Multi-phase Information Diffusion in Social Networks

ICTS Program on Games, Epidemics and Behavior

Swapnil Dhamal

Joint work with Y. Narahari and Prabuchandran K. J.

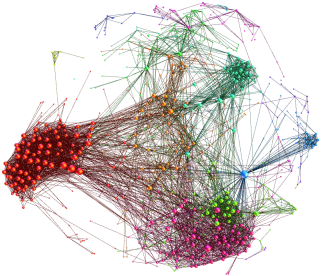
Department of Computer Science and Automation
Indian Institute of Science, Bangalore

June 30, 2015

Overview

Overview

Social Networks



- What is a social network?
  Individuals and some connections among them
- What is a major effect of these connections?
  Individuals influence each other

- What is information diffusion in social networks?
  Diffusion of information through social connections

- How does viral marketing work?
  A few seed nodes are given free products or discounts. They advertise to their friends, who may advertise to their friends, and this process goes on ...

Information Diffusion in Social Networks

- What is information diffusion in social networks?
  Diffusion of information through social connections

- How does viral marketing work?
  A few seed nodes are given free products or discounts. They advertise to their friends, who may advertise to their friends, and this process goes on ... **(But Not For Sure!)**

How about diffusion in multiple phases?

Current Literature and Research Gap

- Influence maximization in a network using single phase [1] [2] [3]
- Basic idea of function maximization using multiple phases [4]

Influence maximization in a network using multiple phases

[1] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In ACM SIGKDD, pages 137–146. ACM, 2003.

[2] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In ACM SIGKDD, pages 199–208. ACM, 2009.

[3] A. Guille, H. Hacid, C. Favre, and D. Zighed. Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 42(1):17–28, 2013.

[4] D. Golovin and A. Krause. Adaptive Submodularity: Theory & Applications in Active Learning and Stochastic Optimization. *JAIR*, 42, 427–486, 2011.

Current Literature and Research Gap

- Influence maximization in a network using single phase [1] [2] [3]
- Basic idea of function maximization using multiple phases [4]

Influence maximization in a network using multiple phases

---

[1] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In ACM SIGKDD, pages 137–146. ACM, 2003.

[2] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In ACM SIGKDD, pages 199–208. ACM, 2009.

[3] A. Guille, H. Hacid, C. Favre, and D. Zighed. Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 42(1):17–28, 2013.

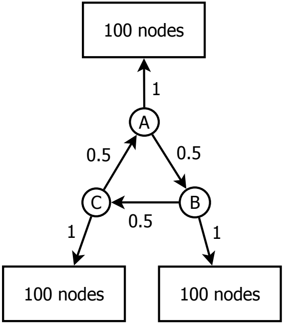
[4] D. Golovin and A. Krause. Adaptive Submodularity: Theory & Applications in Active Learning and Stochastic Optimization. *JAIR*, 42, 427–486, 2011.

Model for Diffusion - Independent Cascade (IC)

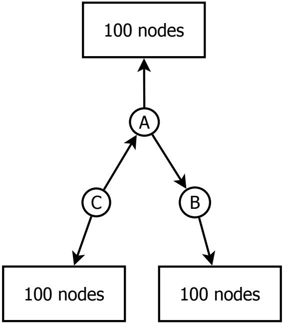
- Graph G - weighted and directed
- p_{uv} = probability with which node u can influence node v
- Diffusion starts synchronously at time step 0 with seed set and proceeds in discrete time steps
- In each time step, recently activated nodes influence their neighbors with probabilities associated with the edges
- Diffusion concludes when no further nodes can be activated

Motivation for Multi-phase Diffusion (An Example)

Total budget = Total number of seed nodes = $k = 2$



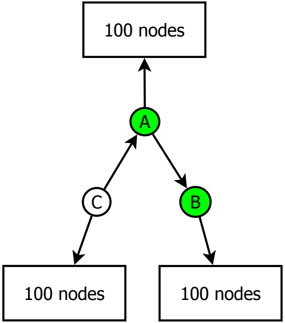
(a) input graph



(b) live graph

Motivation for Multi-phase Diffusion (Case 1)

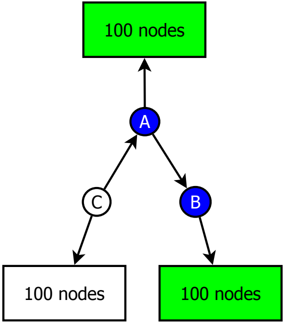
Single phase (budget = 2): A and B selected as seed nodes



$t = 0$

Motivation for Multi-phase Diffusion (Case 1)

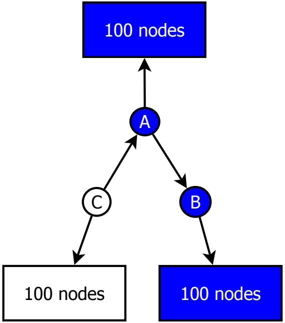
Single phase (budget $= 2$): A and B selected as seed nodes



$t = 1$

Motivation for Multi-phase Diffusion (Case 1)

Single phase (budget = 2): A and B selected as seed nodes



$t = 2$

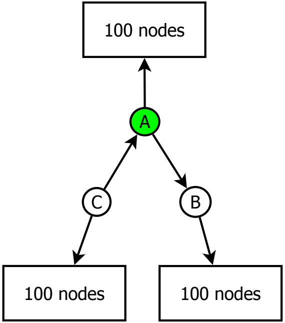
Motivation for Multi-phase Diffusion (Case 1)

Diffusion stops at $t = 1$ with $\mathbf{202}$ influenced nodes

Can we do better?

Motivation for Multi-phase Diffusion (Case 2)

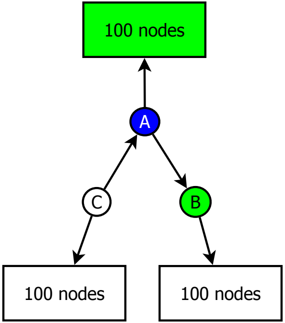
A selected as seed node at time step 0



$t = 0$

Motivation for Multi-phase Diffusion (Case 2)

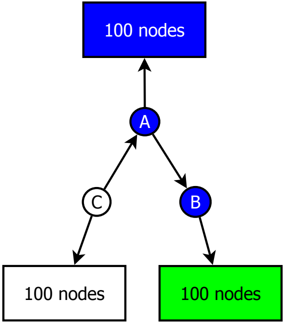
A selected as seed node at time step 0



$t = 1$

Motivation for Multi-phase Diffusion (Case 2)

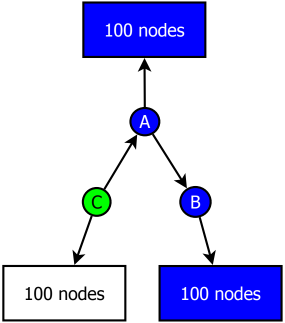
A selected as seed node at time step 0



$t = 2$

Motivation for Multi-phase Diffusion (Case 2)

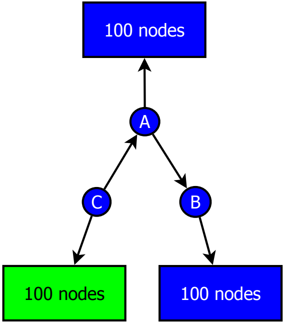
C selected as seed node at time step 3



$t = 3$

Motivation for Multi-phase Diffusion (Case 2)

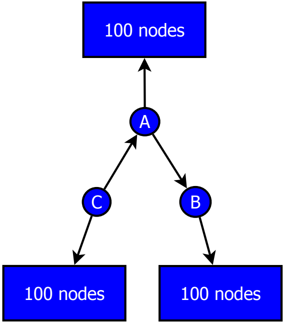
C selected as seed node at time step 3



$t = 4$

Motivation for Multi-phase Diffusion (Case 2)

C selected as seed node at time step 3



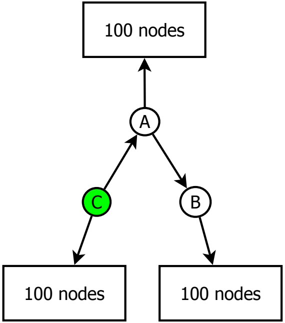
$t = 5$

Diffusion stops at $t = 4$ with **303** influenced nodes

More influenced nodes with the same budget!

Motivation for Multi-phase Diffusion (Case 3)

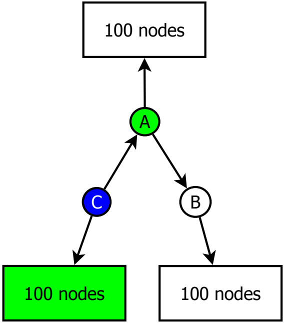
C selected as seed node at time step 0



$t = 0$

Motivation for Multi-phase Diffusion (Case 3)

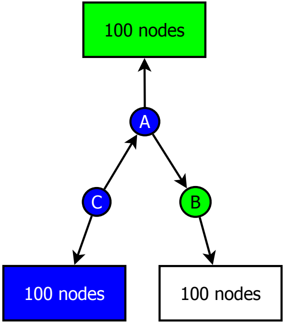
C selected as seed node at time step 0



$t = 1$

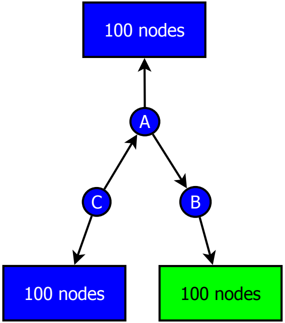
Motivation for Multi-phase Diffusion (Case 3)

C selected as seed node at time step 0



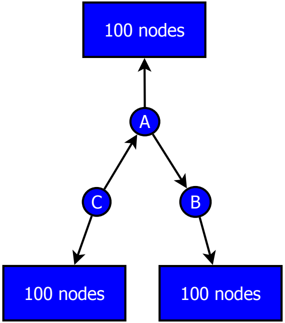
$t = 2$

C selected as seed node at time step 0



$t = 3$

C selected as seed node at time step 0



$t = 4$

Motivation for Multi-phase Diffusion (Case 3)

Diffusion stops at $t = 3$ with $\mathbf{303}$ influenced nodes

Target achieved with a reduced budget!

Challenges in Multi-phase Diffusion

- Advantage: Seed set can be chosen based on the observed spread, thus having more certainty while selecting the seed set
- Disadvantage: The diffusion may be slower, leading to compromise of time, owing to the delayed seed selection

Overview

Problem Statement

Problem Statement

Given the objective of influence maximization in a social network in two phases, what should be the budget split and the delay between the two phases, and how to select the seed nodes?

S. Dhamal, Prabuchandran K. J., and Y. Narahari, A multi-phase approach for improving information diffusion in social networks. In AAMAS, pages 1787–1788. IFAAMAS, 2015.

Two-phase Objective Function

$$\sigma^X(S) = \text{number of nodes reachable from set } S \text{ in live graph } X$$

$$S_1 = \text{seed set for the first phase}$$

$$\mathcal{A}^Y = \text{already influenced nodes },\ \mathcal{R}^Y = \text{recently influenced nodes}$$

$$S_2^{OPT(Y,k_2)} = \text{optimal set of } k_2 \text{ seed nodes, given observation } Y = (X, S_1, d, k_2)$$

$$f(S_1) = \sum_Y p(Y)\{|\mathcal{A}^Y| + \sum_X p(X|Y)\sigma^{X\setminus\mathcal{A}^Y}(\mathcal{R}^Y \cup S_2^{OPT(X,S_1,d,k_2)})\}$$

Two-phase Objective Function

$\sigma^X(S)$ = number of nodes reachable from set S in live graph X

S_1 = seed set for the first phase

$\mathcal{A}^Y$ = already influenced nodes , $\mathcal{R}^Y$ = recently influenced nodes

$S_2^{OPT(Y,k_2)}$ = optimal set of k_2 seed nodes, given observation $Y = (X, S_1, d, k_2)$

$$\begin{aligned}
f(S_1) &= \sum_Y p(Y)\{|\mathcal{A}^Y| + \sum_X p(X|Y)\sigma^{X\setminus\mathcal{A}^Y}(\mathcal{R}^Y \cup S_2^{OPT(X,S_1,d,k_2)})\} \\
&= \sum_Y p(Y) \sum_X p(X|Y)\{|\mathcal{A}^Y| + \sigma^{X\setminus\mathcal{A}^Y}(\mathcal{R}^Y \cup S_2^{OPT(X,S_1,d,k_2)})\} \\
&= \sum_Y p(Y) \sum_X p(X|Y)\sigma^X(S_1 \cup S_2^{OPT(X,S_1,d,k_2)}) \\
&= \sum_X \sum_Y p(Y)p(X|Y)\sigma^X(S_1 \cup S_2^{OPT(X,S_1,d,k_2)})
\end{aligned}$$

$$f(S_1) = \sum_X p(X)\sigma^X(S_1 \cup S_2^{OPT(X,S_1,d,k_2)})$$

Properties of Two-phase Objective Function

- ✓ Non-negative
- ✓ Monotone increasing
- ✓ Subadditive
  $f(S \cup T) \leq f(S) + f(T), \quad \forall S, T \subseteq N$

Properties of Two-phase Objective Function

✓ Non-negative

✓ Monotone increasing

✓ Subadditive
$f(S \cup T) \leq f(S) + f(T), \quad \forall S, T \subseteq N$

✗ Submodular
$f(S \cup \{i\}) - f(S) \geq f(T \cup \{i\}) - f(T), \, \forall i \in N \setminus T, \forall S \subset T \subset N$

✗ Supermodular

Properties of Two-phase Objective Function

✓ Non-negative

✓ Monotone increasing

✓ Subadditive
$f(S \cup T) \leq f(S) + f(T), \quad \forall S, T \subseteq N$

✗ Submodular
$f(S \cup \{i\}) - f(S) \geq f(T \cup \{i\}) - f(T), \, \forall i \in N \setminus T, \forall S \subset T \subset N$

✗ Supermodular

Diminishing returns property (characteristic of submodularity) satisfied in most cases

Approximating Two-phase Objective Function

- Generalized Degree Discount (GDD) heuristic: Until budget is exhausted, select a node having the largest value of

$$w_v = \left(\prod_{x \in \mathcal{X}} (1 - p_{xv}) \right) \left(1 + \sum_{y \in \mathcal{Y}} p_{vy} \right)$$

where $\mathcal{X}$ = in-neighbors of v already selected as seeds and $\mathcal{Y}$ = out-neighbors of v not yet selected as seeds.

Approximating Two-phase Objective Function

$$f(S_1) = \sum_X p(X)\sigma^X(S_1 \cup S_2^{OPT(X,S_1,d,k_2)})$$

✗ Impractical to compute $S_2^{OPT(X,S_1,d,k_2)}$

Approximating Two-phase Objective Function

$$f(S_1) = \sum_X p(X)\sigma^X(S_1 \cup S_2^{OPT(X,S_1,d,k_2)})$$

✗ Impractical to compute $S_2^{OPT(X,S_1,d,k_2)}$

$$g(S_1) \overset{\mathcal{MC}}{=} \sum_X p(X)\sigma^X(S_1 \cup S_2^{GREEDY(X,S_1,d,k_2)})$$

✗ Expensive to compute $S_2^{GREEDY(X,S_1,d,k_2)}$

Approximating Two-phase Objective Function

$$f(S_1) = \sum_X p(X)\sigma^X(S_1 \cup S_2^{OPT(X,S_1,d,k_2)})$$

✗ Impractical to compute $S_2^{OPT(X,S_1,d,k_2)}$

$$g(S_1) \overset{\mathcal{MC}}{=} \sum_X p(X)\sigma^X(S_1 \cup S_2^{GREEDY(X,S_1,d,k_2)})$$

✗ Expensive to compute $S_2^{GREEDY(X,S_1,d,k_2)}$

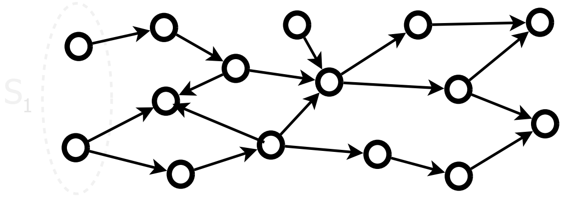
$$h(S_1) \overset{\mathcal{MC}}{=} \sum_X p(X)\sigma^X(S_1 \cup S_2^{GDD(X,S_1,d,k_2)})$$

✓ $$f(\cdot) \overset{\text{theoretically}}{\approx\approx\approx\approx\approx} g(\cdot) \overset{\text{empirically}}{\approx\approx\approx\approx\approx} h(\cdot)$$

A General Algorithm for Two-phase Influence Maximization

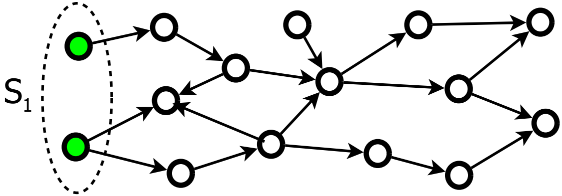
$$k_1, k_2 = \text{budgets for first and second phases}$$
$$d = \text{delay after which the second phase starts}$$
$$h(\cdot) = \text{approximate objective function (farsighted)}$$
$$\sigma(\cdot) = \text{single phase objective function (myopic)}$$

A General Algorithm for Two-phase Influence Maximization



Input: $G = (N, E, \mathcal{P})$, k_1, k_2, d

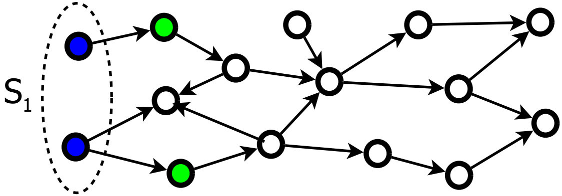
A General Algorithm for Two-phase Influence Maximization



Input: $G = (N, E, \mathcal{P})$, k_1, k_2, d

1: Find set S_1 of size k_1 that maximizes $h(S_1)$ or $\sigma(S_1)$

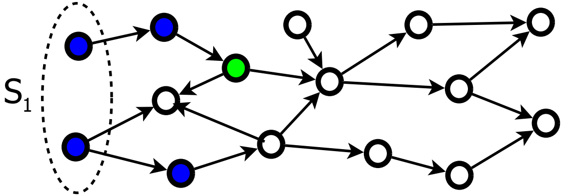
A General Algorithm for Two-phase Influence Maximization



Input: $G = (N, E, \mathcal{P})$, k_1, k_2, d

1: Find set S_1 of size k_1 that maximizes $h(S_1)$ or $\sigma(S_1)$
First phase:
2: Run the diffusion until time step d

A General Algorithm for Two-phase Influence Maximization



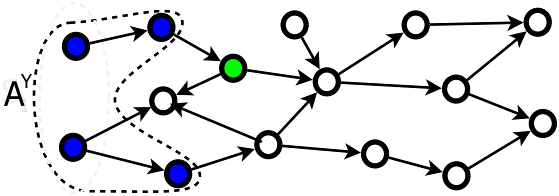
Input: $G = (N, E, \mathcal{P})$, k_1, k_2, d

1: Find set S_1 of size k_1 that maximizes $h(S_1)$ or $\sigma(S_1)$

First phase:

2: Run the diffusion until time step d

A General Algorithm for Two-phase Influence Maximization

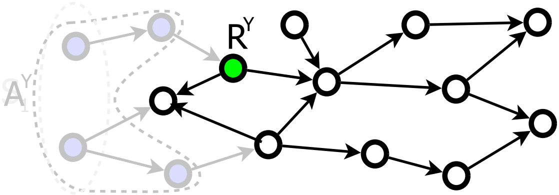


Input: $G = (N, E, \mathcal{P})$, k_1, k_2, d

1: Find set S_1 of size k_1 that maximizes $h(S_1)$ or $\sigma(S_1)$

First phase:

2: Run the diffusion until time step d



Input: $G = (N, E, \mathcal{P})$, k_1, k_2, d

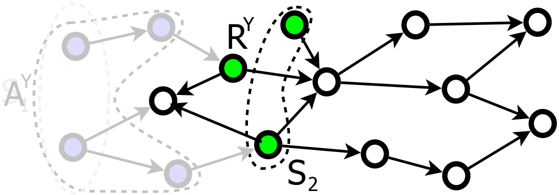
1: Find set S_1 of size k_1 that maximizes $h(S_1)$ or $\sigma(S_1)$

First phase:

2: Run the diffusion until time step d

3: At time step d, form G^d by removing already influenced nodes

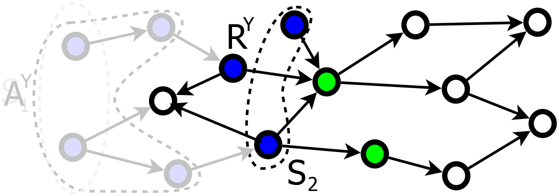
A General Algorithm for Two-phase Influence Maximization



Input: $G = (N, E, \mathcal{P})$, k_1, k_2, d

1: Find set S_1 of size k_1 that maximizes $h(S_1)$ or $\sigma(S_1)$
First phase:
2: Run the diffusion until time step d
3: At time step d, form G^d by removing already influenced nodes
4: Find set S_2 of size k_2 that maximize $\sigma(S_2 \cup \mathcal{R}^Y)$ on G^d (assuming $\mathcal{R}^Y$ already forms the partial seed set)

A General Algorithm for Two-phase Influence Maximization



Input: $G = (N, E, \mathcal{P})$, k_1, k_2, d

1: Find set S_1 of size k_1 that maximizes $h(S_1)$ or $\sigma(S_1)$

First phase:

2: Run the diffusion until time step d

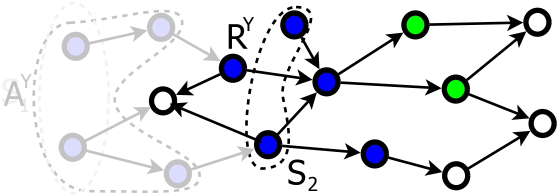
3: At time step d, form G^d by removing already influenced nodes

4: Find set S_2 of size k_2 that maximize $\sigma(S_2 \cup \mathcal{R}^Y)$ on G^d (assuming $\mathcal{R}^Y$ already forms the partial seed set)

Second phase:

5: Continue diffusion until no further nodes can be influenced

A General Algorithm for Two-phase Influence Maximization



Input: $G = (N, E, \mathcal{P})$, k_1, k_2, d

1: Find set S_1 of size k_1 that maximizes $h(S_1)$ or $\sigma(S_1)$

First phase:

2: Run the diffusion until time step d

3: At time step d, form G^d by removing already influenced nodes

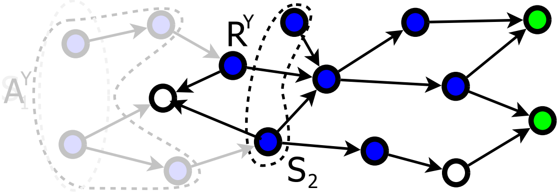
4: Find set S_2 of size k_2 that maximize $\sigma(S_2 \cup \mathcal{R}^Y)$ on G^d (assuming $\mathcal{R}^Y$ already forms the partial seed set)

Second phase:

5: Continue diffusion until no further nodes can be influenced

---

A General Algorithm for Two-phase Influence Maximization



Input: $G = (N, E, \mathcal{P})$, k_1, k_2, d

1: Find set S_1 of size k_1 that maximizes $h(S_1)$ or $\sigma(S_1)$

First phase:

2: Run the diffusion until time step d

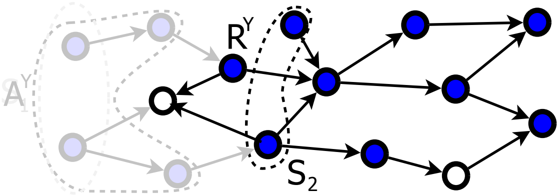
3: At time step d, form G^d by removing already influenced nodes

4: Find set S_2 of size k_2 that maximize $\sigma(S_2 \cup \mathcal{R}^Y)$ on G^d (assuming $\mathcal{R}^Y$ already forms the partial seed set)

Second phase:

5: Continue diffusion until no further nodes can be influenced

A General Algorithm for Two-phase Influence Maximization



Input: $G = (N, E, \mathcal{P})$, k_1, k_2, d

1: Find set S_1 of size k_1 that maximizes $h(S_1)$ or $\sigma(S_1)$

First phase:

2: Run the diffusion until time step d

3: At time step d, form G^d by removing already influenced nodes

4: Find set S_2 of size k_2 that maximize $\sigma(S_2 \cup \mathcal{R}^Y)$ on G^d (assuming $\mathcal{R}^Y$ already forms the partial seed set)

Second phase:

5: Continue diffusion until no further nodes can be influenced

Overview

Results on Expected Influence using Multiple Phases

Observation

Myopic methods perform at par with farsighted, but run orders of magnitude faster

Results on Expected Influence using Multiple Phases

Observation

Myopic methods perform at par with farsighted, but run orders of magnitude faster

Theorem

For any given values of k_1 and k_2, two phase $\succ$ single phase under optimal seed selection

Results on Expected Influence using Multiple Phases

Observation

Myopic methods perform at par with farsighted, but run orders of magnitude faster

Theorem

For any given values of k_1 and k_2, two phase $\succ$ single phase under optimal seed selection

Theorem

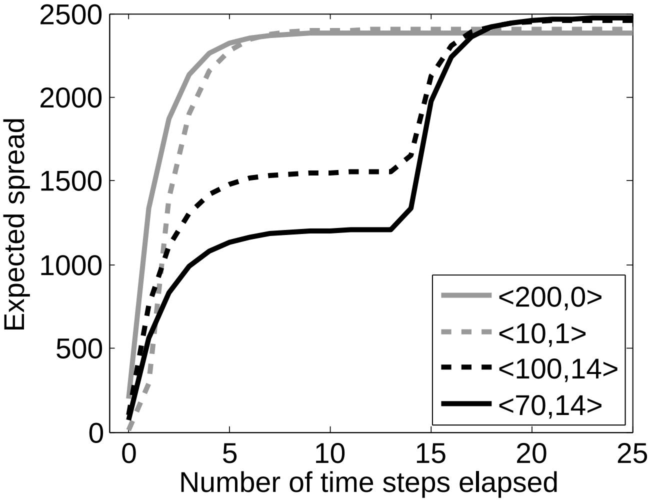
For any given values of k_1 and k_2, two-phase influence is maximized when $d = D$ (longest path in the network)

Improvements for Different Values of k

Model	$k \rightarrow$	50	100	200	300
WC	% Improvement ($k_1 = k_2$)	3.5	1.8	3.5	4.4
	Opt. % improvement	4.5	2.0	4.0	4.5
	Optimizing k_1	15	35	70	105
TV	% improvement ($k_1 = k_2$)	5.0	5.4	5.4	4.8
	Opt. % improvement	6.0	6.0	6.0	5.0
	Optimizing k_1	18	35	70	105

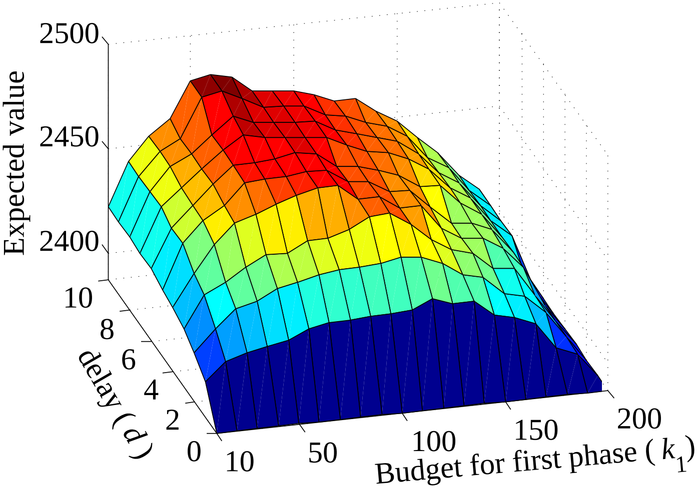
NetHEPT dataset

Progression with Different $< k_1, d >$ Pairs ($k = 200$)



NetHEPT dataset under WC model

Influence for Different $< k_1, d >$ Pairs ($k = 200$)



NetHEPT dataset under WC model

The Other Side of the Coin

Now ...

1. Why not use two-phase diffusion all the time?
2. Why not wait for the first phase to complete its diffusion process before starting the second phase?

[5] P. De Boer, D. Kroese, S. Mannor, and R. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67, 2005.

The Other Side of the Coin

Now ...

1. Why not use two-phase diffusion all the time?
2. Why not wait for the first phase to complete its diffusion process before starting the second phase?

The value of diffusion decays with time
We consider decay function δ^t where $\delta \in [0, 1]$

[5] P. De Boer, D. Kroese, S. Mannor, and R. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67, 2005.

Now ...

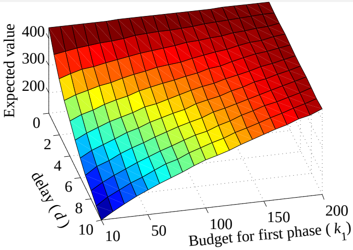
1. Why not use two-phase diffusion all the time?
2. Why not wait for the first phase to complete its diffusion process before starting the second phase?

> The value of diffusion decays with time
> We consider decay function δ^t where $\delta \in [0, 1]$

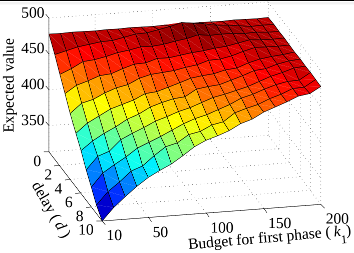
- Can simultaneously optimize over k_1, d, and seed sets using cross entropy method [5]

[5] P. De Boer, D. Kroese, S. Mannor, and R. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67, 2005.

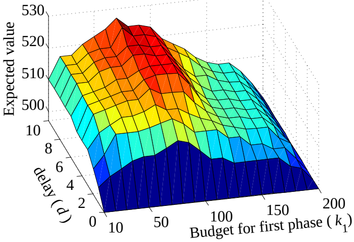
3D plots for Different Values of δ



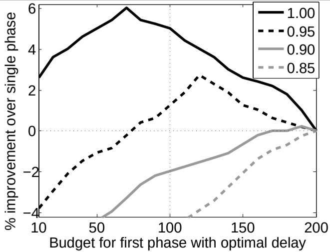
$\delta = 0.85$



$\delta = 0.95$



$\delta = 1$



2D view for best d($\geq$ 1)

NetHEPT dataset under TV model ($k = 200$)

Usage of Golden Section Search

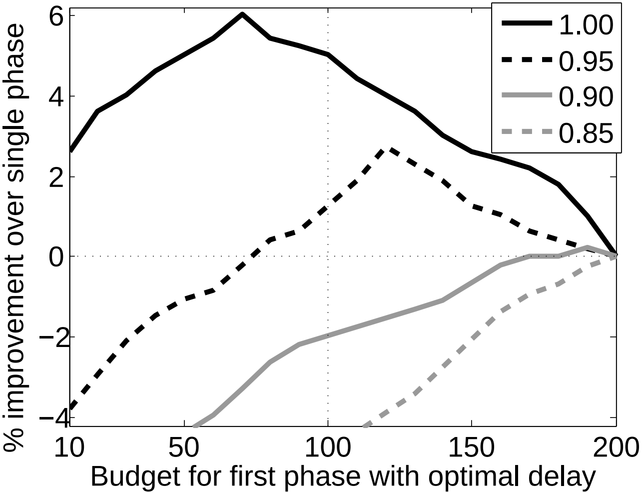


Figure: Unimodal Nature of Plots

Overview

Conclusion

- Strict temporal constraints: single phase diffusion
- Moderate temporal constraints: two-phase diffusion with a short delay while allocating most of the budget to first phase
- No temporal constraints: two-phase diffusion with a long enough delay with $1/3$ budget for the first phase

  Trade-off between
  - the size of the observed diffusion
  - exploitation based on the observed diffusion
  - a skew towards lower values of k_1 because initial highly influential seed nodes contribute to most diffusion

Future Work

- Harness multi-phase diffusion to get a desired expected spread with a reduced budget
- Study under a most realistic (lenient) decay function
- Study diffusion in more than two phases and compare their performances against two-phase diffusion
- Study multi-phase diffusion using other diffusion models
- Study equilibria in a game theoretic setting where multiple campaigns consider the possibility of multi-phase diffusion