Influencer Selection on Social networks based on Information Requirement and Diffusion Cost

Olanrewaju Abdus-Samad Temitope, Ahmad Rahayu, MassudiMahmudin

Abstract: ***Viral marketing is vital to the success of business in this age. Information diffusion on social networks for viral marketing involves selecting a seed set of influencers (nodes) to be infected which leads to an activation process in the network with the aim of infecting a maximum number of nodes. The existing models have selected the influencers based on the node properties (centralities) but do not take into consideration the diffusion cost in spreading the information. In addition, the influencers are selected without considering the need for diffusing information. This study proposes a general additive model that uses a tuneable weight on four centralities in selecting influencers. Our results shed more light on the trade-off between the outreach of information and the diffusion cost incurred. The results demonstrated that selecting the top influencers using a single metrics is not necessarily effective when diffusing information. This study also discovered a positive effect in an increase of the size of the influencers does not always yield an increase in the relative outreach depending on the type of the network.***

Keyword: ***Social networks, centrality, information diffusion, diffusion cost***

I. INTRODUCTION

Social networks are used by businesses to reach out to their clients in the fast-paced business environment. [1]More than 2 million businesses advertise on social media to reach their client. Information diffusion is the process of propagation of information in a system regardless of nature [2]. A central characteristic of social networks is the ability to facilitate rapid information diffusion between large groups of individuals and shape people opinions [3, 4]. The influence ability of social networks has allowed it to be extensively used in marketing, elections, and disaster management [5, 6]. Since social networks are ubiquitous, the process of information diffusion has been actively researched to maximize the spread of information in viral marketing [7]. Previous studies showed that influencers on the social networks are always minute and negligible in comparison to the overall population in the network [8, 9]. However, they are vital for information diffusion and influence maximization where they can be opinion leaders [10, 11]. Therefore, it is critical to determine these influencers and use them to spread the information with low diffusion cost.

Influencers are individuals responsible for information contagion in the overall network [12, 13]. In real scenarios, they are crucial for large information cascade generation [14] which leads to viral information outreach and product adoption [5]. Influencers differ based on what they do and how they influence people [15, 16]which necessitate the need for influencers' selection based on the purpose of information while also reduces diffusion cost. Information diffusion cost function can be defined as the number of times that a message is being spread which is a function of the number of steps in the graph that the message flows through and the average number of times the information was shared at each step[17, 18].

In selecting the best influencers, previous studies have developed algorithms and models using social network metrics such as the network centralities, PageRank and k-core neighbours [19, 20]. A single network centrality measurement usually selects the influencers. Each measurement has their strength and weakness [21].Thus, it would not be advisable to select based on a single metric. Previous research has stated that influencers are selected based on the need for information diffused which enhances the information outreach [7, 16, 22] and minimizes the cost of viral marketing and social commerce. This is important as the requirement for gathering opinions on the social network is different from spreading an advertisement. Not only the requirements differ, but the size of the audience is also dissimilar.

This study aims at investigating information diffusion on social networks by developing and evaluating an algorithm that selects influencers based on the needs or purpose for diffusing information. In doing this, the underlying characteristics of the network would be considered. Additionally, the study would build on the Independent Cascade Model (ICM) model [23, 24] based on the weighted information cascade (WIC) in evaluating influence [24]. The remainder of the paper is structured by giving a brief review of the literature. This was followed by the methodology and the development of the algorithm. The algorithm was then evaluated based on different datasets and the results were presented. Also, the implications of the findings were discussed.

Revised Manuscript Received on March 08, 2019.

Olanrewaju Abdus-Samad Temitope, School of Computing, Universiti Utara Malaysia Sintok, Kedah, Malaysia olanrewaju@ahsgs.uum.edu.my

Ahmad Rahayu, School of Computing, Universiti Utara Malaysia Sintok, Kedah, Malaysia. rahayu@uum.edu.my

MassudiMahmudin, School of Computing, Universiti Utara Malaysia Sintok, Kedah, Malaysia. ady@uum.edu.my

II. LITERATURE REVIEW

Influence maximization problem (IMP) aims at selecting a seed set that will maximize the spread of information in a network. IMP was shown to be an NP-hard problem by [24]. In achieving the maximal spread of information, previous research has made use of greedy and heuristic approaches in selecting the influencers [25]. Previous algorithm demonstrated that the greedy algorithm incur a high computational cost and low scalability [26, 27]. Alternatively, Heuristic algorithms take into consideration the network structural properties and node characteristics in selecting influencers for the network [16].

In selecting influencers based on heuristic algorithms, the degree, betweenness, and closeness centralities are fundamental [16]. Previous studies have used the centralities in identifying influencers [e.g. 21, 28]. The selection of influencers was made for information diffused changes based on the time and the needs of information in a bid to maximize the information outreach [7, 22]. Previous research has highlighted that different network measures were used to identify influencers yield influencers with different characteristics [16, 21]. Furthermore, the nodes structural properties are used to identify influencer for various dissemination needs which lead to variation in the node importance during the information diffusion process[7, 29]. This can be seen in a scenario where a wide outreach is needed, a user with high in-degree is preferred for information dissemination while a user with high betweenness and low clustering coefficient would be ideal for new idea generation. In respect to viral marketing, [30]emphasized the crucial need to select influencers with a minimal seeding cost for the largest outreach.

Several Multi-Criteria Decisions Making (MCDM) approaches such as the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), Analytic Hierarchical Process (AHP), and simple additive weightage (SAW) were employed in selecting influencers. [30] used a mathematical model with weightage to identify influencers. In the model, the parameters include average path length, clustering coefficient, and degree centrality with weightage. Also,[31] assigned a tuneable weightage to identify different influencers in the blogosphere where they were evaluated using in-degree, out-degree, and number of comments as well as PageRank. [32] used the TOPSIS approach in aggregating degree, closeness, and eigenvector centralities to select influencers. It was observed that the nodes selected based on TOPSIS had better influence spread under the SI model than betweenness and eigenvector centralities while the closeness and degree centralities had better spread. However, the authors assumed that all centralities have the same level of importance. [33] extended the TOPSIS model by assigning weightage of importance to the centralities (degree, betweenness, and closeness) where their selected influencers had better influence spread than TOPSIS. [34] used AHP to select influential nodes based on the basic centralities (degree, betweenness, and closeness) with their method outperforming W-TOPSIS. The selected top ten seed had 6,8,7,10 similarities with the clustering coefficient, degree, closeness, and betweenness centralities, respectively. This denotes that the known centralities are still effective in identifying influential nodes in the network. [35] proposed the hybrid degree centrality (HC) which integrates the degree and LeaderRank centrality based on a spreading probability. Their method identified better influencers with a better spread under the SIR model.

In a different approach, the node properties and the social network properties were used to predict the cascade size of the information diffused. [36] modelled the outreach of a message by proposing the message affinity model (MAM) which made use of the node metrics (degree centrality), network metrics (clustering coefficient and average path length), and dynamic parameters (Transmissibility and Fanout Coefficients). [37] predicted the cascade size of a tweet based on the original tweet cascade, retweet cascade, and the prediction of the final cascade size. Network structure affects information diffusion in respect to cascade size, speed of diffusion, and the individual involved. [14] predicted the cascade size based on several features such as the content, root, resharer, structural, and temporal features. [38]used the history of the network to predict the expected number of activations based on the seed set while [7] used Eigenvector centrality in selecting potential influencers by seeding at different activation times in the diffusion process. While previous studies have identified the need for selecting influencers to disseminate information, these studies did not include diffusion cost function. This study improved on previous studies by including DCF which helps in determining the trade-off amongst the influencers. This study stressed that the most popular (degree), or central (closeness), or closest (betweenness) influencers are not necessarily the most influential at every point in time, but an aggregation of influencers based on several centrality measures will perform better at a relatively lower diffusion cost.

III. ALGORITHM

The algorithm was developed based on the GAM model. The loaded graph includes the dataset that was used as previously stated in the methodology section. This was followed by generating all the graph structural properties which were calculated and stored in csv files to reduce computing time. The normalization of the clustering coefficient centrality (CLC) was done by subtracting one from the score. For example, a node of 699 had a CLC of 0.2253061224489796 while a node of 67 had CLC of 1.0. In CLC, the lower the CLC, the better. In a bid to reflect the true scale, the CLC had to be normalized so that 699 has a higher score than 67.

This was followed by the standardization of the centralities. This step is important to avoid the dominance of a particular centrality over the others in the network. For example, the betweenness centrality has large values while the closeness centrality has small values.This leads to the dominance of betweenness centralities in the node rankings. In standardizing this, the values of all centralities were equated in a continuum between 0.0000000001 and 10.

Ten decimal places were used to preserve the distribution in the measures, where the difference between very close measures can be accounted for.

The algorithm is presented below:

1. load the graph
2. generate the graph nodes weight and store it in a dictionary
3. read the pre-generated graph node properties details (betweenness, degree, closeness centrality) from pre-generated csv files
4. select the top 10% of each centrality as potential influencers
5. select common nodes amongst all centralities as final influencers to be considered
6. normalize all the centrality measure to remove the dominance of some centralities over the others
7. compute each node score based on the GAM model
8. based on the ranking select the top 10, 50, 100 and 200 nodes
9. compute the reshare and structural properties based on ICM and WIC
10. compute the DCF
11. End

IV. METHODOLOGY

This study utilizes Simple Additive Weightage (SAW) or more commonly known as General Additive Model in other disciplines. The approach has been extensively used in several fields [39, 40]. This model was employed in this study instead of TOPSIS or AHP due to the independence of the constructs [41] and its flexibility in modelling non-linear relationships were fewer subjective decisions to be made [39-41].However,[40] criticized GAM asit does not always reflect real situation and result may be illogical. This was discounted by [39] that making the summation of the sum to always be one allows for a strict and rational weightage assignment and assessment. This was implemented in selecting the influencers. In summary, a general additive mathematical model was chosen since it is flexible, simple, supports the independence of each parameter, the contribution of each parameter and its coefficient have been checked and an easy way to predict a value based on a set of covariates is provided [42].

In identifying influencers based on the need, node characteristics such as betweenness centrality, closeness centrality, degree centrality, and clustering coefficient were taken into consideration as they are the most revered heuristic centralities. Each of this heuristic centrality with the exception of the clustering coefficient were considered and used by previous studies. The degree centrality is important based on its ability to reach large audience at its first hop, The betweenness centrality is crucial because of its ability in serving as a bridge between two communities which offers the ability to shape opinions and generate new ideas. The closeness centrality would be considered due to the average distance of the node to other nodes in the networks which would lead to quick dissemination of information.

Eigenvector and PageRank centralities were excluded since they evaluate influence based on the nodes neighbour [21]. This is not feasible when selecting influencers for marketing because previous research has identified that they work better on undirected graphs [43] and they assign majority of the scores to few nodes in the graph [8]. Due to this, the clustering coefficient was used in achieving a similar purpose where nodes with little overlap in their immediate neighbourhood were selected to reduce redundancy in the spread of information. The general additive mathematical model was used to evaluate the heuristic centralities when identifying the influencers.

The general additive model is:

$$f(x) = \alpha_1 X_1 + \alpha_2 X_2 + \ldots\ldots + \alpha_n X_n \quad (1)$$

where

α is the coefficient

X is the parameter

In the adoption of the model to this study, Equation 2 was derived:

$$f(u) = \alpha_d D_u + \alpha_b B_u + \alpha_c C_u + \alpha_{cl} Cl_u \quad (2)$$

where

$f(u)$ is the node score

u is the node

α is the coefficient of the heuristic measurement which is a tuneable parameter

*D*is for degree centrality

B is for betweenness centrality

C is for closeness centrality

Cl is for clustering coefficient

The α of each heuristic was selected between 0 and 1 where$\sum \alpha \leq 1$. The α for each heuristic was assigned based on the purpose of information needs i.e. *B* was assigned a high score when interested in idea generation while *D* was higher when immediate wide outreach was crucial. The heuristics werenormalized and standardized to reduce the over dependency of a single centrality in the overall measurement.

Based on the heuristics parameters stated above, influence scores were computed for each node. According to previous studies, influencers in a network are typically within the range of 5-10% of the network [9, 44]. For each centrality measure excluding the clustering coefficient, the top 10% nodes were selected. Clustering coefficient was excluded because of it been weak in predicting influencers as it only considers redundancy and not outreach which is crucial in information outreach. Common nodes in the 10% sample for degree, betweenness and closeness centralities were then taken to improve accuracy of the result and reduce computation time.

Based on the objective of this study, the spread of the information was estimated. In predicting the spread, studies have stressed that influence estimation and cascade size computation are better in small hops and the spread was limited to two hops as done by previous studies [45, 46]. The information diffusion covered both the active and informed nodes (was not activated successfully) as it is vital in assessing the extent of information spread [47].

The information diffusion characteristics were discussed along with highlighted by earlier studies [e.g. 14, 36, 48]. Information diffusion characteristics were evaluated based on the studiesof [14] and [48].

However, the content feature was excluded as this study does not consider content that was shared.Also, the temporal features wereremoved due to the absence of time function in the datasets to evaluate the role played by time differences. The characteristics were further calculated based on the measure developed by categories mentioned by [36] which include:

N_s: number of seed nodes

N_v: number of viral nodes

N_i: number of informed nodes

N: Number of nodes in the network

N_a: Number of active nodes or outreach (Sum of both active and viral nodes)

Original seed nodes features: Nodes degree with much emphasis on out-degree, betweenness centrality, closeness centrality and clustering coefficient

Reshare features: The reshare features are used to study the growth rate of the cascade with respect to the overall network population and how fast it spreads. The transmissibility is growth with respect to overall population in terms of nodes that turn tosecondary spreaders. The fan-out coefficient is the average number of activations made by the seed and viral nodes that lead to the increase in cascade size.

1. Transmissibility is a measure of the message content with respect to how accepted it is. This can be calculated by: $\lambda = \frac{N_v}{N - N_s}$

2. Fan out co-efficient of seed nodes $\gamma_s = {}^{N_a}/_{N_s}$

3. Fan out co-efficient of viral nodes $\gamma_v = {}^{N_a}/_{N_v}$

4. Basic Reproductive Number R_0 or average number of secondary recommendations produced by reached nodes as $R_0 = \lambda\gamma_v$ All were adapted from [36]

Information Diffusion Cost Function: The DCF was evaluated based on the definition by [17, 18]. It is explained by the mathematical equation where:

$$n(\mu) = \sum \inf . p(d) \quad (3)$$

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 [14]. This can be represented as a mathematical function:

$$p(d) = (n)(h) \quad (4)$$

where:

n is the mean number of steps that the information flow through *h*is the mean number of times the message was transmitted at each step.

The diffusion cost function was split into the initial diffusion cost function incurred due to the influencers spreading the message and the overall diffusion cost function equally called DCF which was incurred at the end of the simulation due to the viral nodes. Five datasets collated from the SNAP library [49] were used and are presented in Table 1.

Table. 1Datasets statistics summary

	Wiki-vote	Facebook	Twitter	Epinions	Slashdot
Total number of nodes	7115	4039	81306	75879	77360
Total number of edges	103689	88234	1768149	508837	905468
Average clustering coefficient	0.1409	0.6055	0.5653	0.1378	0.0555
Diameter	7	8	7	14	10

V. RESULTS

A total of 20 simulations were carried out on three datasets to test the proposed GAM model and the results were evaluated. The total weight of the centralities was equal to one. Table 2 demonstrates the first four criteria were based on priorities of individual centrality measure. The fifth criterionwas a uniform weightage amongst all centralities. The sixth to eleventh criteria were based on two centralities having the utmost importance and the other two having less importance. The remaining criteria were assigned randomly with varying importance to the centralities.

Table. 2 List of criteria used

Criteria		B	D	Cl	C	Sum
Closeness priority	Criteria 1	0	0	0	1	1
Clustering priority	Criteria 2	0	0	1	0	1
Degree priority	Criteria 3	0	1	0	0	1
Betweenness Priority	Criteria 4	1	0	0	0	1
Equal weight	Criteria 5	0.25	0.25	0.25	0.25	1
Betweenness and degree priority	Criteria 6	0.4	0.4	0.1	0.1	1

Clustering and degree priority	Criteria 7	0.1	0.4	0.4	0.1	1
Closeness and degree priority	Criteria 8	0.1	0.4	0.1	0.4	1
Closeness and betweenness priority	Criteria 9	0.4	0.1	0.1	0.4	1
Clustering and betweenness priority	Criteria 10	0.4	0.1	0.4	0.1	1
Clustering and closeness priority	Criteria 11	0.1	0.1	0.4	0.4	1
Random	Criteria 12	0.25	0.3	0.3	0.15	1
Random	Criteria 13	0.2	0.05	0.25	0.5	1
Random	Criteria 14	0.05	0.05	0.5	0.4	1
Random	Criteria 15	0.2	0.1	0.1	0.6	1
Random	Criteria 16	0.7	0.2	0.05	0.05	1
Random	Criteria 17	0.15	0.5	0.3	0.05	1
Random	Criteria 18	0.25	0.35	0.1	0.3	1
Random	Criteria 19	0.5	0.05	0.3	0.15	1
Random	Criteria 20	0.15	0.7	0.05	0.1	1

After setting the criteria for simulation, the seed set size was determined. To achieve this, as stated in the algorithm subsection, common influencers to all the centralities except the clustering centralities were chosen. The clustering coefficient was not used because the measurement check for overlaps in the node neighbours. Thus, nodes with little neighbours that have low centralities might rank high and still wield no considerable influence on diffusion process. Four different seed set size were used; 10, 50, 100 and 200; which allowed for the analysis of information to spread on different sizes.

Original seed node features

The influencers were identified based on their original node features which are the degree, betweenness, closeness centralities, and clustering coefficient., it was seen that the prominent nodes in the betweenness, closeness and degree centrality for Wiki-vote dataset at influencer seed set size 10 were featured heavily. It will be expected that a good number of the nodes will be featured in the top ten centralities. Only several nodes from there were chosen with the nodes heavily ranked in the top fifty of the centralities. Nodes such as 58, 59, and 2072 which ranked below the top fifty of individual centrality measure were equally included. In evaluating the nodes selected in the absence of clustering coefficient, three nodes (699, 286, 902) were identified to be common amongst degree, betweenness and closeness centrality. On the inclusion of clustering coefficient, there was no common node in the top ten influencers.

Similar findings were reported for other datasets where there were only three common influencers for Epinions dataset excluding clustering coefficient, and on inclusion, no similarities were found. For the Slashdot dataset, only one influencer was common across the three primary centralities and none after considering clustering coefficient. Based on the findings, it is observed that there is no logical continuum in the nodes ranking. Thus, selecting nodes based a single metric might not be optimal in the scenario where the influencers need to be selected based on the kind of message or the target audience. While it cannot be ruled out that some metrics are more effectual than others, they are still insufficient in specific purposes.

For the seed set size at 50, in the Wiki-Vote dataset, it was observed that some nodes selected were outside the top 200 based on the important metrics. In addition, some important nodes such as node 7 (10^{th} in the betweenness centrality) was excluded. 24 out of the 50 seed influencers, excluding criteria 2 were all ranked within the range of the top 50 important metrics yet it is still not effective. In layman terms, for every 50 influencers selected, 48% can only be relied to perform the duty assigned at a particular point in time for particular information. When excluding clustering coefficient, only 5% of the influencers can perform the required information diffusion. This gives strong support to the central theses argued on. For the Epinions dataset, 14.9% of influencers excluding the clustering co-efficient are effective in diffusing information at every time and 0.2% when clustering coefficient was considered. It was found that approximately 60% of its selected influencers have a major centrality ranked below 100. Nodes such as 277 and 2509 were included due to their closeness and clustering centrality which are where the GAM model derives its strength from. Slashdot result was similar to Epinions, with reliable influencers been 0% on the inclusion of clustering coefficient from 5.6% when it was excluded. Approximately half of the selected influencers have a major centrality ranked below 100. Inclusion of nodes like 220 and 502 was due to the weight assignment that gives all nodes chances to be influencers.

On increasing the seed set to 100 and 200, there was not much of an improvement except that new weaker node made up of the list of influencers with a significant percentage of the nodes not in the top 300 nodes of any of the metrics that were used to measure. Nodes with metrics that were even higher than 600 appeared in the result, i.e., a majority of the nodes with high metrics cannot always be relied on when diffusing information. In the Wiki-vote, for every 200 influencers selected excluding the clustering coefficient, only 47.5% is useful at any given time and 30% effective at every run. The central thesis is still valid here as no marketer will want only 3 out of 10 of his influencers to be able to get the message across at every point in time. However, when the influencers are selected based on mixed criteria, they might be able to reach out to those needed even if they are smaller.

Similar results were found with the Epinions and Slashdot dataset too where the influencer reliability at any point in time excluding criteria 2 was 16.2% and 7.1% respectively, with a lot of weak nodes making up the influencer list.

The influencer results describe that if a marketer aims to reach out to its market based on specific needs, it is insufficient going by a singular metric where he might need to spend more and reach only the set of nodes around him whereas if aggregated metrics were applied and used, he might be able to reach more for less cost which will be explained later.

Re-share feature

The reshare feature as identified in the methodology section was discussed regarding transmissibility λ, fan out co-efficient of seed nodes (FSC), Fan out co-efficient of viral nodes (FVC), and the basic reproductive numberR_0, and most importantly the DCF. A fixed set size of 50 was considered as it is frequently used by existing studies [e.g. 50].

*1) Transmissibility(λ):*Transmissibility is a measure of the message content regarding how accepted it is by the overall population in the network. It was found that the transmissibility of the message does not affect the diffusion process because it was assumed that the message shared was the same across the various criteria evaluated in the study. It was found that message transmissibility increased with the influencer seed size as seen in Figure 1; both at the beginning of the diffusion process and at the end of the diffusion process. Furthermore, the transmissibility of the message was dependent on the extent of spread of the information. This was based on the number of connections the influencer nodes have. Fig 1 was based on the Wiki-Vote dataset, but the remaining datasets exhibited similar behaviour and was not reported individually.

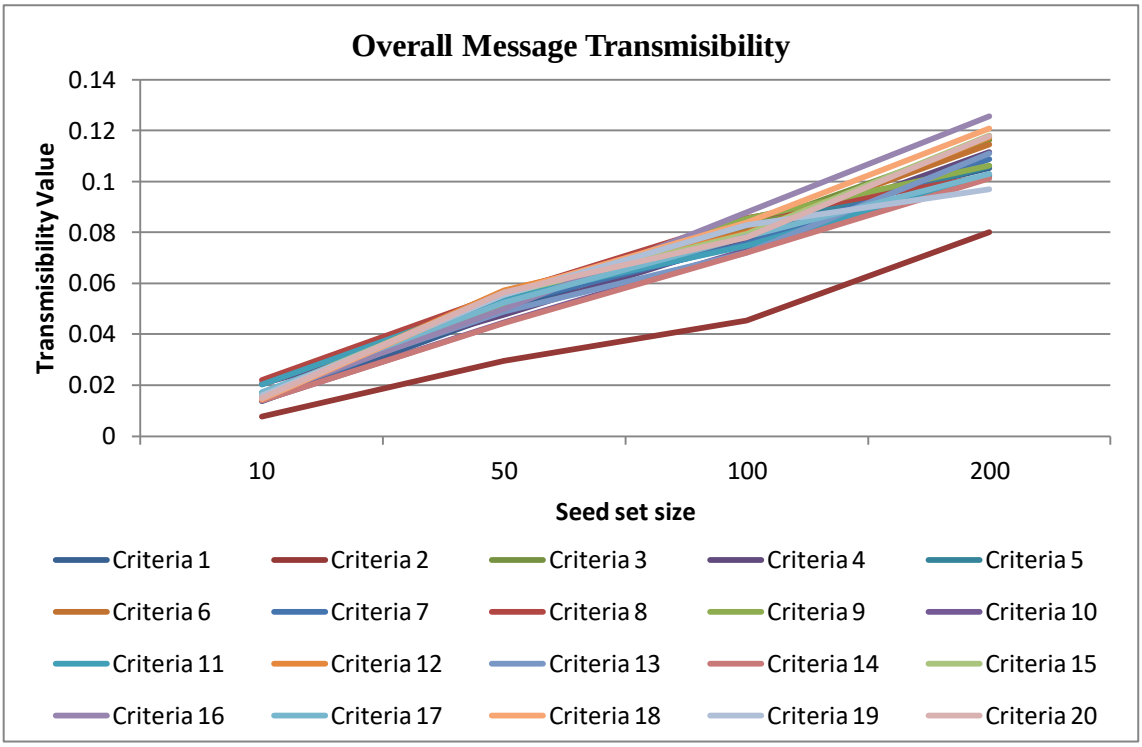

Fig. 1 Overall message transmissibility for Wiki-Vote dataset

2) Fan-out co-efficient of seed nodes (FSC) and fan-out co-efficient of viral nodes (FVC): The Fan-out co-efficient of the seed nodes (FSC) was a factor of the number of activated nodes (the sum of both informed and viral nodes) that the seed node was able to accomplish at the first hop of activation. The criterion that has the highest reach at the first step of activation was based on WIC. Obviously, nodes with high degree and betweenness centrality were able to activate more nodes and have a large FSC. What was surprising was that having a high FSC does not necessarily translate to a large fan-out co-efficient of viral nodes (FVC) as evidenced by the analysis. Intuitively, it was expected that large nodes should know more people and ought to have the effect throughout the diffusion process. Thus, its activated viral nodes ought to lead to more activated nodes as the information diffuses. As the result suggests, this was based on a false premise since redundancy in the network exists during diffusion, i.e. same information is spread amongst the same set of people which might be due to

1. The network hassmall circles cliques e.g. Facebook dataset
2. The network has large circles but disjointed circle cliques e.g. twitter dataset.
3. The network has beenheavy-tailed as it exhibits scale-free network characteristics e.g. Slashdot and Epinions datasets

It was observed that the result was more pronounced in the scale-free networks than small world networks (e.g. Facebook and Twitter). In small world networks, FVC was lower than its scale free because a majority of the nodes in the network are connected in small hops.

This leads to a lower average path length and allowfor a more uniform activation around the network. Despite that, Figure 2 shows that FVC peaked on criterion2 (clustering coefficient) and was the lowest at criterion4(betweenness centrality). This can be attributed to the number of viral nodes that the criteria produced on activation from the seed nodes as the influencers (seed nodes) have lower clustering coefficient.

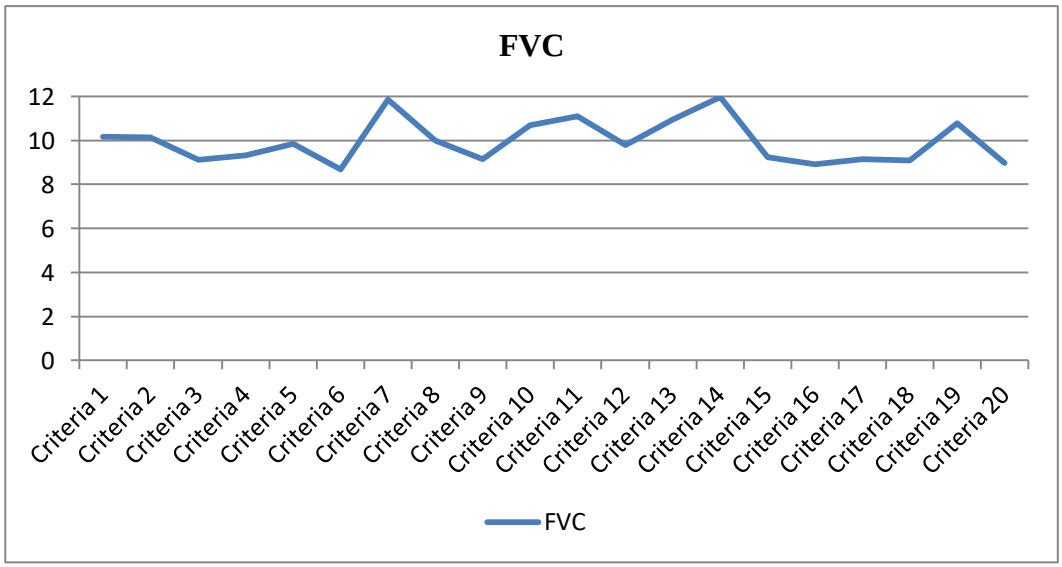

Fig. 2 FVC for Twitter Dataset at seed set size 50

For FSC, it will be higher as it is a factor of the number of nodes that the seed set (influencers) was able to activate. This relies heavily on the centrality of the seed nodes (betweenness centrality) and the number of nodes connected to the seed nodes (degree centrality). This leads to a higher FSC for criteria with high degree and betweenness centrality and lower clustering and closeness centrality as depicted in Fig 3.

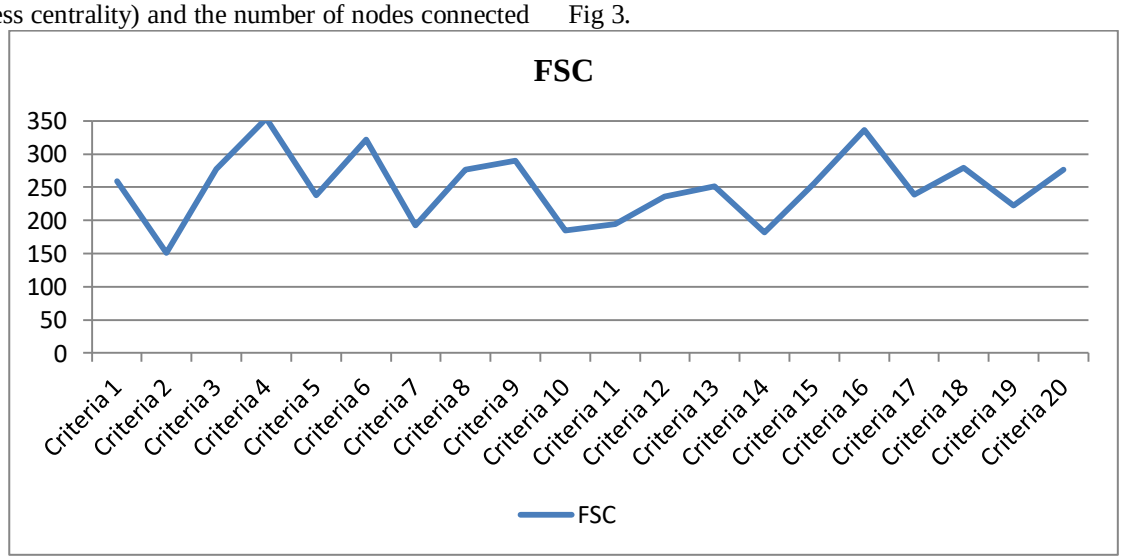

Fig. 3 FSC for Twitter Dataset at seed set size 50

In scale-free networks like Slashdot and Epinions, the difference between the FSC and FVC was more pronounced. This was because the centrality distribution on the network is heavily tailed. The network is not connected as seen in the small world networks and allows for the clustering coefficient to be more pronounced as seen in Table 1. This explains for the steep edges around the criteria with clustering given high importance as shown in Fig 4 and 5. In addition, the same explanation of redundancy, centrality and popularity applies toboth small world and scale-free networks.

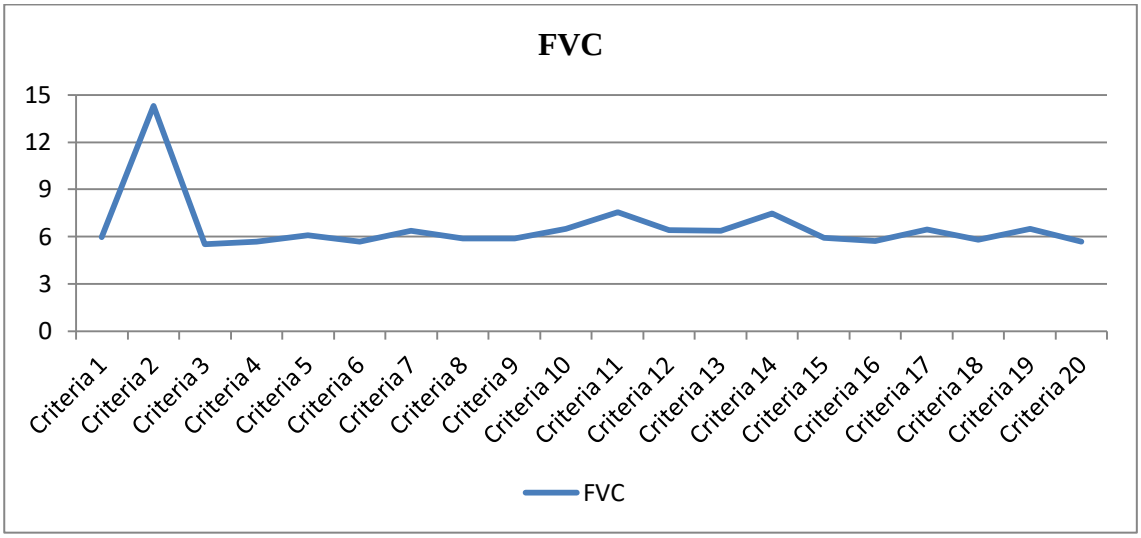

Fig. 4 FVC for Slashdot Dataset at seed set size 50

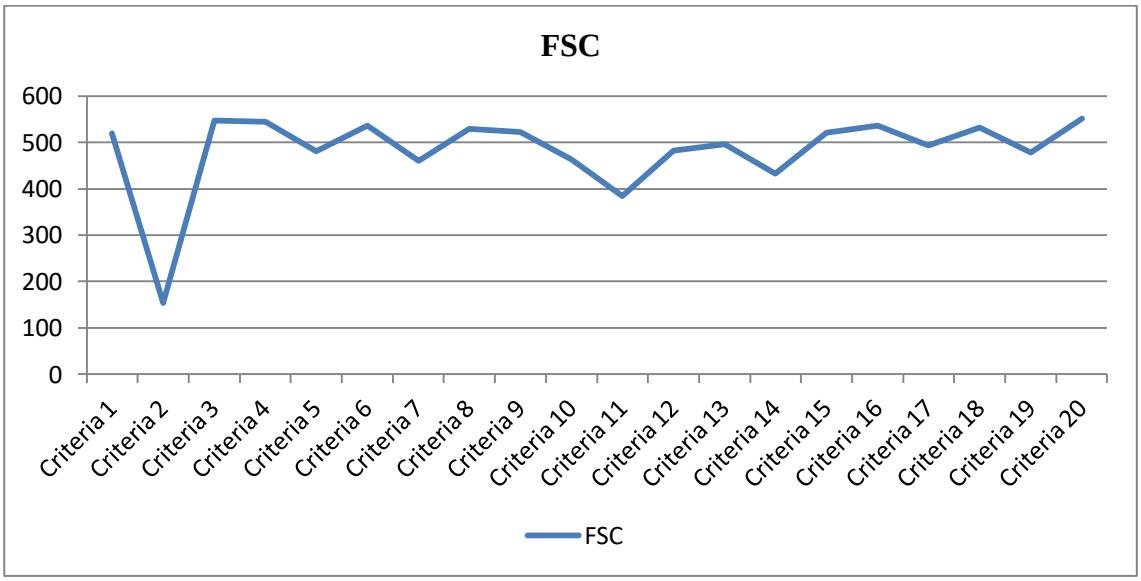

Fig. 5 FSC for Slashdot Dataset at seed set size 50

2) Basic reproductive number (R_0): The basic reproductive number refers to the average number of secondary recommendations produced by reached nodes and was equally studied and found to plot identical graph with the criteria outreach and FSC which was constant across all dataset studied. The reproductive number and outreach werenot surprising as they are both factors of number of activated nodes which spanned through the diffusion network. With respect to FSC which was evaluated at the first hop, it was still similar.

Diffusion Cost Function (DCF)

In explaining DCF, the Wiki-Vote dataset consists of election voters which means that the most influential are those with more votes. It could be observed that the degree and betweenness centralities will be more important. These centralities have higher FSC and outreach (viral nodes + informed nodes) that led to high DCF as compared to the clustering centrality. Looking at the overall diffusion process in Table 3, it was seen that clustering centrality (criteria 2) had higher FVC and spreads to more unique nodes than other centralities as explained above. While it has a low number of outreach (informed nodes and viral nodes), its DCF is relatively low compared to other centralities. Surprisingly, the reproductive number which is a function of the outreach was within range (0.27-0.31), i.e. while it does not have a high FSC, its FVC covered up for the potential weakness of immediate outreach.

To further explain the differences, if a marketer aims to reach a large number of people at the first step, degree centrality (criteria 3) is ideal due to the large FSC, but when aiming at information been diffused beyond the first step, a high FSC with very low FVC will be inappropriate such as criterion 8 and criterion 9 due to the sharp drop of the diffusion cost function (higher than 60%).

In layman terms, if a marketer is aiming to sell a piece of software to the market, going for highly popular people (criterion 4, criterion 3, criterion 8, and criterion 20) might not be ideal as specific software are used by a niche of the population, where it has to go through referrals and word of mouth (WOM) person-to-person recommendation is equal to steps or hops. If the marketer goes for a measure or criteria dominated by major centralities,

a high diffusion cost function will incur with a significant drop throughout the diffusion process (i.e. majority of people will hear but will not spread the information).Thus, the marketer will be at loss. But, in the case of FMCG (fast moving consumer goods),the use of popular metrics might be advisable, but still leads to redundancy of information in the network. Thus, a trade-off calculated balance is preferred. Therefore, the use of the most popular or most connected influencer does not necessarily payoff when considering both DCF and outreach.

The same pattern was found for other datasets where the diffusion cost function of the information sent was very high and the difference between the outreach was relatively low.

This is due to redundancy of information in the network. In the case of high centralities, the information does not move beyond the first set of informers. The information diffusion cost grows rapidly with the influencer size, but surprisingly, the active nodes seem to stabilize across the various simulations regardless of the criteria. This shows that while lower influencers might have a smaller outreach initially, but as their number grow, they have a larger outreach at a similar cost for a smaller set of popular influencers. Another interesting finding is that the dataset used by the research is of various nature. Overall, the findings are consistent baring minor differences which are due to the network type and the individual centrality distribution across the network.

Table. 3 Overall diffusion process for Wiki-Vote dataset

	λ	FSC	FVC	R_0	Outreach	length of viral nodes	length of informed nodes	Initial DCF	Overall DCF	% drop
Criterion 1	0.049	41.04	5.983	0.290	2052	343	2031	95.747	28.935	69.78
Criterion 2	0.029	38.38	9.226	0.272	1919	208	1899	22.997	15.346	33.27
Criterion 3	0.055	43.06	5.535	0.305	2153	389	2130	93.980	36.125	61.56
Criterion 4	0.048	42.24	6.286	0.299	2112	336	2087	84.115	30.026	64.3
Criteria 5	0.053	42.36	5.648	0.300	2118	375	2097	81.363	33.702	58.58
Criterion 6	0.057	43.02	5.337	0.304	2151	403	2113	112.196	37.356	66.7
Criterion 7	0.052	42.9	5.861	0.304	2145	366	2124	95.877	33.737	64.81
Criterion 8	0.056	41.88	5.301	0.296	2094	395	2062	94.537	34.700	63.3
Criterion 9	0.052	41.62	5.670	0.295	2081	367	2055	87.988	31.841	63.81
Criterion 10	0.045	42.52	6.749	0.301	2126	315	2095	72.726	28.524	60.78
Criterion 11	0.054	41.76	5.452	0.296	2088	383	2058	102.286	33.453	67.29
Criterion 12	0.057	43	5.322	0.304	2150	404	2126	109.516	37.414	65.84
Criterion 13	0.049	41.18	5.951	0.291	2059	346	2019	86.290	29.388	65.94
Criterion 14	0.044	40.92	6.516	0.290	2046	314	2015	56.879	26.334	53.7
Criterion 15	0.056	41.52	5.269	0.294	2076	394	2044	100.397	34.019	66.12

Criterion 16	0.050	40.88	5.790	0.289	2044	353	2020	79.563	29.547	62.86
Criterion 17	0.052	43.18	5.835	0.306	2159	370	2121	104.980	34.553	67.09
Criterion 18	0.056	41.32	5.217	0.292	2066	396	2028	105.333	33.863	67.85
Criterion 19	0.055	43.06	5.506	0.305	2153	391	2118	90.417	36.311	59.84
Criterion 20	0.056	43.06	5.410	0.305	2153	398	2126	89.829	36.961	58.85

For instance, Twitter data in Fig 6 show a sharp rise and decline in terms of the diffusion cost. This is because of the small world nature of the network consisting of several egocentric networks and circles. Hence, there will be a high betweenness centrality as specific users will be central to information pathways in each network circle. Thus, information diffusion cost function will be very high for information that goes through selected influencers by betweenness centrality, but the information might not spread beyond their social circle.

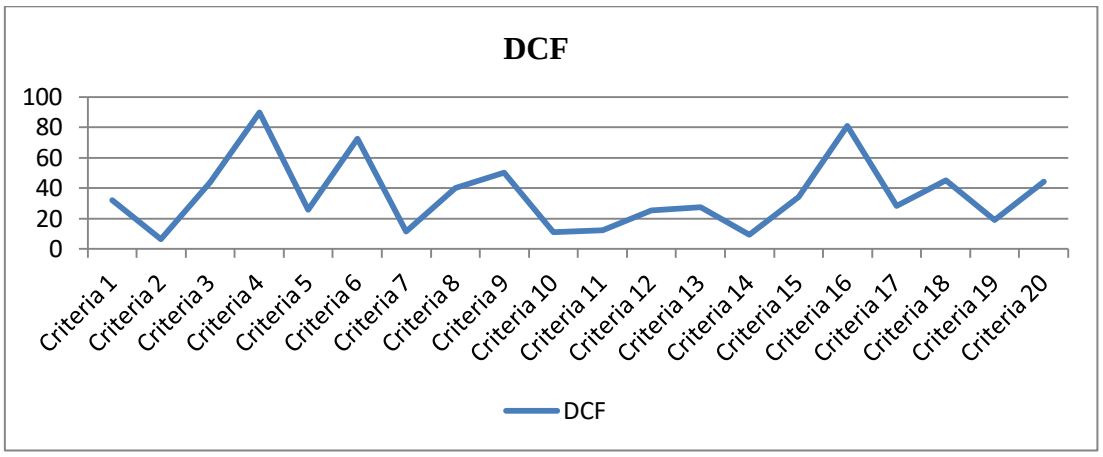

Fig. 6 DCF for Twitter Dataset at seed set size 50

In Fig 7, the sharpness was relatively less steep. In elucidating the differences, criterion13 which places more importance on the diversity (non-redundancy) and the shortest path distance of the influencer in the network had an outreach of 12557 with a DCF of 27.14 as compared to criterion 16 with an outreach of 16835 and DCF of 81, the difference in the outreach is 25.4%, but in the DCF, it was a huge 66.5%. The DCF is more than double, but that does not translate into double outreach which is a loss of information cost. Moreover, criterion 16 information will most likely be redundant amongst its network which leads to high DCF. This is due to the small world network nature of the graph where people know each other in a few hops which is relatively stable. Thus, going for a mix of popularity, centrality and diversity aregoing to give a similar effect at a lower cost than going for the popular or central strictly that leads to high DCF. These results stress that selecting the most popular influencers does not necessarily translate to the most efficient in spreading information.

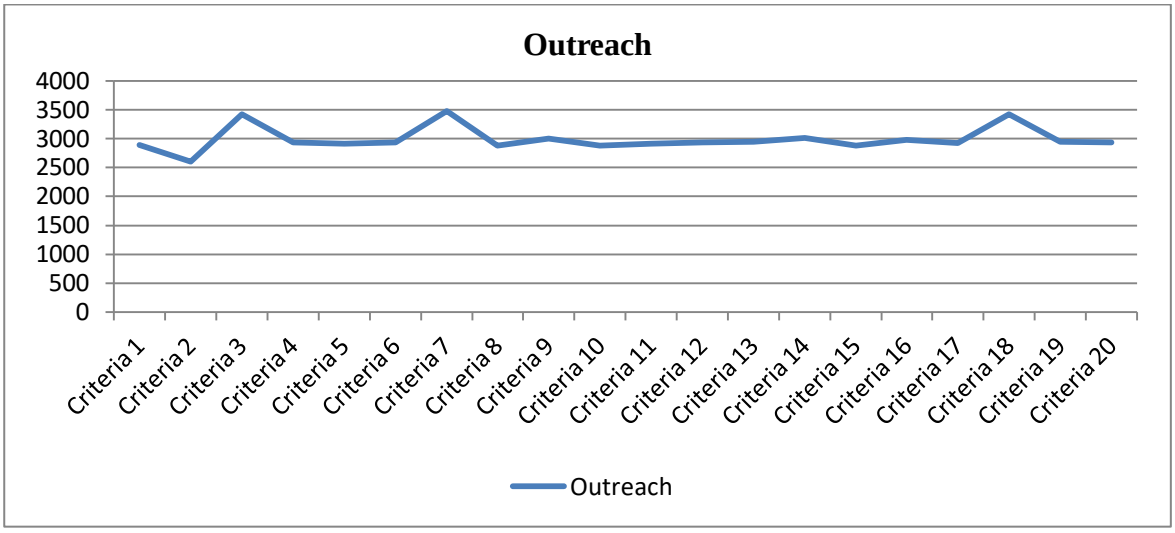

Fig. 7 Outreach for Twitter at seed set size 50

Epinions and Slashdot dataset both exhibited scale-free network properties. It was observed that there was a sharp steepness in their DCF and outreach as shown in Fig 8 and 9 for the Epinions dataset. In the Epinions dataset, which is heavily tailed on betweenness centrality,a comparison between criterion 6 (Central and popular influencers) and criterion11 (Shortest distance and distinct influencers) reveals a 68.4% higher diffusion cost and 25.5% largeroutreach. This shows that having central and popular influencers leads to a high redundancy in the network which increases the DCF with lesser outreach. Similar findings were reported with the Slashdot dataset which is heavily tailed on degree centrality.

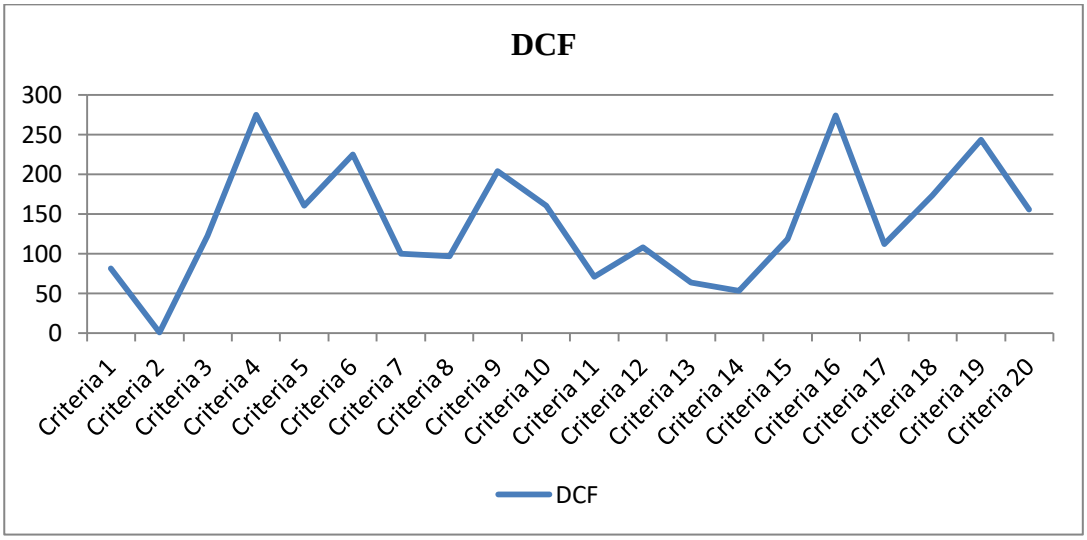

Fig. 8 DCF for Epinions at seed set size 50

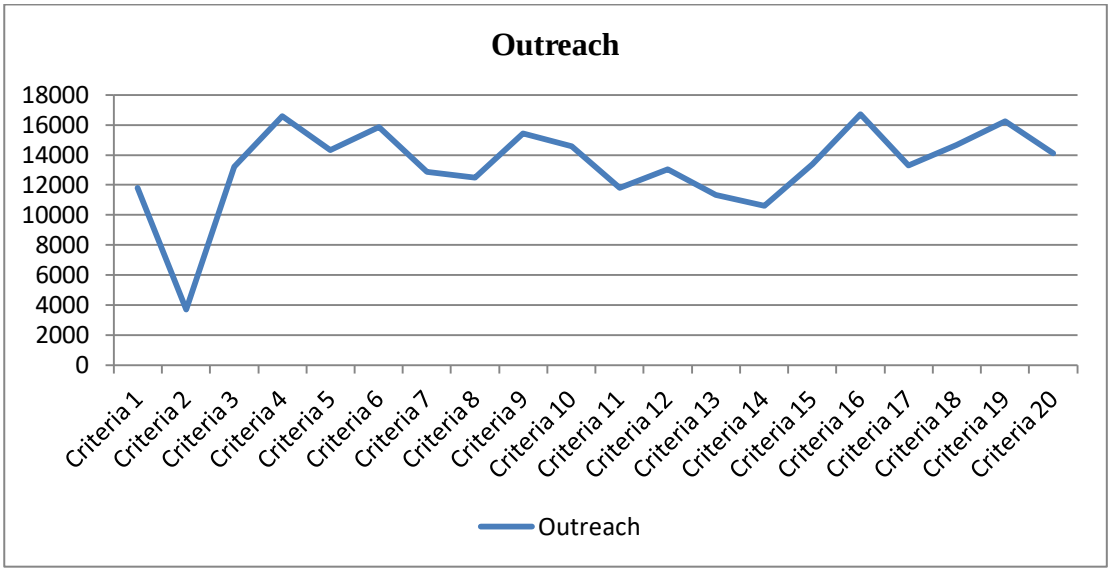

Fig. 9 Outreach for Epinions at seed set size 50

In addition, an increase in the seed set size leads to an increase in the outreach and DCF while the FSC and FVC decreases. This is due to the drop in the overall activation ratio of the influencers i.e. not all the selected influencers activate their neighbours. While this was noticed, the effect it had on the networks differs. For small world networks such asTwitter and Facebook, it was noticed that as the influencer size increases, so do the outreach. Using a seed set size of 200 for the Twitter data, the difference between the DCF and outreach reduced which led to a relative stable DCF percentage drop as compared to the seed set size at 50 as seen in Figure 10. In the scale-free networks which include Epinions and Slashdot, the difference as the seed set increased was minimal.Thus, there was no significant difference in the DCF percentage drop as shown in Figure 11.

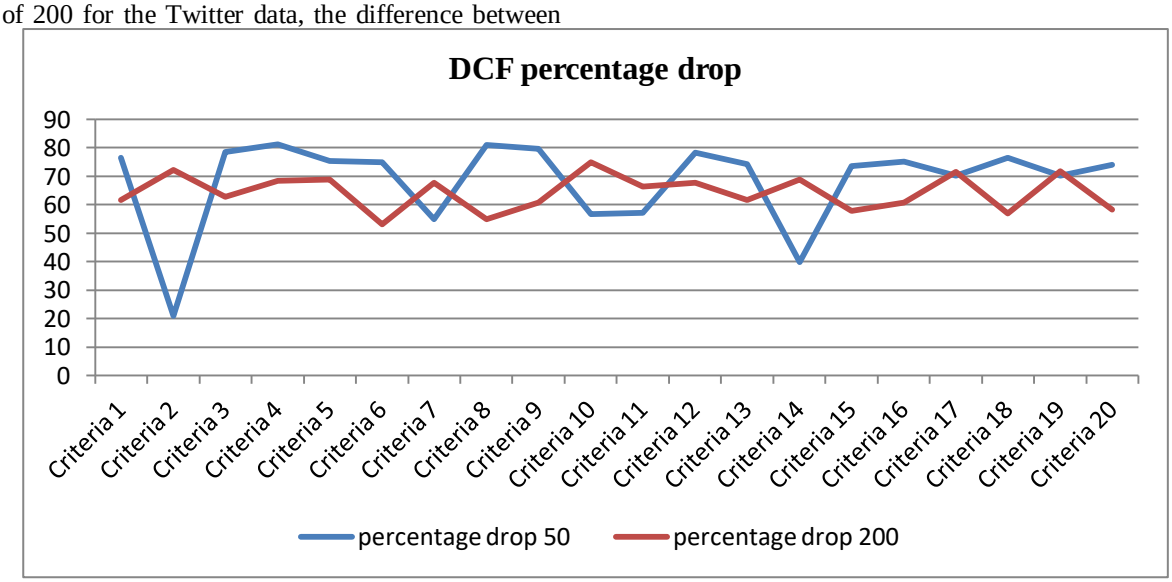

Fig. 10. Percentage drop of information diffusion cost function for Twitter between seed set size 50 and 200

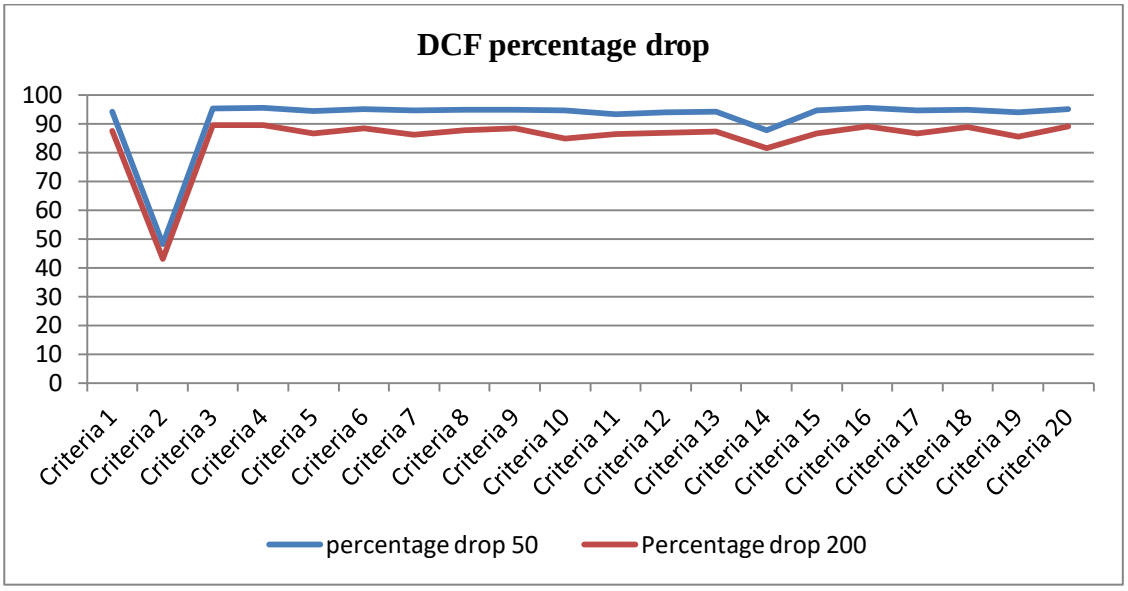

Fig. 11. Percentage drop of information diffusion cost function for Slashdot between seed set size 50 and 200

VI. IMPLICATION

The central argument of this paper is that influencer selection based on traditional approaches (heuristic algorithms) will not always suffice in selecting the best influencers for specific needs. The findings of this paperposit that the combination of various metrics can assist in making a better influencer selection. This study contributed to the body of literature in three ways. Firstly, the study strengthens the need to approach influencer selection in a more diverse way by considering several centralities based on the need of the information to be diffused. This responds to calls on improving influencer selection based on the purpose of diffusing information. This study showed that using the primary centralities individually does not necessarily translate to having the best influencers for the particular need. This was shown by the absence of top influencers in the seed set selection as illustrated in the results. Moreover, the highest outreach based in the FSC or FVC is not necessarily achieved using singular centrality metrics. This advanced our knowledge of selecting influencers which is vital in viral marketing.

Secondly, the role of DCF was taken into consideration. This responds to the call of [5] in influencer selection. DCF made use of where it was seen that of the existence of a trade-off between the outreach of the information and the DCF incurred. It was seen that both degree and betweenness centralities led to high DCF which does not automatically translate to a substantial better outreach at the end of the diffusion process. This is vital to small business owners that engage influencers to promote their products based on social networks where it is essential to consider the type of network (small world or scale-free), the reason for selecting influencers,and balancing the influencer selection.

Thirdly, it is showed that an increase in influencer size is not advisable in all cases. While the work of [16] argued that an increase in seed size allows for better information diffusion, our findings differ. It is noted that an increase in influencers on a scale-free network does not necessarily translate to an increase in outreach or a reduction in DCF. Meanwhile, in a small world network, the investment might eventually payoff as the seed set size increases. Thus, organizations aiming to use influencers as an advertisement strategy can afford to increase their influencer's budget in recruiting more on networks that showsthe small world properties such associal media while it is pointless on scale-free networks. In addition, this study extended the previous work by improving the network structural properties [14] and node metrics [36] in predicting activation at each hop of the diffusion spread [38]. This allowed for weak nodes to be selected as possible influencers.

VII. CONCLUSION

The era of social media growth has witnessed the rapid growth in the number of users on the different social platforms. Hence, there is a crucial need for influencer identification based on the requirement. This study aims at helping marketers in making informed decisions where they can select influencers based on establish properties and establish a trade-off with the cost associated.

Future research is recommended to look into extending the work to target nodes where influencer selection can be estimated to only a set of nodes in the network. This is important as it would allow for streamlining of advertisement on social networks.

REFERENCES

1. M. Lister. (2018). 40 Essential Social Media Marketing Statistics for 2018. Available: https://www.wordstream.com/blog/ws/2017/01/05/social-media-marketing-statistics
2. A. Guille, H. Hacid, and C. Favre, "Predicting the temporal dynamics of information diffusion in social networks," arXiv preprint arXiv:1302.5235, 2013.
3. Q. Wang, Y. Jin, S. Cheng, and T. Yang, "ConformRank: A conformity-based rank for finding top-k influential users," Physica A: Statistical Mechanics and its Applications, vol. 474, pp. 39-48, 2017.

4. Q. Wang, Y. Jin, Z. Lin, S. Cheng, and T. Yang, "Influence maximization in social networks under an independent cascade-based model," Physica A: Statistical Mechanics and its Applications, vol. 444, pp. 20-34, 2016.
5. L. Alsuwaidan, "Toward Information Diffusion Model for Viral Marketing in Business," International journal of advanced computer science and applications, vol. 7, 2016.
6. K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, and S. Nerur, "Advances in social media research: past, present and future," Information Systems Frontiers, vol. 20, no. 3, pp. 531-558, 2018.
7. A. Sela, D. Goldenberg, I. Ben-Gal, and E. Shmueli, "Active viral marketing: Incorporating continuous active seeding efforts into the diffusion model," Expert Systems with Applications, vol. 107, pp. 45-60, 2018.
8. F. Morone and H. a. Makse, "Influence maximization in complex networks through optimal percolation:supplementary information," Current Science, vol. 93, pp. 17-19, 2015.
9. A. S. T. Olanrewaju and R. Ahmad, "Examining the information dissemination process on social media during the Malaysia 2014 floods using Social Network Analysis (SNA)," Journal of Information and Communication Technology, vol. 17, pp. 141-166, 2018.
10. S. Peng, Y. Zhou, L. Cao, S. Yu, J. Niu, and W. Jia, "Influence analysis in social networks: a survey," Journal of Network and Computer Applications, 2018.
11. S. M. H. Bamakan, I. Nurgaliev, and Q. Qu, "Opinion leader detection: A methodological review," Expert Systems with Applications, 2018.
12. W. Chen, L. V. S. Lakshmanan, and C. Castillo, "Information and Influence Propagation in in Social Networks," Synthesis Lectures on Data Management, 2013.
13. 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, ed, 2014, pp. 171-180.
14. J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?," in Proceedings of the 23rd international conference on World wide web, ed, 2014, pp. 925-936.
15. M. Gladwell, The tipping point: How little things can make a big difference. Little Brown, 2006.
16. A. Mochalova and A. Nanopoulos, "On The Role Of Centrality In Information Diffusion In Social Networks," in ECIS, 2013, p. 101.
17. Y. Li, M. Qian, D. Jin, P. Hui, and A. V. Vasilakos, "Revealing the efficiency of information diffusion in online social networks of microblog," Information Sciences, vol. 293, pp. 383-389, 2015.
18. L. Weng, F. Menczer, and Y.-Y. Ahn, "Virality Prediction and Community Structure in Social Networks," Scientific Reports, vol. 3, p. 2522, 2013.
19. S. Pei, L. Muchnik, S. Tang, Z. Zheng, and H. A. Makse, "Exploring the complex pattern of information spreading in online blog communities.," PloS one, vol. 10, p. e0126894, 2015.
20. S. Pei, L. Muchnik, J. Andrade, S. José, Z. Zheng, and H. A. Makse, "Searching for superspreaders of information in real-world social media," Scientific Reports, vol. 4, pp. 1-12, 2014.
21. S. Pei and H. A. Makse, "Spreading dynamics in complex networks," Journal of Statistical Mechanics: Theory and Experiment, vol. 2013, p. P12002, 2013.
22. X. Chen, G. Song, X. He, and K. Xie, "On Influential Nodes Tracking in Dynamic Social Networks," arXiv preprint arXiv:1602.04490, pp. 613-621, 2015.
23. J. Goldenberg, B. Libai, and E. Muller, "Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata," Academy of Marketing Science Review, vol. 9, pp. 1-18, 2001.
24. 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, ed, 2003, pp. 137-146.
25. B. Chang, T. Xu, Q. Liu, and E.-H. Chen, "Study on information diffusion analysis in social networks and its applications," International Journal of Automation and Computing, pp. 1-26, 2018.
26. M. Heidari, M. Asadpour, and H. Faili, "SMG: Fast scalable greedy algorithm for influence maximization in social networks," Physica A: Statistical Mechanics and its Applications, vol. 420, pp. 124-133, 2015.
27. Y. Li, J. Fan, Y. Wang, and K.-L. Tan, "Influence maximization on social graphs: A survey," IEEE Transactions on Knowledge and Data Engineering, 2018.
28. M. Hosseini-Pozveh, K. Zamanifar, and A. R. Naghsh-Nilchi, "A community-based approach to identify the most influential nodes in social networks," Journal of Information Science, vol. 43, no. 2, pp. 204-220, 2017.
29. Y. Feng, B. Bai, and W. Chen, "Information diffusion efficiency in online social networks," in IEEE International Conference on Digital Signal Processing (DSP), 2015, ed, 2015, pp. 1138-1142.
30. F. Stonedahl, W. Rand, and U. Wilensky, "Evolving viral marketing strategies," Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation, pp. 1195-1202, 2010.
31. N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Modeling blogger influence in a community," Social Network Analysis and Mining, vol. 2, pp. 139-162, 2012.
32. Y. Du, C. Gao, Y. Hu, S. Mahadevan, and Y. Deng, "A new method of identifying influential nodes in complex networks based on TOPSIS," Physica A: Statistical Mechanics and its Applications, vol. 399, pp. 57-69, 2014.
33. J. Hu, Y. Du, H. Mo, D. Wei, and Y. Deng, "A modified weighted TOPSIS to identify influential nodes in complex networks," Physica A: Statistical Mechanics and its Applications, vol. 444, pp. 73-85, 2016.
34. T. Bian, J. Hu, and Y. Deng, "Identifying influential nodes in complex networks based on AHP," Physica A: Statistical Mechanics and its Applications, vol. 479, pp. 422-436, 2017.
35. Q. Ma and J. Ma, "Identifying and ranking influential spreaders in complex networks with consideration of spreading probability," Physica A: Statistical Mechanics and its Applications, vol. 465, pp. 312-330, 2017.
36. J. L. Iribarren and E. Moro, "Affinity Paths and Information Diffusion in Social Networks," Social Networks, vol. 33, pp. 134-142, 2011.
37. A. Kupavskii, A. Umnov, G. Gusev, and P. Serdyukov, "Predicting the Audience Size of a Tweet.," in ICWSM 2013 International Conference on Weblogs and Social Media, ed, 2013, pp. 693-696.
38. Y. Lim, A. Ozdaglar, and A. Teytelboym, "A simple model of cascades in networks.," Technical report, LIDS, 2015. 6. 2, 7.1., 2015.
39. A. Pérez-Foguet, R. Giné-Garriga, and M. I. Ortego, "Compositional data for global monitoring: The case of drinking water and sanitation," Science of the total environment, vol. 590, pp. 554-565, 2017.
40. M. Velasquez and P. T. Hester, "An analysis of multi-criteria decision making methods," International Journal of Operations Research, vol. 10, no. 2, pp. 56-66, 2013.
41. G.-H. Tzeng and J.-J. Huang, Multiple attribute decision making: methods and applications. Chapman and Hall/CRC, 2011.
42. R. A. Irizarry, "Additive Models , GAM , and Neural Networks," in Statistical Learning: Algorithmic and Nonparametric Approaches, ed: John Hopkins University, 2006, pp. 210-245.
43. L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, "Vital nodes identification in complex networks," Physics Reports, vol. 650, pp. 1-63, 2016.
44. D. J. Watts and P. S. Dodds, "Influentials, networks, and public opinion formation.," Journal of consumer research, vol. 34, pp. 441-458, 2007.
45. J. R. Lee and C. W. Chung, "A fast approximation for influence maximization in large social networks," in Proceedings of the companion publication of the 23rd international conference on World wide web companion, ed: International World Wide Web Conferences Steering Committee, 2014, pp. 1157-1162.

46. H. Zhang, D. T. Nguyen, H. Zhang, and M. T. Thai, "Least Cost Influence Maximization Across Multiple Social Networks," IEEE/ACM Transactions on Networking, pp. 1-11, 2015.
47. Z. Wang, E. Chen, Q. Liu, Y. Yang, Y. Ge, and B. Chang, "Maximizing the Coverage of Information Propagation in Social Networks," presented at the International Joint Conference on Artificial Intelligence, 2015.
48. C. T. Li, Y.-J. Lin, and M.-Y. Yeh, "Forecasting participants of information diffusion on social networks with its applications," Information Sciences, vol. 422, pp. 432-446, 2018.
49. J. Leskovec and K. Andrej, "Stanford Large Network Dataset Collection," SNAP Datasets, 2014.
50. J. R. Lee and C. W. Chung, "A query approach for influence maximization on specific users in social networks," IEEE Transactions on Knowledge and Data Engineering, vol. 27, pp. 340-353, 2015.