Improved Practical Efficiency for Misinformation Prevention in Social Networks

Michael Simpson
Venkatesh Srinivasan
Alex Thomo

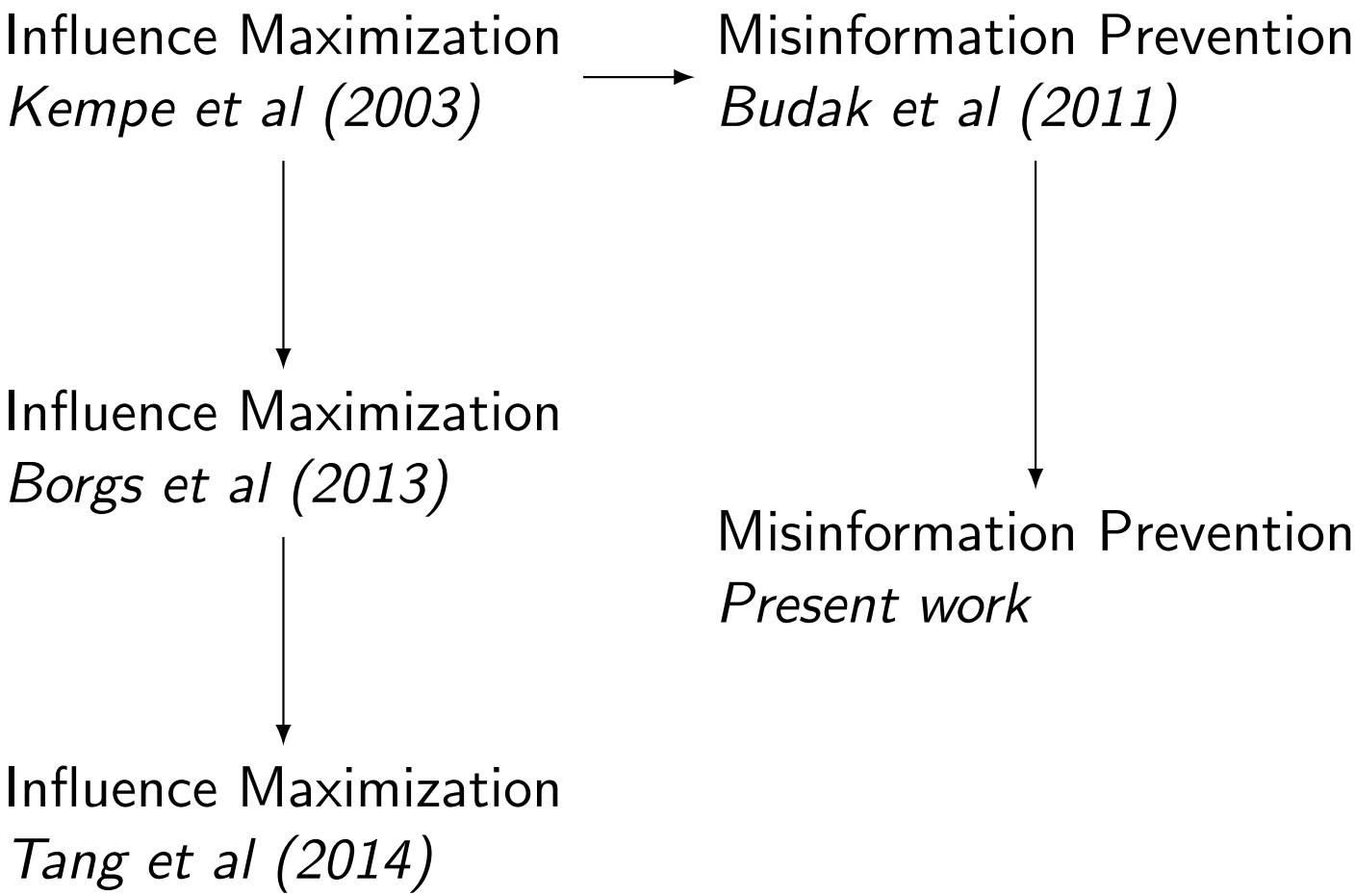
University of Victoria

NWDS 2018

University
of Victoria

Outline

Background



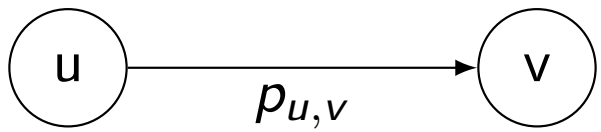
Background

Social networks play a fundamental role as a medium for the spread of information, ideas & influence.

https://phys.org/news/2015-05-rumor-detection-software-ids-disputed-twitter.html

Background: Influence Maximization (2003)

Consider a social network as a graph with edges representing relationships between users and suppose we have estimates for the probabilities that individuals influence one another.



Goal: Adoption of a product by a large fraction of the users in the network by initially targeting a few "influential" members.
Idea: Influential users trigger a cascade of influence leading to many individuals trying the product.
Question: How can we choose the seed set of influential users?

Background: Misinformation Prevention (2011)

- While the ease of information propagation in social networks can be very beneficial, it can also have **disruptive** effects.
- In order for social networks to serve as a reliable platform for disseminating critical information, it is necessary to have tools to limit the effect of misinformation.
- Consider two campaigns propagating through a network: one "good" and one "bad".
- **Question:** What is our objective function?
  - e.g. "save" as many nodes as possible, limit the lifespan of the "bad" campaign, or maximize the adoption of the "good" campaign.

Background: Misinformation Prevention (2011)

- While the ease of information propagation in social networks can be very beneficial, it can also have **disruptive** effects.
- In order for social networks to serve as a reliable platform for disseminating critical information, it is necessary to have tools to limit the effect of misinformation.
- Consider two campaigns propagating through a network: one "good" and one "bad".
- **Question:** What is our objective function?
  - e.g. "save" as many nodes as possible, limit the lifespan of the "bad" campaign, or maximize the adoption of the "good" campaign.
- **Question:** How can we choose a seed set that minimizes the number of users who end adopting the "bad" campaign?

Outline

Background

Influence Maximization
Kempe et al (2003)

→

Misinformation Prevention
Budak et al (2011)

↓

Influence Maximization
Borgs et al (2013)

↓

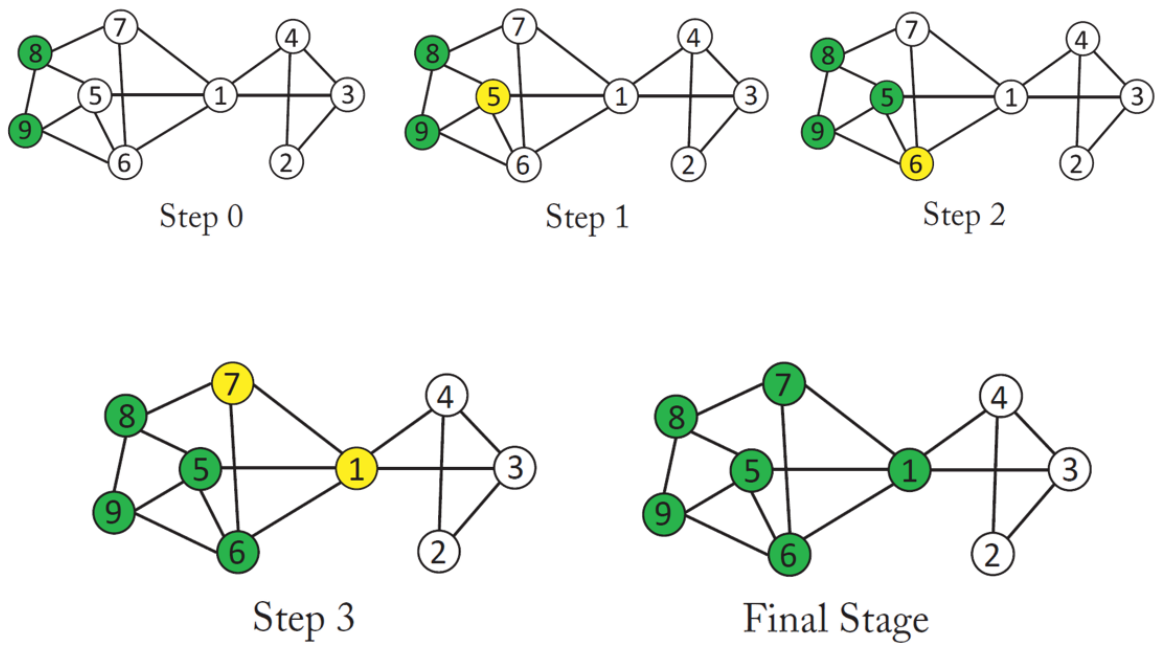
Influence Maximization
Tang et al (2014)

↓

Misinformation Prevention
Present work

Independent Cascade Model (ICM)

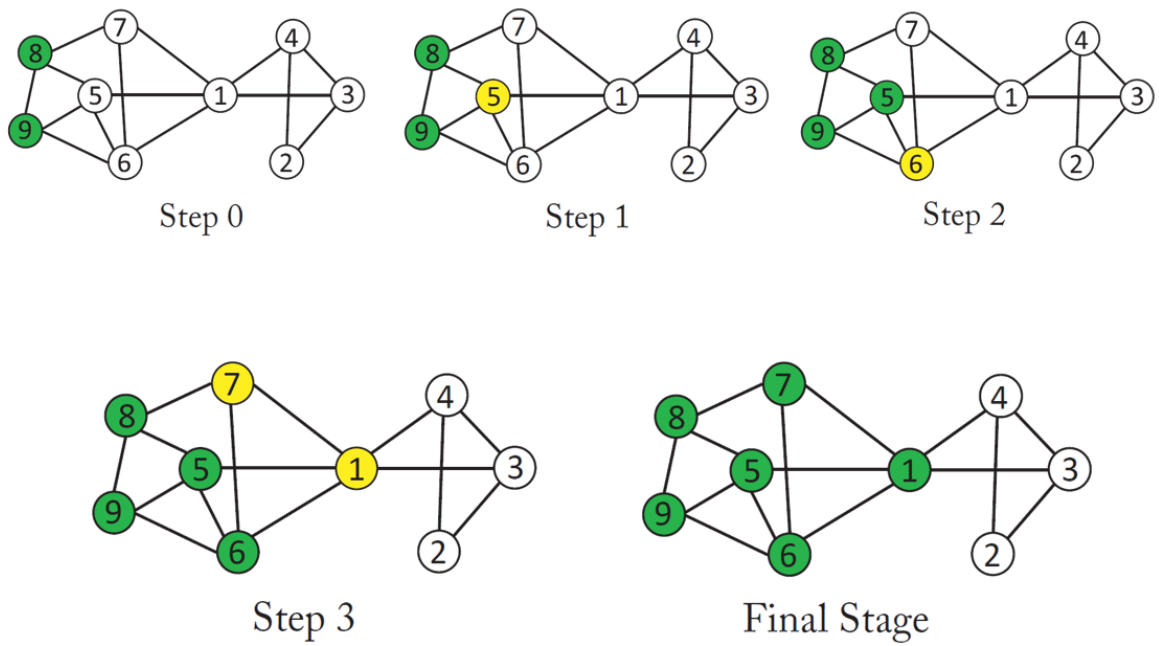
- Seminal work of Kempe, Kleinberg, & Tardos introduce a general model and obtain first provable approximation guarantees.
- Their model considers the diffusion of information through the network in a series of rounds.



http://home.cse.ust.hk/~qyang/621U/

Independent Cascade Model (ICM)

- Formally, assume there is a subset, A_0, referred to as the *seed set* in which the nodes are considered "active".
- In each round, the set of active nodes has a chance to activate neighbouring nodes according to the influence probabilities on the edges.
- Process terminates when no new activations occur from round t to $t+1$.



http://home.cse.ust.hk/~qyang/621U/

Influence Maximization Problem (IM)

- **Influence** of a seed set A_0, denoted $\sigma(A_0)$, is the **expected** number of active nodes at the end of the diffusion process.
- The *Influence Maximization Problem* asks, given a budget k, to find a k-node set of maximum influence (**NP-hard**).

Influence Maximization Problem (IM)

- **Influence** of a seed set A_0, denoted $\sigma(A_0)$, is the **expected** number of active nodes at the end of the diffusion process.
- The *Influence Maximization Problem* asks, given a budget k, to find a k-node set of maximum influence (**NP-hard**).
- Main result of Kempe, Kleinberg, & Tardos is that IM can be approximated to within a factor of $(1 - 1/e - \epsilon)$ via **greedy** approach.

Influence Maximization Problem (IM)

- **Influence** of a seed set A_0, denoted $\sigma(A_0)$, is the **expected** number of active nodes at the end of the diffusion process.
- The *Influence Maximization Problem* asks, given a budget k, to find a k-node set of maximum influence (**NP-hard**).
- Main result of Kempe, Kleinberg, & Tardos is that IM can be approximated to within a factor of $(1 - 1/e - \epsilon)$ via **greedy** approach.
- **Limitation:** in each round of greedy we must estimate the marginal increase in the spread of influence for every node not already in A_0.
  - large number of costly simulations required is a significant computational barrier when considering massive online social networks

Eventual Influence Limitation Problem (EIL)

- Consider two campaigns: a "bad" campaign C and a "limiting" campaign L with seed sets A_C and A_L respectively.
- Let $IF(A_C)$ denote the influence set of C in the **absence** of L, i.e the set of nodes that would adopt campaign C if there were no limiting campaign.

Eventual Influence Limitation Problem (EIL)

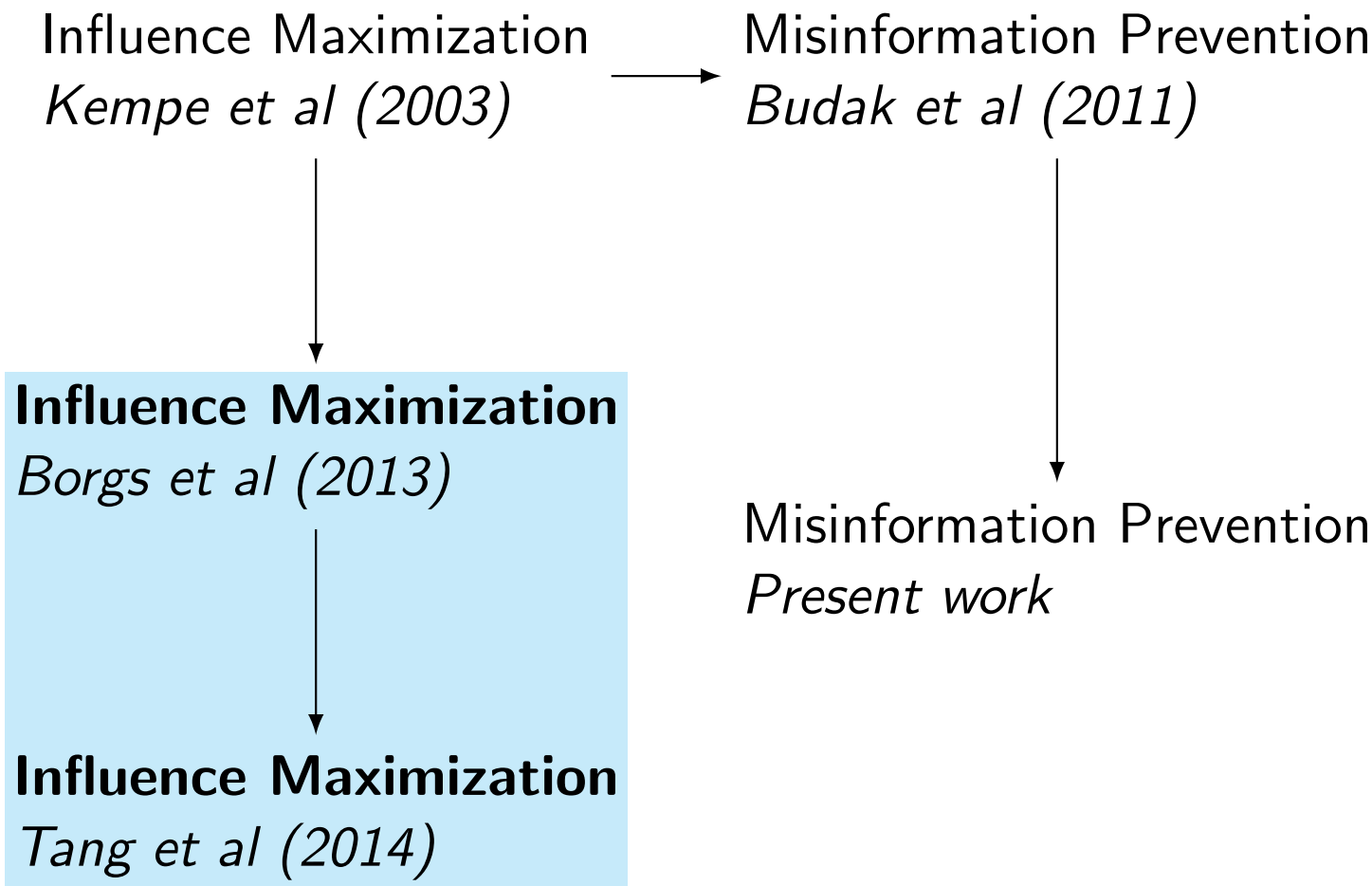
- Consider two campaigns: a "bad" campaign C and a "limiting" campaign L with seed sets A_C and A_L respectively.
- Let $IF(A_C)$ denote the influence set of C in the **absence** of L, i.e the set of nodes that would adopt campaign C if there were no limiting campaign.
- Define the function $\pi(A_L)$ to be the size of the subset of $IF(A_C)$ that campaign L **prevents** from adopting campaign C.

Eventual Influence Limitation Problem (EIL)

- Consider two campaigns: a "bad" campaign C and a "limiting" campaign L with seed sets A_C and A_L respectively.
- Let $IF(A_C)$ denote the influence set of C in the **absence** of L, i.e the set of nodes that would adopt campaign C if there were no limiting campaign.
- Define the function $\pi(A_L)$ to be the size of the subset of $IF(A_C)$ that campaign L **prevents** from adopting campaign C.
- The *Eventual Limitation Problem* asks, for a budget k, to select a k-node set for the limiting campaign L such that the expectation of $\pi(A_L)$ is maximized.
- Budak, Agrawal, & Abbadi are able to show that the **greedy** approach yields the same performance guarantees as it does for IM.

Outline

Background



IM Improvements: Borgs et al

Borgs et al introduced a novel way of viewing the IM problem. Their key insight was instead of asking "Who can I influence?" Asking "**Who could have influenced me?**"

IM Improvements: Borgs et al

Borgs et al introduced a novel way of viewing the IM problem. Their key insight was instead of asking "Who can I influence?" Asking "**Who could have influenced me?**"

In other words: instead of asking, for a node v, which set of nodes can v influence? (i.e. reachability from v)

Asking which nodes could have influenced v? (reverse reachability)

IM Improvements: Borgs et al

Borgs et al introduced a novel way of viewing the IM problem. Their key insight was instead of asking "Who can I influence?" Asking "**Who could have influenced me?**"

In other words: instead of asking, for a node v, which set of nodes can v influence? (i.e. reachability from v)

Asking which nodes could have influenced v? (reverse reachability)

This is a fundamental shift in how to view the Influence Maximization Problem

IM Improvements: Borgs et al

"Who could have influenced me?"

Define the **Reverse Reachable** (RR) set for a node v such that for each node u in the RR set, there is a directed path from u to v in $g \sim G$.

If a node u appears in an RR set generated for a node v, then u should have a **chance to activate** v if we run an influence propagation process on G using $\{u\}$ as the seed set.

IM Improvements: Borgs et al

Idea: If a node u appears in a **large number of random RR sets**, then it should have a high probability to activate many nodes under the IC model; in that case, u's **expected influence** should be large.

Based on this intuition, Borgs' algorithm runs in two steps:

1. Generate a certain number of random RR sets from G.
2. Consider the *maximum coverage* problem of selecting k nodes to cover the maximum number of RR sets generated. Use the standard approach to derive a $(1 - 1/e)$-approximate solution.

IM Improvements: Tang et al

Greedy (Kempe et al) requires $O(kmn)$ time complexity.

IM Improvements: Tang et al

Greedy (Kempe et al) requires $O(kmn)$ time complexity.

Borgs et al propose a threshold-based approach: they keep generating RR sets until the total number of nodes and edges examined during the generation process reaches a pre-defined threshold. This results in a $O(k(m+n)\log^2 n/\epsilon^3)$ time algorithm.

- **Near optimal** since any algorithm that provides same approximation guarantee and succeeds with at least constant probability must run in $\Omega(m+n)$ time.

IM Improvements: Tang et al

Greedy (Kempe et al) requires $O(kmn)$ time complexity.

Borgs et al propose a threshold-based approach: they keep generating RR sets until the total number of nodes and edges examined during the generation process reaches a pre-defined threshold. This results in a $O(k(m+n)\log^2 n/\epsilon^3)$ time algorithm.

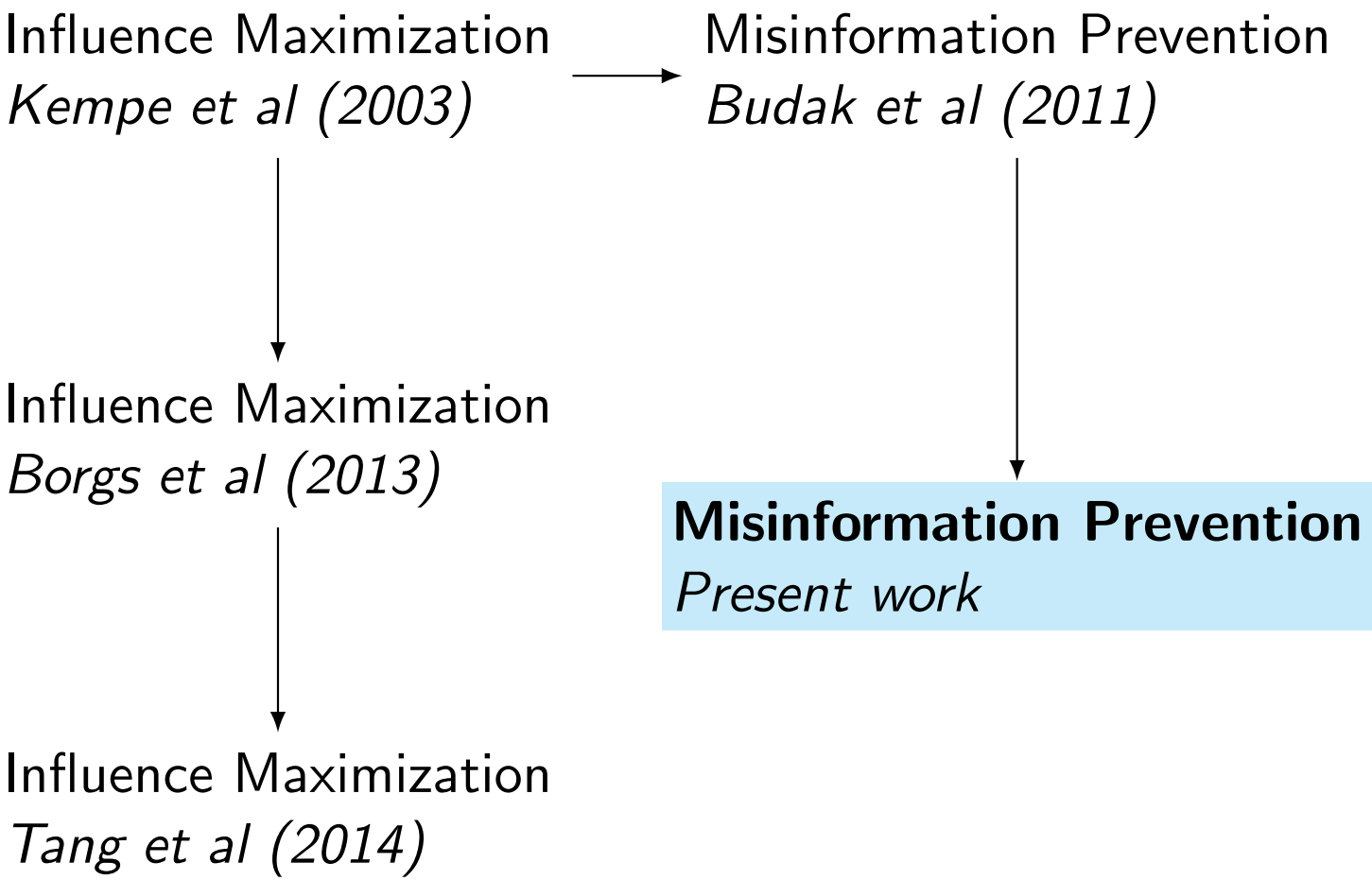
- **Near optimal** since any algorithm that provides same approximation guarantee and succeeds with at least constant probability must run in $\Omega(m+n)$ time.

Tang et al improve this to $O(k(m+n)\log n/\epsilon^2)$ by generating a **fixed** number of RR sets.

- An improvement by a factor of $\log n/\epsilon$.

Outline

Background



Present Work

Our present work seeks to incorporate the new techniques for the IM problem to the misinformation setting of Budak et al. Importantly, this requires adapting the concept of an **RR set** to the multi-campaign setting.

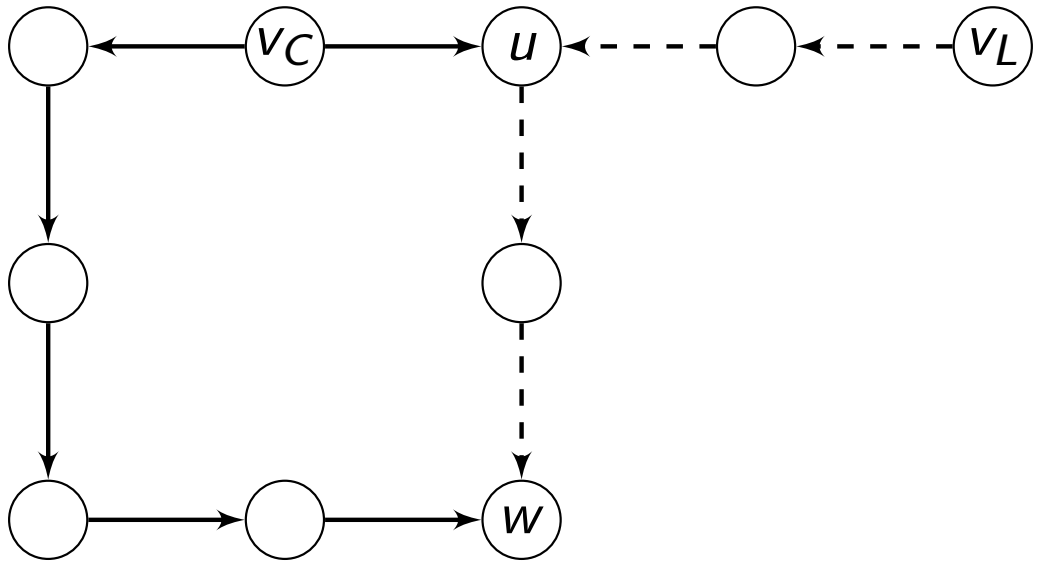
Who could have saved me?

Unlike the IM setting, we must account for the **complicated interactions** that occur during the diffusion of the two campaigns through the graph. Simple shortest path computations **do no suffice**.

Present Work

Who could have saved me?

We must account for the fact that some nodes will be **blocked** by the diffusion of campaign C.



We see that $|SP(v_L, w)| = 4$ and $|SP(v_C, w)| = 5$, but w cannot be saved in the resulting cascade since at timestamp 1 the node u will adopt campaign C.

Present Work

Results:

- We design a **sophisticated BFS-based algorithm** to efficiently compute RR sets in the multi-campaign setting.
- We show that the **proof techniques** of Tang et al can be **successfully applied** to analyze our algorithm for the EIL problem in the multi-campaign setting.
- We use this to construct an approach to solve the EIL problem with a much **stronger asymptotic runtime** than Budak et al.
- Our preliminary experimental results show that our new approach outperforms Budak's greedy approach by a **factor of over 100**.