# Inferring Networks and Estimating Influence in Social Media

# Why is it interesting?

Basic tasks in information diffusion

1. What is the popular topics?
2. What is the network structure?
    a. Inferring underlying cascade given activation sequence. Network structure unknown.
    b. NETINF, NETRATE, INFOPATH
3. How to measure the influence of a set of nodes?
    a. Predict how a diffusion unfolds in existing network
    b. Identify influential nodes, measure the influence
    c. Independent Cascade, ConTinEst

# Agenda

- Background
- Inferring graph structure
- Estimating influence from existing graphs
- Estimating influence from unknown graphs
- Experiments
- Conclusion

# NetInf Algorithm

Propagation likelihood: $P_c(\Delta_{u,v})$

Two models:

- Exponential Model: $P_c(u,v) = P_c(\Delta_{u,v}) \propto e^{-\frac{\Delta_{u,v}}{\alpha}}$
- Power-law Model: $P_c(u,v) = P_c(\Delta_{u,v}) \propto \frac{1}{\Delta_{u,v}^{\alpha}}$

# NetInf Algorithm

For each possible diffusion tree T, we compute $P(c|T)$ :

$$P(c|T) = \beta^q (1-\beta)^r \prod_{(u,v)\in E_T} P_c(u,v)$$

and then the conditional P given the diffusion graph :

$$P(c|G) = \sum_{T\in\mathcal{T}_c(G)} P(c|G)P(T|G) \propto \sum_{T\in\mathcal{T}_c(G)} \prod_{(u,v)\in E_T} P_c(u,v)$$

# NetInf Algorithm

Finally, we compute the likelihood for an entire set of contagion $C = \{c_1, c_2, \ldots, c_n\}$ :

$$P(C|G) = \prod_{c \in C} P(c|G)$$

The sought graph is:

$$\hat{G} = \arg\max_{|G| \leq k} P(C|G)$$

# NetInf Algorithm

- We introduce $\varepsilon$-edges as a low-likelihood "omnipresent" influence

$$P_c^{'}(u,v) = \begin{cases} \beta P_c(u,v), & \text{if } t_u < t_v \text{ and } (u,v) \in E_T \cap E \\ \epsilon P_c(u,v), & \text{if } t_u < t_v \text{ and } (u,v) \in E_T \cap E_\epsilon \\ 1-\beta, & \text{if } t_v = \infty \text{ and } (u,v) \in E \setminus E_T \\ 1-\epsilon, & \text{if } t_v = \infty \text{ and } (u,v) \in E_\epsilon \setminus E_T \\ 0, & \text{otherwise (i.e. if } t_u \geq t_v) \end{cases}$$

# NetInf Algorithm

- We make an approximation and only consider the maximum-likelihood diffusion tree (max-spanning tree)

- NetInf is a greedy algorithm that finds near-optimal solution in polynomial time

$$\begin{aligned}\hat{G} &= \arg\min_{G} F_C(G) \\ &= \sum_{c \in C} \max \sum_{(i,j) \in E_T} log(P_c^{'}(i,j) - log(\epsilon P_c(i,j))\end{aligned}$$

# Influence Estimation

Definition: Given a set of initially infected nodes, how many subsequent follow-ups occur in a specific time window.

Applications: Viral marketing, Spread of news & ideas etc.

Algorithm Elements ([1]):

- Continuous-time Independent Cascade Model ([2])
- Heterogeneous Transmission Functions
- Cohen’s Neighborhood Size Estimation ([3])
- Weibull distribution

# Independent Cascade Model

- Associates each edge in the network with a transmission density function $f_{ji}(\tau_{ji})$.
- Does not require a fixed infection probability for each edge, as time is modeled through a probability density.
- Assumes densities do be independent and differently distributed across edges (heterogeneous).
- Assumes only the neighbor that first infects a node to be the true parent.
- Each cascade induces is a Directed Acyclic Graph (DAG) irrespective of cycles in the network.

# Cohen’s Randomized Algorithm

Used for neighborhood size estimation for single source.
Basically, a modified Djikstra to construct least label list.

- Algorithm: Initialize each node with random label.
  Add node $i$ with smallest label $r_i$ to list.
  Add next node $i'$ if $d_i' < d_i$.
  Generate pairwise ordered list.
  Compute $r_*$ using binary search on list.
- Estimation:

$$|N(s,T)| \approx \frac{m-1}{\sum_{u=1}^{m} r_*^u}$$

# Weibull Distribution

Arguments have been made about exponential and power-law densities for modeling transmission times ([4], [5]).

- The distribution:

$$f(t;\alpha,\beta) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} e^{\left(-\frac{t}{\alpha}\right)^{\beta}} \quad s.t. \, \alpha,\beta > 0$$

- Captures the essence of Rayleigh, power-law and exponential.

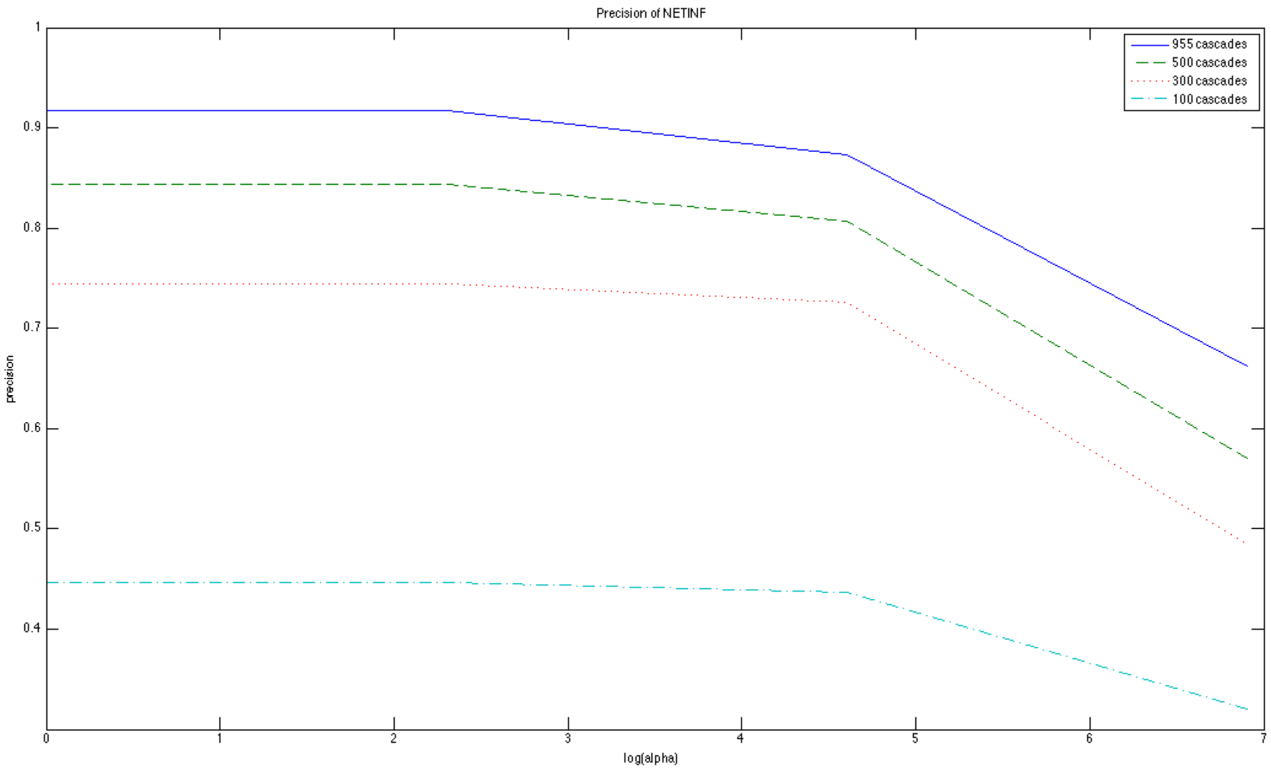
- Is much more flexible than any.

# Influence Estimation When Graph Structure is Unknown

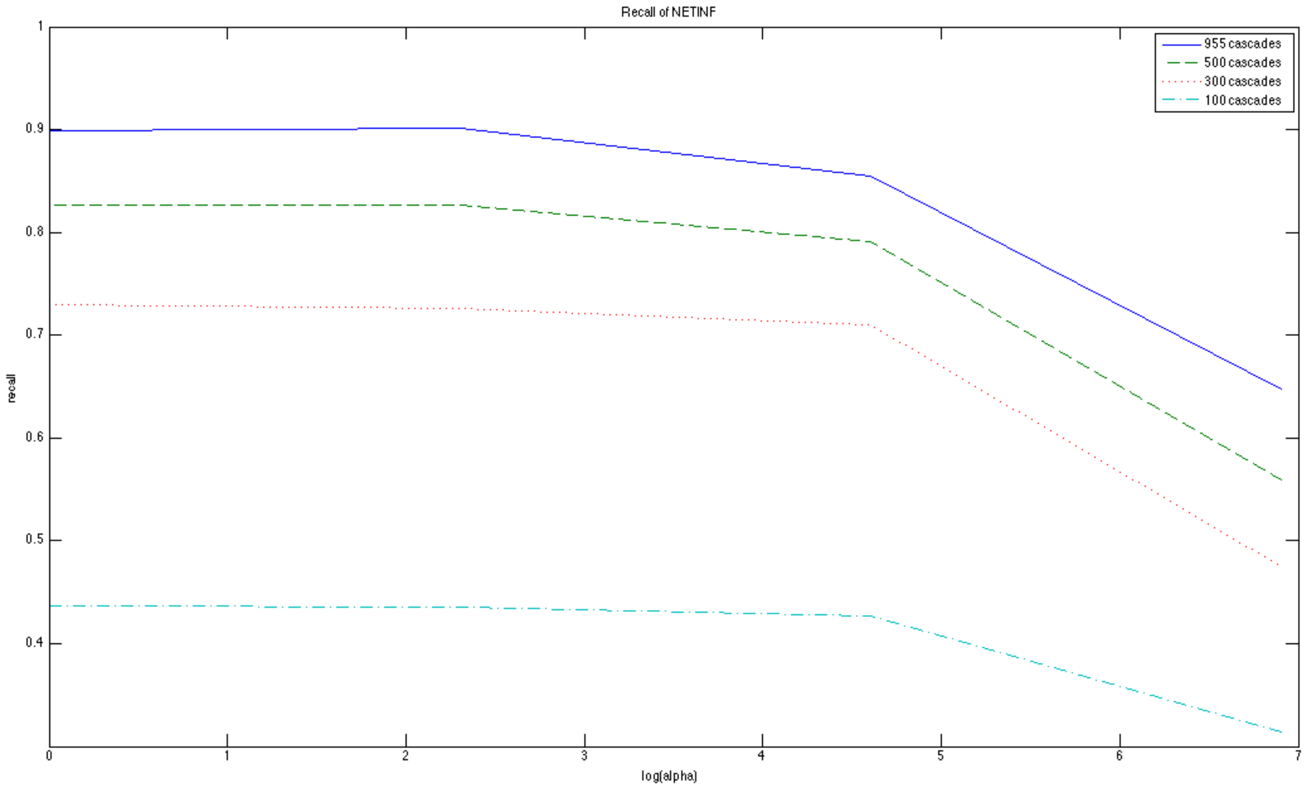
# When Graph Structure is Unknown

- Same assumption about transmission distribution (exponential, weibull)
- Based on contagions, learn graph structure using NETINF
- Learn influence using ConTinEst on estimated graph
- Challenge: How to set the #Edges in NETINF

# Experiments - NetInf

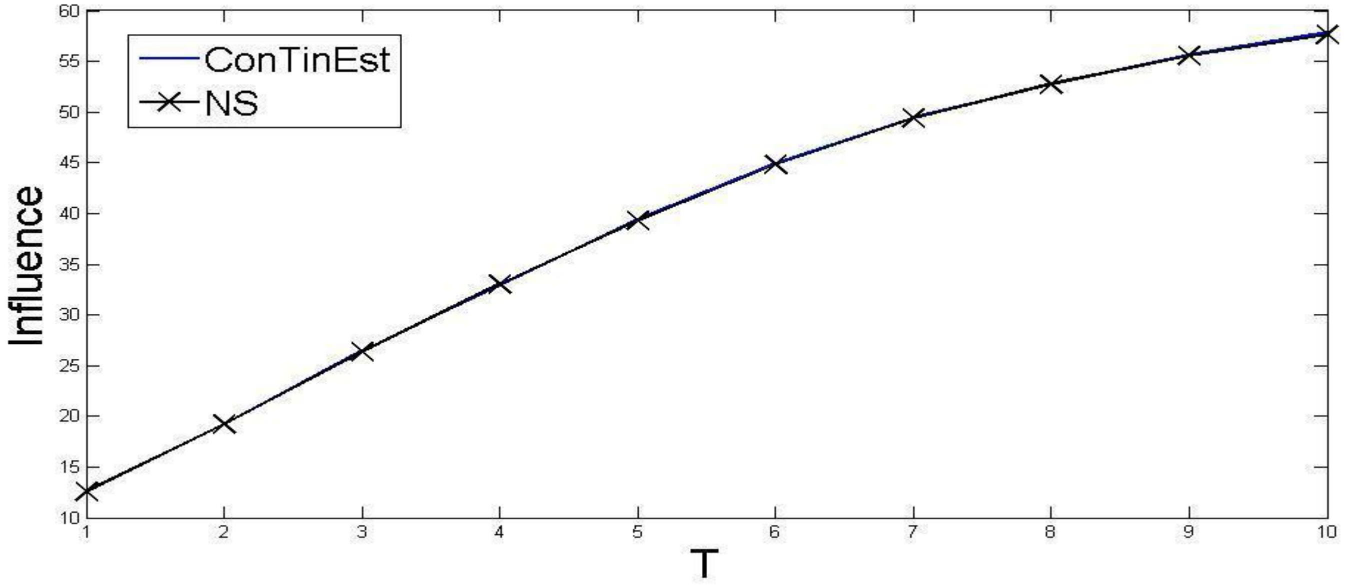


# Experiments - NetInf



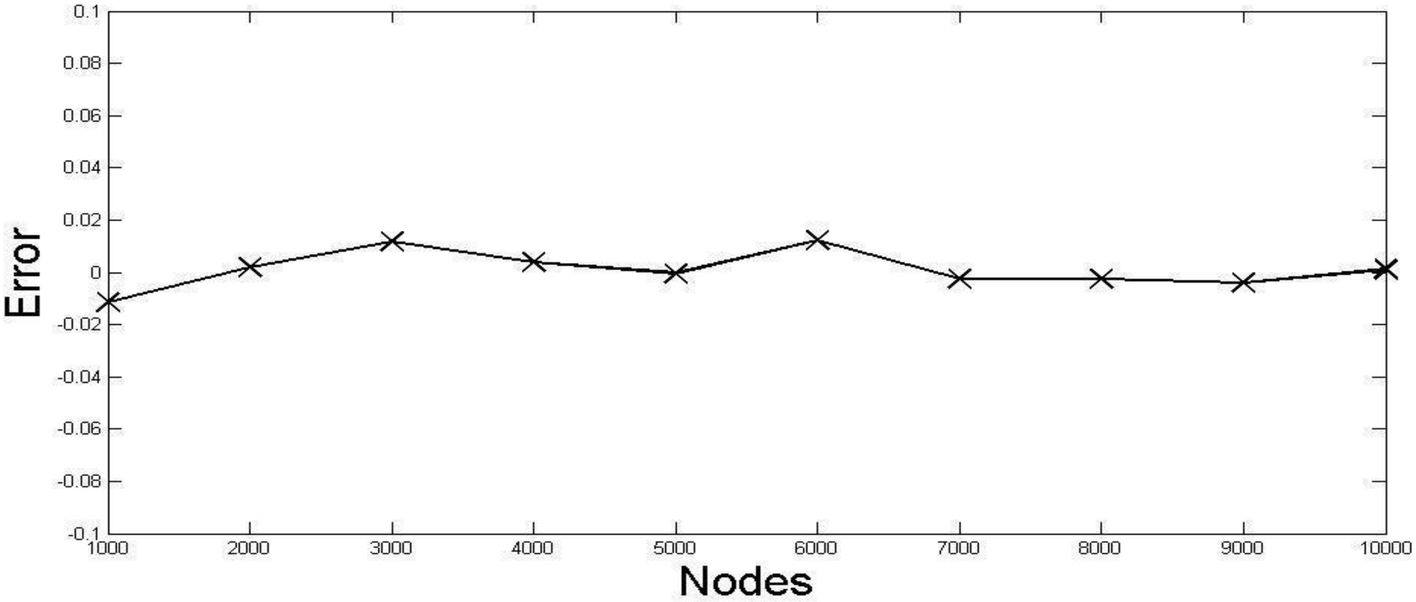
# Experiments – Influence Estimation

- Parameters: Nodes = 4000, N = 10000, M = 5,
- Dataset: MemeTracker Ground Truth



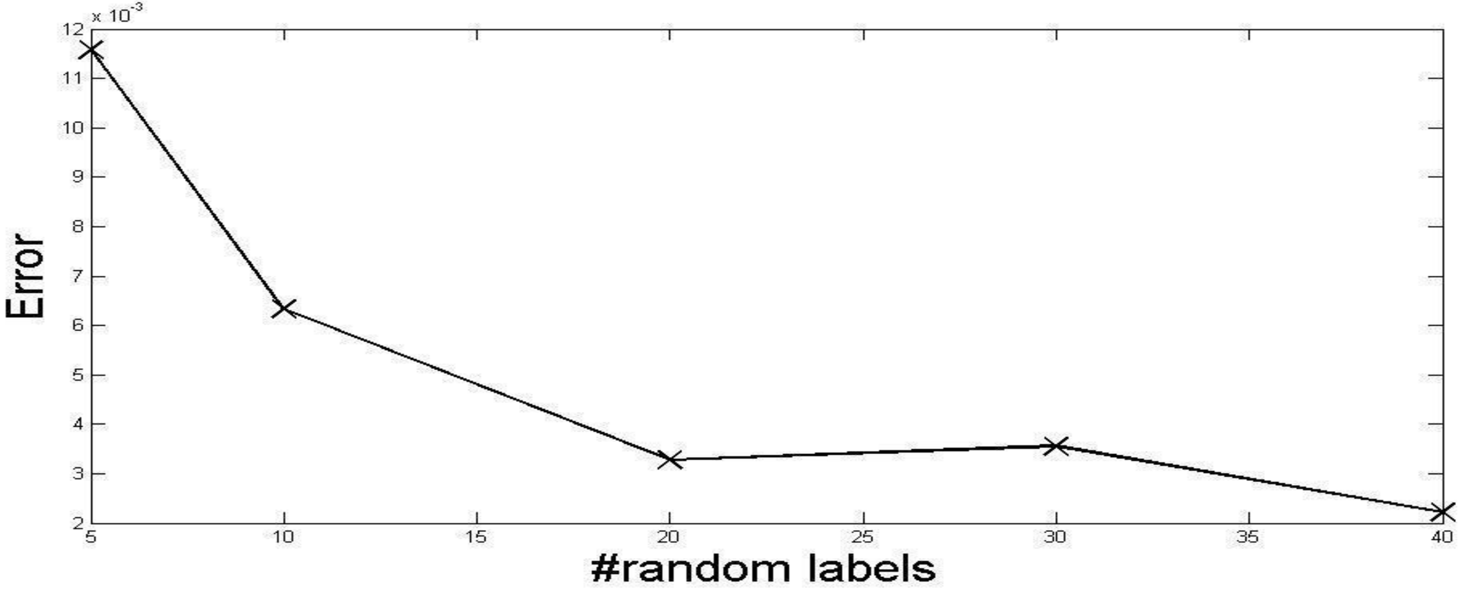
# Experiments – Influence Estimation

- Parameters: T = 10, N = 10000, M = 5
- Dataset: MemeTracker Ground Truth



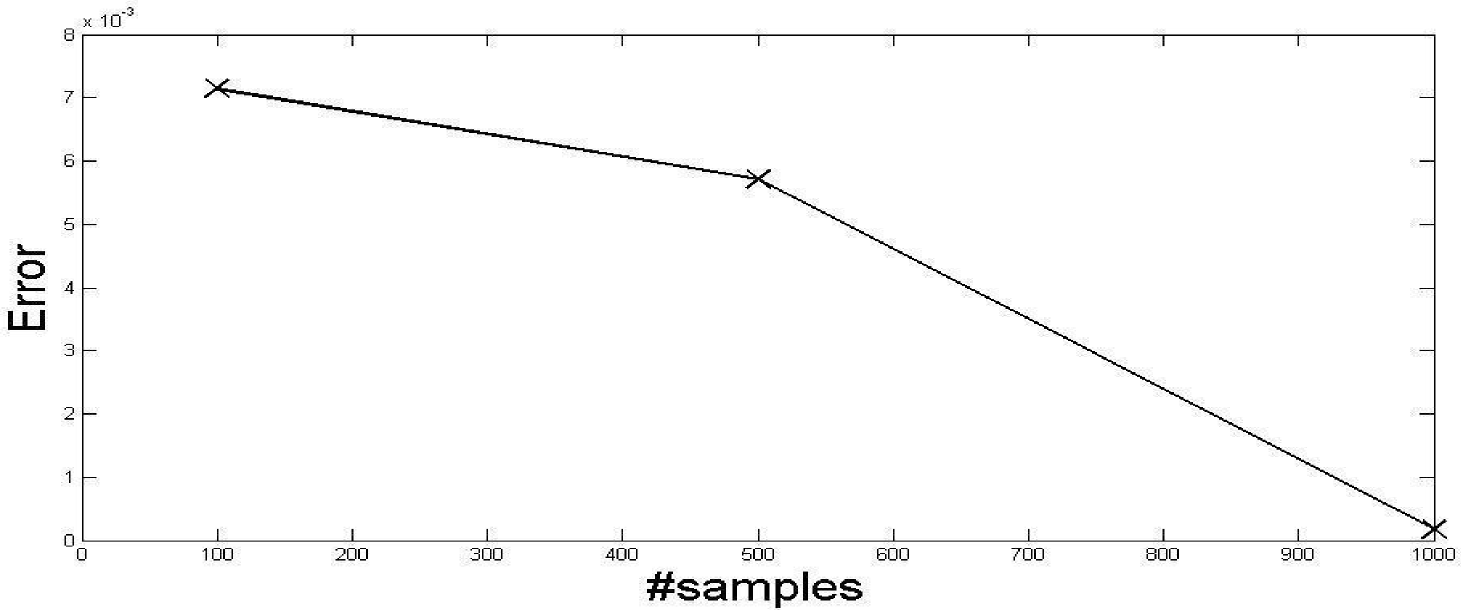
# Experiments – Influence Estimation

- Parameters: Nodes = 1000, T = 10, N = 10000
- Dataset: MemeTracker Ground Truth



# Experiments – Influence Estimation

- Parameters: Nodes = 4000, T = 10, M = 5
- Dataset: MemeTracker Ground Truth

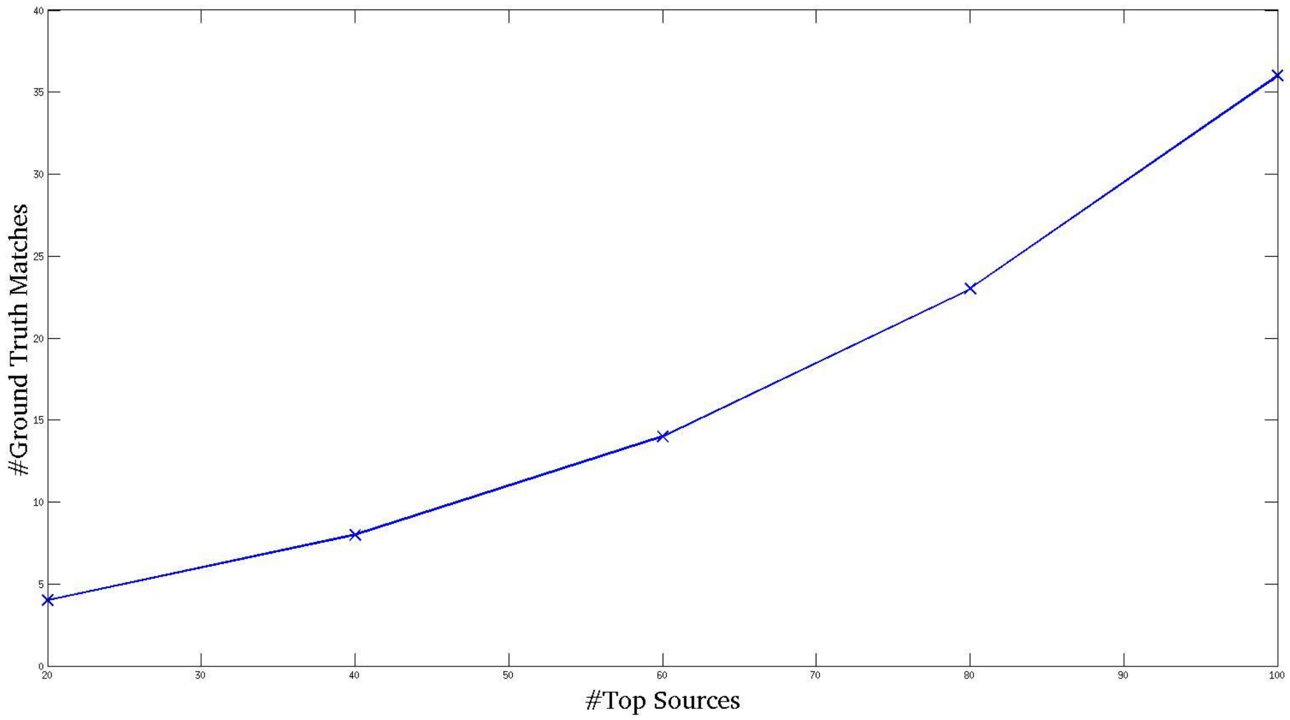


# Experiments – Influence Maximization

- Parameters: Nodes = 1000, N = 10000, M = 5
- Dataset: MemeTracker Ground Truth
- Top 10 sources:
  - http://totallyfuzzy.blogspot.com
  - http://thinkinganimationbook.blogspot.com
  - http://themusicchamber.blogspot.com
  - http://galadarling.com
  - http://drudge.com
  - http://socialitelife.celebuzz.com
  - http://wwwwakeupamericans-spree.blogspot.com
  - http://pr-inside.com
  - http://lockergnome.com
  - http://mashable.com

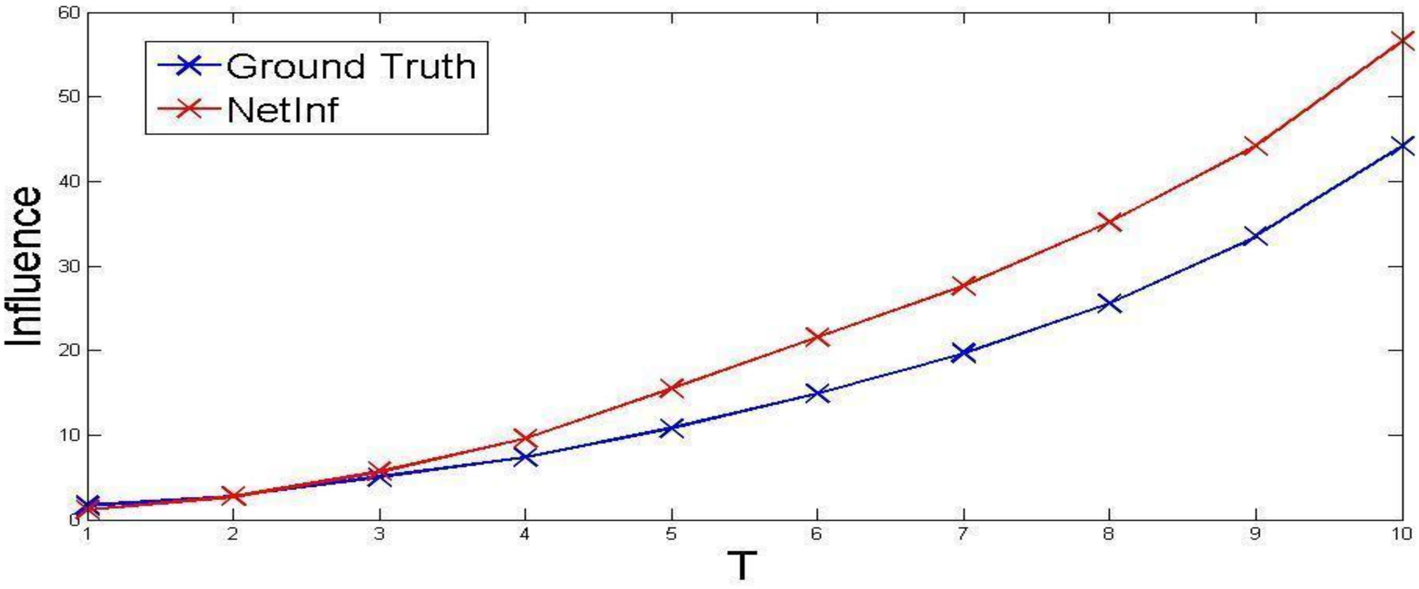
# Experiments – Inference on Estimated Graph

#nodes = 1000    #edges = 2000



# Experiments – Graph Learning/Influence Estimation Integration

- Parameters: Nodes = 1000, N = 10000, M = 5
- Dataset: Kronecker Graphs Ground Truth and Estimated



# Conclusion

- Learning network graph structure and estimating node influence is important in applications such as viral marketing, spread of news, …
- Our experiments show that increasing the <u>number of cascades has a great impact on precision and recall</u> of the graph learned, w.r.t. ground truth
- Our experiments also show that the <u>estimated influence is very close to the ground truth</u> and the relative error decreases on increasing the number of samples and random labels
- Our approach allows us to <u>integrate the graph learning problem with the influence estimation</u> and thereby **eliminate the need to know the ground truth graph, which is the case in most real world application.**

Q A

# References


[1] Du, Nan, et al. "Scalable Influence Estimation in Continuous-Time Diffusion Networks." *Advances in Neural Information Processing Systems*. 2013.

[2] Rodriguez, Manuel Gomez, David Balduzzi, and Bernhard Schölkopf. "Uncovering the temporal dynamics of diffusion networks." *arXiv preprint arXiv:1105.0697* (2011).

[3] Cohen, Edith. "Size-estimation framework with applications to transitive closure and reachability." *Journal of Computer and System Sciences* 55.3 (1997): 441-453.

[4] Barabasi, Albert-Laszlo. "The origin of bursts and heavy tails in human dynamics." *Nature* 435.7039 (2005): 207-211.

[5] Leskovec, Jure, et al. "Patterns of Cascading Behavior in Large Blog Graphs." *SDM*. Vol. 7. 2007.