



Extracting Social Events for Learning Better Information Diffusion Models

Shuyang Lin¹ Fengjiao Wang¹
Qingbo Hu¹ Philip S. Yu^{1,2}

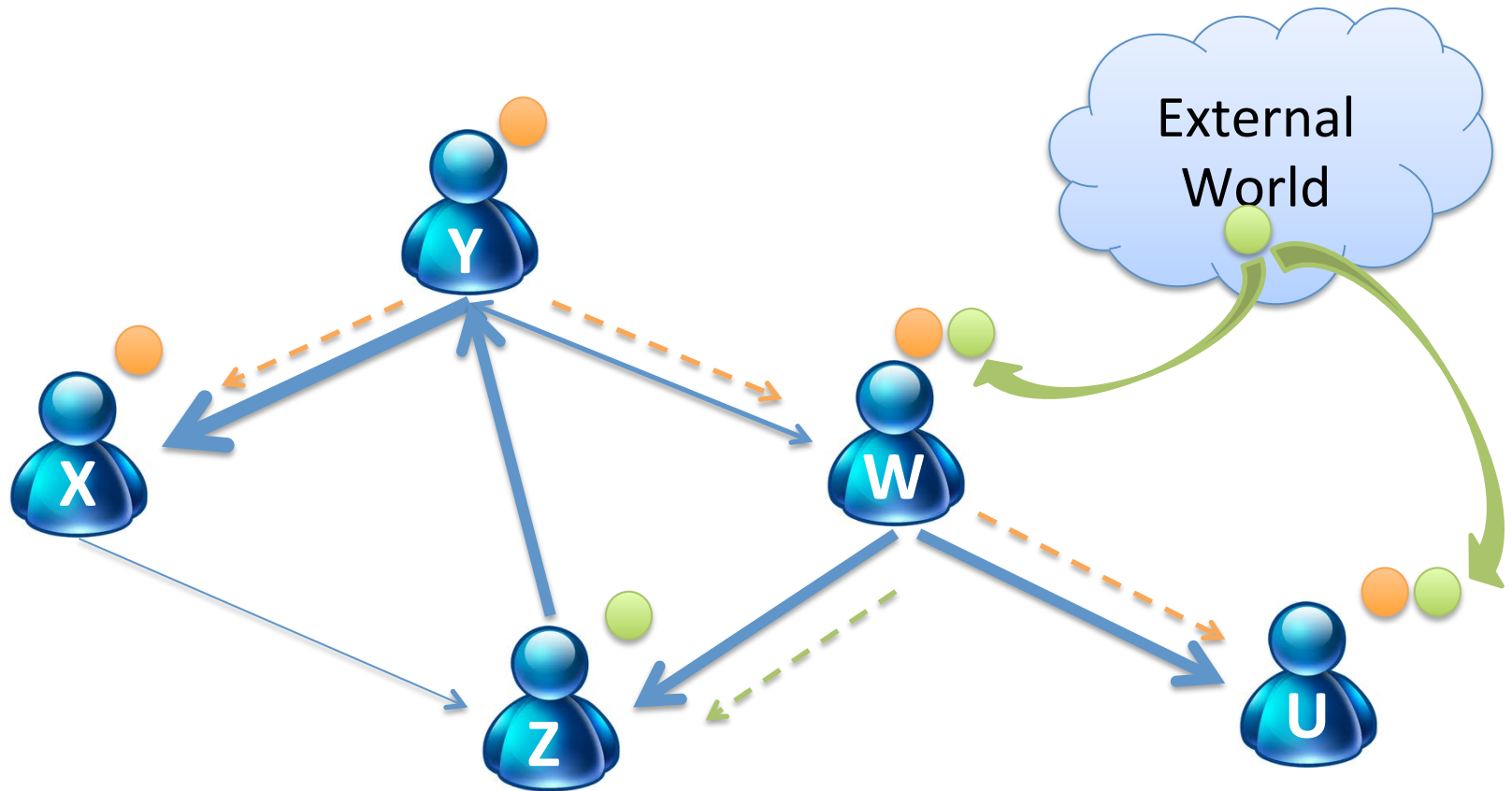
¹Department of Computer Science, University of Illinois at Chicago

²Computer Science Department, King Abdulaziz University

Highlights

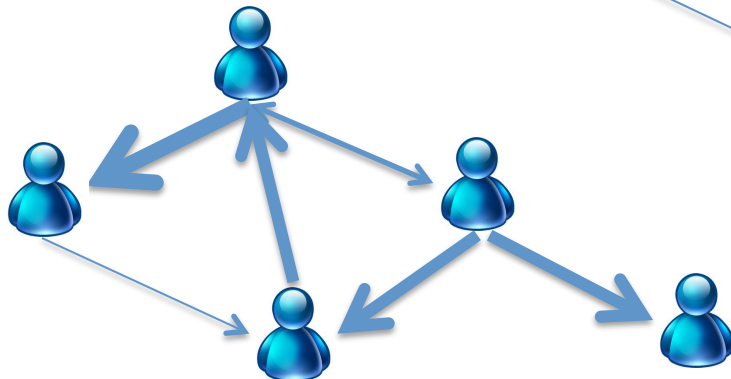
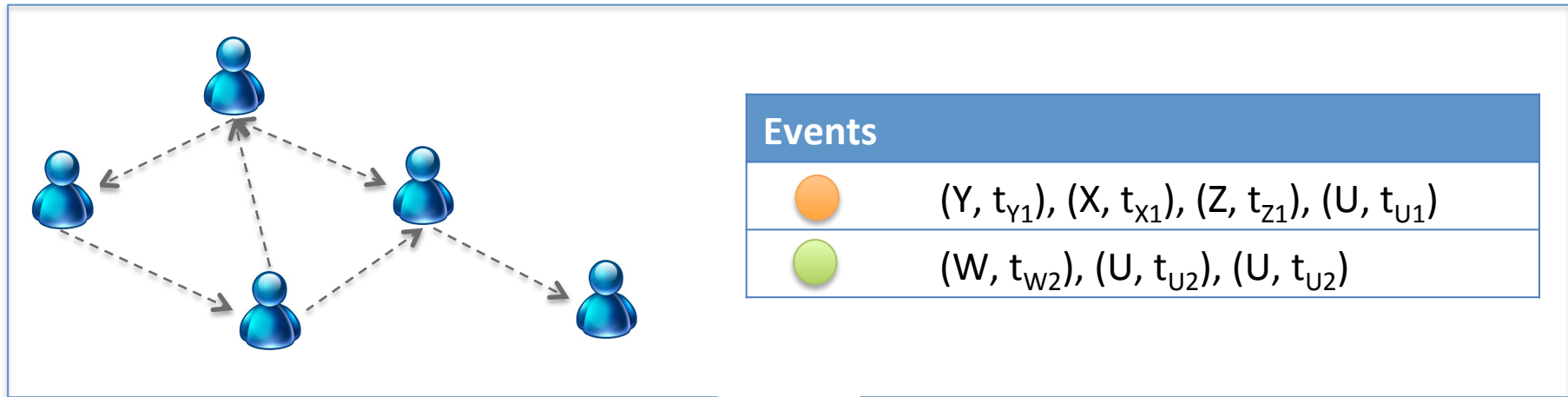
- Idea
 - Distinguish the two different sources of events
 - socially sourced and externally sourced
 - Learning information diffusion models from socially sourced events.
- Model
 - Use a mixture model framework to combine the social and the external influence.
 - An EM-based inference algorithm to overcome the problem of inference dependency

Information diffusion model and events



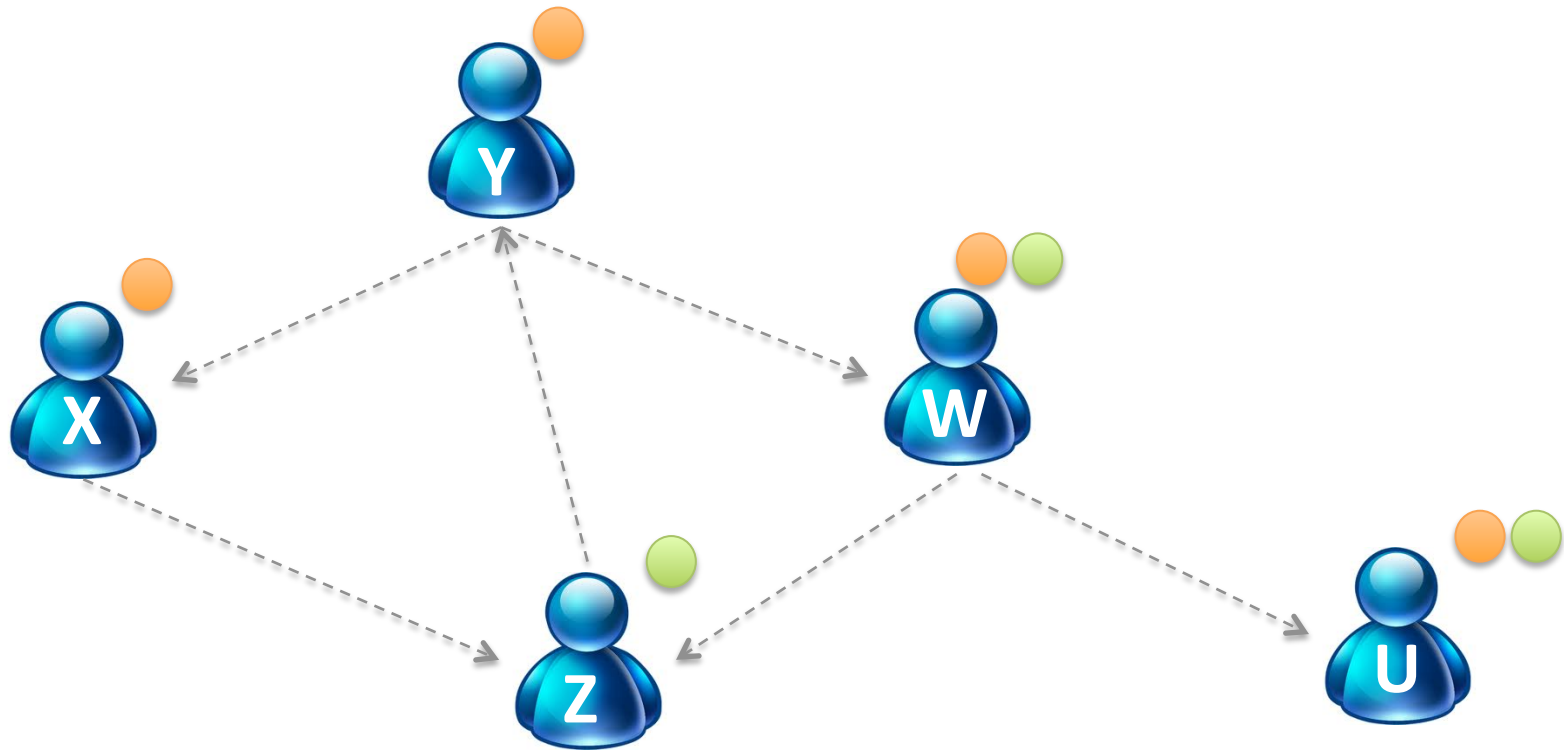
Events	Actions
●	$(Y, t_{Y1}), (X, t_{X1}), (Z, t_{Z1}), (U, t_{U1})$
●	$(W, t_{W2}), (U, t_{U2}), (Z, t_{Z2})$

Learning information diffusion model from events



Given the social network, and a set of events of on it, learn the information diffusion models on it.

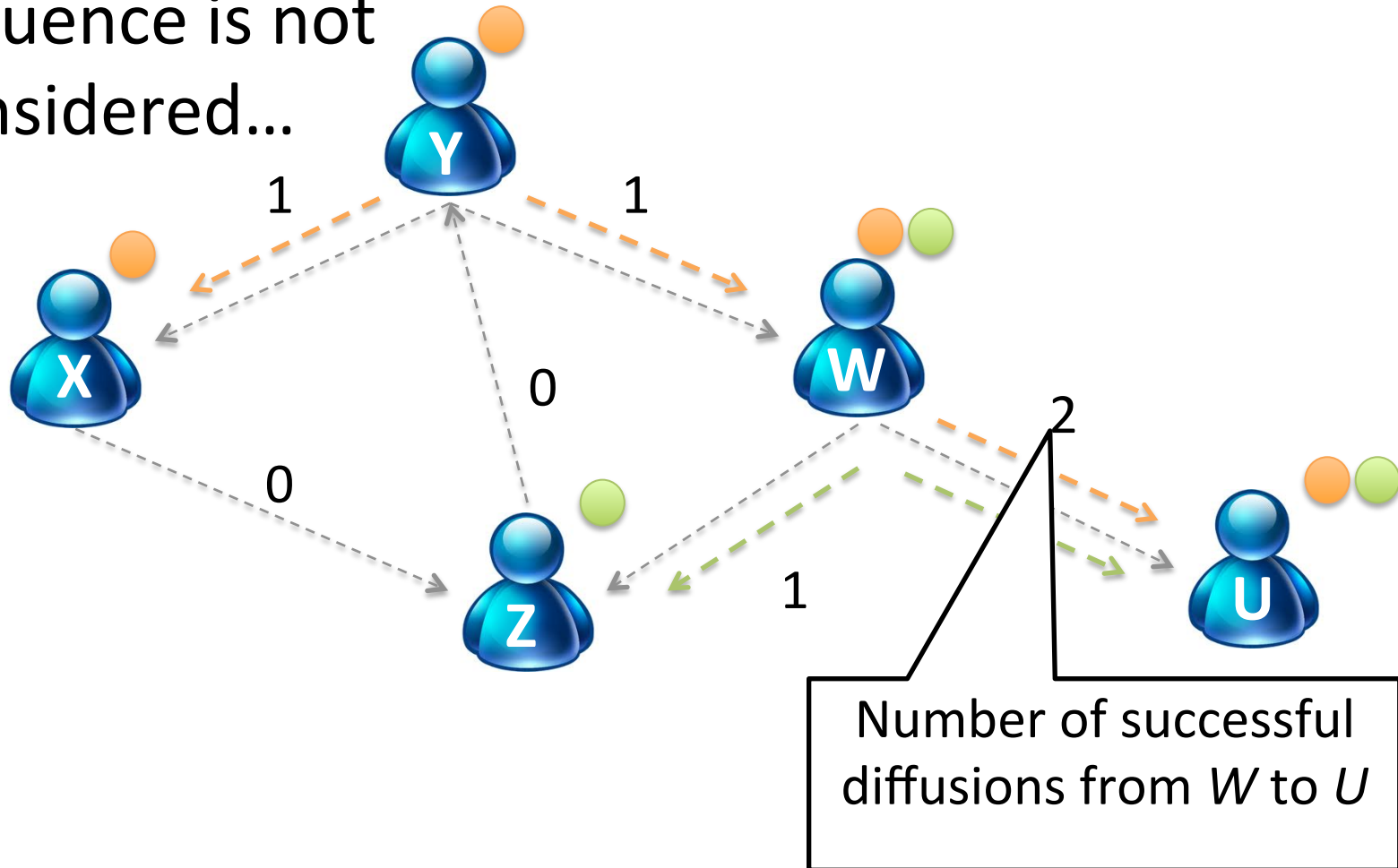
Learning information diffusion model from events



Observed data:
Social network and events

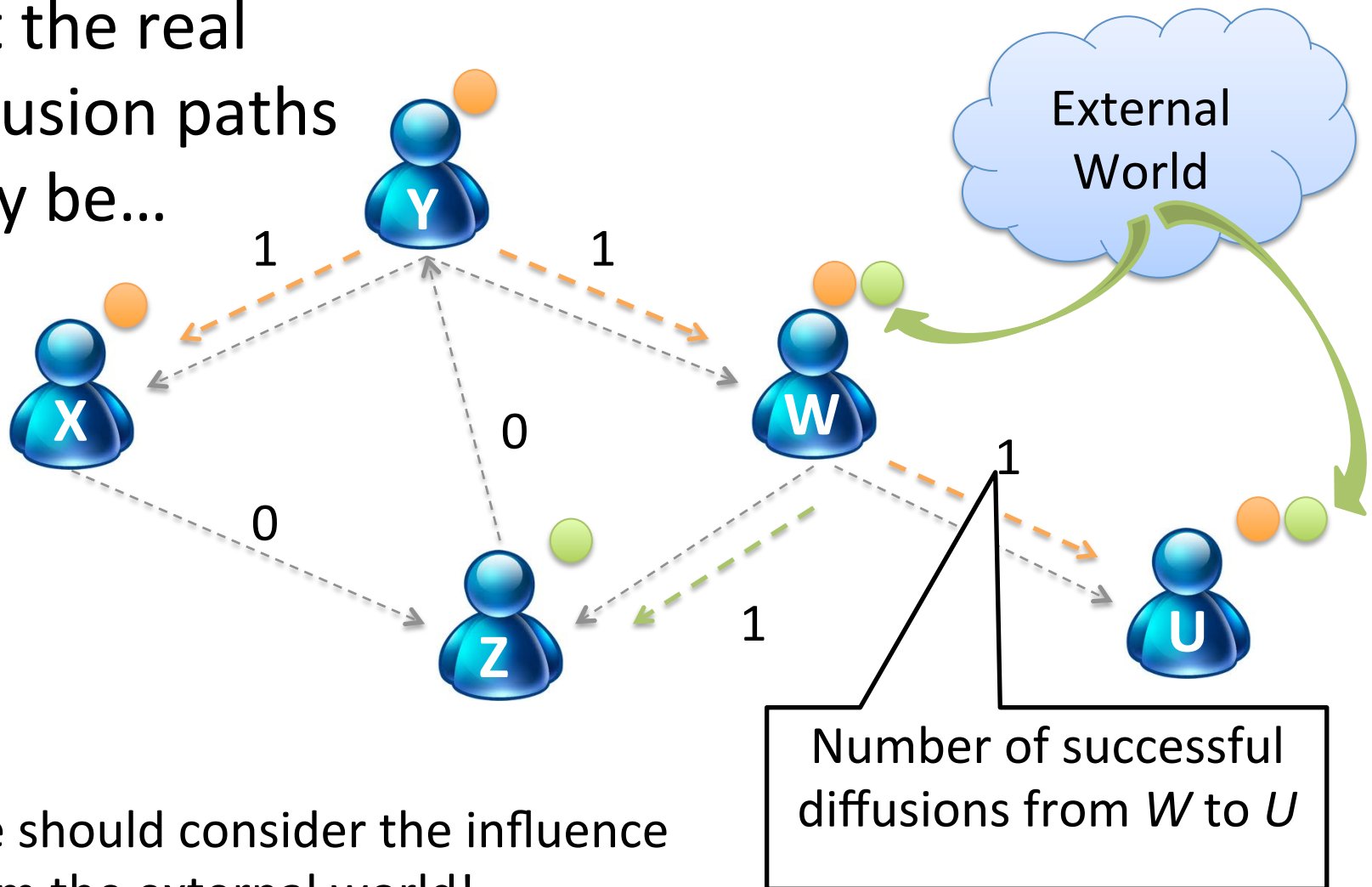
Learning information diffusion model from events

When external Influence is not considered...



Learning information diffusion model from events

But the real diffusion paths may be...



We should consider the influence from the external world!

Social Influence vs. External Influence

Example: two trending Twitter hashtags in 2011.

- #DidYouKnow
 - This hashtag was used in tweets where people talked about surprising facts. It became popular because of social influence among users in the Twitter network.
- #JapanEarthquake
 - This hashtag mainly reflects the external event of 2011 Tōhoku earthquake and tsunami.

Challenge 1: How do we define the sources of influence?

We may define the sources of influence on different levels:

- On the user level
 - Not good. We know it is depend on the events.
- On the event level
 - Not good. Typically each event has multiple sources.
- On the document level
 - Not good enough.


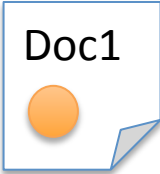



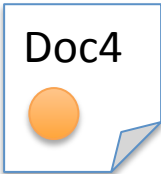
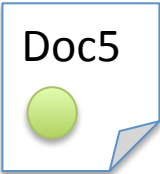
We define the sources of influence on the action level (roughly, every time a user talks about an event).



Challenge 2: Inference dependency

- How can we distinguish the **socially sourced** actions from the **externally sourced** actions?
 - Using the **information diffusion model** and the **external trend model**.
- How can we accurately learn the **information diffusion model** and the **external trend model**?
 - By distinguishing the **socially sourced** actions from **externally sourced** actions

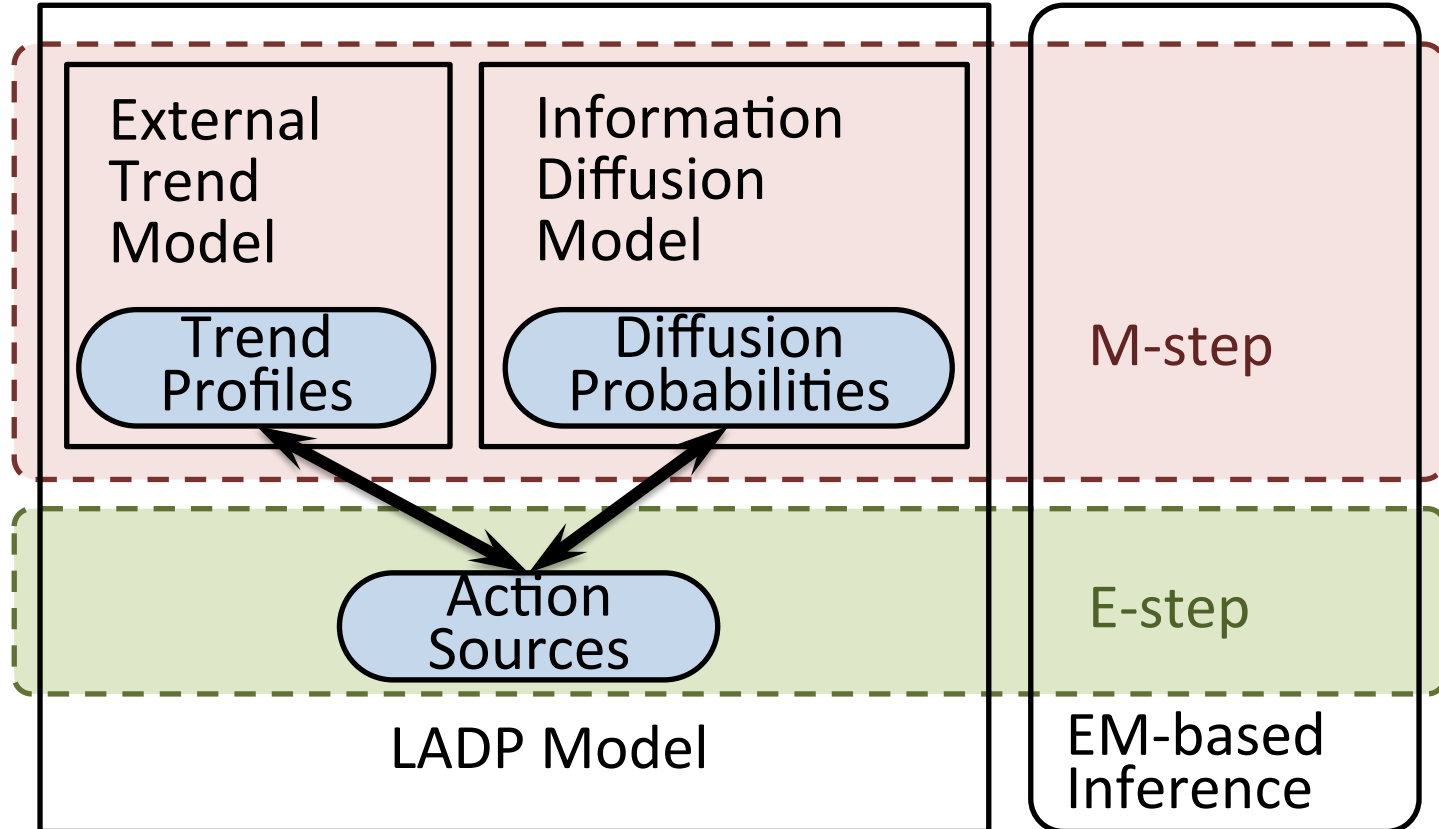
We adopt a mixture model framework, and use a EM-based iterative algorithm for the inference.

Definitions: data stream, event, and actions

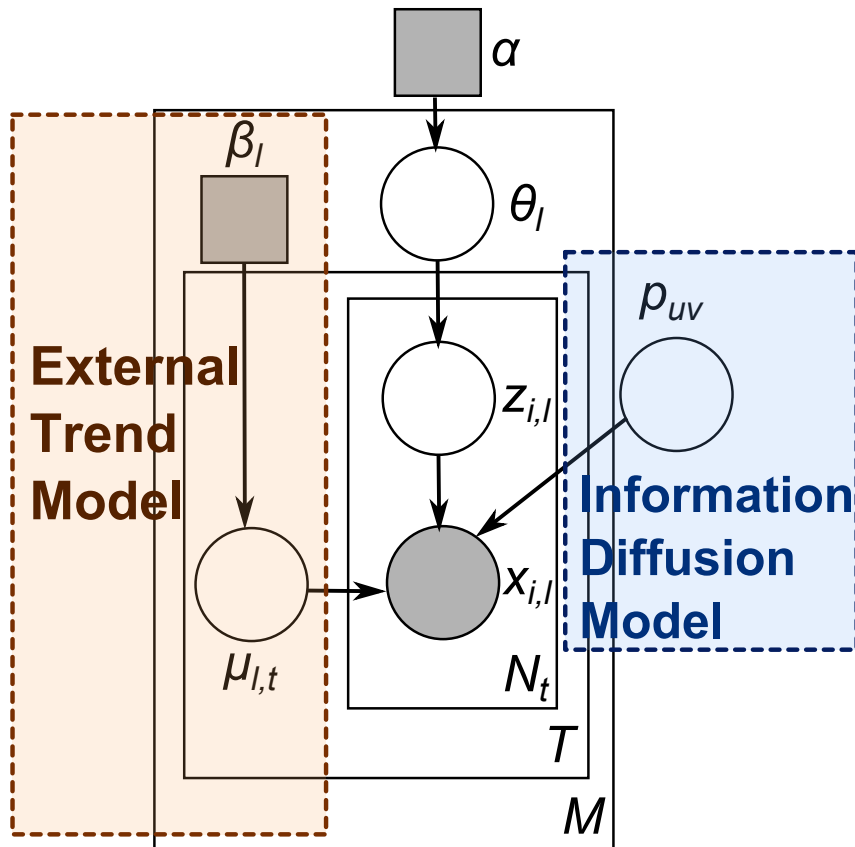
		t_1	t_2	t_3
Data Stream			 	
				

Events	Positive actions	Negative actions
	$(D_1, t_1) (D_2, t_2) (D_4, t_2)$	$(D_3, t_2) (D_5, t_3)$
	$(D_2, t_2) (D_5, t_3)$	$(D_1, t_1) (D_3, t_2) (D_4, t_2)$

The Latent Action Diffusion Path (LADP) Model



The LADP Model



$x_{i,l}$

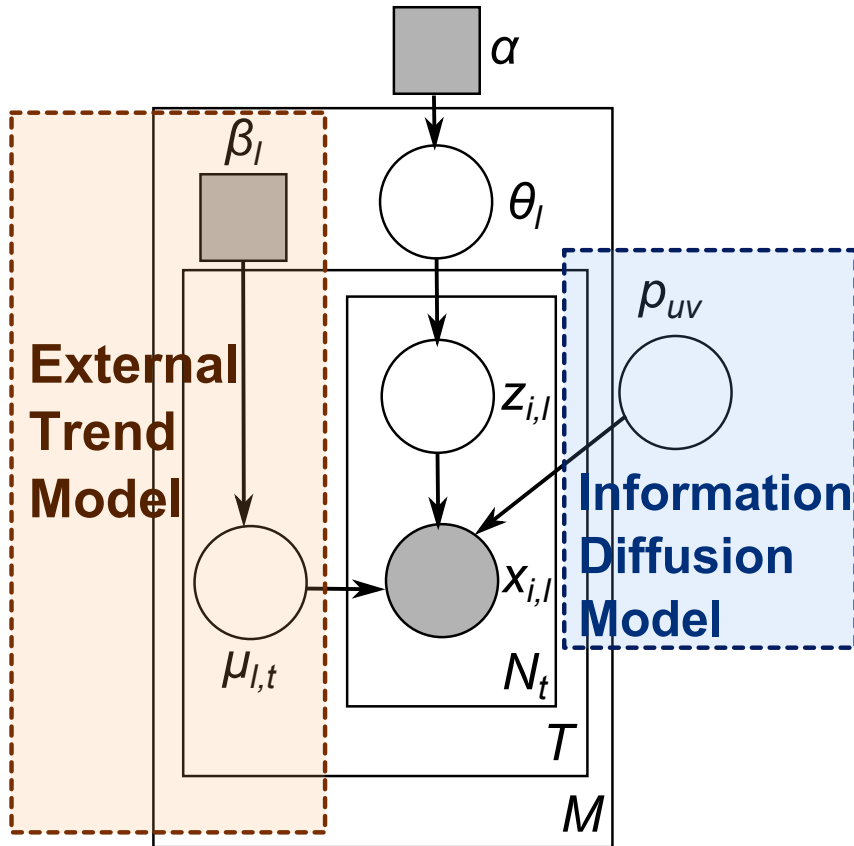
whether the action (i, l) is positive or negative.

$z_{i,l}$

whether action (i, l) is drawn from the information diffusion component.

$$x_{i,l} \sim \begin{cases} \text{Bernoulli}(q_{l,t_{d_i}}, v_{d_i}) & \text{if } z_{i,l} = 1 \\ \text{Bernoulli}(\mu_{l,t_{d_i}}) & \text{if } z_{i,l} = 0. \end{cases}$$

The LADP Model



Information Diffusion Model

$$q_{l,t,v} = 1 - \prod_{d_i \in \mathcal{C}_{t-1}, z_{i,l}=1} (1 - p_{v_{d_i},v})$$

Similar to the Independent Cascade (IC) Model.

$p_{u,v}$ - Diffusion probability with edge (u, v)

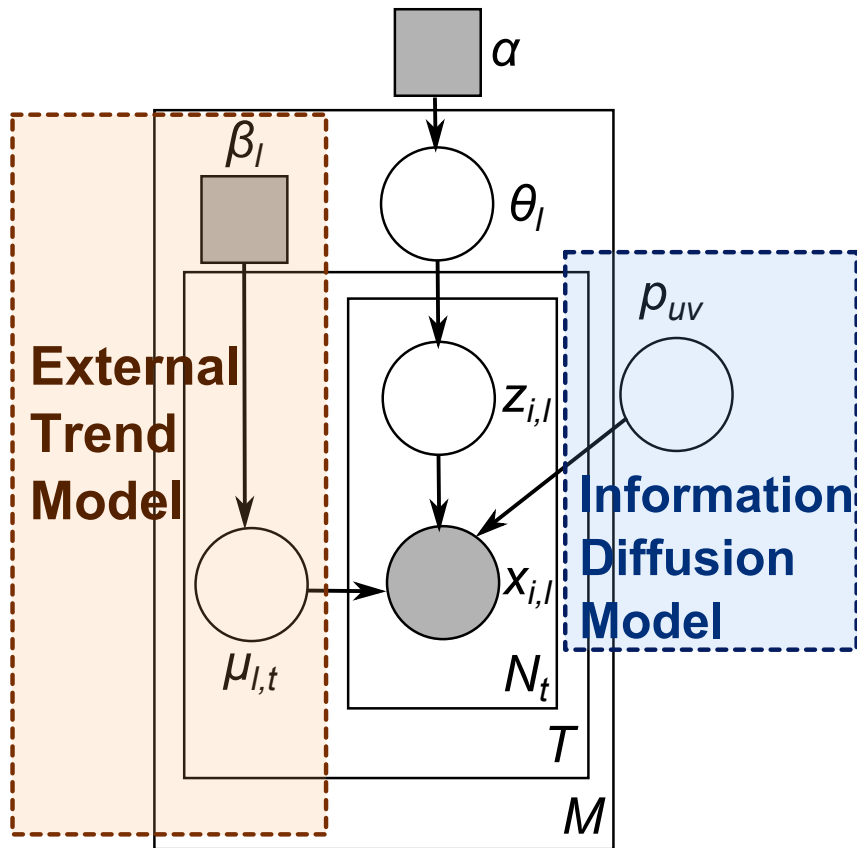
External Trend Model

$$\mu_{l,t} \sim \text{Beta}(\beta_{l,1}, \beta_{l,0})$$

$\beta_{l,1}$ - The average number of positive actions for the term l over all the time steps.

$\beta_{l,0}$ - The average number of negative actions for the term l over all the steps.

EM-based inference algorithm



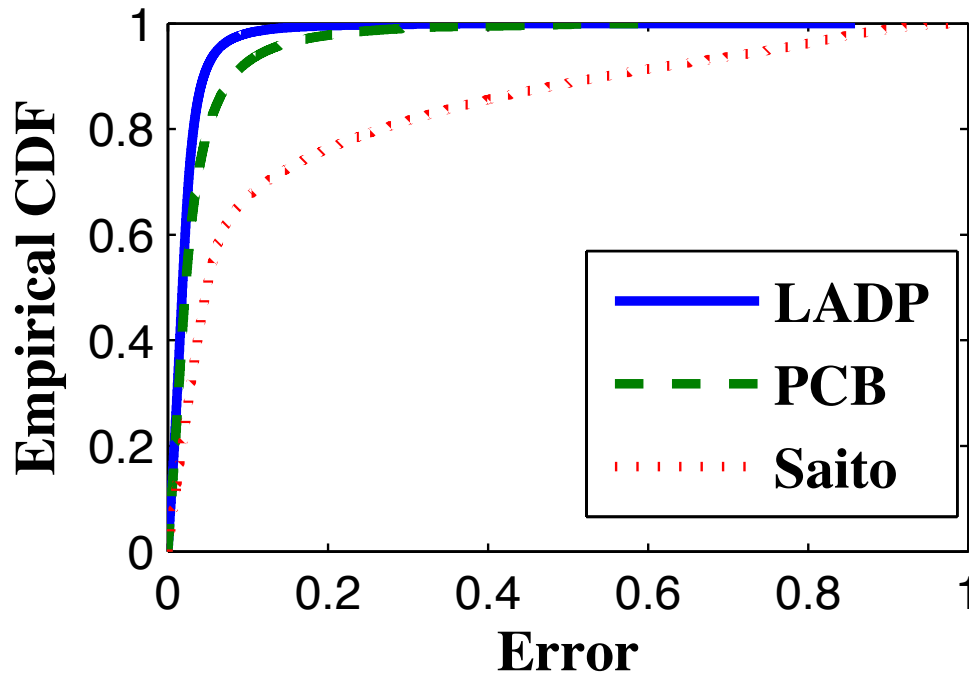
- E-step
 - Calculate the conditional probability of z based on current estimations of parameters.
- M-step
 - Update the estimations of parameters ϑ, μ, p_{uv} .
 - We use approximate estimation of p_{uv} .

Experiment Setup

- Datasets
 - Twitter datasets
 - semi-synthetic dataset
 - Real network with synthetic data stream.
 - Twitter-UIC dataset
 - Real network with real data stream.
 - DBLP datasets
 - 2 communities in the network (data mining and machine learning) and mixture of them.
- Baselines
 - PCB (Goyal 2010)
 - Saito (Saito 2008)

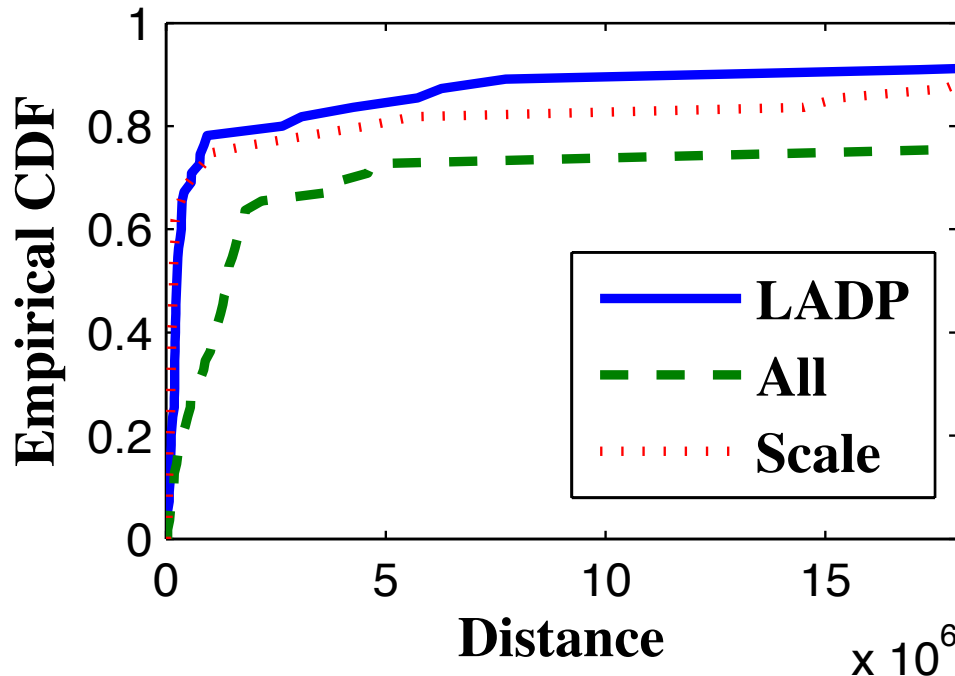
Twitter semi-synthetic Dataset

- How accurate is the inference of diffusion probabilities?



Twitter semi-synthetic Dataset

- How accurate is the inference of socially-sourced portion of events?

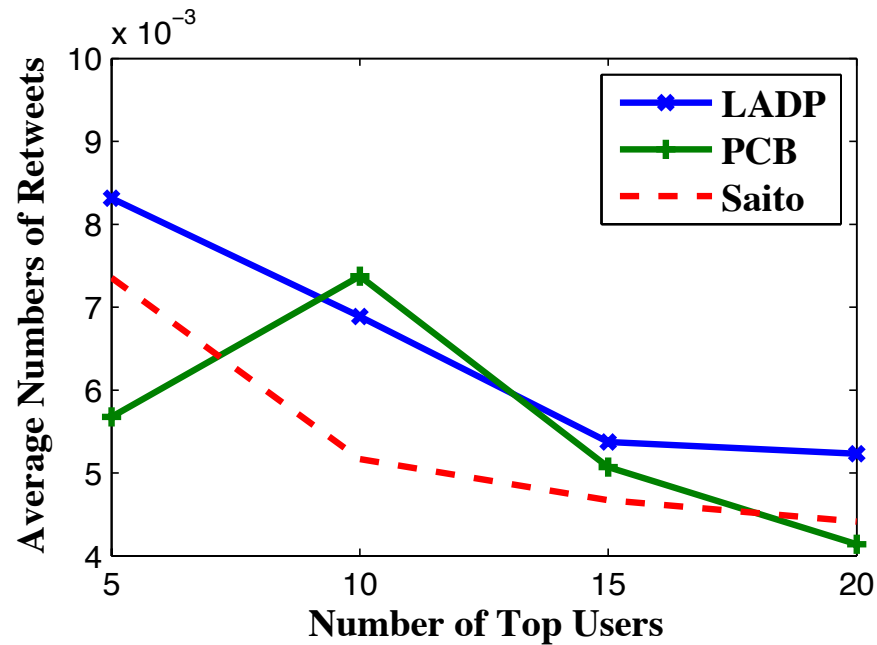


Real Datasets

- How accurate is the inference of information diffusion models?
 - Use the inferred model to decide the most influential nodes in the network.
 - Then evaluate the most influential nodes by known criteria (number of retweets in Twitter, H-index and number of citations in DBLP network)

Twitter network

Evaluation on the most influential nodes.

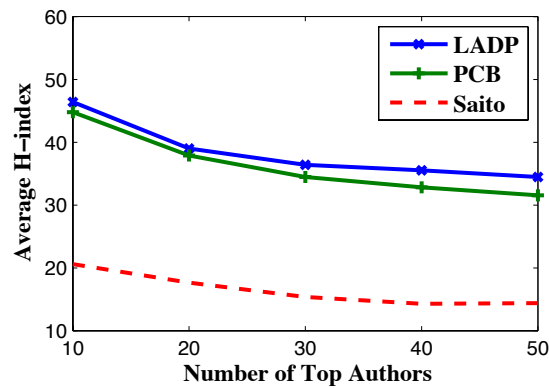


DBLP Datasets

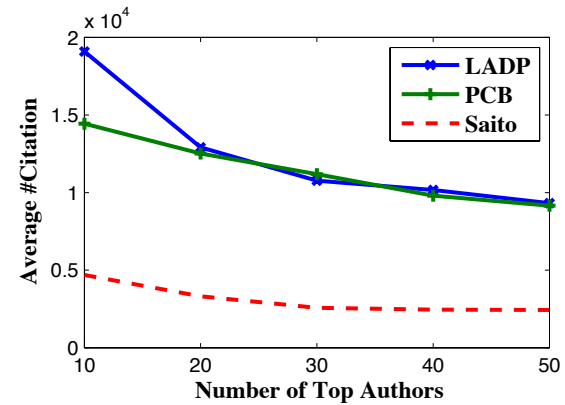
Evaluation on the most influential nodes.

Data mining
community

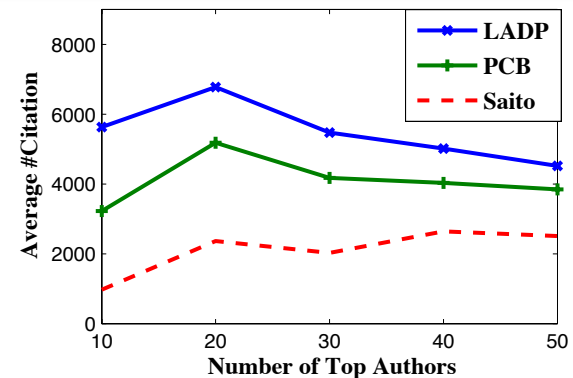
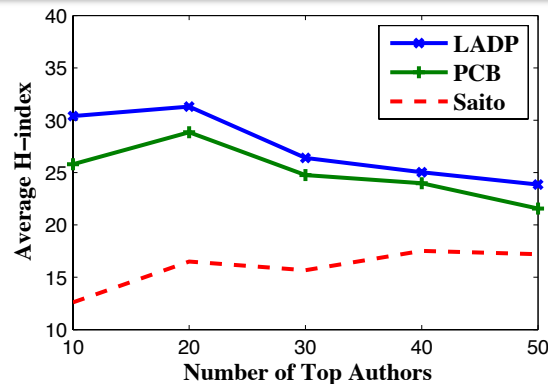
H-index



Number of citation



Machine
learning
community



Inferred socially-sourced portion

All	Inferred socially-sourced portion
1 Chicago	Chicago (-)
2 FF	UIC (↑)
3 UIC	FF (↓)
4 energy	higherEd (↑)
5 higherEd	Illinois (↑)

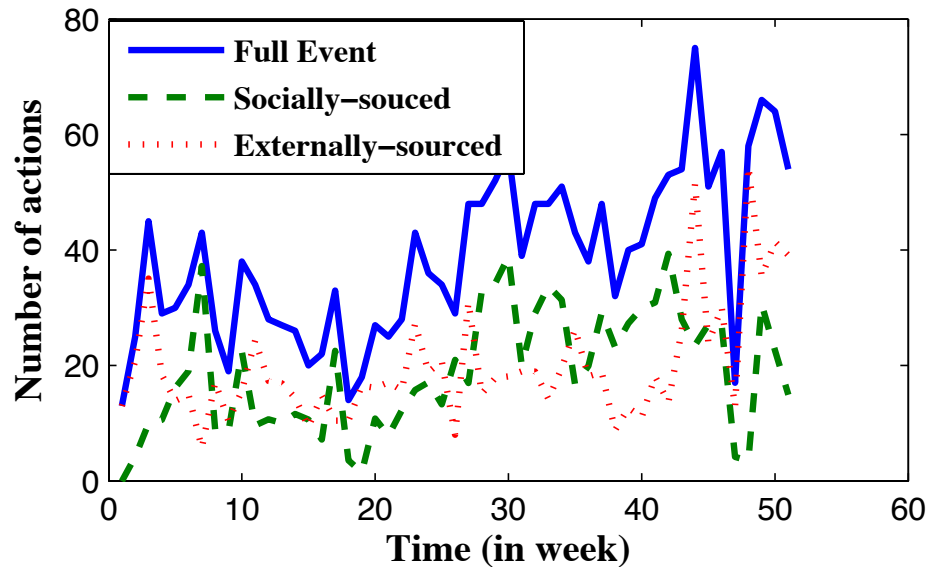
Twitter UIC community

All	Inferred socially-sourced portion
1 Data mining	Data stream (↑)
2 Data streams	Time series (↑)
3 Time series	Data mining (↓)
4 Query processing	Association rules (↑)
5 Association rules	Query processing (↓)

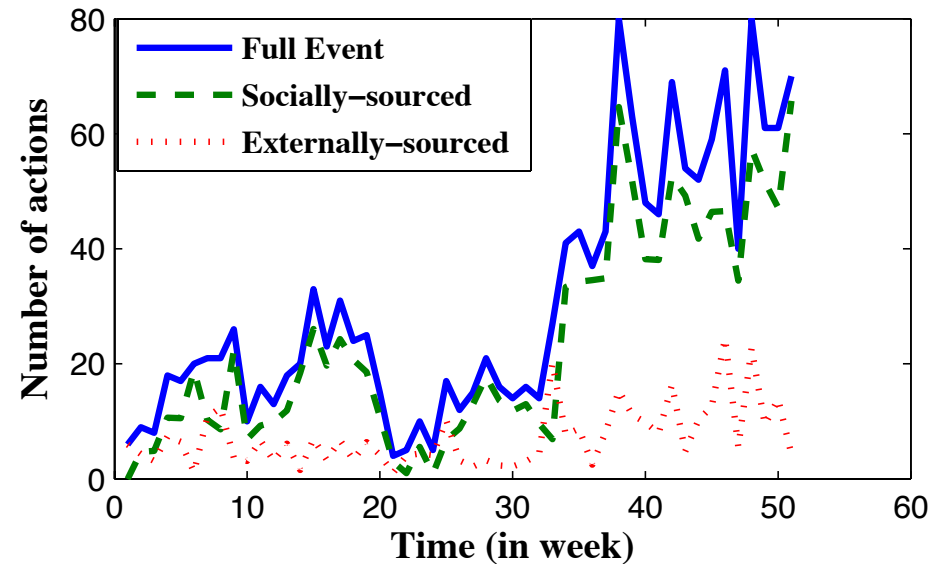
DBLP data mining community

Keywords with higher rankings in the second list are more likely to be related to events in the community.

Case study: two events in Twitter UIC community



FF (Follow Friday)



UIC

The fraction of the event “UIC” that is socially sourced, is larger than that fraction of the event “FF”.

Conclusion

- Extract social sourced portion of events to improve the learning of information diffusion models.
- EM-based iterative algorithm to solve the problem of inference dependency.
- Future work:
 - more sophisticated model for the external trends
 - topic modeling, instead of considering each keyword independently.

Thank you!