Debunking the Myths of Influence Maximization

Akhil Arora[1*] Sainyam Galhotra[2] Sayan Ranu[3]
[1]Text and Graph Analytics, Xerox Research Centre India, Bangalore, India
[2]College of Information and Computer Science, University of Massachusetts, Amherst, USA
[3]Dept. of Computer Science and Engineering, Indian Institute of Technology, Delhi, India
[1]akhil.arora@xerox.com [2]sainyam@cs.umass.edu [3]sayanranu@iitd.ac.in

ABSTRACT

Influence maximization (IM) on social networks is one of the most active areas of research in computer science. While various IM techniques proposed over the last decade have definitely enriched the field, unfortunately, experimental reports on existing techniques fall short in validity and integrity since many comparisons are not based on a common platform or merely discussed in theory. In this paper, we perform an in-depth benchmarking study of IM techniques on social networks. Specifically, we design a benchmarking platform, which enables us to evaluate and compare the existing techniques systematically and thoroughly under identical experimental conditions. Our benchmarking results analyze and diagnose the inherent deficiencies of the existing approaches and surface the open challenges in IM even after a decade of research. More fundamentally, we unearth and debunk a series of incorrect claims made by highly cited papers in the field of IM. Overall, this study establishes that there is no single state-of-the-art technique in IM. At best, a technique is the state of the art in only one aspect.

1. INTRODUCTION

Social networks have become an integral part of our day-to-day lives. We rely on Facebook and WhatsApp to communicate with friends. Twitter is regularly used to disseminate information such as traffic news, emergency services, etc. This reliance on social networks has resulted in wide-spread research in finding solutions to the *influence maximization (IM)* problem [2]. In a social network, each user corresponds to a node and two users are connected through an edge if they *interact*. Interaction between two users may depict friendship, such as in FaceBook, *following* a user, such as in Twitter, or co-authorship of scientific articles, such as in DBLP. Generally, it is assumed that an user u can *directly* influence user v if u interacts with v, i.e., there is an edge from u to v. For example, u positing a positive review on a movie may result in v actually watching the movie. This event may in turn result in v influencing his/her own friends. In other words, indirectly, u can influence any user x of the network if there is a path from u to x. The IM problem is to identify a set of *seed nodes* (or users) so that the total number of users influenced is maximized.

Kempe et al. [2] in their seminal work established that finding an optimal solution for IM is NP-Hard and were the first to prove that a simple GREEDY algorithm can provide the best approximation guarantee in polynomial time. They incorporated the use of three *diffusion models – Independent Cascade* (IC), *Weighted Cascade* (WC) and *Linear Threshold* (LT) for information propagation, which have been almost exclusively followed in majority of the subsequent work. All these models are essentially a function of the edge weights in the social network. The higher the edge weight between u and v, more is the influence of u on v.

Since Kempe et al.'s seminal work [2], almost every year, a new IM technique has been published that claim to be the state-of-the-art. Without doubt, this extensive research has promoted prosperity of the family of IM techniques. However, it also raises several questions that are not adequately addressed. *How to choose the most appropriate IM technique in a given specific scenario? What does it really mean to claim to be the state-of-the-art? More fundamentally, are the claims made by the recent papers true?* To ensure a streamlined growth of the field, it is critical to benchmark the existing techniques in a unified setup across common datasets and answer all of the above questions. We conduct this benchmarking study and firmly establish that several claims from highly cited papers are incorrect, the sampling based evaluation procedure adopted by various techniques could produce highly spurious results, and expose a series of *myths* that could potentially alter the way we approach IM research. For details, we kindly refer the reader to the full version [1] of this work[1].

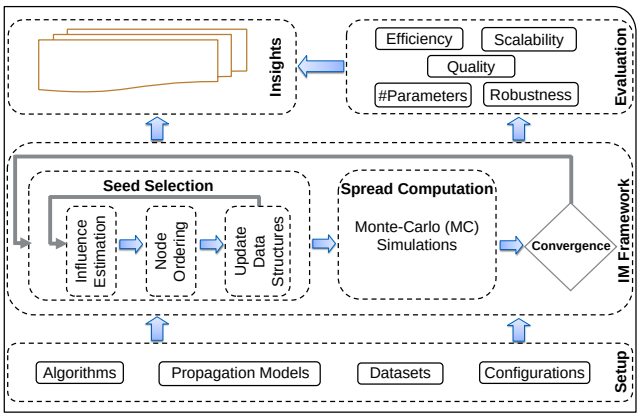


Figure 1: **The proposed benchmarking framework for IM.**

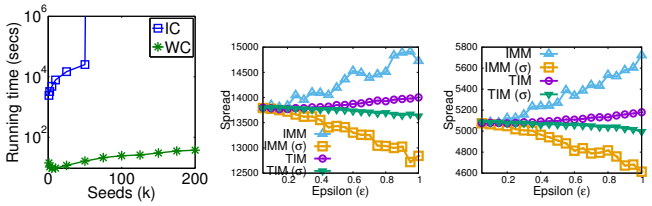
2. BENCHMARKING IM ALGORITHMS

In this paper, we have designed a systematic benchmarking platform for the IM problem. As visible in Fig. 1, the benchmark consists of four core components: (1) **Setup**, including a set of algorithms, real-world datasets, parameter configurations and a diffusion model; (2) **IM Framework**, a generalized IM module with high abstraction of the common workflow of Influence Maximization (details in [1]); (3) **Evaluation**, which provides targeted diagnoses on these algorithms based on our framework, leading to directions of

*The first two authors have contributed equally to this work.

[1]https://www.dropbox.com/s/uzi53ybpbfbdo9s/SIGMOD17_im_benchmarking.pdf?dl=0

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
NEDB'17 MIT, Boston, USA
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.



(a) Orkut (b) DBLP (WC) (c) HepPh (LT)

Figure 2: **(a) Running time of IMM ($\epsilon = 0.5$) under IC ($W(u,v) = 0.1$) and WC. (b-c) Comparison of spreads reported by IMM and TIM$^+$ (denoted as IMM and TIM) with the spread obtained through the classical way of MC simulations (denoted as IMM(σ) and TIM$^+$(σ)) against ϵ.**

improvement over the existing work; and (4) **Insights**, which discusses the key take-away points from the benchmarking study and generally, summarizes the state of the IM field after more than a decade of research.

Using the proposed generic benchmark, we evaluate eleven state-of-the-art IM algorithms across diffusion models, datasets, and parameters. In this process, we have curated the most comprehensive publicly accessible code base for IM algorithms, which can be downloaded from `http://bit.ly/2a9eoo9`. Next, we discuss the key insights resulting from our benchmarking study.

3. MYTHS: AN OVERVIEW

To highlight the ambiguity that plagues the current maze of IM techniques, we provide some concrete examples and debunk several myths (two of which are detailed below while the rest are elaborated in [1]) that have propagated either due to false claims or poorly conducted experiments.

- **WC is not IC:** WC is a specialized instance of the IC diffusion model and not IC itself (differences detailed in [1]). Multiple techniques have claimed to scale well under IC, while in reality, they scale only for WC. To provide a concrete example, consider Fig. 2a, where IMM scales well for the WC model but fails to compute the information spread even for 50 seeds within a reasonable time limit. In fact, it crashes on our machine beyond 50 seeds on the Orkut dataset, while consuming more than 256 GB of RAM.
  While IMM is just one representative technique, several techniques equate WC with IC and incorrectly claim to perform well under the IC diffusion model. As shown in [1], most of the sampling based methods that exploit the idea of *reverse reachability sets* do not scale under IC due to *exorbitantly* high memory-footprint.
- **High spread can be obtained even at high ϵ for TIM$^+$ and IMM:** The spreads reported in both papers are inflated due to the procedure followed to compute them. More specifically, instead of directly computing the spread through MC simulations, TIM$^+$ and IMM extrapolate the influence of the seed nodes on the sampled sub-network over the entire network to compute the spread. Mathematically, let R be the nodes reachable from at least one of the seed nodes on the sampled network, M be the number of sampled nodes, and N be the number of nodes on the entire network. The spread is approximated as $\frac{R}{M} \times N$.
  The consequence of computing the spread through extrapolation is shown in Figs. 2b-2c. As can be seen, the extrapolated spread is significantly higher from the actual expected spread computed through MC simulations. More critically, the extrapolated spread improves with increase in ϵ and goes against the theorem proved by the authors of TIM$^+$ and IMM themselves, which states that the spread is expected to improve with decrease in ϵ. On the other hand, the spread computed by MC simulations follows the theoretical expectations. Thus, it is safe to say that the published spreads are incorrect and inflated.
- **CELF++ is not strictly faster than CELF**.
- **CELF (or CELF++) with 10K MC simulations is not the gold standard for quality**.

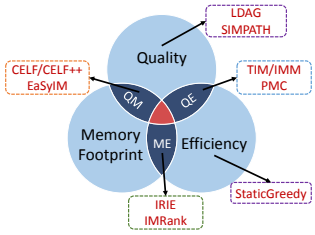


Figure 3: **Summarizing the spectrum of Influence Maximization (IM) techniques based on their strengths.**

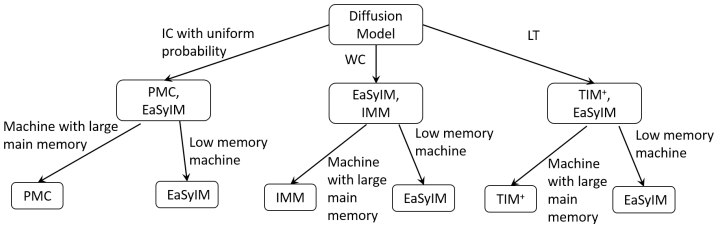


Figure 4: **The decision tree for choosing the most appropriate IM algorithm.**

- **IMM is not strictly faster than TIM$^+$.**
- **SIMPATH is wrongly claimed to be faster than LDAG.**
- **Convergence criteria of IMRank is wrong**.

4. CONCLUDING INSIGHTS

To summarize, a good algorithm for IM stands on three pillars: quality of spread, running time efficiency, and main memory footprint. In addition, it is desirable for the technique to be robust across diffusion models, datasets, and parameters. We benchmark the eleven most promising techniques across all of these features. Fig. 3 summarizes the results. Notice that there is no technique that stands strong on all three pillars. In other words, there is no single state-of-the-art technique for IM.

Several techniques exist that stand on two pillars. Among these, techniques that lie in the "ME" category (memory + efficiency) do not provide a good solution since ensuring quality is of utmost importance. Consequently, in practical scenarios, the choice of the best IM technique is between those that lie in the "QM" and "QE" categories. Towards that goal, Fig. 4 presents the decision tree for choosing the best IM technique given the task and resources at hand. In terms of quality, TIM$^+$, IMM and PMC provide the best spread. These three techniques also provide the fastest performance under LT, WC and IC with uniform weights respectively. Thus, if main memory budget is not a constraint, the choice is between these three techniques. When main memory is scarce, EaSyIM, CELF, CELF++ and IRIE provide alternative solutions. Among these, EaSyIM easily out-performs the other three techniques in memory footprint, while also generating reasonable quality and efficiency. Overall, the choice is between four techniques: IMM, TIM$^+$, EaSyIM, and PMC. Here, we note that a highly promising technique has been published in SIGMOD 2016 [3]. Unfortunately, we could not include the technique in our study due to how recently it is published. Benchmarking any active field of research always risks such issues. Nonetheless, the insights obtained from this study provides the urgent directionality and clarity required to arrest the propagation of erroneous claims and the resultant haphazard growth. We hope our discoveries would lead to a more streamlined advancement in IM research.

5. REFERENCES

[1] A. Arora, S. Galhotra, and S. Ranu. Debunking the Myths of Influence Maximization: An In-Depth Benchmarking Study. In *SIGMOD*, 2017.
[2] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.
[3] H. T. Nguyen, M. T. Thai, and T. N. Dinh. Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks. In *SIGMOD*, pages 695–710, 2016.