Holistic Influence Maximization: Combining Scalability and Efficiency with Opinion-Aware Models

Sainyam Galhotra[1], **Akhil Arora**[1], Shourya Roy

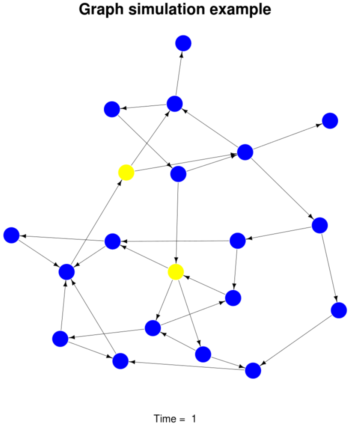
sainyam@cs.umass.edu, aarora@cse.iitk.ac.in
Text & Graph Analytics Group
Xerox Research Centre India, Bangalore

29^{th} June, 2016
ACM SIGMOD Conference, San Francisco, California

[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



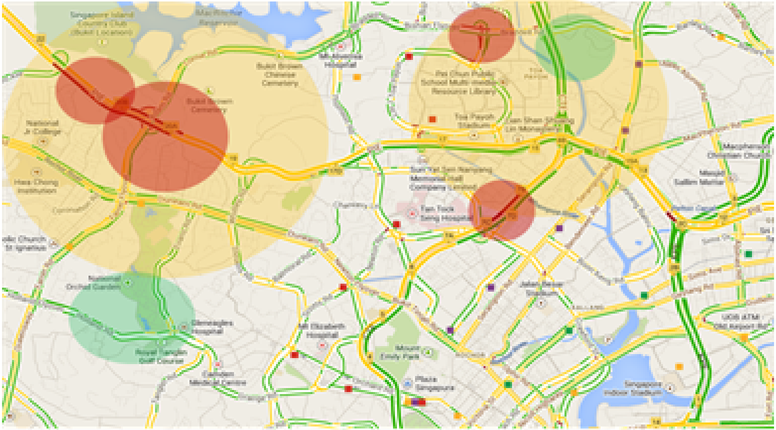
[2]Propagation/Flow/Spread/Diffusion, would be used interchangeably

Need for Modelling??

- Traffic Congestion and its propagation

Need for Modelling??

- Traffic Congestion and its propagation

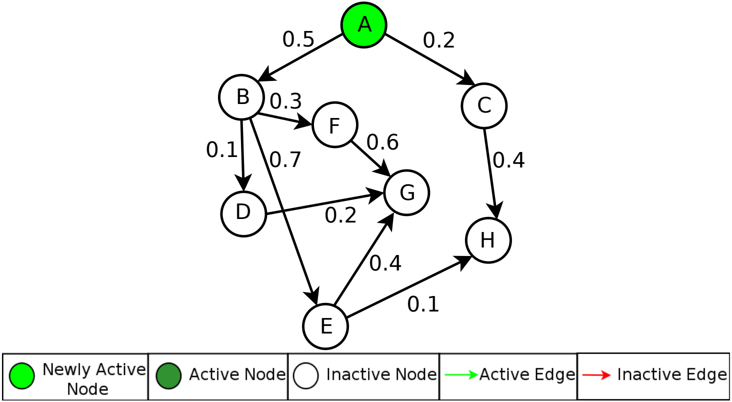


Existing Information Propagation models

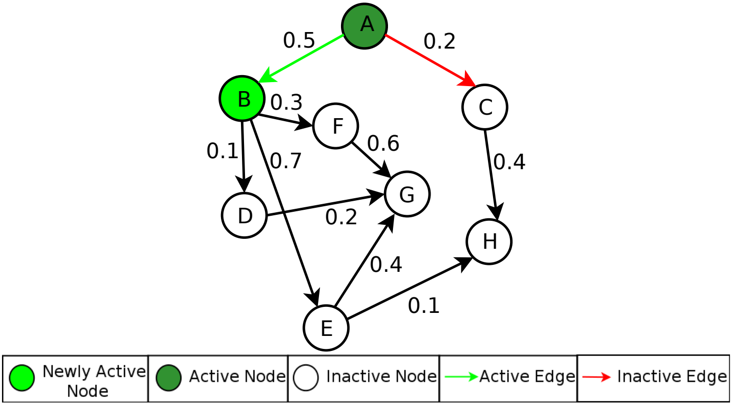
- Independent Cascade (IC) and Weighted Cascade (WC) Models
- Linear Threshold (LT) Model
- Other models – Heat Diffusion etc.

Existing Information Propagation models

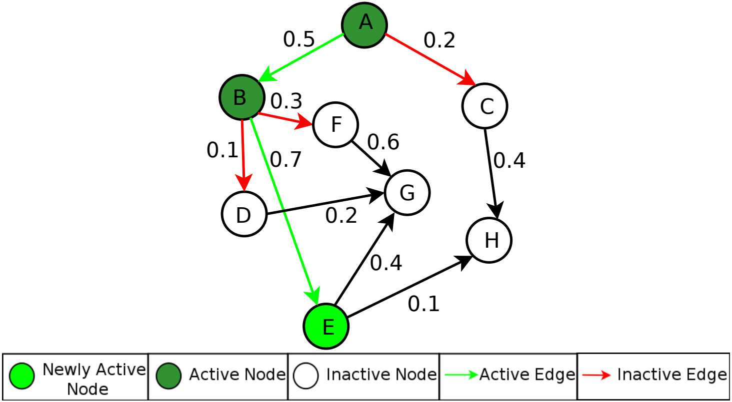
- Independent Cascade (IC) and Weighted Cascade (WC) Models
- Linear Threshold (LT) Model
- Other models – Heat Diffusion etc.



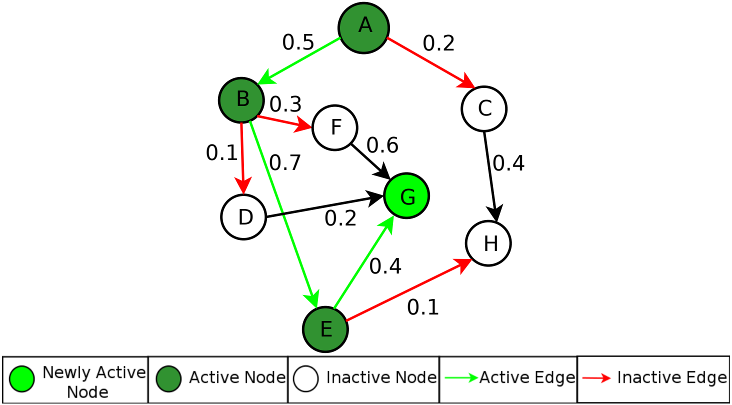
Existing Information Propagation models



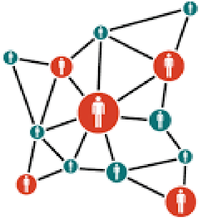
Existing Information Propagation models



Existing Information Propagation models

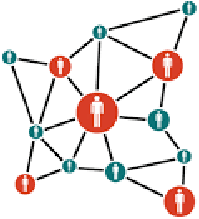


The Influence Maximization (IM) Problem



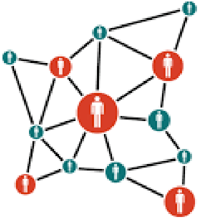
- Input: A graph G, an information-diffusion model $\mathcal{I}$
- Constraints: The budget ($k = |S|$) defining the size of the seed-set

The Influence Maximization (IM) Problem



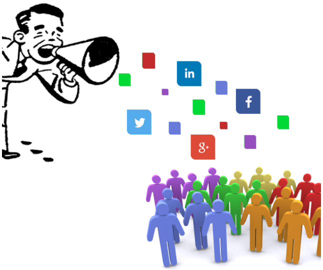
- Input: A graph G, an information-diffusion model $\mathcal{I}$
- Constraints: The budget ($k = |S|$) defining the size of the seed-set
- Task: Identify the set of most-influential nodes in a network
  - Maximize $\sigma(S) = \mathbb{E}[\mathbb{F}(S)]$: Expected number of nodes active at the end, if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

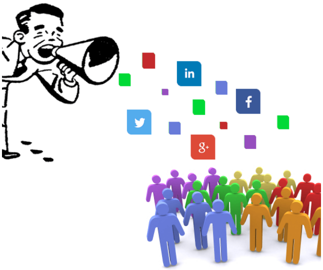
Real-world Applications[3]



- Using the word-of-mouth effect for:

[3] All applications are practical only under a budget constraint

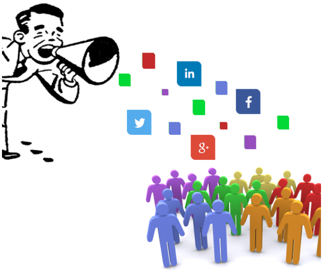
Real-world Applications[3]



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

[3] All applications are practical only under a budget constraint

Real-world Applications[3]



- 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 . . .

[3] All applications are practical only under a budget constraint

Contributions!

- We propose a *holistic* solution to the Influence Maximization problem
- Why do we term our solution holistic?

Contributions!

- We propose a *holistic* solution to the Influence Maximization problem
- Why do we term our solution holistic?
- Answer: We address various aspects of IM
- Diffusion Model:
  - Classical diffusion models [KKT03] fail to capture real-world phenomena. Too simplistic! Results

Contributions!

- We propose a *holistic* solution to the Influence Maximization

Contribution 1: Novel, generic and close to real-world OI model

- Why do we term our solution holistic?
- Answer: We address various aspects of IM
- Diffusion Model:
  - Classical diffusion models [KKT03] fail to capture real-world phenomena. Too simplistic! Results

Contributions!

- We propose a *holistic* solution to the Influence Maximization

Contribution 1: Novel, generic and close to real-world OI model

- Why do we term our solution holistic?
- Answer: We address various aspects of IM
- Diffusion Model:
  - Classical diffusion models [KKT03] fail to capture real-world phenomena. Too simplistic! Results
- Seed Selection Algorithm:
  - Run-time efficiency and efficacy attributes have been extensively studied [KKT03, LKG+07, GLL11, BBCL14, TXS14, CDPW14]
  - However, scalable solutions (catering to both running-time and memory-consumption) are non-existent

Contributions!

- We propose a *holistic* solution to the Influence Maximization

Contribution 1: Novel, generic and close to real-world OI model

- Why do we term our solution holistic?
- Answer: We address various aspects of IM
- Diffusion Model:
  - Classical diffusion models [KKT03] fail to capture real-world phenomena. Too simplistic! Results
- Seed Selection Algorithm:
  - Run-time efficiency and efficacy attributes have been extensively

Contribution 2: Scalable algorithms – EaSyIM and OSIM

W14]

  - However, scalable solutions catering to both running time and memory-consumption) are non-existent

Need for Opinion-Aware IM

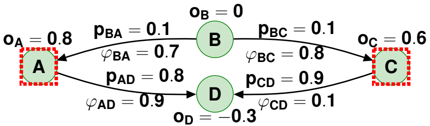


Figure: ***A sample representation of the Twitter network.***

- Perform IM with $k = 1$
- Nodes A and C follow B, while D follows A and C

Need for Opinion-Aware IM

Empirical Analysis

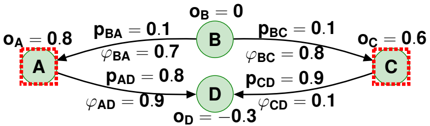


Figure: ***A sample representation of the Twitter network.***

- Perform IM with $k = 1$
- Nodes A and C follow B, while D follows A and C
- IC model:
  - Since $p_{CD} = 0.9$, the expected spread of C under the IC model $\sigma(\mathbf{C}) = \mathbf{0.9}$
  - Similarly for other nodes: $\sigma(A) = 0.8,\ \sigma(B) = 0.3628$ and $\sigma(D) = 0$

Need for Opinion-Aware IM Empirical Analysis

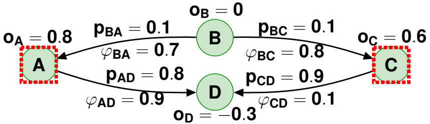


Figure: ***A sample representation of the Twitter network.***

- Perform IM with $k = 1$
- Nodes A and C follow B, while D follows A and C
- IC model:
  - Since $p_{CD} = 0.9$, the expected spread of C under the IC model $\sigma(\mathbf{C}) = \mathbf{0.9}$
  - Similarly for other nodes: $\sigma(A) = 0.8,\ \sigma(B) = 0.3628$ and $\sigma(D) = 0$
  - Thus, C is selected as the seed node

Need for Opinion-Aware IM (Empirical Analysis)

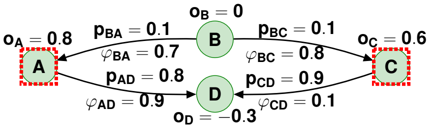


Figure: ***A sample representation of the Twitter network.***

- Perform IM with $k = 1$
- Nodes A and C follow B, while D follows A and C
- IC model:
  - Since $p_{CD} = 0.9$, the expected spread of C under the IC model $\sigma(\mathbf{C}) = \mathbf{0.9}$
  - Similarly for other nodes: $\sigma(A) = 0.8,\ \sigma(B) = 0.3628$ and $\sigma(D) = 0$
  - Thus, C is selected as the seed node
- OI model:
  - Since D agrees with A with a probability of φ_{AD} and disagrees otherwise, the expected opinion-spread of A under the OI model is expressed as $\sigma^{\mathbf{o}}(\mathbf{A}) = p_{AD}\big(\varphi_{AD}(o_D + o_A)/2 + (1 - \varphi_{AD})(o_D - o_A)/2\big) = \mathbf{0.136}$
  - Similarly: $\sigma^o(B) = -0.022564,\ \sigma^o(C) = -0.351$ and $\sigma^o(D) = 0$

Need for Opinion-Aware IM

Empirical Analysis

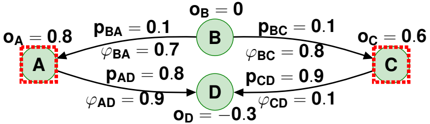


Figure: ***A sample representation of the Twitter network.***

- Perform IM with $k = 1$
- Nodes A and C follow B, while D follows A and C
- IC model:
  - Since $p_{CD} = 0.9$, the expected spread of C under the IC model $\sigma(\mathbf{C}) = \mathbf{0.9}$
  - Similarly for other nodes: $\sigma(A) = 0.8,\ \sigma(B) = 0.3628$ and $\sigma(D) = 0$
  - Thus, C is selected as the seed node
- OI model:
  - Since D agrees with A with a probability of φ_{AD} and disagrees otherwise, the expected opinion-spread of A under the OI model is expressed as
    $\sigma^{\mathbf{o}}(\mathbf{A}) = p_{AD}\big(\varphi_{AD}(o_D + o_A)/2 + (1 - \varphi_{AD})(o_D - o_A)/2\big) = \mathbf{0.136}$
  - Similarly: $\sigma^o(B) = -0.022564,\ \sigma^o(C) = -0.351$ and $\sigma^o(D) = 0$
  - Thus, A is selected as the seed node

Opinion-cum-Interaction (OI) Model

- Second layer on the top of IC/WC and LT to model the propagation and change of opinion
- Models Opinion Spread – The contribution of a newly activated node can be signed
- Two components – opinion ($o_v \in [-1, 1]$) and interaction ($\varphi_{(u,v)} \in [0, 1]$)

Opinion-cum-Interaction (OI) Model

- Second layer on the top of IC/WC and LT to model the propagation and change of opinion
- Models Opinion Spread – The contribution of a newly activated node can be signed
- Two components – opinion ($o_v \in [-1, 1]$) and interaction ($\varphi_{(u,v)} \in [0, 1]$)
- Opinion of a node can be estimated using its own opinions on similar events in the past
- Interaction probabilities (directed), between two nodes, can be estimated by accounting for all of their possible interactions in the past

Opinion-cum-Interaction (OI) Model

- Second layer on the top of IC/WC and LT to model the propagation and change of opinion
- Models Opinion Spread – The contribution of a newly activated node can be signed
- Two components – opinion ($o_v \in [-1, 1]$) and interaction ($\varphi_{(u,v)} \in [0, 1]$)
- Opinion of a node can be estimated using its own opinions on similar events in the past
- Interaction probabilities (directed), between two nodes, can be estimated by accounting for all of their possible interactions in the past
- Effective opinion (o'_v) of an activated node v is dependent upon both, its personal opinion o_v and the effective opinion (o'_u) of all the nodes $u \in V_{(a)}$ (set of nodes activated at previous steps)

The Opinion Maximization (MEO) Problem

Opinion Spread

Sum of opinions of the users in the activated set, when S is the chosen seed set

Effective Opinion Spread

Weighted difference between the opinion spread of the users with positive polarity and the opinion spread of negatively polarised users, in the set of activated nodes.

- Task: Identify the set of most-influential nodes
  - Maximize $\sigma^o(S) = \mathbb{E}[\mathbb{F}^o(S)]$: Expected effective opinion spread obtained at the end, if set S is targeted for initial activation

The Opinion Maximization (MEO) Problem

Opinion Spread

Sum of opinions of the users in the activated set, when S is the chosen seed set

Effective Opinion Spread

Weighted difference between the opinion spread of the users with positive polarity and the opinion spread of negatively polarised users, in the set of activated nodes.

- Task: Identify the set of most-influential nodes
  - Maximize $\sigma^o(S) = \mathbb{E}[\mathbb{F}^o(S)]$: Expected effective opinion spread obtained at the end, if set S is targeted for initial activation
- Result: The opinion spread function σ^o is neither monotonous nor submodular, thus, approximating MEO within any constant ratio is not possible (Proof in the paper)

Outline of our approach

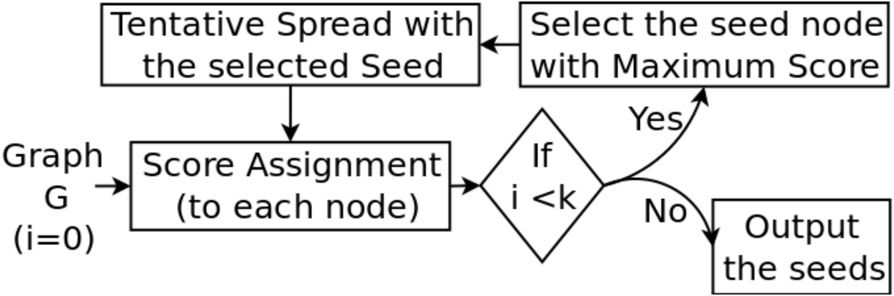


Figure: *Overview of our algorithm*

Seed Selection Algorithms: Intuition

- Observation: Probability of v to get activated by u is dependent on the number of all possible simple paths from u to v in G.

Seed Selection Algorithms: Intuition

- Observation: Probability of v to get activated by u is dependent on the number of all possible simple paths from u to v in G.

- EaSyIM assigns a score to each node (u) of the graph
- Paths of length l from a node u can be calculated as the sum of all paths of length $l-1$ from its neighbors
- $\Delta^l(u)$ ($\forall u \in V$) is defined as the weighted sum of the number of simple paths of length at most (l) starting from u
- Path length $l \leq \mathcal{D}$ (diameter) of the graph is a parameter to control accuracy
- The weight for each path is defined as the product of probabilities $p_{(u,v)}$ of the edges composing that path

Seed Selection Algorithms: Intuition

- Observation: Probability of v to get activated by u is dependent on the number of all possible simple paths from u to v in G.

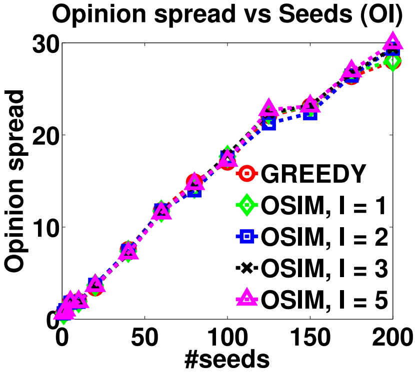
This algorithm can easily be extended to opinion-aware settings (OSIM)

- EaSyIM assigns a score to each node (u) of the graph
- Paths of length l from a node u can be calculated as the sum of all paths of length $l-1$ from its neighbors
- $\Delta^l(u)$ ($\forall u \in V$) is defined as the weighted sum of the number of simple paths of length at most (l) starting from u
- Path length $l \leq \mathcal{D}$ (diameter) of the graph is a parameter to control accuracy
- The weight for each path is defined as the product of probabilities $p_{(u,v)}$ of the edges composing that path

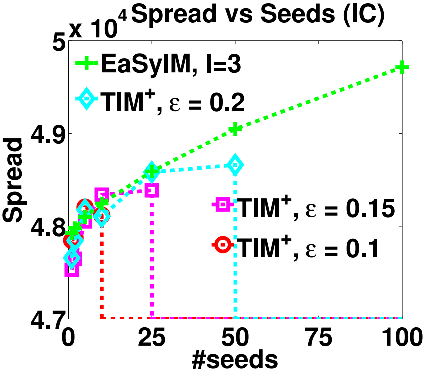
Analysis

- Time Complexity for score assignment – $O((m+n)l)$
- Time taken by EaSyIM/OSIM for selecting k seeds – $O(k(m+n)l)$
- Memory Complexity – $O(n)$
- Approximation Guarantee same as [KKT03] for trees under the IC/WC model and DAGs under the LT model

Quality

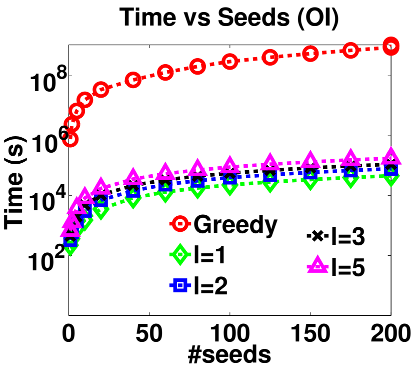


(a) NetHept

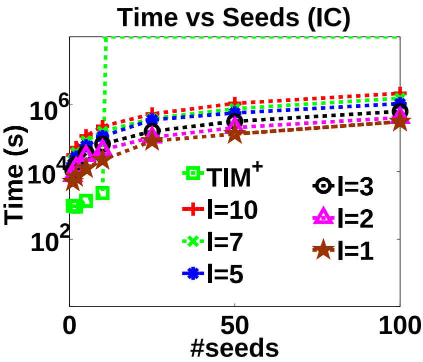


(b) DBLP

Efficiency: Running Time

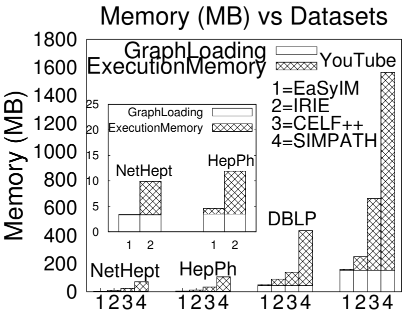


(c) NetHept OI

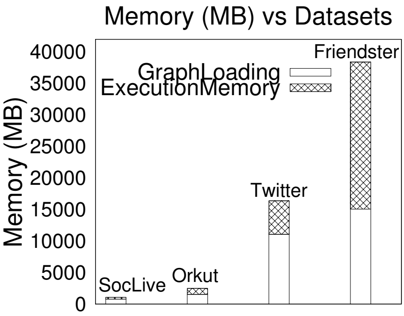


(d) DBLP IC

Scalability: Memory Consumption



(e) EaSyIM: Medium Data



(f) EaSyIM: Large Data

Scalability: SNAP Datasets – Challenges

Dataset	n	m	Type	Avg. Degree	Diameter
NetHEPT	15K	62K	Undirected	4.13	10
HepPh	12K	118K	Undirected	9.87	5.8
DBLP	317K	1M	Undirected	3.3	8
YouTube	1M	3M	Undirected	2.63	6.5
SocLiveJournal	5M	65M	Directed	14.23	6.5

Dataset	**Running Time (min)**			**Memory (MB)**		
	CELF++	**EaSyIM**	**Gain**	**CELF++**	**EaSyIM**	**Gain**
NetHEPT	5352.25	118	**45.35x**	23.26	3.39	6.86x
HepPh	9746.74	230	**41x**	24.60	3.47	7.08x
DBLP	NA	5071.67	**∞**	NA	44.73	∞

Dataset	**Running Time (min)**			**Memory (MB)**		
	TIM$^+$	**EaSyIM**	**Gain**	**TIM$^+$**	**EaSyIM**	**Gain**
DBLP	783.1	2183	0.36x	35234.75	46.5	**758x**
YouTube	NA	5089.5	∞	NA	158.3	∞
socLive	NA	15433.33	**∞**	NA	974.94	∞

Conclusions

- Previous works mostly in the context of opinion-oblivious settings
- First work to propose a holistic solution
  - Realistic and generic information propagation models and optimization objectives
  - Efficient yet Scalable algorithms

Conclusions

- Previous works mostly in the context of opinion-oblivious settings
- First work to propose a holistic solution
  - Realistic and generic information propagation models and optimization objectives
  - Efficient yet Scalable algorithms

THANK YOU!

Conclusions

- Previous works mostly in the context of opinion-oblivious settings
- First work to propose a holistic solution
  - Realistic and generic information propagation models and optimization objectives
  - Efficient yet Scalable algorithms

THANK YOU!
Questions?
Answers!

Need for different Optimization Objectives?

- Maximizing the total spread in a network is not enough
- What about the negative spread? Can hamper product promotion

Need for different Optimization Objectives?

- Maximizing the total spread in a network is not enough
- What about the negative spread? Can hamper product promotion
- Answer: Maximize the positive spread

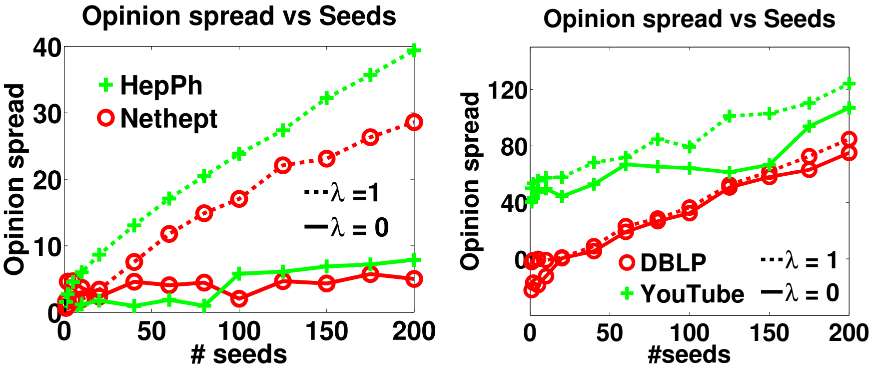
Need for different Optimization Objectives?

- Maximizing the total spread in a network is not enough
- What about the negative spread? Can hamper product promotion
- Answer: Maximize the positive spread
- Even maximizing the positive spread alone is not enough
- Election/Celebrity campaigns can get affected

Need for different Optimization Objectives?

- Maximizing the total spread in a network is not enough
- What about the negative spread? Can hamper product promotion
- Answer: Maximize the positive spread
- Even maximizing the positive spread alone is not enough
- Election/Celebrity campaigns can get affected
- Why not maximize the difference?
- More specifically, Maximize $|\mathbb{F}^{+}(S) - \lambda \times \mathbb{F}^{-}(S)|$, $\lambda \in [-1, 1]$

Optimization Objectives



(g) NetHept and HepPh

(h) DBLP and YouTube

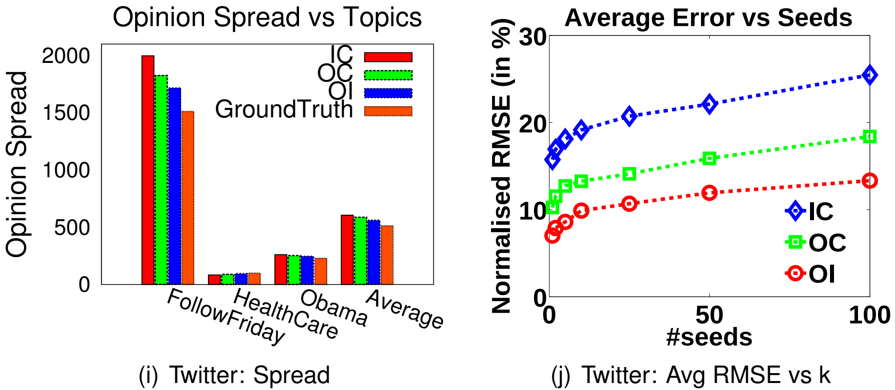
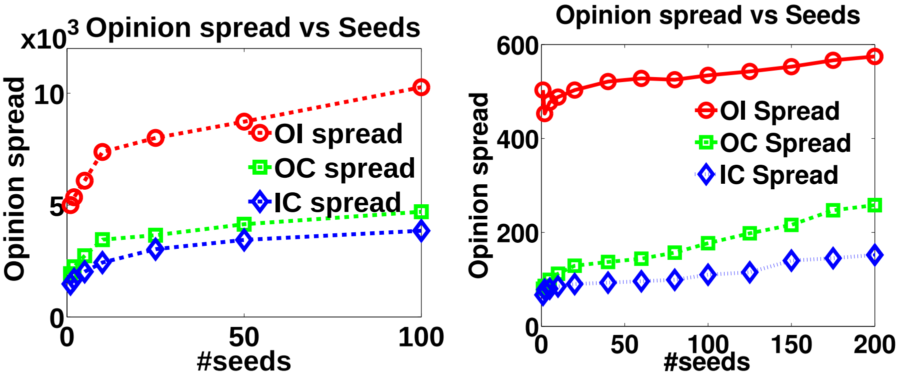


Figure: ***Comparing opinion-spreads under OI, OC and IC with the real-world opinion-spread.***



(a) Twitter (b) PAKDD

Figure: ***Comparing OI with OC and IC: Opinion spread vs k.***

References I

Christian Borgs, Michael Brautbar, Jennifer Chayes, and Brendan Lucier.
Maximizing social influence in nearly optimal time.
In *SODA*, pages 946–957, 2014.

Edith Cohen, Daniel Delling, Thomas Pajor, and Renato F. Werneck.
Sketch-based influence maximization and computation: Scaling up with guarantees.
In *CIKM*, pages 629–638, 2014.

Amit Goyal, Wei Lu, and Laks V.S. Lakshmanan.
Celf++: Optimizing the greedy algorithm for influence maximization in social networks.
In *WWW (Companion Volume)*, pages 47–48, 2011.

David Kempe, Jon Kleinberg, and Éva Tardos.
Maximizing the spread of influence through a social network.
In *KDD*, pages 137–146, 2003.

Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance.
Cost-effective outbreak detection in networks.
In *KDD*, pages 420–429, 2007.

Youze Tang, Xiaokui Xiao, and Yanchen Shi.
Influence maximization: Near-optimal time complexity meets practical efficiency.
In *SIGMOD*, pages 75–86, 2014.