CT-IC: Continuously activated and Time-restricted Independent Cascade Model for Viral Marketing

Wonyeol Lee Jinha Kim Hwanjo Yu

Department of Computer Science & Engineering

POSTECH POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY, Korea

ICDM 2012

Viral Marketing

Influence Maximization Problem

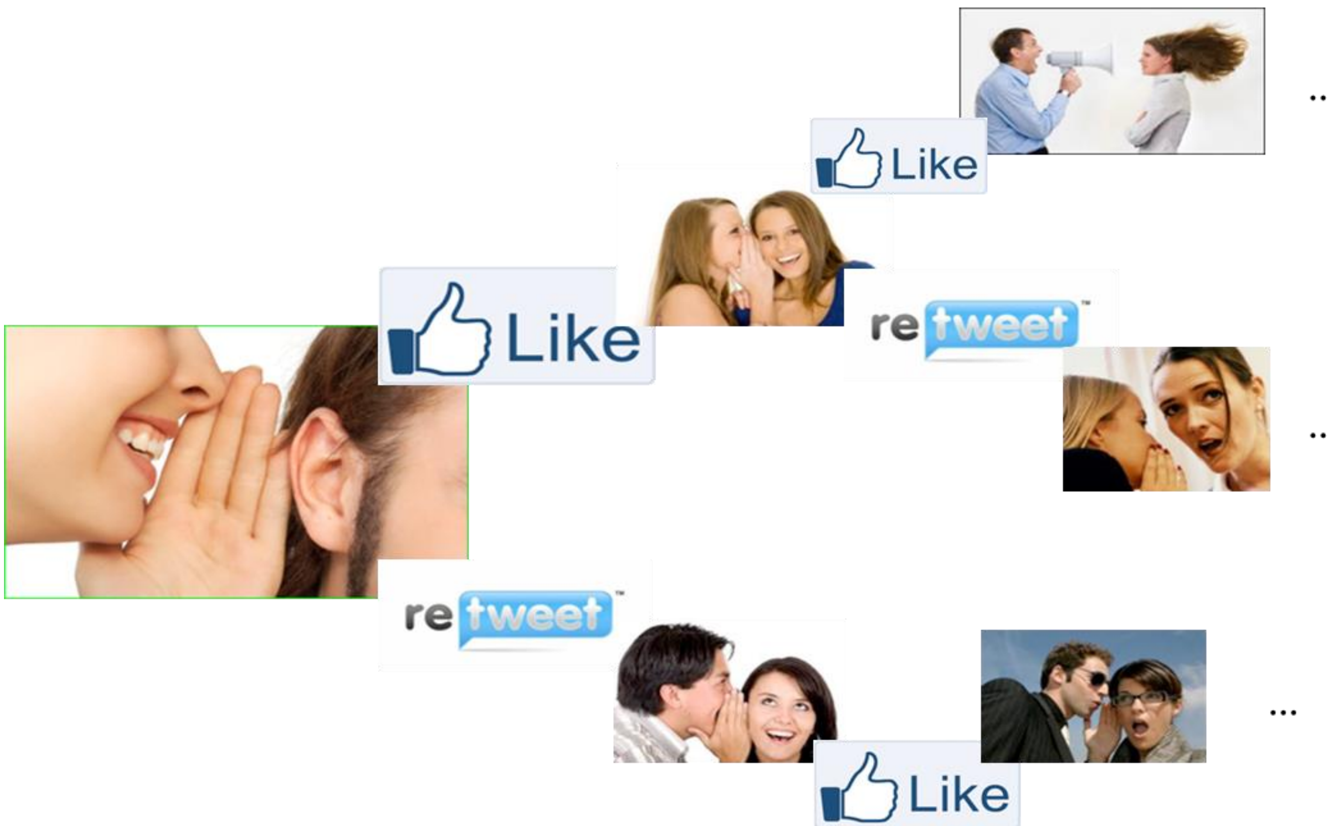
Influence Diffusion Models

Limitations of Existing Models

Introduction & Motivation

Viral Marketing

- Word of mouth effect > TV advertising



Influence Maximization Problem [KDD'03]

$$\sigma(S)$$

the expected number of people influenced by a seed set S

$$\arg \max_{S \subseteq V, |S|=k} \sigma(S)$$

Given a network $G = (V, E)$, and a budget k,
find the k most influential people in a social network

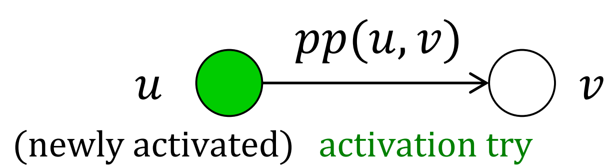
$\sigma(S)$ Depends On …

How influence is propagated through a graph

= Influence Diffusion Model

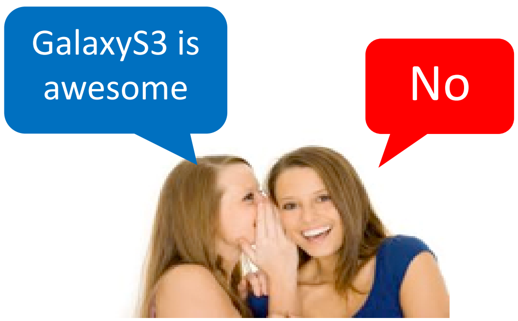
- We need a "*realistic*" diffusion model to apply influence maximization problem to a "*real-world*" marketing.

- Existing diffusion models
  - IC (Independent Cascade) model [KDD'03]
  - LT (Linear Threshold) model [KDD'03]

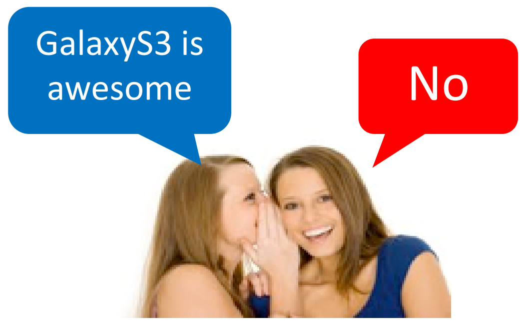


Existing Models Ignore … (1)

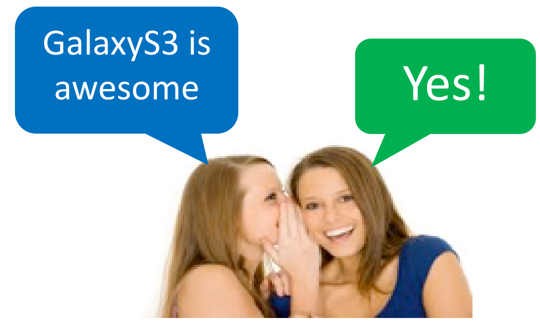
- An individual can affect others *multiple* times.



Yesterday



Today



Tomorrow

  - NOT contained in “IC model.”

Existing Models Ignore … (2)

- Marketing usually has a *deadline*.

Yesterday

Today

Tomorrow

  – NOT contained in “all previous models.”

CT-IC model

Properties of CT-IC model

CT-IPA algorithm

Our Contributions

1) CT-IC model

- We propose a new influence diffusion model "CT-IC" for viral marketing, which generalizes previous models such that
  - An individual can affect others *multiple* times.
  - Marketing has a *deadline*.

$$\boxed{pp_t(u,v)} = pp_0(u,v)\ f_{uv}(t)$$

$$\arg \max_{S \subseteq V, |S|=k} \boxed{\sigma(S,T)}$$

- *An efficient algorithm for influence maximization problem under CT-IC model?*

Greedy Algorithm [KDD'03]

- Influence maximization even under IC model is NP-Hard.

- Greedy algorithm:
  - Repeatedly select the node which gives the most marginal gain of $\sigma(S)$

Algorithm 1 Greedy(G, k)

$S = \phi$
for $i = 1$ **to** k **do**
  $u = \arg\max_{v \in V \setminus S} \sigma(S \cup \{v\}) - \sigma(S)$
  $S = S \cup \{u\}$
end for
return S

- *Theorem:*
  $\sigma(S)$ satisfies non-negativity, monotonicity, submodularity
  $\Rightarrow$ *Greedy* guarantees approximation ratio $(1 - 1/e)$.

- *CT-IC model satisfies these properties?*

2) Properties of CT-IC model

- <u>We prove the *Theorem:*</u> In CT-IC model, $\sigma(\cdot, t)$ satisfies non-negativity, monotonicity, and submodularity.
  - Non-negativity: $\sigma(S, t) \geq 0$
  - Monotonicity: $\sigma(S, t) \leq \sigma(S', t)$ for any $S \subseteq S'$
  - Submodularity: $\sigma(S \cup \{v\}, t) - \sigma(S, t) \geq$
    $\sigma(S' \cup \{v\}, t) - \sigma(S', t)$ for any $S \subseteq S'$
- Thus, *Greedy* guarantees approximation ratio $(1 - 1/e)$ even under CT-IC model.
- *An efficient method for computing $\sigma(S, T)$ under CT-IC model?*

3) CT-IPA algorithm

- Difficulties for computing $\sigma(S,T)$ under CT-IC model
  - Monte Carlo simulation is not scalable. [KDD'10]
  - Evaluating $\sigma(S)$ is #P-Hard even under IC model. [KDD'10]
  - We show that it is difficult to extend *PMIA* (the state-of-the-art algorithm for IC model) to CT-IC model!

- We propose *"CT-IPA"* algorithm (an extension of *IPA* [ICDE'13]) for calculating $\sigma(S,T)$ under CT-IC model.

Lemma 3: The probability that $u \in S$ activates $v \in V \backslash S$ only through a path $p = (u = u_0, u_1, \cdots, u_{l-1}, u_l = v)$ is

$$inf_p(u,v) = [1\ 0\ \cdots\ 0]\left(\prod_{i=0}^{l-1} \mathbf{C}_{u_i u_{i+1}}\right)[1\ 1\ \cdots\ 1]^{\mathrm{Tr}}, \quad (2)$$

where $u_i \in V \backslash S$ for all $i = 1, \cdots, l$, and the order of matrix multiplication is from $i = 0$ to $l - 1$.

Experiments

Dataset

- We use four real networks:

Table I
BASIC INFORMATION OF FOUR REAL DATASET.

Dataset	HEP	PHY	EPINION	AMAZON
Directedness	Undir	Undir	Dir	Dir
# of Nodes	15K	37K	76K	262K
# of Edges	59K	232K	509K	1235K
# of Connected Components	1781	3883	2	1
Average Size of Components	8.6	9.6	38K	262K
θ for CT-IPA	$1/32$	$1/64$	$1/64$	$1/16$

Characteristic of CT-IC model (1)

- Model comparison between IC & CT-IC models:

Table II
TOP-20 SEED NODES OF IC MODEL AND CT-IC MODEL SOLUTION.

(a) On PHY

IC model solution	4840	1568	5192	5120	7387
	12081	2356	10653	4115	23571
	3460	3808	969	809	5567
	2443	3566	5312	6342	3673
CT-IC model solution	4840	5192	5120	1568	809
	4115	2356	3460	23571	12081
	7132	**3842**	10653	**4109**	3673
	6342	**3712**	**2928**	**3982**	**2289**

(b) On AMAZON

IC model solution	17747	222839	25699	18076	168039
	18337	232448	7266	11129	45391
	176067	9657	64815	183084	27562
	59541	14461	238375	114241	1385
CT-IC model solution	17747	176067	**56415**	**51234**	**200657**
	238375	18076	**236670**	**259011**	222839
	6290	**205434**	**143531**	**199539**	**59541**
	25699	**178335**	**82533**	**114241**	**95315**

Characteristic of CT-IC model (2)

- Effect of marketing time constraint T:

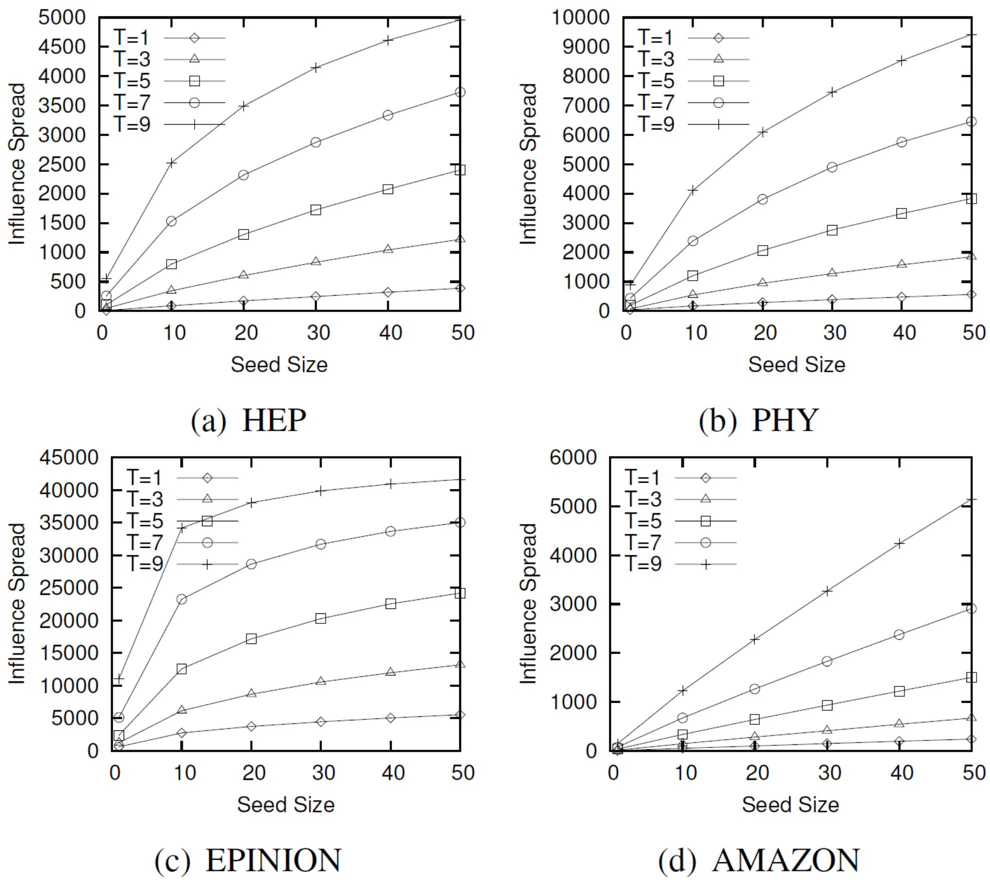


Figure 3. The change of influence spread with respect to T.

Algorithm Comparison (1)

- Comparison of influence spread:

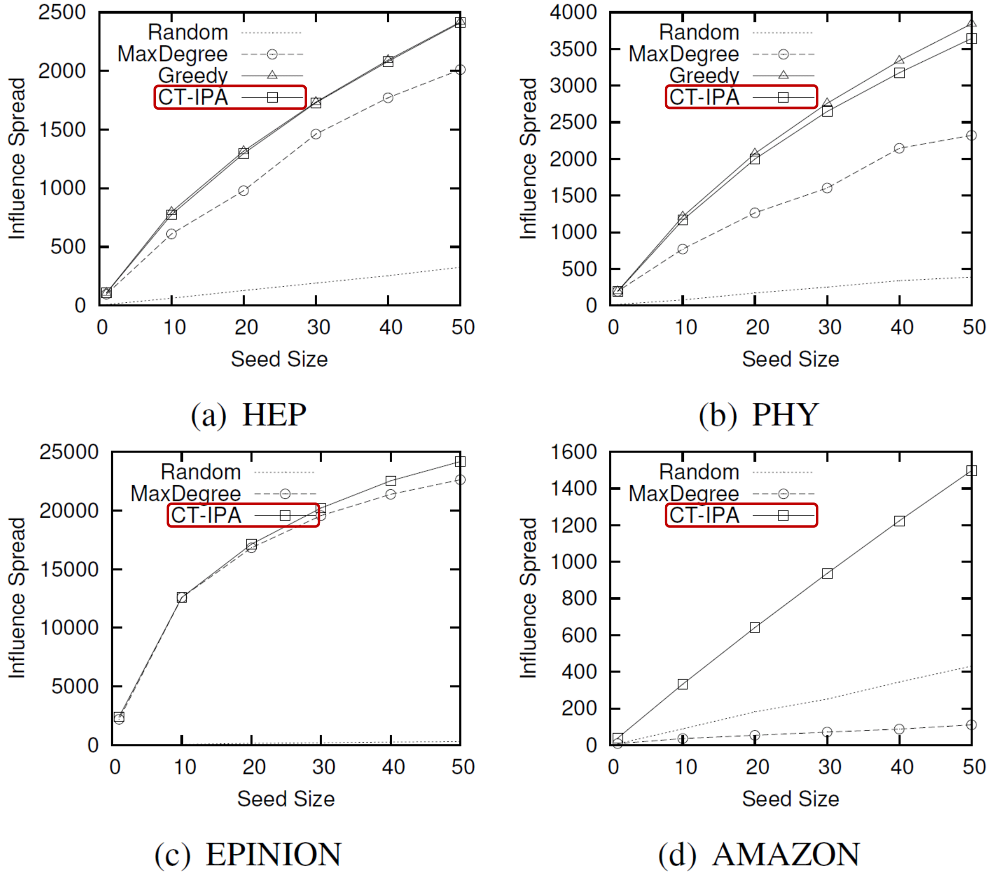


Figure 4. Influence spread of various algorithms.

Algorithm Comparison (2)

- Comparison of processing time:

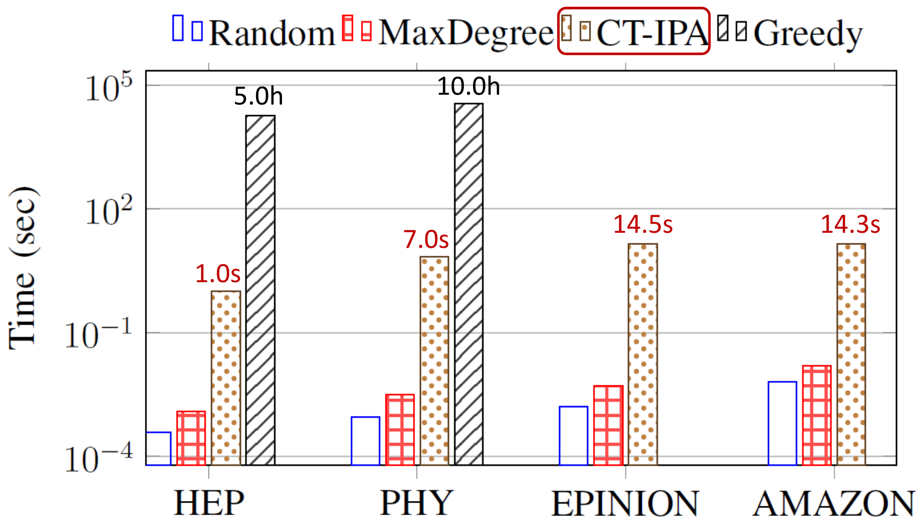


Figure 5. Processing time of various algorithms.

  – *CT-IPA* is four orders of magnitude faster than *Greedy* while providing similar influence spread to *Greedy*.

Conclusion

Conclusion

Existing diffusion models ignore important aspects of real marketing.

1) Propose a *realistic* influence diffusion model "CT-IC" for viral marketing.

2) Prove that CT-IC model satisfies non-negativity, monotonicity, and submodularity.

3) Propose a scalable algorithm *"CT-IPA"* for CT-IC model.

Thank You!

Supplements

CT-IC model & Other Diffusion models

- Relationship between influence diffusion models:

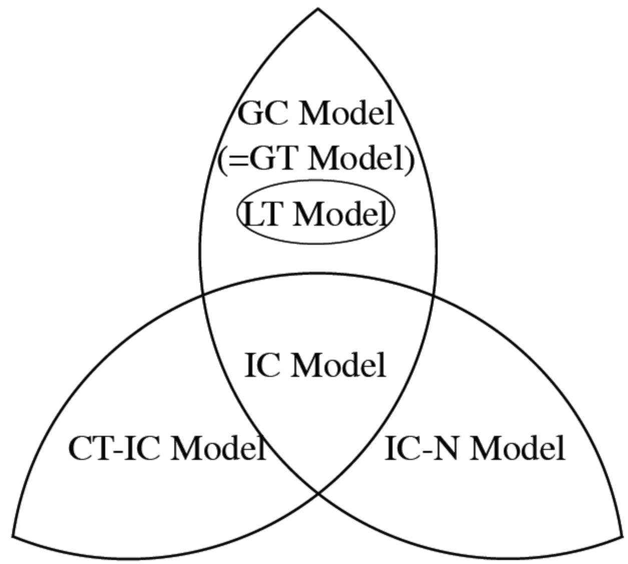


Figure 1. Relationship between influence diffusion models.

Properties of CT-IC model (1)

- Difference between IC & CT-IC models:

  Lemma 1: For any positive k, N, T such that $k < N/4$, $T < (N/4k) - 1 = O(N/k)$, there exists a graph $G = (V, E)$ such that $|V| = N$ and $dr(G, k, T) = \Omega(N/kT)$.

  – Here, given $G = (V, E), k, T$, *difference ratio* $dr(G, k, T)$ is defined by

  $$dr(G,k,T) = \frac{\sigma(S_T^*, T)}{\max\{\sigma(S_I^*, T) | S_I^* \in \mathfrak{S}_I^*\}} \geq 1,$$

  where $\mathfrak{S}_I^*(G,k) = \arg\max\{\sigma_I(S) | S \subseteq V, |S| = k\},$

  $$\mathfrak{S}_T^*(G,k) = \arg\max\{\sigma(S,T)| \; S \subseteq V, |S| = k\}, \; S_T^* \in \mathfrak{S}_T^*.$$

- The Lemma tells us that
  "For some graphs, CT-IC model is largely different from IC model."

Properties of CT-IC model (2)

- Maximum probability path:

Lemma 4: For some graph $G = (V, E)$, $u, v \in V$, $S \subseteq V$, and T, there exists a path $p = (u = u_0, u_1, \cdots, u_l = v)$ and $i \in \{0, \cdots, l-1\}$ such that while p is a maximum probability path, $(u_0, \cdots, u_i)$ or $(u_{i+1}, \cdots, u_l)$ is not a maximum probability path.

  - Here, p^* is called a *maximum probability path* from u to v if

$$p^* \in \arg\max_p \{inf_p(u, v) | p : \text{a simple path from } u \text{ to } v\}.$$

- The Lemma tells us that
  "It is difficult to generalize *PMIA* algorithm into CT-IC model."

Characteristic of CT-IC model

- Model comparison between IC & CT-IC models:

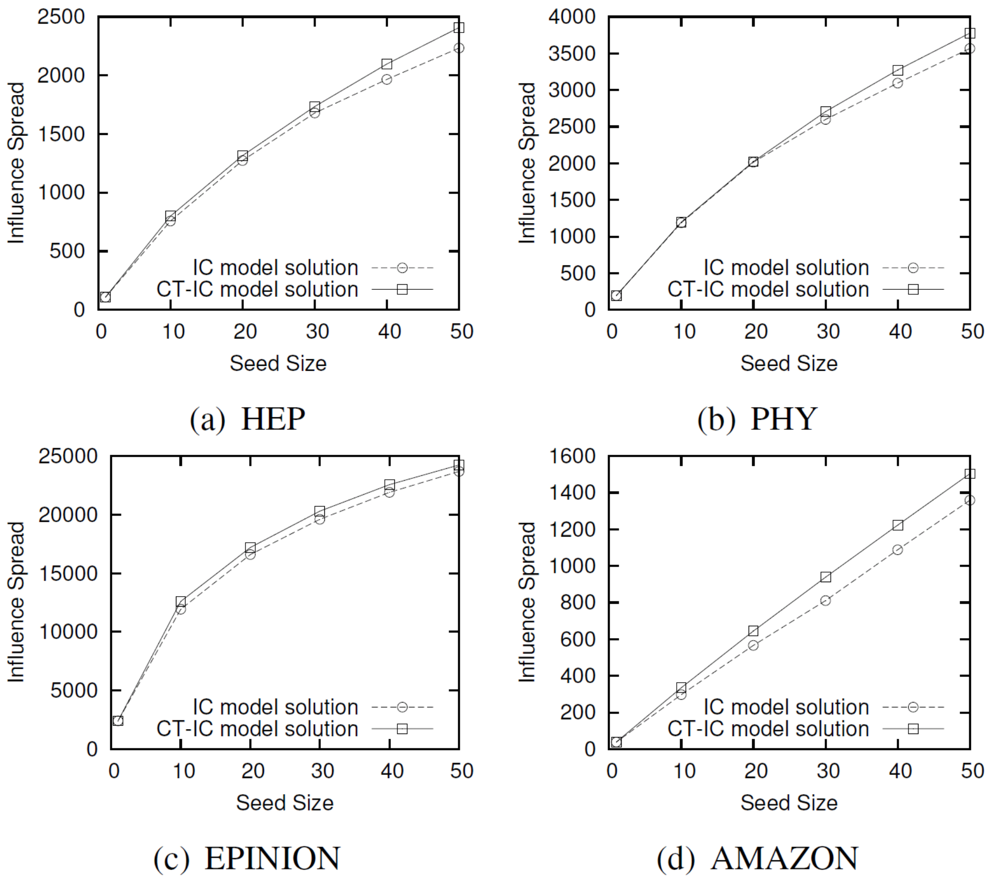


Figure 2. Comparison between IC and CT-IC models.

Exact Computation of Influence Spread (1)

- Case of an arborescence:

Lemma 2: For any $v \in V \backslash S$ and $0 < t \leq T$,

$$ap_S(v,t) = \prod_{u \in N_{in}(v)} \left[1 - \sum_{i=0}^{t-2} ap_S(u,i) g_{uv}(t-2-i)\right] - \prod_{u \in N_{in}(v)} \left[1 - \sum_{i=0}^{t-1} ap_S(u,i) g_{uv}(t-1-i)\right]$$

holds, where $g_{uv}(t) = 1 - \prod_{i=0}^{t} [1 - pp_i(u,v)]$.

where $ap_S(v,t)$ is the probability that v is activated *exactly at* time t by S.

Exact Computation of Influence Spread (2)

- Case of a simple path:

Lemma 3: The probability that $u \in S$ activates $v \in V \backslash S$ only through a path $p = (u = u_0, u_1, \cdots, u_{l-1}, u_l = v)$ is

$$inf_p(u,v) = [1\ 0\ \cdots\ 0] \left(\prod_{i=0}^{l-1} \mathbf{C}_{u_i u_{i+1}} \right) [1\ 1\ \cdots\ 1]^{\mathrm{Tr}}, \quad (2)$$

where $u_i \in V \backslash S$ for all $i = 1, \cdots, l$, and the order of matrix multiplication is from $i = 0$ to $l - 1$.

where $inf_p(u,v)$ is the probability that u activates v in time T along a path p,

$$\mathbf{C}_{uv} = \begin{bmatrix} 0 & c_{uv}^{(1)} & \cdots & c_{uv}^{(T)} \\ 0 & 0 & \cdots & c_{uv}^{(T-1)} \\ 0 & 0 & \cdots & c_{uv}^{(T-2)} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \quad c_{uv}^{(t-i)} = pp_{t-i-1}(u,v) \prod_{j=0}^{t-i-2} (1 - pp_j(u,v)).$$

Exact Computation of Influence Spread (3)

- Case of a simple path: (proof)

By Lemma 2,

$$ap(v,t) = \sum_{i=0}^{t-1} c_{uv}^{(t-i)} ap(u,i)$$

$$= \begin{bmatrix} ap(u,0) \\ ap(u,1) \\ \vdots \\ ap(u,t-1) \end{bmatrix}^{\mathrm{Tr}} \begin{bmatrix} c_{uv}^{(t)} \\ c_{uv}^{(t-1)} \\ \vdots \\ c_{uv}^{(1)} \end{bmatrix},$$

$$c_{uv}^{(t-i)} = pp_{t-i-1}(u,v) \prod_{j=0}^{t-i-2} (1 - pp_j(u,v)).$$

By gathering in a matrix,

$$\begin{bmatrix} ap(v,0) \\ ap(v,1) \\ ap(v,2) \\ \vdots \\ ap(v,T) \end{bmatrix}^{\mathrm{Tr}} = \begin{bmatrix} ap(u,0) \\ ap(u,1) \\ ap(u,2) \\ \vdots \\ ap(u,T) \end{bmatrix}^{\mathrm{Tr}} \begin{bmatrix} 0 & c_{uv}^{(1)} & \cdots & c_{uv}^{(T)} \\ 0 & 0 & \cdots & c_{uv}^{(T-1)} \\ 0 & 0 & \cdots & c_{uv}^{(T-2)} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$\mathbf{AP}(v) = \mathbf{AP}(u)\mathbf{C}_{uv}$$

IPA Algorithm (1)

- Influence spread of a single node u:

$$\hat{ap}_{\{u\},T}(v) = 1 - \prod_{p \in P_{u \to v}} \left(1 - inf_p(u, v)\right)$$

$$\hat{\sigma}(\{u\}, T) = 1 + \sum_{v \in O_u} \hat{ap}_{\{u\},T}(v)$$

where $P_{u \to v} = \{p = (u, \dots, v) | inf_p(u, v) \geq \theta\}$, $O_u = \{w | P_{u \to w} \neq \phi\}$.
Here, θ is a threshold for IPA algorithm.

IPA Algorithm (2)

- Influence spread of a seed set S:

$$\hat{ap}_{S,T}(v) = 1 - \prod_{p \in P_{S \to v}} \left(1 - inf_p(u,v)\right)$$

$$\hat{\sigma}(S,T) = |S| + \sum_{v \in O_S} \hat{ap}_{S,T}(v)$$

where $P_{S\to v} = \{p = (u, \dots, v) | u \in S,\ inf_p(u,v) \geq \theta\}$, $O_S = \{w | P_{S\to w} \neq \phi\}$.
Here, θ is a threshold for IPA algorithm.