Influence Maximisation Towards Target Users on Social Networks for Information Diffusion

Olanrewaju Abdus-Samad Temitope, Rahayu Ahmad, Massudi Mahmudin
School of Computing, Universiti Utara Malaysia
olanrewaju@ahsgs.uum.edu.my

***Abstract*—Influence maximisation has been an area of active research in recent years. This study aims to extend the fundamental influence maximisation problem (IMP) with respect to a set of target users on a social network. It is important to aim at the target users to speed up the rate of information diffusion and reduce the information diffusion cost. In doing so, the MITU algorithm was formulated and compared with state of the art algorithms. Publicly available datasets were used in validating the proposed algorithm. It was found that the MITU identified all target nodes while significantly lowering the information diffusion cost function (IDCF) by up to 79%. The influence overlap problem was equally identified in the heuristic algorithm where the seed set size was reduced by an average of six times. Furthermore, the random influencer selection identifies target nodes better than the betweenness and PageRank centralities. The findings could help organisations to reach target users on social media in the shortest cycle.**

***Index Terms*—Influence Maximization Problem; Information Diffusion; Social Networks Algorithms; Target Users.**

I. Introduction

Due to the ubiquity and pervasiveness of social networks, they are now used by a large number of users as platforms for collaboration, innovation and sharing user related contents, which makes them an essential source of data for social-related research [1]. It affords extensive information dissemination effortlessly, which makes it an ideal arena for information diffusion in viral marketing [2].

Information diffusion is the process of propagation in a system regardless of the nature of the object (audio, video, text)[3]. A central characteristic of social networks is their ability to facilitate rapid information diffusion between large groups of individuals and shape people's opinions [4]. Due to its ability to influence, it has been extensively used in disasters [5] and marketing [6], [7]. In these scenarios, diffusing information always incurs a cost that is called the information diffusion cost function. Information Diffusion Cost Function (IDCF) can be defined as the number of times a message is being spread. It is a function of the number of hops through which the message was passed in the graph and the average number of times the information was shared at each hop [8], [9], with respect to the number of influencers [10].

The primary aim of information diffusion is maximising the spread of information, which leads to the Influence Maximization Problem (IMP), formulated by [11] and further established by [12]. It merely aims to identify, *k*, the set of users that cause the largest contagion on the network [12]. In doing this, the influencers need to be identified in the overall network structure. This procedure further leads to the influencer identification problem, which has generated a lot of research in recent years [13], [14], and more recently towards influencing a set of target users [7], [15]. A high diffusion cost function is good in scenarios where the information is aimed at the general population [8], while the reverse is the case concerning the target population. The importance of influencing target users is crucial because it speeds up the rate of information diffusion and reduces the information diffusion cost.

A target population, according to [7], is a set of users in the network that are selected based on criteria, in a bid to maximise influence towards them. Let us consider an e-commerce network that is comprised of buyers. A seller who wants to maximise the profit on babywear would be wise to target the people who recently talked about babies in their discussions because not all buyers would be interested in the product. This set of people who discussed babywear would form the target population that influences and would be the prime target for influence maximisation. Target population is under-researched, with few available studies in this area [5], [7], [10], [15]. Target users are crucial in viral marketing, because information is diffused towards them, and needs to reach them in the shortest time cycle and with lowest diffusion cost.

This study aims to formulate algorithms that can maximise influence towards target users on social networks, based on the IDCF [8]–[10] and influence spreading paths [7]. The following sections of the paper comprise a review of the relevant literature, formulation of the problem, the methodology and formulation of the algorithms followed by the results and discussion and a conclusion.

II. Related Works

A. Social Networks

Social networks according to [16] are an avenue through which users are able to share data and information that can be in the form of audio, video, picture or text. Users connect with each other by forming edges, which allow information to be transferred. According to [17], social networks can be classified into six broad groups; they are collaborative projects, blogs and microblogs, content communities, Social Networking Sites (SNS), virtual game worlds and virtual second worlds. Due to the rise to prominence of social networks, the rate at which individuals share data about their daily lives is growing at a very fast rate. This condition is largely due to the availability of smartphones [18], which allows for information diffusion on a large scale [19] with a wide outreach [20]. It also makes social media a fertile ground for different activities such as marketing [1], [21] and opinion formation [4], which are driven by information diffusion.

B. *Information Diffusion and the Influence Maximization Problem (IMP)*

Information diffusion maximisation is the main aim when spreading information on social networks. In maximising influence, the IMP was formulated by [11] and further established by[12]. This problem simply aims at identifying the minimal set of influencers that, if influenced, would lead to the largest contagion in the network [12]. Previous research selected influencers based on the overall network structure [8], [14], [22], [23]. Using this approach, high IDCF is accrued, and the information might not reach the intended users. Thus, there is a need to identify influencers concerning the target users [7]. Other recent research has proposed new problems under the IMP such as the Local Influence Maximization Problem [24] and the Information Coverage Maximization Problem (ICMP) [25]. In studying IMP, both the heuristic and greedy algorithms were used. The heuristic algorithms depend on efficient social network metrics, such as the centrality measures and *K*-shell, [2], [26], [5]. This approach is fast [23] but has low influencer identification, and influence spread [14], [23] and does not identify weak nodes as potential influencers.

In contrast, the greedy algorithm relies on the interaction between the nodes in the network. The greedy algorithm was first proposed by [12]. The algorithm takes all nodes in the network into consideration, by computing the influence of each node with respect to the overall influence on the network. This situation leads to the identification of a seed set that leads to the maximal influence spread in the network [12], [23]. Other studies enhanced the greedy algorithm proposed by [12], because it is not feasible on large networks [22], as it takes hours to days to compute the influence spread [14]. [27] were among the pioneers in enhancing the first GA proposed by [12]. They proposed the “Cost-Effective Lazy Forward” (CELF) algorithm, which is based on the submodular property of IMP, to estimate the influence spread and equally reduce computation time. [28] improved on the CELF algorithm under the ICM by limiting the influence spread of the nodes to the first hop, which reduces the time of computation and enhanced the CELF algorithm speed by up to 34%. [29] further enhanced the CELF algorithm through submodularity, where the marginal influence spread of a node was based on the last seed node evaluated. It further enhanced CELF algorithm efficiency by up to 55%.

Other studies have enhanced the greedy algorithm through graph localisation and paths. [30] enhanced scalability by making use of LDAGs (Local directed acyclic graphs) in computing influence spread. Furthermore, [6] proposed the maximum influence arborescence (MIA) model maximum, which was implemented based on the maximum influence path. This study enhanced the speed and scalability of the greedy algorithm by limiting their influencer identification to the first hop only and selecting higher degree nodes as possible influencers. The PMIA was further enhanced by [31], where influence was calculated through considering individual influence paths, excluding insignificant influence paths; paths are kept when there is no cycle, or the influence probability is less than the threshold. [32] proposed the Two-phase Influence Maximization (TIM) algorithm that enhances the greedy algorithm through constant-factor approximation, based on the reverse reachability of searches from the sample nodes. It is fast due to its use of heuristics to reduce processing time but is constrained by a specific seed set size. Due to this limitation, [22] proposed the Sketch-based Influence Maximization and Computation (SKIM) algorithm that makes use of sketched influence paths for nodes; the nodes with maximum influence, based on the sketch, are selected as the seed set to maximise influence spread. The TIM algorithm was further enhanced by exploiting estimation techniques based on martingales, which reduces the large memory footprint and consumption [33].

Under the traditional IMP, the number of active users at the end of the diffusion is sought to be maximised. In viral marketing, not only the active nodes are crucial, but the passive nodes, which get informed during the process of information diffusion are important to maximise profit. Passive nodes are those that are not successfully influenced, and therefore, cannot serve as an influencer to other nodes. Most of the time, these passive nodes are unknown in the network [25] and can jeopardise the efficiency of a viral marketing strategy. Based on this, [25] formulated the Information Coverage Maximization Problem (ICMP) that aims to maximise both the number of active and passive users. The lazy forward algorithm and the degree-based heuristics algorithm were proposed, which maximised the number of active and informed nodes in the network. The greedy algorithm has been widely used in identifying influencers [23] and learning influence probabilities [12]. The greedy algorithm has a good approximation guarantee but has high complexity [22], [23], low scalability [14] and does not identify weak nodes.

The study of influence towards target users is becoming more important due to its essential applicability to viral marketing [2], [34]. One of the pioneering studies toward influence maximisation towards target user(s) was carried out by [24], where they proposed the Local Influence Maximization problem (LIMP). Recent studies have built on this problem in two variants. On the one hand, there is influence maximisation towards a specific user, as done by [24]. [35] developed the IKA (Incremental Katz Approximation) algorithm which aims at suggesting friend recommendations that maximise influence towards a particular user on social networks. On the other hand, there is influence maximisation towards a set of target users. Research that followed this line includes [36], which studied influence maximisation towards a set of target users based on the topic selection. This research was carried out based on maximum influence arborescence (MIA) to compute the influence of nodes in selecting the seed set. [34] proposed the Multiple Acceptance Maximization (MAM) algorithm which aims at maximising the acceptance frequency of target users on social networks by preselecting the most influential seed set. The IMAX query preprocessing algorithm by [24] was based on the ICM. The IMAX and worked with a fixed seed and target node size. This procedure was done using influence spreading paths. The stated previous algorithms weaknesses included their fixed seed set size in [7], [34] and preselection of the seed set in [34]. The algorithm is equally computationally expensive and may not be feasible in real-world scenarios [37].

As stated, previous research on the IMP focused on maximising the number of active nodes at the end of the diffusion process. This study not only aims at doing that but equally stresses the importance of maximising the number of passive nodes, which is crucial to viral marketing. Previous studies [6], [7], [38] limited their influence spread only to the first hop of the seed nodes, which inadvertently leads to a larger seed set size. This condition is not ideal in viral marketing because the target nodes are spread all over the network, and not only is the influence spread important, the information diffusion cost to the nodes is equally required to be lowered while reaching the highest number of active and passive nodes.

III. Problem Formulation

The problem was identified based on the previous literature and has been discussed in the previous sections above. Based on identifying influencers for target users, the Minimal Influencer for Target Users (MITU) problem was formulated. A similar problem was formulated by [7], but it was based on a fixed seed set size and did not take into consideration the IDCF and nodes that are not found or unreachable. The proof of the submodularity and monotonicity of influence maximisation to a set of target users problem can be found in [7].

This problem aims at identifying the minimal seed set of influencers that would have the lowest IDCF while guaranteeing influence propagation to the target users. It is formulated under the independent cascade model (ICM), whereby a node propagates an item of information based on a probability. This propagation probability is derived by obtaining the inverse of the node's in-degree $1/deg^{-}(u)$. Let us consider a weighed directed network $G = (V, E, W)$, in which G is the network structure, $V = \{1, \ldots\ldots, n\}$ is the set of nodes in the network, $E \subseteq V\ X\ V$ is the set of edges in the network and $W = \{\, w_{i,j} \in [0,1] : (i,j) \in E\}$ is the set of activation weights calculated based on the WIC model. The probability of information propagation from node (u, v) is $p(u, v)$ where p is the activation probability derived from W. This study aims at identifying influencers toward a target user set $k \subset V$. The influencer set S $(S \subseteq V)$ would be chosen based on the nodes with the IDCF. The target nodes, k, identified would be influenced by nodes in S, which can be either through a direct or indirect influence.

The IDCF would be calculated based on the definition by [8]–[10]. It can be explained using mathematical equations that can be broken down into separate equations, where the diffusion function can be represented as:

$$n(\mu) = \sum \inf . p(d) \qquad (1)$$

where:

$n(\mu)$ is the diffusion cost function.
$\sum \inf$ is the total number of influencers.
$p(d)$ is the total path distance.

The total path distance can be represented as a function of the number of times the information was diffused and the mean number of steps it passed through [8]. This can be represented as a mathematical function, where:

$$p(d) = (n)(h) \qquad (2)$$

where:

n is the mean number of steps that the information flows through;

h is the mean number of times the message was transmitted in each step.

IV. Algorithms

In carrying out the study, the greedy algorithm was designed using the ICM approach. Equally, the heuristic algorithms for node centrality (degree, betweenness, and PageRank) would be modified to suit the target users.

In doing this, the greedy algorithm was broken into sub-algorithms for implementation. Algorithm one identified the target node component which was used in identifying influencers. Furthermore, nodes that had no in-degree were removed, since they cannot be influenced and, if considered, would lengthen the time of execution. The pseudocode of the algorithm is:

```
1  Input: Graph $G = (V, E)$; target users $\{K\}$
2  Output: reachable target users $T$, components $C$
   select components of node
3  for $k \in K$:
4      if $deg^{-}(k)$ >= 1:
5          $T$.append($k$)
6          $c$ = components of node $k$
7  return $T$,$C$
```

After identification of the reachable target nodes, there would be a need to identify possible influencers. In implementing this, the influencer paths were used. A random-number generator based on the Merssene-Twister algorithm was used to evaluate activation probability of the nodes. This algorithm is different from those in [6], [7], where the influence spread was limited to the first hop, The MITU algorithm goes beyond the first hop on line 15 of the algorithm. Algorithm two was used to identify possible influencers and passive nodes. The algorithm pseudocode is:

```
1  Input: Graph $G = (V, E)$; hops $h$, reachable target users
   $\{T\}$, components $C$
2  Output: Possible influencer tuple $dict(t, h, i)$, possible
   influencer list $p$
3  Dictionary $dict$ to hold possible influencer tuple is
   created
4  List possible influencer is created to hold the
   influencers
5  for $t \in T$:
6      Get the tree structure of $t$ based on the BFS and
         component $C$ is made up of edges $E$
7      While hop < $h$:
8          If $hop = 0$
9              for e $\in E$:
10                 Random number $rand$
11                 Edge weight $w$ of $e$
12                 If $rand < w(e)$:
13                     add $t$, $h$, $e[1]$ to $dict$, $ed[1]$ + $\{p\}$
14         If $hop > 1$:
15             for e $\in E[1]$ in previous hop:
16                 Get edges $Ed$ where $e[1] == ed[2]$:
17                 for $ed \in Ed$:
18                     If $rand < w(ed)$:
19                         add $t$, $h$, $ed[1]$ to $dict$, $ed[1]$ + $\{p\}$
20 return $dict$, $p$
```

Based on the identified possible influencers, the algorithm further had to identify the influencers that are able to influence the maximum set of users in the target. This technique was implemented based on the following algorithm:

```
input possible influencer list $P$, possible influencer tuple $dict$
Output =influencer dictionary $influencer\_dict(p, pl, n)$
for $p \in P$:
    $pl = 0$ //path length
    $n = 0$ //total target user seen
    for $d \in dict$:
        if $p == d[2]$:
            $pl += d[0]$
            $n+=1$
    add $p, pl, n$ to $influencer\_dict$
    break
return $influencer\_dict$
```

Based on the influencer dictionary, the influencers that influences the maximum numbers of nodes were selected until the target users were completely identified or cannot be reached. In doing this, the algorithm was implemented:

```
input influencer dictionary $influencer\_dict(p, pl, n)$, Target user $\{T\}$
Output seed set $\{S\}$, Information diffusion cost function $IDCF$
sort $influencer\_dict$ based on $n$ and $pl$
for $d \in$ sorted$(influencer\_dict)$:
    for $t \in T$ where $p$ is an influencer and $T \neq \{\emptyset\}$:
        $S \cup \{p\}$
IDCF = length {S} * total path distance
return $\{S\}, IDCF$
```

In summary, the overall greedy algorithm pseudocode is given below:

```
Input: Graph $G = (V, E)$; hops $h$, target users $\{T\}, S = \{\emptyset\}$
Output: seed set $\{S\}; IDCF$
Removal of unreachable nodes from $T$
Get nodes that are in the same component as $c \in T$
Get possible influencers for $t \in T$ based on $c$ using BFS
    on inward links, $h$ and $WC$ (for ICM)
Based on successive activation path group possible
    influencers based on (influencer, hops, target node)
Possible influencer (s) tuple re-arranged and sorted
    based on the number of target nodes found and
    minimal path length
Influencers are identified by taking nodes with the
    most target nodes and the shortest distance
Based on sorted tuple; $S \cup \{s\}$ while the reached
    target nodes $t$ are removed from $T$ until $t = \{\emptyset\}$
Calculate IDCF: length$(S)$ * total path distance
return $S, IDCF$
```

Based on previous studies that considered influence spread based on paths, which was limited to the first hop of the influencers [6], [7], [38], for evaluation purposes, the MITU algorithm was modified to calculate influence spread of its seed nodes based on the first hop. The influencer spread was done based on the Monte-Carlo simulation to achieve the average influence spread. Furthermore, the influencers out-edges were considered if the target user has a path to the influencer. In doing this, the algorithm pseudocode of MITU (1 hop) is given below:

```
input $G = (V,E)$, seed set $\{S\}$, target user $\{T\}$, number of simulation $n$
output influence spread $\{sp\}$
$simulation = 0$
while $simulation < n$:
    $sp = 0$
    for $S \in T$:
        sp+=1
        remove $S$ from $T$
    for $s \in S$:
        for $t \in T$:
            if $rand < w(t,s)$:
                $sp += 1$
                $t - \{T\}$
$sp = sp/rs$
return $sp$
```

The greedy algorithm was proposed due to its selection of influencers based on the nearest influencer and lowest IDCF. As its seed set size is not specified, it may be very high; this leads to the tradeoff of maximum influence spread or minimal set of influencers. In resolving this, an algorithm was further put in place to select the *k* seed set size, as specified, or the number of maximized influence nodes at the end of the diffusion process. This helped in comparing the final influence spread based on a specific *k* seed set size across multiple algorithms. This is crucial in viral marketing, where the tradeoffs need to be considered.

The heuristic algorithm was modified with respect to target users. This required the simulation to be initially done with respect to all the identified influencers. The execution of each path of the influencer was stopped when the activation sequence could not be guaranteed based on the WC and ICM. This was done to speed up the algorithm. The algorithm was enhanced to identify the best set of influencers from the initial seed set size, because of the influence overlap. The best set of influencers equally diffuses information to the same set of target users as the initial seed set does, but in a shorter time span and lower IDCF.

Heuristic algorithm based on ICM and WC:

```
Input: Graph $G = (V, E)$; hops $h$, target users $\{T\}$, seed set $\{S\}$
Output: optimal influencers $\{X\}$, found nodes $\{n\}$, optimal $IDCF$

Removal of unreachable nodes from $T$
For all $s \in S$; Get all reachable nodes based on the
    activation path sequence using BFS based on inward
    links
Based on successive activation path; group possible
    influencers based on $(i, pl, t)$
For $t \in T$ found; $n + \{m\}$; get the summation of
    paths for $s \in S$
Calculate IDCF: length($S$) * total path distance
6. Sort influencers based on the number of target
    nodes found and minimal path length
Based on sorted_tuple $X \cup \{s\}$; path_length +
    Minimal path length while $t - \{T\}$ until $T \neq \{\emptyset\}$
Calculate optimal IDCF: length($X$) * total path
    distance
```

V. Methodology

The study was carried out in five steps. It began by formulating the research problem (MITU) which was done in the preceding section. This was followed by the formulation of algorithms. Greedy algorithms were formulated based on the ICM approach. The formulated algorithms were simulated on various datasets. The algorithms were then evaluated with respect to state-of-the-art algorithms such as the PMIA [6], IMAX [7] and IRIE [38], and their parameters were used based on [6], [7]as the results for the algorithms were supplied by the author in that study. Four heuristic algorithms (degree, betweenness, PageRank centralities, Random) were equally used as a baseline for comparison under the ICM approach. In estimating the influence probability, the weighted cascade $1/deg^{-}(u)$ was used, as used by previous studies [7], [14], [22], [23].

In identifying the target users, 10% of the nodes in the network were randomly chosen, with the heuristic algorithms starting with an initial seed set size of 50, as done in [7]. The heuristic and greedy algorithms were reported because of the optimisation achieved in them in terms of IDCF, running time and final seed set size, which served as an improvement over [7]. The IDCF of the enhanced heuristic algorithms was at most times below a tenth of that of the heuristic algorithm itself. Meanwhile, the seed set size was 6–7 times lower than the initial seed set size, and the running time was also lower.

Table 1
Dataset Statistics Summary

Dataset	Node	Edge	Degree
Wiki-Vote	7K	104K	14.6
Epinions	76K	509K	6.7
Slashdot	77K	906K	11.7

VI. Results and Discussion

This section presents the simulation results derived from the experiment. In reporting the findings, the first section will explain the findings in the light of the previous research. Initially, MITU found all the target nodes in the dataset; however, the number of influencers were very large in the Epinions (803) and Slashdot (511) datasets. In a bid to make the findings comparable with the previous research, which had their seed set size at 50, the seed set size used was limited to 50.

As shown in Figure 1, the MITU (1 hop), which considers influence spread based on the first hop, like the PMIA, IMAX and IRIE, had a low influence spread, because the influencers were selected as a function of the entire target users which made it have a very poor spread. In contrast, since the PMIA, IMAX and IRIE only limit their influence spread to their first hop, the MITU outperforms them, as the influence spread of the seed set went beyond the first hop, and equally the all target users were aimed to be influenced. While all approaches were based on influence spreading paths, the MITU seed set influence path went beyond the first hop, and on the average, less than three hops, which further confirmed the findings of [37]. Based on its ability to go beyond the first hop, the number of its passive nodes was significantly higher, with close to 50% of the overall nodes in the network. This implies that, while an item of information can be diffused towards a set of targeted users, a significant number of the general users would be informed of the information, while equally maximising the number of target users that were influenced.

The algorithms could not be compared based on the IDCF, because of the result of the IRIE, PMIA, and IMAX simulations were provided by [7] and were not diffusion cost function oriented. Furthermore, the algorithms could not be compared based on running time, due to the difference in the language in which they were implemented, as MITU was implemented in Python while IRIE, PMIA, IMAX were implemented in C++. Moreover, the system configuration of the computer used in the simulation was equally different. IRIE, PMIA, and IMAX were implemented on an Intel(R) i7-990X 3.46 GHz CPU machine with 48GB RAM, while MITU was implemented on an Intel(R) i7-3537U 2.00GHz, 8GB RAM computer, which makes it difficult to evaluate the running time.

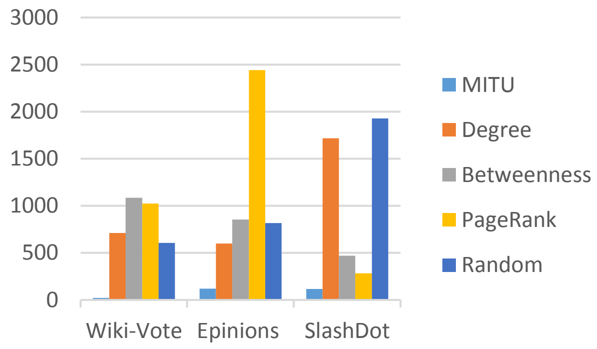

Figure 1: Influence Spread

The comparison of the MITU algorithm with the heuristic algorithm to evaluate its diffusion cost function and influence spread will be discussed in the following paragraphs.

Figure 2 shows the seed set size. For the heuristic algorithms, the initial seed set size was 50 and 0 for the greedy algorithms. On completion of the simulation, Wiki-Vote had a seed set size of 16, Epinions (803) and Slashdot (511). In order for the findings to be comparable with other heuristic algorithms, the top 50 influencers were selected for both Epinions and Slashdot. The seed set size in this study was not fixed, which is in contrast to [24] where it was fixed at 50. In viral marketing, the seed set size need not be fixed to maximise the outreach. The MITU had a smaller seed set size based on Wiki-Vote dataset, which was 68% smaller, while with the other two datasets it was over a 1000% larger. This is due to the lower diameter (shortest path distance) of Wiki-Vote. Moreover, Wiki-Vote has a higher clustering coefficient, which means the nodes are more tightly connected and a lower number of influencers is needed to cause a large contagion, which is unlike the case with Slashdot and Epinions.

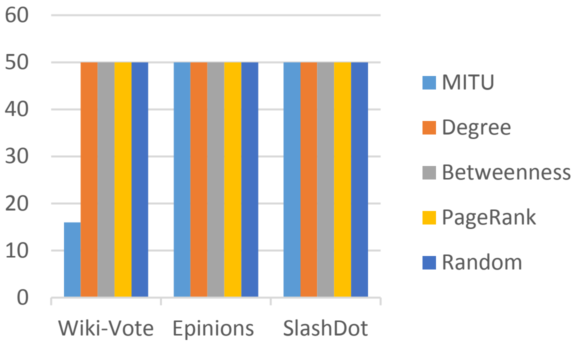

Figure 2: Seed set size

The heuristic algorithm seed set size was fixed at 50. On completion of the simulation, the enhanced heuristic algorithm was run to see if the same influence spread would be achieved. The enhanced heuristic algorithm seed set size was reduced on average by 88%. This was due to the existence of influence overlap, which has also been highlighted by previous studies [39], [44]. Recognizing the problem, this is one of the first few studies that aim at reducing influence overlap while still trying to accomplish the same outreach.

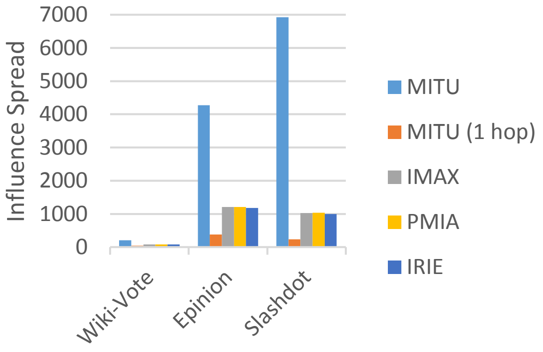

Figure 3: Found target users

Figure 3 shows the found target nodes. It was seen that the MITU was able to reach all reachable target nodes, while the heuristic algorithm had a very poor influence spread. This can be explained because probably the target users were not located on the first hop distance to the influencers identified by the heuristic algorithms. Another possible explanation could be the existence of components and communities in the network, which leads to a break in the activation sequence [39]. In comparison, the proposed greedy algorithm is superior, as it takes the component of each node into consideration and optimal influencers are found based on that.

It was seen that the random influencer selection outperforms the Betweenness, PageRank, and degree centrality in reaching target nodes. The degree centrality was only better on the Wiki-Vote dataset, due to its better clustering and shorter diameter, which shortens the distance between influencers and target nodes. The contrast could be seen on the Slashdot and Epinions, with larger diameters, and the nodes are more separated. PageRank algorithm had the worst influence spread; this is not surprising as previous studies have shown that it is better adapted to web pages and does not identify influencers that spread information [44]. The betweenness centrality was surprising, due to its wide application in identifying influencers [44], [45] and this might be a pointer that it is not effective in identifying them.

With respect to IDCF, Figure 4, based on Eq. (1), shows that the proposed greedy algorithm had a high IDCF because it identified all the nodes. In a comparative sense, the greedy algorithm IDCF was low because it identified all nodes. The IDCF of the random algorithm, which identified more users, was a bit lower than that of the greedy algorithm but it still had a high IDCF. Other algorithms had very low influence spread and relatively high IDCF, as the distances between the influencer and target nodes were wide.

The IDCFs of the greedy algorithms were lower because they chose influencers concerning the target users. Therefore the influencers were not necessarily the best with respect to the overall network. The influencers, on average, were less than three hops away from the target users. The heuristic algorithm did badly because the influencers were pre-selected as a function of the overall network; this made it difficult to identify all target nodes. Equally, degree centrality had very high IDCF because it is based on the number of its neighbour nodes, leading to more messages being transmitted but it remains factually ineffective.

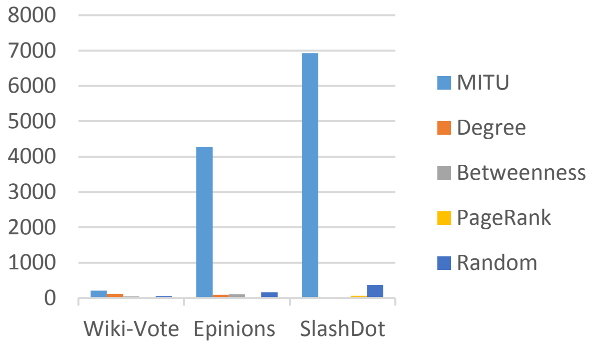

Figure 4: Comparative diffusion cost function

Based on the runtime, Figure 5 shows that the heuristic algorithm did better. This has been established in previous studies [23], but its weaknesses include its high IDCF, the total number of found nodes and the influencer seed set size. The greedy algorithm takes considerable time in establishing components, identifying potential influencers, sorting them based on the total number of found nodes and path length. All these make it more time-consuming than the heuristic algorithms.

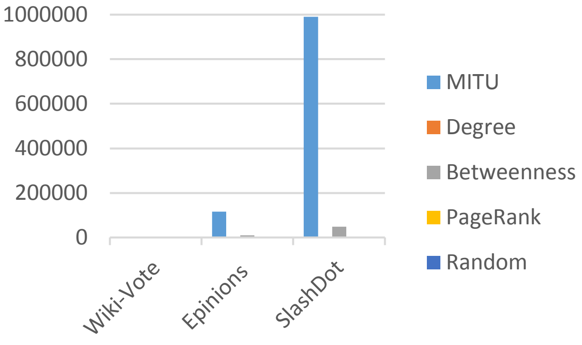

Figure 5: Runtime(s)

This research contributes to the literature in three ways. Firstly, it is one of the few studies that includes the concept of IDCF as a prerequisite for influencer identification with respect to the target user. It extended the research of [7] by incorporating the mathematical equations of IDCF derived from [8]–[10] into an influence maximisation algorithm. This makes it possible to identify influencers better without a fixed seed set size, which that is in contrast to [7]. In doing this, it incorporated the idea of informed (passive) nodes suggested by [25]. This is very crucial in viral marketing, where the primary aim transcends only reaching the right target audience with minimal cost, and also aims at equally maximising the size of the audience that knows about the information for reference. While it is believed that PMIE, IRIE and IMAX might have lower diffusion, their ability to reach lower target nodes makes them not ideal for the viral marketing scenario.

Secondly, it responded to the call of [2] by developing an algorithm that takes into consideration the diffusion cost, especially with respect to a set of target users. It made use of the ICM (sender-centric) [2] for the influencer identification. It was found that the ICM greedy had lower diffusion cost function and influenced more users, but its slower with respect to time. It further validated the of three steps of influence identified by [37], it was found that influencers to target users were on the average less than three hops away from them so as to mount indirect influence. This made weak nodes potential influencers, which is against the concept of heuristic-based influencer selection or the greedy based influencer selection with respect to the overall network. Furthermore, due to the more number of hops, the number of passive users increased which is crucial to viral marketing. This made it have an edge with respect to previous algorithm [7] where passive nodes were not considered, and moreover, their influence spread was limited to one hop. Passive nodes are crucial to viral marketing [25] as they can help enhance further contagion in the network.

Thirdly, the heuristic algorithm had low influence spread, which has been highlighted by previous studies [14]. This is due to the presence of influence overlap. The influence overlap problem has been identified by [22], and this study further confirms it and equally tried to reduce it. In doing that, an enhanced heuristic algorithm was developed which reduces the seed set size by an average of six times. The enhanced heuristic algorithm selects fewer influencers, which maximizes influence to the same set of users as the heuristic algorithm with little influence overlap. This leads to diffusion of information in shorter cycles. Thus, this study is one of the few studies that aims at reducing the influence overlap problem in the heuristic algorithm.

Furthermore, it helps in redefining the notion of influencers. Based on the findings, it was found that influencers selected at random outperformed those based on the centralities algorithm. While the PageRank algorithm has already been stated not to be helpful in identifying influencers on a social network [26], its performance in terms of the degree and betweenness centralities was surprising. While such poor performance can be attributed to the social network structure, such as clustering and shortest path distance diameter, which would work well based on the overall network. It cannot be said of target users that are distributed over the network, where the betweenness centrality of the influencers might not be helpful if the target users are not near the centre of the social network, nor the degree centrality if they are not situated near to the nodes with high degrees. This leads to questioning the efficacy of centralities measures in identifying influencers for target users. While previous research has highlighted their low efficiency which is a problem fundamental to heuristic algorithms [14], it was not expected that random selection of influencers would be better in all simulations. Previous research has discussed influencers with the overall users, as those that are most popular or more central in the information pathway [26], [5]. However, this is negated here, where it is found that influencers towards a set of users are not necessarily the most influential but are the most effective in diffusing information to the set of target users, while equally having a high number of informed (passive) users. The findings of this study can help organisations in streamlining and selecting their influencers while trying to maximise their outreach to their target users. It would also help in cost reduction, while equally leading to faster information dissemination.

VII. Conclusion

In conclusion, in this era of big data, there is a need to identify influencers on social networks that are specifically engineered towards the target users. This would help in reducing the number of times information was spread while maximising information outreach and influence. This would help drive innovations, viral marketing and customer-based outreach (B2C), while equally reducing the amount of generated information, especially in this time of information overload, which might have little or no benefit towards the aim it is meant to fulfil. In the future, this study aims to validate the algorithm on more data sets and other study algorithms. Equally, an algorithm to suggest possible influencers based on the unreachable nodes could be formulated.

References

[1] L. Yafeng, W. Feng, and M. Ross, “Business Intelligence from Social Media: A study from the VAST box office challenge,” *Bus. Intell. Anal.*, 2014.

[2] L. Alsuwaidan, “Toward Information Diffusion Model for Viral Marketing in Business,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 2, 2016.

[3] A. Guille, H. Hacid, and C. Favre, “Predicting the temporal dynamics of information diffusion in social networks,” *arXiv Prepr. arXiv1302.5235*, 2013.

[4] M. Ramos, J. Shao, S. D. S. Reis, C. Anteneodo, J. S. Andrade, S. Havlin, and H. A. Makse, “How does public opinion become extreme?” *Sci. Rep.*, vol. 5, p. 10032, 2015.

[5] A. T. Olanrewaju, R. Ahmad, and K. F. Hashim, “Information Exchanges of Social Media Evangelists during Flood: A Social Network Analysis,” *J. Teknol.*, vol. 78, no. 9–3, pp. 49–55, 2016.

[6] W. Chen, C. Wang, and Y. Wang, “Scalable influence maximization for prevalent viral marketing in large-scale social networks,” *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD ’10*, p. 1029, 2010.

[7] J. R. Lee and C. W. Chung, “A query approach for influence maximization on specific users in social networks,” *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 2, pp. 340–353, 2015.

[8] Y. Li, M. Qian, D. Jin, P. Hui, and A. V. Vasilakos, “Revealing the efficiency of information diffusion in online social networks of microblog,” *Inf. Sci. (Ny).*, vol. 293, no. September, pp. 383–389, 2015.

[9] L. Weng, F. Menczer, and Y.-Y. Ahn, “Virality Prediction and Community Structure in Social Networks,” *Sci. Rep.*, vol. 3, p. 2522, 2013.

[10] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, “Information Diffusion in Online Social Networks: A Survey,” *ACM SIGMOD*, vol. 42, no. 2, pp. 17–28, 2013.

[11] 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, 2001, pp. 57–66.

[12] D. Kempe, J. Kleinberg, and T. Eva, “Maximizing the Spread of Influence through a Social Network,” in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 137–146.

[13] F. Morone and H. a. Makse, “Influence maximization in complex networks through optimal percolation:supplementary information,” *Curr. Sci.*, vol. 93, no. 1, pp. 17–19, 2015.

[14] M. Heidari, M. Asadpour, and H. Faili, “SMG: Fast scalable greedy algorithm for influence maximization in social networks,” *Phys. A Stat. Mech. its Appl.*, vol. 420, pp. 124–133, 2015.

[15] M. Mirbabaie, C. Ehnis, S. Stieglitz, and D. Bunker, “Communication roles in public events – A case study on Twitter communication,” in *Information Systems and Global Assemblages. (Re)Configuring Actors, Artefacts, Organizations*, 2014, pp. 207–218.

[16] T. Tang, M. Hämäläinen, A. Virolainen, and J. Makkonen, “Understanding user behavior in a local social media platform by social network analysis,” *Proc. 15th Int. Acad. MindTrek Conf. Envisioning Futur. Media Environ. - MindTrek ’11*, p. 183, 2011.

[17] A. M. Kaplan and M. Haenlein, “Users of the world, unite! The challenges and opportunities of Social Media,” *Bus. Horiz.*, vol. 53, pp. 59–68, 2010.

[18] S. Philpott, “I ’ ll have what she’s having ! The importance of Social Network Analysis,” 2010.

[19] Q. Liu, B. Xiang, E. Chen, H. Xiong, F. Tang, and J. Xu Yu, “Influence Maximization over Large-Scale Social Networks : A Bounded Linear Approach,” in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 2014, pp. 171–180.

[20] L. Palen, S. R. Hiltz, and S. B. Liu, “Online forums supporting grassroots participation in emergency preparedness and response,” *Communications of the ACM*, vol. 50. p. 54, 2007.

[21] W. Chen, F. Li, T. Lin, and A. Rubinstein, “Combining Traditional Marketing and Viral Marketing with Amphibious Influence Maximization,” in *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 2015, pp. 779–796.

[22] E. Cohen, D. Delling, T. Pajor, and R. F. Werneck, “Sketch-based Influence Maximization and Computation: Scaling up with Guarantees,” *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manag. (CIKM ’14)*, no. Ic, pp. 629–638, 2014.

[23] S. Cheng, H.-W. Shen, J. Huang, W. Chen, and X.-Q. Cheng, “IMRank: Influence Maximization via Finding Self-Consistent Ranking,” in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014, pp. 475–4.

[24] J. Guo, P. Zhang, C. Zhou, Y. Cao, and L. Guo, “Personalized influence maximization on social networks,” in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, 2013, pp. 199–208.

[25] Z. Wang, E. Chen, Q. Liu, Y. Yang, Y. Ge, and B. Chang, “Information Coverage Maximization in Social Networks,” *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 0, no. 0, pp. 2104–2110, 2015.

[26] S. Pei and H. A. Makse, “Spreading dynamics in complex networks,” *J. Stat. Mech. Theory Exp.*, vol. 2013, no. 12, p. P12002, 2013.

[27] 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*, 2007, p. 420–429).

[28] W. Chen, Y. Wang, and S. Yang, “Efficient influence maximization in social networks,” in *international conference on Knowledge discovery and data mining, KDD 2009*, 2009, vol. 67, no. 1, p. 199.

[29] A. Goyal, W. Lu, and L. V. S. Lakshmanan, “CELF ++ : Optimizing the Greedy Algorithm for Influence Maximization in Social Networks,” in *Proceedings of the 20th international conference companion on World wide web*, 2011, pp. 47–48.

[30] W. Chen, Y. Yuan, and L. Zhang, “Scalable influence maximization in social networks under the linear threshold model,” *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 88–97, 2010.

[31] J. Kim, S. K. Kim, and H. Yu, “Scalable and parallelizable processing of influence maximization for large-scale social networks?,” in *Proceedings - International Conference on Data Engineering*, 2013, pp. 266–277.

[32] 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*, 2014, pp. 75–86.

[33] Y. Tang, S. Yanchen, and X. Xiaokui, “Influence Maximization in Near-Linear Time : A Martingale Approach,” in *SIGMOD ’15 Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 2015, pp. 1539–1554.

[34] C. Chang, “On Influence Maximization to Target Users in the Presence of Multiple Acceptances,” in *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2015, pp. 1592–1593.

[35] S. Kim, “Friend Recommendation with a Target User in Social Networking Services,” in *31st IEEE International Conference Data Engineering Workshops (ICDEW)*, 2015, pp. 235–239.

[36] S. Chen and K. Tan, “Online Topic-Aware Influence Maximization,” *Proc. VLDB Endow.*, vol. 8, no. 6, pp. 666–677, 2015.

[37] Y. Qin, J. Ma, and S. Gao, “Efficient influence maximization under TSCM: a suitable diffusion model in online social networks,” *Soft Comput.*, pp. 1–12, 2016.

[38] K. Jung, W. Heo, and W. Chen, “IRIE: Scalable and Robust Influence Maximization in Social Networks,” in *IEEE 12th International Conference on Data Mining (ICDM)*, 2012, pp. 918–923.

[39] F. Liu, “Constrained Opinion Leader Influence in an Electoral Campaign Season: Revisiting the Two-Step Flow Theory With Multi-Agent Simulation,” *Adv. Complex Syst.*, vol. 10, no. 2, pp. 233–250, 2007.