Community Level Topic Diffusion

Zhiting Hu[1,3] , Junjie Yao[2], Bin Cui[1] , Eric Xing[1,3]

[1]Peking Univ., China

[2]East China Normal Univ., China

[3]Carnegie Mellon Univ.

OUTLINE

- ❑ **Background**
- ❑ **Model: COLD**
- ❑ **Diffusion Prediction & Analysis**
- ❑ **Experimental Results**
- ❑ **Conclusion**

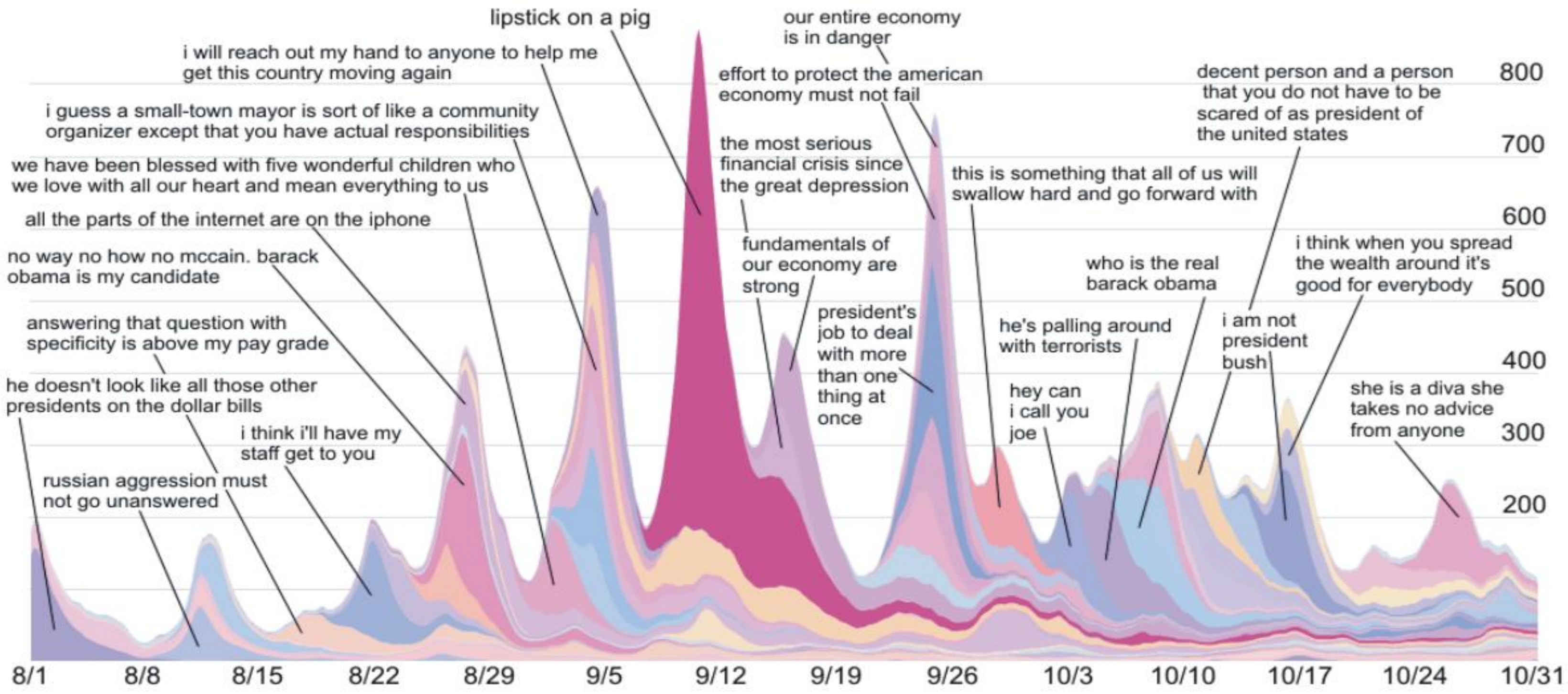
Social Media

- Facebook/Twitter/Weibo/ …
- *One in every five* people in the world uses Facebook (2014)
- Every day around *500 million* tweets are tweeted on Twitter (2015)

http://www.statisticbrain.com

Rich Temporal Dynamics

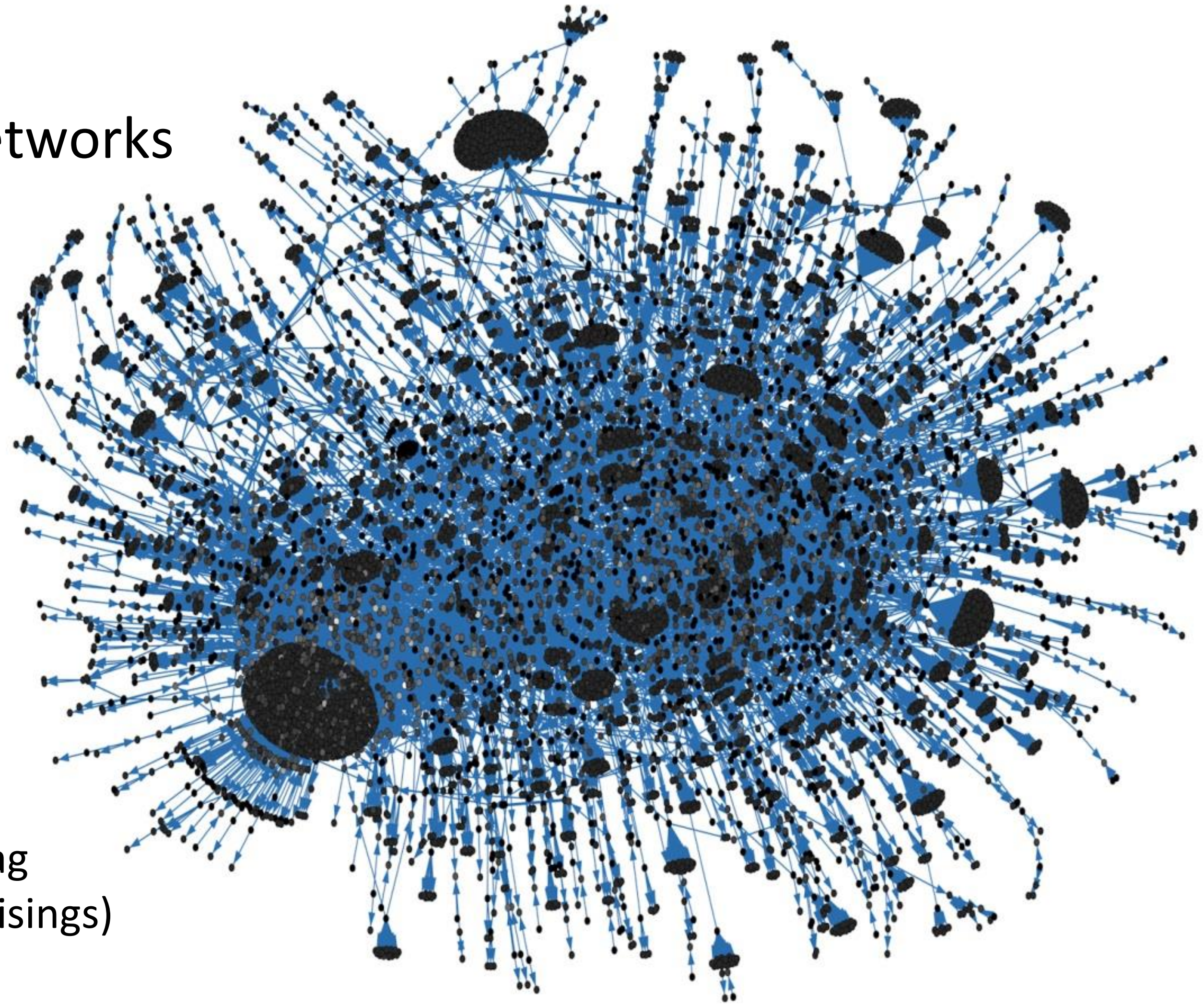
- Popular topics vary over time



J. Leskovec et al., “Meme-tracking and the Dynamics of the News Cycle”. KDD’09

Rich Temporal Dynamics

- Popular topics vary over time
- Messages forwarded across social networks

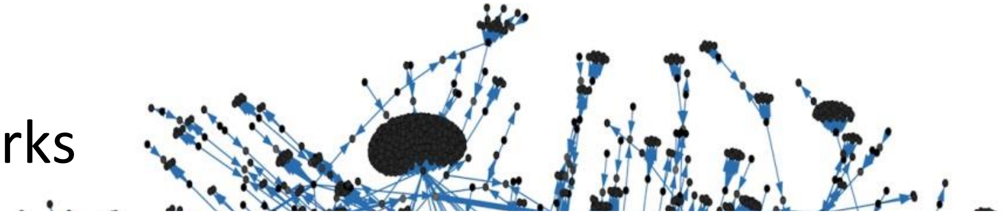


Retweet network of #Egypt hashtag
(the Arab Spring and the 2011 uprisings)

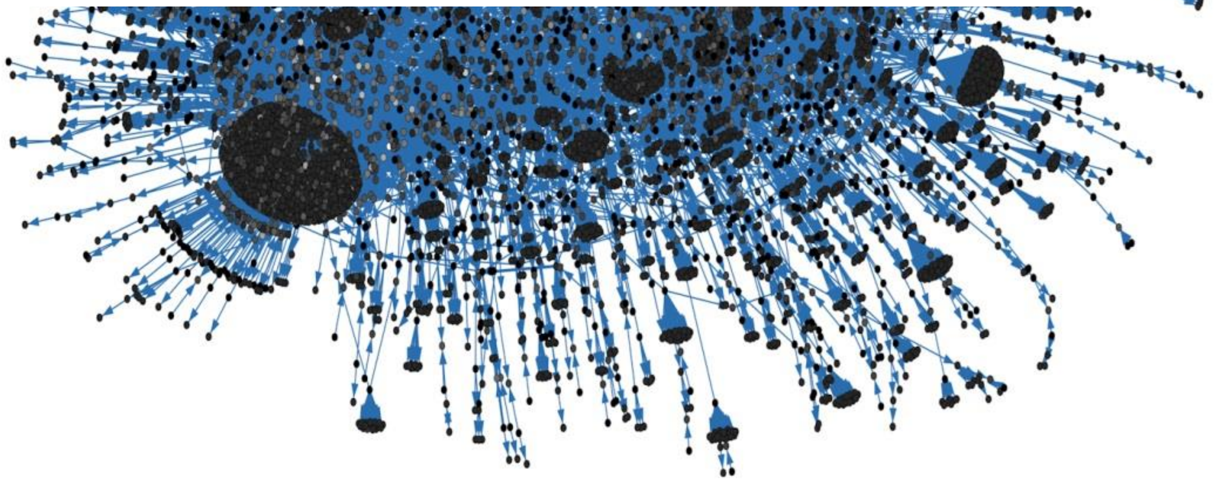
http://phys.org/news/2012-04-tweets-die-network-competition-attention.html

Rich Temporal Dynamics

- Popular topics vary over time
- Messages forwarded across social networks

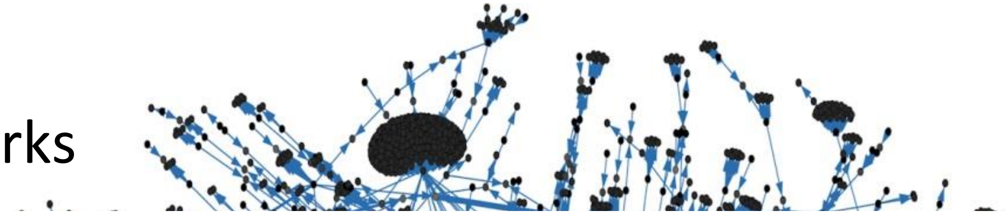


Who says **W**hat to **W**hom in **W**hat channel with **W**hat effect?

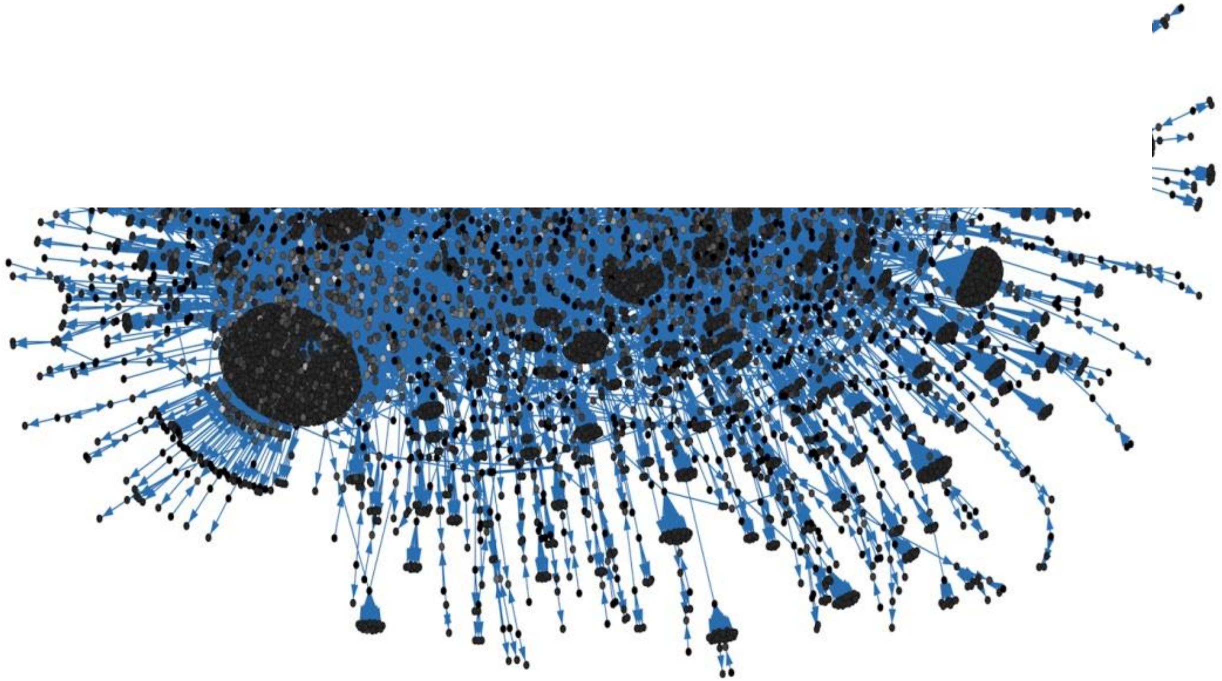


Rich Temporal Dynamics

- Popular topics vary over time
- Messages forwarded across social networks



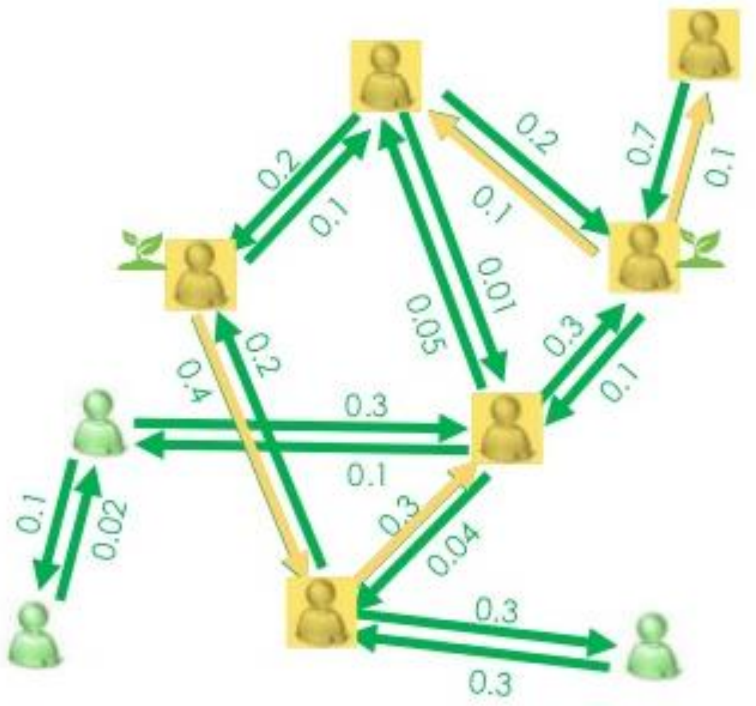
Who says What to Whom in What channel with What effect?



- Online marketing
- Information/friend RecSys
- Search
- ....

Previous Work

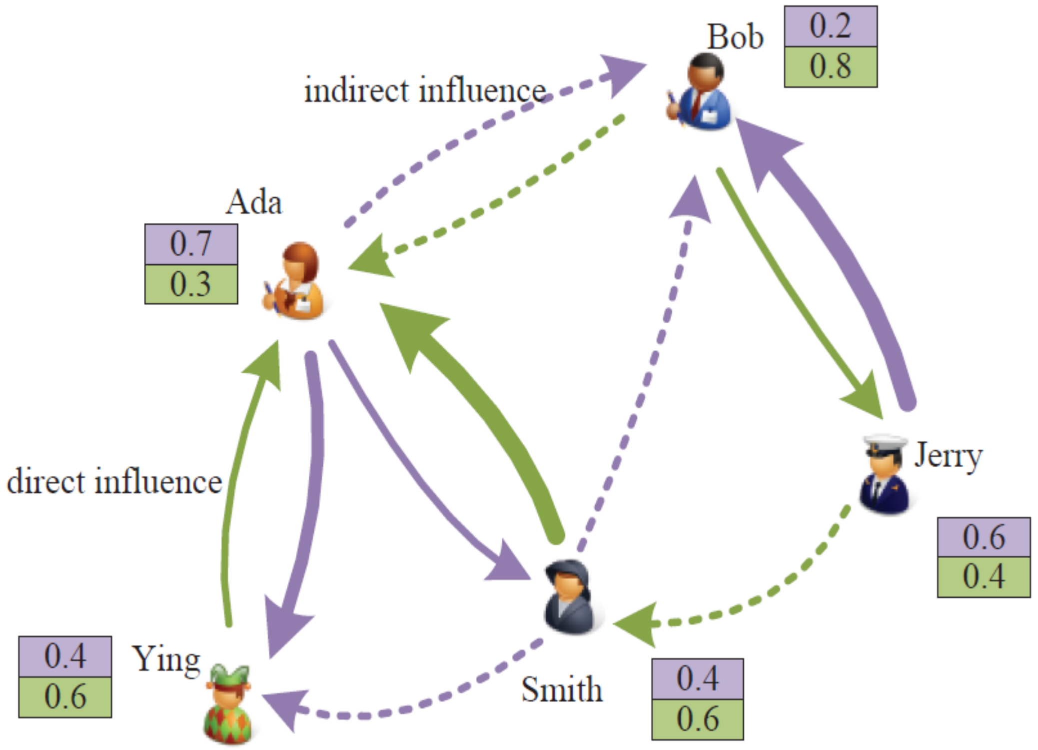
- Information propagation across networks
  - Models interactions between *individuals*, and structured topologies



J. Goldenberg et al., Independent cascade model

Previous Work

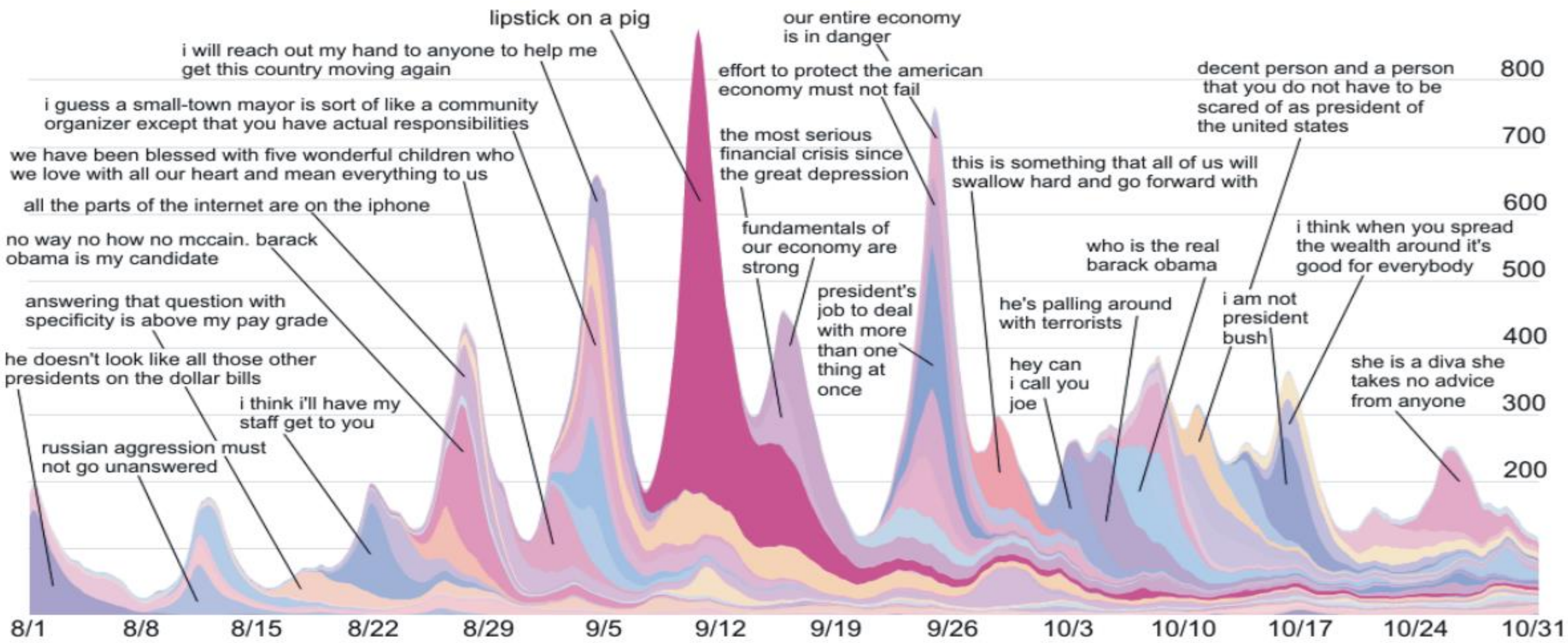
- Information propagation across networks
  - Models interactions between *individuals*, and structured topologies



L. Liu. “Mining Topic-level Influence in Heterogeneous Networks”. CIKM’10

Previous Work

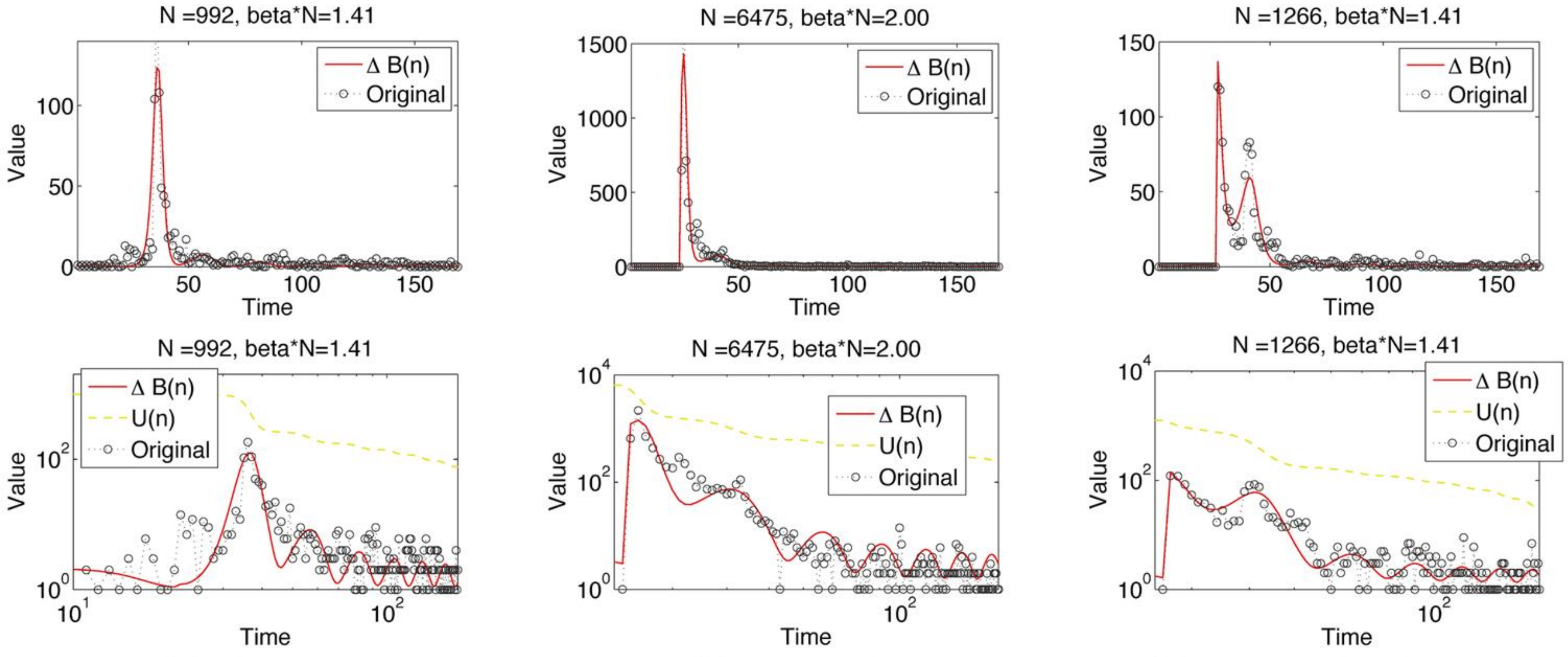
- Temporal topic modeling
  - Captures *aggregated* temporal trends of online content



J. Leskovec et al., “Meme-tracking and the Dynamics of the News Cycle”. KDD’09

Previous Work

- Temporal topic modeling
  - Captures *aggregated* temporal trends of online content

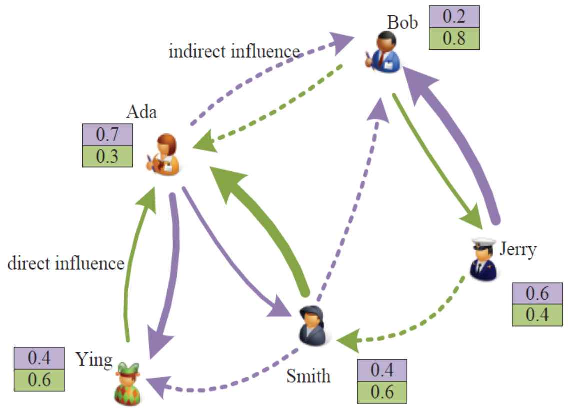


(a) #assange (b) #stevejobs (c) #arresteddevelopment

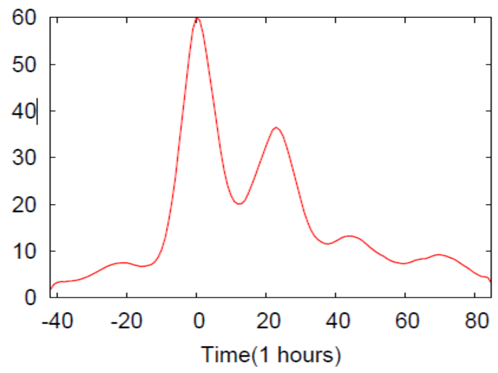
Y. Matsubara et al., “Rise and Fall Patterns of Information Diffusion: Model and Implications”. KDD’13

Limitations

- Individual-level diffusion
  - volatile individual behaviors
  - hard to accurately uncover



- Aggregated information dynamics
  - cannot reveal detailed dissemination process



Can we unify these different lines?

Can we unify these different lines?

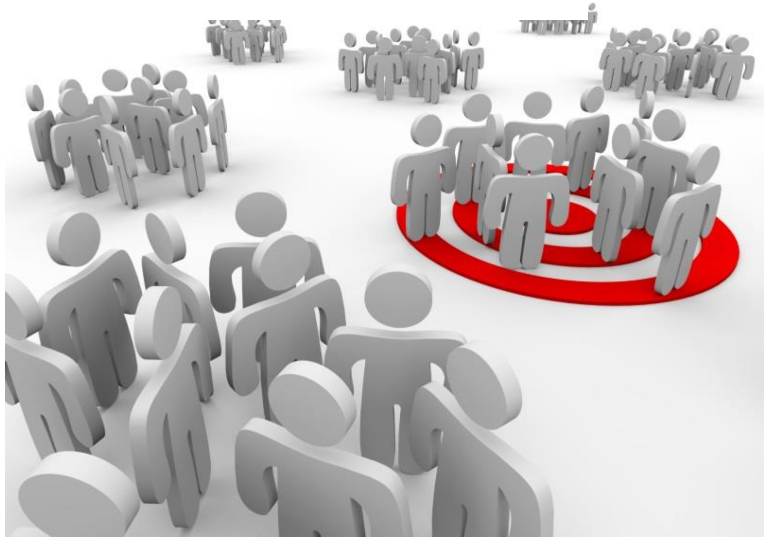
Community level diffusion extraction

- modeling diffusion patterns of topics across different *communities*

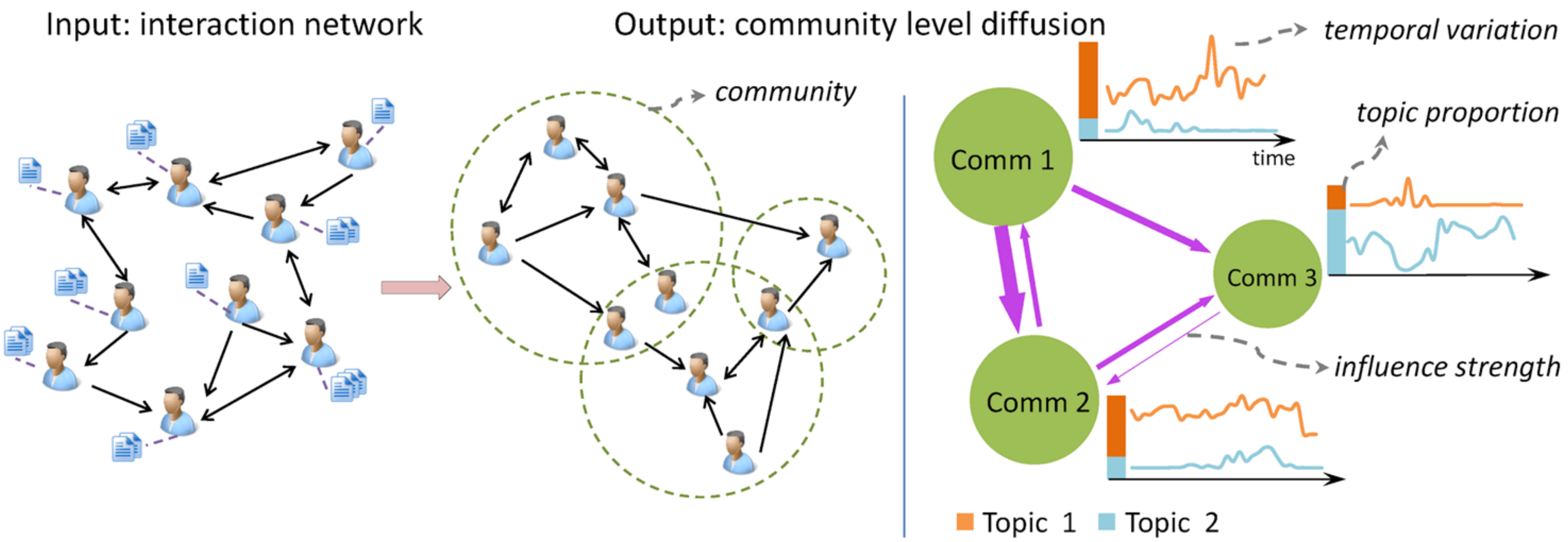
Community level diffusion extraction

Community
- provides the basis for user engagement in social networks
- ``Strength of Weak Ties'' theory
  - a critical role of inter-community interactions in online diffusion.

- Collective user behaviors
  - more predictable than those of individuals

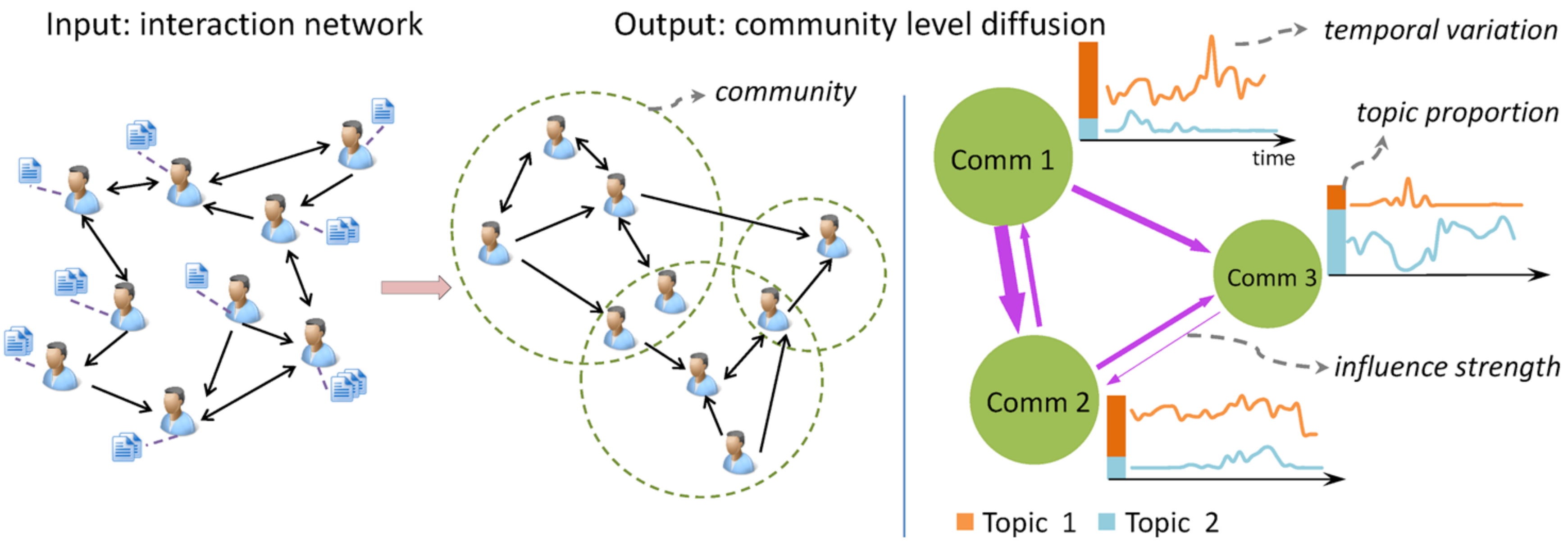


Community level diffusion extraction



- Input
  - an interaction network among users
  - user-generated content over time
- Goal: to uncover
  - hidden communities, their interests in different topics
  - topic temporal variation within communities
  - Influence strength between communities

Community level diffusion extraction



- Input
  - an interaction network among users
  - user-generated content over time
- Goal: to uncover
  - hidden communities, their interests in different topics
  - topic temporal variation within communities
  - Influence strength between communities

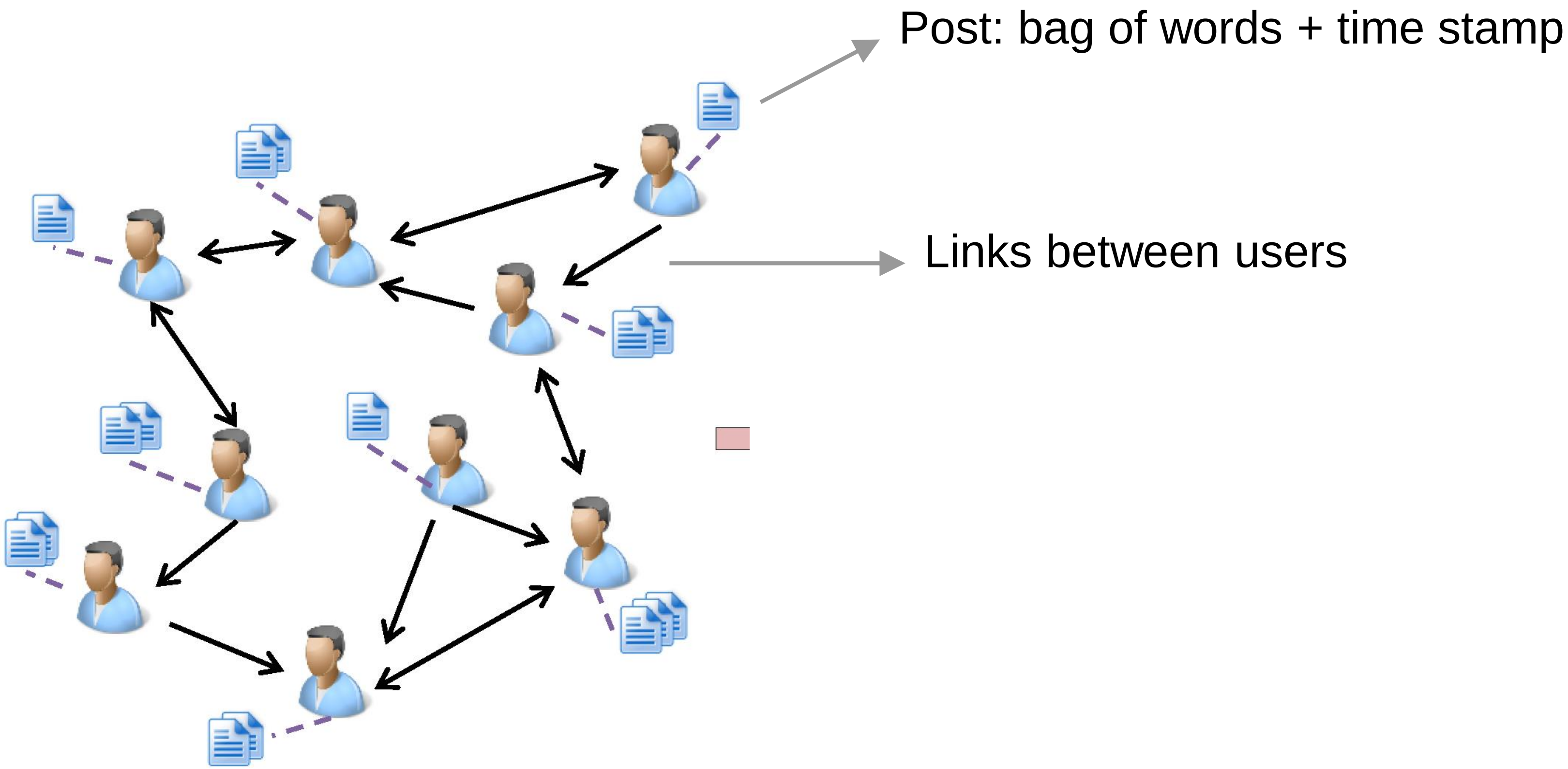
- Compact community-level representation
- more accurate prediction and analysis

OUTLINE

- Background
- **Model: COLD**
  - Diffusion Prediction & Analysis
- Experimental Results
- Conclusion

Problem Formulation

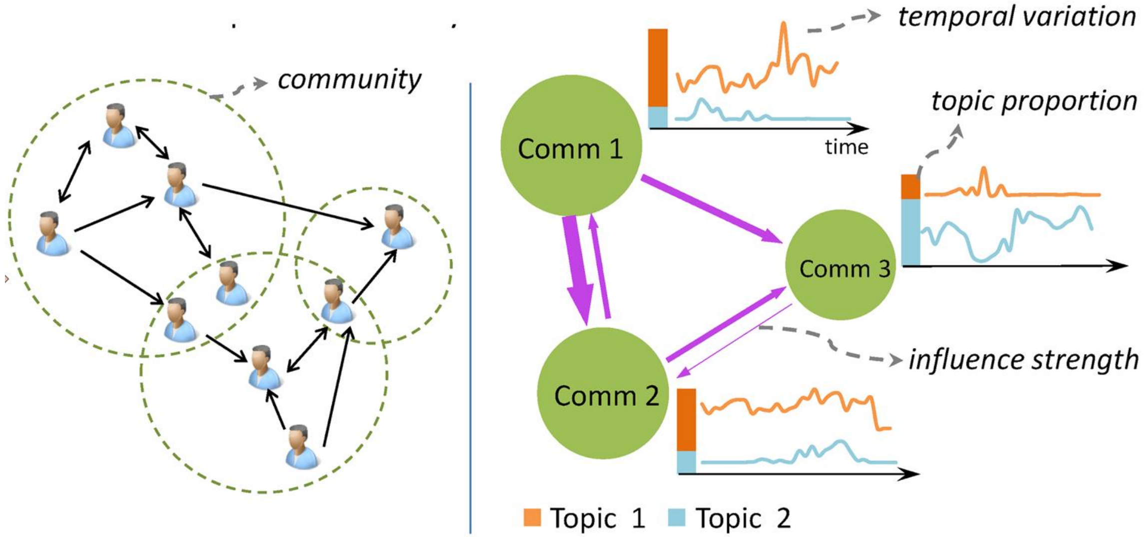
- Consider an interaction network among users
- Two types of data (user behaviors)
  - text data (posting)
  - network data (social interaction)



Problem Formulation (cont.)

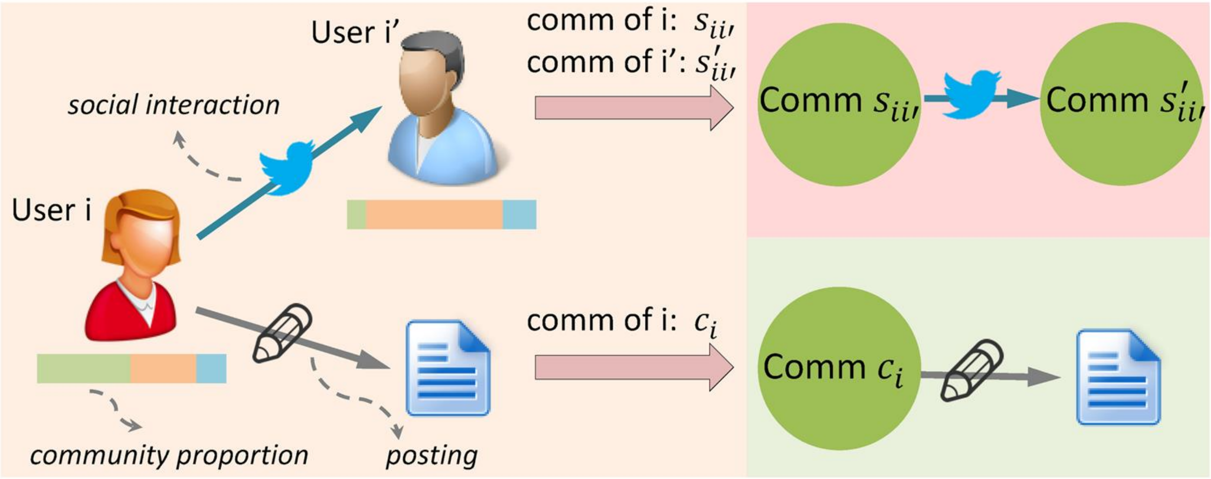
Assume:

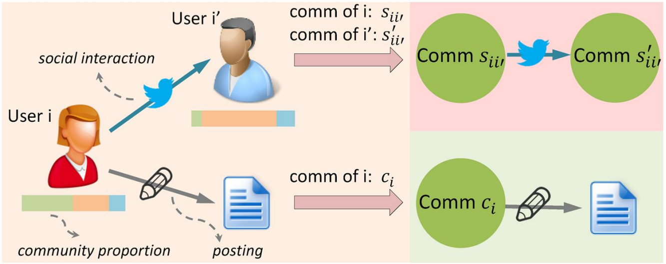
- C Communities
  - membership: each user i has a multinomial distribution over communities: $\boldsymbol{\pi}_i$
  - interest: each community c has a multinomial distribution over topics: $\boldsymbol{\theta}_c$
- K Topics
  - content: a multinomial distribution over words: $\boldsymbol{\phi}_k$
  - variation: a multinomial distribution over time stamps in each community c: $\boldsymbol{\psi}_{kc}$
- Community level influence strength
  - For each topic k, the diffusion probability between two communities c and c': $\zeta_{kcc'}$



<u>CO</u>mmunity <u>L</u>evel <u>D</u>iffusion (COLD) Model

- Two types of user behaviors: posting & social interaction
- Each user assumes a community membership when taking a behavior
- The behavior is then explained by the corresponding community-specific context



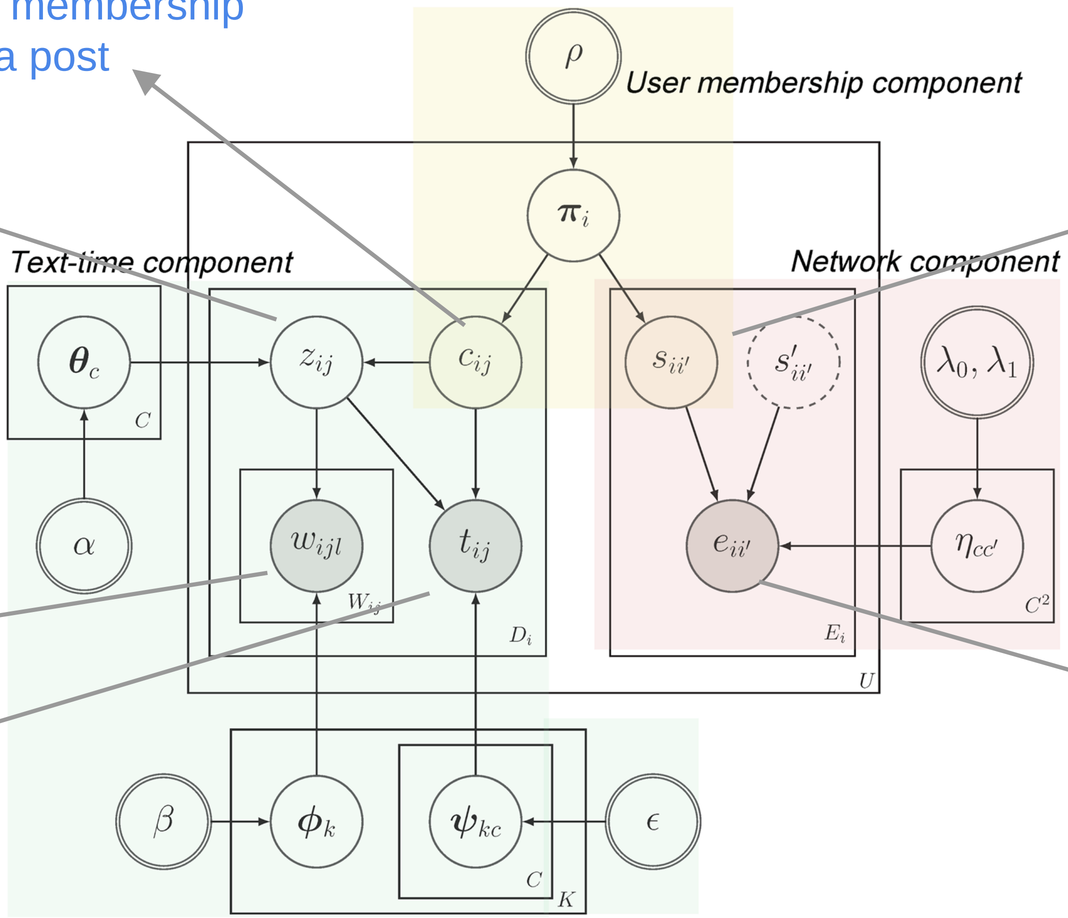


Generative Model

1. community membership when writing a post

2. topic of the post

i) community memberships when interacting



3. words of the post

4. time stamp of the post

ii) link between the two users

Approximate Inference

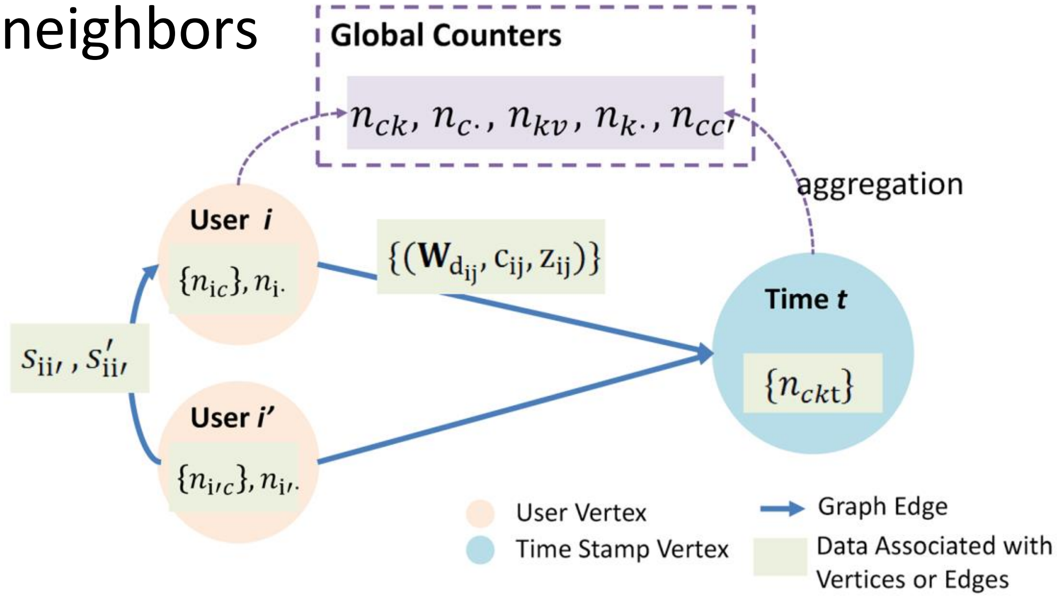
- Gibbs Sampling
  - iteractively samples latent variables
    - community memberships for posting/interaction
    - topics for posts
  - constructs the distributions of interest based on the samples
    - users’ membership distribution
    - communities’ interest distribution
    - topics’ temporal distribution
    - influence strength between communities

- Time Complexity
  - O(#tokens + #links)
  - linear to the data size

Parallel Implementation

- Implement the Gibbs Sampler based on GraphLab
  - simliar to the GraphLab implementation of LDA
- Construct a bipartite graph
  - users ⇔ time stamps
- Sufficient statistics of sampler are stored globally or locally

- Gather:
  - nodes collect sufficient statistics from neighbors
- Apply:
  - nodes update their own data
- Scatter:
  - nodes do sampling



OUTLINE

- ❑ Background
- ❑ Model: COLD
  - ❑ **Diffusion Prediction & Analysis**
- ❑ Experimental Results
- ❑ Conclusion

Community level diffusion

- Topic-sensitive influence strength of community c on c':

$$\zeta_{kcc'} = \theta_{ck}\theta_{c'k}\eta_{cc'}$$

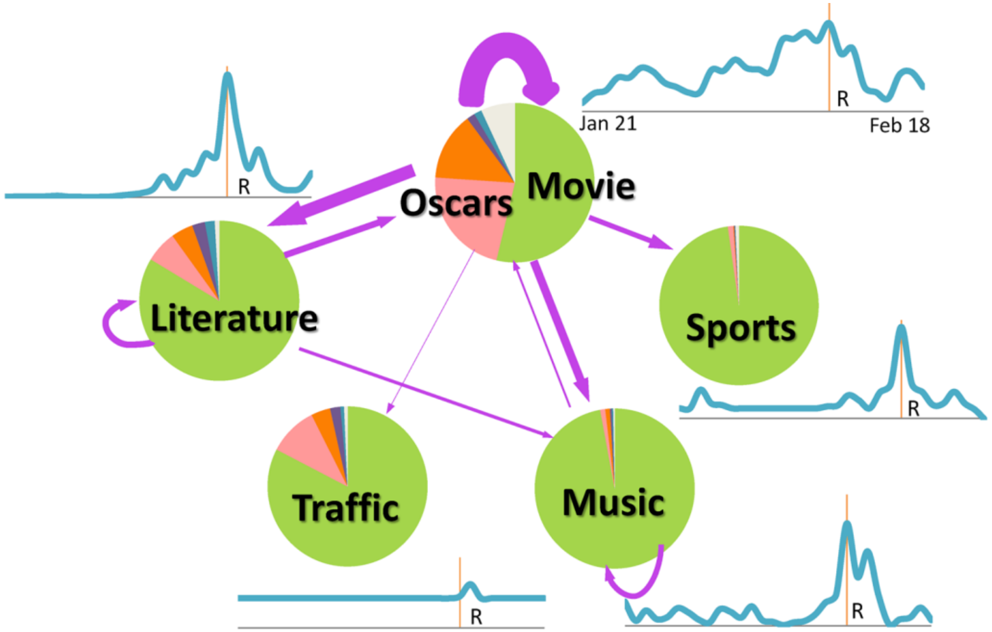
interest level of community c on topic k

interest level of community c' on topic k

(general) influence strength of community c on c'

Community level diffusion

- Topic-sensitive influence strength of community c on c’:



Diffusion Prediction

- Predict whether a post will propagation from one individual to another
  - Given:
    - the words of the post d
    - its author i
    - another user i’
  - Goal:
    - infer the probability user i’ retweets the post d from user i
- Previous methods model the individual level probability directly
  - volatility of individual's actions
  - sparsity of individual's records
- Ours: community members' *collective behavior* patterns
  - stable and predictable

Diffusion Prediction (cont.)

- Given:
  - the words of the post d
  - its author i
  - another user i’
- Goal:
  - infer the probability user i’ retweets the post d from user i: $P(i, i’, d)$

- Infer the topic of post based on its words and author: $P(k | d, i)$
- The influence of user i on user i’ on topic k: $P(i, i’ | k)$
- Combine the above:
  - P(i, i’, d) = \sum_k P(k | d, i) P(i, i’ | k)

- Time complexity:
  - O(K*|d|)
  - K: #topics, |d|: length of the post

OUTLINE

- ❑ Background
- ❑ Model: COLD
  - ❑ Diffusion Prediction & Analysis
- ❑ **Experimental Results**
- ❑ Conclusion

Setup

- Two datasets crawled from Weibo.com
  - Data-1: 53K users, 11M posts, 91M words; 2.7M links
  - Data-2: 0.52M users, 14M posts, 112M words; 10M links

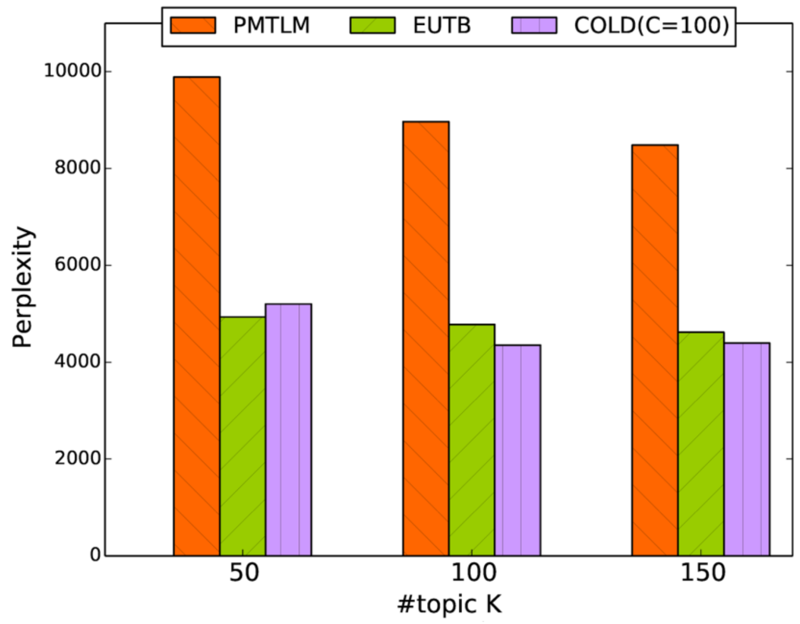
- Baselines & Tasks

	features			tasks			
	text	social	time	topic ext	comm detec	temp modl	diff pred
PMTLM [39]	•	•		•	•		
MMSB [1]		•			•		
EUTB [37]	•	•	•	•		•	
Pipeline	•	•	•	•	•	•	
WTM [31]	•	•					•
TI [20]	•	•		•			•
COLD	•	•	•	•	•	•	•

Table 2: Feature and Task Comparison of Different Methods

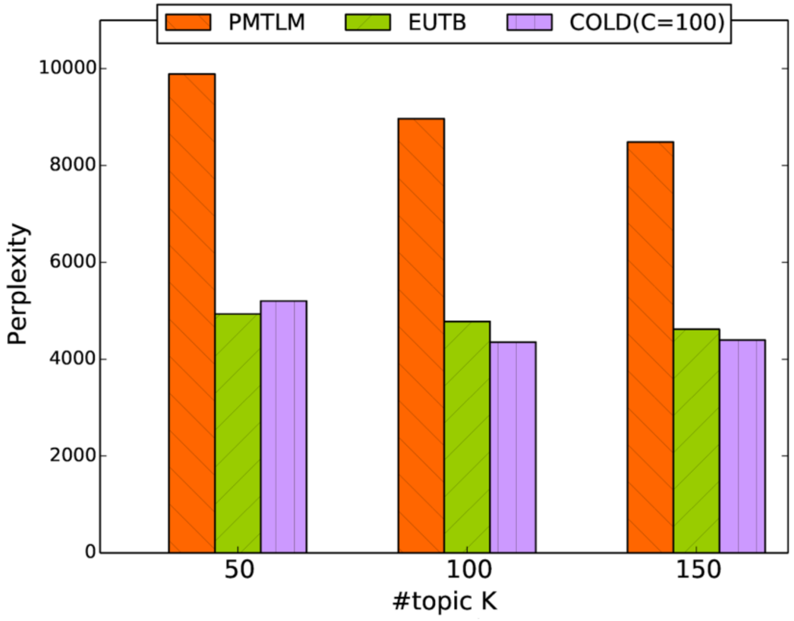
Task 1: Topic Extraction

- Topic perplexity (the lower the better)
  - the predictive power of a probabilistic model
  - proportional to the cross-entropy between the word distribution learned by the model and the actual distribution in test set



Task 1: Topic Extraction

- Topic perplexity (the lower the better)
  - the predictive power of a probabilistic model
  - proportional to the cross-entropy between the word distribution learned by the model and the actual distribution in test set




Sports

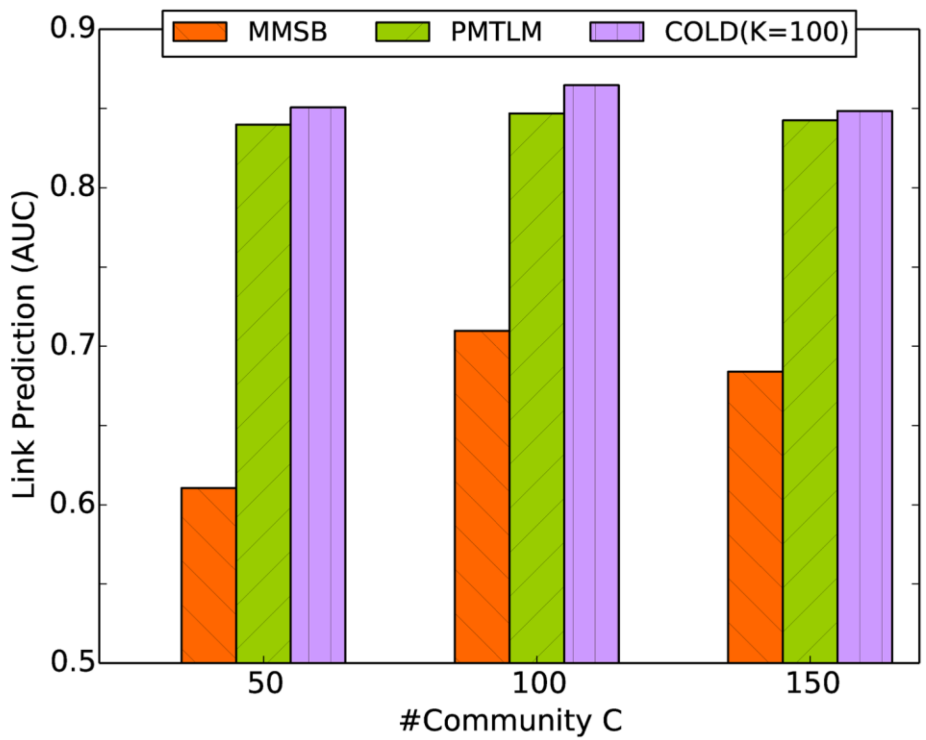

Oscars2013


New Year


IT

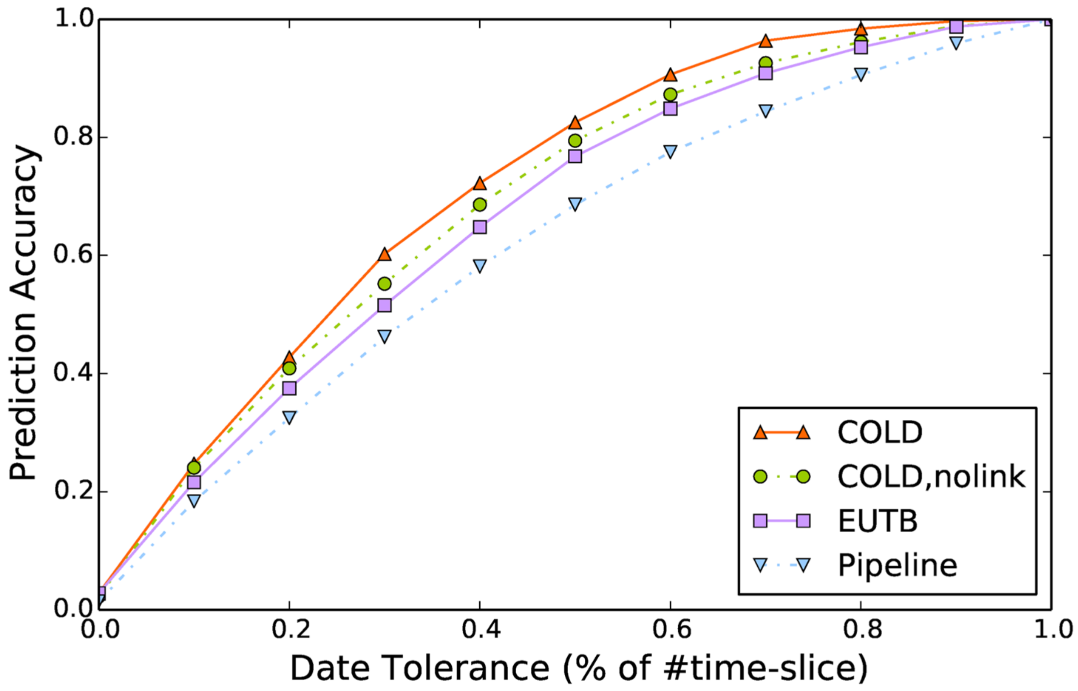
Task 2: Community Detection

- Link Prediction
  - widely-used when no ground truth of community memberships is available
  - AUC: the higher the better:
    - the probability that a randomly chosen true positive link is ranked above a randomly chosen true negative link



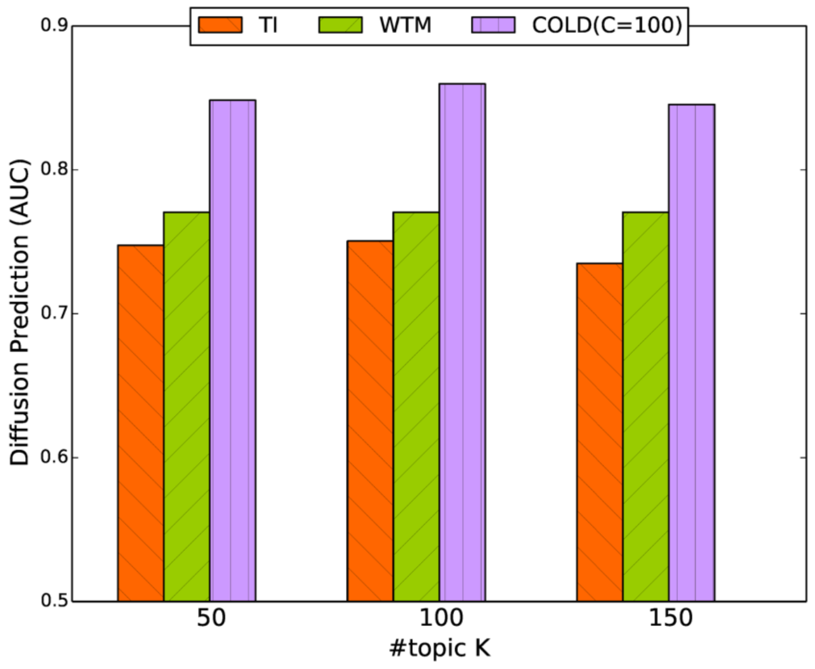
Task 3: Temporal Modeling

- Time-stamp Prediction
  - Estimate the time stamp of a post given its words and author
  - Accurarcy: the higher the better



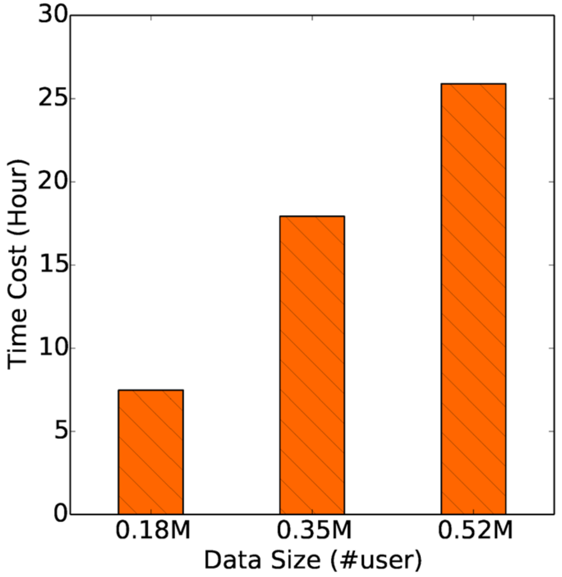
Task 4: Diffusion Prediction

- Diffusion prediction
  - Predict whether a post by a user will be retweeted by another user
  - AUC: the higher the better

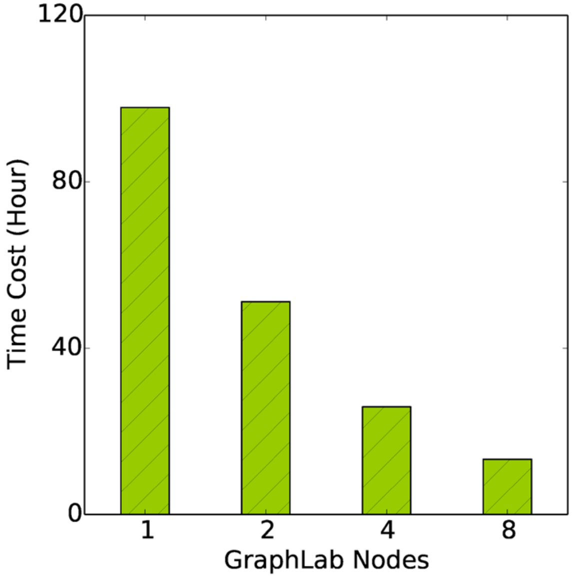


Efficiency

- Training time



(a) Three Subsets, 4 Nodes.



(b) Whole Dataset, # nodes.

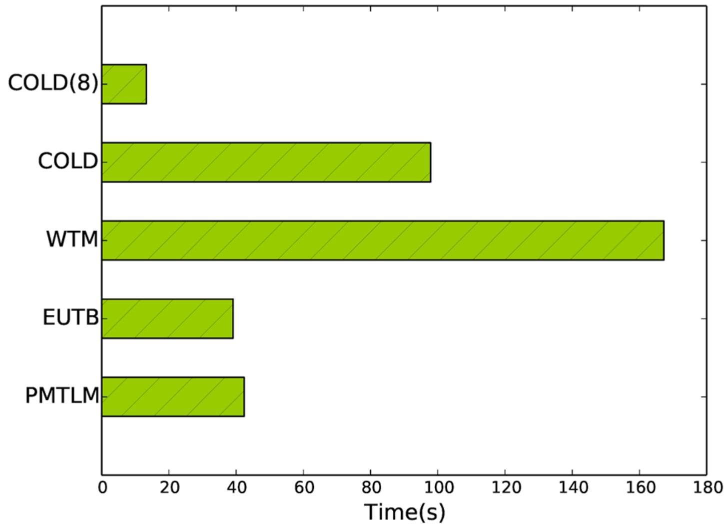
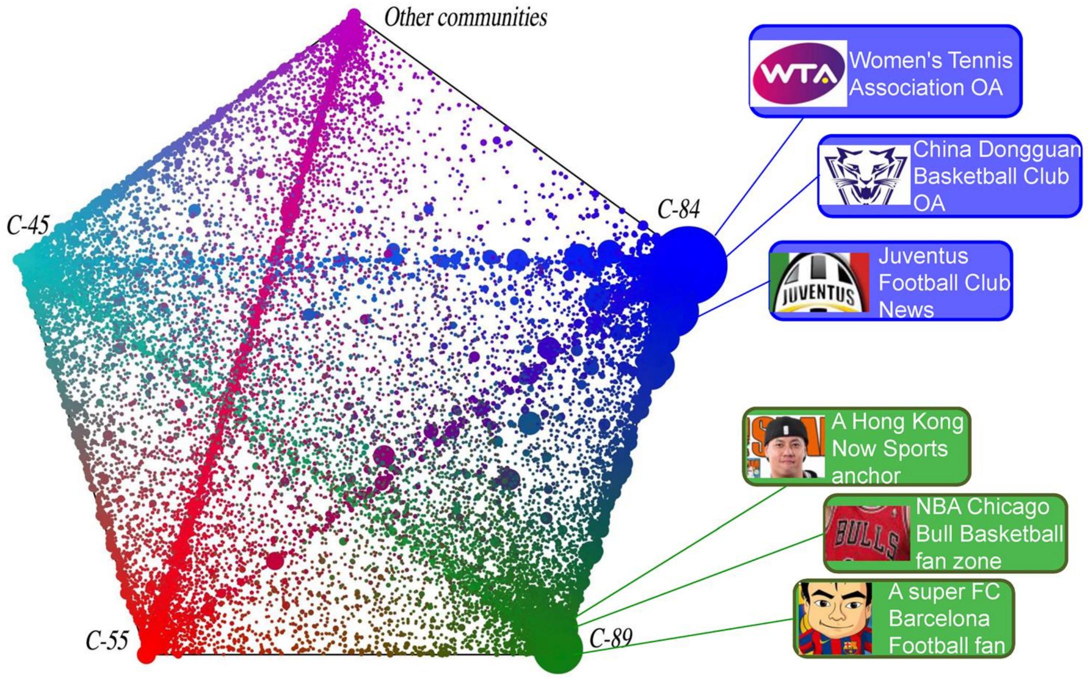


Figure 14: Training Time (C=K=100). "COLD (8)" is the distributed implementation on 8 nodes.

A cluster of Linux machines, each with 8 2.4GHz CPU cores and 48G memory.

Other Application

- identify *influential communities*
  - compute the influence degree of each community
    - setting the single community as the seedset
    - applying the Independent Cascade on the community-level diffusion graph



Conclusions

- Novel Perspective
  - the problem of community level diffusion
- COLD Model
  - a latent model to uncover
    - the hidden topics and communities
    - the community-specific temporal diffusion.
  - parallel implementation
- Prediction & Exploration
  - An effective diffusion prediction approach leveraging community level patterns
  - Other tasks, e.g., community detection, influential community identification, etc.