Influence Maximization in Dynamic Social Networks and Graphs

Gkolfo I. Smani, Vasileios Megalooikonomou

Abstract—Influence and influence diffusion have been studied extensively in social networks. However, most existing literature on this task are limited on static networks, ignoring the fact that the interactions between users change over time. In this paper, the problem of maximizing influence diffusion in dynamic social networks, i.e., the case of networks that change over time is studied. The DM algorithm is an extension of Matrix Influence (MATI) algorithm and solves the Influence Maximization (IM) problem in dynamic networks and is proposed under the Linear Threshold (LT) and Independent Cascade (IC) models. Experimental results show that our proposed algorithm achieves a diffusion performance better by 1.5 times than several state-of-the-art algorithms and comparable results in diffusion scale with the Greedy algorithm. Also, the proposed algorithm is 2.4 times faster than previous methods.

Keywords—Influence maximization, dynamic social networks, diffusion, social influence.

I. INTRODUCTION

RECENTLY, social networks are playing a fundamental role in information propagation, since more and more people prefer to publicize their views or ideas on the networks. One of the main research interests is to understand the way of influence and information spread in social networks. For example, a company wants to market a new product through the "word of mouth" effect in the social network. It wishes to find and persuade a small subset of users (seed users) to adopt the product so as to trigger a large number of further adoptions via social influence. At first, we need to understand the influence diffusion by answering questions such as, how to be selected the seed users so that the total number of triggered users to adopt the product can be maximized (a.k.a. influence maximization) [1], [9], [16], [17].

A natural problem for social influence is how to find the initial users that will eventually influence the largest numbers of users, which is known as influence maximization (IM). Given a social network G and an integer k, IM's goal is to select k seed users in G in hope that adopting a promoted product or idea can maximize the expected number of final adopted users through word-of-mouth effect [1], [7], [16]. Initially proposed by Kempe et al. [6], the problem of IM has been intensively studied by a number of subsequent projects, improvements, or modifications from multiple aspects, including estimation of influence size, adaptive seeding, boosting seeding, and many others.

The main task in IM lies in estimating the expected size of

G. I. Smani and V. Megalooikonomou are with the Department of Computer Engineering and Informatics, University of Patras, 26504 Greece (e-mail: gsmani@upatras.gr, vasilis@ceid.upatras.gr).

influenced users of each alternative seed set based on each user's activation probabilities, referring to the probability that a user successfully influences his social neighbors after being influenced. The influences among users are quantified by those activation probabilities [1], [16], [17]. While existing literature works well in finding the most influential seed users, they are all limited to the assumption that the number of nodes in the network, along with the edges between them, are fixed during influence diffusion. Consequently, it violates real practices as many realistic social networks usually develop over time.

In this paper, we study the problem of IM on dynamic social networks which are changing over time, and specifically under the LT and IC models. According to both, at any discrete time step a user can be either active or inactive (for example, has adopted the product or not) and the information propagates until no more users can be activated.

The main contributions of this work can be summarized as follows:
- The proposed DM algorithm is an efficient IM algorithm for dynamically changing networks.
- DM on large scale real-world graphs under the LT and IC models performs better than several alternative methods in terms of influence and computation time, and achieves comparable results to MC Greedy algorithm in terms of influence.

II. RELATED WORK

IM aims at a set of k users that maximize the influence spread over a network. Previous efforts on IM can be generally categorized into static methods and dynamic methods. In the case of static methods, there has been a vast amount of literature.

A Monte-Carlo simulation method is proposed by Kempe [6], which estimates $\sigma(S)$ repeating Monte-Carlo simulation, where S is the set of seed nodes and $\sigma(S)$ is the average number of infected vertices. Chen et al. [2] proposed prefix excluding maximum influence arborescence (PMIA) model to find seed vertices focusing on the paths with high information diffusion ratio. Chen et al. [3] also suggested Degree Discount based on node degree where the nodes which are adjacent to the selected node, are given penalty. When node v is selected as a seed node and u is its neighbor, it is possible that v propagates information to u, so selecting nodes other than u as seed nodes is better for information diffusion.

Two categories of methods have been proposed for the IM problem in dynamic networks: Monte-Carlo simulation-based methods and heuristic-based methods. The previous method is

World Academy of Science, Engineering and Technology
International Journal of Humanities and Social Sciences
Vol:16, No:11, 2022

proposed by Habiba and Berger-Wolf [5]. The method estimates the scale of propagation σ(·) by repeating Monte-Carlo simulation in the case of static networks. Since σ(·) is a monotonic and submodular function also in dynamic networks, this method achieves large-scale propagation [11], [12]. However, this method's computational cost is high as in static networks [11]. Osawa et al. [12], [14] proposed a heuristic method for calculating σ(·) faster. After $\sigma(S)$ is computed, seed nodes are obtained by greedy algorithm as in the method by Monte-Carlo simulation. Also, Murata and Koga [10] proposed three methods, Dynamic Degree Discount, Dynamic CI and Dynamic RIS, as extensions of static network methods to dynamic network methods, based on the node degree, the degree of distant nodes, and the reachable nodes, respectively.

III. PRELIMINARIES AND PROBLEM STATEMENT

A social network is typically modeled as a directed graph $G = (V, E)$, consisting of $|V|$ users represented as nodes and $|E|$ directed edges reflecting the relationship between users. An influence weight $p_{u,v} \in [0,1]$ is also associated with each edge $(u, v) \in E$, and represents the probability that a node u will affect node v. We assume that T (u)={τ₁, τ₂, …, τ_M} represents the set of all possible paths that exist in the graph starting from node u and leading to "leaf" nodes and are generated by the Depth-First Search (DFS) algorithm. Each path τ_i consists of a sequence of nodes: τ_i = {n_{i1}, n_{i2}, …, n_{iN}}. M is the number of all possible paths from a node u and N represents the number of the nodes and the index of the terminal node of path τ_i [15].

Let $p_{l,l+1}^{\tau}, 1 \le l \le N - 1$, represents the influence probability between two sequential nodes in path τ. Then F(τ_i) = {f_{i1}, f_{i2}...., f_{iN}} represents the cumulative probability path for every path τ_i starting from a node u to be active (i.e., a path is considered active if each one of its edges is active). Each f_{ij} is equal to $\prod_{l=1}^{j-1} p_{l,l+1}^{\tau_i}$ if $j \ge 1$, and 1 otherwise [15].

Let $\Psi(u,v) = \{\psi_1, \psi_2, ..., \psi_L\}$ denote the set of all passible unique paths from a node u to a node v, where each path $\psi_i = \{n_{i1}, n_{i2}, ..., n_{iN}\}$ consists of a sequence of nodes, L is the number of all possible paths between nodes u and v, and N represents the number of nodes of path ψ_i, with L≤M. Respectively, $\Phi(\psi_i) = \{\varphi_{i1}, \varphi_{i2}, ..., \varphi_{iN}\}$ denotes the probability for every path ψ_i between nodes u and v, and is calculated in the same way as f_{ij} [15].

Goyal et al. [4] showed that the spread of a set S of nodes is the sum of the spread of each individual $u \in S$ on the subgraphs induced by the set $V - S + u$:

$$\sigma(S) = \sum_{v \in V} \sum_X Pr[X]I(S, v, X) = \sum_{v \in V} A(S, v)$$
$$= \sum_{u \in S} \sigma^{V-S+u}(u) \qquad (1)$$

where X is a possible *live-edge* graph, Pr[X] is the sampling probability of X, I(S, v, X) is an indicator function equal to 1 if there is a live path in X from S to v and 0 otherwise, A(S,v) is the probability of the single node v to be influenced by S, and $\sigma^{V-S+u}(u)$ denotes the total influence of u in the subgraph

induced by the set $V - S + u$ $((V \backslash S) \cup \{u\})$.

Under the LT model, the calculation of influence after a node x addition to a set of nodes S is given by:

$$\sigma(S + x) = \sigma(S) + \sigma(x) - \sum_{y \in S} \Omega(x, y) - \sum_{y \in S} \Omega(y, x) \; (2)$$

under the IC model the following heuristics [13] are used:
1. for each path originating from node x or a node of seed set S, we keep the subpaths before finding a node of $S \cup \{u\}$.
2. $\sigma(S + x)$ is equal to the sum of the influence probabilities that correspond to each of these subpaths.

Also, the forward cumulative influence Ω(u,v) corresponds to the influence of node u to v and to the nodes that can be found right after v in the paths T(u) of node u.

In this paper, we model a dynamic social network $G = \{G^1, G^2, ..., G^T\}$ as a set of network snapshots evolving over time. We assume that the nodes remain the same while the edges in each network snapshot change through time. This is used as assumption in other papers as well [3], [11]. Each snapshot graph $G^t = (V, E^t)$ is modeled as a directed network which includes edges appearing during time $t = 1,2, ..., T$. Moreover, an influence weight $p_{u,v}^t \in [0,1]$ is also associated with each directed edge $(u, v) \in E^t$, and represents the probability of node u to influence node v at time t.

Our goal is to discover a set of seed sets, $S^t, t = 1,2, ..., T$, whose size is k, such that it maximizes the influence $\sigma(S^t)$.

Table I presents the notations used in this paper.

TABLE I
NOTATIONS

Notation	Description
p$^t_{u,v}$	Influence weight on directed edge (u,v) at t time-step
σ(S^t)	Influence of a set of nodes St to the graph
A^t(u,v)	Influence of node u to node v at time-step t
Ω^t(u,v)	Forward cumulative influence of node u to node v, at time-step t
T ^{t}(u)={τ$^t_1$,τ$^t_2$, … ,τ^t_M}	Set of all possible paths starting from node u, at time-step t
τ^t_i = {n_{i1}, n_{i2}, …, n_{iN}}	Path consisting of N nodes starting from node u, at time-step t
F (τ^t_i) = {f_{i1}, f_{i2}, …., f_{iN}}	Cumulative probability path for path τi, at time-step t
Ψ^t(u,v)={ψ$^t_1$, ψ$^t_2$, …. ,ψ^t_L}	Set of all possible paths between nodes u and v, at time-step t
ψ^t_i={n_{i1}, n_{i2}, …., n_{iN}}	Path between nodes u and v, at time-step t
Φ^t(ψ$_i$)={φ_{i1}, φ_{i2},….,φ_{iN}}	Cumulative probability path for path ψti , at time-step t

IV. PROPOSED METHOD

In this section, we presented a method for the IM Problem in Dynamic Networks: the DM algorithm. The proposed method is an extension of static method MATI [14]. Based on the functions A, Ω and influence function σ, we use functions A^t of activation probability and the forward cumulative influence Ω^t at time stamp t, for $t = 1,2, ..., T$.

World Academy of Science, Engineering and Technology
International Journal of Humanities and Social Sciences
Vol:16, No:11, 2022

Algorithm 1: DM LT

Input: G^0, k, T ▷k = number of seed nodes, T = max time-step
Initialize: $S^t = 0, \forall t = 1,2,\ldots,T$
for $t = 0$ to T do
 if $t == 0$ then
   Calculate A^t, Ω^t
 Calculate Q^t
 for $i = 1$ to k do
   $s, \sigma(s) = Q^t.top()$
   $S^t = S^t \cup \{s\}$
   $U = V \backslash S^t$
   for each $u \in U$ do
       $\sigma(u) = Q^t(u)$
       for each $v \in S^t$ do
               $\sigma(u) -= \Omega^t(v,u)$
               $\sigma(u) -= \Omega^t(u,v)$
               $Q^t.add((u, \sigma(u)))$
   Update G^t to G^{t+1}
   Update A^t, Ω^t
return $S = \{S^1, S^2, \ldots, S^T\}$

Algorithm 2: DM IC

Input: G^0, k, T ▷k = number of seed nodes, T = max time-step
Initialize: $S^t = 0, \forall t = 1,2,\ldots,T$
for $t = 0$ to T do
 if $t == 0$ then
   Calculate A^t
 Calculate Q^t
 for $i = 1$ to k do
   $s, \sigma(s) = Q^t.top()$
   $S^t = S^t \cup \{s\}$
   $U = V \backslash S^t$
   for each $u \in U$ do
       $\sigma(S^t + u) = |S^t \cup \{u\}|$
       $\sigma(u) += influence(S^t)$
       $\sigma(u) += influence(S^t \cup \{u\})$
       $Q^t.add((u, \sigma(S^t \cup u) - \sigma(S^t)))$
   Update G^t to G^{t+1}
   Update A^t
return $S = \{S^1, S^2, \ldots, S^T\}$

We name these algorithms DM LT (Dynamic MATI under LT), shown in Algorithm 1, and DM IC (Dynamic MATI under IC) shown in Algorithm 2.

V. EXPERIMENTS

In this section, we present results of experiments conducted on real-world dynamic networks to test the performance of proposed algorithms. The dynamic networks we used in the experiments are listed below.

- EmailEUCore: The network was generated using email data from a large European research institution. It includes anonymized information about all incoming and outgoing email between members of the institution. The e-mails represent communication only between institution members and not with of the rest of the world. A directed edge (u, v, t) means that person u sent an email to person v at time t [19].

- Email-dnc: The directed network of emails in the 2015 Democratic National Committee email leak. Each edge in the dataset denotes that an email has been sent from a person to another person [18].

- High school dynamic contact networks: This dataset contains the temporal network of contacts between students in a high school in Marseilles, France. In case of multiple active contacts in a given interval, multiple lines start with the same value of time t which is measured in seconds [20].

- Primary school temporal network data: This data set contains the temporal network of contacts between the children and teachers. In case of multiple active contacts in a given interval, multiple lines start with the same value of time t which is measured in seconds [20].

- Hospital ward dynamic contact network: This dataset contains the temporal network of contacts between patients, 29 patients and 46 health-care workers (HCWs) and among HCWs in a hospital ward in Lyon, France, from December 6, 2010 at 1:00 pm to December 10, 2010 at 2:00 pm. If multiple contacts are activated in a given interval, we see multiple lines starting with the same value of t. Time is measured in seconds [20].

- CollegeMsg: This dataset is comprised of private messages sent on an online social network at the University of California, Irvine. Users could search the network for others and then initiate conversation based on profile information. An edge (u, v, t) means that user u sent a private message to user v at time t [19].

- WikiTalk: This is a temporal network representing Wikipedia users editing each other's Talk page. A directed edge (u, v, t) means that user u edited user v's talk page at time t [19].

Table II shows the number of nodes and edges of each dataset.

TABLE II
DATASETS

Datasets	Nodes	Edges
EmailEuCore	986	12216 (Temporal), 24929 (Static)
Email-dnc	92	9800
High school dynamic contact network	327	188,508
Primary school temporal network data	242	125,775
Hospital ward dynamic contact network	75	32,424
CollegeMsg	1899	59835 (Temporal), 20296 (Static)
WikiTalk	1140149	7833140 (Temporal), 3309592 (static)

A. Evaluation

We compared the performance of the proposed algorithms with the following ones:
- Dynamic Degree Discount [8]
- Dynamic CI [8]
- Dynamic RIS [8]
- Osawa [10]
- MC Greedy [3]

The results of information propagation for different seed

World Academy of Science, Engineering and Technology
International Journal of Humanities and Social Sciences
Vol:16, No:11, 2022

sizes k are shown in Fig. 1 with fixed threshold $\theta = 0.1$. The x-axis shows the size of the seed set, while the y-axis shows the number of propagated vertices. Values of x-axis are $\frac{k}{|V|} *$ 100, i.e., the percentage of seed vertices to all vertices in the network. Values of y-axis are $\frac{\sigma(S)}{|V|} * 100$, i.e., the percentage of influence $\sigma(S)$ to all vertices in the network. As shown in

Fig. 1, MC Greedy achieves the highest diffusion in all datasets. Diffusion of the proposed methods are inferior by 5% compared to MC Greedy, although it is better than the other ones (Dynamic Degree Discount, Dynamic CI, Dynamic RIS, Osawa). Specifically, DM - IC performs higher diffusion by 1.5% than several state-of-the-art algorithms.

(a) Email dnc   (b) EmailEuCore   (c) CollegeMsg

(d) WikiTalk   (e) Primary School
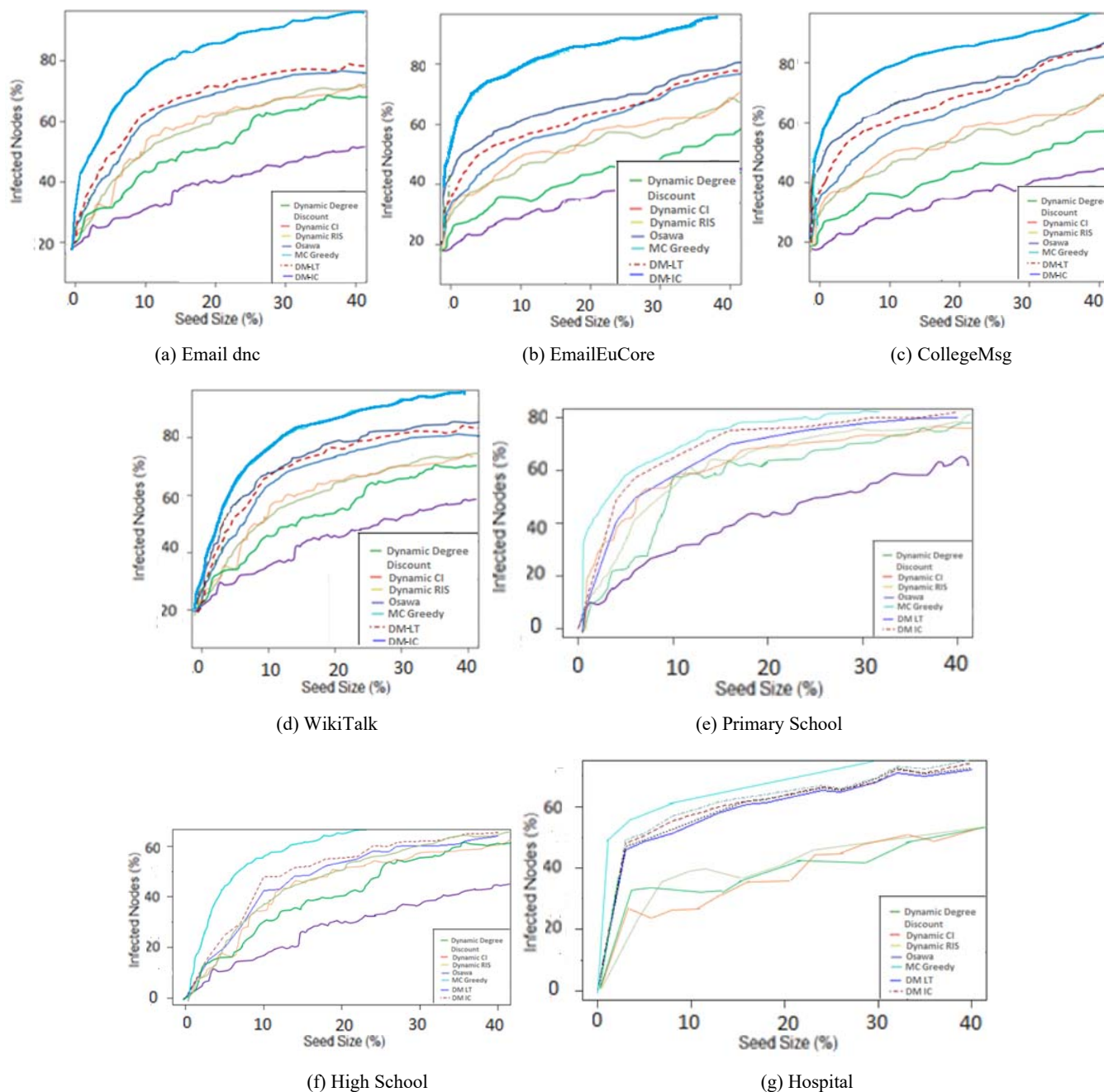
(f) High School   (g) Hospital

Fig. 1 Comparison of DM-LT and DM-IC with state-of-the-art algorithms in terms of % of infected nodes with varying seed size for different datasets

Fig. 2 shows the computational time needed for θ set to 0.1 and for varying seed nodes. X-axis shows the size of seed

vertices and y-axis shows the computational time in log-scale.

World Academy of Science, Engineering and Technology
International Journal of Humanities and Social Sciences
Vol:16, No:11, 2022

(a) Email dnc   (b) EmailEuCore   (c) CollegeMsg

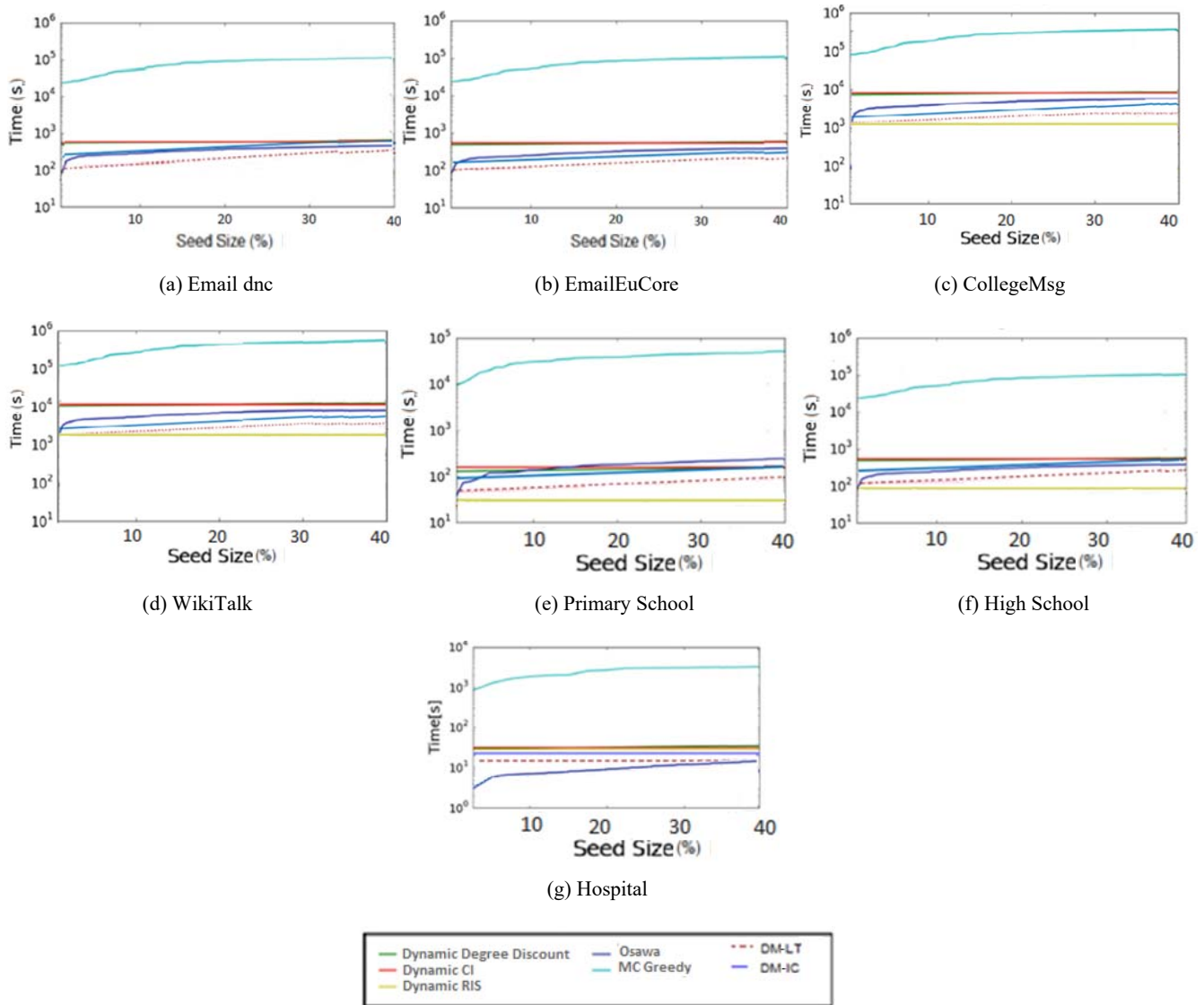(d) WikiTalk   (e) Primary School   (f) High School

(g) Hospital

Fig. 2 Comparison of DM-LT and DM-IC with state-of-the-art algorithms in terms of computation time when the seed size k changes for different datasets

Fig. 2 shows that for all datasets, while all other methods including the proposed ones can compute seed vertices in realistic time, MC Greedy needs several hours to compute seed vertices. This shows that MC Greedy is intractable in realistic time for large scale networks. The computational time of proposed methods, DM LT and DM-IC, is about the same for the Primary School dataset. Regarding the comparison with proposed methods, Dynamic RIS and Dynamic Degree Discount, except the Primary School dataset, where Dynamic RIS is faster and the Hospital dataset, where Dynamic Degree Discount and DM-IC are almost the same, DM-IC is faster than them or the same.

VI. CONCLUSION

We proposed DM, a method for the IM problem on dynamic networks which extends methods in static networks. Based on the experiments performed for comparing with previous methods, the proposed method performs better in terms of diffusion, with the exception of MC Greedy which achieves better diffusion. However, computational time of the proposed method is better by 5% than MC Greedy. As future work, it is planned to investigate the IM problem in dynamic networks in case of different community subgraphs, as well as the case of partially observed dynamic graphs. In addition, we intend to study the IM problem in dynamic networks using other models.

REFERENCES

[1] Borgs C., Brautbar M., Chayes J., and Lucier B., Maximizing social influence in nearly optimal time. In SODA, pp. 946-957, 2014.
[2] Chen S., Wang C., Wang Y., Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining -KDD 2010, pp. 1029-1038, 2010.
[3] Chen S., Wang Y., Yang Y., Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining -KDD 2009, pp.

World Academy of Science, Engineering and Technology
International Journal of Humanities and Social Sciences
Vol:16, No:11, 2022

199-207, 2009.

[4] Goyal Amit, Wei Lu, and Laks V. S. Lakshmanan. "SIMPATH: An Efficient Algorithm for Influence Maximization Under the Linear Threshold Model." In: ICDM. 2011.

[5] Habiba and Berger-Wolf TY. Maximizing the extent of spread in a dynamic network. DIMACS Technical Report, 2007.

[6] Kempe D., Kleinberg J., and Tardos E. Maximizing the spread of influence through a social network. In KDD, pp. 137-146, 2003.

[7] Li G., Chen S., Feng J., Tan K.-L., and Li W.-S. Efficient location-aware influence maximization. In SIGMOD, pp. 87-98, 2014.

[8] Li H., Bhowmick S. S., Sun A., and Cui J. Conformity-aware influence maximization in online social networks. The VLDB Journal, 24(1):117-141, 2015.

[9] Li Y., D. Zhang, and K.-L. Tan. Real-time targeted influence maximization for online advertisements. PVLDB, 8(10):1070-1081, 2015.

[10] Murata T., Koga H. Extended methods for influence maximization in dynamic networks. Comput Soc Netw 5, 8, 2018.

[11] Nguyen H. T., Thai M. T., and Dinh T. N. Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks. In SIGMOD, pp. 695-710, 2016.

[12] Osawa S, Murata T. Selecting seed nodes for influence maximization in dynamic networks. In: Proceedings of the 6th workshop on complex networks (CompleNet 2015), studies in computational intelligence. Berlin: Springer; pp.91–8, 2015

[13] Rong-Hua Li, Jeffrey Xu Yu, and Rui Mao. "Efficient core maintenance in large dynamic graphs." In: IEEE Transactions on Knowledge and Data Engineering 26.10, pp. 2453–2465, 2014.

[14] Rossi M., Shi B., Tziortziotis N., Malliaros F., Giatsidis C., Vazirgiannis M., MATI: An efficient algorithm for influence maximization in social networks. PLoS ONE. 13. 1. 10.1371/journal.pone.0206318, 2018.

[15] Tang Y., Shi Y., and Xiao X. Influence maximization in near-linear time: A martingale approach. In SIGMOD, pp. 1539-1554, 2015.

[16] Tang Y., Xiao X., and Shi Y. Influence maximization: near-optimal time complexity meets practical efficiency. In SIGMOD, pp. 75-86, 2014.

[17] Wang X., Zhang Y., Zhang W., and Lin X. Distance-aware influence maximization in geo-social network. In ICDE, pp. 1-12, 2016.

[18] http://networkrepository.com

[19] https://snap.stanford.edu/data

[20] http://www.sociopatterns.org