Scalable Influence Estimation in Continuous-Time Diffusion Networks

Nan Du [1]

Joint work with Le Song [1], Manuel Gomez Rodriguez[2] and Hongyuan Zha[1]

[1]Georgia Institute of Technology

[2]Max Planck Institute for Intelligent Systems

Motivation

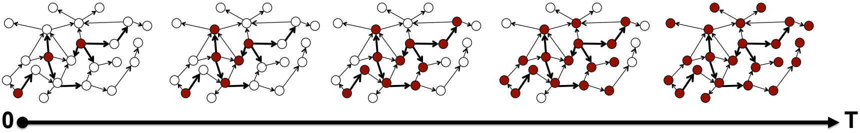
- Networks abstract interactions among entities (media sites, people, organizations, etc.)
- Propagation of news, product reviews, virus, etc. takes place over
  - Information networks
  - Social networks
  - Traffic networks
  - Communication networks

Propagation traces can be extracted from various data sources.

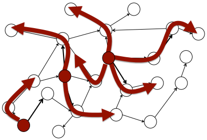
Motivation

- Question : If a new piece of information is released in a few nodes, can it spread, in 1 month, to a million nodes ?
- Question : How can we optimize the selection of the earlier nodes to trigger, *within a time window T*, the largest expected number of follow-ups ?

time-sensitive viral marketing

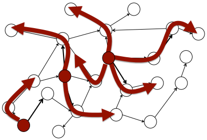


1 Continuous-time diffusion process.

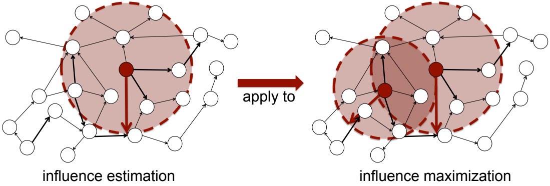


Outline

1 Continuous-time diffusion process.

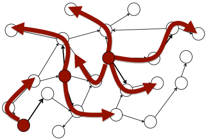


2 Efficient influence estimation and maximization.

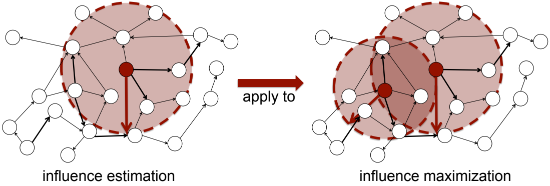


Outline

1 Continuous-time diffusion process.



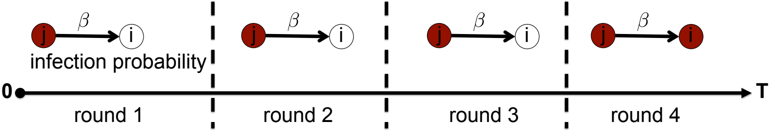
2 Efficient influence estimation and maximization.



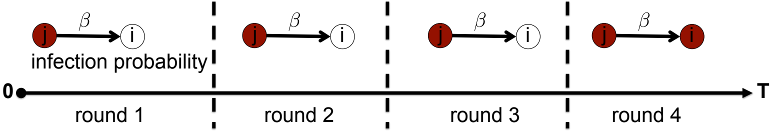
3 Experimental evaluation with synthetic and true data.

Continuous vs. Discrete Time Diffusion Model

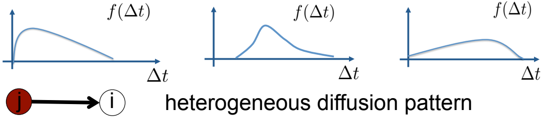
- Traditionally, diffusion has been modeled as discrete steps (or rounds).



- Traditionally, diffusion has been modeled as discrete steps (or rounds).



- However, real time is continuous.

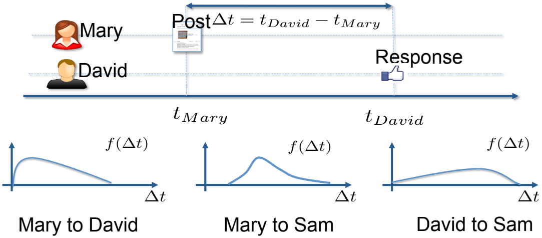


  - how long is each round ?
  - how to aggregate events within one round ?

Continuous-time Independent Cascade Model

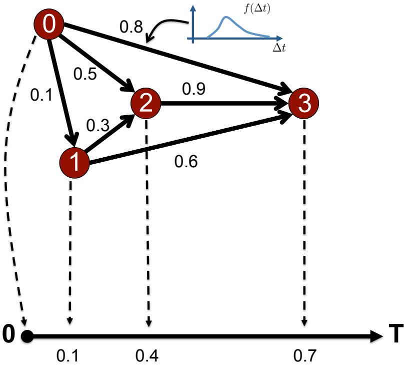
- Node set $\mathcal{V}$: people, media-sites, organizations, etc.
- Edge set $\mathcal{E}$: relations, channels, etc.
- Infection : an event occurs to a node, e.g., adopting a product.
- Pairwise conditional density over time

$$f_{ji}(t_j|t_i) = f_{ji}(t_i - t_j)$$



Continuous-time Independent Cascade Model

- node 0 is the source $\mathcal{A} = \{0\}$.
- node 0 influences out-going neighbors with $f(\Delta t)$.
- node 1 is infected at $t_1 = 0.1$.
- both node 0 and 1 influence node 2.
- node 1 first infects node 2 since $0.4 < 0.5$.
- node 3 is infected at 0.7 by node 1.



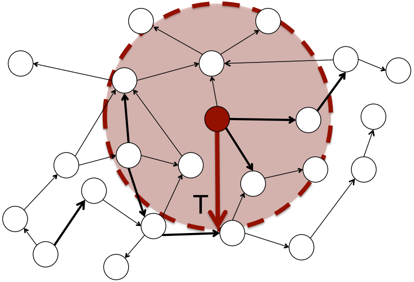
Node's Perspective

- For each node i, let random variable t_i represent the infection time.

Node's Perspective

- For each node i, let random variable t_i represent the infection time.
- The influence value of sources $\mathcal{A}$ by time T is

$$\sigma(\mathcal{A}, T) = \mathbb{E}\left[\sum_{i\in\mathcal{V}} \mathbb{I}\left\{t_i \leq T\right\}\right] = \sum_{i\in\mathcal{V}} \Pr\left\{t_i \leq T\right\}$$



influence estimation

Node's Perspective

- Directed graphical model representation

$$p\left(\{t_i\}_{i\in\mathcal{V}}\right) = \prod\nolimits_{i\in\mathcal{V}} p\left(t_i | \{t_j\}_{j\in\pi_i}\right), \;\; \pi_i \text{ is the set of parents}$$

NODE'S PERSPECTIVE

- Directed graphical model representation

$$p\left(\{t_i\}_{i\in\mathcal{V}}\right) = \prod\nolimits_{i\in\mathcal{V}} p\left(t_i | \{t_j\}_{j\in\pi_i}\right), \ \pi_i \text{ is the set of parents}$$

- Marginalization

$$\Pr\{t_i \leq T\} = \int_0^\infty \cdots \int_{t_i=0}^{T} \cdots \int_0^\infty \left(\prod\nolimits_{j\in\mathcal{V}} p\left(t_j | \{t_l\}_{l\in\pi_j}\right)\right)\left(\prod\nolimits_{j\in\mathcal{V}} dt_j\right)$$

Node's Perspective

- Directed graphical model representation

$$p\left(\{t_i\}_{i\in\mathcal{V}}\right) = \prod\nolimits_{i\in\mathcal{V}} p\left(t_i|\{t_j\}_{j\in\pi_i}\right), \;\; \pi_i \text{ is the set of parents}$$

- Marginalization

$$\Pr\{t_i \leq T\} = \int_0^{\infty}\cdots\int_{t_i=0}^{T}\cdots\int_0^{\infty}\left(\prod\nolimits_{j\in\mathcal{V}} p\left(t_j|\{t_l\}_{l\in\pi_j}\right)\right)\left(\prod\nolimits_{j\in\mathcal{V}} dt_j\right)$$

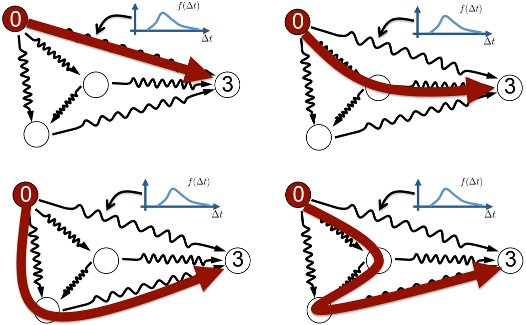
  - Need to integrate all possible configurations of cascades where $t_i < T$.
  - No closed form solution for general heterogeneous transmission function.
  - Hard to approximate.

Edge's Perspective

- Mutually independent transmission times $\tau_{ji} = t_i - t_j$

Edge's Perspective

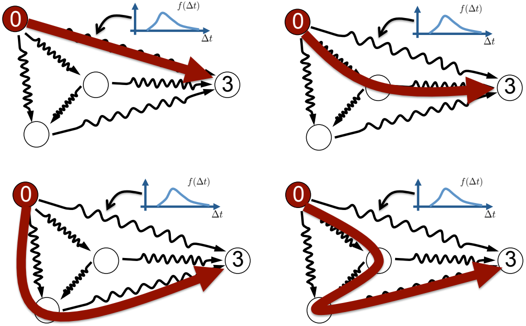
- Mutually independent transmission times $\tau_{ji} = t_i - t_j$
- A network with stochastic edge weights (inter-infection time)



shortest path property

Edge's Perspective

- Mutually independent transmission times $\tau_{ji} = t_i - t_j$
- A network with stochastic edge weights (inter-infection time)



shortest path property

- t_3 equals to the length of the shortest path from t_0.

Node's vs. Edge's Perspective

- Node's perspective

$\sigma(\mathcal{A}, T) = \sum_{i \in \mathcal{V}} \Pr\{t_i \leq T\}$

$\Pr\{t_i \leq T\} = \int_0^\infty \cdots \int_{t_i=0}^{T} \cdots \int_0^\infty \left(\prod_{j \in \mathcal{V}} p\left(t_j | \{t_l\}_{l \in \pi_j}\right) \right) \left(\prod_{j \in \mathcal{V}} dt_j \right)$

- Edge's perspective

$\sigma(\mathcal{A}, T) = \mathbb{E}_G \left[\sum_{i \in \mathcal{V}} \mathbb{I}\{t_i \leq T\} \right]$

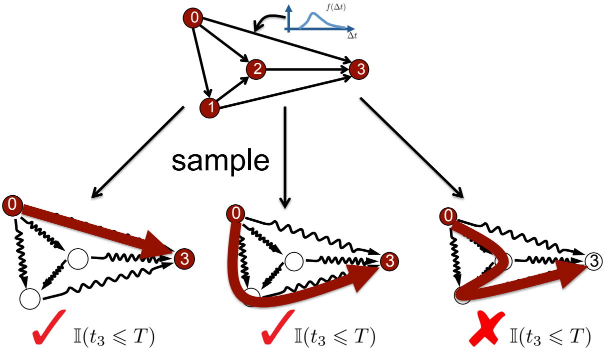
$G \backsim \prod_{(j,i) \in \mathcal{E}} f_{ji}(\tau_{ji})$

Naive Simulation

- Given G, t_i is the length of the shortest path.
- Check whether $t_i \leqslant T$ on many samples.

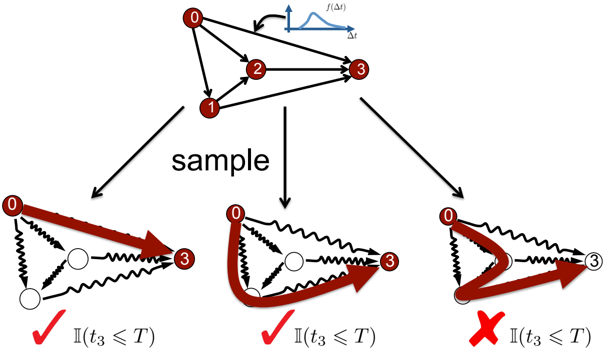
Naive Simulation

- Given G, t_i is the length of the shortest path.
- Check whether $t_i \leqslant T$ on many samples.



Naive Simulation

- Given G, t_i is the length of the shortest path.
- Check whether $t_i \leqslant T$ on many samples.



- Average the counts across n samples.

$$\sigma(\mathcal{A}, T) \approx \frac{1}{n}\left(\sum_{i \in \mathcal{V}} \mathbb{I}\left\{t_i \leq T | G_1\right\} +, \ldots, + \sum_{i \in \mathcal{V}} \mathbb{I}\left\{t_i \leq T | G_n\right\}\right)$$

Naive Simulation

- Using shortest path is not scalable.

Naive Simulation

- Using shortest path is not scalable.
- Influence Estimation of a single source j
  - $\sigma(\{j\}, T)$
  - Compute all shortest paths from j to the other nodes.

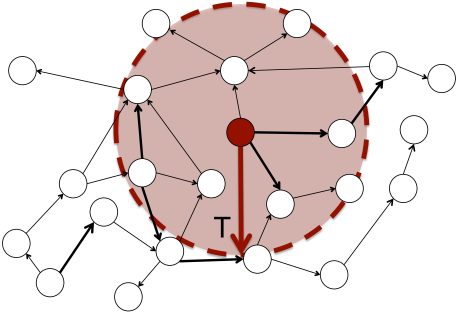
Naive Simulation

- Using shortest path is not scalable.
- Influence Estimation of a single source j
  - $\sigma(\{j\}, T)$
  - Compute all shortest paths from j to the other nodes.
- Which source is the best ?
  - Chose j with the largest $\sigma(\{j\}, T)$
  - Try source $j = 0, \ldots, |\mathcal{V}| - 1$, $O(|\mathcal{V}|^2)$
- Quadratic in network size
  Can not deal with large networks !

Neighborhood Size Estimation

- Given a sampled network G and source node j, estimate

  $|\mathcal{N}(\{j\}, T)| = |\{i : t_i \leqslant T\}|$ the size of neighborhood within distance T.



influence estimation

Neighborhood Size Estimation

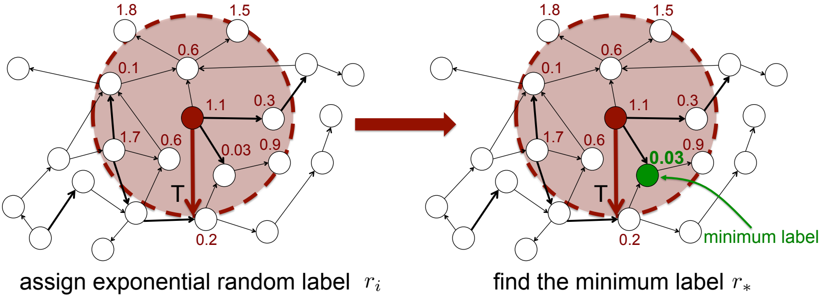
Key Fact

Given a set of n i.i.d random variable $X^i \backsim e^{-x}$*, the minimum* $X_* \backsim ne^{-nx}$.

Neighborhood Size Estimation

Key Fact

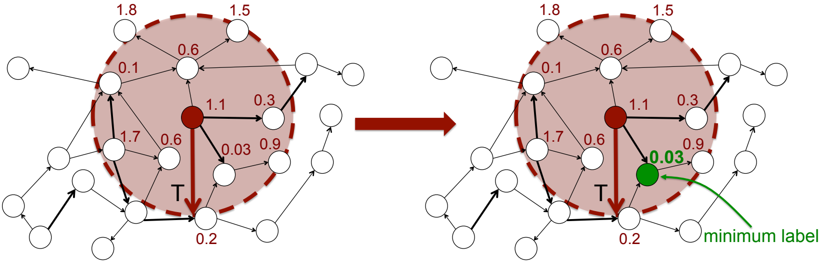
Given a set of n i.i.d random variable $X^i \backsim e^{-x}$, *the minimum* $X_* \backsim ne^{-nx}$.



Neighborhood Size Estimation

Key Fact

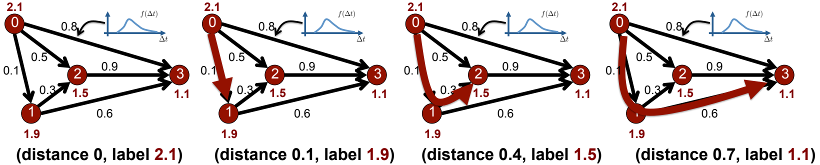
Given a set of n i.i.d random variable $X^i \backsim e^{-x}$, *the minimum* $X_* \backsim ne^{-nx}$.



assign exponential random label r_i find the minimum label r_*

- Find m such least labels, $\{r_*^u\}_{u=1}^m$ to estimate $|\mathcal{N}(\{j\}, T)| \approx \frac{m-1}{\sum_{u=1}^m r_*^u}$, convert counting problem to estimation problem !

- Each node holds a least-label list



Node 0's least-label list

- (Cohen 97) gives the smart algorithm to calculate the least-label list for each node in $\tilde{O}(|\mathcal{E}|)$.
- Estimator $|\mathcal{N}(\{j\}, T)| \approx \frac{m-1}{\sum_{u=1}^{m} r_*^u}$ is unbiased with variance $O(\frac{1}{m-2})$.
- Sample $m \ll \min(|\mathcal{V}|, |\mathcal{E}|)$ to select the single best source linearly in network size !

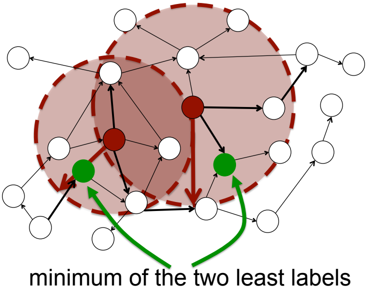
Multiple Sources

- Multiple sources $\mathcal{A}$

$$\mathcal{N}(\mathcal{A}, T) = \bigcup_{s \in \mathcal{A}} \mathcal{N}(s, T).$$

- Reuse least-label list for each single source $s \in \mathcal{A}$

$$r_* = \min_{i \in \mathcal{A}} \ \min_{j \in \mathcal{N}(i,T)} r_j$$



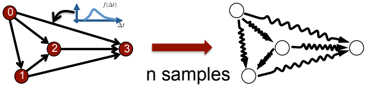
Overall Algorithm ConTinEst

1. Sample n sets of random transmission times

$$\{\tau_{ij}^{l}\}_{(j,i)\in\mathcal{E}} \sim \prod_{(j,i)\in\mathcal{E}} f_{ji}(\tau_{ji})$$

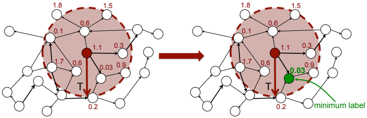
Overall Algorithm ConTinEst

1. Sample n sets of random transmission times



$$\{\tau_{ij}^{l}\}_{(j,i)\in\mathcal{E}} \sim \prod_{(j,i)\in\mathcal{E}} f_{ji}(\tau_{ji})$$

2. Given a set of $\{\tau_{ij}^{l}\}_{(j,i)\in\mathcal{E}}$, sample m sets of random labels

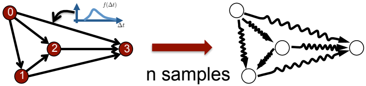


assign exponential random label r_i find the minimum label r_*

$$\{r_i^{u}\}_{i\in\mathcal{V}} \sim \prod_{i\in\mathcal{V}} \exp(-r_i)$$

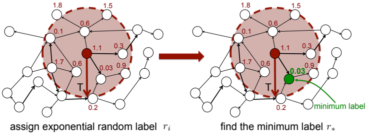
Overall Algorithm ConTinEst

1. Sample n sets of random transmission times



$$\{\tau_{ij}^{l}\}_{(j,i)\in\mathcal{E}} \sim \prod_{(j,i)\in\mathcal{E}} f_{ji}(\tau_{ji})$$

2. Given a set of $\{\tau_{ij}^{l}\}_{(j,i)\in\mathcal{E}}$, sample m sets of random labels



$$\{r_i^{u}\}_{i\in\mathcal{V}} \sim \prod_{i\in\mathcal{V}} \exp(-r_i)$$

3. Estimate $\sigma(\mathcal{A}, T)$ by sample averages

$$\sigma(\mathcal{A}, T) \approx \frac{1}{n}\sum_{l=1}^{n}\left((m-1)/\sum_{u_l=1}^{m} r_*^{u_l}\right)$$

Overall Algorithm ConTinEst

Theorem

Draw the following number of samples for the set of random transmission times

$$n \geqslant \frac{C\Lambda}{\epsilon^2} \log\left(\frac{2|\mathcal{V}|}{\delta}\right),$$

where Λ depends on $\mathcal{A}$ and T, and for each set of random transmission times, draw m set of random labels. Then $|\widehat{\sigma}(\mathcal{A}, T) - \sigma(\mathcal{A}, T)| \leqslant \epsilon$ uniformly for all $\mathcal{A}$ with $|\mathcal{A}| \leqslant C$, with probability at least $1 - \delta$.

Overall Algorithm ConTinEst

Theorem

Draw the following number of samples for the set of random transmission times

$$n \geqslant \frac{C\Lambda}{\epsilon^2} \log\left(\frac{2|\mathcal{V}|}{\delta}\right),$$

where Λ depends on $\mathcal{A}$ and T, and for each set of random transmission times, draw m set of random labels. Then $|\widehat{\sigma}(\mathcal{A}, T) - \sigma(\mathcal{A}, T)| \leqslant \epsilon$ uniformly for all $\mathcal{A}$ with $|\mathcal{A}| \leqslant C$, with probability at least $1 - \delta$.

- Implications : influence of larger source set $\mathcal{A}$ at the longer time window T requires more samples in the worst case.

Overall Algorithm ConTinEst

Theorem

Draw the following number of samples for the set of random transmission times

$$n \geqslant \frac{C\Lambda}{\epsilon^2} \log\left(\frac{2|\mathcal{V}|}{\delta}\right),$$

where Λ depends on $\mathcal{A}$ and T, and for each set of random transmission times, draw m set of random labels. Then $|\widehat{\sigma}(\mathcal{A}, T) - \sigma(\mathcal{A}, T)| \leqslant \epsilon$ uniformly for all $\mathcal{A}$ with $|\mathcal{A}| \leqslant C$, with probability at least $1 - \delta$.

- Implications : influence of larger source set $\mathcal{A}$ at the longer time window T requires more samples in the worst case.
- In practice : small $m = 5$ achieves good performance. Inaccuracy is canceled out due to large outer-loop n samples.

Influence Maximization

- We seek to solve

$$\mathcal{A}^* = \operatorname{argmax}_{|\mathcal{A}| \leqslant C} \ \sigma(\mathcal{A}, T)$$

which is NP-hard in general.

Influence Maximization

- We seek to solve

$$\mathcal{A}^* = \operatorname{argmax}_{|\mathcal{A}| \leqslant C} \ \sigma(\mathcal{A}, T)$$

  which is NP-hard in general.
- $\sigma(\mathcal{A}, T)$ is a non-negative, monotonic, submodular function.

Influence Maximization

- We seek to solve

$$\mathcal{A}^* = \operatorname{argmax}_{|\mathcal{A}| \leqslant C}\ \sigma(\mathcal{A}, T)$$

  which is NP-hard in general.
- $\sigma(\mathcal{A}, T)$ is a non-negative, monotonic, submodular function.
- Greedy algorithm achieves at least a fraction $(1 - 1/e)$ of the optimal value (OPT)

Influence Maximization

- We seek to solve

$$\mathcal{A}^* = \operatorname{argmax}_{|\mathcal{A}| \leqslant C} \ \sigma(\mathcal{A}, T)$$

  which is NP-hard in general.
- $\sigma(\mathcal{A}, T)$ is a non-negative, monotonic, submodular function.
- Greedy algorithm achieves at least a fraction $(1 - 1/e)$ of the optimal value (OPT)

Theorem

Suppose the influence $\sigma(\mathcal{A}, T)$ for all $\mathcal{A}$ with $|\mathcal{A}| \leq C$ are estimated uniformly with error ϵ and confidence $1 - \delta$, the greedy algorithm returns a set of sources $\widehat{\mathcal{A}}$ such that $\sigma(\widehat{\mathcal{A}}, T) \geq (1 - 1/e)OPT - 2C\epsilon$ with probability at least $1 - \delta$.

Experimental Evaluation

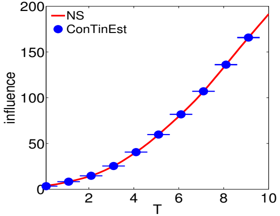
- Synthetic dataset
  - Generate network structure.
  - Weibull pairwise transmission function with randomly chosen parameters.
  - Accuracy of estimated influence (compared to simulation).
  - Quality of selected sources.
  - Scalability.

Experimental Evaluation

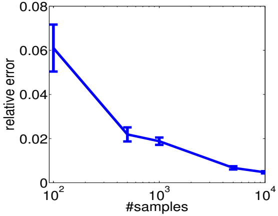
- Synthetic dataset
  - Generate network structure.
  - Weibull pairwise transmission function with randomly chosen parameters.
  - Accuracy of estimated influence (compared to simulation).
  - Quality of selected sources.
  - Scalability.
- Real dataset
  - MemeTracker data (172m news articles 08/2009-09/2009).
  - Infer network structures from hyperlink cascade data.
  - Accuracy of estimated influence (compared to real value).
  - Quality of selected sources on real data.

Synthetic Dataset

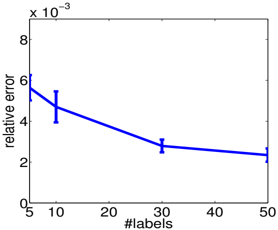
Accuracy of the estimated influence (highest out-degree node)



(a) Influence vs. time



(b) Error vs. #samples

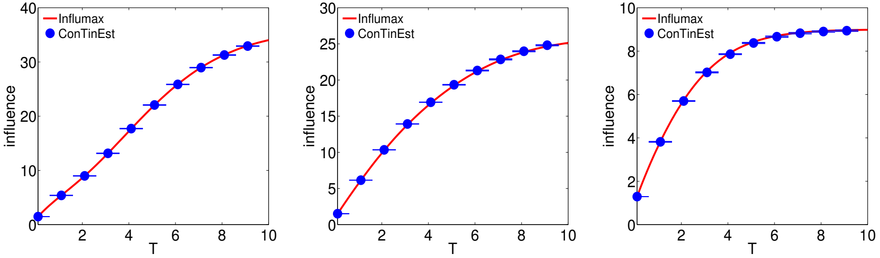


(c) Error vs. #labels

- 100,000 samples for naive simulation (NS).
- $m(\#lables) << n(\#samples)$ still achieves good accuracy.
- accuracy does not depend on network structure (1024 nodes, 2048 edges).

Synthetic Dataset

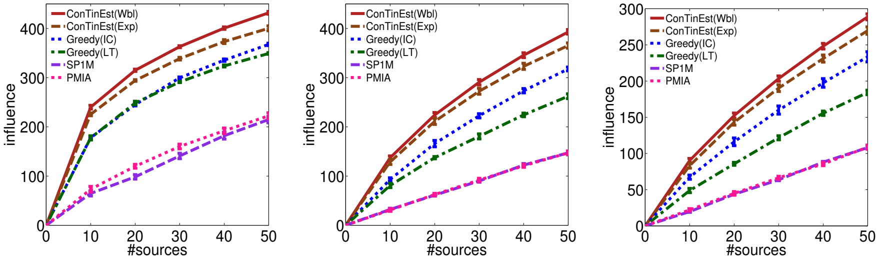
Accuracy of the estimated influence (highest out-degree node)



(a) Core-periphery (b) Random (c) Hierarchical

- ConTinEst is close to Influmax (sparse small networks, exponential transmission functions).
- accuracy does not depend on network structure (128 nodes, 141 edges).

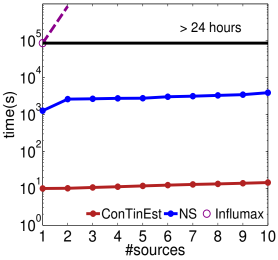
Quality of the selected nodes for influence maximization



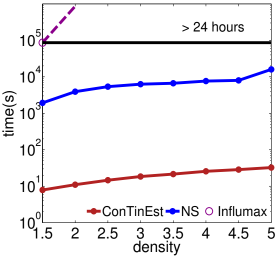
(a) Core-periphery (b) Random (c) Hierarchical

- CONTINEST typically outperforms competitive methods by 20%.
- Performance does not depend on network structure (1024 nodes, 2048 edges).

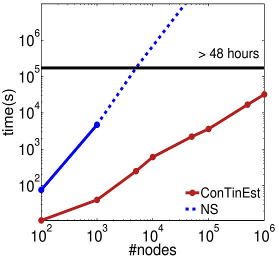
Scalability of influence maximization



(a) # sources
Small network



(b) network density
Small network



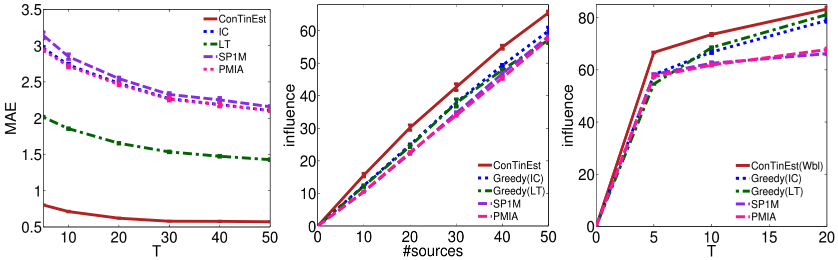
(c) network size
Up to one million nodes

- Small network : 128 nodes.
- Large network : up to 1 million nodes, with density 1.5.
- Our algorithm : sample 10K networks, 5 random labels.

Real Dataset

- 10,967 hyperlink cascades.
- Use 80% cascades for learning continuous-time diffusion model.
- Select sources based on the learnt model.
- Evaluate influence of the sources using 20% test cascades.
- Compared to discrete-time diffusion models and scalable heuristics.

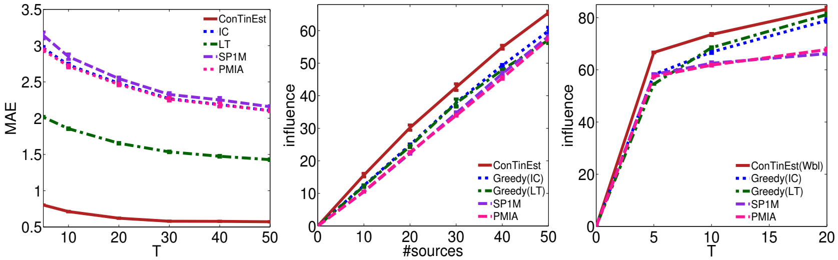
Real Dataset



(a) Estimation error (b) #sources (c) time

- ConTinEst achieves the lowest MAE error.

Real Dataset



(a) Estimation error (b) #sources (c) time

- ConTinEst achieves the lowest MAE error.
- ConTinEst produces the set of sources with the largest true influence within short time window.

Conclusion

- a novel view of the influence estimation problem in continuous-time diffusion networks.
  - very little assumptions about transmission functions.
  - only depends on temporal cascades induced by diffusion.

Conclusion

- a novel view of the influence estimation problem in continuous-time diffusion networks.
  - very little assumptions about transmission functions.
  - only depends on temporal cascades induced by diffusion.
- an efficient randomized algorithm improving :
  - the accuracy of the estimated influence (the lowest MAE in real data).
  - the quality of selected sources (the largest influence within short time period).
  - the scalability (scaling up to millions of nodes in practice).

Conclusion

- a novel view of the influence estimation problem in continuous-time diffusion networks.
  - very little assumptions about transmission functions.
  - only depends on temporal cascades induced by diffusion.
- an efficient randomized algorithm improving :
  - the accuracy of the estimated influence (the lowest MAE in real data).
  - the quality of selected sources (the largest influence within short time period).
  - the scalability (scaling up to millions of nodes in practice).
- natural follow-up : product / advertisement allocation with more realistic constraints.