International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Targeted Influence Maximization Based on Cloud Computing over Big Data in Social Networks

Kavita Joshi[1], Prof. Roshani S. Golhar[2], Dr. Neerajkumar Sathawane[3], Dr. Ashutosh Mathur[4], Pravin Ganpatrao Gawande[5], Milind S. Patil[6], Santosh Gore[7]

Submitted: 17/01/2024 **Revised**: 25/02/2024 **Accepted**: 03/03/2024

Abstract: This research focuses on cloud computing-based targeted impact maximization in social networks. Most influence maximisation operates currently in use to identify the Top-k, which is a node in a network that is recognised or chosen according to specific standards, including the parameter "k." Users are expected to maximise the spread of influence under the assumption that the effect diffusion possibilities on connections are fixed, and these works assume an understanding of the whole networking graph. In practical settings, however, marginal probability tends to vary depending on a range of issues and can be influenced by incoming information. Therefore, the greedy algorithm is used. These approaches aim to detect a seed collection that increases the anticipated impact distribution across users with target audiences who are pertinent for specified subjects. The MIA model is used to locate the subgraph in a network where a certain collection of nodes can have the greatest impact on other nodes, which results in influence coverage and effectiveness. In the meantime, privacy and computational concerns make it challenging to access all network data. Additionally, current impact maximization techniques that take target users into account do not address cloud computing, which results in our algorithm consistently outperforming other scalable heuristics in influence spread across all size ranges, outperforming greedy algorithms by up to 100%-260%. This study suggests a cloud-based targeted influence maximization strategy to achieve this goal.

Keywords*: Influence Maximization (IM), Maximum Influence Arborescence (MIA), Independent Cascade (IC) Model*

1. Introduction

In today's world, online social networks are essential for the dissemination of knowledge, concepts, and influence among individuals[1]. A social network is an association of individuals or businesses linked together through interactions or relationships. These relationships may be founded on common interests, professions, or relationships. There are billions of nodes and limits in the modern generation of social networks. They are now the most informative resources for big data[2]. The sources of big data are diverse and span multiple areas, including social media, the Internet of Things, corporate transactions, academic data, public and government databases, and more. For both academia and industry, managing and mining large amounts of social data is a difficulty[3].

Social media has been rapidly utilized as a medium for marketing and promotion. To promote a product, for example, a marketer in viral marketing attempts to identify a seed set of customers who have significant influence over the product being promoted. A marketer's goal is to increase the number of clients who ultimately decide to adopt the product while working within a set budget for options.

This is the problem of classical influence maximization[4][5].

Through the Internet, cloud computing offers users access to computational resources like servers, storage, databases, and software[6]. It does away with the necessity for companies to purchase physical infrastructure by letting them lease resources from cloud service providers.

Users can access data and applications remotely thanks to this adaptable and scalable paradigm, which encourages accessibility and cost-effectiveness. Certain users could demonstrate higher levels of activity, influence, or importance as a result of their network friends or behaviour[7]. Supposedly powerful users have significant function in social networks and may be essential in facilitating the rapid and broad dissemination of messages[8]. The challenge of determining a category of people in the social world who have substantial and widespread influence over other users is called the Influence Maximization (IM) problem. It's also well known that this is a difficult topic, especially since determining influence requires a criterion.

[1]Asst.Prof., Dr.D.Y.Patil Institute of Engineering, Management and Research Akurdi Pune, 411035, India, kvtjoshi68@gmail.com
[2]Research Scholar, Sun Rise University Alwar, (Rajasthan) India, roshni.golhar@gmail.com
[3]Associate Professor, MIMA Institute of Management, Pune, MH, India, niraj_51413@rediffmail.com
[4]Assistant Professor, Symbiosis Centre for Management Studies, Pune Symbiosis International, Pune, ashutosh.mathur@scmspune.ac.in
[5]Savitribai Phule Pune University, Vishwakarma Institute of Information Technology, Pune, pravin.gawande@viit.ac.in
[6]Assistant Professor E&TC Engineering Department, Vishwakarma Institute of Information Technology, Pune 4, milind.patil@viit.ac.in
[7]Director, Sai Info Solution, Nashik, Maharashtra, India, https://orcid.org/0000-0003-1814-59131 ,sai.info2009@gmail.com

This research, driven by these earlier discoveries, focuses on cloud computing-based targeted influence maximisation with partial network information. The objective identify the Top-k significant nodes that are pertinent to the supplied themes to maximise the influence spread among the target audience[9]. For the maximum coverage problem, the Greedy technique is utilised to determine the Top-k Seed Runners. A greedy algorithm's main objective is to select the optimal solution among those that are now available, without going back or re-evaluating choices made in the past[10]. When solving optimisation issues the aim is to determine the most effective approach given a set of limitations, greedy algorithms are frequently employed. The term "Maximum Influence Arborescence" (MIA) describes the subgraph in a network where a specific collection of nodes can have the greatest impact on other nodes. Put more simply, it looks for the connections between the most significant nodes in the network and its subgroups of nodes that have the greatest influence or information spreading across the network[11]. Experiments have been performed on multiple real-world social networks. Despite their significantly increased efficiency, the experimental results show that the strategies can continue to take a few minutes on a system with millions of nodes[12]. The Greedy algorithm and MIA model perform well in the influence spreads and require less running time than other algorithms. The experiment results validate the influence coverage and effectiveness of the proposed algorithms.

The following is a summary of our contributions:

- It explores the Independent Cascade model, the Influence Maximization challenge as a variant of the classic influence maximization problem is maximizing influence coverage in an evolving social network [13].
- It provides the greedy technique and MIA model, which is an effective approach. For optimisation issues where finding the best solution given a set of constraints is the goal, we suggest the greedy approach and MIA model.
- The MIA model's influence spread is sub-modular, meaning it has a diminishing marginal return property. As a result, any advanced ratio calculation is NP-hard[14], and the greedy technique selects the node with the most margin influence spread in every round, resulting in an impact extended inside (1 - 1/e) of the MIA model's best result. Because (a) the minimal effect spread on the arborescent structures can computed efficiently through recursive, and (b) once the seed with the greatest impact on spread has been selected, it just needs to improve the nearby arborescent structures. associated with it to select the next seed. Also, create an ongoing update strategy to accelerate the upgrading process.
- It assesses the results using a real, extensive social network. Our conceptual findings are supported by the experiment data, which also demonstrate that greedy algorithms and the MIA model improve concerning influence coverage and effectiveness.

2. Related Work

The initial people near examine the algorithmic problem of influence maximization are Domingos and Richardson [15][16][17]. However, their approaches are probabilistic. Kempe, Kleinberg, and Tardos initially defined it as a discrete optimization issue. [18]. In addition to the subjects we have just covered, they also research a variety of other subjects like mixed marketing techniques in impact maximization and generalizations of influence cascade models. As mentioned, their greedy algorithms' scalability is the primary flaw in their work.

To address this issue, several recent research were conducted. As evidenced by their experimental results, Leskovec et al. [19] provide a “lazy-forward” efficiency in the process of choosing new seeds that significantly decreases the number of decisions on the influence distribution of nodes and speeds up the process by up to 700 times. As demonstrated in [20], finding the 50 most important nodes in a system with tens of thousands of nodes still takes hours, despite the "lazy-forward" efficiency being considerable[21].

Kimura and Saito propose influencing cascade (IC) models that utilize the shortest path and offer effective techniques for computing below these models, the influence spreads [22]. The fundamental difference between their approach and ours is that (a) they employ basic shortest paths on the graph, those are unrelated to spread possibilities, in place of maximum influence paths, and (b) it does not use localized structures, like our arborescence., so each round they must perform international calculations to choose the subsequent seed. As a result, the algorithms they use lack our level of efficiency.

The search for scalable and effective influence maximization methods is carried out by this study, which is an extension of [20]. In [20], we investigate two approaches to increase efficiency: creating new heuristic algorithms and refining the greedy algorithm from [18]. While there has been some development in the first direction, it is not enough to suggest that this course will be easy to follow. The second path produces extremely effective new degree discount heuristics with a passably respectable spread of influence. The primary problem is that the degree discount algorithms rely on the uniform IC model, which is rarely the practice case and assumes that diffusion probability on all edges is equal[23]. Our recent work represents a significant phase in breaking this restriction; our fresh heuristic approach maintains a solid balance between efficacy and efficiency while working for the broader IC model. In comparison to [20], we carry out a lot more experiments on a wider range of graphs and our data demonstrate that the MIA heuristic regularly outperform the degree discount heuristic throughout the board[24].

3. Propose Methodology

This research aims to identify a seed combination that maximises the impact among an aim group of users who are thinking about different themes. In this section, effective algorithms for maximising focused influence are shown. Some concepts are taken from to address targeted influence maximisation utilising partial network information.

Independent Cascade (IC) Model and Influence Maximization (IM) Problem

This study examined the researcher investigates communal impact using the well-known Independent Cascade (IC) model[18]. The directed system is how the social network is represented in the IC model. G=(V,E), where V stands for the people and E for the social linkages that connect them. Additionally, a propagation probability $\mathcal{P}_{u,v}^{G}$, which represents the degree to which each edge $(u,v) \in E$ has an impact on v, is linked to each individual u. To keep the notations simple when G is evident based on the setting, it only uses $\mathcal{P}_{u,v}$.

A straightforward and understandable diffusion process is described by the Independent Cascade model. The diffusion mechanism occurs in discrete time increments, as follows, beginning with a seed set S that is active at the beginning (having adopted a behaviour). Step t + 1 is the attempt made by a node u to activate all of its dormant neighbours when it becomes active in step t. It is successful with the known probability $\mathcal{P}_{u,v}$ for each neighbour v. If it is successful, v becomes active; if not, v stays dormant. It is not allowed to attempt any more activation attempts after you have completed all of these[25].

The impact function indicates the anticipated number of working nodes at the final stage of the diffusion process $\sigma(S)$, is what we describe as the effect coverage of seed set S. According to the IC model, the Influence Maximization (IM) problem seeks to maximize the influence function $\sigma(S)$ by locating a seed set $S \subseteq V$ of maximum size k. Formally, an optimization problem

$$S^* = \underset{|S| \leq k}{argmax}\, \sigma(S)$$

is specified as the IM problem. Despite the NP-hardness of the influence maximization problem in the IC model, as demonstrated by Kempe et al. [18], the IC model's advantageous attributes enable the use of an approximate approach to identify the important nodes: According to the Independent Cascade model, the influence purpose $\sigma(S)$ is submodular and monotone [18][26].

3.1. Greedy Algorithm

The greedy technique is a way of solving problems where choices are made at every step based just on what is best, without taking future effects into account. Making locally optimal decisions in the hopes of locating a global optimum characterizes it. The greedy algorithm chooses the best alternative available at each step, which could provide a series of decisions that don't always result in the best answer overall. Although Greedy technique are frequently effective and simple to use, they cannot always yield the best answer to challenging issues[27]. They are frequently employed in tasks involving scheduling, minimum spanning trees, and shortest path algorithms, where selecting the best local option at each stage results in a workable solution.

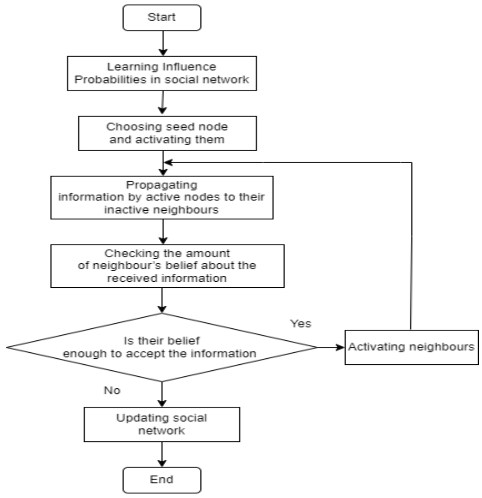

Fig 1: Influence maximization in a greedy algorithm.

Using the greedy approach, continuous influence maximization involves selecting nodes from a collection of to maximize the propagation of influence over time. The technique adds nodes iteratively based on their projected influence contribution, starting with an empty set. When a predetermined threshold is reached, such as a budget limit or an insignificant gain, it ends. The estimation of influence dispersion and node selection determine the complexity. The goal is to effectively choose nodes for the network's maximal continuous influence propagation[28].

The algorithm for targeted influence maximisation is given in this section. give rise to Nemhauser et al.'s [23] straightforward greedy method (Algorithm 1) for maximising monotone submodular equations. The process iteratively selects the node with the largest marginal gain is added to the present seed set until the budget k is reached. Established by [23], this technique roughly calculates the best result for the Influence Maximization issue using a factor of the $(1 - 1/e)$.

Algorithm 1 Greedy (*k*, *f*)

1: initialize $S = \emptyset$

2: for *i*=1 and *k* do

3: select $u = argmax_{w \in V \setminus S}\big(f(S \cup \{w\}) - f(S)\big)$

4: $S = S \cup \{u\}$

5: end for

6: output S

Function $\sigma_I(\cdot)$ is demonstrated in [18] to be submodular and monotone, with $\sigma_I(\emptyset) = 0$. As a result, the impact maximization problem is solved by the algorithm Greedy (k, σI) with an approximate ratio of $(1 - 1/e)$. One significant problem, though, is that provided a set S, there is no effective method to compute $\sigma_I(S)$. While Kempe et al. state that it is open to find an efficient algorithm to compute $\sigma_I(S)$ [18], It demonstrates that the process is #P-hard. [29], as demonstrated through a decrease after the count issue s-t connection within the graph [30].

3.2. Maximum Influence Arborescence (MIA) model

In network influence maximization problems, the maximum influence arborescence (MIA) model is employed. It is the process of locating the subgraph in a network where a certain collection of nodes can have the greatest impact on other nodes. To put it another way, it looks for the connections between the most significant nodes in a system and the subsets that have the greatest influence or information spreading across the network. To find important nodes for focused interventions or marketing efforts, the MIA model is frequently used in social network research and viral marketing techniques.

The propagation probability of path $P = \langle u = p1, p2, \dots, pm = \upsilon\rangle$, is defined as

$$pp(P) = \prod_{i=1}^{m-1} pp(pi, pi + 1).$$

Given that it must activate each node along the path, it follows that the possibility that u will activate υ over path P is $pp(P)$. We suggest using the Maximum Influence Path (MIP) to quantify the impact of moving from one node to a different one to estimate the real projected influence within the social network. $\mathcal{P}$(G, u, υ) is a collection of all routes in a graph G that connect u and υ.

Definition of maximum influence path

The maximum influence path $MIP_G(u, \upsilon)$ for the graph, G has been defined as follows:

$$MIP_G(u, \upsilon) = {}^{argmax}_{P} \{pp(P) | P \in \mathcal{P}(G, u, \upsilon)\}.$$

$MIP_G(u, \upsilon)$ is always unique because the connections are fragmented in a determined and constant manner, and every subpath in $MIP_G(u, \upsilon)$ that connects x to y is also the $MIP_G(x, y)$. We indicate $MIP_G(u, \upsilon) = \emptyset$ if $\mathcal{P}(G, u, \upsilon) = \emptyset$.

Observe that $MIP_G(u, \upsilon)$ is just the shortest length among u and υ in weighted graph G for every edge (u, υ) we can translate the spread of probabilities in the graph $pp(u, \upsilon)$ to a distance weight $-\log pp(u, \upsilon)$ on the edge. As a result, the shortest paths and shortest-path arborescences directly correlate to the greatest influence paths, and then to the maximal influence arborescence, making it possible for effective algorithms like the Dijkstra algorithm to compute them. We suggest using greatest Influence In-Arborescence (MIIA), which is the combination of most impact routes to υ,[3] (The union of maximum impact paths to a node does not include undirected cycles meanwhile we continuously break ties in extreme impact paths, proving that arborescence is present) for a specific node υ in the chart to compute the impact that other nodes between the network have on υ. To calculate υ's influence on other nodes, we additionally describe Maximum Influence Out-of-Arborescence (MIOA).

Definition of Maximum Influence In (Out) Arborescence (MIIA)

The maximal influence in-arborescence of a node $\upsilon \in V, MIIA(\upsilon, \theta)$, for an influence threshold θ is as follows:

$$\mathrm{MIIA}(\upsilon, \theta) = \cup_{u \in V, pp(\mathrm{MIP}_G(u,\upsilon)) \geq \theta} \mathrm{MIP}_G(u, \upsilon).$$

Maximum Influence Out-Arborescence is:

$$\mathrm{MIOA}(\upsilon, \theta) = \cup_{u \in V, pp(\mathrm{MIP}_G(u,\upsilon)) \geq \theta} \mathrm{MIP}_G(\upsilon, u)$$

Algorithm 2: $ap\big(u, S, \mathrm{MIIA}(\upsilon, \theta)\big)$

1: If $u \in S$ $then$

2: $ap(u) = 1$

3: $else\ if\ N^{in}(u) = \emptyset\ then$

4: $ap(u) = 0$

5: $else$

6: $ap(u) = 1 - II_{w \in N^{in}(u)} \big(1 - ap(w).pp(w, u)\big)$

7: $end\ if$

Pretentious that the effect from S to υ is exclusively developed along edges in $\mathrm{MIIA}(\upsilon, \theta)$, we can approximate the IC model with an array of seeds S in G, the in-arborescence $\mathrm{MIIA}(\upsilon, \theta)$ to few $\upsilon \notin S$. It can precisely compute the chance that υ is active given S using this approximation. Assume that the activation probability of every node u in $\mathrm{MIIA}(\upsilon, \theta)$ is $ap\big(u, S, \mathrm{MIIA}(\upsilon, \theta)\big)$, which represents the likelihood that u is activated when the influence propagates in $\mathrm{MIIA}(\upsilon, \theta)$ and the seed set is S. Let $\mathrm{N}^{in}\big(u, \mathrm{MIIA}(\upsilon, \theta)\big)$N be the set of u's in-neighbours in $\mathrm{MIIA}(\upsilon, \theta)$ When it is evident from the context, $\mathrm{MIIA}(\upsilon, \theta)$ and S in the preceding notations may be omitted. Next, using Algorithm 2, one can compute $ap\big(u, S, \mathrm{MIIA}(\upsilon, \theta)\big)$ recursively.

3.3. Greedy Algorithm is more Efficient

The selection of the subsequent seed with the highest progressive influence spread is the only crucial step within the greedy algorithm. Using the provided seed set S, assume the highest impact in-arborescence $\text{MIIA}(v,\theta)$ of size t. The initiation possibility $ap\big(u, S \cup \{w\}, \text{MIIA}(v,\theta)\big)$ for each $w \in \text{MIIA}(v,\theta)$ must be computed to choose the next seed u. If we simply apply Algorithm 2 for computing each $ap\big(u, S \cup \{w\}, \text{MIIA}(v,\theta)\big)$, this will take O(t2) time. To compute $ap\big(u, S \cup \{w\}, \text{MIIA}(v,\theta)\big)$'s for every $w \in \text{MIIA}(v,\theta)$ in O(t) time, we now provide a batch updating approach. The following lemma is easily derived from Algorithm 2 of 6 lines.

Lemma 1 (Influence Linearity): Illustrates how we use the linear link between $ap(u)$ and $ap(v)$) within $\text{MIIA}(v,\theta)$. Think about a node u in $\text{MIIA}(v,\theta)$. $ap(v) = \alpha(u,v) \cdot ap(u) + \beta(v,u)$, where $\alpha(u,v), \beta(v,u)$ are variables are interdependent of *ap(u)*, If we understand the activating probability *ap(u)* and *ap(v)* as variables and other $ap(w)$'s as constants, where w is any node in $\text{MIIA}(v,\theta)$ other than u and v.

3.4. Prefix excluding MIA model (PMIA)

The max in-fluence path from u to v is the only influence propagation path that is taken into account in the basic MIA model. Take two seeds, s1 and s2, in the case when $\text{MIP}_G(s_2,v) \subset \text{MIP}_G(s_1,v)$. In the fundamental MIA model, the action of s1 on v is blocked by s2 in the middle, meaning that s2 alone determines the probability that v is activated and is unaffected by s1. Researchers favour an MIA model where a seed's influence is not inhibited by other seeds to obtain a more accurate approximation of the IC model. Taking into account maximal influence paths while avoiding other seeds is a logical technique to expand the fundamental MIA model.

In this paradigm, the general Algorithm 1 is also functional. It's unclear, though, how to apply it effectively in a way that is comparable to MIA Algorithm. This section examines a version of the previously mentioned extension that permits an effective greedy algorithm. This expansion is known as the PMIA (prefix excluding MIA) model. It makes sense that the seeds of the PMIA model have an order, determined by a greedy algorithm's method of selection. The greatest influence paths that a given seed s can take to reach other nodes should steer clear of every seed in the prefix preceding s. The consequence of the largest impact on (out)-arborescence for the PMIA model is the primary technical distinction, particularly if one wants to create a successful greedy algorithm within the framework of the MIA algorithm.

3.5. Algorithm in PMIA model

The modifications required to align MIA algorithm with the PMIA model are now presented. The calculation of $\text{PMIIA}(v,\theta,S)$ and $\text{PMIOA}(v,\theta,S)$ is the main problem. Meanwhile, we just want to delete S from the graph, computing $\text{PMIOA}(v,\theta,S)$ is a rather straightforward process. Thus, we may compute $\text{PMIOA}(v,\theta,S)$ using the Dijkstra method on graph G(S).

For every node v ∈ V \ S, people keep the collection of useless seeds $IS(v,S)$ to compute $\text{PMIIA}(v,\theta,S)$ efficiently. The calculation of $\text{PMIIA}(v,\theta,S)$ can be done using $IS(v,S)$. We traverse the inner edges from v to begin a Dijkstra algorithm. The Dijkstra algorithm ceases this branch and does not proceed on s's in-neighbours whenever it encounters a seed node, s. Upon the completion of the Dijkstra method, we eliminate every node $IS(v,S)$ from the calculated in-arborescence.

All nodes v in $\text{PMIOA}(u,\theta,S)$ must have their $IS(v,S)$updated when a new seed, u, is chosen. To accomplish this, examine the collection of active seeds $\big(in\ S \backslash IS(v,S)\big)$ that the u has stopped in $\text{PMIIA}(v,\theta,S)$. To be thorough, we also provide the PMIA algorithm, which is the PMIA model's efficient greedy algorithm. PMIA is identical to MIA model, except that all MIIAs and MIOAs are swapped out for PMIIAs and PMIOAs, and these PMIIAs and PMIOAs are recomputed each time the seed set is altered.

4. Experiment Analysis

The targeted influence maximisation, algorithms are tested directly against the most advanced methods in this section. This part includes the findings from our dataset trials, a new proposal probability type of the IC model, and more heuristic techniques.

4.1. Dataset

For testing, two datasets are used. The first, called NetPHY, was used in [20] and has 37,154 nodes and 231,584 edges. It is sourced from the "Physics" section of arXiv. The second, called DM, is based on data mining research and has 680 nodes and 1689 edges [31]. It was acquired from ArnetMiner. Basic network statistics are shown in Table 1. Furthermore, scalability studies use fake data to create networks of different sizes.

Table 1: Dataset Characteristics

Dataset	NetPHY	DM	NetHEPT
#Node	38K	680	17K
#Edge	176K	1689	32K
Average Degree	13.5	5.15	4.57
Maximal Degree	287	64	65
#Component Connection	3889	1	1781

Largest Size Component	19877	680	6804
Average Size Component	10.23	680	9.6

4.2. Propagation probabilities

Propagation probabilities are produced using an additional model, which is explained below. Furthermore, for the TRIVA-LENCY model, we employ a distinct set of parameters.

- TAP model: A topical affinity propagation (TAP) approach for calculating propagation probabilities depending upon both structure and topic information in the group is introduced by a model in [31]. There is variation in this probability. Propagation probabilities are computed for the DM dataset using topical data that is currently available. Since precise subject information is lacking in the NetHEPT dataset, a uniform topic distribution is assumed. The size of the NetPHY dataset prevents the TAP algorithm from being used.
- TRIVALENCY Model: Replace the 0.2, 0.02, and 0.002 in the main paragraph with the probability values 0.3, 0.06, and 0.0012.
- WC Model: Network edges are given weights by the weighted cascade model, which indicates the degree of effect between nodes. By incorporating different degrees of impact in social networks, these weights calculate activation probabilities as information travels. When information dispersion is influenced by connection strength, it is helpful.

4.3. Algorithms

The following extra algorithms are included for evaluation.

- Degree: A straightforward heuristic that determines which k nodes in a graph have the biggest out-degrees
- Weighted Degree: The total of the propagation probabilities on all of a node's outgoing edges is its weighted degree. The k nodes with the highest weighted degrees are chosen by this heuristic.
- SPM: By combining the shortest-path approach from [22] with lazy-forward optimization from [19], SP1M is an enhanced version of SPM. For calculating influence, it takes into account both the shortest pathways and those that need an extra hop from S to node v.
- PageRank: The widely used algorithm for page ranking on the internet [32]. A hyperlinked group of documents, such as web pages, can have each element given a numerical weight by the link analysis method PageRank, which determines the documents' relative relevance in the network.

5. Results and Discussion

The influence spread maximizes and becomes more efficient when a greedy and heuristic approach is used; the results are shown below.

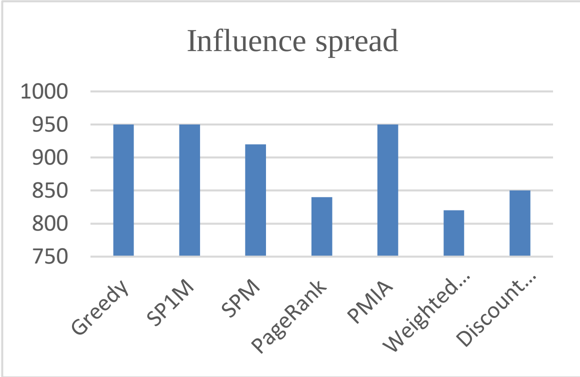

Fig 2: The influence spread for several WC model algorithms on the NetHEPT dataset.

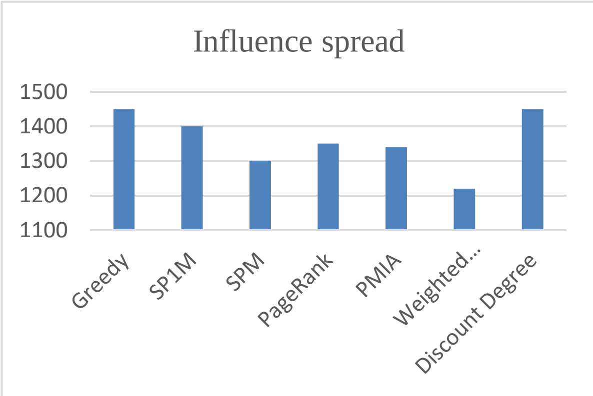

Fig 3: The influence spread for several WC model algorithms on the NetPHY dataset

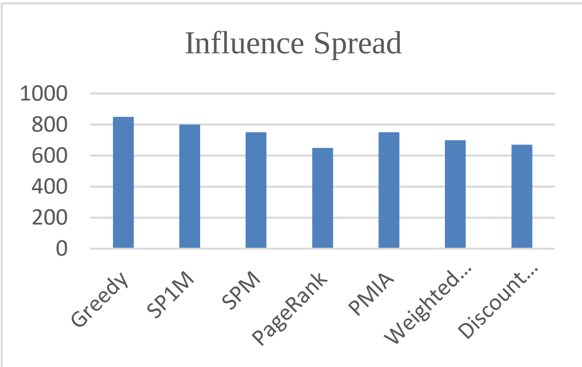

Fig 4: The influence spread for several TAP model algorithms on the NetHEPT dataset.

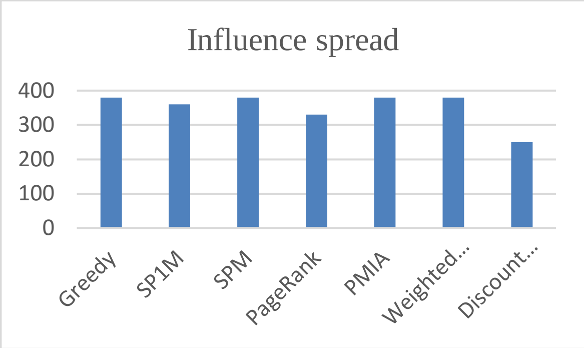

Fig 5: The influence spread for several TAP model algorithms on the DM dataset.

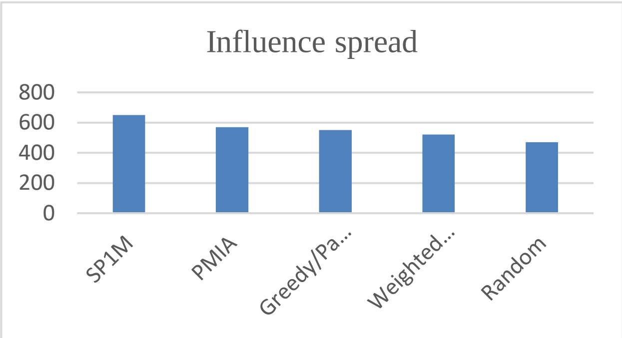

Fig 6: Influence distribution for different algorithms in the NetHEPT dataset's TRIVANLENCY model with three probabilities (0.3, 0.06, and 0.0012).

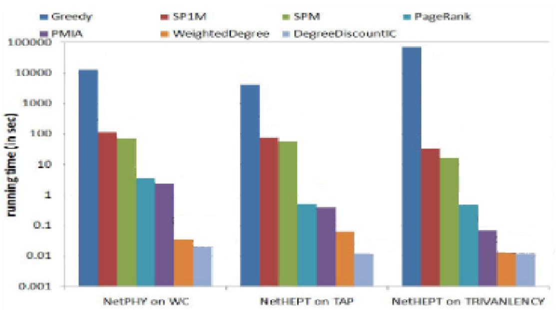

Fig 7: Algorithm execution times over three datasets.

Table 2: Execution time on different algorithms

Algorithms	NetPHY on WC	NetHEPT on TAP	NetHEPT on TRIVANLENCY
Greedy	3.5hr	1.2hr	Greater than 20hr
SP1M	1.9min	1.3min	34.3s
SPM	1.2min	54.7s	17.2s
PageRank	3.6s	0.5s	0.5s
PMIA	2.5s	0.4s	66.6ms
Weighted Degree	34.9ms	62.2ms	12.9ms
Degree Discount	19.3ms	12.2ms	11.3ms

Result on Influence spreading: The influence spread findings are displayed in Figures 2–5, together with the outcomes of the algorithms we tried for the main text. The majority of the findings are straightforward and consistent with the conclusion stated in the main paragraph. In general, PMIA outperforms the other heuristics, including the innovative ones we reviewed here, and performs consistently over all datasets and propagating models, equal or surpassing the efficiency of Greedy and SPM/SP1M. Figure 6 deserves particular attention because it demonstrates how noticeably inferior Greedy performs to PMIA. An issue of that Greedy is too impatient, which means we need to lower the number of trials from 20,000 to 200 to estimate influence spread accurately. For more information on why Greedy is sluggish, go to the running time section. This suggests that decreasing the number of simulations won't be enough to quickly accelerate Greedy. It's also important to note that Weighted-Degree does fairly well, Figures 4 and 5 show the two TAP model-related exams close to PMIA. The TAP model is expected to generate an impact model in which the bulk of influences are, in fact, only transferred during a single phase, whereas Weighted Degree only takes into account influence propagated among one-step neighbours. Weighted Degree is not as consistent as PMIA, though, as evidenced by its poorer performance in other tests.

Execution time: NetPHY uses the WC model; NetHEPT uses the TAP model; and NetHEPT uses the TRIVALENCY model are the three tests that are compared in Figure 7 and Table 2 described in terms of running times. Lazy forward optimization is less successful when greedy performs worse in the TRIVALENCY model because of sharp drops in the minimal impact spread for succeeding seed candidates. The reason behind PMIA's quick execution time during the third test (NetHEPT on TRIVALENCY) is that a higher θ value results in smaller arborescences. Running time could be further enhanced by adjusting θ. Despite this, by taking into account overlapping impacts among seeds, PMIA performs better in influence spread than Weighted Degree. Thus, PMIA performs better in influence spread than Weighted Degree even though it has maintenance costs associated with arborescence data structures and frequent updates.

6. Conclusion

This research examines the study on influence maximization in cloud computing over big data in social networks utilizing the Maximum Influence Arborescence (MIA) model and the greedy method. Despite being computationally efficient, the greedy algorithm's myopic nature may prevent it from always producing the best results. It does, however, provide a useful starting point for comparison. However, the MIA model provides a more advanced method by considering the network's overall structure and selecting significant nodes according to their overall influence. Empirical investigation revealed that the MIA model performs better than the greedy algorithm when it comes to maximizing influence over vast data in social networks. The MIA model can effectively manage large-scale datasets by utilizing cloud computing resources, which makes it appropriate for real-world applications.

References

[1] A. Ramlatchan, M. Yang, Q. Liu, M. Li, J. Wang, and Y. Li, "A survey of matrix completion methods for recommendation systems," *Big Data Mining and Analytics*, vol. 1, no. 4. pp. 308–323, 2018. doi: 10.26599/BDMA.2018.9020008.

[2] S. Gore, Y. Bhapkar, J. Ghadge, S. Gore, and S. K. Singha, "Evolutionary Programming for Dynamic Resource Management and Energy Optimization in Cloud Computing," in *Proceedings of 3rd International Conference on Advanced Computing Technologies and Applications, ICACTA 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICACTA58201.2023.10393769.

[3] B. N. Tiwari and R. K. Thakur, "On stability of thermodynamic systems: a fluctuation theory perspective," *Eur. Phys. J. Plus*, vol. 138, no. 6, Jun. 2023, doi: 10.1140/epjp/s13360-023-04000-6.

[4] L. Qi, Y. Chen, Y. Yuan, S. Fu, X. Zhang, and X. Xu, "A QoS-aware virtual machine scheduling method for energy conservation in cloud-based cyber-physical systems," *World Wide Web*, vol. 23, no. 2, pp. 1275–1297, Mar. 2020, doi: 10.1007/s11280-019-00684-y.

[5] X. Wang, W. Wang, L. T. Yang, S. Liao, D. Yin, and M. J. Deen, "A Distributed HOSVD Method with Its Incremental Computation for Big Data in Cyber-Physical-Social Systems," *IEEE Trans. Comput. Soc. Syst.*, vol. 5, no. 2, pp. 481–492, 2018, doi: 10.1109/TCSS.2018.2813320.

[6] C. Zhang, M. Yang, J. Lv, and W. Yang, "An improved hybrid collaborative filtering algorithm based on tags and time factor," *Big Data Min. Anal.*, vol. 1, no. 2, pp. 128–136, 2018, doi: 10.26599/BDMA.2018.9020012.

[7] R. K. Thakur, H. Kumar, S. Gupta, D. Verma, and R. Nigam, "Investigating the Hubble tension: Effect of cepheid calibration," *Phys. Lett. Sect. B Nucl. Elem. Part. High-Energy Phys.*, vol. 840, 2023, doi: 10.1016/j.physletb.2023.137886.

[8] N. Mahankale, S. Gore, D. Jadhav, G. S. P. S. Dhindsa, P. Kulkarni, and K. G. Kulkarni, "AI-based spatial analysis of crop yield and its relationship with weather variables using satellite agrometeorology," in *Proceedings of 3rd International Conference on Advanced Computing Technologies and Applications, ICACTA 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICACTA58201.2023.10392944.

[9] W. Yu, G. Cong, G. Song, and K. Xie, "Community-based Greedy algorithm for mining top-K influential nodes in mobile social networks," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 1039–1048. doi: 10.1145/1835804.1835935.

[10] G. Song, Y. Li, X. Chen, X. He, and J. Tang, "Influential Node Tracking on Dynamic Social Network: An Interchange Greedy Approach," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 2, pp. 359–372, 2017, doi: 10.1109/TKDE.2016.2620141.

[11] S. Gore, D. D. Jadhav, M. E. Ingale, S. Gore, and U. Nanavare, "Leveraging BERT for Next-Generation Spoken Language Understanding with Joint Intent Classification and Slot Filling," in *Proceedings of 3rd International Conference on Advanced Computing Technologies and Applications, ICACTA 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICACTA58201.2023.10393437.

[12] S. Gore, A. S. Deshpande, N. Mahankale, S. Singha, and D. B. Lokhande, "A Machine Learning-Based Detection of IoT Cyberattacks in Smart City Application," 2023, pp. 73–81. doi: 10.1007/978-981-99-6568-7_8.

[13] C. C. Aggarwal, S. Lin, and P. S. Yu, "On influential node discovery in dynamic social networks," in *Proceedings of the 12th SIAM International Conference on Data Mining, SDM 2012*, Society for Industrial and Applied Mathematics Publications, 2012, pp. 636–647. doi: 10.1137/1.9781611972825.55.

[14] L. Sun, W. Huang, P. S. Yu, and W. Chen, "Multi-round influence maximization," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Jul. 2018, pp. 2249–2258. doi: 10.1145/3219819.3220101.

[15] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery (ACM), 2001, pp. 57–66. doi: 10.1145/502512.502525.

[16] L. C. Freeman, "Centrality in social networks conceptual clarification," *Soc. Networks*, vol. 1, no. 3, pp. 215–239, 1978, doi: 10.1016/0378-8733(78)90021-7.

[17] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery (ACM), 2002, pp. 61–70. doi: 10.1145/775047.775057.

[18] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003, pp. 137–146. doi: 10.1145/956750.956769.

[19] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. Vanbriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007, pp. 420–429. doi: 10.1145/1281192.1281239.

[20] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 199–207. doi: 10.1145/1557019.1557047.

[21] R. K. Thakur, S. Gupta, R. Nigam, and P. K. Thiruvikraman, "Investigating the Hubble Tension Through Hubble Parameter Data," *Res. Astron. Astrophys.*, vol. 23, no. 6, 2023, doi: 10.1088/1674-4527/acd0e8.

[22] M. Kimura and K. Saito, "Tractable models for information diffusion in social networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2006, pp. 259–271. doi: 10.1007/11871637_27.

[23] S. Pangaonkar, R. Gunjan, and V. Shete, "Recognition of Human Emotion through effective estimations of Features and Classification Model," in *2021 International Conference on Computing, Communication and Green Engineering, CCGE 2021*, 2021. doi: 10.1109/CCGE50943.2021.9776405.

[24] A. Suman, P. Suman, V. Jaiswal, and S. Padhy, "Intelligent Hardware for Preventing Road Accidents Through the Use of Image Processing," in *ACM International Conference Proceeding Series*, 2023, pp. 313–321. doi: 10.1145/3607947.3608012.

[25] P. Sahane, S. Pangaonkar, and S. Khandekar, "Dysarthric Speech Recognition using Multi-Taper Mel Frequency Cepstrum Coefficients," in *2021 International Conference on Computing, Communication and Green Engineering, CCGE 2021*, 2021. doi: 10.1109/CCGE50943.2021.9776318.

[26] M. Streeter and D. Golovin, "An online algorithm for maximizing submodular functions," in *Advances in Neural Information Processing Systems 21 - Proceedings of the 2008 Conference*, 2009, pp. 1577–1584. Accessed: Feb. 22, 2024. [Online]. Available: https://proceedings.neurips.cc/paper/2008/hash/5751ec3e9a4feab575962e78e006250d-Abstract.html

[27] V. Jaiswal, V. Sharma, and D. Bisen, "Modified Deep-Convolution Neural Network Model for Flower Images Segmentation and Predictions," *Multimed. Tools Appl.*, 2023, doi: 10.1007/s11042-023-16530-3.

[28] S. Pangaonkar and R. Gunjan, "A consolidative evaluation of extracted EGG speech signal for pathology identification," *Int. J. Simul. Process Model.*, vol. 16, no. 4, pp. 300–314, 2021, doi: 10.1504/IJSPM.2021.118846.

[29] L. G. Valiant, "The Complexity of Enumeration and Reliability Problems," *SIAM J. Comput.*, vol. 8, no. 3, pp. 410–421, Aug. 1979, doi: 10.1137/0208032.

[30] K. Jain and V. V. Vazirani, "An approximation algorithm for the fault tolerant metric facility location problem," *Algorithmica (New York)*, vol. 38, no. 3, pp. 433–439, Dec. 2003, doi: 10.1007/s00453-003-1070-1.

[31] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 807–815. doi: 10.1145/1557019.1557108.

[32] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine," *Comput. Networks*, vol. 56, no. 18, pp. 3825–3833, 2012, doi: 10.1016/j.comnet.2012.10.007.