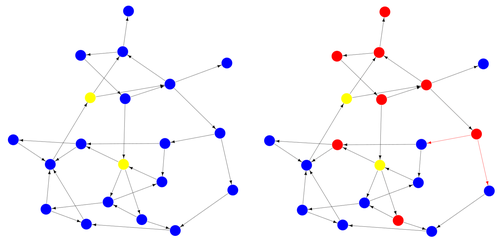
ASIM: Scalable Influence Maximization under the IC Model

Sainyam Galhotra, Akhil Arora, Srinivas Virinchi, and Shourya Roy
Xerox Research Centre India, Bangalore
{sainyam.galhotra, akhil.arora, srinivas.virinchi, shourya.roy}@xerox.com

Motivation

- Need for Influence based Modelling??
- Answer: Interpret Real-world processes
  - Spread of Diseases
  - Traffic Congestion and its propagation
  - many more ...

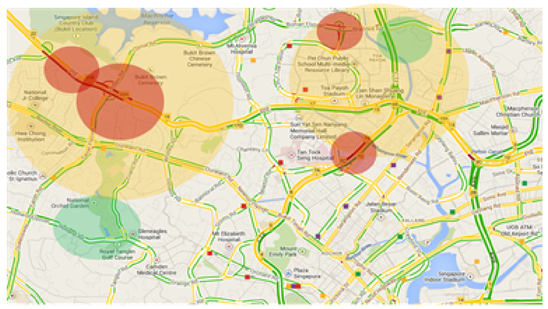


Influence Maximization

- **Task:** Identify the most-influential set of nodes
- **Constraints:** Budget ($\mathcal{B} = |S|$)
- $\mathbb{F}(\mathbf{S})$: Expected number of nodes active at the end, if set S is targeted for initial activation
- More formally, Given a budget $\mathcal{B}$, select a set S of $\mathcal{B} = |S|$ nodes, so as to maximize $\mathbb{F}(S)$

Possible Practical Applications

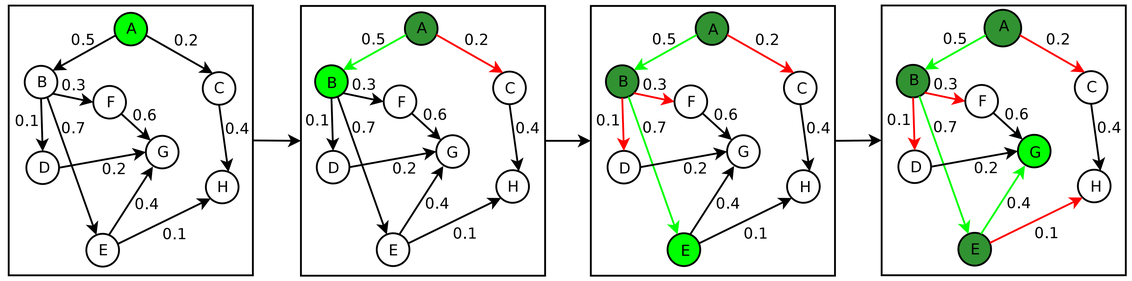
- **Viral Marketing**, Viral Ad-Targeting [VLDB'15]
- Blog Selection [KDD'07]
- **Detect and Prevent outbreaks** [KDD'07]
  - Quickly identify sources of infection spread
- Identifying key-sources for spreading traffic congestion and prevent congestion from going viral



Propagation Models

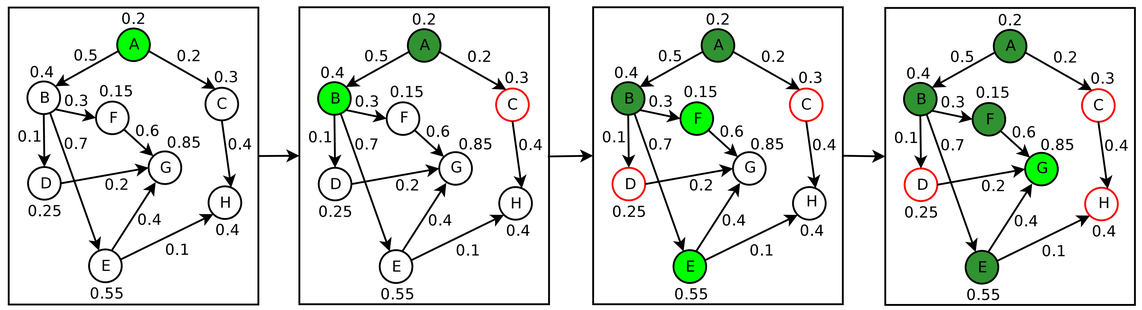
- Independent Cascade (IC) Model
- Weighted Cascade (WC) Model
- Linear Threshold (LT) Model

IC and WC Models



- **Edge Weights:**
  - For IC – $p_{(u,v)} = 0.1$ (generally) or any small constant $\in [0, 1]$
  - For WC – $p_{(u,v)} = 1/in\_deg(v)$

LT Model

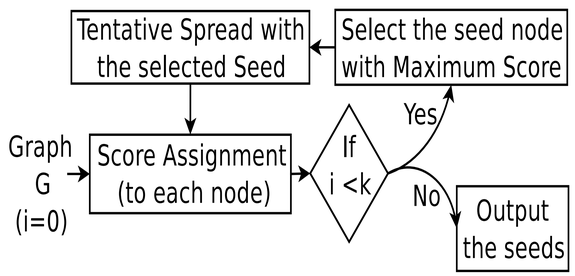


- Each node possesses an **activation threshold** $\theta_v \in [0, 1]$
- **Edge Weights:** $w_{(u,v)} \in [0, 1]$

Unaddressed Challenges!

- Extensive study on **run-time** efficiency and efficacy – [1, 2, 3], KDD'07, SODA'14, CIKM'14
- **Scalable** solutions (catering to both running-time and memory-consumption) are **non-existent**

Overview of our Approach



ASIM

- **ASIM** assigns a score to each node (u) of the graph
- **Intuition:** The probability of a node v to get activated by a seed node u is dependent upon all possible simple paths from u to v in G.

- $Score_d[u]$ $(\forall u \in V)$ is defined as the weighted sum of the number of simple paths of length at most $d \mid d \leq \mathcal{D}$ starting from u

Analysis

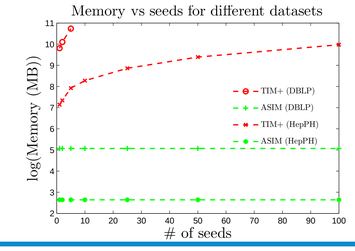
- **Time Complexity** (score assgn) – $\mathbf{O(d(m+n))}$
- Total time taken for k seeds – $\mathbf{O(kd(m+n))}$
- **Memory Complexity** – $\mathbf{O(n)}$

Result Summary

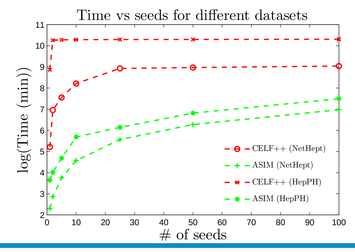
Dataset	Running Time (min)			Memory (MB)		
	CELF++	ASIM	Gain	CELF++	ASIM	Gain
NetHEPT	5352.25	648.33	**8.25x**	23.26	10.5195	2.2x
HepPh	9746.74	1355	**7.2x**	24.60	14.0391	1.75x
DBLP	88216.69	13166.67	**6.7x**	138.19	159.09	0.87x

Dataset	Running Time (min)			Memory (MB)		
	TIM	ASIM	Gain	TIM	ASIM	Gain
DBLP	783.1	6500	0.12x	35234.75	159.09	**221x**
YouTube	NA	19666.67	∞	NA	553.94	∞

Memory Consumption



Running Time



Future Directions

- Extension to the LT model
- Prove approximation guarantees

References

[1] A. Goyal, W. Lu, and L. V. Lakshmanan. Celf++: Optimizing the greedy algorithm for influence maximization in social networks. In *WWW (Companion Volume)*, pages 47–48, 2011.

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

[3] Y. Tang, X. Xiao, and Y. Shi. Influence maximization: Near-optimal time complexity meets practical efficiency. In *SIGMOD*, pages 75–86, 2014.