, and information propagation on Twitter. The proposed approach provides a solid generic solution for promoting products and services in online social networks like Twitter.
- An influence probability model is proposed. The influence probability during a specific time period is

978-1-4673-8999-0/17/$31.00 ©2017 IEEE

calculated according to users' action history including tweet, favourite, mention/reply, and retweet.

- An information diffusion model is proposed to capture the major characteristics of information spread on Twitter. This model inherits the classic *independent cascade model* and has the capability to adopt the assumption that a user can have multiple chances to be influenced by others in a considered time period.
- A heuristic algorithm is designed for influence maximization on Twitter. Experimental results show that this algorithm achieves better influence spread than classic heuristics, and has the influence spread quite close to that of the well-known improved greedy algorithm but uses less than one-thirtieth of its running time.

The remainder of the paper proceeds as follows. Section II reviews existing models and algorithms for influence maximization. In Section III, we specify the influence maximization problem on Twitter; propose the influence probability model and information diffusion model; and develop a heuristic algorithm for the selection of seed users. Section IV provides the details of a set of experiments and discusses the results. In Section V, we conclude the paper with a discussion on the future work.

II. Related Work

Information diffusion is the process that information spreads out among users in a social network as time goes on. A variety of diffusion models have been proposed by researchers. *Linear Threshold Model* (LT) and *Independent Cascade Model* (IC) [2] are two well-known fundamental models in studying information diffusion in social networks.

In the Independent Cascade model, the diffusion probability on each edge of the network must be specified. Most of existing works assume that edge weights (i.e. the probabilities of influence between users) are given as input. Kempe et al. [2] assign each edge of the co-authorship graph a uniform probability (1% or 10%) in their experiments. Chen et al. [5] use three pre-determined propagation probabilities of $p = 0.01$, 0.02 and 0.05 in their work. Another popular model is referred to as "weighted cascade" [2], [5] in which the probability of user u influencing v is assigned with the value $1/d_v$, where d_v denotes how many people user v follows. Although some recent studies develop their methods to learn the influence probabilities from historical data [3], [4], it is still lacking of enough research with details to consider the specific features of a social network when modelling influence probabilities.

Domingos and Richardson [1], [6] are early researchers to study the influence propagation in social networks. They investigate how to identify the influential users for potential marketing applications. In order to make viral marketing plans, they propose a probabilistic model to calculate the customer's influence. Subsequently, Kempe et al. [2] formalise the problem as a discrete optimization problem. They prove that the influence maximization problem is NP-hard for both Linear Threshold Model and Independent Cascade Model, and they present a simple greedy algorithm which approximates the optimum to within a factor of $(1 - 1/e)$ (where e is the base of the natural logarithm). This greedy algorithm needs a long running time and it is not suitable for large-scale social networks.

Several recent studies have tried to improve the original greedy algorithm to tackle this efficiency issue. Leskovec et al. [7] exploit the submodularity property to develop an efficient approximation algorithm, called CELF (Cost-Effective Lazy Forward selection). In their experiments, CELF algorithm is up to 700 times faster than standard greedy algorithm. To further reduce the running time, Goyal et al. [8] propose an extension of CELF, called CELF++, which is 35-55% faster than CELF. However, these algorithms are still not efficient enough for large-scale networks or complicated diffusion models.

This work studies an influence maximization problem on Twitter. The proposed approach aims to reflect the real situation of information propagation in the context of Twitter. The influence probabilities are calculated according to the action history of users. A diffusion model is proposed to simulate the influence propagation process. We also compare our algorithm with the improved greedy algorithm and several existing heuristic algorithms.

III. Methodology

A. Influence Maximization Problem

The influence maximization problem can be described as follows. A social network is represented by a directed graph $G = (V, E)$, where the nodes V represent users, and the directed edges E represent social ties between users. We are also given a budget k, which is a integer. The goal of influence maximization is to find k users (seed nodes) in the social network so that the spread of influence (defined as the excepted number of influenced users) could be maximized.

Based on the general definition of influence maximization problem above and the specific characteristics of Twitter network, we have the assumptions as follows.

- *Influence*: As a general term, *influence* means "change in a person's cognition, attitude, or behaviour, which has its origin in another person or group" [9]. When the term *influence* is used in the research community of OSNs, many researchers have provided their own explanations about *influence* [10], [11] in their interested contexts. In this paper, this term is referred to as "the ability to let someone know something, or pass information to others". We consider u influencing v if v gets the information from u.
- *Influence Probability*: A directed edge $(u, v) \in E$ between users u and v represents the probability of u influencing v, which is denoted as $p_{u,v} \in (0, 1)$. This

probability will be calculated according to the *action history* on Twitter, including individual user's actions and interactions between users. More details will be provided in Section III-B.

- *Information Diffusion*: We assume that the information diffusion can be simulated as a process with multiple discrete steps. A user can have multiple chances to be influenced by activated neighbours during the considered time period. At step t, the nodes which were active at step $t-1$ remain active, and other inactive nodes might be activated based on our probability model. More details will be provided in Section III-C.
- *Information Maximization*: We specify influence maximization as the problem of selecting a set of users in order to maximize the influence spread within a specific time period. In this work, we simulate the information diffusion with N discrete steps during this period.

In this work, we utilize influence maximization techniques to support the development of marketing campaigns on Twitter. The maximal information propagation is the goal of the proposed approach.

B. Influence Probabilities

Given a graph of a social network $G = (V, E)$, each directed edge $(u, v) \in E$ is labelled with a weight $p_{u,v}$, representing the influence probability with which u will succeed in activating his neighbour v. We assume that the *action history* is given. The *action history* includes information of individual user's actions and interactions between users. Let A_u denote the total number of actions user u performs and R denote the set of interaction types. $I(u, v, a)$ is a function to calculate the number of interaction $a \in R$ with which user v reacts to user u.

On Twitter, users deliver messages by posting tweets. After other users read a tweet, they can respond to the tweet by means of favouriting, replying or retweeting. We assume user u is likely to influence user v only in a fixed-size time-frame T since u posts a message, and the influence probability does not change over time. In the influence maximization problem, we consider A_u as the total number of tweets the user u posts in a certain time period T. The *action history* contains three kinds of interactions (denoted by R), which are *favourite*, *mention/reply*, and *retweet*.

If user v reacts to user u, it means u has successfully passed the information to v, say u has influenced v. The *influence factor* ($infl(u, v) \geq 0$) from user u to v is defined as the ratio of $v \to u$ reactions to the total actions performed by u.

$$infl(u, v) = \sum_{a \in R} I(u, v, a) \;/\; A_u \tag{1}$$

The influence probability $P_{u,v}$ $(0 \leq P_{u,v} < 1)$ is calculated based on the *influence factor* as:

$$P_{u,v} = 1 - \exp(-infl(u, v)) \tag{2}$$

We assume that there is always a small influence probability between connected users even if there are no historical interactions between them. If $P_{u,v} < 0.01$, $P_{u,v}$ will be set to 0.01. This constant value 0.01 has also been used in [2], [5]. In our work, the influence time-frame T is divided equally into N slots. The diffusion process moves one step forward in each time slot. $P_{u,v}$ is calculated as follows.

$$P_{u,v} = 1 - (1 - p_{u,v})^N \tag{3}$$

where $p_{u,v}$ is the probability of u influencing v at each step of propagation. Based on Eq. 3, $p_{u,v}$ is:

$$p_{u,v} = 1 - (1 - P_{u,v})^{1/N} \tag{4}$$

C. Information Diffusion Model

In cascade models, when a node u first becomes active, it has a single chance to influence its inactive neighbour v, with a probability $p_{u,v}$. If u succeeds in activating v at step t, then v can make an attempt to influence its inactive neighbours at step $t+1$. The diffusion process stops until every active node has tried its single chance and there are no more activations.

Based on the problem definition in Section III-A, we propose a *R-J cascade model* for the information diffusion process. There are two modifications in *R-J cascade model* compared with the basic independent cascade model. Firstly, a user u always has the chance to activate his inactive neighbour v at each step, which means that the attempt from u to v can be *repeatable*. In the context of Twitter, a user can obtain the information (or to say, read the tweet) from others whom he/she follows any day after the information / tweet is posted. Secondly, if an inactive user v has a set of activated neighbours denoted by S, we predict whether v will be activated based on a *joint influence probability* denoted by $p_v(S)$. At each step, the user v will become active if any of his/her active neighbours succeeds in activating v. Thus, the joint influence probability is calculated as below:

$$p_v(S) = p_{w,v} + (1 - p_{w,v}) * p_v(S \setminus \{w\})$$

where $w \in S$. The user v will be activated unless all his active neighbours fail to activate v. This formula can be expressed as follows.

$$p_v(S) = 1 - \prod_{u \in S} (1 - p_{u,v})$$

The algorithm for estimating the expected influence spread with *R-J cascade model* is provided as Algorithm 1.

Algorithm 1 RJCascade

Input: SS, r, N, $p_{u,v}$
Output: s
initialize activated set $AS = SS$
create a table jp to store the joint probability for each user
for each user $u \in V$ **do**
    $p_u(S) \leftarrow 0$
end for
for each user $u \in AS$ **do**
    find the follower set FS of user u
    for each user $v \in FS$ **do**
        $p_v(S) \leftarrow p_v(S) + (1 - p_v(S)) * p_{u,v}$
    end for
end for
$s \leftarrow 0$
for $i \leftarrow 1, r$ **do**
    reset the table jp to initial values
    for $j \leftarrow 1, N$ **do**
        initialize an inactive user set $IS = \varnothing$, who might be activated at this step
        find all the edges $(u, v) \in E$ where $u \in AS$ and $v \notin AS$, and add v into IS
        for each user $u \in IS$ **do**
            generate a random value $r \in (0, 1)$
            if $r < p_u(S)$ **then**
                $AS \leftarrow AS \cup \{u\}$
                find the follower set FS of user u
                for each user $v \in FS$ **do**
                    $p_v(S) \leftarrow p_v(S) + (1 - p_v(S)) * p_{u,v}$
                end for
            end if
        end for
    end for
    $s \leftarrow s$ + number of users in AS
end for
$s \leftarrow s/r$

D. Approximation Algorithms

Kempe et al. [2] develop a greedy algorithm to solve their identified influence maximization problem and have obtained the best result for expected influence spread comparing with existing approximation algorithms. While dealing with a large real-world social network, this greedy algorithm is inefficient and it is infeasible to get the results in an acceptable period of time on a normal computer.

Kempe et al. [2] prove that the influence function $f(\cdot)$ has the properties of monotonicity and submodularity. The submodularity property means that the marginal gain from adding a user u to a set S is equal to or greater than the marginal gain from adding the same user to a superset of S. In this work, we use the similar idea in the CELF algorithm [7] and present an improved greedy algorithm based on our problem definition in Section III-A. The details of the algorithm is described in Algorithm 2.

Algorithm 2 ImprovedGreedy

Input: $G = (V, E)$, k, $RJCascade$
Output: SS, s
initialize $SS = \varnothing$
for each user $v \in V$ **do**
    $spread \leftarrow RJCascade(v)$
    Add the tuple <*user id, spread*> into $slist$
end for
Sort $slist$ in descending order based on *spread* values
Add the first user in $slist$ into SS
$s \leftarrow$ *spread* of the first user in $slist$
remove the first user in $slist$
for $i \leftarrow 2, k$ **do**
    $s' \leftarrow RJCascade(SS \cup$ the first user in $slist)$
    if $s' - s >=$ *spread* of the second user in $slist$ **then**
        $SS \leftarrow SS \cup$ the first user in $slist$
        remove the first user in $slist$
        $s \leftarrow s'$
    else
        $\Delta \leftarrow s' - s$
        spread of the first user in $slist \leftarrow s' - s$
        $x \leftarrow 2$
        while $\Delta <$ *spread* of the x-th user in $slist$ **do**
            $s' \leftarrow RJCascade(SS \cup$ the x-th user in $slist)$
            spread of the x-th user in $slist \leftarrow s' - s$
            if $s' - s > \Delta$ **then**
                $\Delta \leftarrow s' - s$
            end if
            $x \leftarrow x + 1$
        end while
        Sort $slist$ in descending order based on *spread* values
        $SS \leftarrow SS \cup$ the first user in $slist$
        remove the first user in $slist$
        $s \leftarrow s + \Delta$
    end if
end for

Heuristic algorithms have been developed to tackle the efficiency issue in solving the influence maximization problems. The *high-degree* heuristic algorithm selects the seed nodes based on their degrees (in descending order), i.e. the number of followers on Twitter. Experimental results in [2] show that this algorithm can achieve the performance close to that of the greedy algorithm, outperforming several existing algorithms.

The *distance centrality* is another commonly used influence measure in sociology. It has been evaluated in [2], [5]. The distances from one node to other nodes are measured.

The node with shorter average distance to other nodes is regarded at more central position in the social network. Nodes at more central positions are more influential and they will be selected as seed nodes.

We propose an *influence index* heuristic algorithm. The algorithm aims to obtain the expected spread results close to the greedy algorithms with much less computation time. Different from the measures mentioned above, we use *influence index* to represent the overall influence power of a user. It is calculated as the sum of the influence probabilities from one user to others. For example, user u has only one outgoing linkage with an influence probability 100%, user v has ten outgoing linkages with an influence probability 10% on each edge. In this case, user u and v are considered to have equivalent influence power. In solving the influence maximization problem, we give priority to the users with larger *influence index* value when selecting seed nodes.

IV. Experiment and Analysis

A. Experiment Setup

We build up an experimental dataset collected from Twitter. This dataset includes the social network data associated with users from a real city *Darwin* in Australia. Firstly, we capture all the users whose location profiles include the word "*Darwin*" (5,276 users as of July 21, 2015). Then we check through the details of user profiles and filter out users who are not from the *Darwin* city in Australia. For example, in some cases, the word "*Darwin*" indicates a city or town in another country rather than Australia, or it actually means a person's name. Finally, the identified social network includes 3,292 users from the city of *Darwin* in Australia.

In the social network graph $G = (V, E)$, if user u has a follower v, there is a directed edge from u to v. There are 23,605 following relationships (i.e. directed edges) and 1,158 isolated users in the *Darwin* community. Furthermore, there are 39 connected components and maximum vertices in a connected component are 2,048.

The dataset includes tweets posted by the users from Darwin city in 30 days. According to our influence probability model, the accumulative influence probability of these 30 days is calculated based on the ratio of reactions to total actions. The probability of each day is calculated by Eq. 4 with $N = 30$. The influence propagation is simulated by setting one day as one step. The influence maximization goal is to find k seed users ($k = 1, 2, ..., 20$) in order to maximize the influence coverage after 30 days. To compare the performance of different algorithms, we run the following algorithms against a dataset from Twitter.

- **ImprovedGreedy:** The improved greedy algorithm proposed in Section III-D.
- **InfluenceIndex:** A heuristic algorithm based on users' sum of influence probability to others, defined as *influence index* in Section III-D.

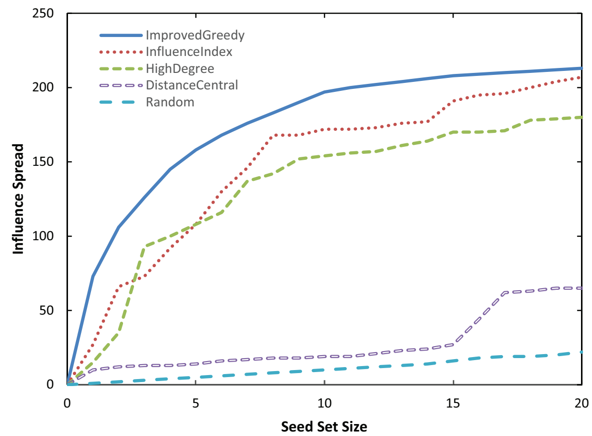

Figure 1. Influence spread achieved from the seed sets selected by different algorithms, with our proposed diffusion model

- **HighDegree:** A simple heuristic algorithm based on users' follower count, which is known as "degree centrality" in sociology literature.
- **DistanceCentral:** A heuristic algorithm based on users' average distance to other users in the whole network.
- **Random:** Seed users are randomly selected.

B. Results and Discussion

In the experiments for the five influence maximization algorithms, we simulate the diffusion process 100 times (i.e. set $r = 100$) based on the cascade model proposed in Section III-C, and calculate the average of the results as the expected influence spread. Fig. 1 shows the influence spreads of these algorithms, with different seed set sizes ranging from 1 to 20. The performances of **HighDegree**, **DistanceCentral** and **Random** are similar to the experimental results in [2]. The simple **Random** algorithm is a baseline and performs quite poor. In the experiments of [2], the influence spread of **DistanceCentral** algorithm is close to that of **HighDegree** algorithm. **DistanceCentral** algorithm performs worse against our dataset. It is only slightly better than the **Random** algorithm. This means that the performance of **DistanceCentral** algorithm can change when different datasets and diffusion models are applied.

Our proposed **InfluenceIndex** algorithm and **HighDegree** algorithm can achieve significantly better influence spread than **DistanceCentral** algorithm. The overall performance of **InfluenceIndex** algorithm is better than that of **HighDegree** algorithm. The influence index is a more effective indicator for evaluating a user's influence than the number of followers. Compared with **ImprovedGreedy** algorithm, when seed set size is 20, the influence spreads of **InfluenceIndex** and **HighDegree** are 2.8% and 15.5% smaller respectively.

The curve for **ImprovedGreedy** algorithm becomes almost horizontal when the seed set size is bigger than 15. From the 16th seed node, the influence spread only increases

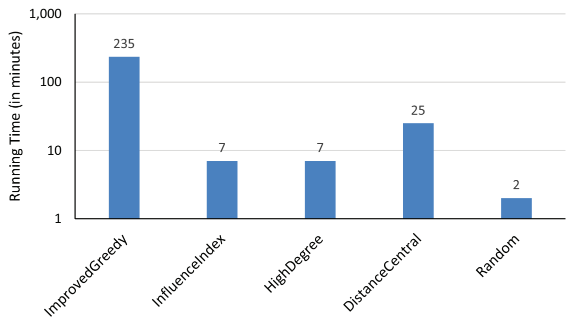

Figure 2. Running time (in minutes) for different algorithms, with our proposed diffusion model

by 1 when adding a new seed node. If we assume the cost of targeting a new seed user is equal to the profit of obtaining a new influenced user, it makes no sense to expand the seed set size after it reaches 15. The number 15 is approximately 0.5% of the total number of users (3,292) in our dataset. This percentage (0.5%) can be used as a benchmark for determining the seed set size in a social network.

Fig. 2 reports the running times of different algorithms with our proposed diffusion model when the seed set size is 20. Although **ImprovedGreedy** algorithm achieves the best influence spread, its running time is very long (nearly 4 hours). It is impractical to use **ImprovedGreedy** algorithm when dealing with large-scale social networks. Comparing with **ImprovedGreedy** algorithm, our proposed **InfluenceIndex** algorithm can reduce the running time significantly (more than 30 times faster) and achieve a quite close influence spread. **DistanceCentral** algorithm takes a longer time because it is time-consuming to calculate users' average distance to others in the whole network.

V. Conclusion and Future Work

This paper proposes an influence maximization approach with detailed description of influence probabilities, diffusion model, and heuristic algorithm in the context of Twitter. The proposed approach can cover a wide range of marketing campaign scenarios on Twitter. Influence probabilities are calculated based on users' action history. An information diffusion model is proposed to simulate the information spread on Twitter. The model covers the specific situation that a user can have multiple chances to be influenced by others in a considered time period. A concise algorithm is developed based on heuristic principles.

A set of experiments are carried out with real Twitter data in Darwin (a city of Australia). Experimental results show the effectiveness of the proposed influence maximization approach. The developed heuristic algorithm has the capability to achieve much better influence spread than existing heuristic-based solutions. Comparing with the well-known improved greedy algorithm, our algorithm can obtain close influence coverage but save about 97% running time.

In the future, we will develop methods to specify communities in online social networks and study influence maximization by utilising the specific features of these communities. We will recruit more large-sized datasets to analyse the characteristics of real-world network structures and investigate the scalability of seed selection algorithms.

References

[1] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 61–70.

[2] D. Kempe, J. Kleinberg, and É. 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*. ACM, 2003, pp. 137–146.

[3] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *Knowledge-based intelligent information and engineering systems*. Springer, 2008, pp. 67–75.

[4] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 241–250.

[5] 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*. ACM, 2009, pp. 199–208.

[6] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 57–66.

[7] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 420–429.

[8] A. Goyal, W. Lu, and L. V. Lakshmanan, "Celf++: optimizing the greedy algorithm for influence maximization in social networks," in *Proceedings of the 20th international conference companion on World wide web*. ACM, 2011, pp. 47–48.

[9] B. H. Raven, "Social influence and power," DTIC Document, Tech. Rep., 1964.

[10] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring user influence in twitter: The million follower fallacy." *ICWSM*, vol. 10, no. 10-17, p. 30, 2010.

[11] E. T. Rosenman, "Retweetsbut not just retweets: Quantifying and predicting influence on twitter," Ph.D. dissertation, Bachelors thesis, applied mathematics. Harvard College, Cambridge, 2012.