Influence Maximization Revisited: The State of the Art and the Gaps that Remain

Akhil Arora
EPFL Lausanne
akhil.arora@epfl.ch

Sainyam Galhotra
UMass Amherst
sainyam@cs.umass.edu

Sayan Ranu
IIT Delhi
sayanranu@cse.iitd.ac.in

ABSTRACT

The steady growth of graph data from social networks has resulted in wide-spread research on the influence maximization (IM) problem. This results in extension of the state-of-the-art almost every year. With the recent *explosion* in the application of IM in solving *real-world* problems, it is no longer a theoretical exercise. Today, IM is used in a plethora of real-world scenarios, with *OnePlus*[1] series of mobile phones, *Hokey Pokey*[2] ice-creams, and *galleri5 influencer marketplace*[3] being the most prominent industrial use-cases. Given this scenario, navigating the *maze* of IM techniques to get an in-depth understanding of their utilities is of *prime* importance. In this tutorial, we address this *paramount* issue and solve the *dilemma* of "*Which IM technique to use and under What scenarios*"? "*What does it really mean to claim to be the state-of-the-art*"?

This tutorial builds upon our benchmarking study [1], and will provide a concise and intuitive overview of the most important IM techniques, which is usually lost in the technical literature. Specifically, we will unearth a series of *incorrect claims* made by prominent IM papers, disseminate the inherent *deficiencies* of existing approaches, and surface the *open challenges* in IM even after a decade of research.

1 MOTIVATION

Influence maximization (IM) has been one of the most actively studied areas of data management research over the past decade. With this, almost every year, a new IM technique has been published that claims to be the state-of-the-art. However, IM is no longer a theoretical problem. We rely on *Facebook* and *WhatsApp* to communicate with friends. *Twitter* is used to disseminate information such as traffic-news, emergency-services, etc. IM is used by companies to publicize their products or shape opinions (Ex: *OnePlus*, *galleri5*, and *HokeyPokey* [19]).

On the academic front, researchers are interested in classical IM [3, 7–10, 16–18, 21, 24, 25, 31] as well as more application-specific models such as IM under *competition* [22], *time and opinion-aware* IM [6, 12] etc. Undoubtedly, this extensive research has promoted prosperity of the family of IM techniques. However, it also raises several questions that are not adequately addressed. Given this widespread applicability, it is important to understand the following questions from a *neutral standpoint*.

[1]https://oneplusstore.in
[2]http://www.hokeypokey.in
[3]https://galleri5.com/aboutus

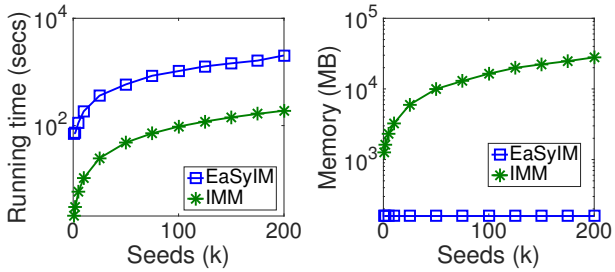

Figure 1: Comparing IMM ($\epsilon = 0.5$) with EaSyIM ($iter = 100$) under IC ($W(u, v) = 0.1$) on the YouTube dataset.

- *Which IM technique should one use given the resources in hand? How to choose the most appropriate IM technique in a given specific scenario?*
- *What does it mean to claim to be the state-of-the-art? More fundamentally, Is there really a single state-of-the-art technique as is often claimed?*
- *Are the claims made by the recent papers true?*
- *What are the unsolved challenges in the field?*

To highlight the ambiguity that plagues the current maze of IM techniques, we provide a concrete example[4] to motivate the need for answering the questions stated above.

What does it mean to be the state of the art? While many techniques claim to be the state of the art, in reality, they are often the state of the art in only one aspect of the IM problem. Consider Figs. 1a-1b, where EaSyIM [12] and IMM [30] scale better with respect to memory and running time respectively. Thus, neither technique can be termed as better than the other.

1.1 Relevance and Timeliness

• First, EDBT is an appropriate platform to present a tutorial on IM from a *neutral standpoint* since reproducibility tests and benchmarking have always been a key area of interest of the database community at large. Moreover, as shown in Fig. 2, of late database conferences have become the venue of choice for authors conducting research in the field of influence maximization, since ensuring scalability (while maintaining quality guarantees) has become central to the problems identified in this area.

• Second, to ensure a streamlined growth of the field, this tutorial, in addition to surveying existing IM techniques, serves as a *timely* and *relevant* avenue to *disseminate answers of the questions stated above to the data management community*. Overall, the tutorial will build upon our benchmarking study [1] and present our musings over almost a decade long literature (along with the recent advances) in the field of IM from top publication venues, and *unravel* many interesting and unknown avenues in the well-studied area of influence maximization.

[4]A detailed analysis on more such examples may be found in [1].

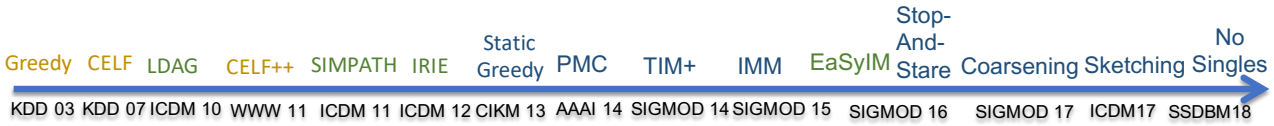

Figure 2: Timeline showcasing the Evolution of IM techniques.

• Lastly, as mentioned previously, IM is no longer a theoretical problem. It is regularly used by companies to publicize their products or shape opinions. *OnePlus*, *galleri5*, and *HokeyPokey* [19] for example rely completely on IM through social networks. To this end, the insights presented in this tutorial would definitely be advantageous to a broad audience at EDBT/ICDT ranging from theorists to researchers who are more interested in understanding and harnessing the practical power of IM in social networks.

2 INTENDED AUDIENCE, PREREQUISITE KNOWLEDGE AND LENGTH

The tutorial is aligned to the general area of data management and the web, thereby being relevant for a broad audience at EDBT: including students, academic researchers, and industrial experts specifically interested in benchmarking, data mining, social-network analysis, large-scale analytics, and performance tuning. No prior knowledge beyond basic probability and graph theory is expected. Familiarity with information-diffusion concepts would help, but not needed. The tutorial is self-contained and possesses introduction of most of the foundational concepts.

The key take away would be knowledge of the gaps including mis-claims and myths, leading to some of the never unraveled aspects of the IM problem, thereby enabling a more streamlined advancement in IM research. Since IM is a hot topic, we expect around 50 participants.

3 OUTLINE OF THE TUTORIAL

3.1 Introduction (20 minutes)

The first part of the tutorial will involve the formal definition of the IM problem along with an analysis of the fundamental information diffusion models: IC and LT [18]. The other aspect here would be to motivate the importance of IM, by citing real-world applications, in order to bridge the gap between theoretical models and real-world information diffusion.

Moving ahead, we will explain in detail the various challenges faced in designing effective solutions for the IM problem. We will analyze various properties of IM, namely – NP-hardness, submodularity etc., and also present the scenarios where exact estimation of influence is possible.

3.2 Summary of IM Algorithms (30 minutes)

First, we will present a categorized overview [1] of existing IM algorithms. This will enable the attendees to grasp the broad spectrum of IM techniques as portrayed in Table 1 in an intuitive and concise manner. Next, we will delve into a detailed description of each category.

Given that the IM problem is NP hard, Kempe et al. [18] leverage submodularity to propose a Greedy algorithm that provides the best approximation on the *quality* of obtained *spread*. Later, CELF [21] and CELF++ [15] were proposed to maintain the same quality of *spread* with an attempt to improve the efficiency by applying several optimizations over the Greedy algorithm.

Next, we will present the *heuristics* IMRank/IRIE [8, 17] and LDAG/SIMPATH [7, 16] that improve the *efficiency* and *scalability* aspect of IM, and perform well for the WC and LT models respectively. The caveat with these techniques is that they work well in practice, however lack any theoretical backing on the *quality* of the obtained *spread*. We will also introduce our techniques – *ASIM* [13] and *EaSyIM* [12], which are better both empirically and theoretically when compared to other heuristics.

Lastly, we will present a recent class of techniques that use *sampling* [9, 24, 25, 30, 31] to portray superior *efficiency* while retaining *quality* guarantees. These techniques either maintain reverse reachable (RR) sets of nodes or snapshots of cascades, and try to estimate influential nodes by sampling nodes from the original network. The caveat here is that most of these techniques are not scalable owing to their *exorbitantly* high memory footprint [1].

Table 1 summarizes the techniques discussed above, while stating their key highlights and the respective state-of-the-arts. This will enable the attendees to understand the representative techniques in the literature and the different aspects they address. It will also facilitate the attendees to appreciate as to why "*One Size Doesn't Fit All!*".

Finally, we will analyze why there does not exist any algorithm capable of simultaneously excelling in all the three fronts: (1) efficiency, (2) scalability, and (3) quality?

3.3 Myths, Mis-Claims and Insights (20 minutes)

In continuation to the overview of the techniques, here, we present our findings and provide recommendations to answer the question(s) posed by us in Section 1. We firmly establish that several *claims* from highly cited papers are *incorrect* (our experiments have been marked *SIGMOD Reproducible*), the evaluation procedure adopted by various techniques could produce misleading results, and expose a series of *myths* that could potentially alter the way we approach IM research. All the *insights* would be supported by *empirical* results, presented to the audience using a *python interface* to the publicly available implementations[5] of the discussed techniques.

3.4 Open Challenges and Future Directions (20 minutes)

The last part of the tutorial would focus towards *summarizing* the key insights discussed previously to eventually *shortlist* the best technique(s) and the corresponding scenarios in which they are the best. To this end, a decision-tree (Fig. 3b) will be presented as a tool to the audience. Next, we would delve into a detailed discussion on the open

[5]For details please visit our project page: https://sigdata.github.io/infmax-benchmark

Type	Theoretical Guarantee?	Highlights	State-of-the-Art
Greedy and Optimizations	Yes	Superior Quality and Scalability but low Efficiency	CELF/CELF++ [15, 21]
Heuristics	No	Superior Efficiency and Scalability at the cost of Quality	*EaSyIM*, IRIE, & LDAG/SIMPATH [7, 12, 16, 17]
Sampling Snapshots/RR sets	Yes	Superior Efficiency and Quality at the cost of Scalability	PMC, Stop-and-stare, Coarsening, Sketching, and NoSingles [23–27]

Table 1: The spectrum of IM techniques.

challenges in the field of information propagation, thereby providing a streamlined view of future research directions.

- The most important research direction is the development of a *scalable* and *efficient* algorithm with error *guarantees* (Fig. 3a), which still remains as the holy grail of influence maximization. While recent efforts by Popova et al. [27], Ohsaka et al. [26], and Nguyen et al. [23] are steps in this direction to improve scalability of the class of memory-intensive sampling algorithms [10, 24, 30, 31], more work is needed to achieve true scalability. To this end, there is a need for a generic framework inspired by classical data management systems which are shown to perform well for managing graph data [11], or the use of modern data management technologies that rely on distribution and parallelization to improve scalability and efficiency.
- Another compelling and novel research direction lies in validating the correctness/effectiveness of the classical information diffusion models proposed in [18] using real-world social media data capturing cascades from retweets or mentions as ground truth. The advantage of this exercise would be two fold: (1) Exploring the most unfathomed area in the field provides tremendous scope for advancement of the state-of-the-art, and (2) Curating a benchmark dataset for all the follow-up research. Such efforts would also attract further research on scalably learning influence probabilities from real-world interaction data extending on the works of [14] and [20].
- The challenges involved in scaling up influence maximization to massive networks under classical information diffusion models also cascade to the recent research activities around development of sophisticated diffusion models like opinion-aware [12], topic-aware [5] etc., which too are deprived of scalable algorithms. To this end, there is a need for devising a generic and unified framework for scaling up influence maximization under classical and various sophisticated real-world scenarios.

4 RELATED TUTORIALS

Multiple tutorials have been presented in the broad field of information propagation and IM at major data-centric venues [2, 4, 28, 29]. However, all these tutorials possess a common theme, i.e., each of them have provided an overview of models for information diffusion in networks and associated algorithms for influence analysis. While [2, 4, 29] were based on an algorithmic and data-mining perspective of the broad area of information diffusion, [28] focused on machine learning methods, specifically encircling the problems of network inference, influence estimation and control. In sum, the main focus of these tutorials was towards dissemination of the mathematical, technological, and algorithmic innovations to all-and-sundry, thereby enabling a step forward for sound analysis of research problems in the field of information propagation.

The proposed tutorial has the following key differentiations:

- First, the state-of-the-art has never been discussed from a neutral standpoint. More fundamentally, all the previous tutorials have given overviews of the existing literature, however, this exercise has not been from a perspective of a critic. We are the first to present some of the *highly controversial and ground-breaking discoveries*, thereby unraveling several *mis-claims and myths* in the existing IM research. This in-depth focus of our tutorial enables a more streamlined advancement in IM research with possible redefinition of the state-of-the-art.
- Second, IM is still a niche problem and provides many avenues to devise real world models and scalable algorithms capable of tackling competitive, time aware, opinion aware settings and many more. The knowledge gathered from this tutorial would facilitate more informed extensions to these settings.

5 TUTOR BIOGRAPHY AND EXPERTISE

Akhil Arora is a doctoral researcher at EPFL. His research interests include large scale data mining, databases, and machine learning. He is a recipient of the prestigious "EDIC Doctoral Fellowship" for the academic year 2018-19, and the "Most Reproducible Paper" award at SIGMOD 2018. He has published his research in prestigious data mining and database conferences, served as a reviewer, and co-organized workshops in these conferences. Further information available at https://www.cse.iitk.ac.in/~aarora.

Sainyam Galhotra is a graduate student at UMass Amherst. His research interests include graph analysis, data mining and integration. He is the recipient of the "Best Paper Award" at FSE 2017 and the "Most Reproducible Award" at SIGMOD 2018. He is the first recipient of the "Krithi Ramamritham Scholarship" at UMass for contribution to research in databases. He has published in top data mining, database and machine learning conferences. Further information available at https://people.cs.umass.edu/~sainyam.

Sayan Ranu is an assistant professor in the Computer Science department at IIT Delhi. His research interests include graph mining, spatio-temporal data analytics, and bioinformatics. He was a recipient of the Best Paper Award at WISE 2016 and the "Most Reproducible Paper" award at SIGMOD 2018. Sayan regularly serves in the program committees of conferences and journals including KDD, ICDE, WWW, ICDM, TKDE, VLDB Journal. Further information available at http://www.cse.iitd.ac.in/~sayan/.

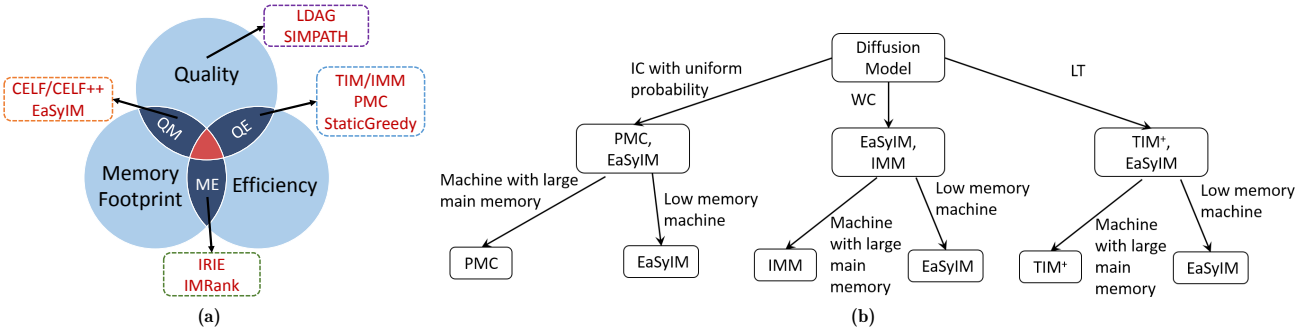

Figure 3: (a) Summarizing the spectrum of Influence Maximization (IM) techniques based on their strengths. (b) The decision tree for choosing the most appropriate IM algorithm.

5.1 Tutorials given by Authors

The authors possess adequate experience of delivering tutorials at reputed venues as indicated below:

- **Akhil Arora, Sainyam Galhotra, Sayan Ranu**, Shourya Roy: "Influence Maximization Revisited", COMAD 2018.
- **Sayan Ranu** and Ambuj Singh: "Indexing and mining topological patterns for drug discovery", in EDBT 2012.
- **Sayan Ranu** and Ambuj Singh: "Topological Indexing and Mining of Chemical Compounds", in BCB 2011.

5.2 Previous Edition of this Tutorial

An overview of the state-of-the-art IM techniques was presented at ACM CoDS-COMAD 2018. The current proposal for EDBT 2019 would include the following extensions:

- We will build upon our benchmarking framework to present detailed insights about the state-of-the-art IM algorithms in real-world scenarios. We will unravel several *myths and ambiguities that plague the current maze of IM techniques.*
- We will discuss in detail about the *open-challenges* that remain in the field of influence maximization and provide concrete pointers to important research questions in order to facilitate streamlined advancement of the field.

REFERENCES

[1] Akhil Arora, Sainyam Galhotra, and Sayan Ranu. 2017. Debunking the Myths of Influence Maximization: An In-Depth Benchmarking Study. In *SIGMOD*.

[2] Cigdem Aslay, Laks Lakshmanan, Wei Lu, and Xiaokui Xiao. 2018. Influence Maximization in Online Social Networks. In *WSDM*.

[3] Christian Borgs, Michael Brautbar, Jennifer Chayes, and Brendan Lucier. 2014. Maximizing Social Influence in Nearly Optimal Time. In *SODA*.

[4] Carlos Castillo, Wei Chen, and Laks V. S. Lakshmanan. 2012. Information and Influence Spread in Social Networks. In *KDD*.

[5] Shuo Chen, Ju Fan, Guoliang Li, Jianhua Feng, Kian-lee Tan, and Jinhui Tang. 2015. Online Topic-aware Influence Maximization. *PVLDB* 8 (2015).

[6] Wei Chen, Wei Lu, and Ning Zhang. 2012. Time-Critical Influence Maximization in Social Networks with Time-Delayed Diffusion Process. *CoRR* abs/1204.3074 (2012).

[7] Wei Chen, Yifei Yuan, and Li Zhang. 2010. Scalable influence maximization in social networks under the linear threshold model. In *ICDM*.

[8] Suqi Cheng, Huawei Shen, Junming Huang, Wei Chen, and Xueqi Cheng. 2014. IMRank: influence maximization via finding self-consistent ranking. In *SIGIR*.

[9] Suqi Cheng, Huawei Shen, Junming Huang, Guoqing Zhang, and Xueqi Cheng. 2013. Staticgreedy: solving the scalability-accuracy dilemma in influence maximization. In *CIKM*.

[10] Edith Cohen, Daniel Delling, Thomas Pajor, and Renato F. Werneck. 2014. Sketch-based Influence Maximization and Computation: Scaling up with Guarantees. In *CIKM*.

[11] Jing Fan, Adalbert Gerald Soosai Raj, and Jignesh M. Patel. 2015. The Case Against Specialized Graph Analytics Engines. In *CIDR*.

[12] Sainyam Galhotra, Akhil Arora, and Shourya Roy. 2016. Holistic Influence Maximization: Combining Scalability and Efficiency with Opinion-Aware Models. In *SIGMOD*.

[13] Sainyam Galhotra, Akhil Arora, Srinivas Virinchi, and Shourya Roy. 2015. ASIM: A Scalable Algorithm for Influence Maximization Under the Independent Cascade Model. In *WWW*.

[14] Amit Goyal, Francesco Bonchi, and Laks V. S. Lakshmanan. 2011. A Data-based Approach to Social Influence Maximization. *PVLDB* 5 (2011).

[15] Amit Goyal, Wei Lu, and Laks V.S. Lakshmanan. 2011. CELF++: Optimizing the Greedy Algorithm for Influence Maximization in Social Networks. In *WWW*.

[16] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. 2011. Simpath: An efficient algorithm for influence maximization under the linear threshold model. In *ICDM*.

[17] Kyomin Jung, Wooram Heo, and Wei Chen. 2012. IRIE: Scalable and Robust Influence Maximization in Social Networks. In *ICDM*.

[18] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the Spread of Influence Through a Social Network. In *KDD*.

[19] V. Kumar, Vikram Bhaskaran, Rohan Mirchandani, and Milap Shah. 2013. Creating a Measurable Social Media Marketing Strategy: Increasing the Value and ROI of Intangibles and Tangibles for Hokey Pokey. *Marketing Science* 32, 2 (2013).

[20] Konstantin Kutzkov, Albert Bifet, Francesco Bonchi, and Aristides Gionis. 2013. STRIP: Stream Learning of Influence Probabilities. In *KDD*.

[21] Jure Leskovec, Andreas Krause, Carlos Guestrin, and Christos Faloutsos. 2007. Cost-effective Outbreak Detection in Networks. In *KDD*.

[22] Hui Li, Sourav S. Bhowmick, Jiangtao Cui, Yunjun Gao, and Jianfeng Ma. 2015. GetReal: Towards Realistic Selection of Influence Maximization Strategies in Competitive Networks. In *SIGMOD*.

[23] Hung T. Nguyen, Tri P. Nguyen, NhatHai Phan, and Thang N. Dinh. 2017. Importance Sketching of Influence Dynamics in Billion-Scale Networks. In *ICDM*.

[24] Hung T. Nguyen, My T. Thai, and Thang N. Dinh. 2016. Stop-and-Stare: Optimal Sampling Algorithms for Viral Marketing in Billion-scale Networks. In *SIGMOD*.

[25] Naoto Ohsaka, Takuya Akiba, Yuichi Yoshida, and Kenichi Kawarabayashi. 2014. Fast and Accurate Influence Maximization on Large Networks with Pruned Monte-Carlo Simulations. In *AAAI*.

[26] Naoto Ohsaka, Tomohiro Sonobe, Sumio Fujita, and Ken-ichi Kawarabayashi. 2017. Coarsening Massive Influence Networks for Scalable Diffusion Analysis. In *SIGMOD*.

[27] Diana Popova, Naoto Ohsaka, Ken-ichi Kawarabayashi, and Alex Thomo. 2018. NoSingles: a space-efficient algorithm for influence maximization. In *SSDBM*.

[28] Manuel Rodriguez and Le Song. 2015. Diffusion in Social and Information Networks: Research Problems, Probabilistic Models and Machine Learning Methods. In *KDD*.

[29] Jimeng Sun and Jie Tang. 2014. Models and algorithms for social influence analysis. In *WWW*.

[30] Youze Tang, Yanchen Shi, and Xiaokui Xiao. 2015. Influence Maximization in Near-Linear Time: A Martingale Approach. In *SIGMOD*.

[31] Youze Tang, Xiaokui Xiao, and Yanchen Shi. 2014. Influence Maximization: Near-optimal Time Complexity Meets Practical Efficiency. In *SIGMOD*.