Influence Maximization in Online Social Networks

Resa Brockman
Division of Science and Mathematics
University of Minnesota, Morris
Morris, Minnesota, USA 56267
brock245@morris.umn.edu

ABSTRACT

Online Social networks are an increasingly important part of our culture. They are now one of the dominant ways in which some people communicate, and the rate of that communication can be almost instantaneous. For that reason, the spread and diffusion of information throughout a network is an interesting phenomenon to understand. It especially can be a useful tool for marketing purposes where, specifically, the *influence maximization* problem is relevant. The goal of influence maximization is to find any given number of nodes (people) in a network that could spread some specific information to as large a portion of the network as possible. Solutions for this problem have been proposed since about 2003, and already several good approximation algorithms are in use. Current research mostly aims to improve results with novel techniques that focus on estimating more accurate *influence probabilities* between nodes. Other research in the area aims to include more information such as the susceptibility of certain people to certain information. Still other research aims to find trendsetters of a certain expertise in a network. There are many ways in which we can understand how information spreads in an online social network, and this information can be used as an advantage in influencing an entire online network of people.

Keywords

influence, influence maximization, social network, data mining, event cascade, trendsetters, susceptibility, information diffusion, similarity analysis

1. INTRODUCTION

Online Social networks have become a large part of modern culture, but are still a relatively new way for communities to interact. Understanding how they can affect our society is important and relevant. The spread of influence in a network via the *word-of-mouth effect* is particularly worth understanding in the interest of both marketing, and in understanding how information is spread and received throughout an online community. In this paper, the *influence maximization (IM)* problem will be defined, and current research into novel ways of improving the accuracy of existing solutions will be discussed.

This work is licensed under the Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/us/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.
UMM CSci Senior Seminar Conference, April 2016 Morris, MN.

2. BACKGROUND

Finding a way to spread specific information to as large a portion of a network as possible is an interesting and relevant problem. This issue has been defined as influence maximization [2]. The article *Online Influence Maximization* gives the following explanation of the problem:

> Given a promotion budget, the goal of IM is to select the best *seed nodes* from an *influence graph.* An influence graph is essentially a graph with *influence probabilities* among nodes representing social network users.

In essence, the end goal is to find a given number of seed nodes, which are specific users within the network with the highest expected *influence spread*, to introduce a product or idea to (for example, giving them free samples of a product) who will then cause a chain reaction of spreading the word to many of the nodes around them, who will spread the word to more nodes, etc. Influence probabilities are a way of quantifying the likelihood of the spread of information between two nodes [2]. Once a node has adopted the information or idea being spread, it is considered *active* [2, 1].

Influence maximization is based on previous information-diffusion problems that have been studied in the social sciences in the context of understanding the adoption of medical and agricultural technologies across a region [1]. The *independent cascade model* has been adopted from these other problems and is commonly used when working on influence maximization problems [2, 1]. It is sometimes also referred to as an event cascade model, or an information diffusion model. In its simplest form the independent cascade model starts with k number of active nodes in an influence graph. The active nodes will then attempt to influence adjacent nodes with the connecting influence probabilities of success. If success is achieved, the adjacent node has now been activated, and begins to attempt to influence its adjacent nodes. At the end of the simulation the amount of activated nodes are counted. Finding the best k seed nodes is an NP-hard problem, meaning the only way to find the absolute best solution is to explore every possible solution. Exploring every possible solution would be generally unfeasible due to time constraints, so the first step to getting a solution is to create an approximation algorithm [1].

Many effective approximation algorithms already exist as this problem has been well studied since early this millennium. Several approaches have been developed using methods such as greedy hill-climber techniques, which involve choosing an initial working solution algorithm and having a program that modifies the algorithm in some way, and then checking to see if the new algorithm is a better solution by comparing algorithm results to previous results from running the model. If it is better, the new algorithm is now the working solution. The current working solution is then modified again, and checked for improvement again. This process is continued for a specified amount of time, or until the rate of return begins to diminish on improvements to the algorithm.

There are several current IM algorithms that are well developed and used. The most recent algorithm, called *Two-phase Influence Maximization* (TIM) was developed in 2014 and will be discussed in this paper as a modern example of an IM algorithm.

3. METHODS

One of the most recent algorithms designed for influence maximization is TIM, which will be discussed here. Then, I will describe other current IM research, which is finding creative ways to improve results outside of designing new algorithms.

3.1 TIM

Prior to TIM, existing algorithms for IM had to attempt to balance approximation guarantees with practical efficiency, often leaving much to be desired in one way or another. The creators of the TIM algorithm aimed to create an algorithm with greatly reduced running time, while still achieving the same quality of results as existing algorithms. This would make large networks viable options for IM as the computing time needed for IM on the order of thousands of nodes can take days, but in reality, these techniques aim to be applied to billions of nodes, so time becomes an important limiting factor.

One of the original solution algorithms to the IM problem was published in 2003 by Kempe et. al., and was simple and effective. However, the run-time of a data set of even a few thousand was days long [1]. Since then, many algorithms have been proposed and used for IM, but they often sacrifice effectiveness of solutions for faster computation time.

TIM stands for *Two-phase Influence Maximization.* It is shown to return a solution equivalently good to Kempe's original algorithm in near-linear expected time. The time complexity of TIM is nearly optimal under the independent cascade model.

TIM is given G (the social network), k (the number of desired solution seed nodes), and two parameters l (a constant most often set to 1) and ε. ε is calculated using G and another variable r which is almost always set to 10000 based on recommendations in previous literature.

In the first phase G and k are fed as input into Algorithm 1 which returns the mean of the expected spread of influence, KPT^*. Then, also using n - the size of the network, TIM computes θ:

$$\theta = \lambda / KPT^*$$

where the equation for λ is:

$$\lambda = (8 + 2\epsilon)n \cdot (l \log n + \log\tbinom{n}{k} + \log 2) \cdot \epsilon^{-2}$$

Once the mean of the expected spread of influence is calcuated, the second phase can begin. TIM gives G, k, and θ as input to Algorithm 2, whose output S is the final result, a size-k node set of optimal seed nodes.

The parameter θ is designed to minimize the expected running time while ensuring solution quality. An RR set (seen in Algorithm 2) is a *Reverse-Reachable* set. That is, for each node, it is the set of other nodes in G that have a path to that node. A random RR set is an RR set from a randomly sampled single node from G. RR sets are called reverse-reachable sets because they are determined by starting at a node, removing and storing the identity of each connected node, and then going forward from each of the connected nodes until there are no more connections to follow. This, in essence, finds all the nodes that could potentially reach the node in question (thus reverse-reachable) but can also be considered the nodes that the initial node could influence.

In summary, Algorithm 1 takes in a network and desired number of seed nodes, and provides the expected spread of influence per node. This is used to calculate how many nodes should be randomly selected to ensure that enough potential quality seed nodes will be selected. This information is then used in Algorithm 2 which takes a random sample of nodes and returns the desired number of best seed nodes.

Algorithm 1 KPT Estimation (G,k)

Input: G (the network), k (desired number of seed nodes)
Output: KPT (mean of the expected spread of influence)

for $i = 1$ to $\log_2 n - 1$ **do**
    Let $C_i = (6l \log n + 6 \log(\log_2 n)) \cdot 2^i$
    Let $sum = 0$
    for $j = 1$ to c_i **do**
        Choose a random node and generate its RR set
        $k(R) = 1 - (1 - (\frac{w(R)}{m})^k)$ //determining the size of the RR set
        $sum = sum + k(R)$
    end for
    if $\frac{sum}{c_i} > \frac{1}{2^i}$ **then**
        return $KPT^* = n \cdot sum/(2 \cdot c_i)$
    end if
end for
return $KPT^* = 1$ //All nodes are expected to be influenced by any single node (extremely small or interconnected network)

TIM+ was also developed by the authors as a more generalized algorithm that does not assume use of the influence cascade model.

In a test round, it was found that when k = 50, εis greater than or equal to 0.2, and $l = 1$, TIM requires less than one hour to process a network with 41.6 million nodes and 1.4 billion edges. It is believed by the TIM authors that this is the first result in the literature that demonstrates reasonably efficient IM on a billion-edge graph[5]. Figure 1 shows the running time of TIM and TIM+ compared to two common IM algorithms.

Algorithm 2 Node Selection (G,k, θ)

Input: G (the network), k (desired number of seed nodes), θ (computed in 1)
Output: S (solution seed nodes)

Initialize a set $R = \emptyset$
Generate θ random RR sets and insert them into R
Initialize a node set $S = \emptyset$
for $j = 1$ to k **do**
    Identify the node v_j that covers most RR sets in R
    Add v_j into S
    Remove from R all RR sets covered by v_j
end for
return S

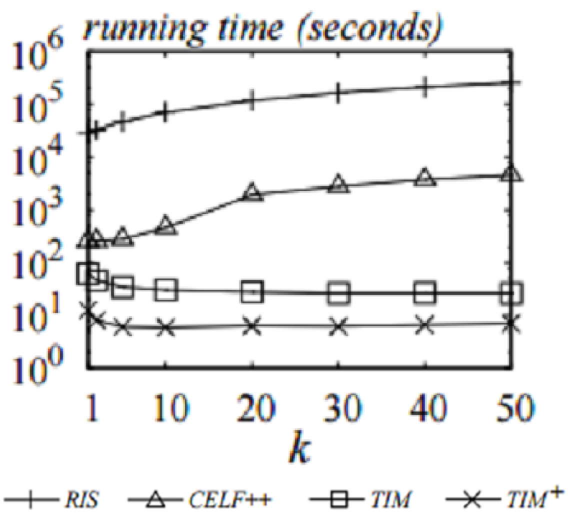

Figure 1: Run time of TIM and TIM+ compared to two previously preferred IM algorithms, RIS and CELF++. Taken from [5].

3.2 Dealing with Incomplete Data

As described in the background section, influence maximization results depend upon influence probabilities between nodes. This information is often determined using action logs of users' past activities in the network. However, this information can be both difficult to acquire and even more difficult to process, especially in large networks. In a 2015 paper *Online Influence Maximization*, a method is discussed for solving IM problems with missing or incomplete influence probability data using a multiple-trial approach. Any existing IM algorithm can be used with this method [2]

In summary, the approach (called *Online Influence Maximization* by the authors) begins the same way as most other IM approaches: by selecting a specific starter set of seed nodes and beginning a cascade from those nodes. However, as the cascade is run, user-feedback is used to update influence information. Then, another cascade round is performed, using the updated influence maximization information to choose new seed nodes.

In detail, first a selection of S starting nodes are selected using an existing IM algorithm. The influence probabilities between nodes are modeled as an average of a Beta distribution with some given initial α and β such that between two nodes the influence probability is the expected value of the

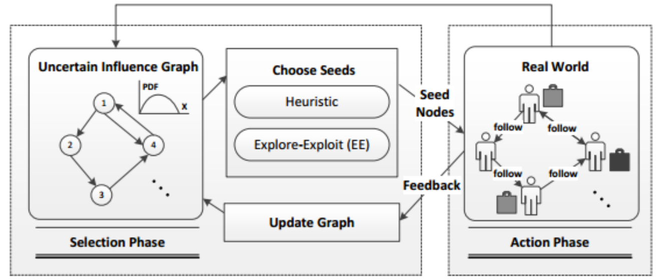

Figure 2: A graphic summary of the multi-trial approach. Taken from [2].

probability density function of a beta distribution with α and β indicating prior belief or knowledge, such that, while such information is missing, $\alpha = \beta = 1$.

$$\text{influence probability} = E[P \sim B(\alpha, \beta)].$$

Once the initial seed nodes have been selected, a real-world trial is started where an attempt is made to activate the seed nodes (perhaps by sending them an advertising message.) once the trial is complete, feedback information, which can consist of things like Retweets or likes, is gathered. This information is used to determine if a node was successfully activated. If a node was activated by another node, the influence probability between those two nodes is updated by adding 1 (success) to α :

$$\text{updated influence probability} = E[P \sim B(\alpha + 1, \beta)].$$

If a node was not activated (failure) the influence probability between nodes is updated by adding 1 to β :

$$\text{updated influence probability} = E[P \sim B(\alpha, \beta + 1)].$$

New seed nodes can be chosen using the updated influence graph, and another trial can be run. This can continue as long as the budget for trials does not run out, the improvements made each trial are not trivial, or the marketing campaign continues. This method for dealing with incomplete or missing influence data has been shown to find almost the same solution as is found when all the influence data is known, in a preliminary trial. [2]

3.3 Better Results Using Similarity Analysis

Similarity analysis is another field of study with social networks. According to the article *Influence and Similarity on Heterogeneous Networks* similarity analysis is:

> ...proposing methods for measuring nodes similarities, based on the network structure and node features.[6]

The article, by Wang et alia, goes on to propose using similarity findings to improve IM results and vice versa. They begin by taking a given network and splitting it into an influence network and a separate similarity network. A similarity network is similar to an influence network (or influence graph) but the nodes are connected with a similarity score instead of an influence probability. The two networks are then connected back together by what the paper refers to as *information tunnels*. The researchers named this configuration an *influence similarity (IS) network*.

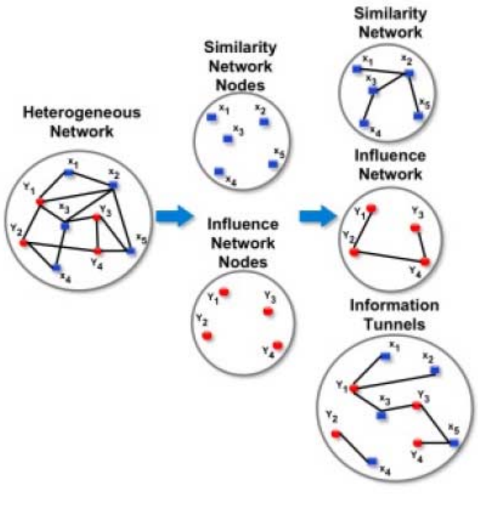

Figure 3: The separation of the similarity network and the influence network. Taken from [6].

This technique is application dependent, so before any work can be done, one type of node in the network needs to be fixed as the type of nodes on which similarity findings will be done, and another type of node needs to be chosen for the influence maximization findings. The categorization should be meaningful to the specific application.

The influence probability between two nodes, using this technique, is based on those nodes connections in the similarity network. This paper specifically deals with heterogeneous networks – networks with different types of nodes, like most social networks, therefore, the similarity scores between nodes are asymmetric. So a node Y could be more similar to a node Z, than node Z is to that Y. Traditionally, similarity score is computed by considering the number of common neighbors of the two nodes as a starting point for their similarity score. This starting score is then updated with an iterative process that takes into account the updated similarity values of the other nodes in the network so that two nodes' similarity depends on other nodes' similarities as well.

The change proposed by Wang et al. is to use influence probability information of connected nodes to weight the similarity score. So if a node W has a high influence probabilities with other nodes, the similarity score of nodes connected to W in the similarity network will be weighted to be higher than their original similarity scores. If the influence probability between nodes is extremely low, then the similarity scores will be weighted to be lower than its original scores. Then, the reverse will be done between the two networks. If nodes have high similarity scores, then connected nodes' influence probabilities will be weighted to be higher than they originally were. Lower similarity scores between nodes will weight influence probabilities between nodes to be lower. Using the new influence probabilities between nodes, seed nodes are chosen. These seed nodes are shown to be better than seed nodes chosen using uniform or poorly informed influence probabilities between nodes.

The technique was tested on a citation network. First, a paper citation network was used as a similarity network.

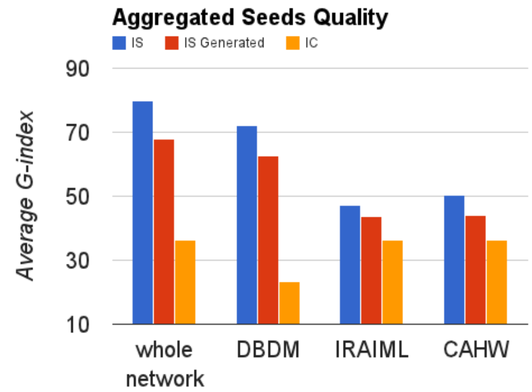

Figure 4: The g index scores of the IS network on comparison of seed nodes found using randomly assigned influence probabilities (IS Generated), and constant influence probabilities (IC) on different sections of a network. Taken from [6].

Then, an influence network was created using citation information. For example, if Author A cites Author B in their paper, then Author B has influence on Author A. Then similarity scores and influence probabilities were found using the new technique, and, finally, seed nodes were chosen from the influence network. These seed nodes were compared with seed nodes chosen from a network with uniform influence probabilities, as well as seed nodes chosen from a network with randomly generated influence probabilities. As can be seen in Figure 4, the seed nodes chosen using the IS technique are shown to have a higher *g-index* than the comparison seed nodes. A g-index is used for academic authors, and takes into account the number of papers they have published, the number of times their papers have been cited, and how exceptional some of their papers have been (so that it is better to have a few truly well cited articles and some lesser cited articles than to have many articles that are all cited about the same). A high rating on the g-index is a good indicator of influence for an author, and therefore it makes sense that seed nodes with a higher g-index are better choices. [6]

The paper also went on to show that the similarity analysis done using influence probabilities was better than current similarity analysis methods.

3.4 Including Susceptibility Information

The 2015 paper *Virus Propagation in Multiple Profile Networks* poses an interesting argument about the reality of information spreading through a network [3]. It points out that how information spreads throughout a network partially depends on how much each node or person in the network personally cares about that information. The example the paper gives is that if a new PS4 game becomes available, PS4 fans are very likely to share this information (e.g. by Retweeting information about it, or sharing links on Facebook.) Meanwhile, nodes that do not play PS4 games are much less likely to share or spread this information. The paper refers to this phenomenon as the *susceptibility* or *sensitivity* of nodes towards certain information. Higher sensitivity indicates that a node will become activated more easily in response to that information. Lower sensitivity indicates that it will be much harder to activate those nodes, but not impossible.

This paper looks at influence maximization in terms of a virus which is attempting to infect as much of a social network as possible. It uses an SIS model, which is an epidemiological model where each node can be in any of the two states: susceptible to infection or infected. As implied by our previous discussion, we use this epidemiological model in the general setting of information diffusion, meaning that a virus may as well correspond to a piece of information diffused on a social network subject to the rules posed by the SIS model. The paper's authors believe that this is the first time information diffusion in a network has been approached in such a way. This is assuming a heterogeneous network in relation to the virus where nodes have different levels of sensitivity against the virus. All nodes have the potential for infection even with a low sensitivity. The SIS model is also used to model no-immunity viruses such as influenza (common flu or cold).

The authors split nodes into two types, those with high sensitivity to the virus or information, and those with low sensitivity. Note that this model also includes a *healing rate* where over an amount of time a node can *heal* or become inactive after being activated. Nodes with high sensitivity (easy to infect) have a slower healing rate than those with low sensitivity.

This study also looks at two types of influence graphs *clique graphs* and *arbitrary graphs*. In clique graphs all nodes are connected to all other nodes. In this section, however, we will only focus on the results found for arbitrary graphs. It is also important to point out that rather than focusing on which nodes to infect at the beginning of a campaign, the authors were looking for a fixed point of equilibrium, which is, basically, how much of the network is infected when the infection is neither growing nor declining. They were more interested in the behavior of the network than in solving a specific influence maximization problem. However, their findings can be used in future influence maximization studies or techniques where the sensitivity of nodes is a factor in the spread of information. [3]

Although this paper did not have much in the way of applicable results, the authors did conclude that exploring a network with included susceptibility information was very complex and difficult, but also relevant for future work with information diffusion in a network. They suggest that demographic data, such as age, could be used to determine if a node has high or low sensitivity to an idea or product, and that such information should be taken into account when exploring how specific information spreads in a network. For example, the current main demographic for buyers of video games are teens and young adults. In an influence graph or network, influence probabilities could be weighted so that teen and young adult nodes are more susceptible to activation.

3.5 Finding Trendsetters

So far, we have reviewed the current fastest and most reliable IM algorithm, potential solutions for unknown influence information, improving results with the help of similarity scores, and the idea of using susceptibility information when determining the spread of influence. Now we will look at finding a specific type of influential node: *trendsetters*. Trendsetters are people who adopt and spread new ideas before these ideas become popular (which they eventually do). Trendsetters may not necessarily be well known or well connected, but their ideas still spread successfully through a network. Trendsetters are also early adopters, so they may be more susceptible to ideas in their area of expertise. In this way, they may also be relevant in any future susceptibility studies like the one discussed above. Trendsetters are relevant, as past research using number of Twitter followers has shown that a greater number of connections does not necessarily indicate a greater ability to influence [4].

Trendsetters must be two things: they must have a specific area of interest in which they have managed to spread content, and they must be early adopters of the information or products in that specific area. In this way, time is especially important to the problem of discovering trendsetters; we must be able to tell who will adopt ideas the fastest. The paper, *Finding Trendsetters in Information Networks*, suggests a novel approach to identifying trendsetters by including timing information in a given social graph [4]. In this research, rather than using influence graphs along with algorithms to determine the most influential nodes in a network, a social graph of nodes connected to other nodes with edges is used with the goal of ranking nodes in order of how much of a trendsetter they are. This is called their *Trendsetter Rank (TR)*.

Rather than creating a theoretical model of a given network and estimating a solution, this study was looking at past events and finding the trendsetters using Twitter data. They then compared the findings using TR (which includes time-of-adoption data) with another influential user ranking system called *Page Rank (PR)*.

One test study done was on an Iran-election timeline, which was a very important topic in 2009. The main hashtag used to talk about this was #iranelection, and other related tags included #iran and #tehran. It is very important to have a way to categorize what trends you are looking for, and Twitter makes this easy with hashtags. This sort of information could also be gleaned using a set of related memes, pins on Pinterest, or similar shared articles as examples for other social networks. Figure 5 shows the best result using the new TR method compared to the currently used PR method. The red dot shows the trendsetter chosen using TR, the green x shows the trendsetter using PR.

Another example is the #musicmonday which became popular on twitter and encouraged users to share music on Mondays. An example just like this that springs to mind is #tbt or #throwbackthursday where social media participants share old pictures of themselves on Thursdays. The researchers had Twitter data since the beginning of #musicmonday, making it a prime candidate for analysis. Their results compared with PR results can be seen in Figure 6. Again, the red dot shows the trendsetter chosen using TR, and the green x shows the trendsetter using PR.

Notice how in both graphs TR chooses earlier adopters than traditional PR, and the ideas of those adopters spread noticeably well throughout the network. Those are the defining characteristics of trendsetters. PR found the top trendsetters in both of these cases to be much more well known and famous members of society, but you can see from the graphs that the trend was already growing and spreading before those members began using the specific hashtags.

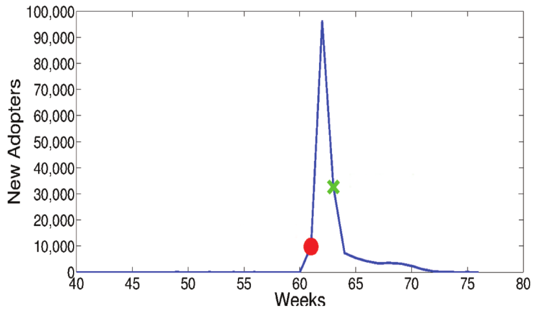

Figure 5: 2009 Iran Election Timeline. The red dot shows the trendsetter chosen using TR, where the green x shows the trendsetter chosen using PR. Note how much earlier in the timeline the TR chosen trendsetter adopted the hashtag. Taken from [4].

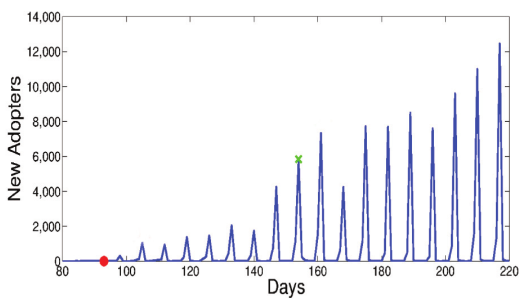

Figure 6: #musicmonday Twitter adoption timeline. The red dot shows the trendsetter chosen using TR, where the green x shows the trendsetter chosen using PR. Again, the TR chosen trendsetter shows up much earlier in the timeline. Taken from [4].

4. CONCLUSION

Influence maximization is an interesting problem in today's increasingly online-oriented culture. It is a well funded area of research due to its implications for marketing efforts, however, it has other potential uses in areas such as information or political campaigns.

Influence maximization is a complicated problem. Trying to turn a full social network containing a diverse group of people with complex social connections into something that can be understood as some sort of graph with nodes and influence probabilities is no small task. In spite of existing solutions, there is still room for progress.

It is likely that some improvements to existing algorithms will be made to improve the computation time of influence maximization problems, in the same way TIM did. But a lot of new research has started to focus on including more relevant data into influence maximization infrastructures in an attempt to improve results and to start with a more realistic understanding of the data.

Including information that takes into account the similarity of nodes and not just the number of connections each node has is important. Also important is considering what specific information is attempting to be spread, and how susceptible each node is to adopting or to continue spreading that information is also important. It is also useful to look at who is a trendsetter and who is not.

More information about the users of a social network is created every day. This information is complicated to work with and add into models, but ultimately, it will greatly improve our understanding of these networks as well as improve the results of computations of things like influence maximization. The goal of influencing other people in a society will never go away, but the ways we do so will continue to evolve.

5. ACKNOWLEDGMENTS

This writing would not have been possible without the guidance and support of several people, in particular K.K. Lamberty as a direct adviser to the paper as well as a professor of the Senior Seminar Class along with Elena Machkosova. Feedback from a previous University of Minnesota - Morris alumni, Kirbie Dramdahl, was also particularly helpful, as were my peers who helped significantly with the LaTeX formatting of this paper.

6. REFERENCES

[1] 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.

[2] S. Lei, S. Maniu, L. Mo, R. Cheng, and P. Senellart. Online influence maximization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 645–654, New York, NY, USA, 2015. ACM.

[3] A. Rapti, S. Sioutas, K. Tsichlas, and G. Tzimas. Virus propagation in multiple profile networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 975–984, New York, NY, USA, 2015. ACM.

[4] D. Saez-Trumper, G. Comarela, V. Almeida, R. Baeza-Yates, and F. Benevenuto. Finding trendsetters in information networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1014–1022, New York, NY, USA, 2012. ACM.

[5] Y. Tang, X. Xiao, and Y. Shi. Influence maximization: Near-optimal time complexity meets practical efficiency. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 75–86, New York, NY, USA, 2014. ACM.

[6] G. Wang, Q. Hu, and P. S. Yu. Influence and similarity on heterogeneous networks. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1462–1466, New York, NY, USA, 2012. ACM.