Source Aware Budgeted Information Maximization

Rithic Kumar N*,
Department of Computer Science,
Indian Institute of Information Technology,
Design & Manufacturing (IIITDM) Kancheepuram
Email: *coe18b044@iiitdm.ac.in

Yayati Gupta†, Sanatan Sukhija‡
École Centrale School of Engineering,
Mahindra University
Hyderabad, India
Email: †yayati.gupta@mahindrauniversity.edu.in,
‡sanatan.sukhija@mahindrauniversity.edu.in

***Abstract*—The paper proposes a more general framework for budgeted influence maximization. We propose a novel cost function that considers the potential seed nodes' properties and the firm interested in maximizing the influence. A greedy algorithm, maximizing the influence to cost ratio, is then used to select a balanced set of seed nodes. We also show that the edge weights play an important role in determining the influential power of nodes. Further, the edge weights for a network can be efficiently predicted with the help of link prediction heuristics like resource allocation metrics and the Adamic-Adar index.**

***Index Terms*—Budgeted influence maximisation, Influence propagation models, Core-Periphery structure, k-shell decomposition, $H-$ index, Independent Cascade model**

I. Introduction

People influence each other by transmitting information/ideas through word of mouth, social media, websites and blogs etc. Such influences lead to product purchases, trend acceptances, and behavior adoption. Starting spreading from one/many people (aka seeds) to their friends, from their friends to their friends' friends and so on; an information/idea quickly reaches many other people. The total number of people reached by the end greatly depends on "who the seeds were?". In digital marketing, the seeds leading to the largest influence cascade are called influencers. Such influencers receive hefty payments from digital marketing firms for products' advertisements. By some estimates, "influencer marketing" has become a \$2 billion venture. The problem of finding influencers (Influence Maximisation abbreviated as IM) is formally addressed with the help of a friendship network, $G(V,E)$ where nodes, $V(G)$, represent people and edges, $E(G)$, represent the friendship connections between them. Thereafter, information spreading models like linear threshold, independent cascade or their approximations [1] are used to determine the spread of influence. Such models compute the function $\sigma : S \rightarrow T$, where S and T are subsets of $V(G)$ ($|T| \geq |S|$). T is the set of people who are finally influenced when the influence propagation starts from the seed set S. The goal of influence maximisation is to find a set S of some given size k which maximises $|\sigma(S)|$. Since the problem (polynomial time reducible to set cover problem) is NP complete, various approximation algorithms, heuristics, meta-heuristics and community based frameworks have been proposed to solve it.

A. Challenges in framing the IM problem

In solving the IM problem, the major challenge is the in general unavailability of any kind of data other than the network $G(V,E)$. These challenges lead to a number of assumptions in the basic framework of the problem. Below we mention the most common assumptions.

1) Knowledge of strength of connections: Because of the unavailability of the real world data, the strength of friendship connections between pairs of people are unknown. In most of the studies, either the edge weights are sampled randomly [2] from the range $[0,1]$ or are assumed to be equal for all edges [3]. This does not reflect the true nature of real world heterogeneous networks. In reality, influence propagation between two people significantly depends on the strength of dyadic tie (friendship) between them [4].

2) Symmetric connections: Even for an undirected friendship network, the strength of a tie need not be symmetric [5], i.e. if A and B are friends, they might influence each other differently owing to the difference in their social status/influence. It has been shown that influential people are, in general, less susceptible [6] and vice versa.

3) Uniformly sampled influence costs for seeds: A generalisation of the traditional Influence Maximisation is the Budgeted Influence Maximisation (BIM) [7] which considers the fact that the seeding cost/effort for each node in the network need not be equal. In this generalisation, the required number of seeds, k, is replaced with a given budget B. Further, a parameter $c(u)$ (cost of persuading/activating the node u) is defined for each $u \in V(G)$. The objective then is to maximise the influence, $|\sigma(S)|$ keeping the total seeding cost, $\Sigma_{u \in V(G)} c(u)$, less than B. Again, because of the unavailability of the real world data, the seeding costs are either randomly sampled [7] or are based on centrality measures like the indegree or pagerank [8] of a node.

4) Fixed seeding costs of nodes irrespective of the interested firm: A fairly new concept in digital marketing is the categorisation of influencers as mega, macro, mid-tier, micro and nano [9]. These categories are ordered in a decreasing order of influential power. The influential power in this categorisation is simply based on the number of followers. However, a number of recent research articles and blogs emphasize the use of advanced network centrality measures rather than simply using the indegree. Higher the influence,

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/\$31.00 © 2022 IEEE

greater the efforts/cost required to persuade them. Since a not so established brand will find it difficult to target a mega influential, mid-tier brands are suggested to target micro and nano influencers [9], [10]. This implies that the influence difference between the brand and the influencer is an important factor in defining the seeding effort. Another recent concept in digital marketing is employing the power of personalisation whereby brands are suggested to build strong long-term relationships with the influencers [11]. This concept demonstrates the importance of social distance (geodesic distance in the network) in persuading an influencer. The closer you are to an influencer in the network, the better. The above reasonings offer the following premise for our work.

"The seeding cost not only depends on who the influencer is, but also on who you are and where you are in the social network."

We propose an IM framework which in addition to the influencers (seeds) also considers the interested firm/brand/marketer (referred as source). Unlike all the previous studies, we consider the source to be a node in the network. It is imperative to assume that the interested firm/marketer is also a part of the network since today most of the firms/brands/marketers have a strong presence on social media.

B. Contributions

The main aim of our work is to propose a more realistic and generalised framework for BIM with no explicit data other than the network. Later, we also lift the requirement of the knowledge of whole network. The major contributions of our work are listed below.

1) We first employ link prediction heuristics in computing bidirectional edge weights for an undirected network. We show that the popular link prediction heuristics like resource allocation metric and Adamic-Adar index perform fairly well in predicting the edge weights. We do not assume the requirement of any past data.
2) As discussed before, the current literature in digital marketing classifies influencers in 5 categories simply based on the number of followers (indegree). We propose the use of coreness centrality to classify influencers in $q(G)$ categories of decreasing influence for a given network G. We show that the edge weights can't be ignored in such generalised categorisation.
3) The most important and final contribution of our work is to define the seeding costs of nodes with respect to the structural position of source in the network. Stated formally; instead of finding one optimal set S which maximises $\sigma(S)$ and minimises $\sum_{v \in S} c(v)$, our aim is to find $S_u : u \in V(G)$ which maximises $\sigma(S_u)$ and minimises $\sum_{v \in S_u} c(u, v)$, where u is the source node and $c(u, v)$ is the seeding cost/effort required by u to persuade v . We propose a geodesic distance and h-index based method to compute $c(u, v)$.

II. Preliminaries

In our work, we use independent cascade model to find the influence function $\sigma(.)$. We use the well established notions of core-periphery structure [12] and k-core decomposition [13] to quantify the influential power of nodes in a network. The 3 terms used above are described below.

A. Independent Cascade Model

In Independent Cascade model, each edge $E(u, v) \in E(G)$ is assigned a value, $0 \leq p(u, v) \leq 1$, which denotes the probability with which v becomes active given that u has turned active. In every iteration, each of the recently activated node tries to activate its neighbours in accordance with the probability associated with the corresponding edge. If a node gets active in iteration t, it gets only iteration $t{+}1$ to activate its neighbours. By this rule, every node gets exactly one chance to infect its neighbours during the entire process. The process stops on reaching an iteration where no new node is infected. An example of simulation of IC Model is shown in Figure 1. To model real life scenarios, we consider the influence probability to be decaying with time (number of iterations in the IC model). The decay factor at time t is defined as

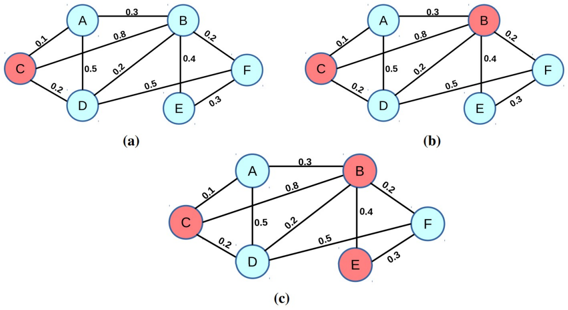


Fig. 1: Simulation of independent cascade model. Blue nodes represent inactive nodes and pink nodes are the active nodes. The fraction against an edge $E(u, v)$ represents its probability of activation.

$$\delta_t = \frac{1}{\sqrt{t+1}}$$

where $t = 0, 1, 2 \ldots$

So the diffusion probability for an edge $E(u, v)$ at time t is $\delta_t \times p[u, v]$.

B. Core-Periphery Structure

Core-Periphery structure is an interesting mesoscale property found in complex networks [12]. There are few nodes in a network which are structurally central, together called the core of the network. The nodes in the core are well connected to each other as well as with the other nodes in the network, called the periphery nodes. It is observed that most of the paths between the periphery nodes pass through the core. Hence, core nodes are responsible for binding the network together. Stephen and Borgatti [12] were the first ones to model this structure. They modeled, what they call, the ideal core-periphery structure. In an ideal core-periphery structure, the nodes in the network are divided in two categories: core and periphery. Each core node is connected to every other node in the network, be it a core or a periphery. Each periphery node is connected only to the core nodes. No two of the periphery nodes are connected to each other. An example of such a structure is shown in Figure 2.

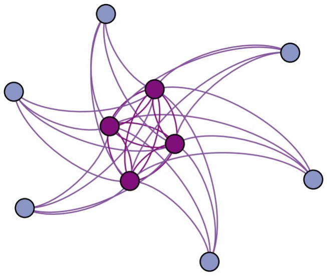


Fig. 2: An Ideal Core-Periphery Structure: The nodes in purple color are core nodes and other are periphery nodes.

C. k-core Decomposition

k-core, G^k of a network is defined as the maximal connected subgraph of G in which each node has degree $\geq k$. A node is said to be present in shell number s iff it is a part of $s-core$, s $core$, $s-2$ $core$,... 1 $core$. k-core decomposition (aka k-shell decomposition) determines the shell number of nodes in a network by recursively pruning the nodes as explained below.

The pruning proceeds in iterations. In iteration i, the nodes having degree $\leq i$ are removed. If the removal of these nodes results in more nodes having degree $\leq i$, then those nodes are also removed in the iteration i. This recursive pruning continues in the iteration i till all nodes in the network have degree $> i$.
The algorithm starts with $i = 1$ and hence continues pruning here till all nodes have degrees > 1. Then i is incremented and pruning continues recursively till all nodes have degrees > 2. The increment in i and pruning continues till the graph becomes empty, i.e., all the nodes in the network are removed. The shell number $shell(u)$ of a vertex u is the iteration number i in which u was removed.

The k-core and shell numbers of nodes in an example network are shown in Figure 3.

D. H-index

We use h^2 [14] index as an approximation of the shell number of a node. Consider a series, $A(x) = (x_1, x_2, \ldots, x_n)$, then we say that $y = H(A(x))$, if y is the maximum integer such that there exist at least y elements in $A(x)$ each of which is no less than y [15]. Assume the degree of a node u be represented as $d(u)$. Let $N(u) = \{v_1, v_2, \ldots, v_{d(u)}\}$, be the set of neighbours of u, having degrees $d(v_1), d(v_2), \ldots, d(v_{d(u)})$, respectively. Then h-index of u using H-operator is given by

$$h(u) = H(d(v_1), d(v_2), \ldots, d(v_{d(u)})) \qquad (1)$$

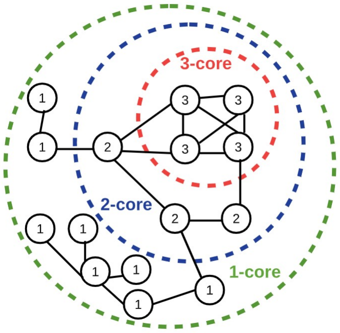


Fig. 3: k-core and k-shell decomposition. The node labels represent their shell numbers.

This process of calculation of H-index of nodes can be defined recursively, leading to a family of H-indices. Let

$$h^1(u) = H(h(v_1), h(v_2), \ldots, h(v_{d(u)})) \qquad (2)$$

Similarly

$$h^k(u) = H(h^{k-1}(v_1), h^{k-1}(v_2), \ldots, h^{k-1}(v_{d(u)})) \qquad (3)$$

It has been shown that when the global information of a network is unavailable, h^2 index can be used as a reliable approximation for the shell number of a node [14].

III. Estimating Bidirectional Edge Probabilities

Notation	Meaning
$G(V, E)$	A graph G where V is the set of vertices and E is the set of edges.
$\Gamma_{out}(u)$	Successors (neighbors of u connected by an outgoing edge from u) of a node $u \in V(G^D)$.
$\Gamma_{in}(u)$	Predecessors (neighbors of u connected by an incoming edge to u) of a node $u \in V(G^D)$.
$CN(u, v)$	$\Gamma_{out}(u) \cap \Gamma_{in}(v)$

TABLE I: Notations used

For our experiments, we consider unweighted, undirected networks. However, each edge is considered as bidirectional so as to differentiate between the direction of infuence spreading. Hence, for a given undirected network, $G(V, E)$, we create a directed network $G^D(V^D, E^D)$ such that $V^D(G^D) = V(G)$ and

$$E^D(G^D) = \bigcup_{(u,v)\in E(G)} \{(u, v), (v, u)\}$$

To predict the edge weights, we consider the below mentioned link prediction heuristics for an edge $E(u, v)$ [16].

1) Inverse of indegree,

$$I(u, v) = |\Gamma_{in}(v)|$$

2) Resource Allocation,

$$RA(u, v) = \sum_{z \epsilon CN(u,v)} \frac{1}{|\Gamma_{out}(z)|}$$

3) Adamic - Adar index,

$$AA(u,v) = \sum_{z \epsilon CN(u,v)} \frac{1}{log(|\Gamma_{out}(z)| + \varepsilon)}$$

where ε is a small number added to avoid the denominator to be zero

4) Leicht-Holme-Newman Index,

$$LHI(u,v) = \frac{|CN(u,v)|}{|\Gamma_{out}(u)| \times |\Gamma_{in}(v)|}$$

5) Jaccard Coefficient,

$$Jaccard(u,v) = \frac{|CN(u,v)|}{|\Gamma_{out}(u) \cup \Gamma_{in}(v)|}$$

To compare the above heuristics with actual edge weights (the ground truth), we first normalise the heuristic values as well as the edge weights from 0 to 1 using min-max normalisation, as mentioned below.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

The above normalisation also translates the edge weights into their equivalent edge diffusion probabilities. Next, we compare the actual and predicted weights with the help of mean absolute error (MAE[1]).

$$MAE = \frac{1}{|E(G^D)|} \sum_{e \in E(G^D)} |w_e - \hat{w}_e|$$

,where w_e and $\hat{w}_e$ respectively represent the normalised ground truth and normalised predicted weight of the edge e. The comparison results are reported in Table II. It is seen that the Adamic-Adar index and resource allocation metric perform the best. Further, it is seen that the link prediction heuristics perform extremely well for the human communication networks (last 3 rows).

IV. A Generalised Categorisation of Influencers

The weight $w(u,v)$ of an edge $E(u,v)$ is a natural representative of influence propagation probability from u to v. Therefore, to observe the impact of edge weights/probabilities on information diffusion and user categorisation, we simulate independent cascade model with

1) edge diffusion probabilities as normalised edge weights predicted from 5 link prediction heuristics.
2) Equal probability $p \in \{0.25, 0.5, 0.75\}$ for all edges

We simulate independent cascade model starting from each node $u \in V(G)$. For each node, the simulation is done x times[2]. For each of these x times, we obtain $|\sigma(\{u\})|$. The influential power of node u,

$$\kappa(u) = \frac{1}{x} \sum_{i=1}^{x} |\sigma(\{u\})|$$

The core (shell) number of a node is shown to be a better predictor of the influential power as compared with other centrality measures. Hence, we categorise the nodes based on their shell number as computed by the k-shell decomposition algorithm. Assume $\lambda(u)$ to represent the shell number of a node u, then we define the influential capacity of a shell s as

$$\kappa(s) = \frac{\sum\limits_{\{u: u \in V(G), \lambda(u) = s\}} \kappa(u)}{|\{u \in V(G), \lambda(u) = s|\}}$$

Since k-shell decomposition requires the knowledge of the entire network, various influence approximation heuristics like h^2 index, sum of degrees of neighbors etc. can also be used. We have used the shell number for higher accuracy.

The shell-wise influential powers for various edge weight metrics are shown in Figure 4. With non-varying edge weights; initially, the influential power increases with an increase in the shell number. However, it stabilises quickly at a shell much before the core (the innermost shell comprising the most influential people). Hence, on ignoring the varying edge weights; even in the large networks, we find only a handful of user influence categories. Constant edge weights lead to an extremely large number of people having as high influential power as the core nodes. Gupta et. al. has referred such nodes as "Pseudo-cores" [17] in their work. Our experiments demonstrate the disappearance of pseudo-cores when the edge weights are considered in the diffusion process. This shows that the consideration of varying edge weights is important in framing the IM problem and properly categorising the users based on their influential power.

V. Proposed Cost Function for BIM

We formulate the cost function as $c(s,S)$, where s is the source and S is the seed set. Further, cost of selecting a single seed node, u, is represented as $c(s,u)$. Hence, $c(s,S) = \sum\limits_{v \in S} c(s,v)$. Our proposed cost function is based on two important metrics in network science.

1) Geodesic Distance: Being closer to a seed node in the network reduces the seeding effort. Hence

$$c(s,v) \propto e^{dist(s,v)} \tag{4}$$

We consider the cost to increase exponentially with the distance because of small average path length in the real-world networks. With the advent of online social networks, the "degrees of separation" has shrunken from 6 (offline social network in 1967) to 3.57 (online social network in 2016). Today, we can reach a larger number of people in lesser number of steps, but the number of people at each step has increased dramatically. Hence, it becomes increasingly difficult to influence people as one moves away from his/her social circle.

2) Influential powers of seed and the source: Higher the influence difference between the seed and the source, greater the seeding effort. Hence, using h^2 indices as approximations of the shell numbers of the nodes,

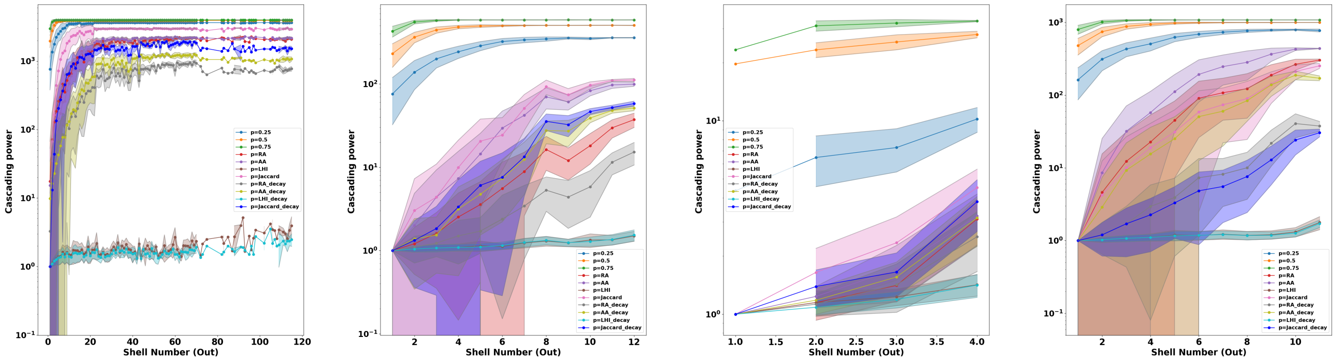
$$c(s,v) \propto h^2(v) - h^2(s) + max_{w \in V(G)}(h^2(w)) \tag{5}$$

[1] One can also use other binary comparison metrics.

[2] x=20 for our experiments.

Dataset	$P(u,v) = 1/d_in(v)$	RA	AA	LHI	Jaccard
OC_links_w.txt	0.08802367306	0.02875541807	0.02283834326	**0.02026267239**	0.02805961572
celegans_n306.txt	0.1129833215	**0.03895540717**	0.04010538201	0.03910858261	0.06336657119
USairport500.txt	0.115072135	**0.04549573447**	0.07709290887	0.06918787321	0.1610725297
Openflights.txt	0.1159152076	0.05050898399	**0.04649292223**	0.05223146929	0.1189160933
USairport_2010.txt	0.06420394069	0.03147297619	**0.01995794608**	0.02405235168	0.1982655843
Freemans_EIES-3_n32.txt	0.0754741649	0.1289793154	0.1983897109	**0.05230991099**	0.3388098307
lkml.txt	0.09449739269	0.006111997109	**0.004170866435**	0.004182685629	0.05793790077
soc-redditHyperlinks-body.txt	0.1402697141	0.005117659033	**0.002867841732**	0.005485788915	0.03133831493
soc-redditHyperlinks-title.txt	0.1085843674	**0.002996738955**	0.004052272382	0.004891667201	0.01913004215

TABLE II: Mean absolute error in the predicted edge probabilities. The least error for each dataset is highlighted in bold font.



(a) Induced subgraph of Facebook

(b) Weblog

(c) Zachary Karate Club Network

(d) Email Network

Fig. 4: Comparison of Cascading power vs Shell Numbers

Dataset (G)	$\lvert V(G)\rvert$	$\lvert E(G)\rvert$
Facebook	4039	88234
Weblog	644	2280
Animal Interaction	446	2139
Email	1134	5451
Zachary	34	78
OClinks_w.txt	1899	20296
Celegans_n306	306	2345
Freemans_EIES-3_n32	32	460
Linux Kernel Mailing List (lkml)	63400	1644660
Openflights	7976	30501
soc-redditHyperlinks-body	35776	137821
soc-redditHyperlinks-title	54075	234792
USairport_2010	1574	28236
USairport500	500	2980

TABLE III: Dataset used for determining importance of probability on cascade size

The last term has been added so that the cost remains positive.

Combining the above two factors,

$$c(s,v) = e^{dist(s,v)}(h^2(v) - h^2(s) + max_{w \in V(G)}(h^2(w))) \quad (6)$$

Since the defined influence costs vary based on the network scale, we define the total budget for a network G as

$$B = f * max_{u \in V(G)}(cost(u,s)) \quad (7)$$

where f is the budget factor.

We find the optimal seed set S for varying budget factors. For good accuracy and efficiency, we use a balanced seed selection strategy similar to that of Han et. al. 2014 [18]. The idea is to find a balanced set S of nodes comprising both, the highly influential (billboard) as well as the most budget friendly nodes (handbill).

Since simulating independent cascade model for large scale networks is a computationally intensive process, we use the Global Structure Model (GSM) heuristic [19] for the estimating the influence power of a node.

$$GSM(u) = e^{\frac{\lambda(u)}{|V|}} * \sum_{v \neq u, v \epsilon V} \frac{\lambda(v)}{dist(u,v)} \quad (8)$$

The greedy approximation for BIM using GSM heuristic is shown in Algorithm 1.

A. Results and Comparison

We select the least outdegree node from each shell as source node. Thereafter, the greedy BIM algorithm 1 finds the set of most influential seeds given the budget B. Three different budget factors, $\{0.5, 1.5, 2\}$, are used. To the best of our knowledge, most of the previous cost functions do not consider the source node. Hence, for comparison purposes, we modify the most commonly used cost functions in literature to take the source node in account. Previous methods to assign costs

Algorithm 1 Greedy BIM using GSM

```
Require: : $G$: Graph ,$k$: number of seeds to be chosen $s$: source node, $B$: budget
  $S \leftarrow \phi$
  $cost \leftarrow 0$
  while $|S| <= k$ do
      $C \leftarrow \{u | u \in V \setminus S \ \& \ cost(s,u) + cost \le B\}$
      $seed \leftarrow \{u | u \in C, \frac{GSM(u)}{cost(s,u)} = max_{v \in V(G)} \frac{GSM(v)}{cost(s,v)}\}$
      $cost \leftarrow cost + cost(seed, s)$
      $S \leftarrow S \cup seed$
  end while
```

are based on the degree [20] [18] of the seed node as listed below

$$c(v) = |\Gamma_{out}(v)| \tag{9}$$

The above cost function can be normalised as following.

$$c(v) = \frac{|\Gamma_{out}(v)| \times |V(G)|}{\sum\limits_{w \in V(G)} |\Gamma_{out}(w)|} \tag{10}$$

We further generalise the cost function (taking seed node s) in account.

$$c(s,v) = |\Gamma_{out}(v)| * e^{d(v,s)} \tag{11}$$

where $dist(s,v)$ represents the shortest path length between v and s. We further normalise the formulation as

$$c(s,v) = \frac{e^{dist(v,s)} \times |\Gamma_{out}(v)| \times |V(G)|}{\sum\limits_{w \in V(G)} |\Gamma_{out}(w)|} \tag{12}$$

The above formulations, GSM based on outdegree (eq. 11) and its scaled variant (eq. 12) are used as baseline measures. Figures 5, 6 and 7 shows the results corresponding to budget factors 0.5, 1.5 and 2 respectively. The X-axis represents the shell number of the source node s and Y-axis represents the final influence, $\sigma(s,S)$. The set S for a given s is obtained using the above 2 baseline measures for cost function as well as the proposed cost function which also takes influence difference in account in addition to the geodesic distance.

It is evident from Figures 5, 6 and 7 that the proposed cost function (GSM-h2 in the plots) leads to more influence in comparison to the baseline approaches.

Based on a modification of role classification algorithm followed in Han et. al, we have categorised the seed nodes in a number of categories. Billboard nodes (top 20% of nodes sorted based on outdegree) are further classified into king (top 30% of billboard nodes sorted based on outdegree) and seignior (remaining 70% billboards). Handbill (non-billboard) nodes with zero outdegree are said to be leaf nodes. A handbill node having a billboard node as outneighbour is called butterfly. Rest of the them are called civilians. The distribution of these seed roles is visualized in the Figure 8. It is seen that the proposed cost function enables a more balanced seed selection approach over the degree based functions.

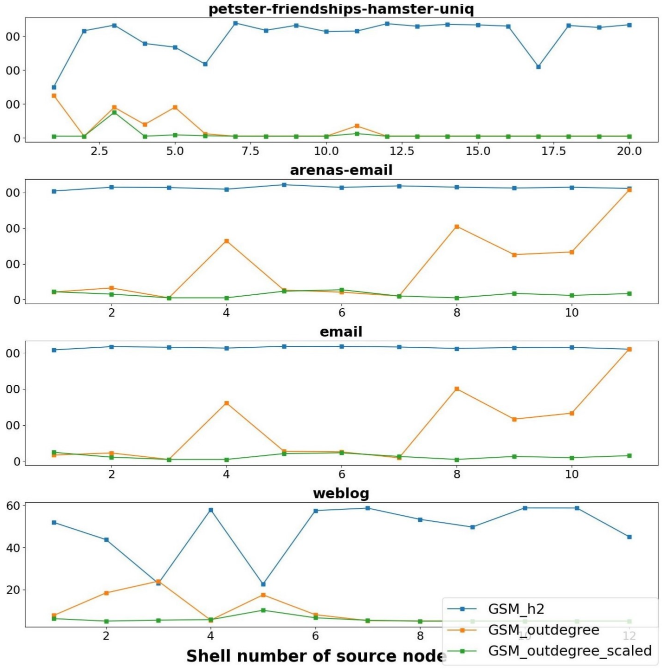


Fig. 5: $\lambda(s)$ vs. $\sigma(s,S)$ for $f = 0.5$

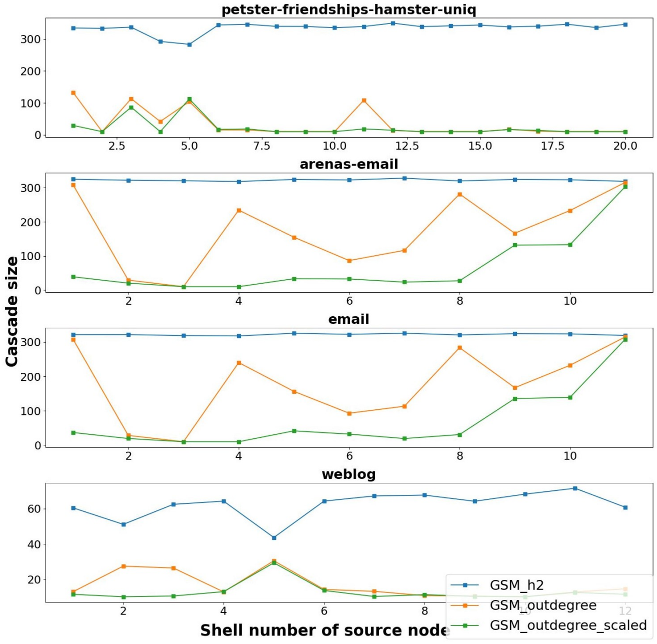


Fig. 6: $\lambda(s)$ vs. $\sigma(s,S)$ for $f = 1.5$

VI. Literature Review

In its most conventional form, Influence Maximisation (IM) aims at finding the superspreaders of influence in a network. To compute the influence function, independent cascade model is used. However, this model requires the knowledge of the influence probability, $p(u,v)$, for each edge $E(u,v)$ in the network $G(V,E)$. Such data is commonly unavailable. Hence, several action log [21], likelihood maximization [22] and IM

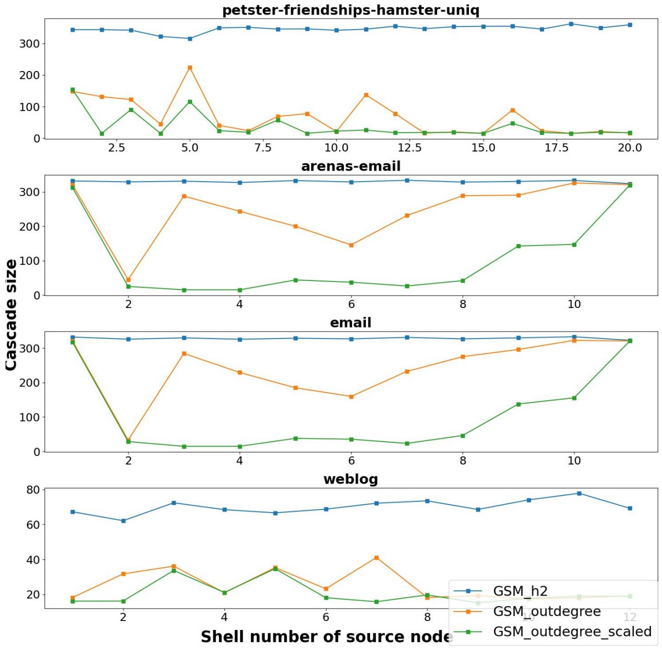


Fig. 7: $\lambda(s)$ vs. $\sigma(s, S)$ for $f = 2$

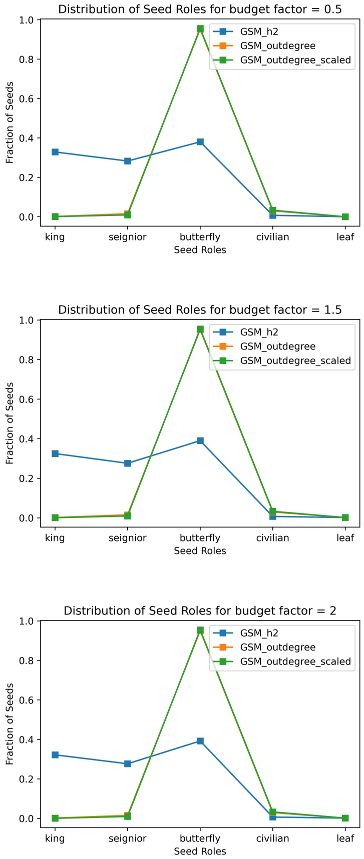


Fig. 8: Distribution of seed roles for different budgets

bandits [23] based methods have been proposed to predict edge weights but they require the knowledge of some amount of past data. There are several other simple heuristics which either sample edge weights uniformly or deduce them based on simple heuristics without needing any additional data. Wang et. al. [24] proposed a variation of the IC model for directed networks, called weighted independent cascade model where edge weight for an edge $E(u, v)$ is reciprocal of v's indegree. Tang et. al. [25] proposed a random model where edge weights are uniformly sampled from $[0, 1]$. Constant model employed by Kempe et. al. in their seminal work [1] suggests .01 to be the most stable probability value. Goldenberg et. al. proposed trivalency model [26] which chooses edge weights uniformly at random from the set $[0.001, 0.01.0.1]$ [27] used a metric based on degrees of nodes and common neighbors to find edge wights. In our work, we show that simple link prediction heuristics like resource allocation and Adamic-Adar coefficient perform fairly well in predicting the edge weights from the structure of the network.

Among various generalisations and specialisations of the IM problem, Budgeted Influence Maximisation (BIM) [7], Targeted Influence Maximisation (TIM) [28], Multiple Influence Maximisation (MIM) [29], Context Aware Influence Maximisation (CAIM) [30], Location Aware Influence Maximisation (LAIM) [31], Competitive Influence Maximisation (CIM) [32] etc. have been proposed.

For BIM, several metrics have been used in literature to compute the seeding costs of the nodes. One most commonly used method is uniform sampling [7]. Xu et. al. [8] used a pagerank based metric. Lei et. al. [33] used the basic idea that the seeding cost of a node is proportional to its influence. However, all these methods assumed the seeding costs to be non-varying, i.e., seeding cost of each node will remain the same irrespective of the interested firm which wants to advertise its product. However, it is often seen in digital marketing that the interpersonal connections of a firm with an influencer helps in better advertising as well as reducing the budget requirements [34]. This shows that in addition to how influential the influencer is, the social distance of a firm from an influencer also decided the seeding cost. Mid-range business finds it difficult to collaborate with the biggest brands. Hence, they are often suggested to make use of micro-influencers. We introduce a similar concept formally, which considers the importance of both, the influencer (seed) as well as the interested firm (source).

VII. Conclusion

With the ongoing advancements in digital marketing, various marketing experts suggest mid-tier firms to target nano and micro influencers rather than mega influencers. We define this concept formally by introducing the importance of source nodes in defining the cost function for influence maximisation. We show that the consideration of the influence difference between the source and the potential seed nodes leads to a better and more balanced seed selection. In addition to reframing the cost function, our work highlights the importance of considering edge weights in the influence propagation process. Consideration of edge weights lead to a better categorisation of users based on their influetial power. We also show that the popular link prediction heuristics, resource allocation metric and Adamic-Adar index are good predictors for the edge weights.

References

[1] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '03. New York, NY, USA: Association for Computing Machinery, 2003, p. 137–146.

[2] J. Zhang, D. Chen, Q. Dong, and Z. Zhao, "Identifying a set of influential spreaders in complex networks," *CoRR*, vol. abs/1602.00070, 2016. [Online]. Available: http://arxiv.org/abs/1602.00070

[3] C. V. Pham, H. V. Duong, and M. T. Thai, "Importance sample-based approximation algorithm for cost-aware targeted viral marketing," *CoRR*, vol. abs/1910.04134, 2019.

[4] M. S. Granovetter, "The strength of weak ties," *American journal of sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[5] S. J. Brams, H. Mutlu, and S. L. Ramirez, "Influence in terrorist networks: From undirected to directed graphs," *Studies in Conflict & Terrorism*, vol. 29, no. 7, pp. 703–718, 2006.

[6] S. Pei, L. Muchnik, J. S. Andrade Jr, Z. Zheng, and H. A. Makse, "Searching for superspreaders of information in real-world social media," *Scientific reports*, vol. 4, no. 1, pp. 1–12, 2014.

[7] H. Nguyen and R. Zheng, "On budgeted influence maximization in social networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 6, pp. 1084–1094, 2013.

[8] X. Xu, Y. Zhang, Q. Hu, and C. Xing, "A balanced method for budgeted influence maximization." in *SEKE*, 2015, pp. 250–255.

[9] J. Park, J. M. Lee, V. Y. Xiong, F. Septianto, and Y. Seo, "David and goliath: When and why micro-influencers are more persuasive than mega-influencers," *Journal of Advertising*, vol. 50, no. 5, pp. 584–602, 2021.

[10] M. Au-Yong-Oliveira, A. S. Cardoso, M. Goncalves, A. Tavares, and F. Branco, "Strain effect-a case study about the power of nano-influencers," in *2019 14th Iberian conference on information systems and technologies (CISTI)*. IEEE, 2019, pp. 1–5.

[11] L. McCormick, "The benefits of building long-term relationships with influencers."

[12] S. P. Borgatti and M. G. Everett, "Models of core/periphery structures," *Social Networks*, vol. 21, no. 4, pp. 375–395, 2000.

[13] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature Physics*, vol. 6, no. 11, pp. 888–893, aug 2010.

[14] A. Saxena, R. Gera, and S. Iyengar, "Estimating degree rank in complex networks," *Social Network Analysis and Mining*, vol. 8, no. 1, pp. 1–20, 2018.

[15] L. Lü, T. Zhou, Q.-M. Zhang, and H. E. Stanley, "The h-index of a network node and its relation to degree and coreness," *Nature communications*, vol. 7, no. 1, pp. 1–7, 2016.

[16] Z. Li, L. Ji, S. Liu, and J. Li, "A new link prediction in directed networks based on attributes fusion," in *2020 IEEE International Conference on Smart Internet of Things (SmartIoT)*, 2020, pp. 161–167.

[17] Y. Gupta, D. Das, and S. Iyengar, "Pseudo-cores: the terminus of an intelligent viral meme's trajectory," in *Complex Networks VII*. Springer, 2016, pp. 213–226.

[18] S. Han, F. Zhuang, Q. He, and Z. Shi, "Balanced seed selection for budgeted influence maximization in social networks," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2014, pp. 65–77.

[19] A. Ullah, B. Wang, J. Sheng, J. Long, N. Khan, and Z. Sun, "Identification of nodes influence based on global structure model in complex networks," *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021.

[20] H. T. Nguyen, T. N. Dinh, and M. T. Thai±, "Cost-aware targeted viral marketing in billion-scale networks," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*. IEEE Press, 2016, p. 1–9. [Online]. Available: https://doi.org/10.1109/INFOCOM.2016.7524377

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

[22] S. Yang and V.-A. Truong, "Online learning of independent cascade models with node-level feedback," *arXiv preprint arXiv:2109.02519*, 2021.

[23] Q. Wu, Z. Li, H. Wang, W. Chen, and H. Wang, "Factorization bandits for online influence maximization," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 636–646.

[24] Y. Wang, H. Wang, J. Li, and H. Gao, "Efficient influence maximization in weighted independent cascade model," in *International Conference on Database Systems for Advanced Applications*. Springer, 2016, pp. 49–64.

[25] Y. Tang, Y. Shi, and X. Xiao, "Influence maximization in near-linear time: A martingale approach," in *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, 2015, pp. 1539–1554.

[26] 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, no. 3, pp. 1–18, 2001.

[27] N. Trivedi and A. Singh, "Efficient influence maximization in social-networks under independent cascade model," *Procedia Computer Science*, vol. 173, pp. 315–324, 2020.

[28] C. Song, W. Hsu, and M. L. Lee, "Targeted influence maximization in social networks," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016, pp. 1683–1692.

[29] H. Sun, X. Gao, G. Chen, J. Gu, and Y. Wang, "Multiple influence maximization in social networks," in *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, 2016, pp. 1–8.

[30] A. K. Singh, L. Kailasam, T. Pradhan, and D. Gupta, "Context-aware influential nodes tracking in online social networks," 2021.

[31] T. Zhou, J. Cao, B. Liu, S. Xu, Z. Zhu, and J. Luo, "Location-based influence maximization in social networks," in *Proceedings of the 24th ACM international on conference on information and knowledge management*, 2015, pp. 1211–1220.

[32] S. Chakraborty, S. Stein, M. Brede, A. Swami, G. de Mel, and V. Restocchi, "Competitive influence maximisation using voting dynamics," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2019, pp. 978–985.

[33] L. Zhang, Y. Liu, F. Cheng, J. Qiu, and X. Zhang, "A local-global influence indicator based constrained evolutionary algorithm for budgeted influence maximization in social networks," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 2, pp. 1557–1570, 2021.

[34] M. Haenlein, E. Anadol, T. Farnsworth, H. Hugo, J. Hunichen, and D. Welte, "Navigating the new era of influencer marketing: How to be successful on instagram, tiktok, & co." *California management review*, vol. 63, no. 1, pp. 5–25, 2020.