Debunking the Myths of Influence Maximization

Akhil Arora[1], **Sainyam Galhotra**[1], Sayan Ranu

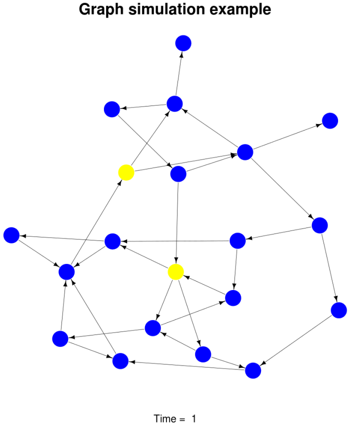
sainyam@cs.umass.edu
University of Massachusetts, Amherst

January 27, 2017
NEDB, 2017

[1]The first two authors have contributed equally to this work.

Information Propagation[2]: Need for Modelling??

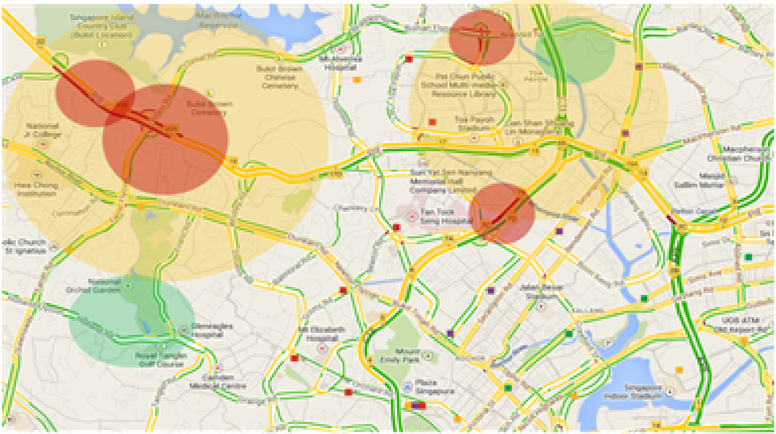
- Many real-world processes can be interpreted using concepts from information propagation
- For example: Spread of Diseases



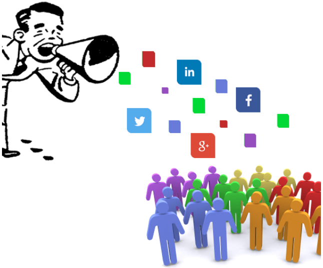
[2] Propagation/Flow/Spread/Diffusion, would be used interchangeably

Need for Modelling??

- Traffic Congestion and its propagation

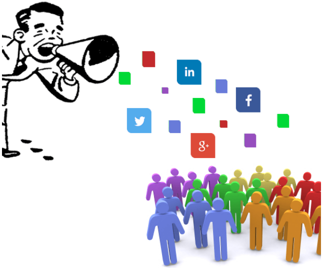


Other Applications



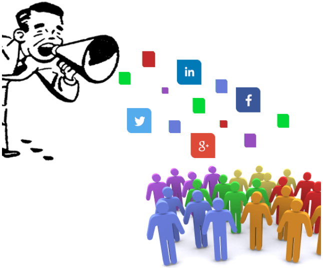
- Using the word-of-mouth effect for:

Other Applications



- Using the word-of-mouth effect for:
  - Viral Marketing: Product/Topic/Event promotion
  - Managing Celebrity/Political campaigns

Other Applications



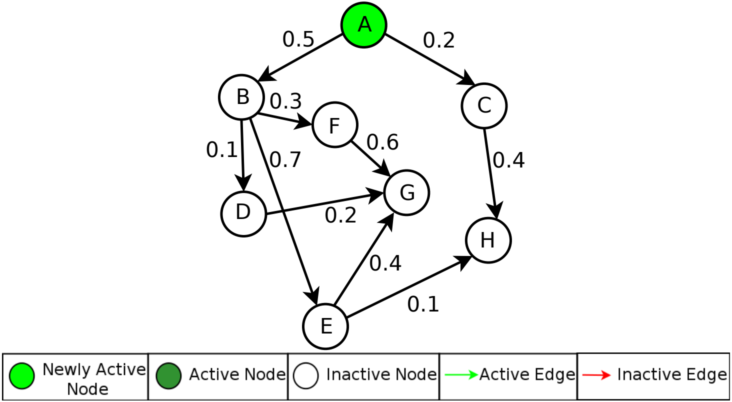
- Using the word-of-mouth effect for:
  - Viral Marketing: Product/Topic/Event promotion
  - Managing Celebrity/Political campaigns
- Detect and Prevent Outbreaks/Epidemics/Rumours
- Many more . . .

Existing Information Propagation models

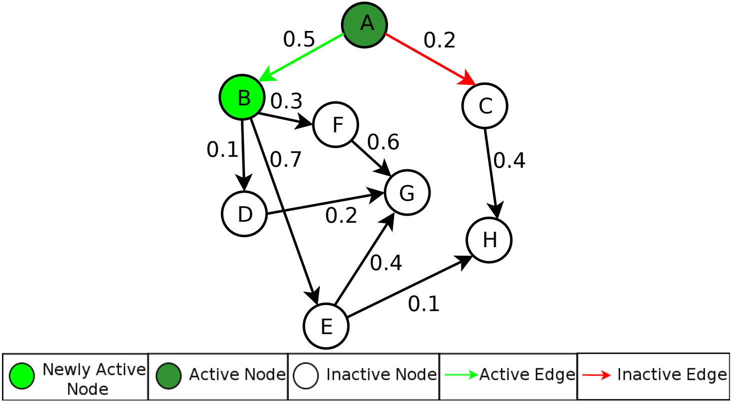
- Independent Cascade (IC) and Weighted Cascade (WC) Models
- Linear Threshold (LT) Model
- Other models – Heat Diffusion etc.

Existing Information Propagation models

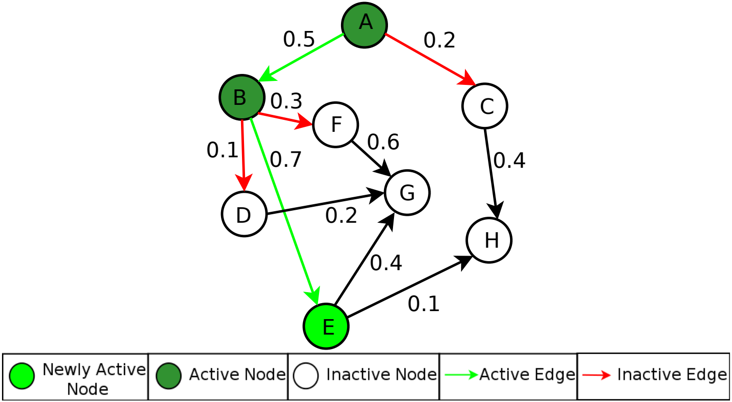
- Independent Cascade (IC) and Weighted Cascade (WC) Models
- Linear Threshold (LT) Model
- Other models – Heat Diffusion etc.



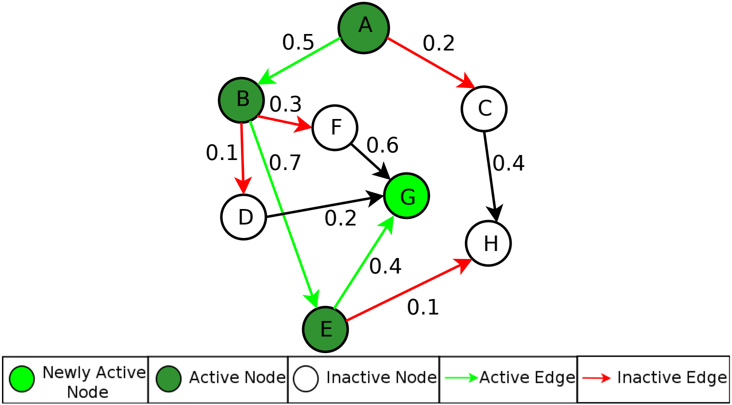
Existing Information Propagation models



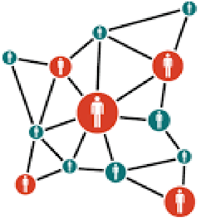
Existing Information Propagation models



Existing Information Propagation models

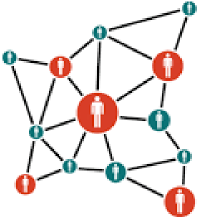


The Influence Maximization (IM) Problem



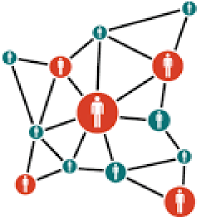
- Input: A graph G, an information-diffusion model $\mathcal{I}$
- Constraints: The budget ($k = |S|$) defining the size of the seed-set

The Influence Maximization (IM) Problem



- Input: A graph G, an information-diffusion model $\mathcal{I}$
- Constraints: The budget ($k = |S|$) defining the size of the seed-set
- Task: Identify the set of most-influential nodes in a network
  - Maximize $\sigma(S) = \mathbb{E}[\mathbb{F}(S)]$: Expected number of nodes active at the end, if set S is targeted for initial activation

The Influence Maximization (IM) Problem



- Input: A graph G, an information-diffusion model $\mathcal{I}$
- Constraints: The budget ($k = |S|$) defining the size of the seed-set
- Task: Identify the set of most-influential nodes in a network
  - Maximize $\sigma(S) = \mathbb{E}[\mathbb{F}(S)]$: Expected number of nodes active at the end, if set S is targeted for initial activation
- Tractability: The IM problem is NP-hard. Need for Approximate Solutions!
- The spread function σ is Monotone and Submodular, thus, a simple GREEDY algorithm provides the best possible $(1 - 1/e)$ approximation

Need for benchmarking? Wide variety of techniques

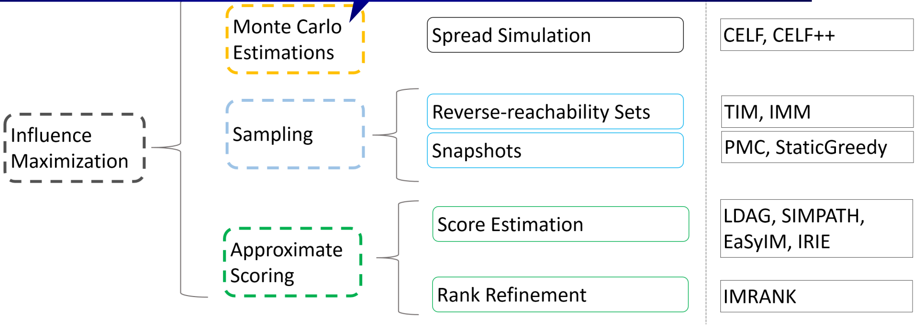
Influence Maximization

Categories	Problem Solving Perspective	Representatives
Monte Carlo Estimations	Spread Simulation	CELF, CELF++
Sampling	Reverse-reachability Sets	TIM, IMM
	Snapshots	PMC, StaticGreedy
Approximate Scoring	Score Estimation	LDAG, SIMPATH, EaSyIM, IRIE
	Rank Refinement	IMRANK

Need for benchmarking? Wide variety of techniques

MC Simulation

1. **Run MC Simulation from each node to estimate its spread.**
2. **Exploit submodularity to prune out nodes with low spread**



Need for benchmarking? Wide variety of techniques

MC Simulation

1. **Run MC Simulation from each node to estimate its spread.**
2. **Exploit submodularity to prune out nodes with low spread**

Sampling

Store a DAG for a sample of nodes and use it to estimate influ-ence

tives

++

Influence Maximization

Sampling

Reverse-reachability Sets — TIM, IMM

Snapshots — PMC, StaticGreedy

Approximate Scoring

Score Estimation — LDAG, SIMPATH, EaSyIM, IRIE

Rank Refinement — IMRANK

Need for benchmarking? Wide variety of techniques

MC Simulation

1. **Run MC Simulation from each node to estimate its spread.**
2. **Exploit submodularity to prune out nodes with low spread**

Sampling

Store a DAG for a sample of nodes and use it to estimate influence

Approximate Scoring

Estimate the influence of the nodes using heuristics as exact computation is #P - hard

Influence

Sampling

Scoring

Rank Refinement

IMRANK

Need for benchmarking? : Ambiguities

- Existing Literature: Use IC, WC interchangeably
- Actual scenario: Varied behaviour in terms of the spread of seed nodes, efficiency and scalability aspects of different techniques.

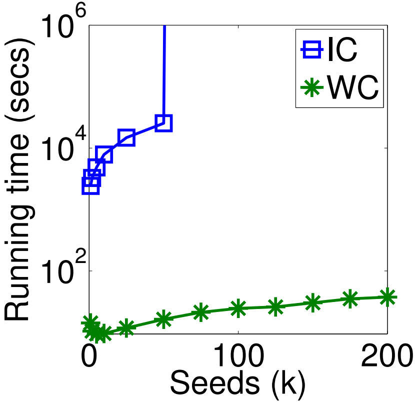
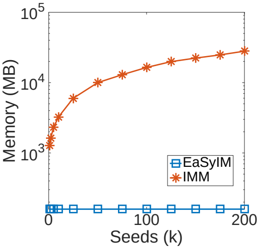


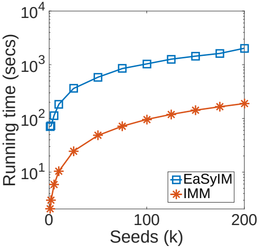
Figure: *IMM ($\epsilon = 0.5$) for Orkut dataset*

Need for benchmarking? : Ambiguities

- State-of-the-art technique in one aspect behaves the worst in another aspect of the problem.



(a) Memory



(b) Running Time

Important Questions

- How to choose the most appropriate IM technique in a given specific scenario?

Important Questions

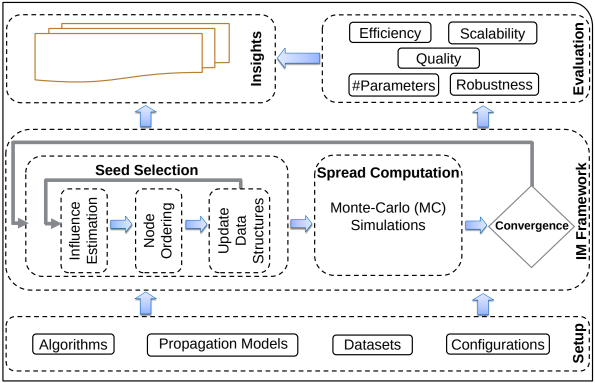
- How to choose the most appropriate IM technique in a given specific scenario?
- What does it really mean to claim to be the state-of-the-art?

Important Questions

- 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?
- Are the claims made by the recent papers true?

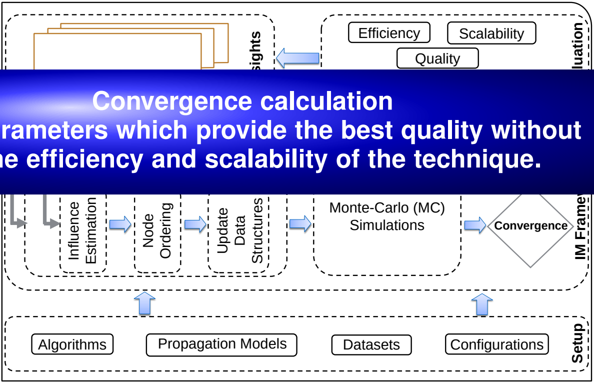
Our Framework

- Generic framework applicable on all techniques.
- Unified approach to tune the external parameters.



Our Framework

- Generic framework applicable on all techniques.
- Unified approach to tune the external parameters.



Convergence calculation

Select the parameters which provide the best quality without hampering the efficiency and scalability of the technique.

Myths

- IMM is always faster than TIM^{+}?

against the states of the art under several popular diffusion models, using real social networks with up to 1.4 billion edges. Our experimental results show that the proposed algorithm consistently outperforms the states of the art in terms of computation efficiency, and is often orders of magnitude faster.

Myths

- IMM is always faster than TIM^+?

> against the states of the art under several popular diffusion models, using real social networks with up to 1.4 billion edges. Our experimental results show that the proposed algorithm consistently outperforms the states of the art in terms of computation efficiency, and is often orders of magnitude faster.

Model	ϵ (TIM^+)	ϵ **(IMM)**	**Time (**TIM^+**)**	**Time (IMM)**	Gain
IC	0.05	0.05	8582.23	829.6	10.3x
LT	0.35	0.1	0.79	1.2	0.65x

Table: ***Comparison of convergence parameter and running time (secs) for IMM and*** TIM^+ ***over HepPH dataset for 200 seeds***

Myths

- CELF++ is the fastest IM technique in the MC estimation paradigm?

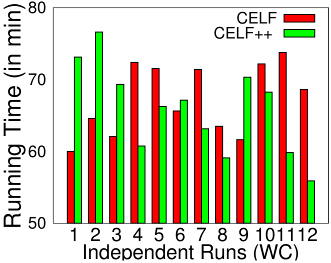
5, 1, 3]. Leskovec et al. [6] proposed the CELF algorithm for tackling the second. In this work, we propose CELF++ and empirically show that it is 35-55% faster than CELF.

Categories and Subject Descriptors H.2.8 [Database

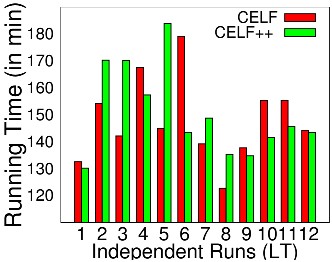
Myths

- CELF++ is the fastest IM technique in the MC estimation paradigm?

> 5, 1, 3]. Leskovec et al. [6] proposed the CELF algorithm for tackling the second. In this work, we propose CELF++ and empirically show that it is 35-55% faster than CELF.
>
> **Categories and Subject Descriptors** H.2.8 [Database



(e) Nethept (WC)



(f) Nethept (LT)

Myths

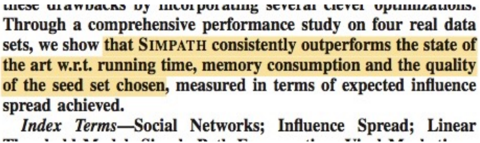
- SIMPATH is faster the LDAG?

Through a comprehensive performance study on four real data sets, we show that SIMPATH consistently outperforms the state of the art w.r.t. running time, memory consumption and the quality of the seed set chosen, measured in terms of expected influence spread achieved.

Index Terms—Social Networks; Influence Spread; Linear

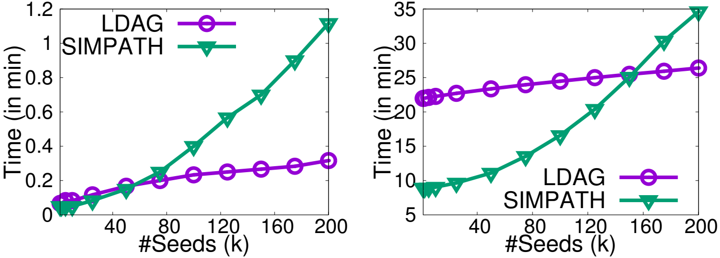
Myths

- SIMPATH is faster the LDAG?



Through a comprehensive performance study on four real data sets, we show that SIMPATH consistently outperforms the state of the art w.r.t. running time, memory consumption and the quality of the seed set chosen, measured in terms of expected influence spread achieved.

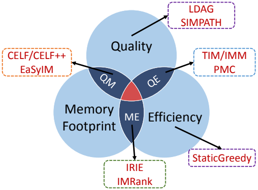
Index Terms—Social Networks; Influence Spread; Linear



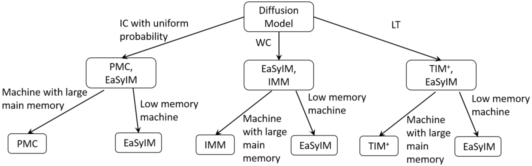
(i) Nethept (j) DBLP

Conclusions

- No technique is the best on all aspects of IM.



(k) Qualitative categorization of IM techniques



(l) Which technique to choose & when?

Thanks!

- For more details, please refer :
  A. Arora, S. Galhotra, S. Ranu. Debunking the Myths of Influence Maximization : An In-Depth Benchmarking Study. SIGMOD 2017