Data-Driven Behavioral Analytics:
Observations, Representations and Models

Meng Jiang (UIUC)
Peng Cui (Tsinghua)
Jiawei Han (UIUC)

http://www.meng-jiang.com/tutorial-cikm16.html

Tutorial in CIKM 2016, October 24, Indianapolis, IN

What is Behavior?

- *Definition*. **Interactions** made by **individuals** in conjunction with **themselves** or their **environment**. (*Wikipedia*)

Behavioral Analysis

- *Significance*. What can we discover from behavioral data?
  - *Ex*. Given every phone call/message between military leaders, scientists, businesspersons, find…

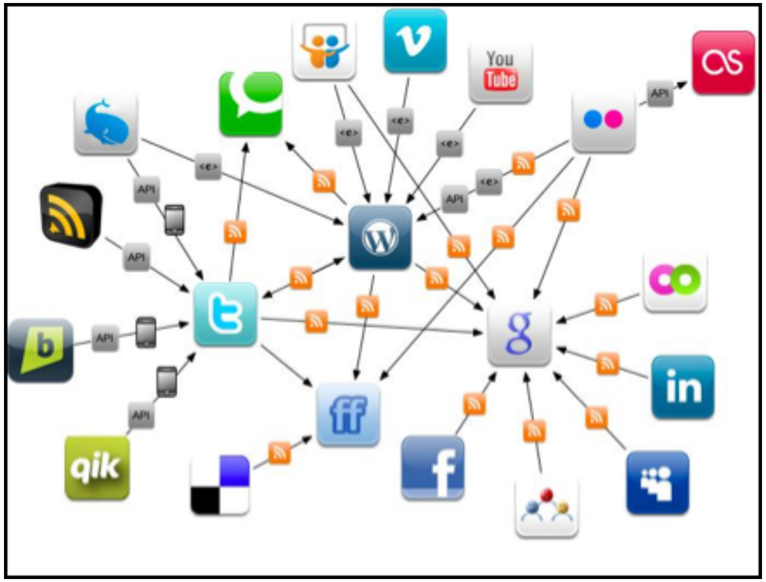
Observations	Who, what, where, when, why, how… (scientific view)
Representations	Graph, network, matrix, tensor… (mathematical view)
Models	Prediction, recommendation, anomaly detection… (application view)

Why Behavioral Analysis Today?

- *Today*. The human behaviors are broadly recorded in an unprecedented level. Insights of sciences and society?

Physical World

Online Applications



Basic Research Areas

- Six Disruptive Basic Research Areas
  - Engineered Materials (metamaterials and plasmonics)
  - Quantum Information and Control
  - Cognitive Neuroscience
  - Nanoscience and Nanoengineering
  - Synthetic Biology
  - Computational Modeling of Human and Social Behavior

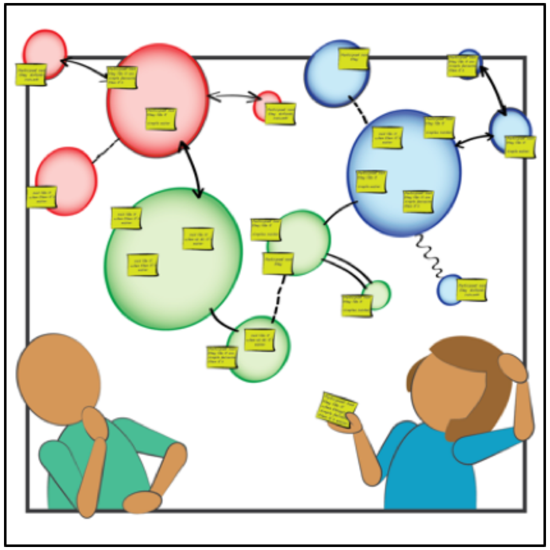
VI. Computational Models of Human Behavior

A fundamental understanding and predictive capability of human behavior dynamics from individuals to societies.

- ***Enabled capabilities***
  - *Predictive models supporting strategic, operational, and tactical decision making and planning*
  - *Real time cultural situational awareness*
  - *Immersive training and mission rehearsal*
  - *Cross cultural coalition building*
- ***Key research challenges:***
  - *Conflicting theories*
  - *Data management and fusion*
  - *Mathematical complexity*
  - *Validation of models*

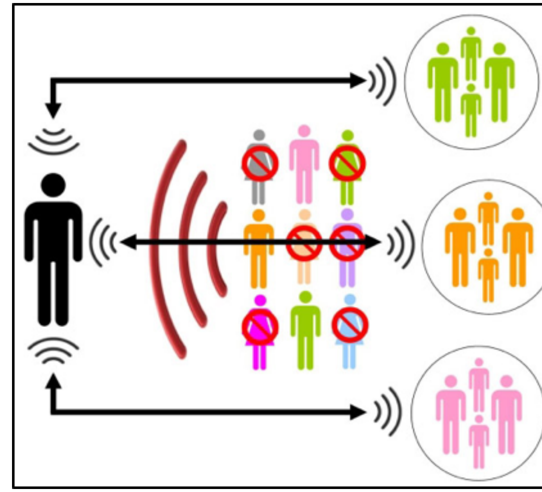
- ***Measures of success***
  - *Early success of simple models*
  - *Success of social network analysis*
  - *Prediction of crowd tipping points*

Challenges in Behavioral Analysis



Content
(preference)

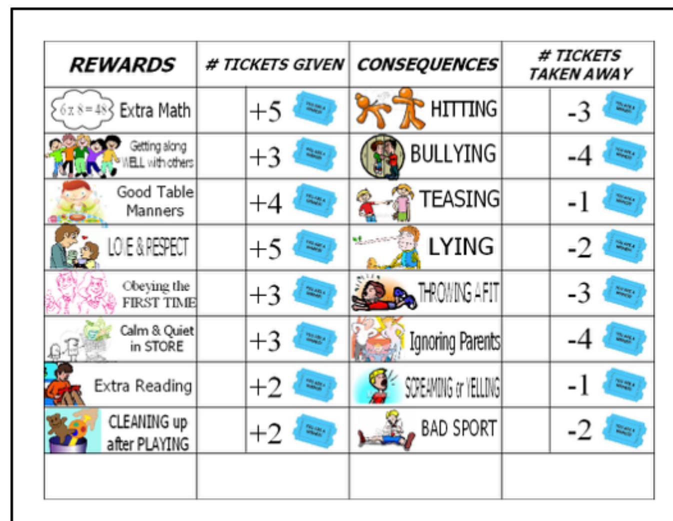
Social context
(influence)



Behavioral
Analysis

Spatiotemporal context

Intention
(suspiciousness)

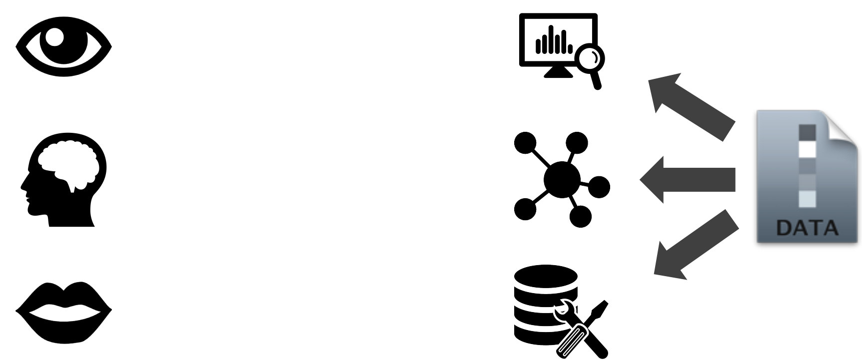


REWARDS	# TICKETS GIVEN	CONSEQUENCES	# TICKETS TAKEN AWAY
Extra Math	+5	HITTING	-3
Getting along WELL with others	+3	BULLYING	-4
Good Table Manners	+4	TEASING	-1
LOVE & RESPECT	+5	LYING	-2
Obeying the FIRST TIME	+3	THROWING A FIT	-3
Calm & Quiet in STORE	+3	Ignoring Parents	-4
Extra Reading	+2	SCREAMING or YELLING	-1
CLEANING up after PLAYING	+2	BAD SPORT	-2

Methodology: Why Data-Driven?

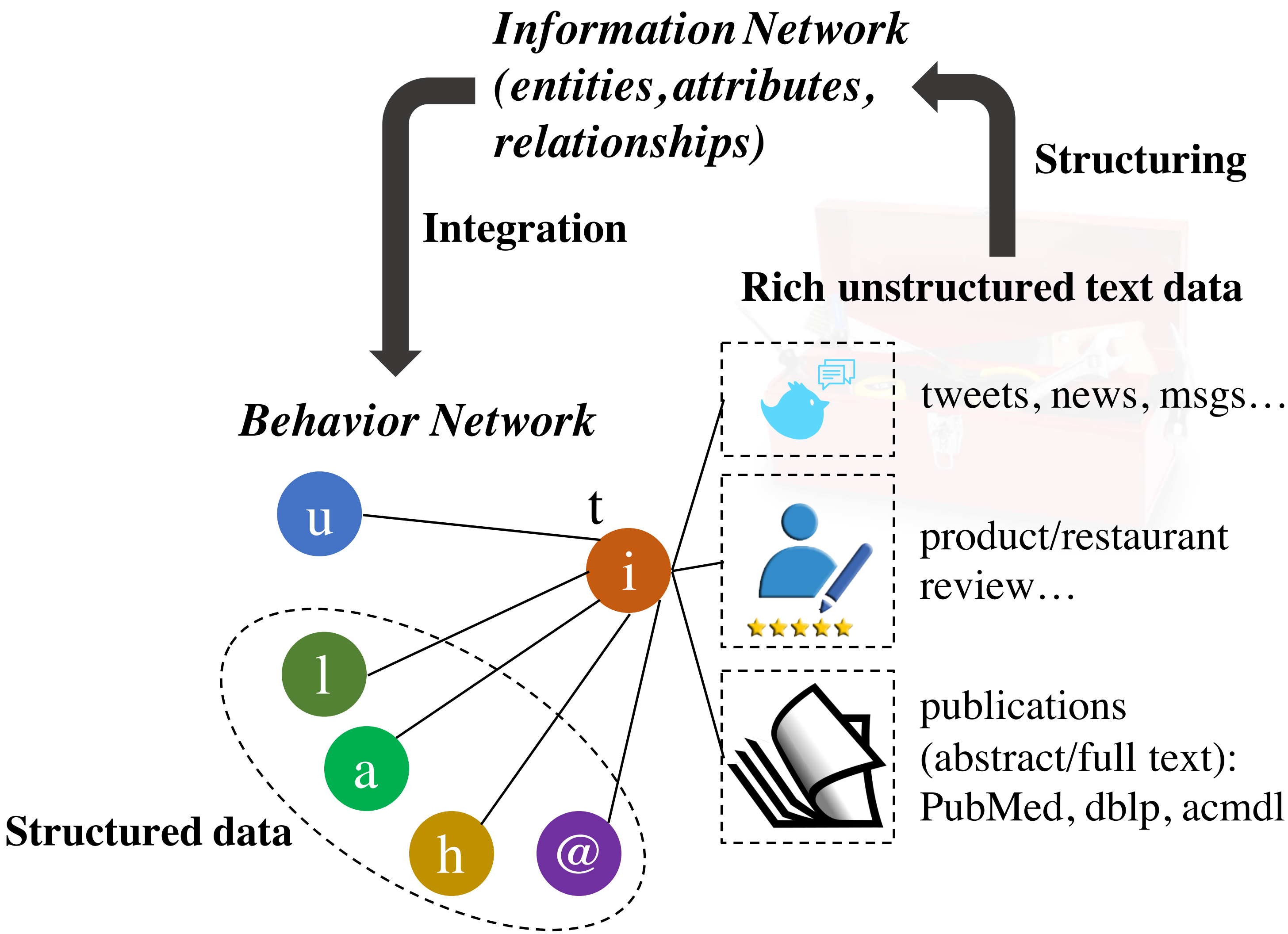
Experience-Driven **Data-Driven**

Observation → Representation → Models/Algorithms



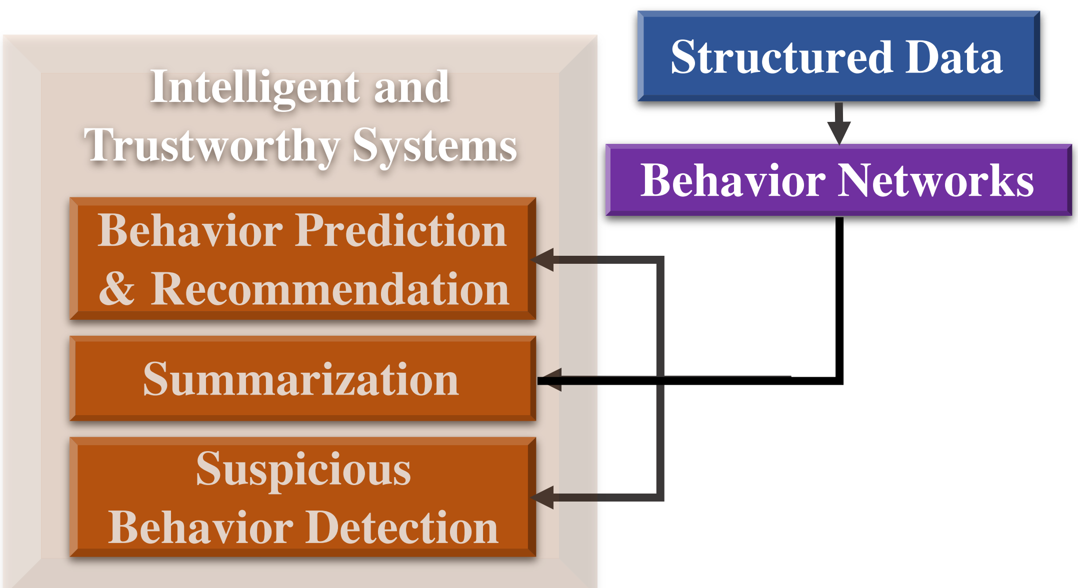
- *Applications*. Recommender systems, fraud/spam detection.
- *Representation*. **B**ehavior **N**etwork for interaction.
  - **Nodes:** users/authors, items (*e.g.*, products, tweets, papers), *etc*.
  - **Links:** (interaction) following, purchasing, tweeting, publishing, *etc*.
  - **Node attributes:** user profiles, item properties/features, *etc*.
  - **Link attributes:** similarity, distance, weight, *etc*.

Data to Network to Knowledge



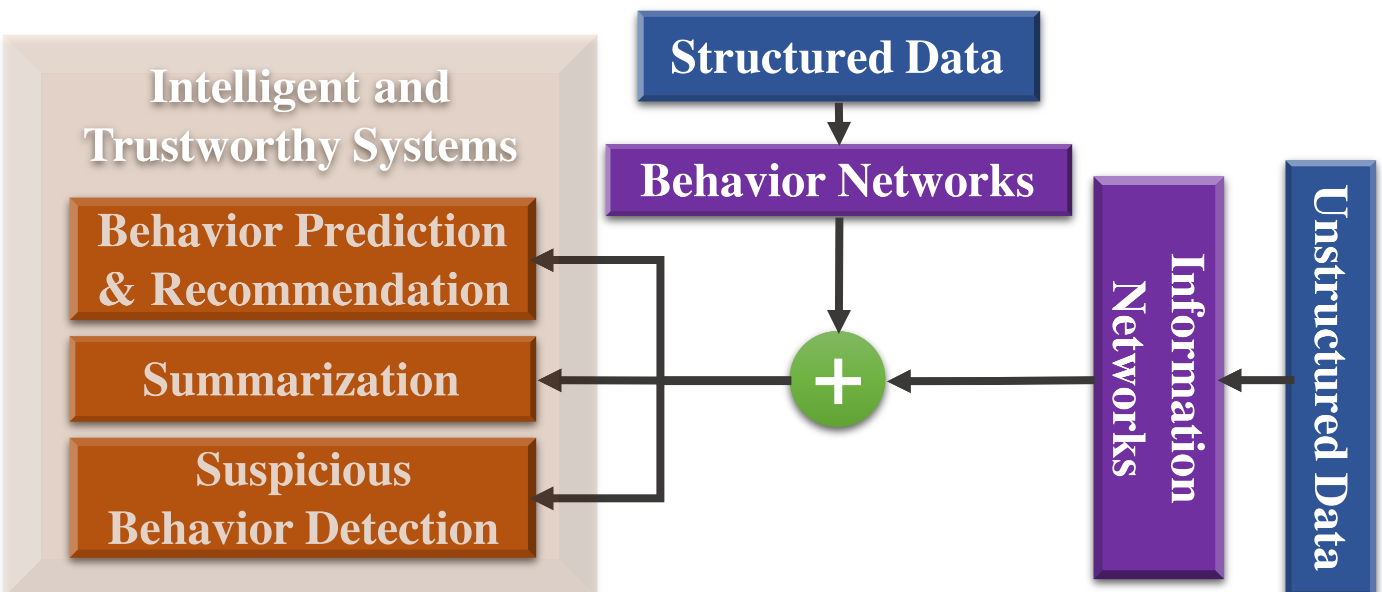
Outline: Data-Driven Behavioral Analytics

- Mining **behavior networks** with **social and spatiotemporal contexts** to support intelligent and trustworthy systems
  - Mining for **behavior prediction and recommendation**
  - Mining for **suspicious behavior detection**



Outline: Data-Driven Behavioral Analytics

- Mining **behavior networks** with **social and spatiotemporal contexts** to support intelligent and trustworthy systems
  - Mining for **behavior prediction and recommendation**
  - Mining for **suspicious behavior detection**
- Structuring behavioral content and integrating behavioral analysis with **information networks**

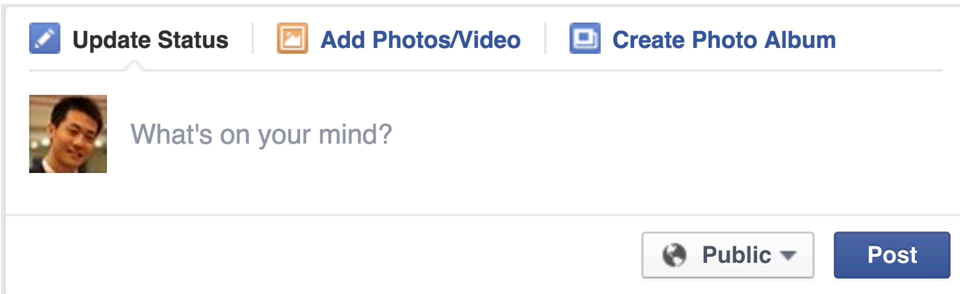


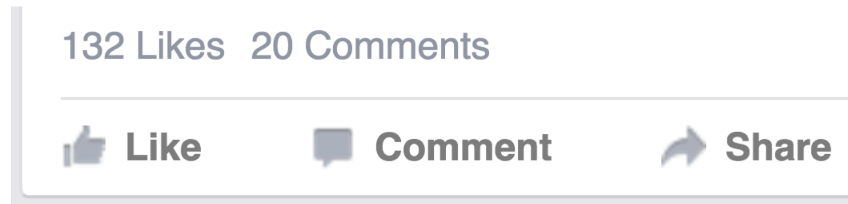
I. Mining behavior networks with social and spatiotemporal contexts

I.1. Behavior prediction and recommendation

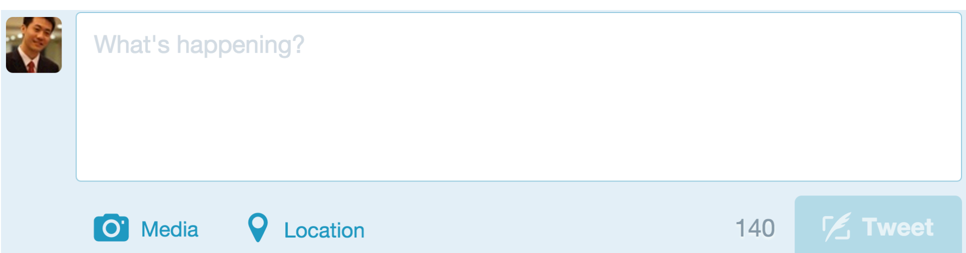
Behavior in Social Networks

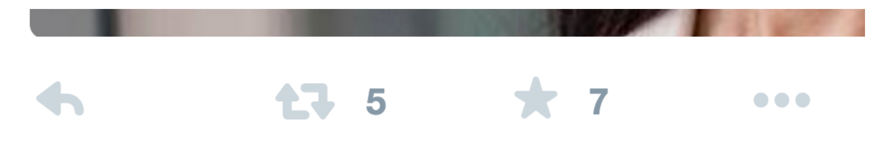
- **Facebook:** Post, Like, Comment, Share



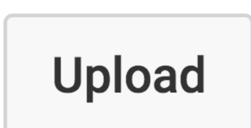


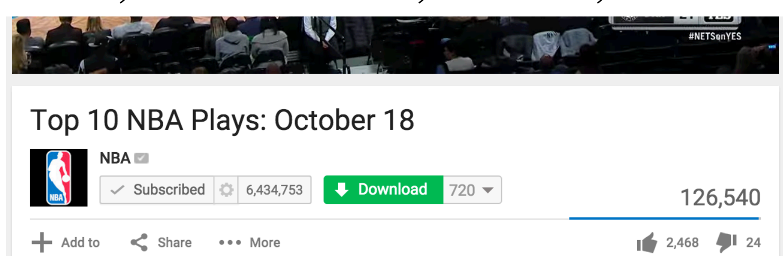
- **Twitter:** Post, Reply, Retweet, Favorite





- **YouTube:** Upload, Subscribe, Download, Share, Comment

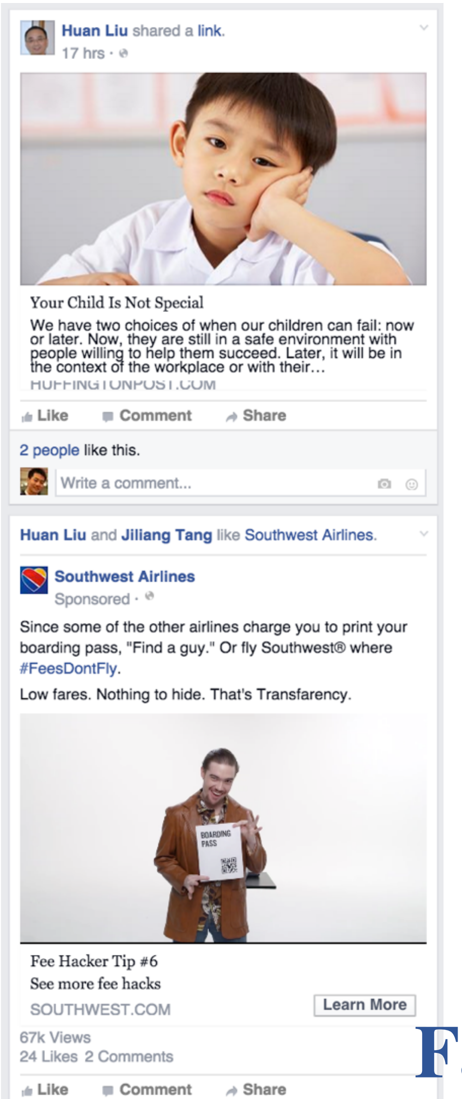




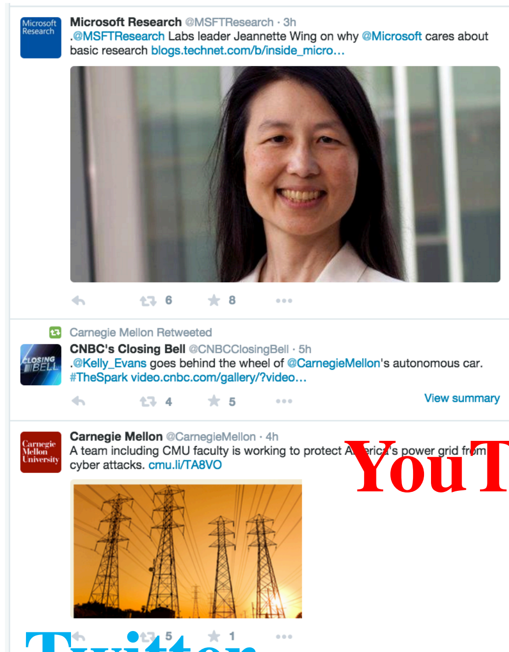
Behavior in Social Networks

Like
Reply
Share
Favorite
Retweet
Comment
Subscribe
Download
Add to
Send
Pin it
Visit
……

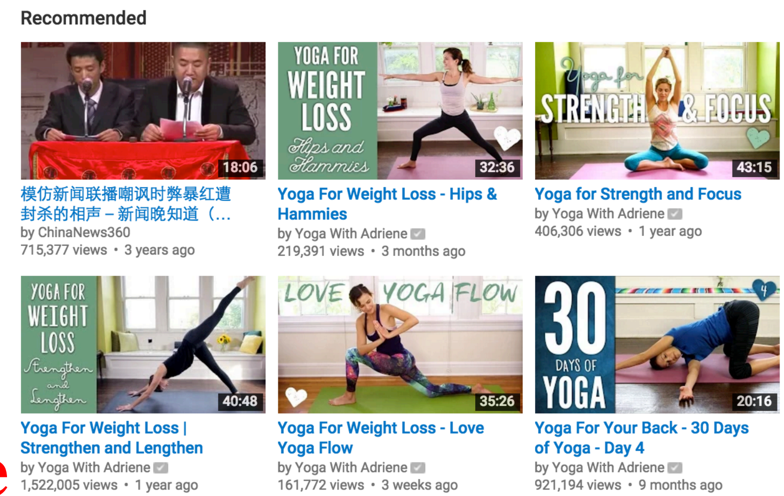
Social Recommender Systems



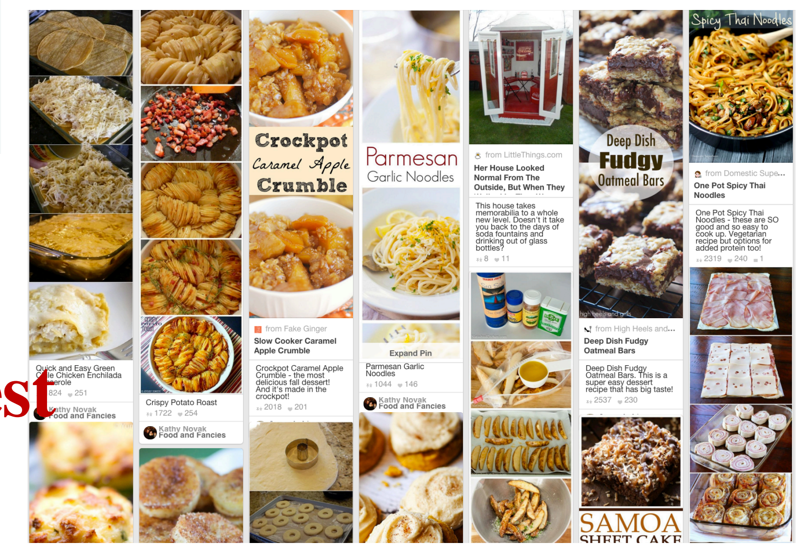
Facebook



Twitter



YouTube

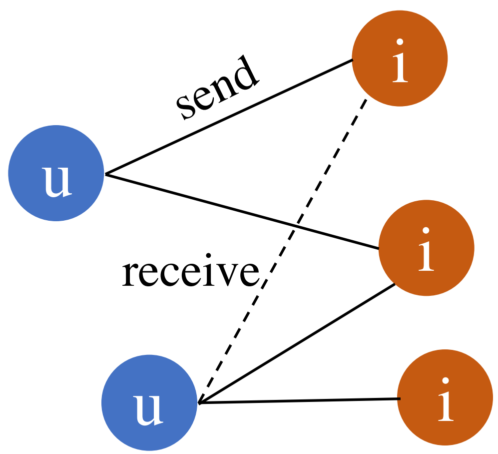


Pinterest

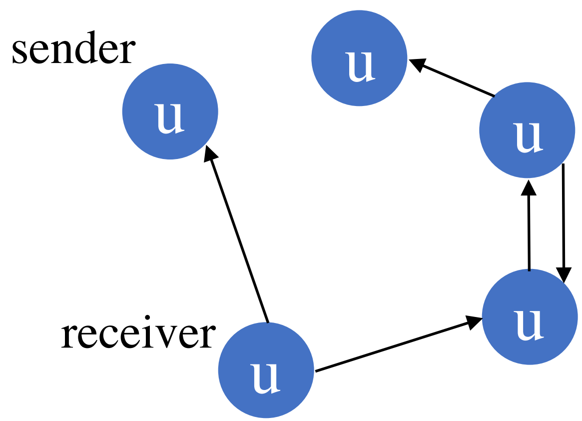
Social Recommender Systems

- April 20, 2011: **Tencent Weibo** visited Tsinghua University
  - Low *conversion rate* (< 6%): #retweets per feed request
  - Can we build a ***social recommender system***?
  - **Given**

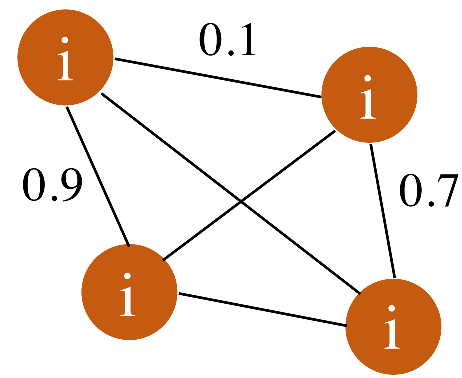
User-item behavior network



User-user social network



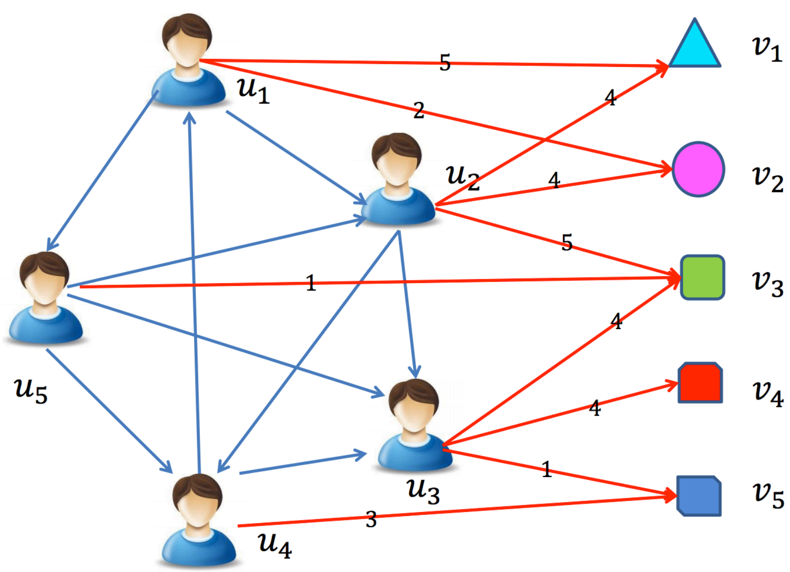
Content similarity (topic level) [Blei *et al.*]



- **Predict** which tweet/item a user will retweet.

Traditional Recommender Systems

- Assumed that users are independent and identically distributed (user-movie, user-book, *etc*.)



	v_1	v_2	v_3	v_4	v_5
u_1	5	?	2	?	?
u_2	4	4	5	?	?
u_3	?	?	4	4	1
u_4	?	?	?	?	3
u_5	?	?	1	?	?

Traditional Recommender Systems

- Content-based recommender (e.g., TF-IDF)
  - For textual information (e.g., news, documents)
  - *Limitation:* limited content analysis, over-specialization
- Collaborative filtering based recommender
  - Memory-based CF (e.g., PCC, similarity)
  - Model-based CF (e.g., factorization based)
  - *Limitation:* data sparsity, cold-start problem
- Hybrid recommender system

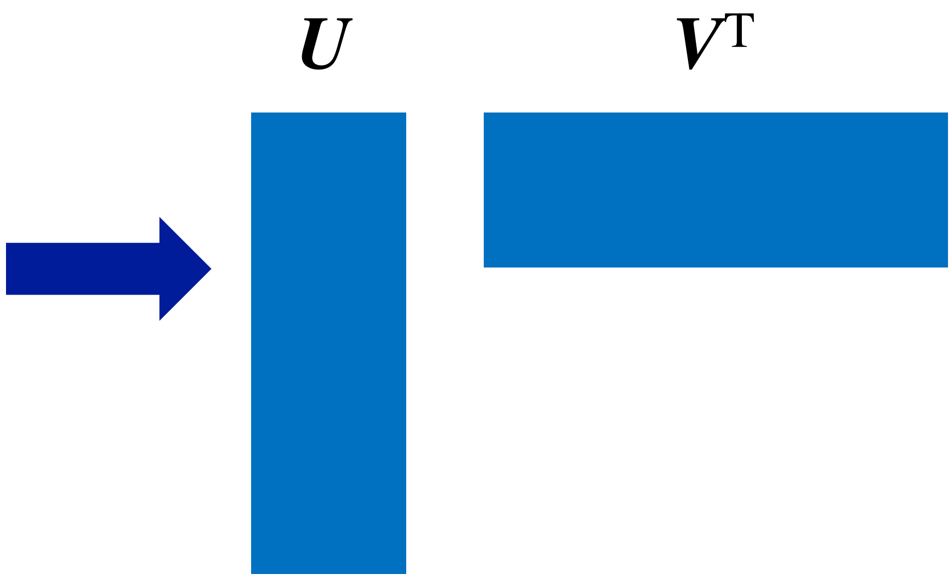
Tang et al. **Social Recommendation: A Review.** *Social Network Analysis and Mining*, 2013. Springer.

Matrix Factorization (MF) based CF

- Low-rank MF on the user-item rating matrix ***R***
- User preference vector ***U***
- Item characteristic vector ***V***

R

	v_1	v_2	v_3	v_4	v_5
u_1	5	?	2	?	?
u_2	4	4	5	?	?
u_3	?	?	4	4	1
u_4	?	?	?	?	3
u_5	?	?	1	?	?



Koren. **Factorizatoin Meets the Neighborhood: A Multifaceted Collaborative Filtering Model.** *KDD*, **2008**.

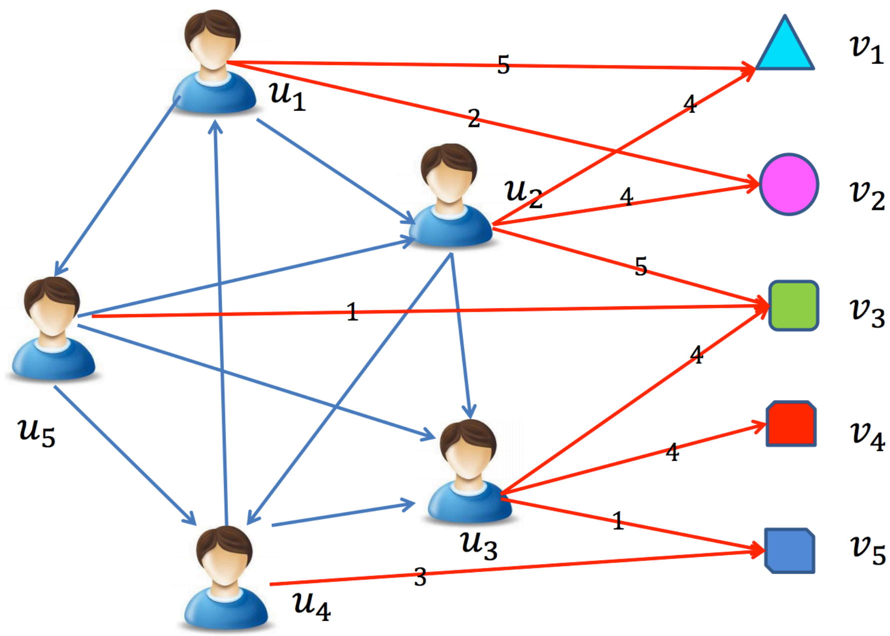
Matrix Factorization (MF) based CF

- Low-rank MF on the user-item rating matrix ***R***
- User preference vector ***U***
- Item characteristic vector ***V***
- Observed weight matrix ***W***

$$\min_{\mathbf{U},\mathbf{V}} \sum_{i=1}^{n} \sum_{j=1}^{m} \boxed{\mathbf{W}_{ij}} (\mathbf{R}_{ij} - \mathbf{U}_i \mathbf{V}_j^\top)^2 + \boxed{\alpha(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2)}$$

avoid over-fitting,
controlled by the parameter

Social Recommendation



Social relations

	u_1	u_2	u_3	u_4	u_5
u_1	0	1	0	0	1
u_2	0	0	1	1	0
u_3	0	0	0	0	0
u_4	1	0	1	0	0
u_5	0	1	1	1	0

	v_1	v_2	v_3	v_4	v_5
u_1	5	?	2	?	?
u_2	4	4	5	?	?
u_3	?	?	4	4	1
u_4	?	?	?	?	3
u_5	?	?	1	?	?

Memory based Social Recommender

- TidalTrust

rating
(user i, item m)

$$r_{sm} = \frac{\sum_{i \in S} t_{si} r_{im}}{\sum_{i \in S} t_{si}}$$

rating
(user s, item m)

trust from social relation
(user s, user i)

Golbeck. **Personalizing applications through integration of inferred trust values in semantic web-based social networks.** *Semantic Network Analysis Workshop*, 2005.

Memory based Social Recommender

- MoleTrust

rating (user u, item i)

average rating (user u)

average rating (user a)

$$p_{a,i} = \overline{r}_a + \frac{\sum_{u=1}^{k} w_{a,u}(r_{u,i} - \overline{r}_u)}{\sum_{u=1}^{k} w_{a,u}}$$

predicted rating (user a, item i)

trust from social relation (user a, user u)

Massa et al. **Trust-aware recommender systems.** *RecSys*, 2007.

Memory based Social Recommender

- **TrustWalker**

similarity measure (item i, item j)

$$P(Y_{u,i} = j) = \frac{sim(i,j)}{\sum_{l \in RI_u} sim(i,l)}$$

probability of user u's random walk from item i to item j

Pearson correlation of (item i, item j)

$$sim(i,j) = \frac{1}{1 + e^{-\frac{|UC_{i,j}|}{2}}} \times corr(i,j)$$

common user set of (item i, item j)

Jamali et al. **TrustWalker: A Random Walk Model for Combining Trust-based and Item-based Recommendation.** *KDD*, 2009.

Model based Social Recommender

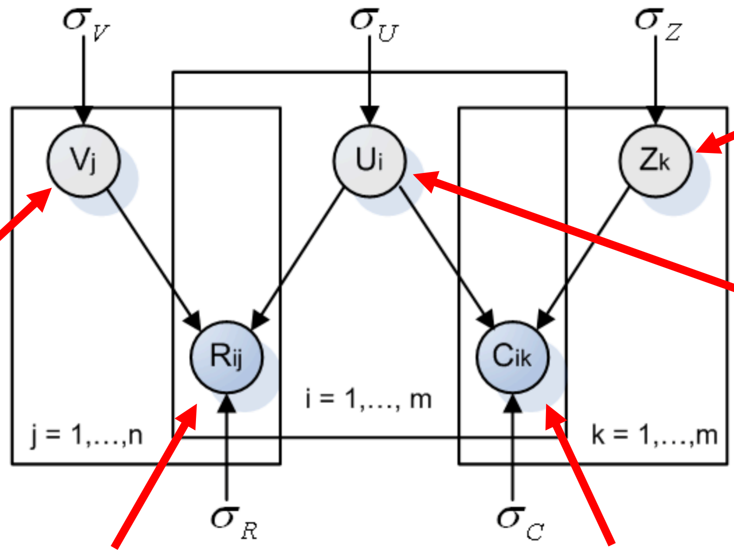
- Optimization methods such as gradient based methods can be applied to find a well-worked optimal solution.
- MF has a nice probabilistic interpretation with Gaussian noise.
- MF is very flexible and allows us to include prior knowledge.

$$\begin{aligned} & \textit{Social Recommendation CF} \\ = \ & \textit{Basic CF} + \textit{Social Information Model} \end{aligned}$$

Tang et al. **Social Recommendation: A Review.** *Social Network Analysis and Mining*, 2013. Springer.

Model based Social Recommender

- **SoRec**



factor vector

item vector

user vector

***R*:** user-item rating matrix

	v_1	v_2	v_3	v_4	v_5
u_1	5	?	2	?	?
u_2	4	4	5	?	?
u_3	?	?	4	4	1
u_4	?	?	?	?	3
u_5	?	?	1	?	?

***C*:** user-user social matrix

	u_1	u_2	u_3	u_4	u_5
u_1	0	1	0	0	1
u_2	0	0	1	1	0
u_3	0	0	0	0	0
u_4	1	0	1	0	0
u_5	0	1	1	1	0

Ma et al. **Social Recommendation Using Probabilistic Matrix Factorization.** *CIKM*, 2008.
Improving Recommender Systems by Incorporating Social Contextual Information. *TIS*, 2011.

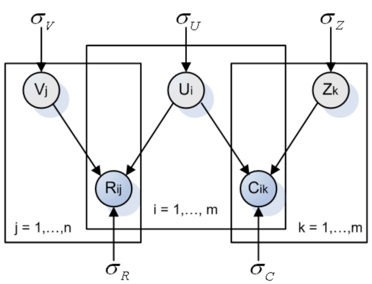
Model based Social Recommender

- SoRec

$$p(\cancel{C}|U,V,\sigma_R^2)=\prod_{i=1}^{m}\prod_{j=1}^{n}\underline{\mathcal{N}}\left[\left(r_{ij}|g(U_i^TV_j),\sigma_R^2\right)\right]^{I_{ij}^R}$$

R

Gaussian distribution



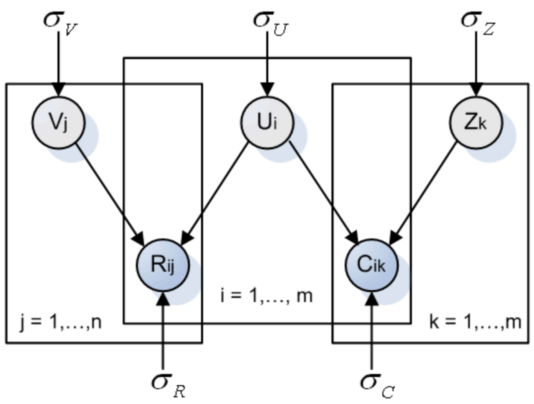
Logistic function

Observed

$$p(C|U,Z,\sigma_C^2)=\prod_{i=1}^{m}\prod_{k=1}^{m}\mathcal{N}\left[\left(c_{ik}|\underline{g(U_i^TZ_k)},\sigma_C^2\right)\right]^{\underline{I_{ik}^C}}$$

Model based Social Recommender

- **SoRec**



behavioral term

social term

$$\mathcal{L}(R, C, U, V, Z) =$$

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}^{R}\underline{(r_{ij} - g(U_i^T V_j))^2} + \frac{\lambda_C}{2}\sum_{i=1}^{m}\sum_{k=1}^{m} I_{ik}^{C}\underline{(c_{ik}^{*} - g(U_i^T Z_k))^2}$$

$$\underline{+\frac{\lambda_U}{2}\|U\|_F^2 + \frac{\lambda_V}{2}\|V\|_F^2 + \frac{\lambda_Z}{2}\|Z\|_F^2}, \quad (9)$$

regularization terms

Model based Social Recommender

- **SoRec**

Gradient Descent Methods

deviate of Logistic function

$$\frac{\partial \mathcal{L}}{\partial U_i} = \sum_{j=1}^{n} I_{ij}^{R} \underline{g'(U_i^T V_j)}(g(U_i^T V_j) - r_{ij})V_j$$

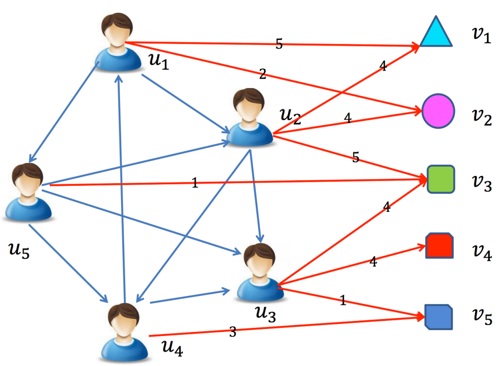
$$+ \lambda_C \sum_{j=1}^{m} I_{ik}^{C} \underline{g'(U_i^T Z_k)}(g(U_i^T Z_k) - c_{ik}^{*})Z_k + \lambda_U U_i,$$

$$\frac{\partial \mathcal{L}}{\partial V_j} = \sum_{i=1}^{m} I_{ij}^{R} \underline{g'(U_i^T V_j)}(g(U_i^T V_j) - r_{ij})U_i + \lambda_V V_j,$$

$$\frac{\partial \mathcal{L}}{\partial Z_k} = \lambda_C \sum_{i=1}^{m} I_{ik}^{C} \underline{g'(U_i^T Z_k)}(g(U_i^T Z_k) - c_{ik}^{*})U_i + \lambda_Z Z_k, (10)$$

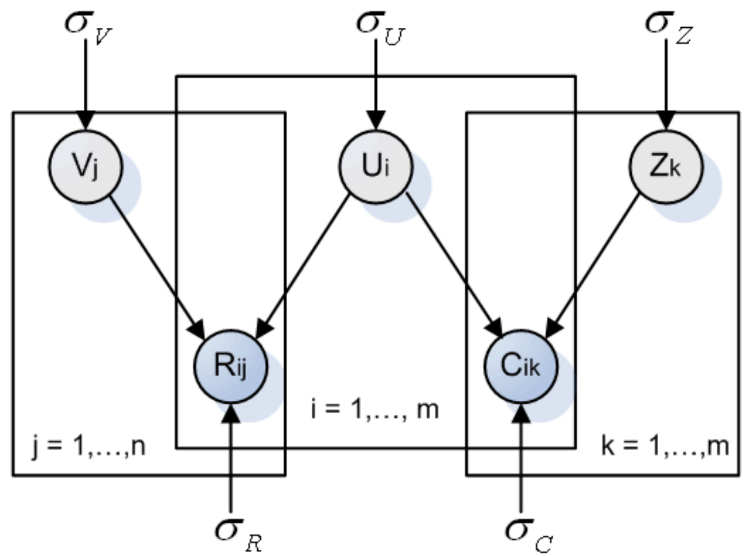
Model based Social Recommender

❑SoRec



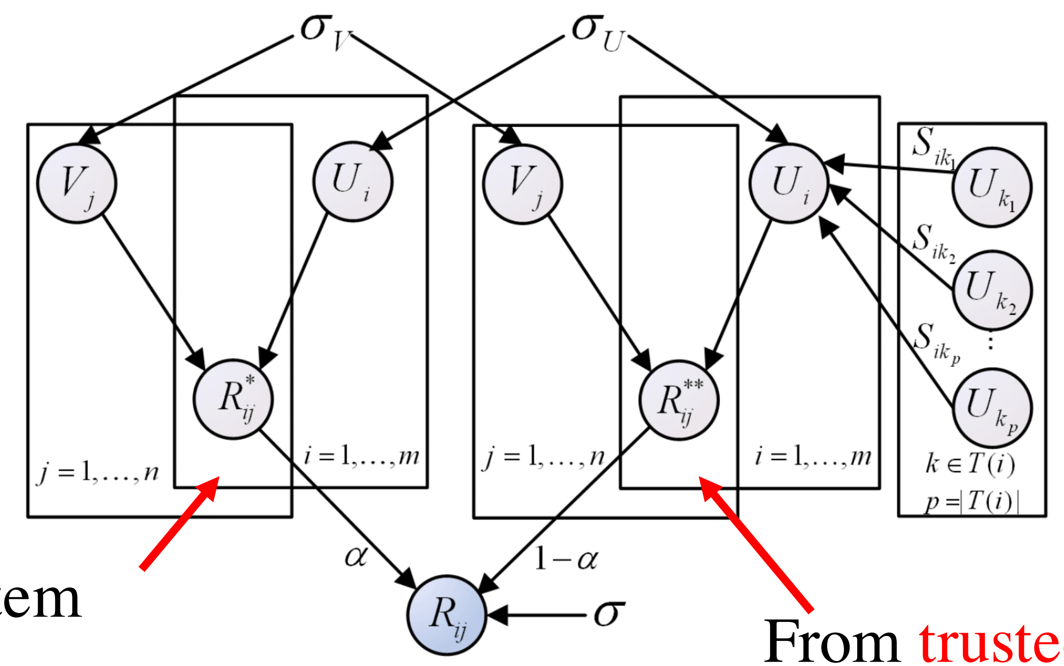
	u_1	u_2	u_3	u_4	u_5
u_1	0	1	0	0	1
u_2	0	0	1	1	0
u_3	0	0	0	0	0
u_4	1	0	1	0	0
u_5	0	1	1	1	0

	v_1	v_2	v_3	v_4	v_5
u_1	5	?	2	?	?
u_2	4	4	5	?	?
u_3	?	?	4	4	1
u_4	?	?	?	?	3
u_5	?	?	1	?	?



Model based Social Recommender

- Replacing social with trust
- **"Social Trust" Ensemble** for Epinion data



From user-item matrix

Operation "plus"

From trusted friends

Ma et al. **Learning to Recommend with Social Trust Ensemble.** *SIGIR*, 2009.
Learning to Recommend with Explicit and Implicit Social Relations. *TIST*, 2011.

Model based Social Recommender

- "Social Trust" Ensemble

From user-item matrix

From trusted friends

$$\mathcal{L}(R,S,U,V) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}^{R}(R_{ij} - g(\underline{\alpha U_i^T V_j} + \underline{(1-\alpha)\sum_{k\in\mathcal{T}(i)} S_{ik}U_k^T V_j}))^2 + \frac{\lambda_U}{2}\|U\|_F^2 + \frac{\lambda_V}{2}\|V\|_F^2, \quad (13)$$

Model based Social Recommender

- "Social Trust" Ensemble

Gradient Descent Methods

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial U_i} = & \alpha \sum_{j=1}^{n} I_{ij}^R g'(\alpha U_i^T V_j + (1-\alpha) \sum_{k \in \mathcal{T}(i)} S_{ik} U_k^T V_j) V_j \\
& \times (g(\alpha U_i^T V_j + (1-\alpha) \sum_{k \in \mathcal{T}(i)} S_{ik} U_k^T V_j) - R_{ij}) \\
& + (1-\alpha) \sum_{p \in \mathcal{B}(i)} \sum_{j=1}^{n} I_{pj}^R g'(\alpha U_p^T V_j + (1-\alpha) \sum_{k \in \mathcal{T}(p)} S_{pk} U_k^T V_j) \\
& \times (g(\alpha U_p^T V_j + (1-\alpha) \sum_{k \in \mathcal{T}(p)} S_{pk} U_k^T V_j) - R_{pj}) S_{pi} V_j + \lambda_U U_i, \\
\frac{\partial \mathcal{L}}{\partial V_j} = & \sum_{i=1}^{m} I_{ij}^R g'(\alpha U_i^T V_j + (1-\alpha) \sum_{k \in \mathcal{T}(i)} S_{ik} U_k^T V_j) \\
& \times (g(\alpha U_i^T V_j + (1-\alpha) \sum_{k \in \mathcal{T}(i)} S_{ik} U_k^T V_j) - R_{ij}) \\
& \times (\alpha U_i + (1-\alpha) \sum_{k \in \mathcal{T}(i)} S_{ik} U_k^T) + \lambda_V V_j, \qquad (14)
\end{aligned}
$$

Model based Social Recommender

- **SoReg**

Average-based regularization:

Regularize with the average of friends' tastes

$$\min_{U,V} \mathcal{L}_1(R,U,V) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}(R_{ij} - U_i^T V_j)^2$$

$$\boxed{+ \frac{\alpha}{2}\sum_{i=1}^{m} \|U_i - \frac{\sum_{f\in\mathcal{F}^+(i)} Sim(i,f)\times U_f}{\sum_{f\in\mathcal{F}^+(i)} Sim(i,f)}\|_F^2,}$$

$$+ \frac{\lambda_1}{2}\|U\|_F^2 + \frac{\lambda_2}{2}\|V\|_F^2. \tag{8}$$

Information loss: Friends may have diverse tastes!!!

Ma et al. **Recommender Systems with Social Regularization.** *WSDM*, 2011.

Model based Social Recommender

- SoReg

Individual-based regularization:
Regularize with friends individually

$$
\begin{aligned}
\min_{U,V} \mathcal{L}_2(R,U,V) &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}(R_{ij} - U_i^T V_j)^2 \\
&+ \boxed{\frac{\beta}{2}\sum_{i=1}^{m}\sum_{f\in\mathcal{F}^+(i)} Sim(i,f)\|U_i - U_f\|_F^2} \\
&+ \lambda_1\|U\|_F^2 + \lambda_2\|V\|_F^2. \qquad (11)
\end{aligned}
$$

Related Work

	Behavior	Content	Social	Trust
Collaborative filtering (CF) [Herlocker *et al*. TOIS; Koren KDD]	✔			
Content-based filtering with CF [Balabanovic *et al*.; Liu *et al*. CIKM;]	✔	✔		
SoRec [Ma *et al*. CIKM, TIS] SoReg [Ma *et al*. WSDM]	✔		✔	
Trust-based methods [Massa *et al*. RecSys; Jamali *et al*. KDD; Ma *et al*. SIGIR, TIST]	✔			✔

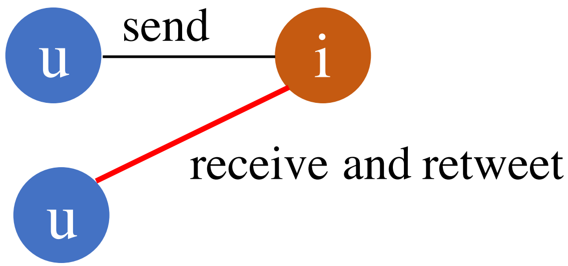
- **Q:** What are the **factors** of users' decisions on retweeting? Can we **observe** them from the data? How to **integrate** the information for accurate prediction?

Observation: Social Contextual Factors

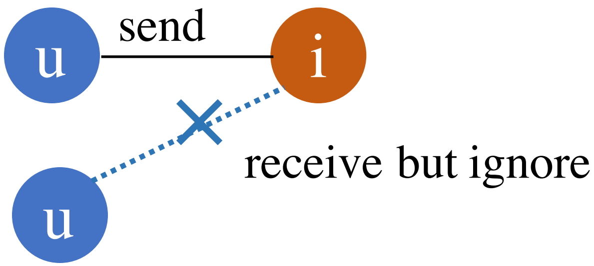
- Will Michelle Obama share this message?
- Please list your reasons.

Jiang et al. **Social Contextual Recommendation.** *CIKM*, 2012.
Social Recommendation with Contextual Information. *TKDE*, 2014.

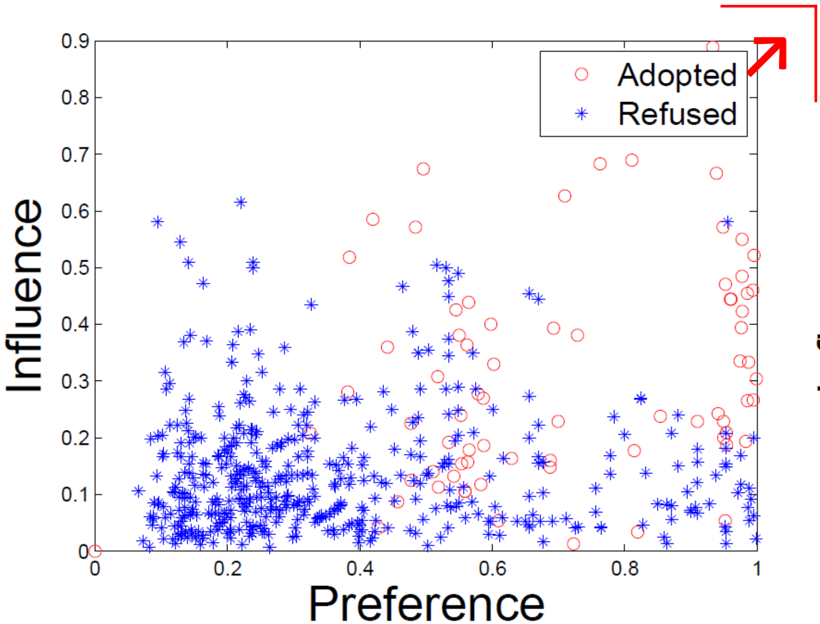
Observation: Social Contextual Factors



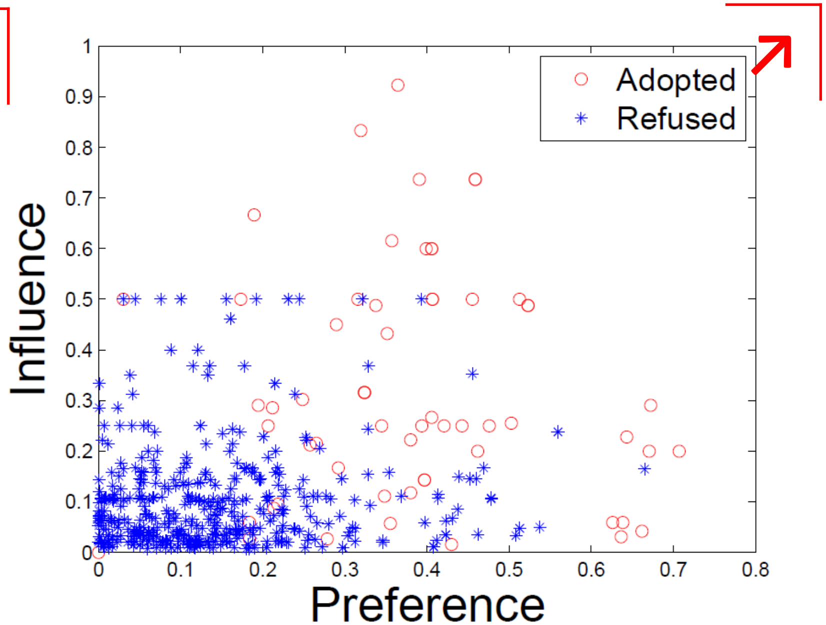
adopted



refused

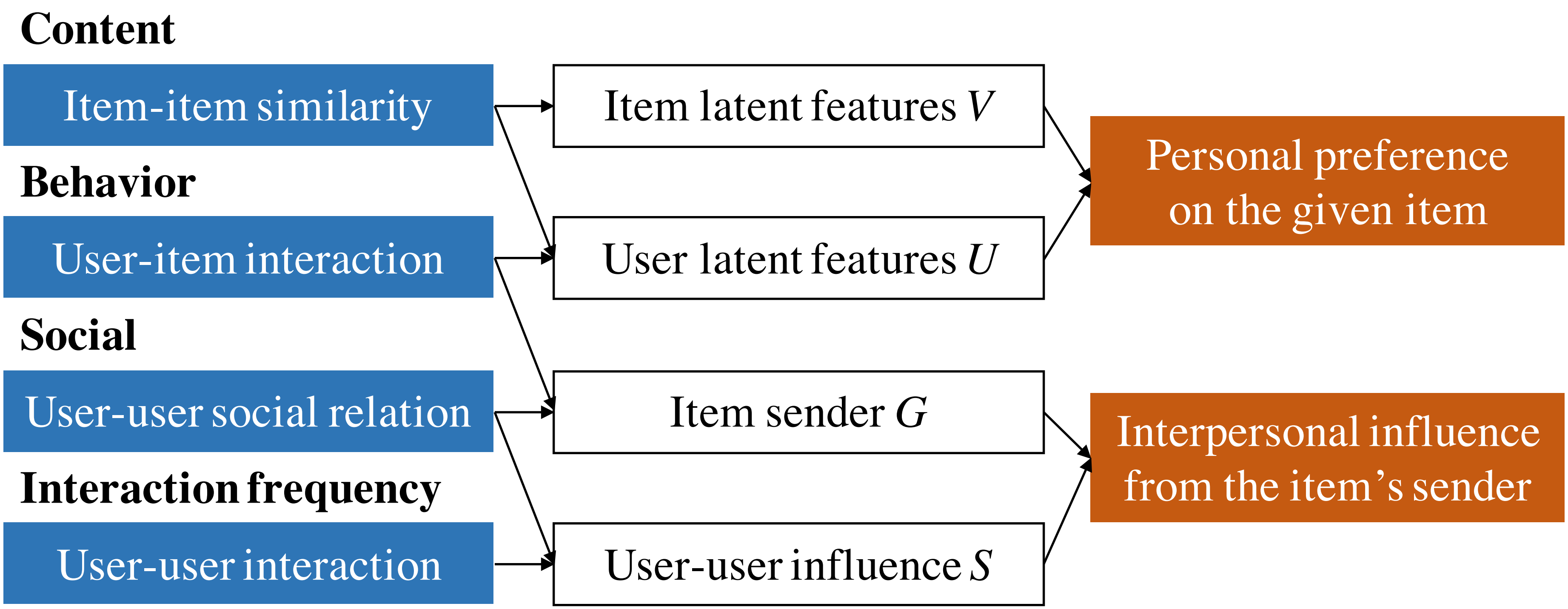


China's Facebook: Renren

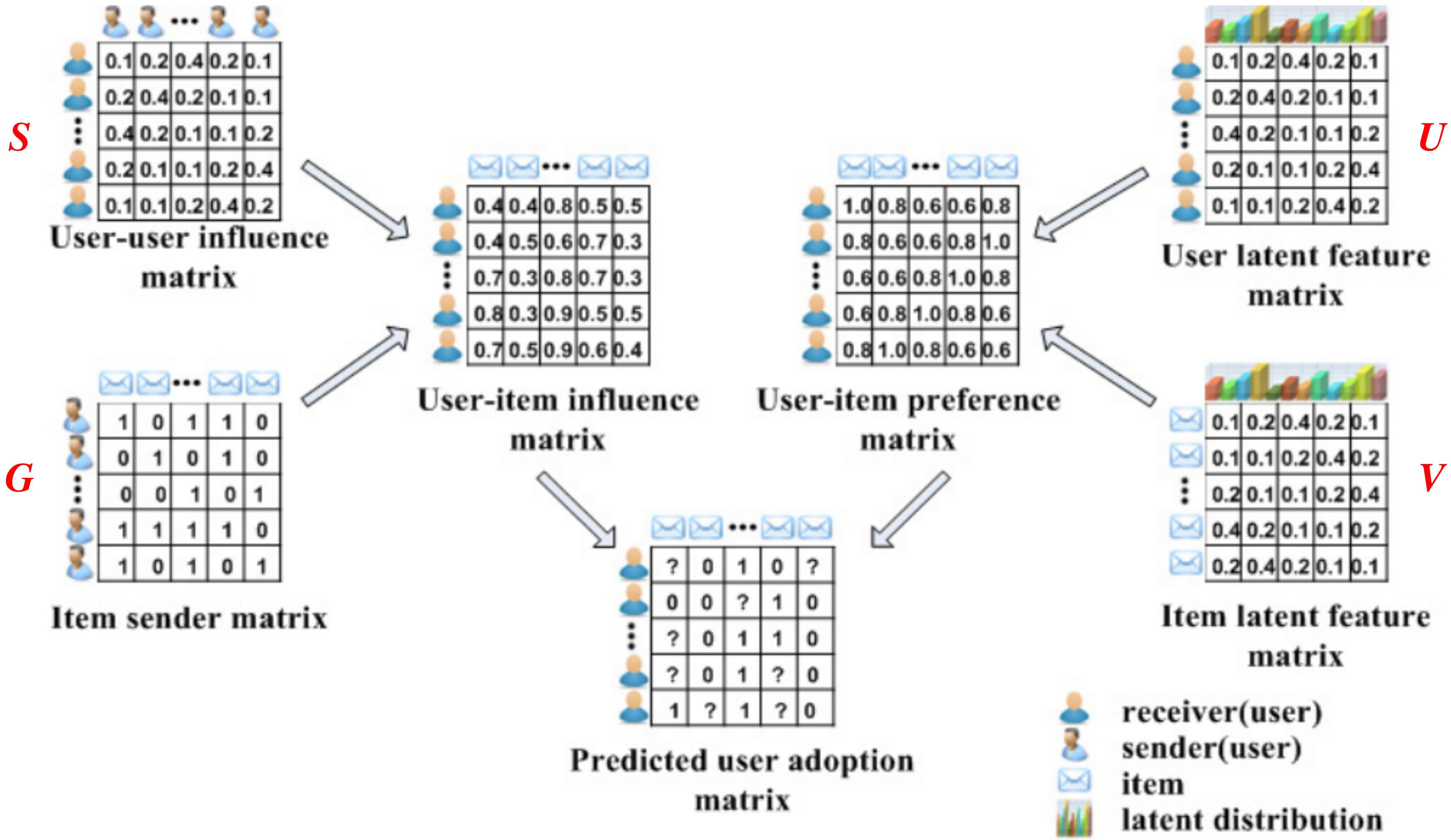


China's Twitter: Tencent Weibo

Representation: From Contextual Information to Contextual Factors



Model: ContextMF



Model: ContextMF

behavior influence preference

$$P(\mathbf{R}|\mathbf{S},\mathbf{U},\mathbf{V},\sigma_R^2) = \prod_{i=1}^{M}\prod_{j=1}^{N}\mathcal{N}(\mathbf{R}_{ij}|\mathbf{S}_i\mathbf{G}_j^\top \odot \mathbf{U}_i^\top\mathbf{V}_j,\sigma_R^2)$$

behavior interaction frequency/trust

item content

$$\begin{aligned}\mathcal{J} = &\, ||\mathbf{R} - \mathbf{S}\mathbf{G}^\top \odot \mathbf{U}^\top\mathbf{V}||_F^2 + \alpha||\mathbf{W} - \mathbf{U}^\top\mathbf{U}||_F^2 \\ &+\beta||\mathbf{C} - \mathbf{V}^\top\mathbf{V}||_F^2 + \gamma||\mathbf{S} - \mathbf{F}||_F^2 \\ &+\delta||\mathbf{S}||_F^2 + \eta||\mathbf{U}||_F^2 + \lambda||\mathbf{V}||_F^2\end{aligned}$$

social relation

Model: ContextMF

- Gradient descent method

$$
\begin{aligned}
\frac{\partial \mathcal{J}}{\partial \mathbf{S}} &= 2\Big(-\mathbf{R}(\mathbf{G}\odot\mathbf{V}^\top\mathbf{U}) + (\mathbf{S}\mathbf{G}^\top\odot\mathbf{U}^\top\mathbf{V})\mathbf{G} \\
&\quad +\gamma(\mathbf{S}-\mathbf{F}) + \delta\mathbf{S}\Big) \\
\frac{\partial \mathcal{J}}{\partial \mathbf{U}} &= 2\Big(-\mathbf{V}\mathbf{R}^\top + \mathbf{V}(\mathbf{G}\mathbf{S}^\top\odot\mathbf{V}^\top\mathbf{U}) - 2\alpha\mathbf{U}\mathbf{W} \\
&\quad +2\alpha\mathbf{U}\mathbf{U}^\top\mathbf{U} + \eta\mathbf{U}\Big) \\
\frac{\partial \mathcal{J}}{\partial \mathbf{V}} &= 2\Big(-\mathbf{U}\mathbf{R} + \mathbf{U}(\mathbf{S}\mathbf{G}^\top\odot\mathbf{U}^\top\mathbf{V}) - 2\beta\mathbf{V}\mathbf{C} \\
&\quad +2\beta\mathbf{V}\mathbf{V}^\top\mathbf{V} + \lambda\mathbf{V}\Big)
\end{aligned}
$$

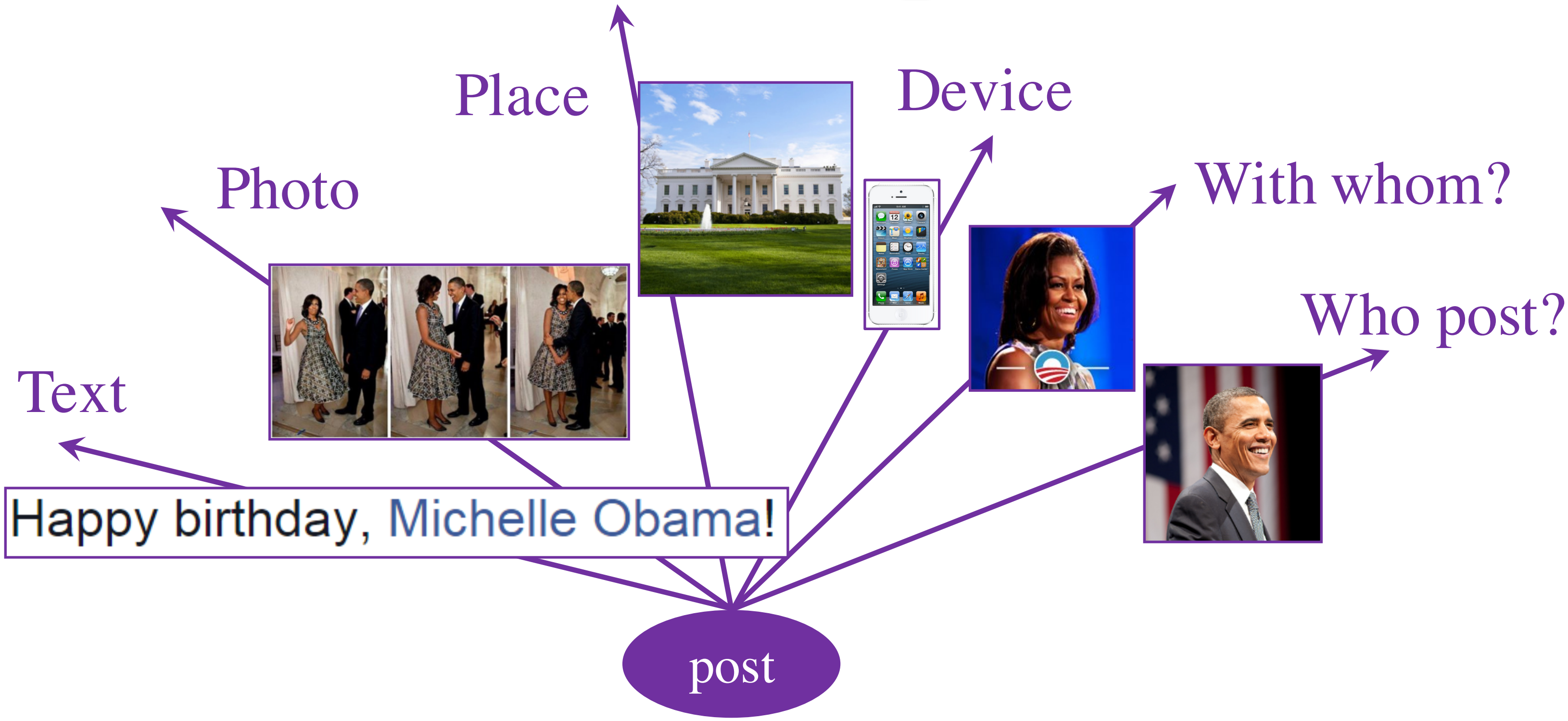
Experimental Results

Method	MAE	RMSE	$\hat{\tau}$	$\hat{\rho}$
Renren Dataset				
Content-based [1]	0.3842	0.4769	0.5409	0.5404
Item CF [25]	0.3601	0.4513	0.5896	0.5988
FeedbackTrust [22]	0.3764	0.4684	0.5433	0.5469
Influence-based [9]	0.3859	0.4686	0.5394	0.5446
SoRec [19]	0.3276	0.4127	0.6168	0.6204
SoReg [20]	0.2985	0.3537	0.7086	0.7140
Influence MF	0.3102	0.3771	0.6861	0.7006
Preference MF	0.3032	0.3762	0.6937	0.7036
Context MF	**0.2416**	**0.3086**	**0.7782**	**0.7896**
Tencent Weibo Dataset				
Content-based [1]	0.2576	0.3643	0.7728	0.7777
Item CF [25]	0.2375	0.3372	0.7867	0.8049
FeedbackTrust [22]	0.2830	0.3887	0.7094	0.7115
Influence-based [9]	0.2651	0.3813	0.7163	0.7275
SoRec [19]	0.2256	0.3325	0.7973	0.8064
SoReg [20]	0.1997	0.2962	0.8390	0.8423
Influence MF	0.2183	0.3206	0.8179	0.8258
Preference MF	0.2111	0.3088	0.8384	0.8453
Context MF	**0.1514**	**0.2348**	**0.8570**	**0.8685**

vs. SoReg [TIST'11]	Renren	Tencent Weibo
MAE	↓19.1%	↓24.2%
RMSE	↓12.8%	↓20.7%
Kendall's	↑9.82%	↑2.1%
Spearman's	↑10.6%	↑3.1%

- **Deployed** in Weibo News Feed. Improved conversion rate from 5.78% to 8.27% (relatively **43%**).
- #citations = **149**

Observation: Spatial Context



Jan. 18
Birthday party
@ White house

Jiang et al. **Flexible Evolutionary Multi-faceted Analysis for Dynamic Behavioral Pattern Discovery.** *KDD*, 2014.

Observation: Temporal Context

Photo

Text

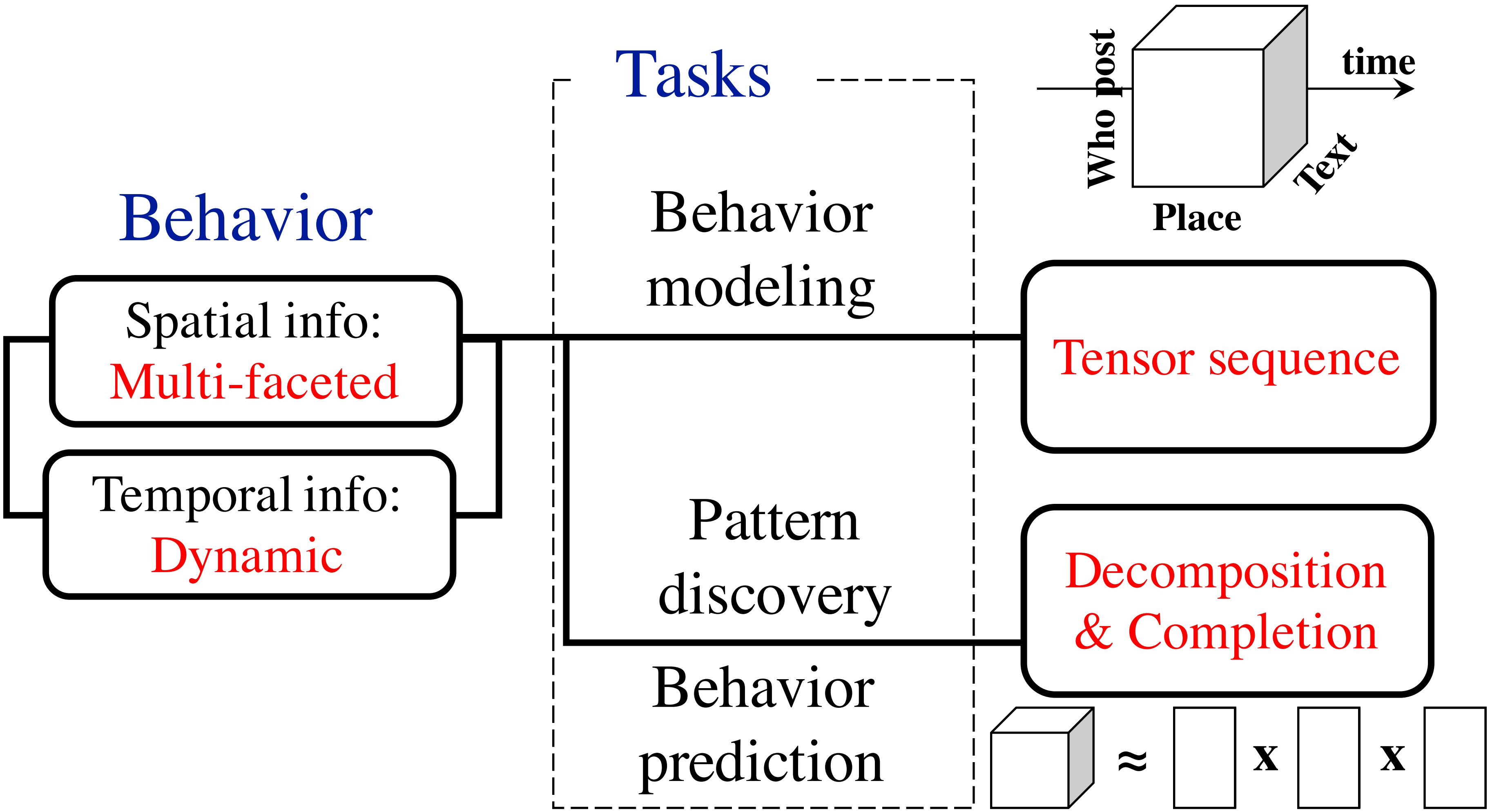
Happy birthday, Michelle Obama!

post → post → time

Jan. 18
Birthday party
@ White house

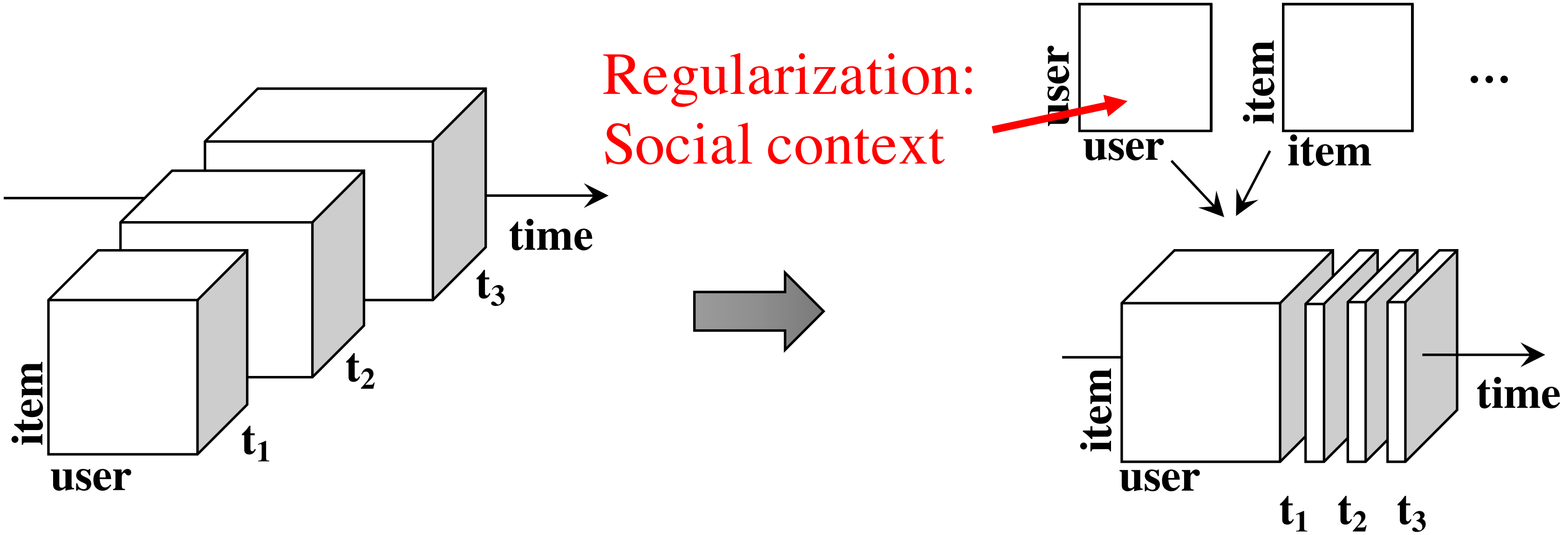
March 17
St. Patrick's Day
@ Bar

Representation: Tensor Sequence



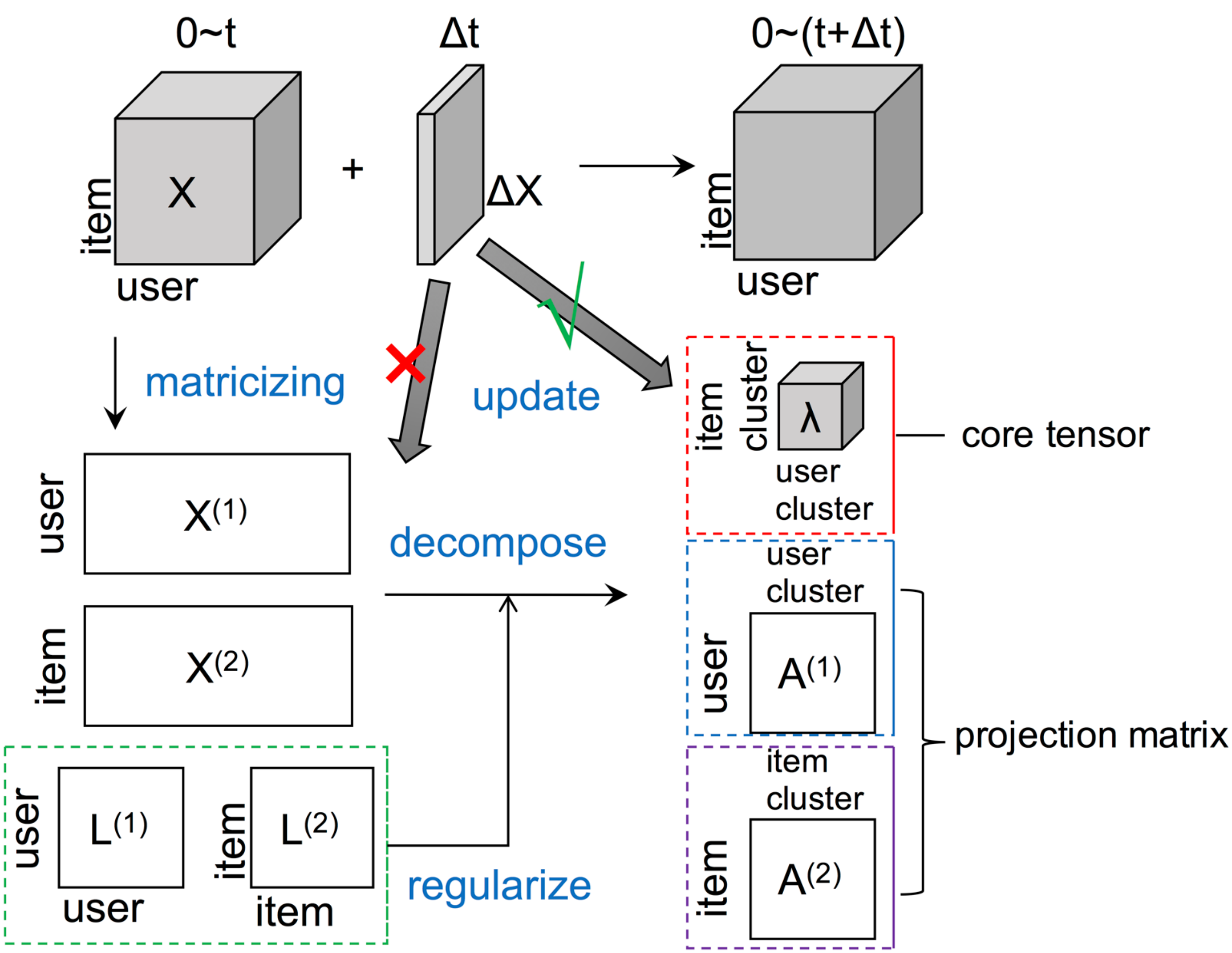
Challenges: Sparsity and Complexity

- Addressing sparsity: *Flexible regularization with auxiliary data*
- Addressing high complexity: *Incremental updates for projection matrix*



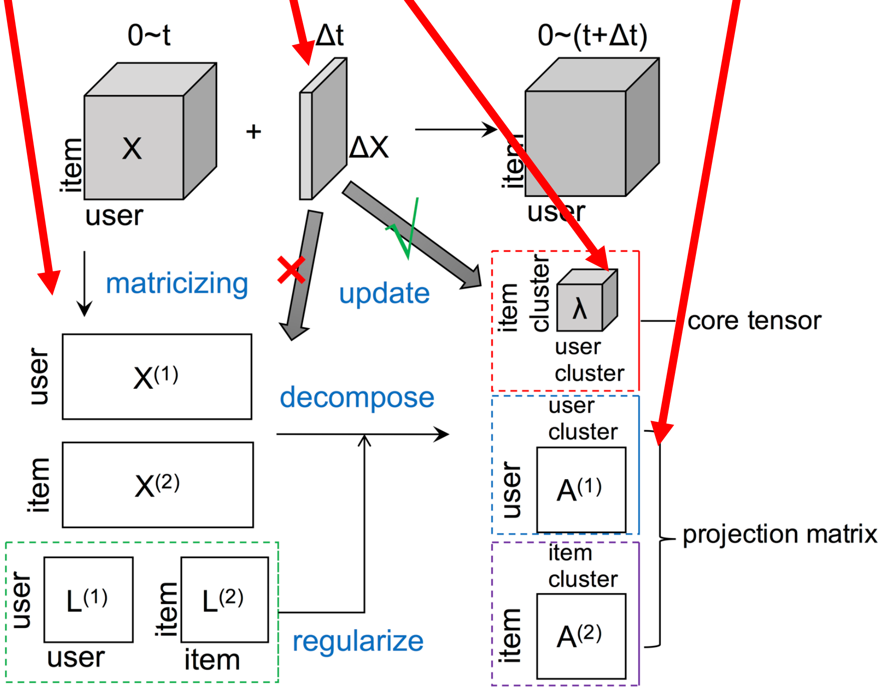
Model: FEMA

Flexible Evolutionary Multi-faceted Analysis



Tensor Perturbation Theory

$$[(\mathbf{X}^{(\mathbf{m})} + \Delta\mathbf{X}^{(\mathbf{m})})(\mathbf{X}^{(\mathbf{m})} + \Delta\mathbf{X}^{(\mathbf{m})})^\top + \mu^{(m)}\mathbf{L}^{(\mathbf{m})}]$$
$$\cdot(\mathbf{a}_\mathbf{i}^{(\mathbf{m})} + \Delta\mathbf{a}_\mathbf{i}^{(\mathbf{m})}) = (\lambda_i^{(m)} + \Delta\lambda_i^{(m)})(\mathbf{a}_\mathbf{i}^{(\mathbf{m})} + \Delta\mathbf{a}_\mathbf{i}^{(\mathbf{m})})$$



Algorithm: FEMA

Approximation

Require: $\mathcal{X}_t, \Delta\mathcal{X}_t, \mathbf{A}_t^{(m)}|_{m=1}^M, \lambda_t^{(m)}|_{m=1}^M$
for $m = 1, \ldots, M$ **do**
for $i = 1, \ldots, r^{(m)}$ **do**
Compute $\Delta\lambda_{t,i}^{(m)}$ using

$$\boxed{\Delta\lambda_i^{(m)} = \mathbf{a}_i^{(m)\top}(\mathbf{X}^{(m)}\Delta\mathbf{X}^{(m)\top} + \Delta\mathbf{X}^{(m)}\mathbf{X}^{(m)\top})\mathbf{a}_i^{(m)}}$$

and compute

$$\lambda_{t+1,i}^{(m)} = \lambda_{t,i}^{(m)} + \Delta\lambda_{t,i}^{(m)};$$

Compute $\Delta\mathbf{a}_{t,i}^{(m)}$ using

$$\boxed{\Delta\mathbf{a}_i^{(m)} = \sum_{j\neq i}\frac{\mathbf{a}_j^{(m)\top}(\mathbf{X}^{(m)}\Delta\mathbf{X}^{(m)\top} + \Delta\mathbf{X}^{(m)}\mathbf{X}^{(m)\top})\mathbf{a}_i^{(m)}}{\lambda_i^{(m)} - \lambda_j^{(m)}}\mathbf{a}_j^{(m)}}$$

and compute

$$\mathbf{a}_{t+1,i}^{(m)} = \mathbf{a}_{t,i}^{(m)} + \Delta\mathbf{a}_{t,i}^{(m)} \text{ and } \mathbf{A}_{t+1}^{(m)} = \{\mathbf{a}_{t+1,i}^{(m)}\};$$

end for
end for
$\mathcal{Y}_{t+1} = (\mathcal{X}_t + \Delta\mathcal{X}_t)\prod_{m=1}^M \times_{(m)} \mathbf{A}_{t+1}^{(m)\mathrm{T}};$
return $\mathbf{A}_{t+1}^{(m)}|_{m=1}^M, \lambda_{t+1}^{(m)}|_{m=1}^M, \mathcal{Y}_{t+1}$

Bound Guarantee

core tensor

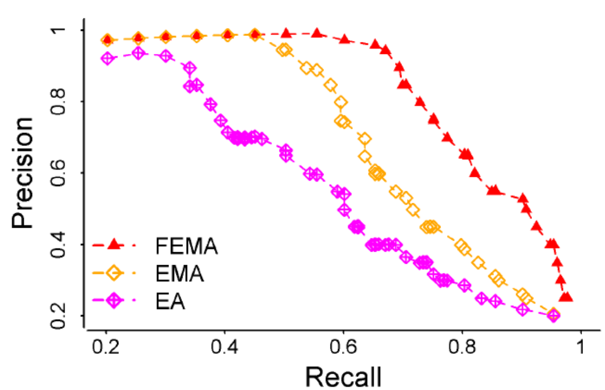
$$|\Delta\lambda_i^{(m)}| \leq 2(\lambda^{\max}_{\mathbf{X}^{(m)\top}\mathbf{X}^{(m)}})^{\frac{1}{2}}\|\Delta\mathbf{X}^{(m)}\|_2$$

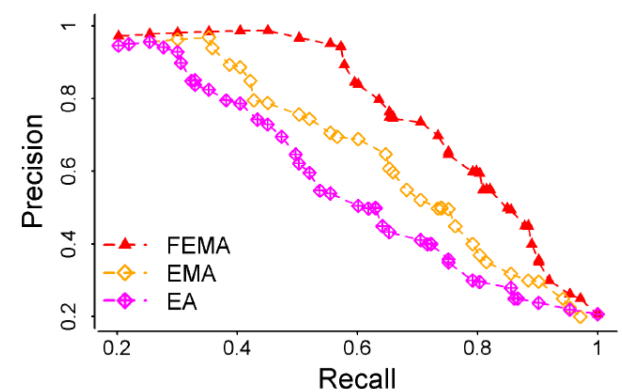
$$|\Delta\mathbf{a}_i^{(m)}| \leq 2\|\Delta\mathbf{X}^{(m)}\|_2\sum_{j\neq i}\frac{(\lambda^{\max}_{\mathbf{X}^{(m)\top}\mathbf{X}^{(m)}})^{\frac{1}{2}}}{|\lambda_i^{(m)} - \lambda_j^{(m)}|}$$

projection matrix

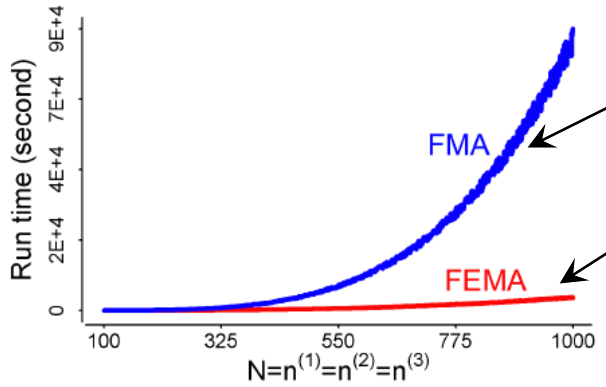
Results: FEMA > EMA > EA

	Microsoft Academic Search		Tencent Weibo mentions “@”	
	MAE	RMSE	MAE	RMSE
FEMA X L	**0.735**	**0.944**	**0.894**	**1.312**
EMA X	0.794	1.130	0.932	1.556
EA X	0.979	1.364	1.120	1.873
Precision vs Recall				





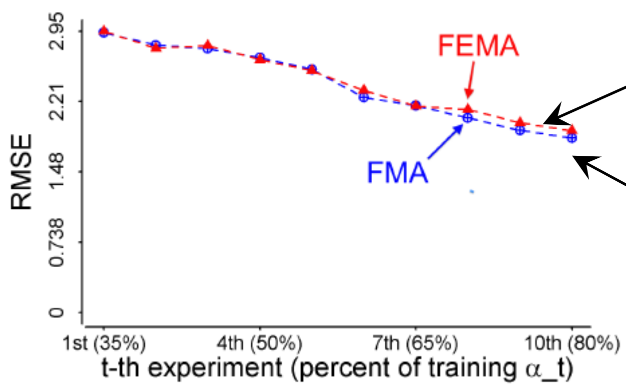
Results: Efficiency



Time vs Num. objects N

Re-decomposition:
Re-compute projection matrices

Evolutionary analysis:
Use ΔX to update λ and a



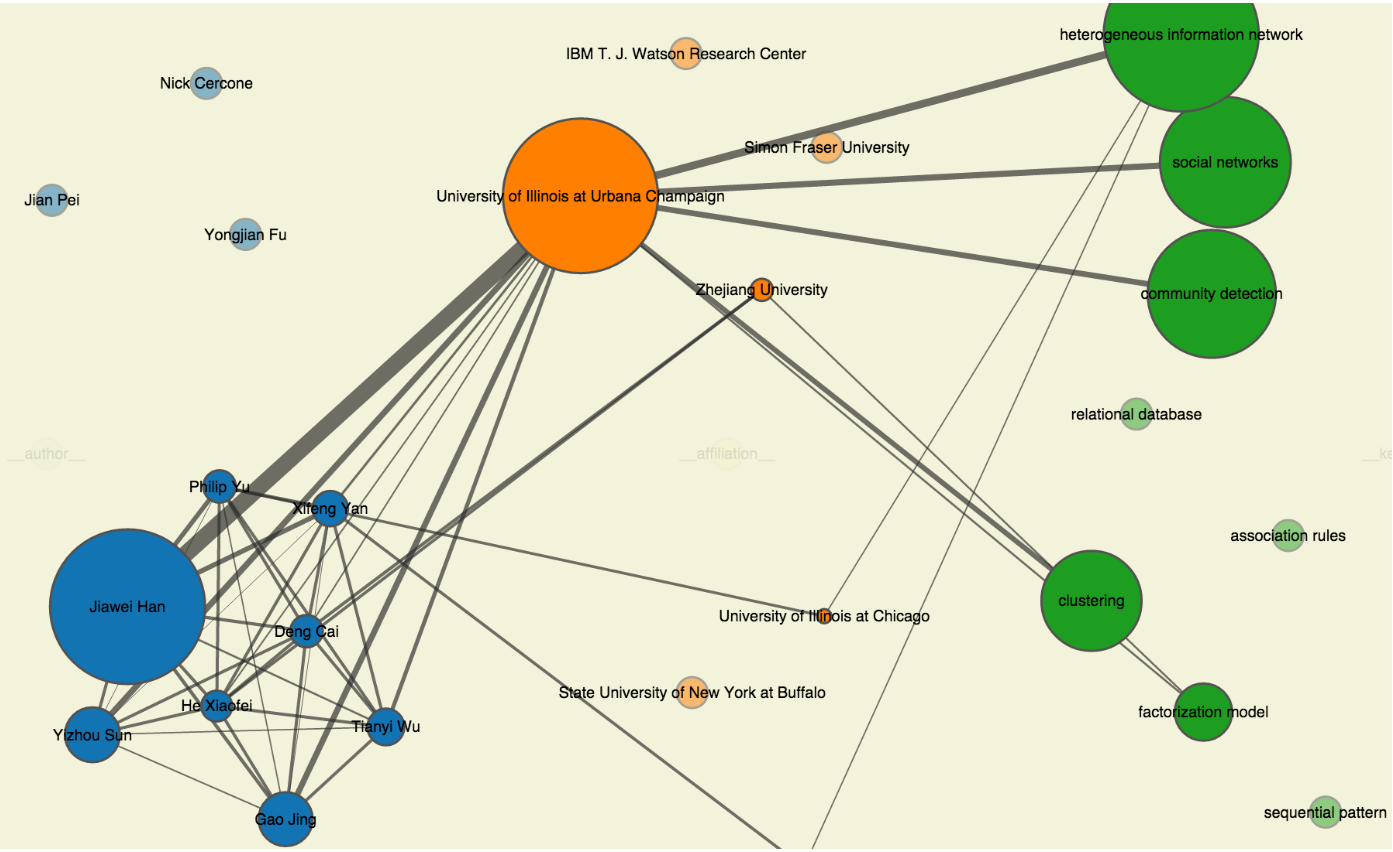
The loss is small.

Evolutionary analysis:
Use ΔX to update λ and a

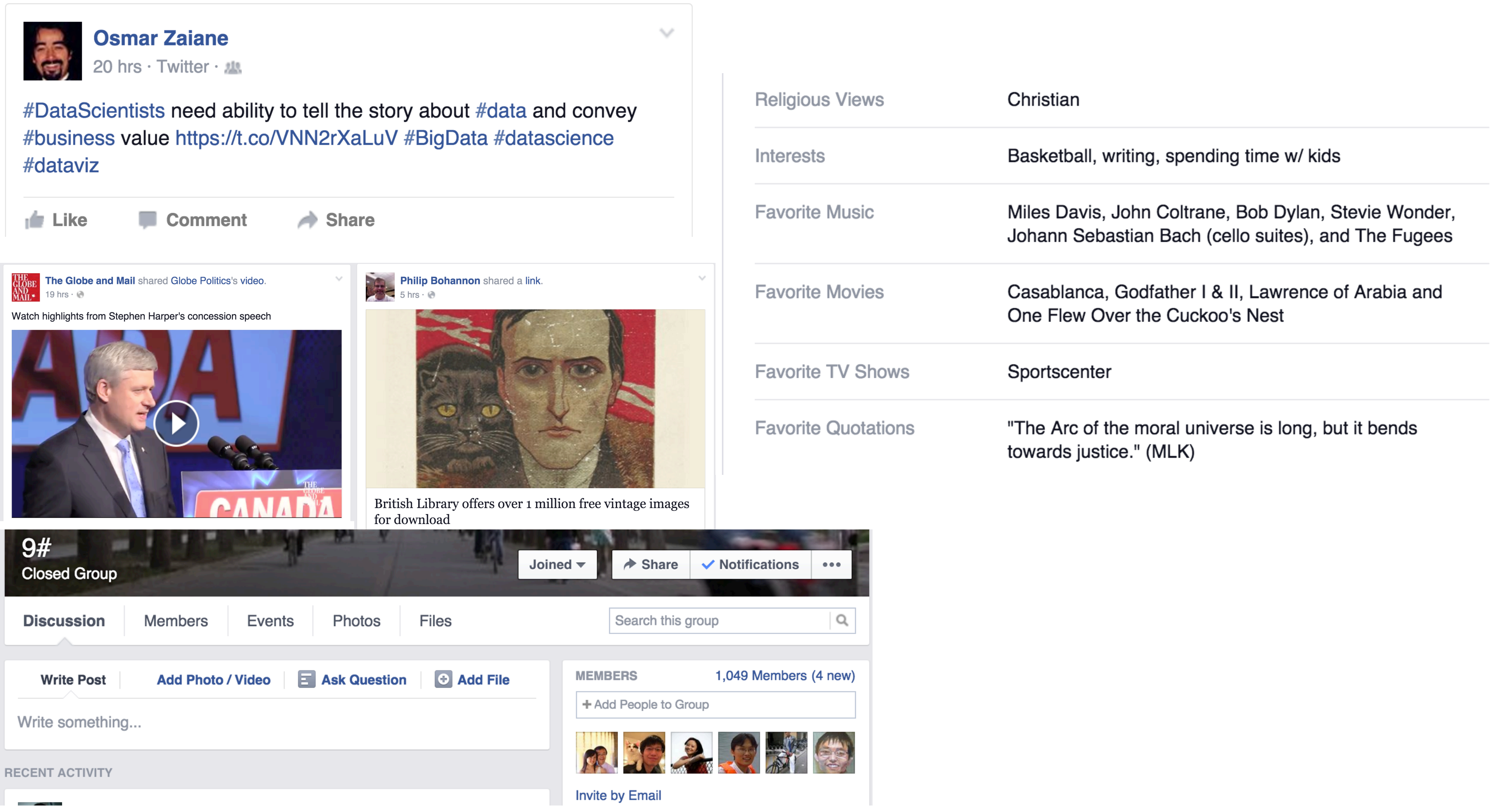
Re-decomposition:
Re-compute projection matrices

Demo: Author@Affiliation#Keyword

http://www.meng-jiang.com/demos/fema/mas/

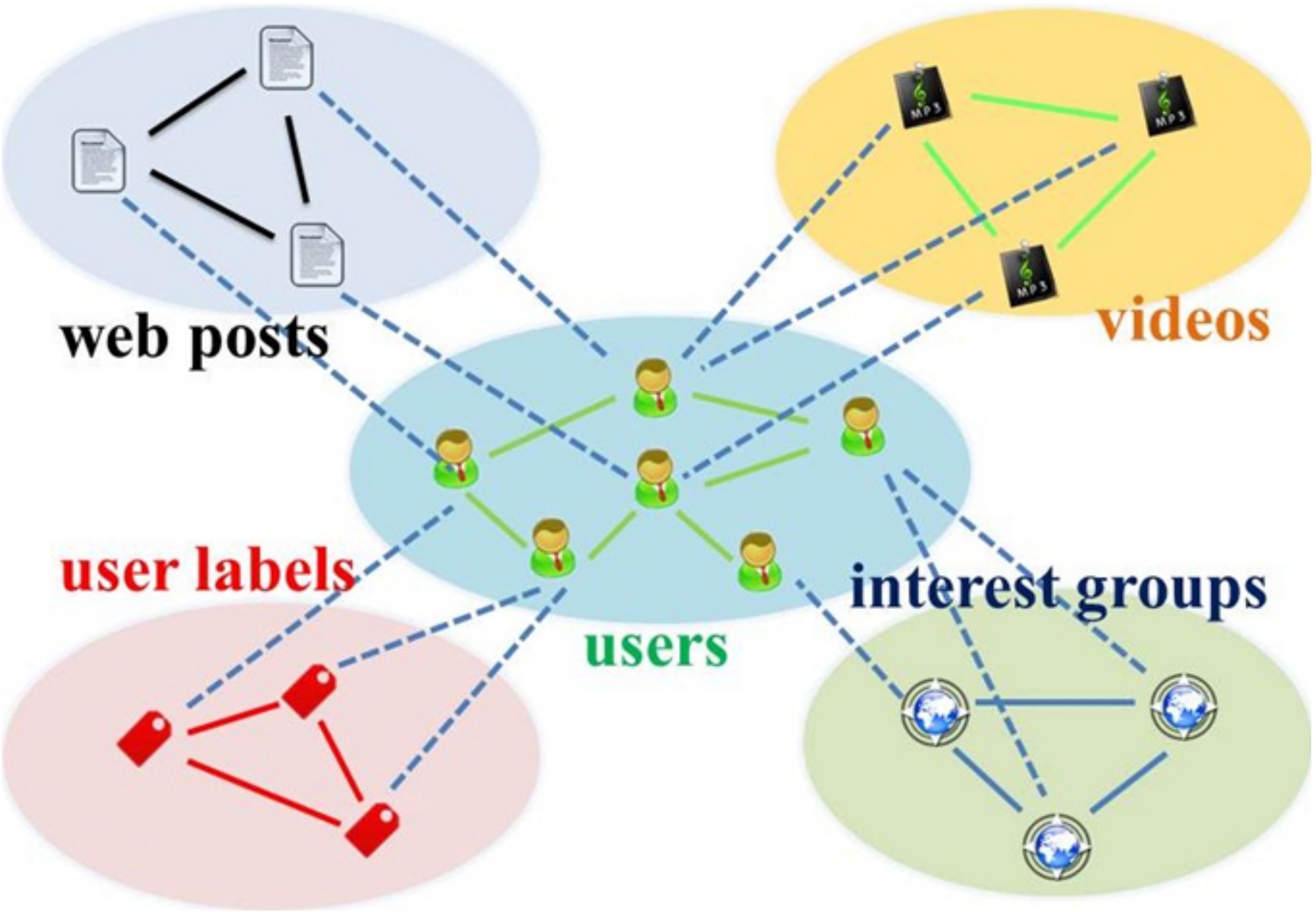


Observation: Multiple Domains

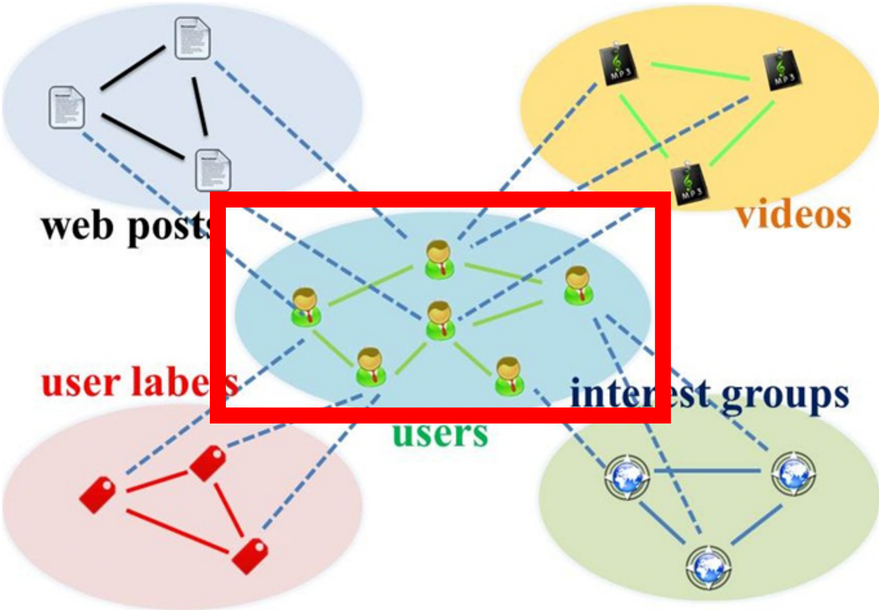


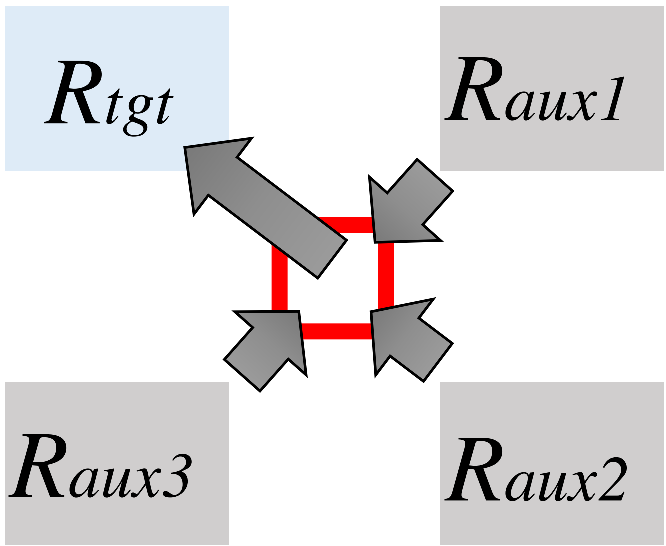
Jiang et al. **Social Recommendation across Multiple Relational Domains.** *CIKM*, 2012.
Social Recommendation with Cross-Domain Transferable Knowledge. *TKDE*, 2015.

Representation: Star-Structured Graph



Representation: Social Bridge

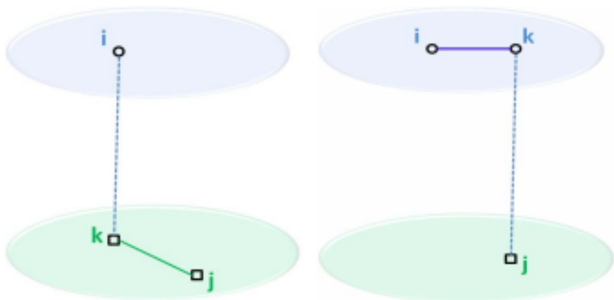




Bridge: Tie strength

Algorithm: Hybrid Random Walk

- Updating cross-domain links



(a) Route 1 (b) Route 2

$$p_{ij}^{(\mathcal{UP})^+} = \delta \sum_{u_k \in \mathcal{U}} r_{ik}^{(\mathcal{U})} p_{kj}^{(\mathcal{UP})^+} + (1-\delta) \sum_{p_k \in \mathcal{P}} p_{ik}^{(\mathcal{UP})^+} r_{kj}^{(\mathcal{P})}$$

$$p_{ij}^{(\mathcal{UP})^-} = \delta \sum_{u_k \in \mathcal{U}} r_{ik}^{(\mathcal{U})} p_{kj}^{(\mathcal{UP})^-} + (1-\delta) \sum_{p_k \in \mathcal{P}} p_{ik}^{(\mathcal{UP})^-} r_{kj}^{(\mathcal{P})}$$

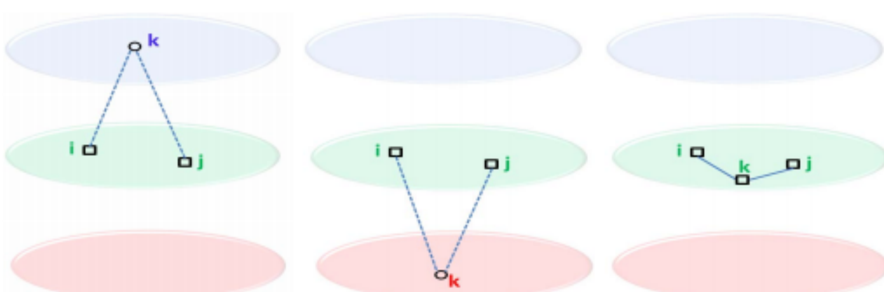
$$p_{ij}^{(\mathcal{UT})^+} = \eta \sum_{u_k \in \mathcal{U}} r_{ik}^{(\mathcal{U})} p_{kj}^{(\mathcal{UT})^+} + (1-\eta) \sum_{t_k \in \mathcal{T}} p_{ik}^{(\mathcal{UT})^+} r_{kj}^{(\mathcal{T})}$$

$$\mathbf{P}^{(\mathcal{UP})^+}(t+1) = \delta \mathbf{R}^{(\mathcal{U})}(t)\mathbf{P}^{(\mathcal{UP})^+}(t) + (1-\delta)\mathbf{P}^{(\mathcal{UP})^+}(t)\mathbf{R}^{(\mathcal{P})}$$

$$\mathbf{P}^{(\mathcal{UP})^-}(t+1) = \delta \mathbf{R}^{(\mathcal{U})}(t)\mathbf{P}^{(\mathcal{UP})^-}(t) + (1-\delta)\mathbf{P}^{(\mathcal{UP})^-}(t)\mathbf{R}^{(\mathcal{P})}$$

$$\mathbf{P}^{(\mathcal{UT})^+}(t+1) = \eta \mathbf{R}^{(\mathcal{U})}(t)\mathbf{P}^{(\mathcal{UT})^+}(t) + (1-\eta)\mathbf{P}^{(\mathcal{UT})^+}(t)\mathbf{R}^{(\mathcal{T})}$$

- Updating within-domain links

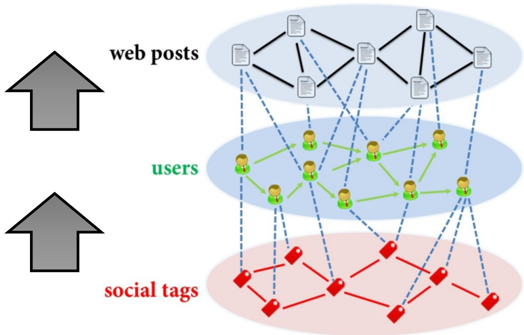


(a) Route 1 (b) Route 2 (c) Route 3

$$r_{ij}^{(\mathcal{U})} = \tau^{(\mathcal{P})}\Big(\mu \sum_{p_k \in \mathcal{P}} p_{ik}^{(\mathcal{UP})^+} p_{jk}^{(\mathcal{UP})^+} + (1-\mu) \sum_{p_k \in \mathcal{P}} p_{ik}^{(\mathcal{UP})^-} p_{jk}^{(\mathcal{UP})^-}\Big) + \tau^{(\mathcal{T})} \sum_{t_k \in \mathcal{T}} p_{ik}^{(\mathcal{UT})^+} p_{jk}^{(\mathcal{UT})^+} + \tau^{(\mathcal{U})} \sum_{u_k \in \mathcal{U}} r_{ik}^{(\mathcal{U})} r_{kj}^{(\mathcal{U})} \tag{12}$$

$$\mathbf{R}^{(\mathcal{U})}(t+1) = \tau^{(\mathcal{P})}\Big(\mu \mathbf{P}^{(\mathcal{UP})^+}(t)\mathbf{P}^{(\mathcal{UP})^+}(t)^T + (1-\mu)\mathbf{P}^{(\mathcal{UP})^-}(t)\mathbf{P}^{(\mathcal{UP})^-}(t)^T\Big) + \tau^{(\mathcal{T})}\mathbf{P}^{(\mathcal{UT})^+}(t)\mathbf{P}^{(\mathcal{UT})^+}(t)^T + \tau^{(\mathcal{U})}\mathbf{R}^{(\mathcal{U})}(t)\mathbf{R}^{(\mathcal{U})}(t)^T \tag{13}$$

Results



Comparing with Random Walk with Restarts Models

Algorithm	MAE	Precision	Recall	F1	Kendall's $\hat{\tau}$
HRW	**0.227±1.5e-3**	**0.711±1.3e-3**	0.921±1.4e-3	**0.802±1.1e-3**	**0.792±2.5e-3**
BRW-R_U-P (TrustWalker)	0.276±1.1e-3	0.657±7.6e-4	**0.935±9.8e-4**	0.772±7.6e-4	0.774±1.6e-3
BRW-R_U	0.282±5.3e-3	0.655±4.0e-3	0.921±1.2e-2	0.765±7.7e-3	0.725±2.8e-3
BRW-W_U-P	0.292±1.1e-3	0.666±7.0e-4	0.900±5.2e-4	0.765±6.6e-4	0.725±8.5e-4
BRW-W_U (ItemRank)	0.318±1.4e-3	0.671±1.5e-3	0.713±2.4e-3	0.691±1.2e-3	0.661±2.2e-3
BRW-P	0.438±2.6e-4	0.571±3.4e-4	0.499±4.2e-4	0.532±3.2e-4	0.606±2.3e-4

Comparing with Social Recommendation Baselines

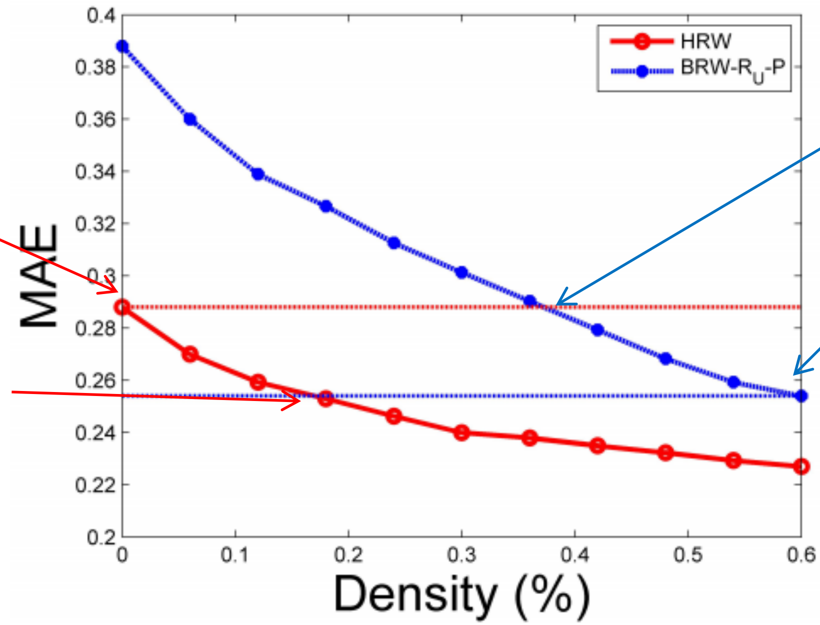
Algorithm	MAE	Precision	Recall	F1	Kendall's $\hat{\tau}$
HRW	**0.227±1.5e-3**	**0.711±1.3e-3**	0.921±1.4e-3	**0.802±1.1e-3**	**0.792±2.5e-3**
BRW-R_U-P (TrustWalker) [10]	0.276±1.1e-3	0.657±7.6e-4	0.935±9.8e-4	0.772±7.6e-4	0.774±1.6e-3
BRW-W_U (ItemRank) [8]	0.318±1.4e-3	0.671±1.5e-3	0.713±2.4e-3	0.691±1.2e-3	0.661±2.2e-3
MCF [5]	0.352±2.3e-4	0.592±1.8e-3	**0.951±6.0e-4**	0.730±1.3e-3	0.582±4.3e-4
CF [22]	0.506±3.4e-4	0.552±1.5e-3	0.589±7.2e-4	0.570±1.0e-3	0.540±5.2e-4

Results: Insight

- Knowledge transfer from auxiliary domains improves cold-start users' behavior prediction
  - Using aux. (label) data, saving ***60-70%*** tgt. (post) data

0 user-post
100% user-label

18% user-post
100% user-label



35% user-post
0 user-label

60% user-post
0 user-label

Observation: Multiple Platforms

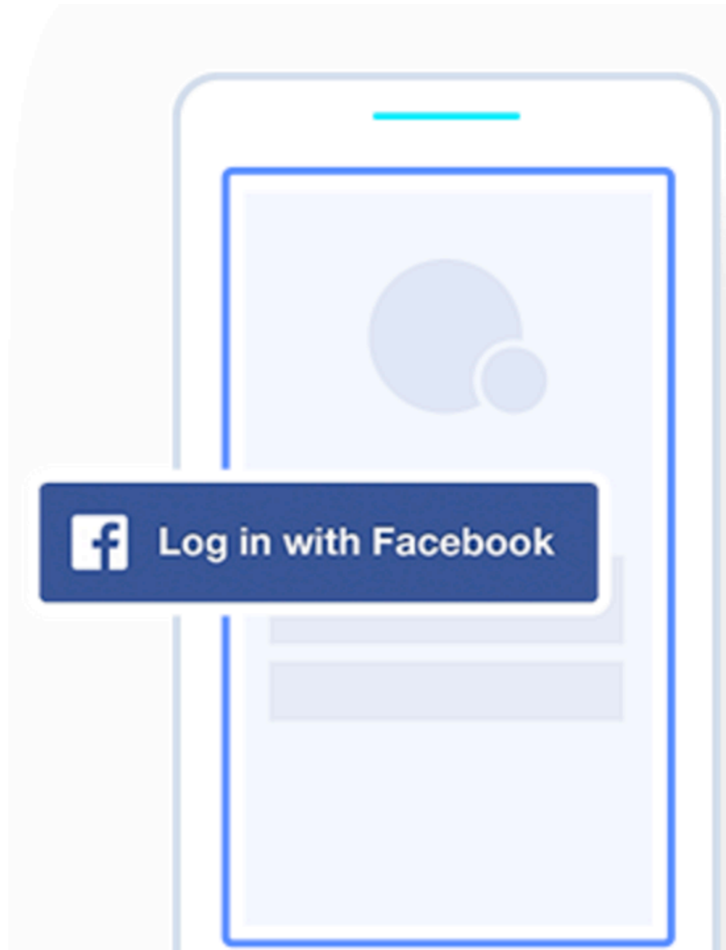
Jiang et al. **Little is Much: Bridging Cross-Platform Behaviors through Overlapped Crowds.** *AAAI*, 2016.

Observation: Cross-Platform

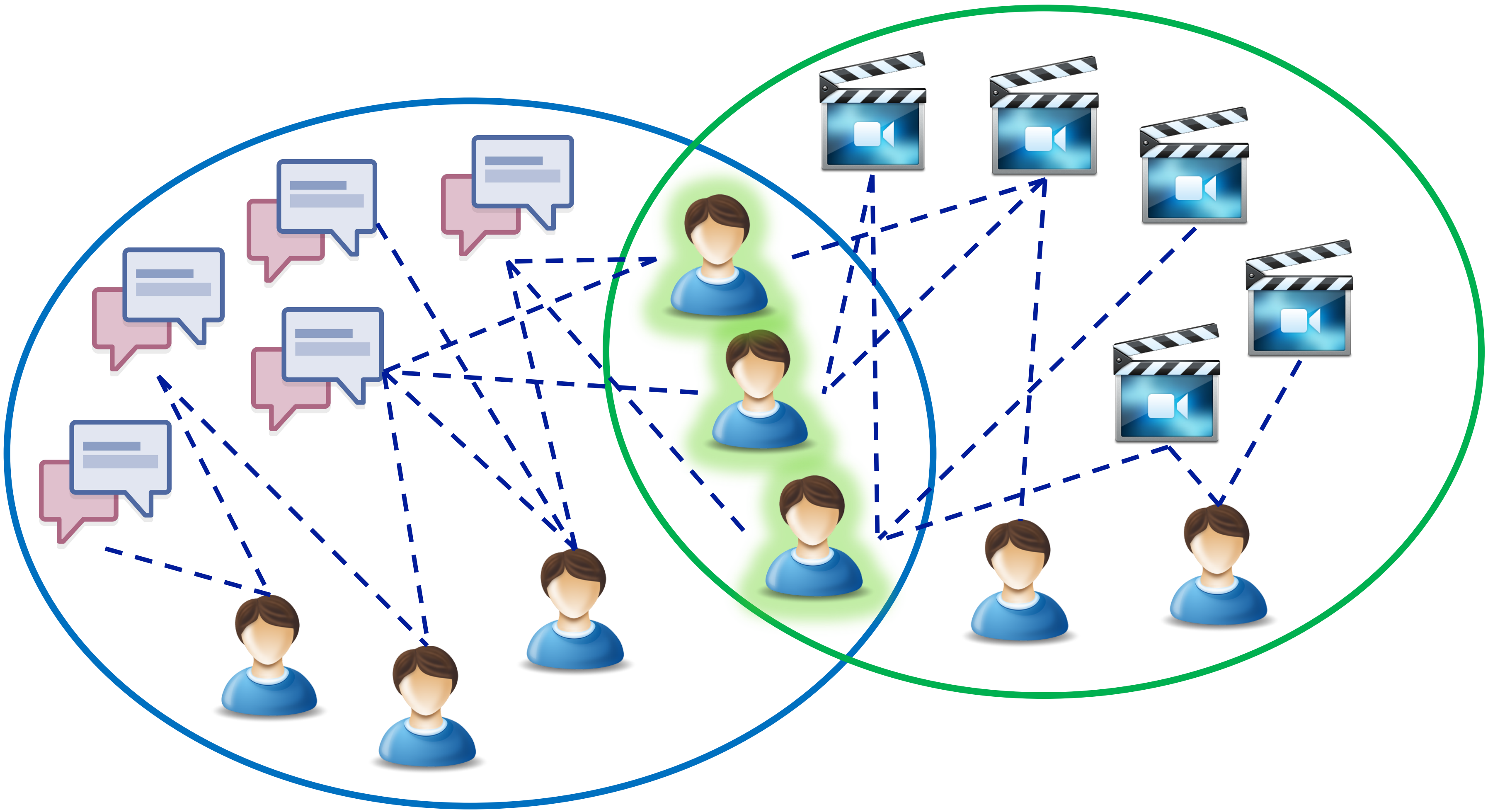
Add Facebook Login to Your App or Website

Facebook Login for Apps is a secure, fast and convenient way for people to log into your app or website.

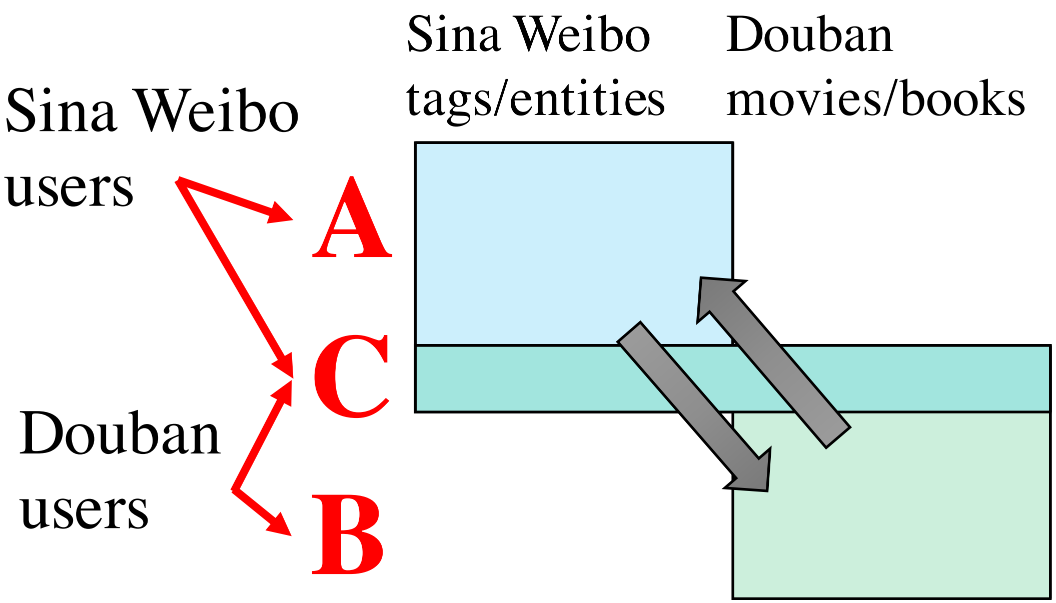
- iOS
- Android
- Websites or mobile websites
- More platforms

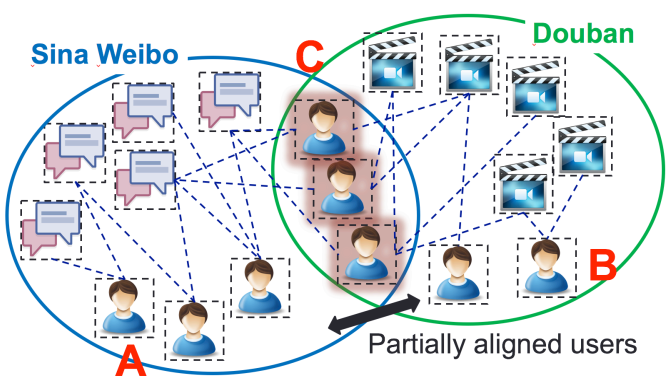


Observation: Partially Overlapped Crowds



Representation: When NO Transfer

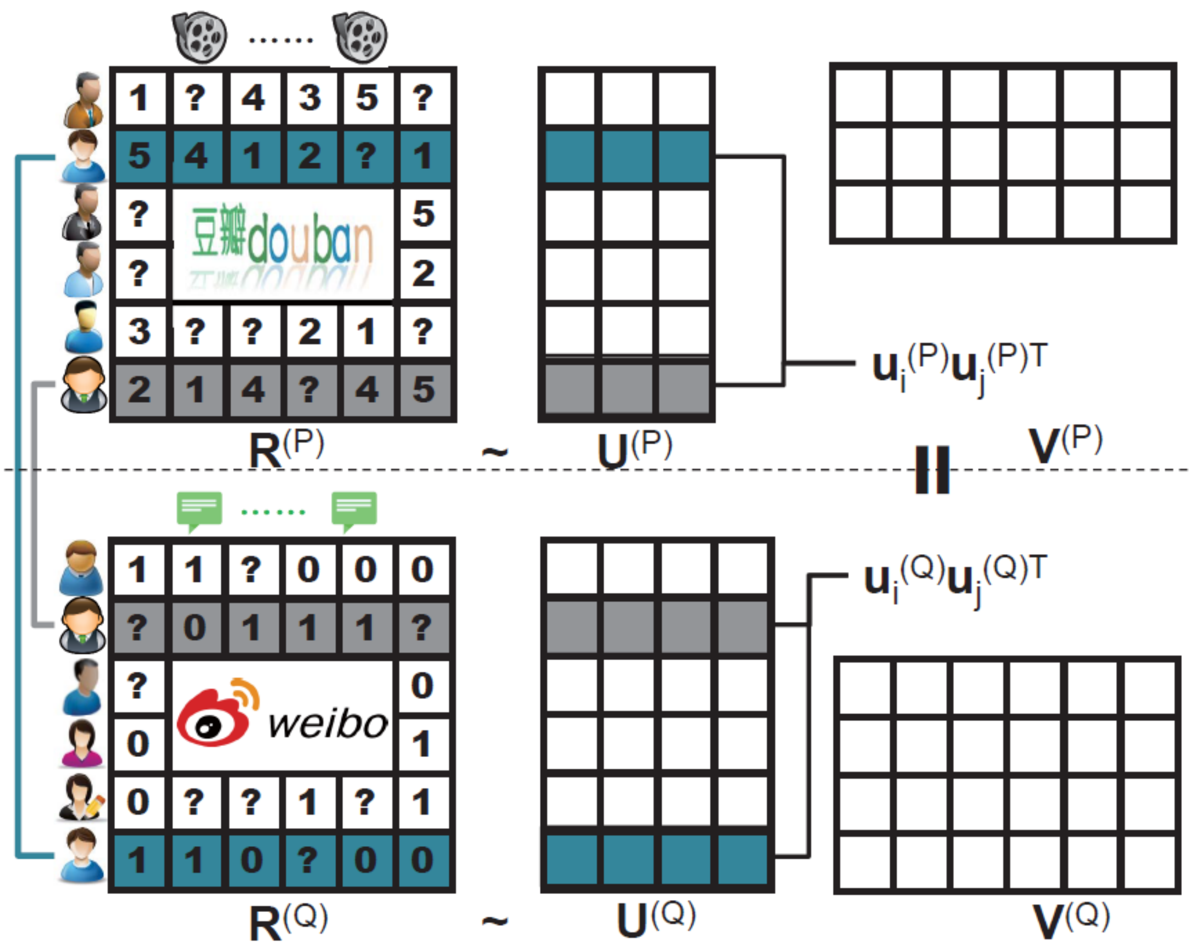




User set	Weibo tweet entity to Douban movie	
	RMSE	MAP
A	Auxiliary platform data!	
C	**0.779**	**0.805**
B	1.439	0.640

User set	Douban book to Weibo social tag	
	RMSE	MAP
A	0.429	0.464
C	**0.267**	**0.666**
B	Auxiliary platform data!	

Algorithm: XPTrans



Algorithm: XPTrans

- Input
  - Tgt./Aux. platform P/Q;
  - Behavior data R(P)/R(Q);
  - Observation W(P)/W(Q);
  - Overlapping indicator W(P,Q),
- Output
  - User latent representation U(P)/U(Q);
  - Item latent representation V(P)/V(Q);
  - Missing values in R(P)

- Objective function

Target platform

Auxiliary platform

$$\mathcal{J} = \sum_{i,j} W_{i,j}^{(P)} \left(R_{i,j}^{(P)} - \sum_{r} U_{i,r}^{(P)} V_{r,j}^{(P)} \right)^2 + \lambda \sum_{i,j} W_{i,j}^{(Q)} \left(R_{i,j}^{(Q)} - \sum_{r} U_{i,r}^{(Q)} V_{r,j}^{(Q)} \right)^2 + \mu \sum_{i_1,j_1,i_2,j_2} W_{i_1,j_1}^{(P,Q)} W_{i_2,j_2}^{(P,Q)} \left(A_{i_1,i_2}^{(P)} - A_{j_1,j_2}^{(Q)} \right)^2$$

Overlapping user similarity (Pair-wise regularization)

Results: Leveraging Auxiliary Platform Data

NO Transfer

User set	Weibo tweet entity to Douban movie	
	RMSE	MAP
A	Auxiliary platform data!	
C	0.779	0.805
B	**1.439**	**0.640**

User set	Douban book to Weibo social tag	
	RMSE	MAP
A	**0.429**	**0.464**
C	0.267	0.666
B	Auxiliary platform data!	

Transfer via the Same Latent Space

User set	Weibo tweet entity to Douban movie	
	RMSE	MAP
A		
C	0.757	0.811
B	**1.164 (-19%)**	**0.702 (+9.7%)**

User set	Douban book to Weibo social tag	
	RMSE	MAP
A	**0.411 (-4.2%)**	**0.487 (+5.0%)**
C	0.256	0.681
B		

Results: Leveraging Different Latent Spaces

Transfer via the Same Latent Space

User set	Weibo tweet entity to Douban movie	
	RMSE	MAP
A		
C	0.757	0.811
B	**1.164**	**0.702**

User set	Douban book to Weibo social tag	
	RMSE	MAP
A	**0.411**	**0.487**
C	0.256	0.681
B		

Transfer via Different Latent Spaces

User set	Weibo tweet entity to Douban movie	
	RMSE	MAP
A		
C	0.715	0.821
B	**0.722 (-38%)**	**0.820 (+17%)**

User set	Douban book to Weibo social tag	
	RMSE	MAP
A	**0.374 (-11%)**	**0.533 (+12%)**
C	0.236	0.705
B		

Results: Where Amazing Happens

NO Transfer

User set	Weibo tweet entity to Douban movie	
	RMSE	MAP
A	Auxiliary platform data!	
C	0.779	0.805
B	**1.439**	**0.640**

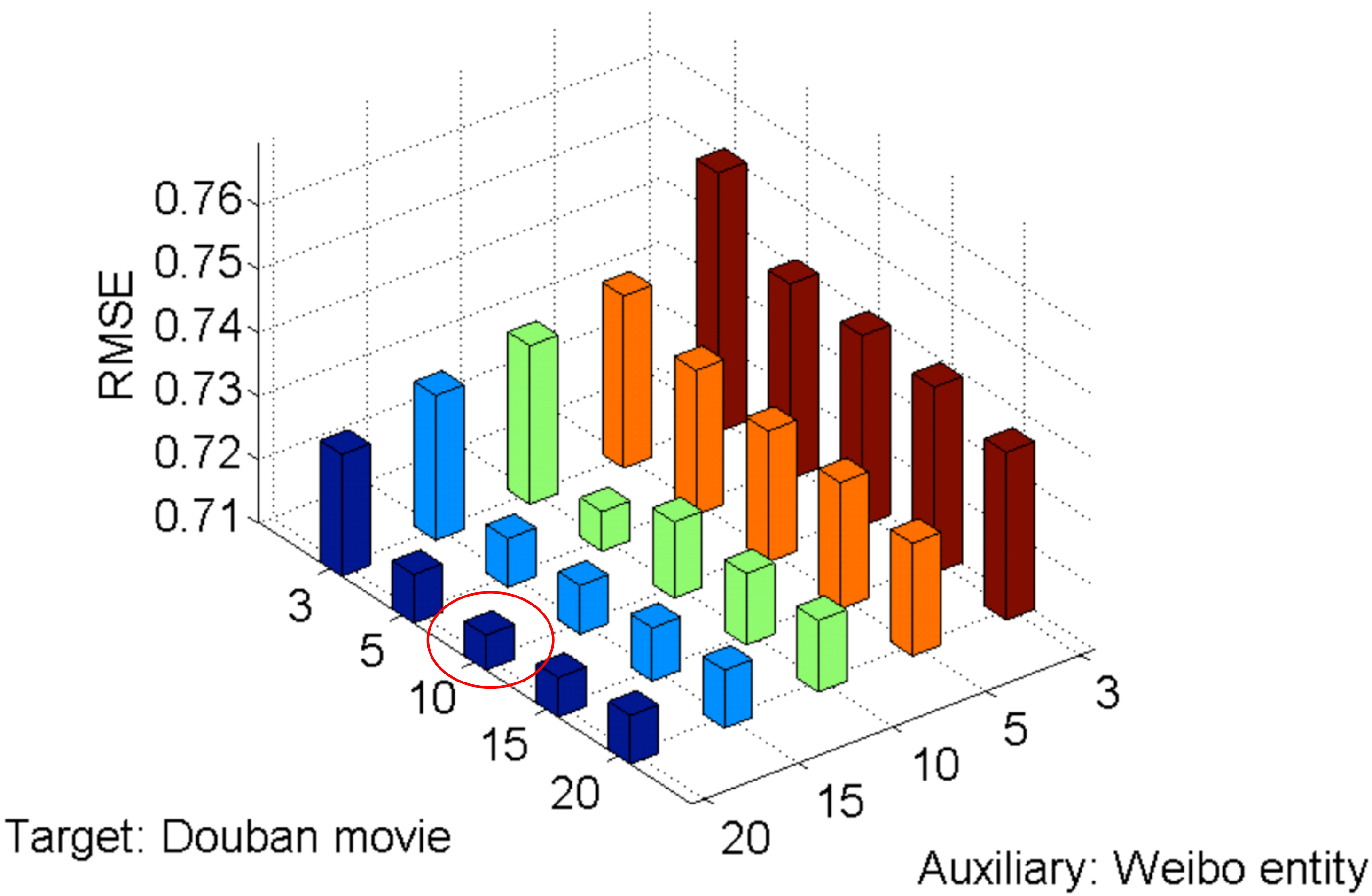
User set	Douban book to Weibo social tag	
	RMSE	MAP
A	**0.429**	**0.464**
C	0.267	0.666
B	Auxiliary platform data!	

Transfer via Different Latent Spaces

User set	Weibo tweet entity to Douban movie	
	RMSE	MAP
A		
C	0.715	0.821
B	**0.722**	**0.820**

User set	Douban book to Weibo social tag	
	RMSE	MAP
A	**0.374**	**0.533**
C	0.236	0.705
B		

Results: Different Sizes of Latent Spaces



Summary

- Like, Reply, Share, Retweet, Favorite, Comment …
- Memory based social recommenders
  - TidalTrust, MoleTrust, TrustWalker
- Model based social recommenders
  - SoRec, "Social Trust" Ensemble, SoReg
- **Observations, Representations, Models**
  - **ContextMF:** Social contexts (preference & influence)
  - **FEMA:** Spatiotemporal contexts (multidimensional)
  - **HybridRW:** Cross-domain behavior modeling
  - **XPTrans:** Cross-platform behavior modeling

I. Mining behavior networks with social and spatiotemporal contexts

I.2. Suspicious behavior detection

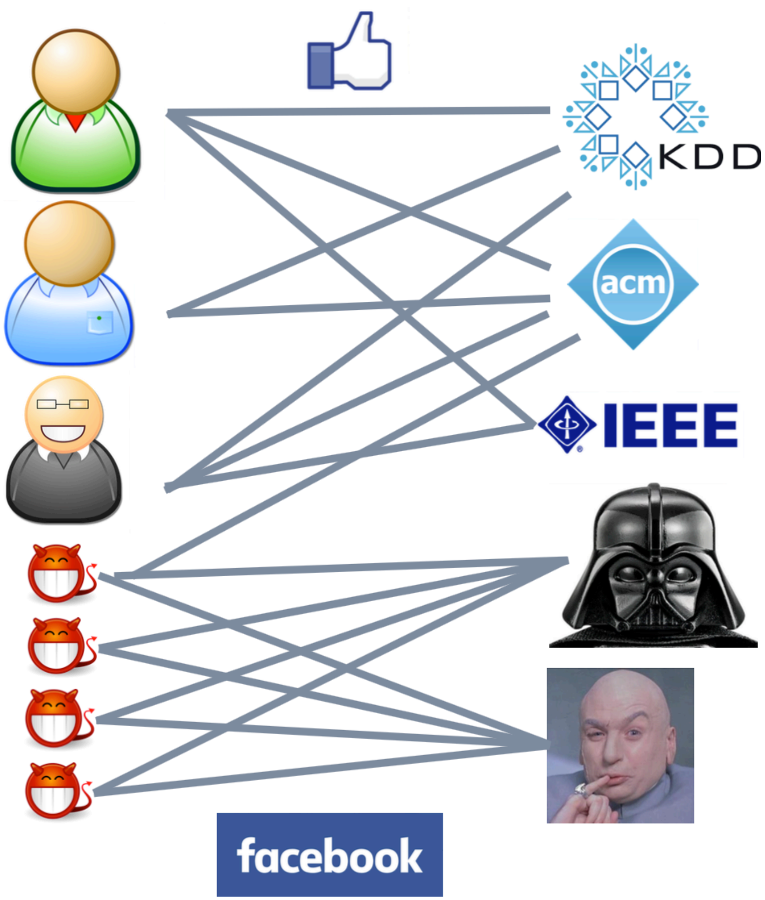
Ill-gotten Facebook Likes

25,000 Facebook Likes	50,000 Facebook Likes	100,000 Facebook Likes	200,000 Facebook Likes
$265	$525	$1,000	$1,750
Lifetime Replacement Warranty	Lifetime Replacement Warranty	Lifetime Replacement Warranty	Lifetime Replacement Warranty
Dedicated 24/7 Customer Service	Dedicated 24/7 Customer Service	Dedicated 24/7 Customer Service	Dedicated 24/7 Customer Service
100% Risk Free, Try Us Today	100% Risk Free, Try Us Today	100% Risk Free, Try Us Today	100% Risk Free, Try Us Today
Order starts within 24 - 48 hours	Order starts within 24 - 48 hours	Order starts within 24 -48 hours	Order starts within 24 -48 hours
Order completed within 22 days	Order completed within 35 days	Order completed within 35 days	Order completed within 35 days

Suspicious Behavior Detection

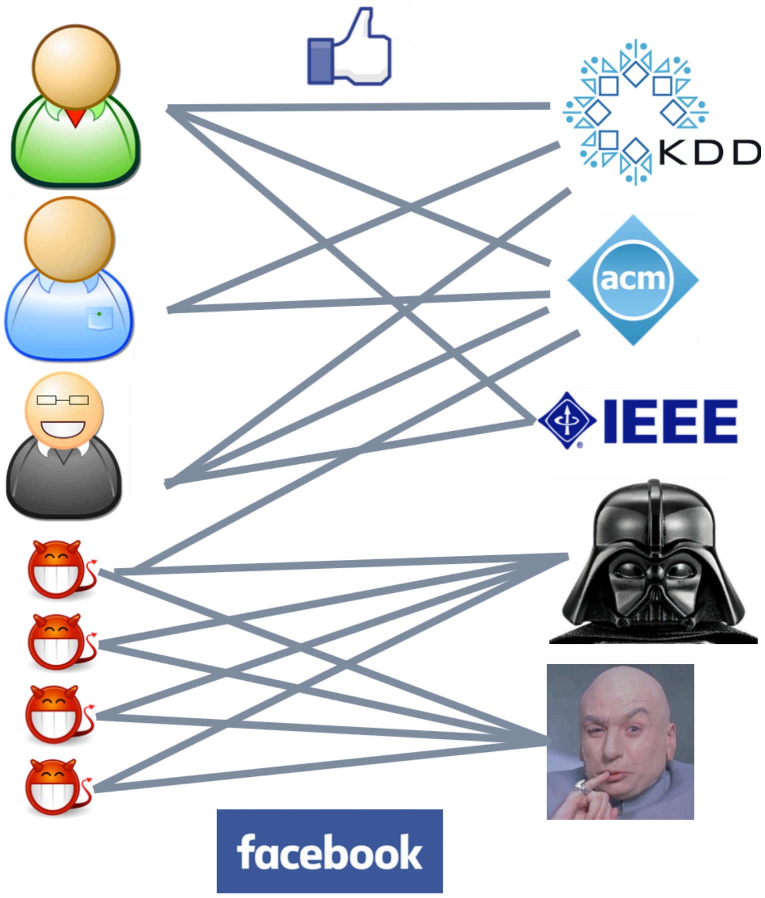
Meng Jiang, Peng Cui and Christos Faloutsos.
Suspicious Behavior Detection: Current Trends and Future Directions.
IEEE Intelligent Systems (ISSI), 2016.

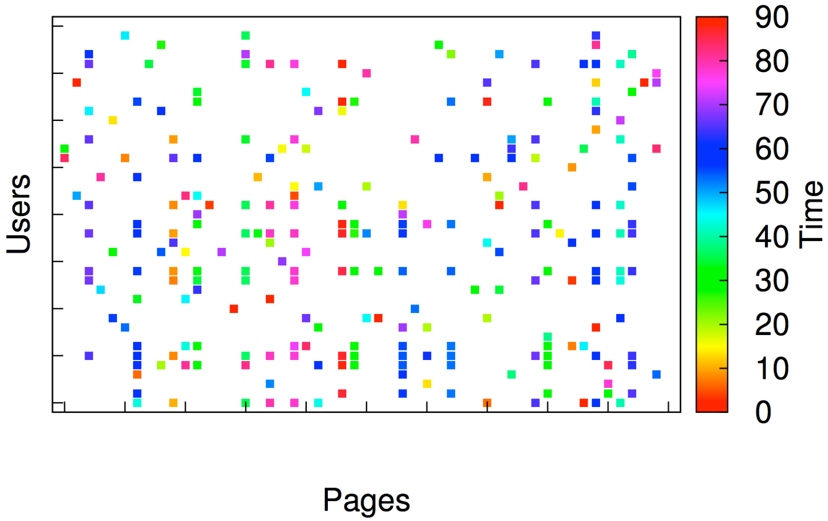
Ill-gotten Facebook Likes



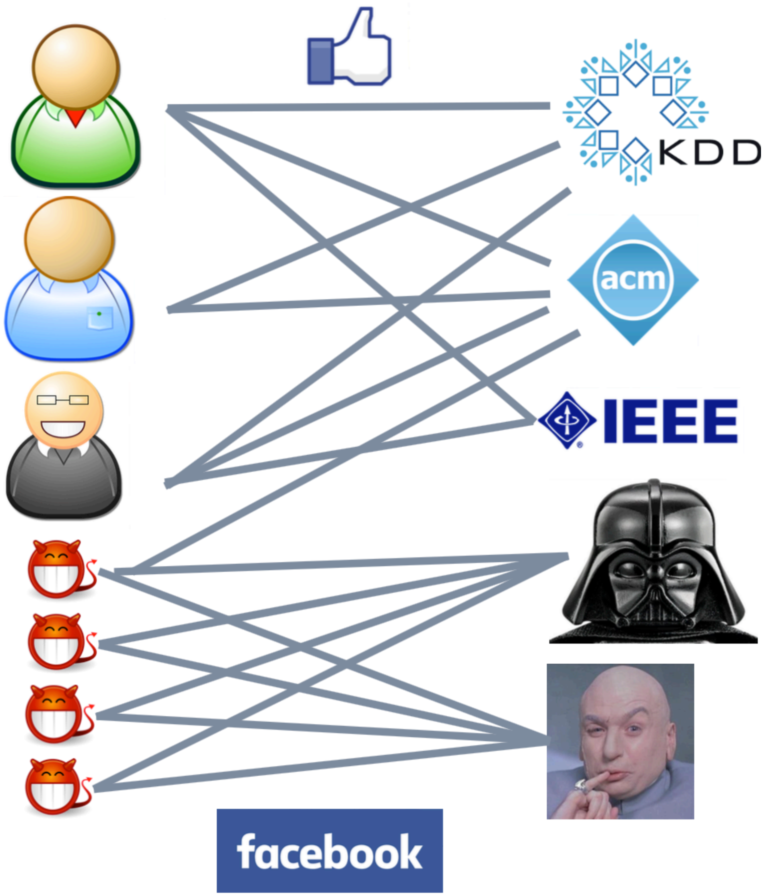
Beutel et al. **CopyCatch: Stopping Group Attacks by Spotting Lockstep Behavior in Social Neworks.** *WWW*, 2013.

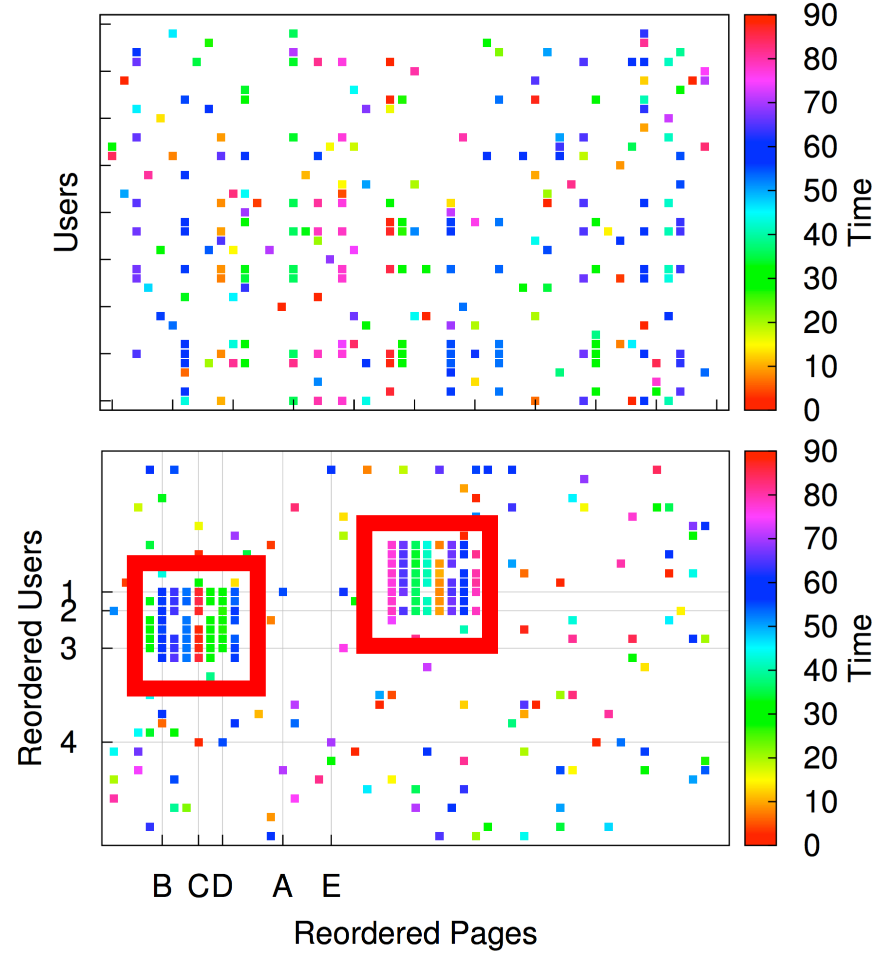
Observation: Graphical View





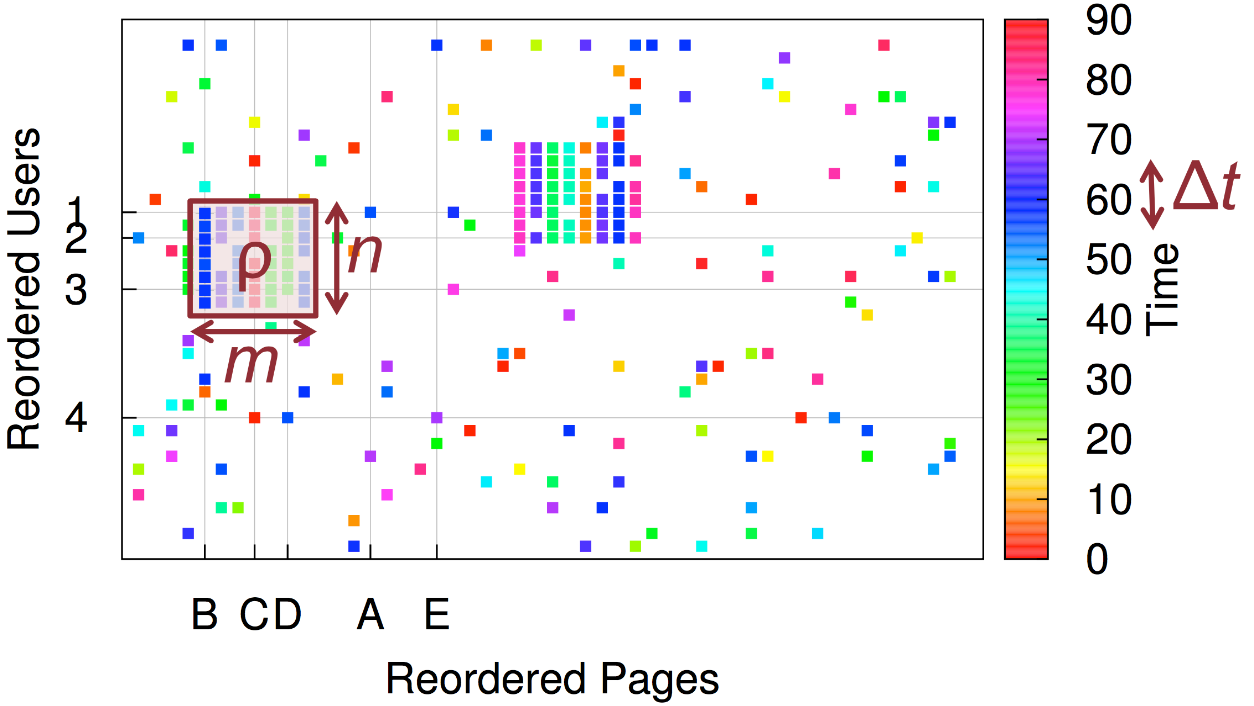
Observation: Reorder Matrix



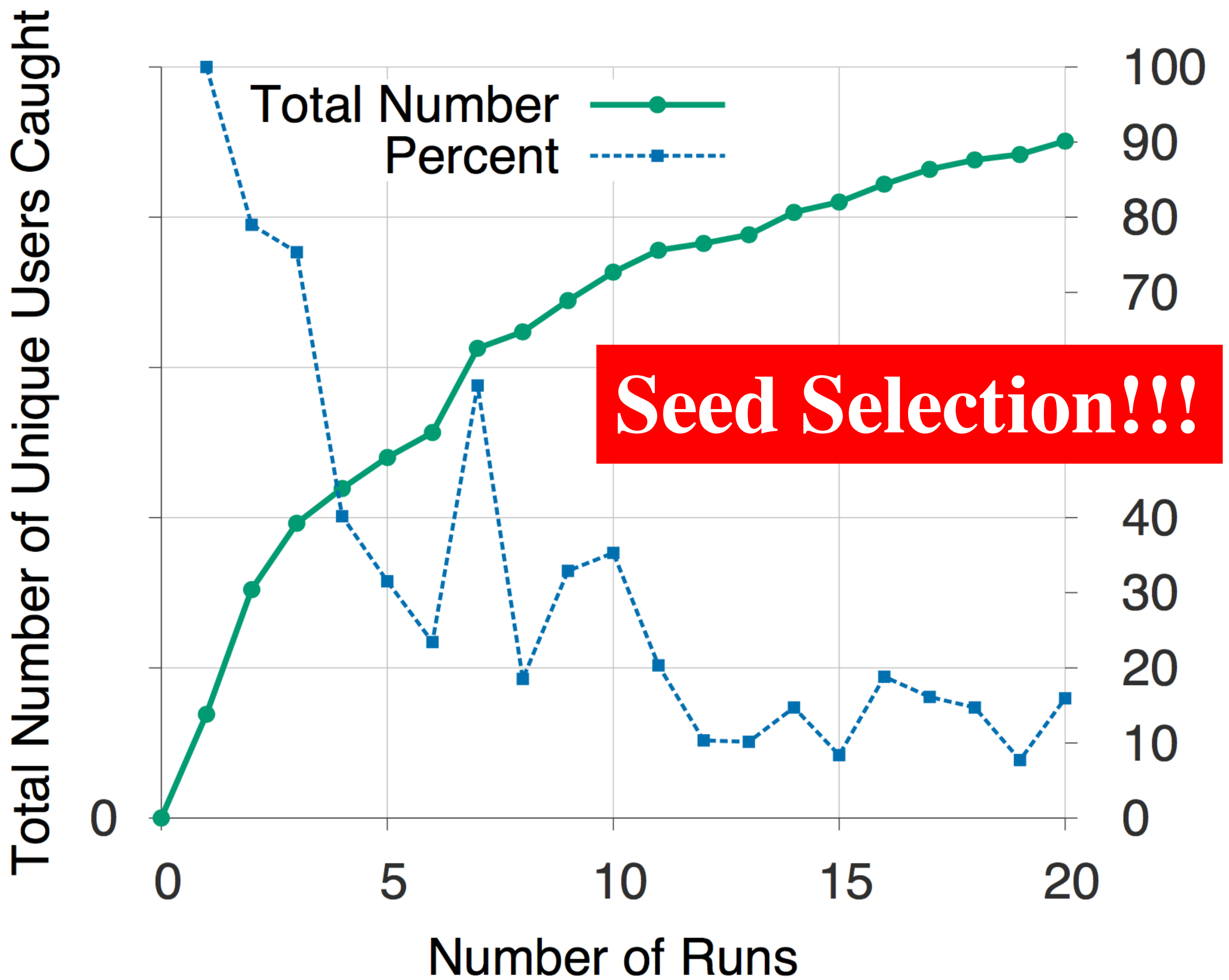


Algorithm: Seed + Search

- CopyCatch
  - "Near Bipartite Core": n users, m Pages, ϱ, Δt



Experimental Result

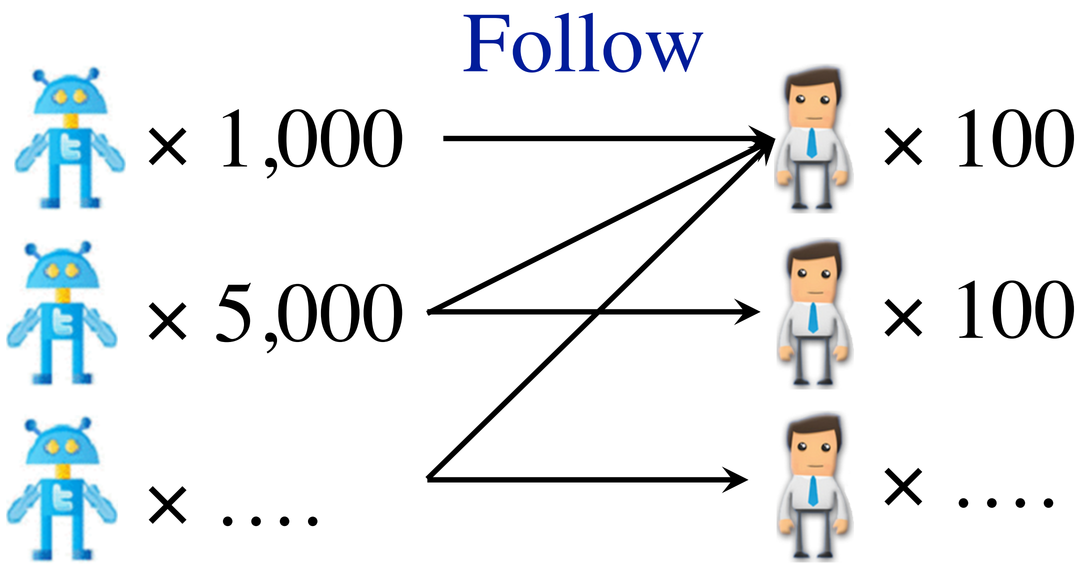


Serious Problem in Weibo

Experience-driven approaches:

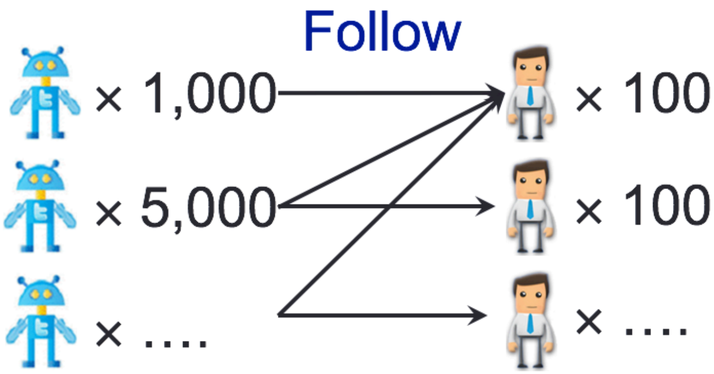
features of #followees, #hashtags, #URLs…

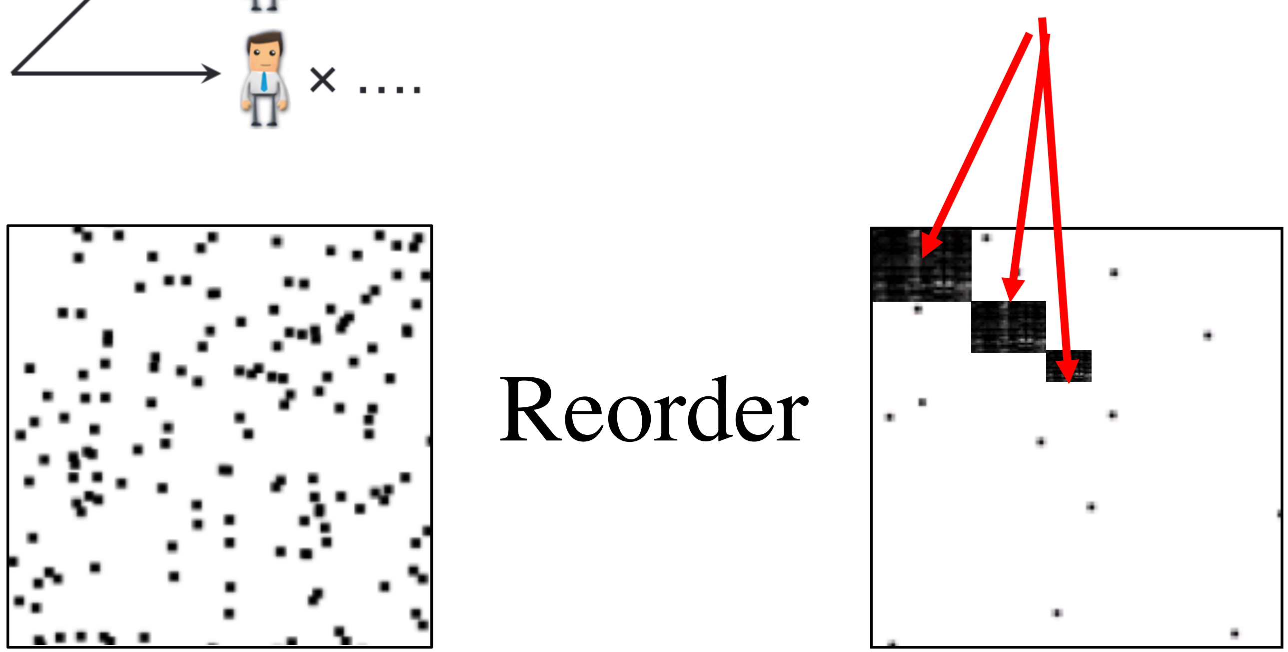
Zombie Followers



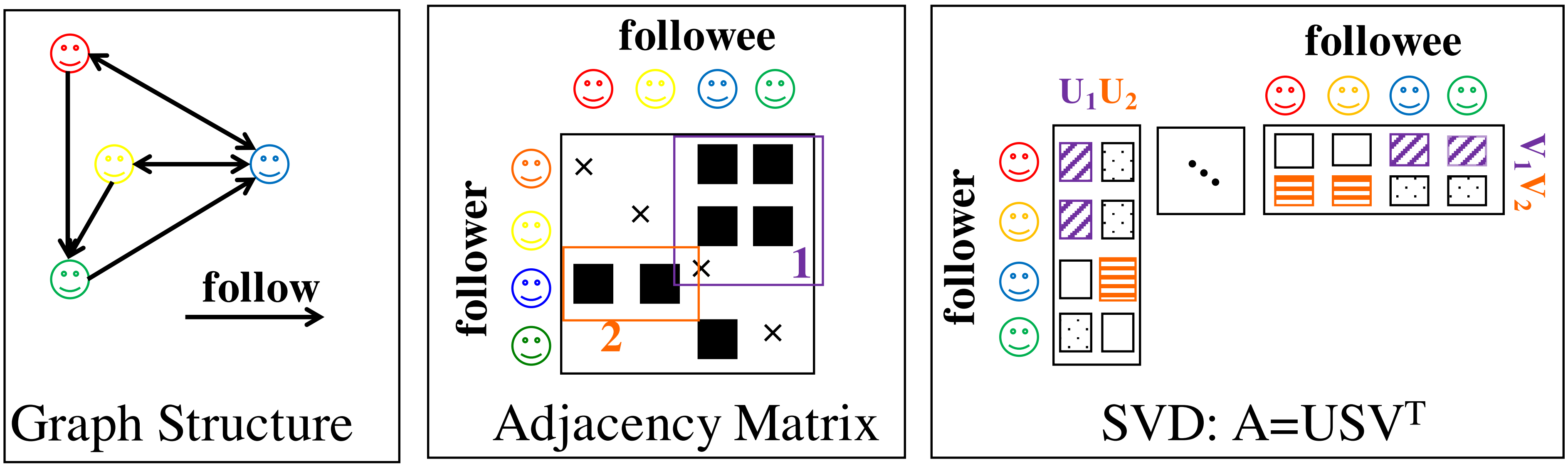
Jiang et al. **Inferring Strange Behavior from Connectivity Pattern in Social Networks.** *PAKDD*, 2014.

Observation: Reorder Matrix

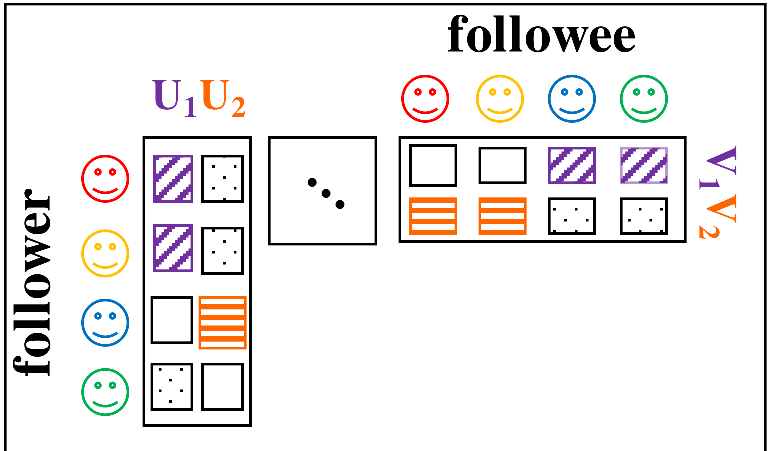




Representation: SVD Reminder

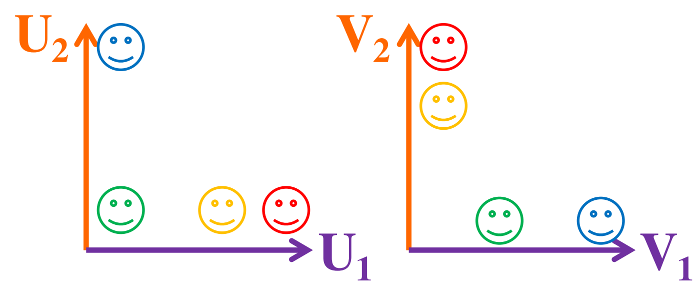


Representation: Spectral Subspace



SVD: $A=USV^T$

Pairs of singular vectors:

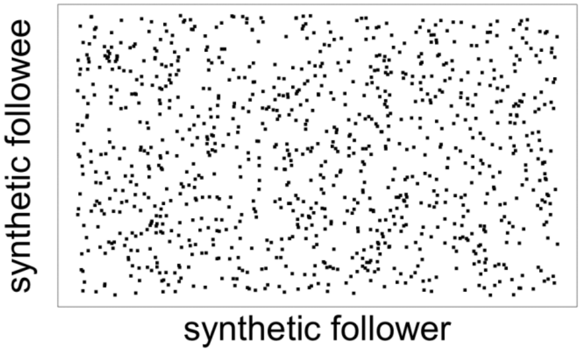


"Spectral Subspace Plot"

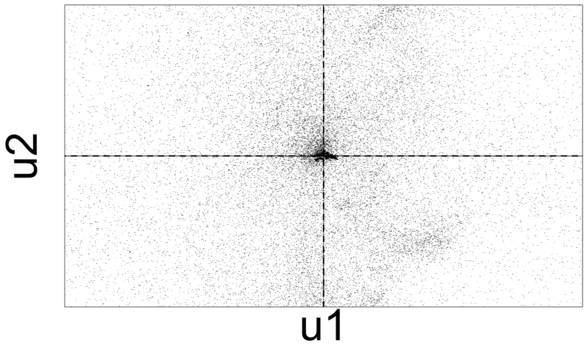
Spectral Subspace Plot: Case #0

- NO lockstep behavior: Scatter

Adjacency Matrix



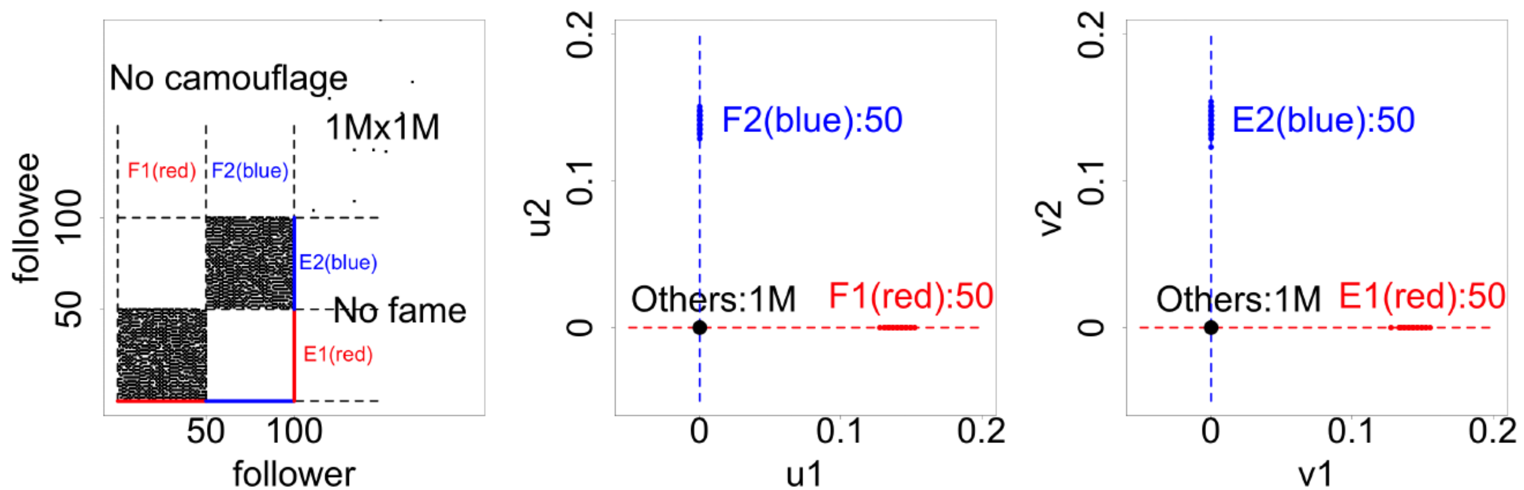
Spectral Subspace Plot



Spectral Subspace Plot: Case #1

- Non-overlapping lockstep: "Rays"

Adjacency Matrix | Spectral Subspace Plot

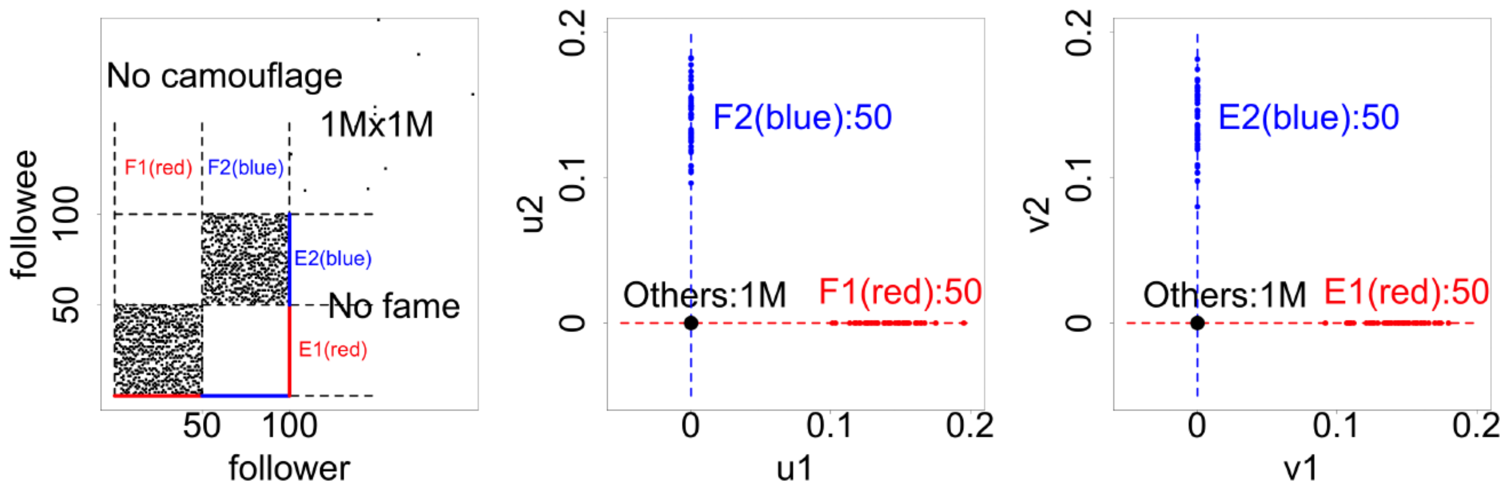


Rule 1 (short "rays"): two blocks, high density (90%), no "camouflage", no "fame"

Spectral Subspace Plot: Case #2

- Non-overlapping: Low density, Elongation

Adjacency Matrix　　　　Spectral Subspace Plot

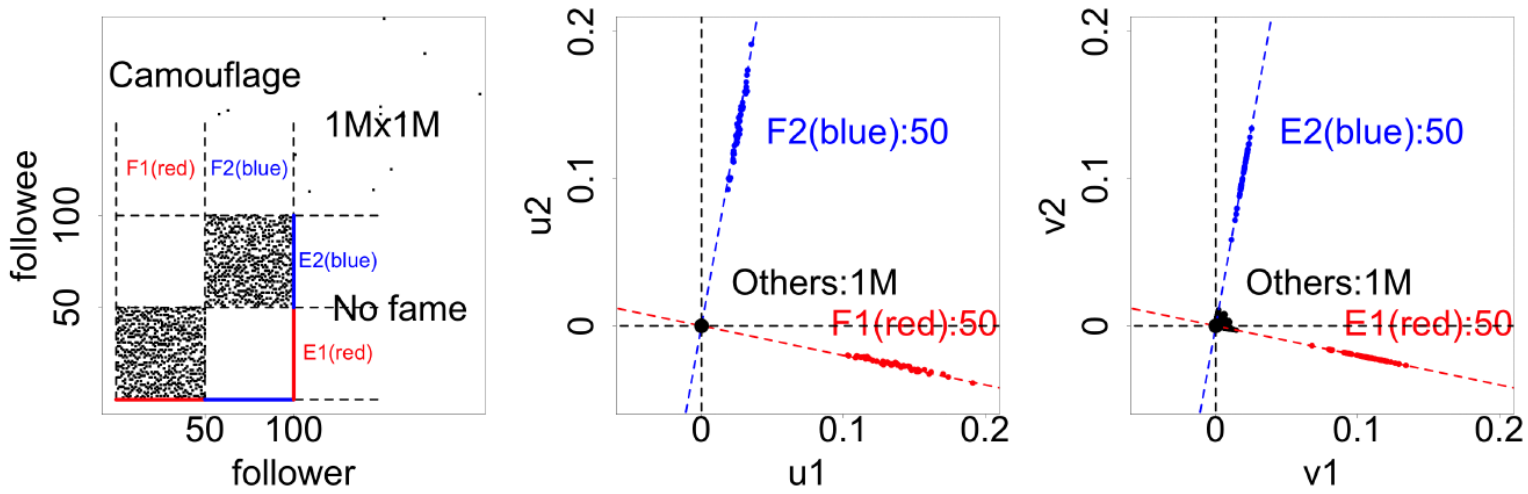


Rule 2 (long "rays"): two blocks, low density (50%), no "camouflage", no "fame"

Spectral Subspace Plot: Case #3

- Non-overlapping: Camouflage/Fame, Tilting

Adjacency Matrix | Spectral Subspace Plot

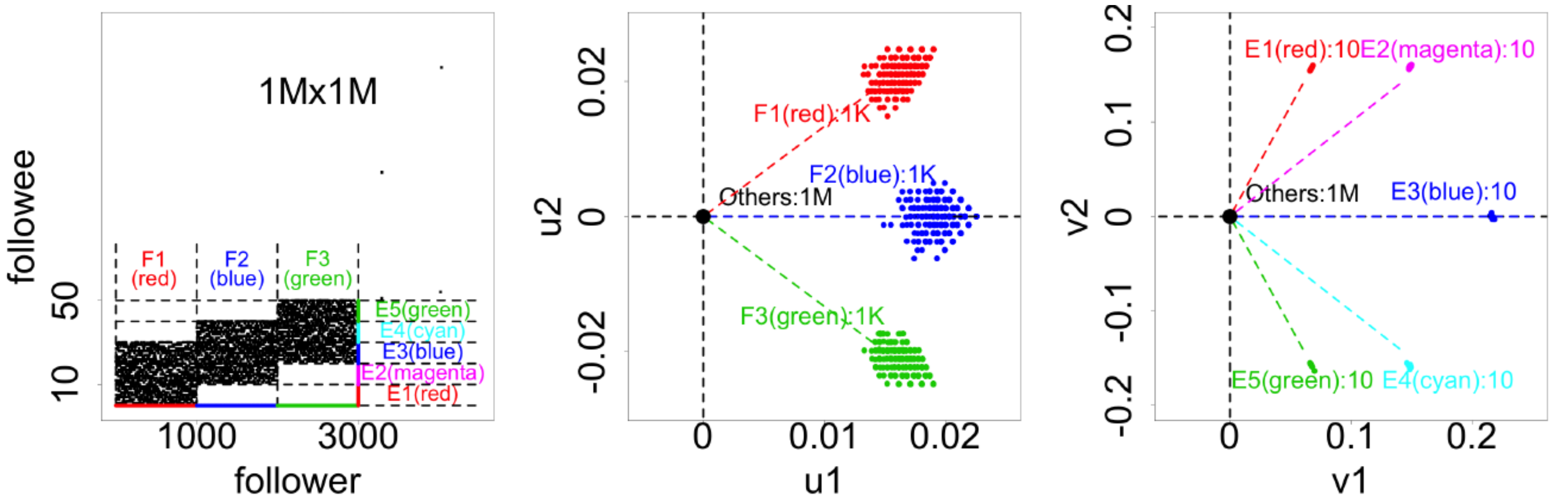


Rule 3 (tilting "rays"): two blocks, with "camouflage", no "fame"

Spectral Subspace Plot: Case #4

- Overlapping: "Staircase", "Pearls"

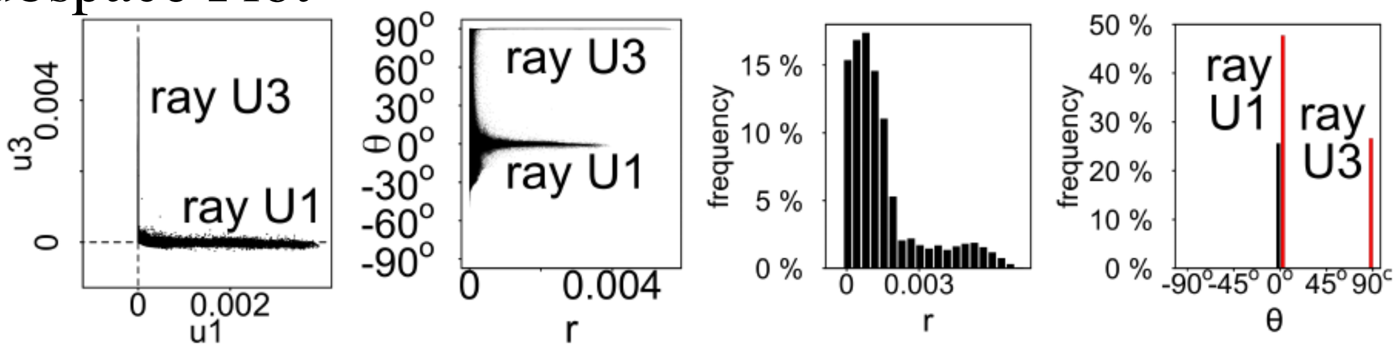
Adjacency Matrix Spectral Subspace Plot



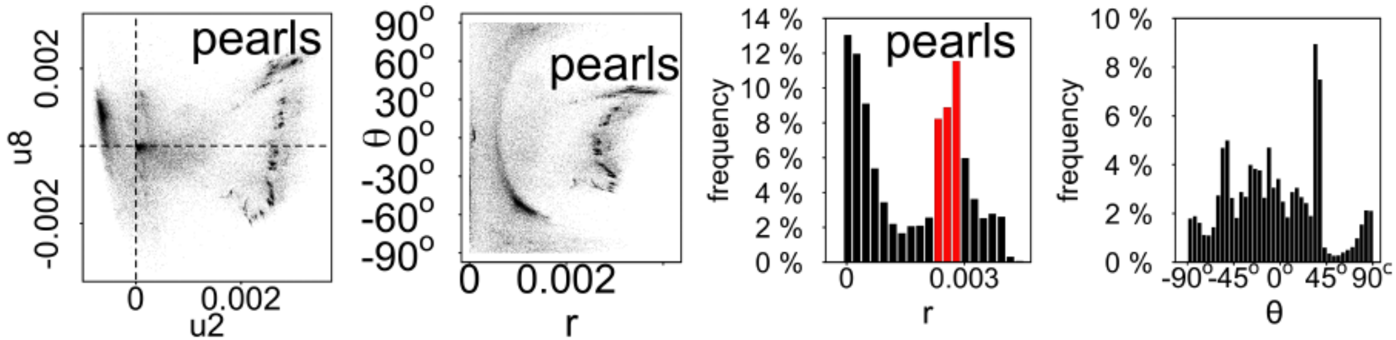
Rule 4 ("pearls"): a "staircase" of three partially overlapping blocks.

Algorithm: Reading & LockInfer

Spectral Subspace Plot | Polar Coordinate Transform | Histograms



“rays” show two apparent spikes on θ frequency at 0^o and 90^o

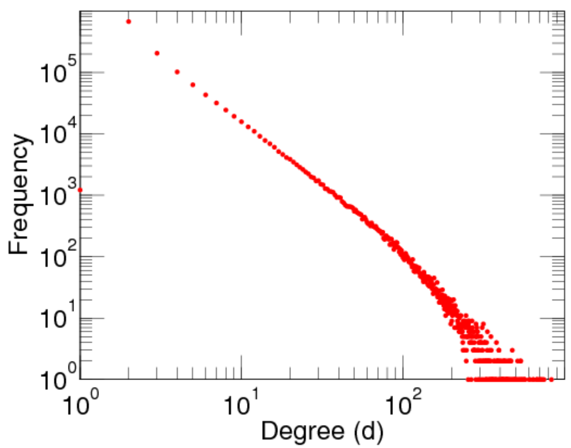


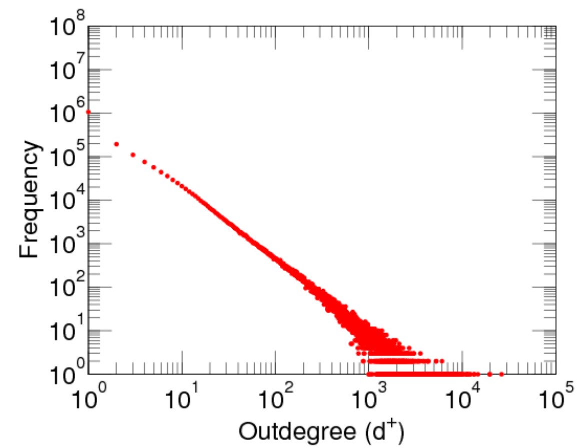
“pearls” show a spike on r frequency at a much-greater-than-zero value

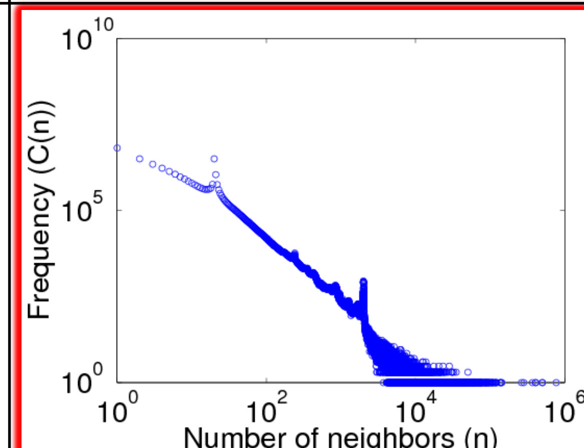
High precision but low recall!!!

Out-Degree Distributions

- **Power-law distribution** [Faloutsos *et al*. SIGCOMM; Broder *et al*. Computer Networks; Chung *et al*. PNAS]

<table>
<tr><th>DBLP
Author-publication</th><th>Flickr
User-user</th><th>Twitter
Who-follows-whom</th></tr>
<tr><td>


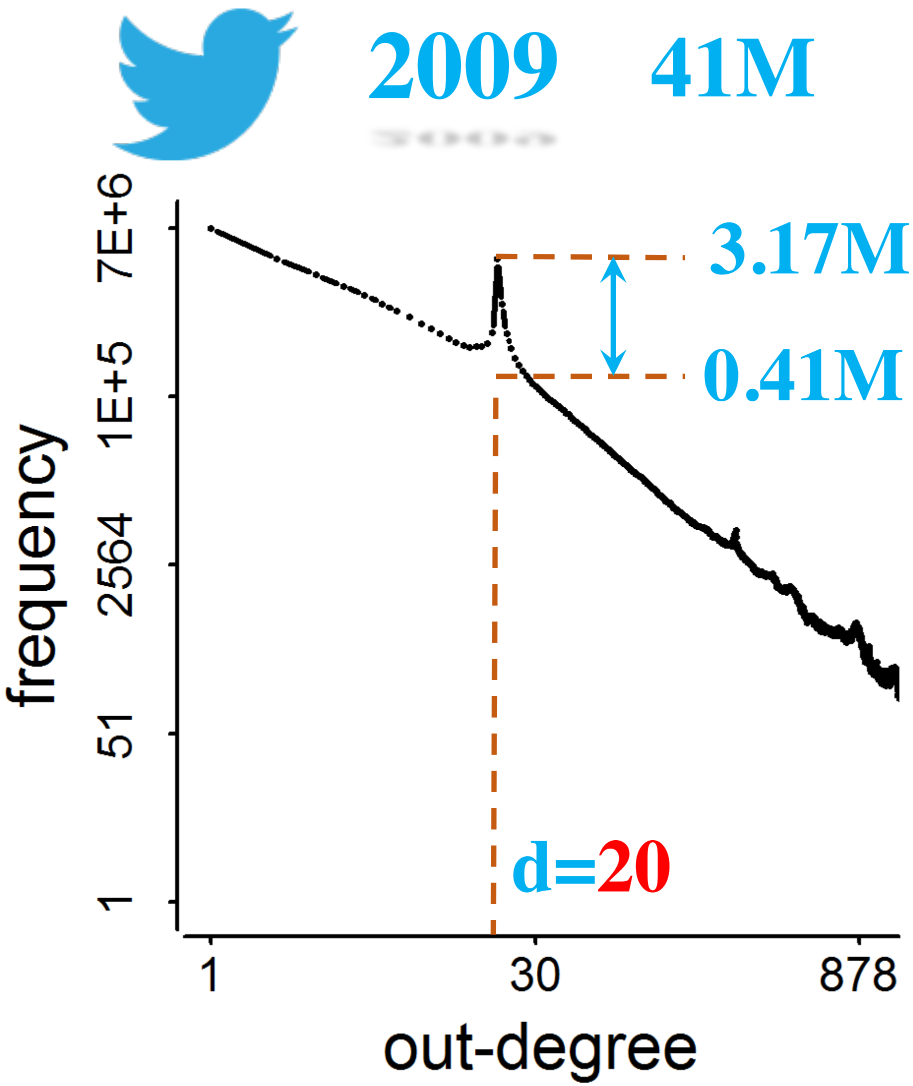
</td><td>


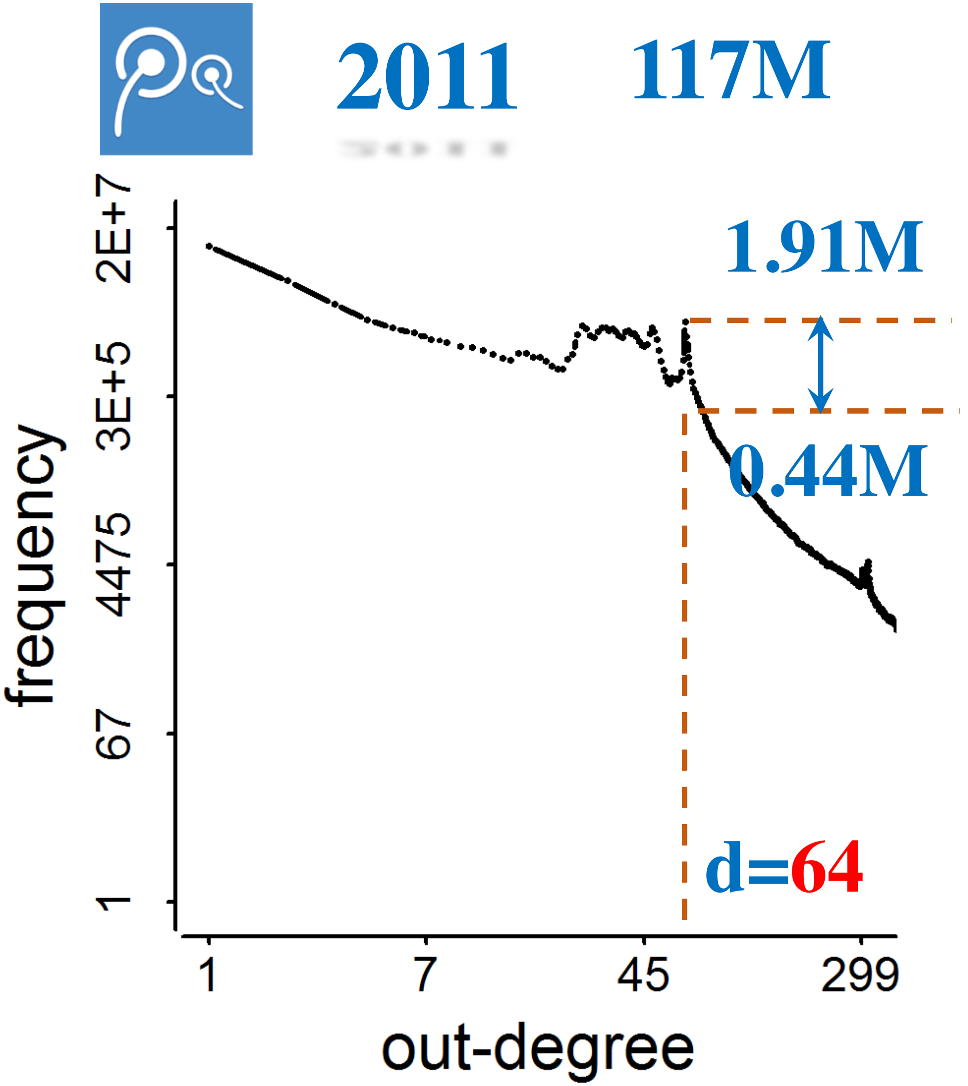
</td><td>


</td></tr>
</table>

[konect.uni-koblenz.de/networks/]

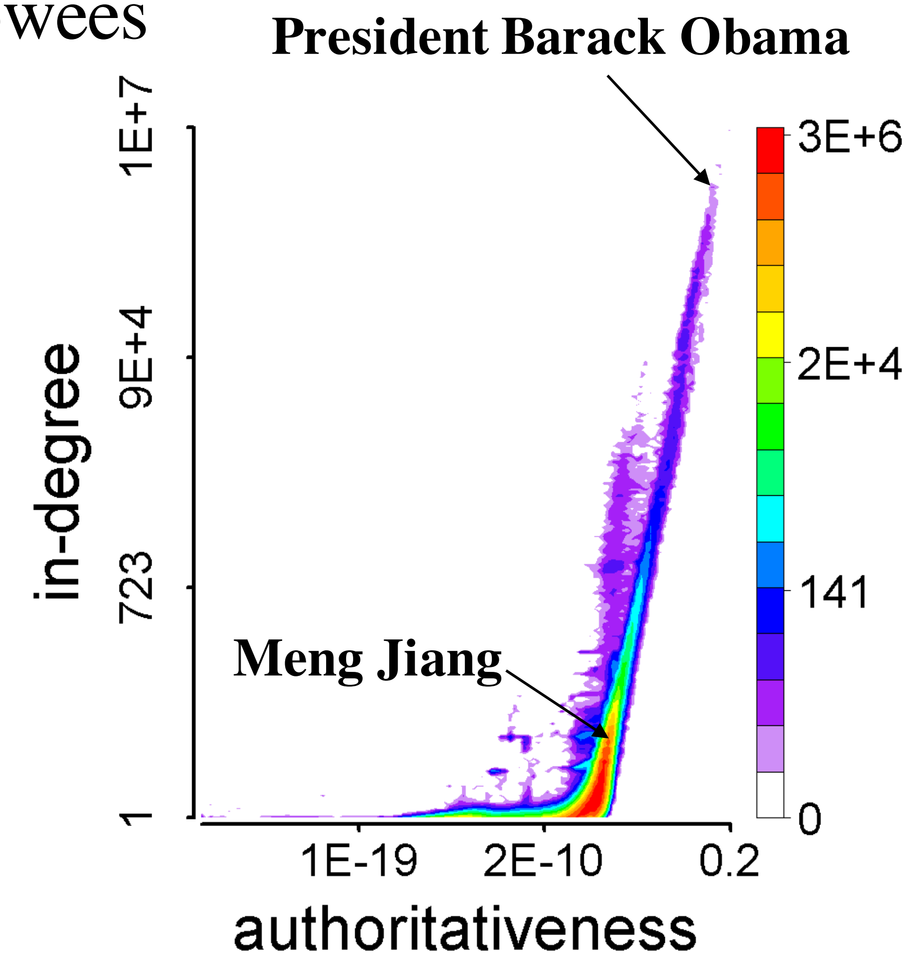
Spikes!


2009 41M
frequency
7E+6
1E+5
2564
51
1
3.17M
0.41M
d=20
1
30
878
out-degree


2011 117M
frequency
2E+7
3E+5
4475
67
1
1.91M
0.44M
d=64
1
7
45
299
out-degree

Observation: How They Behave

- Feature space of followees [Kleinberg. JACM]



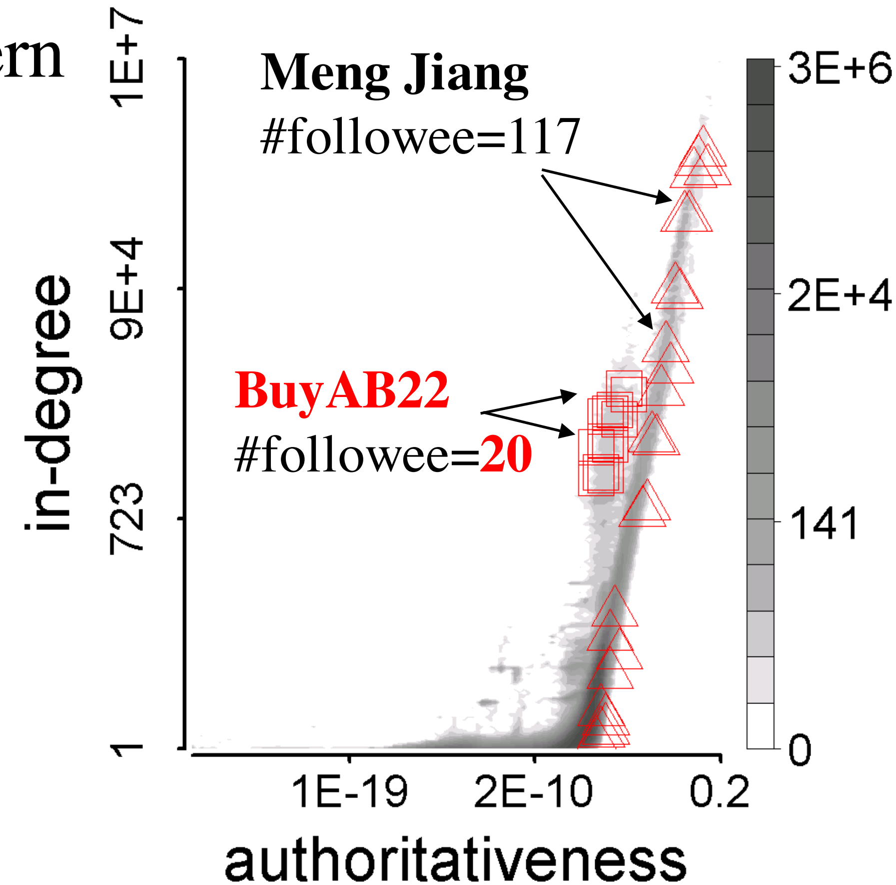
Observation: How They Behave

- Who are their followees?
- Their behavioral pattern
  - **Synchronized**

Similar with each other

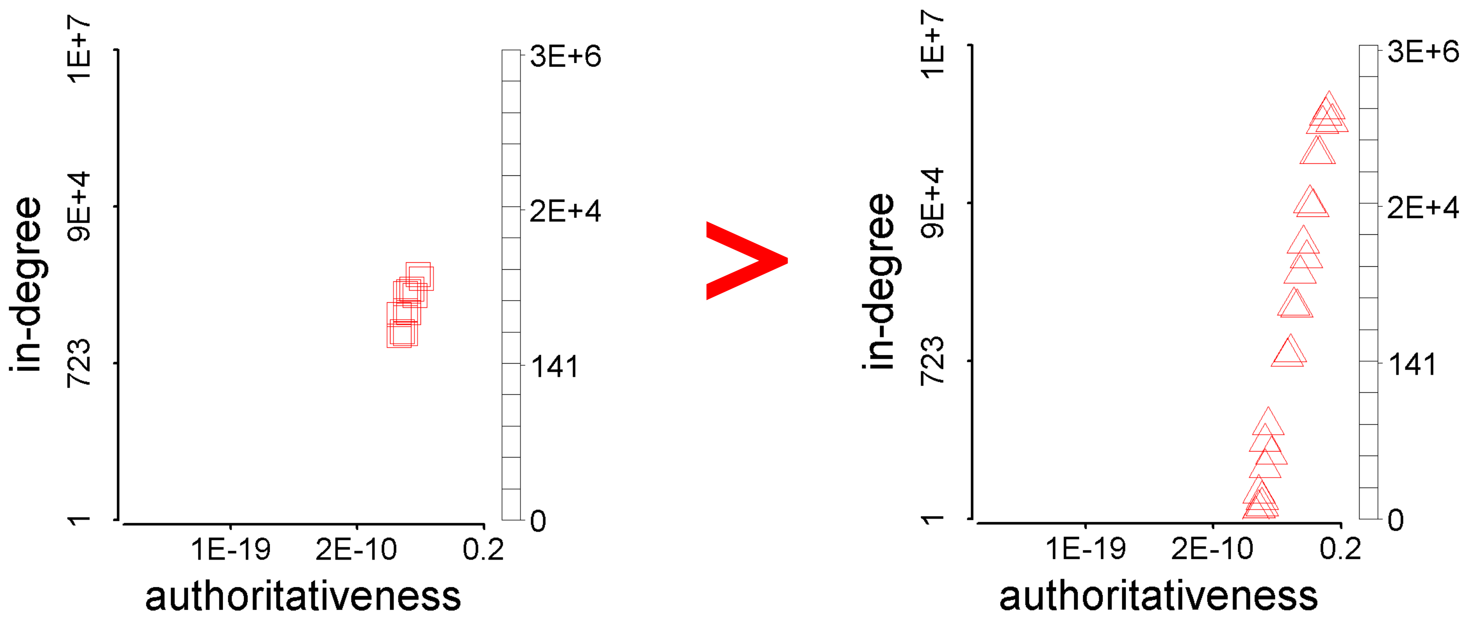
  - **Abnormal**

Different from the majority



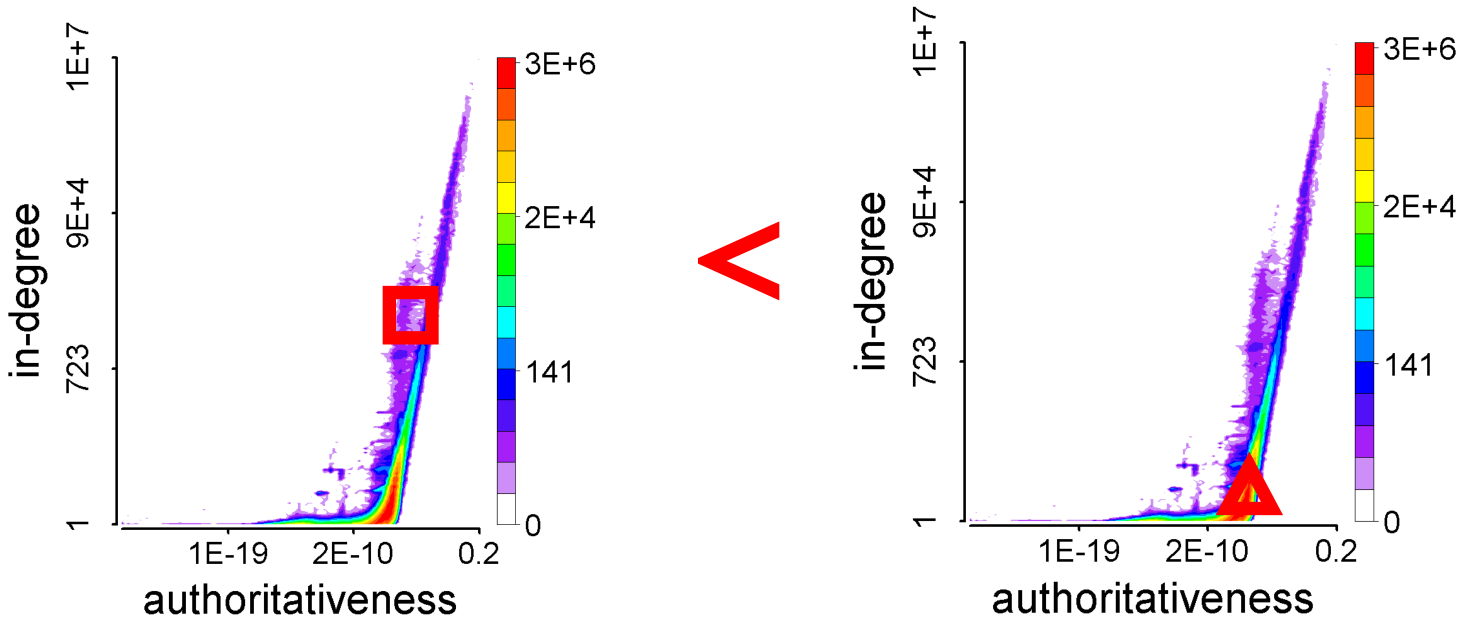
Represent Synchronicity

$$sync(u) = \frac{\sum_{(v,v')\in\mathcal{F}(u)\times\mathcal{F}(u)} \mathbf{p}(v)\cdot\mathbf{p}(v')}{d(u)\times d(u)}$$

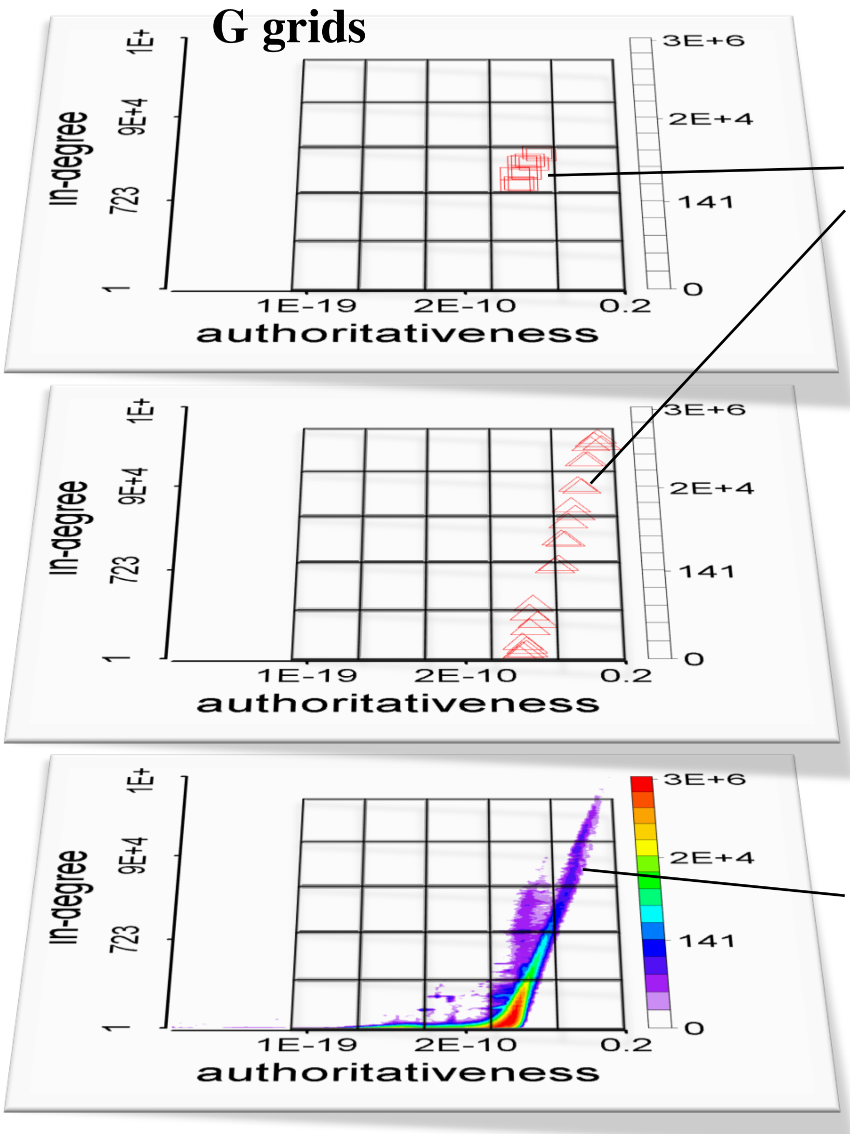


Represent Normality

$$norm(u) = \frac{\sum_{(v,v')\in\mathcal{F}(u)\times\mathcal{U}} \mathbf{p}(v)\cdot\mathbf{p}(v')}{d(u)\times N}$$



Theorem: Synchronicity vs. Normality



fp_g: #foreground points in grid **g**

$\sum fp_g = F = d(u)$ (#followees of **u**)

Given normality

$$n = \sum(fp_g/F)\,(bp_g/B) = \sum f_g b_g,$$

find minimal synchronicity

$$s = \sum(fp_g/F)\,(fp_g/F) = \sum f_g^2$$

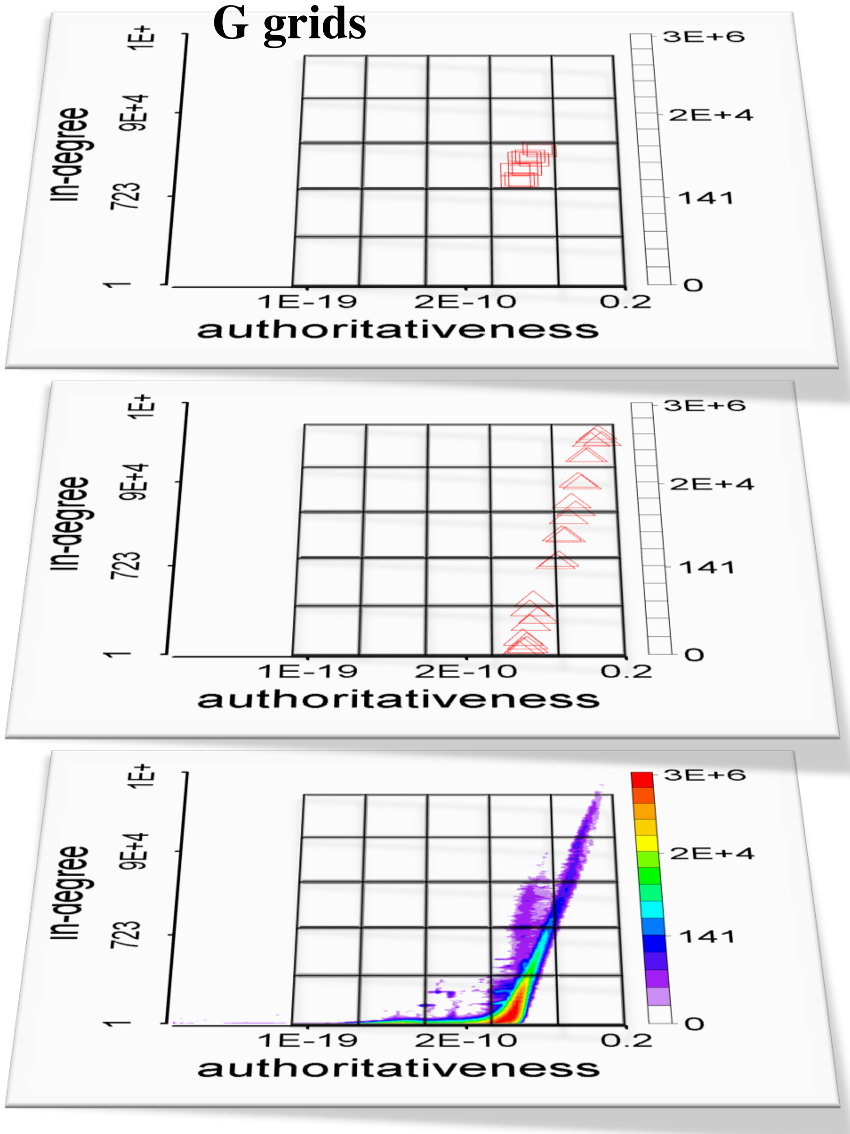
where

$$\sum f_g = 1, \sum b_g = 1$$

bp_g: #background points in grid **g**

$\sum bp_g = B = N$ (#all users)

Theorem: Synchronicity vs. Normality



Solution.

Lagrange multiplier:

minimize $s(f_g) = \sum f_g^{\,2}$

subject to $\sum f_g = 1, \sum f_g b_g = n$

Lagrange function:

$F(f_g, \lambda, \mu) = (\sum f_g^{\,2}) + \lambda(\sum f_g - 1) + \mu(\sum f_g b_g - n)$

Gradients:

$$\begin{cases} \nabla_{f_g} F = 2 f_g + \lambda + \mu\, b_g = 0 \\ \nabla_{\lambda} F = \sum f_g - 1 = 0 \\ \nabla_{\mu} F = \sum f_g b_g - n = 0 \end{cases}$$

$$\begin{cases} 2 + \lambda\, G + \mu = 0 \xleftarrow{\ \Sigma\ } \\ 2\, n + \lambda + \mu\, s_b = 0 \xleftarrow{\ \times b_g\ \Sigma\ } \\ 2\, s_{min} + \lambda + \mu\, n = 0 \xleftarrow{\ \times f_g\ \Sigma\ } \end{cases}$$

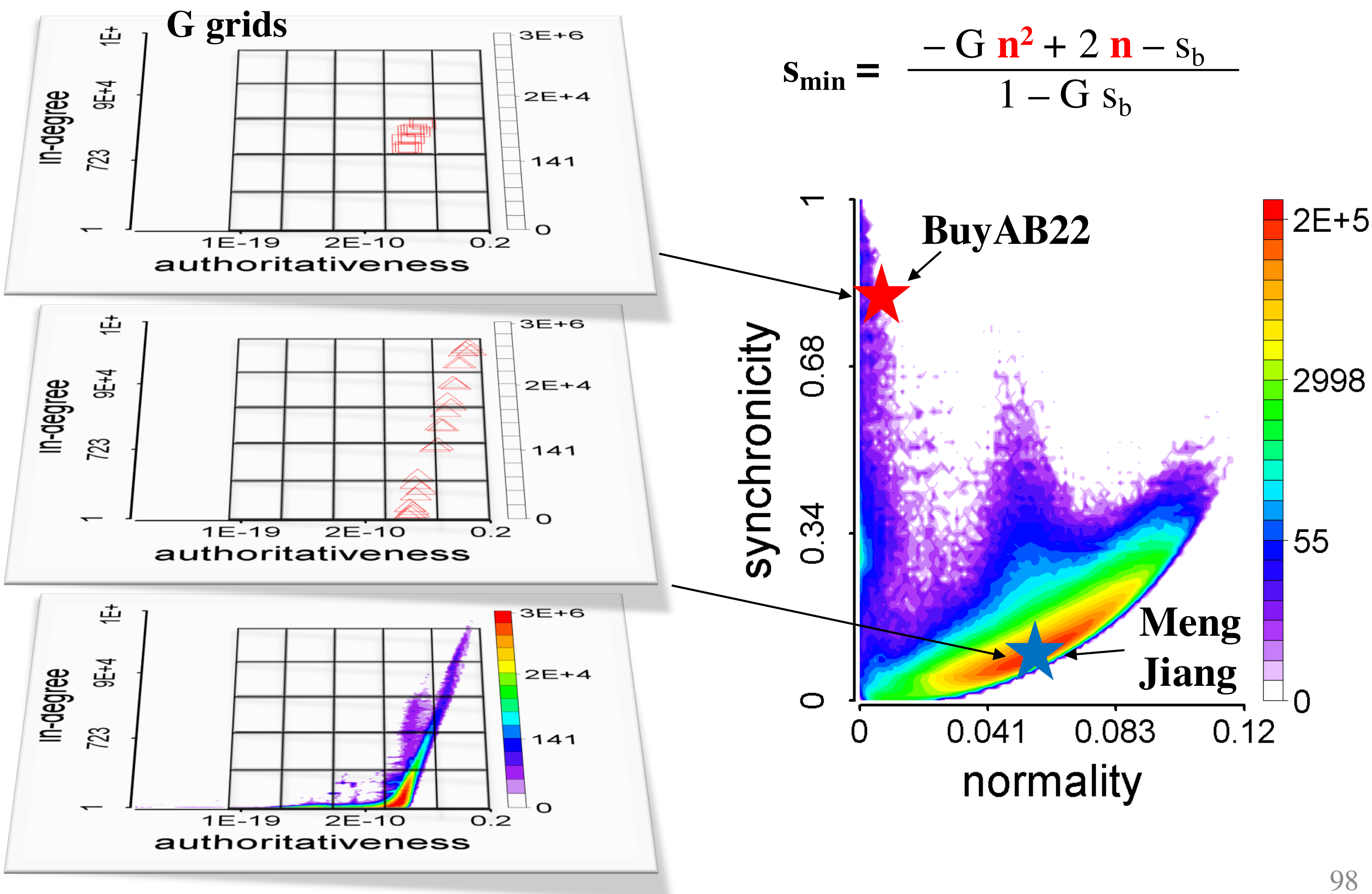
where $s_b = \sum b_g^{\,2}$.

Therefore,

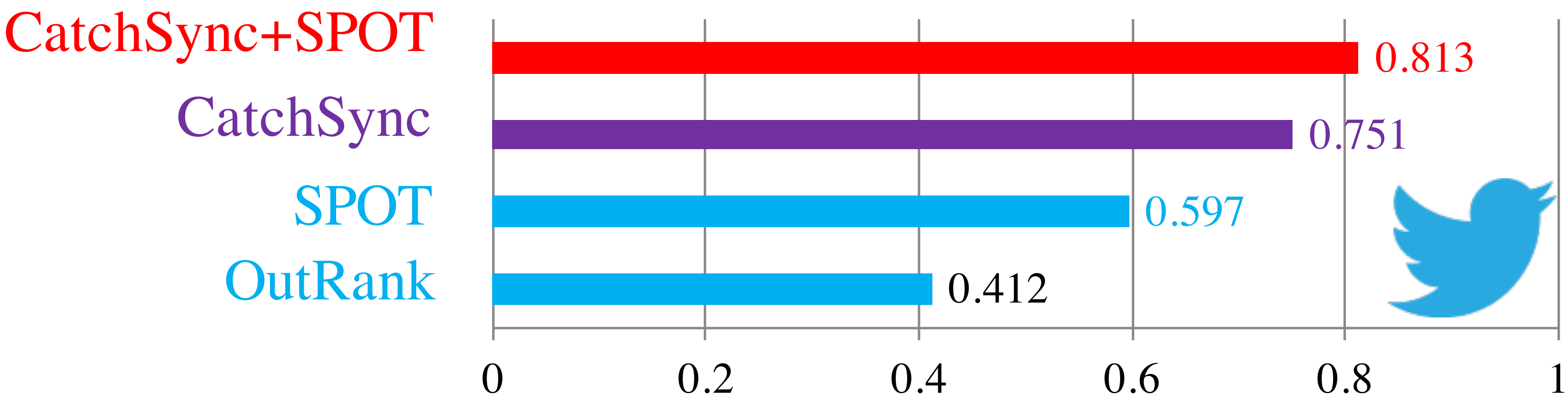
$$\mathbf{s_{min}} = \frac{-\,G\,\mathbf{n^2} + 2\,\mathbf{n} - s_b}{1 - G\, s_b}$$

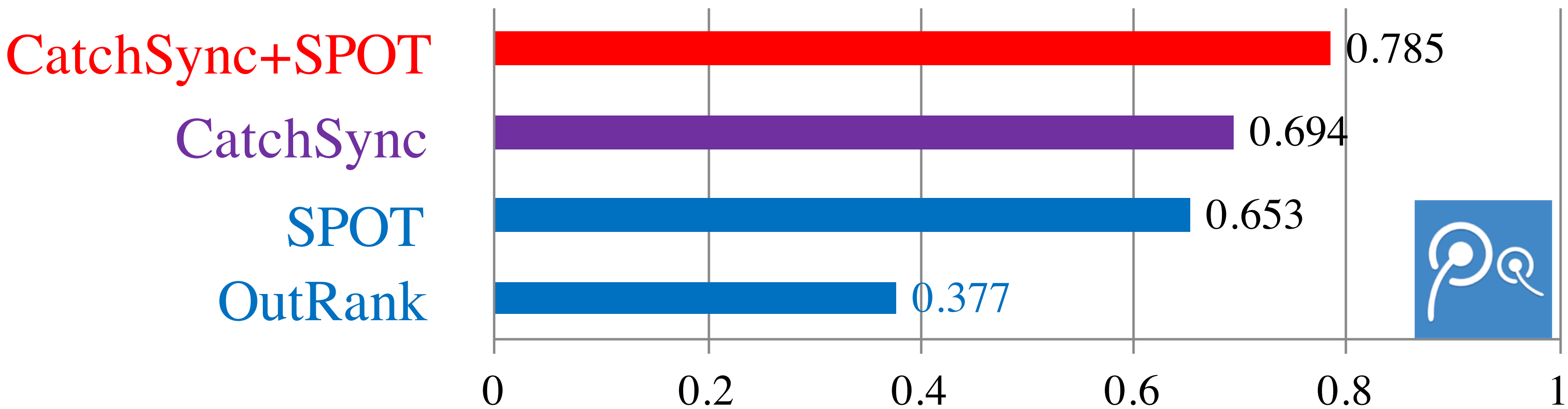
CatchSync Algorithm

$$s_{min} = \frac{-G\,\mathbf{n}^2 + 2\,\mathbf{n} - s_b}{1 - G\,s_b}$$

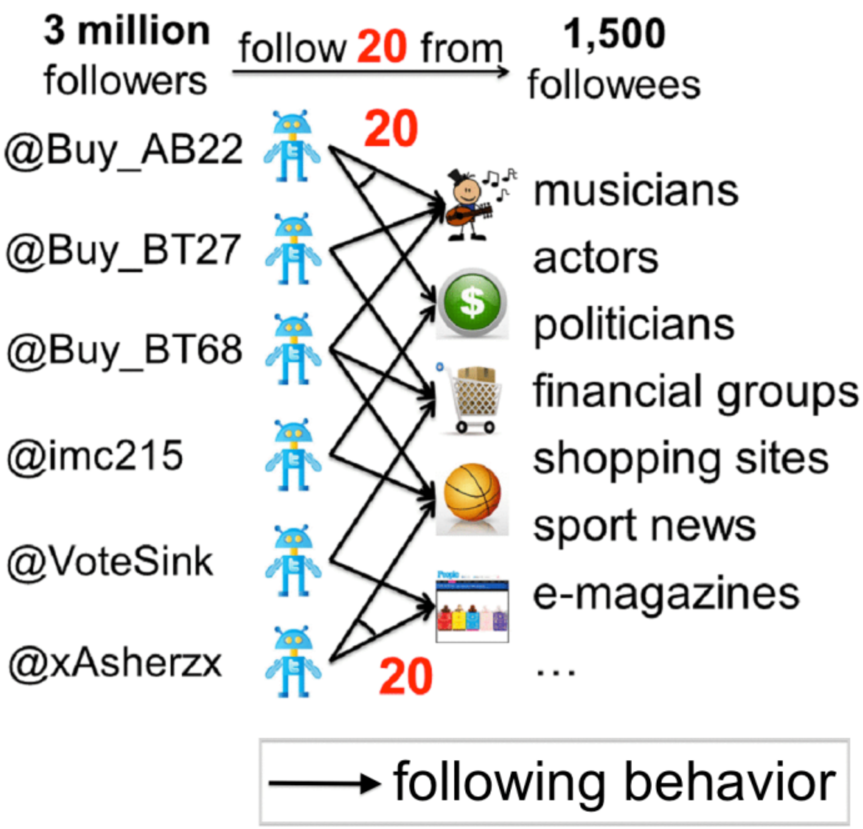


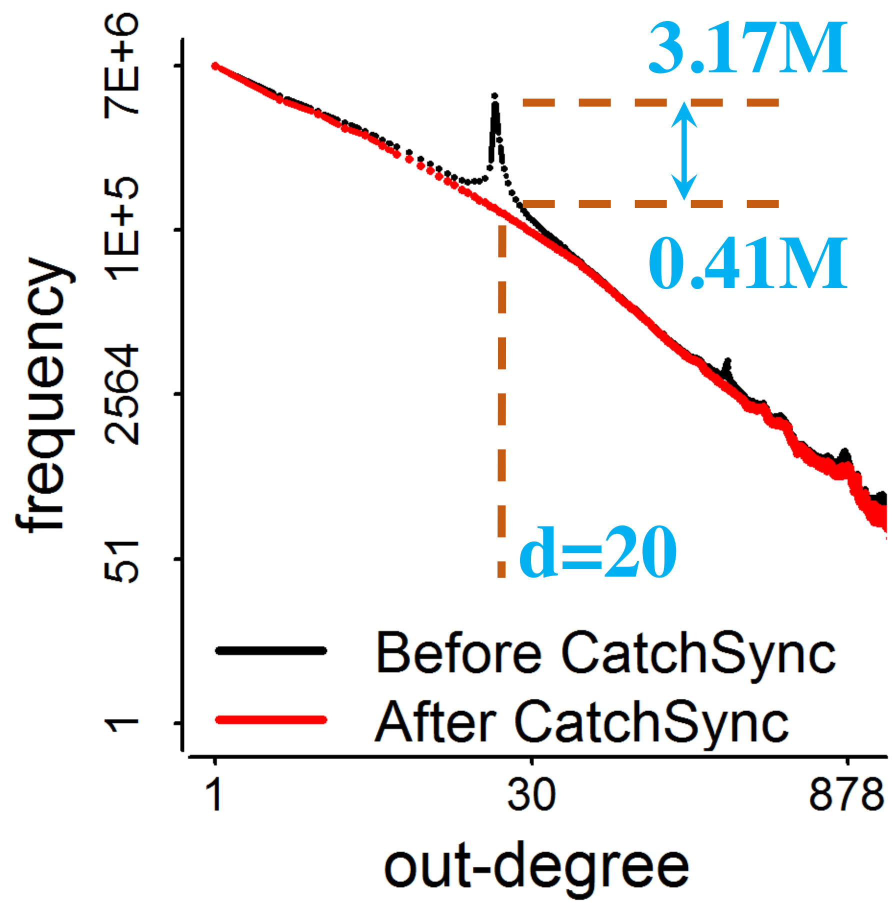
Experimental Results





Experimental Results





Impact

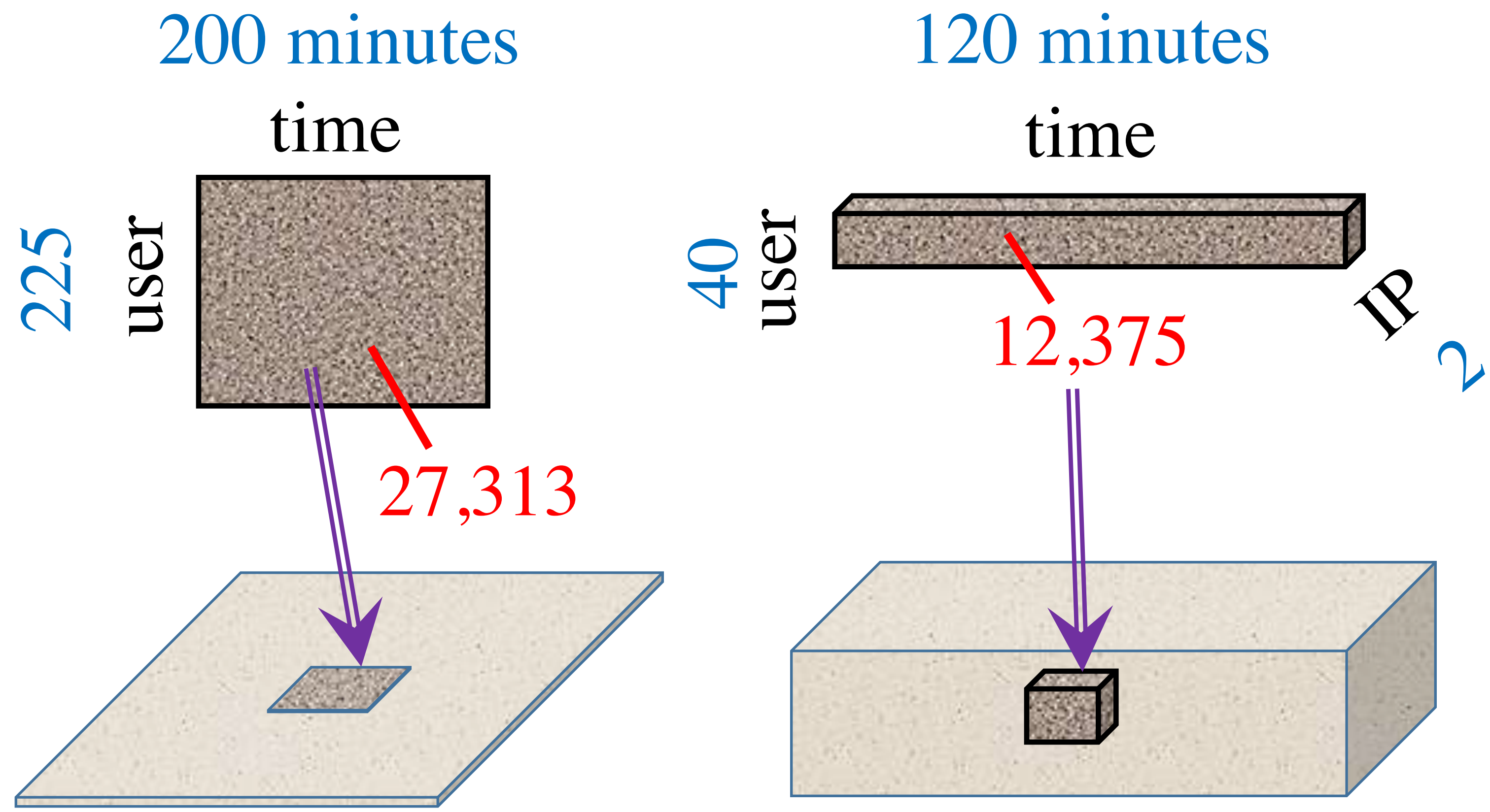
- **M. Jiang**, P. Cui, A. Beutel, C. Faloutsos and S. Yang. "CatchSync: Catching Synchronized Behavior in Large Directed Graphs" in **KDD'14 Best Paper Finalist**, Aug 2014. (#citations = **36**)
- Taught in
  - **CMU 15-826:** Multimedia Databases and Data Mining
  - **UMich EECS 598:** Graph Mining and Exploration at Scale
  - **ASONAM'16 Tutorial:** "Identifying Malicious Actors on Social Media" by S. Kumar, F. Spezzano, V.S. Subrahmanian
- Deployed in Weibo? Unfortunately, in July 2014…

Observation: Spatiotemporal Contexts

Dataset	Dimension/Mode				Mass
Weibo's Retweeting	User	Root ID	IP	Time (min)	#retweet
	29.5M	19.8M	27.8M	56.9K	211.7M
Weibo's Trending (Hashtag)	User	Hashtag	IP	Time (min)	#tweet
	81.2M	1.6M	47.7M	56.9K	276.9M
Network attacks (LBNL)	Src-IP	Dest-IP	Port	Time (sec)	#packet
	2,345	2,355	6,055	3,610	230,836

Jiang et al. **A General Suspiciousness Metric for Dense Blocks in Multimodal Data.** *ICDM*, 2015.

Dense Block Indicates Suspiciousness

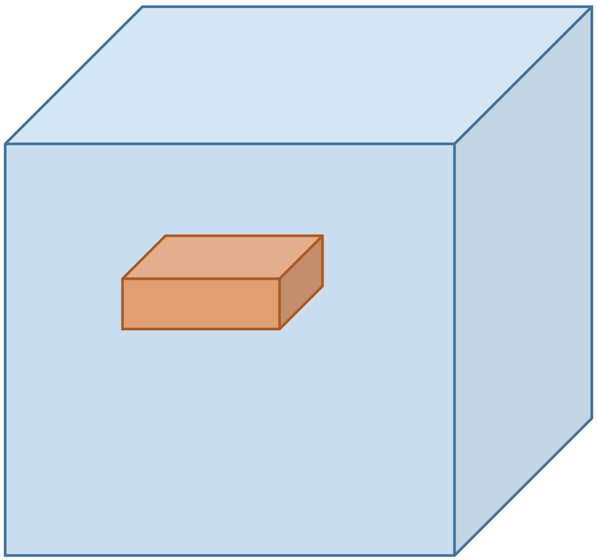


Q: Which is more suspicious?

We need a metric to evaluate the suspiciousness.

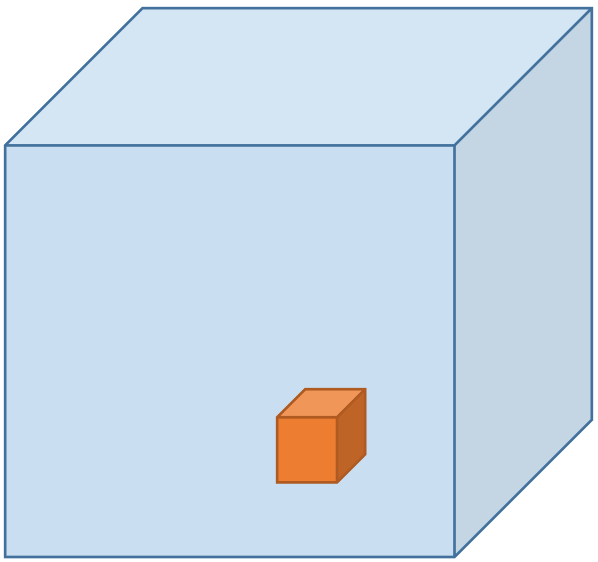
Criteria for Suspiciousness Metric

What properties are required of a good metric?



$N_1 \times N_2 \times N_3$

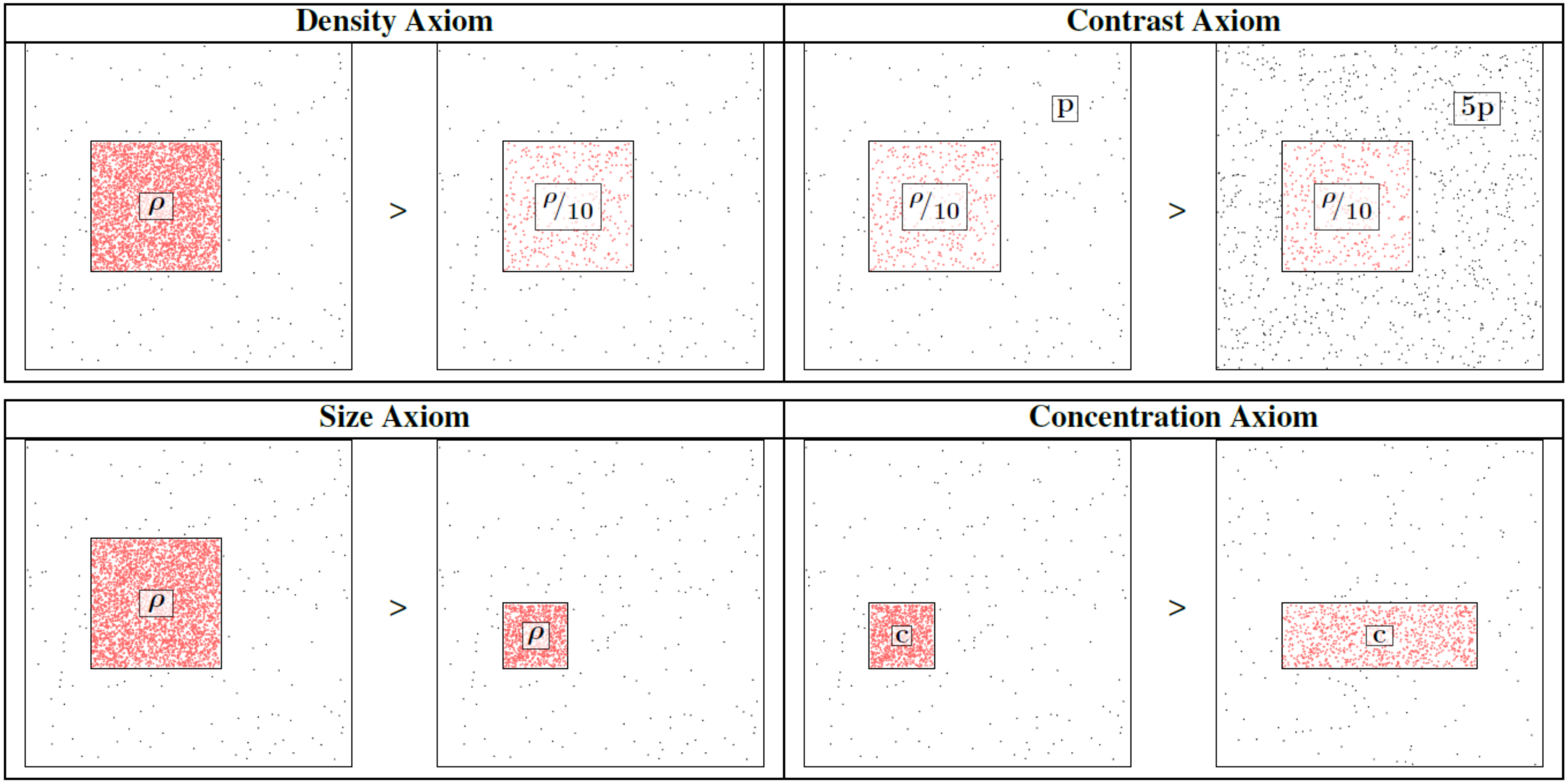
Count data with total “mass” C



$$f\left(\begin{matrix} n_1 \times n_2 \times n_3 \\ \text{mass } c \\ \text{density } \varrho \end{matrix}\right) \quad \text{vs} \quad f\left(\begin{matrix} n'_1 \times n'_2 \times n'_3 \\ \text{mass } c' \\ \text{density } \varrho' \end{matrix}\right)$$

Axioms: 1 to 4

$$c_1 > c_2 \iff f(\mathbf{n}, c_1, \mathbf{N}, C) > f(\mathbf{n}, c_2, \mathbf{N}, C)$$

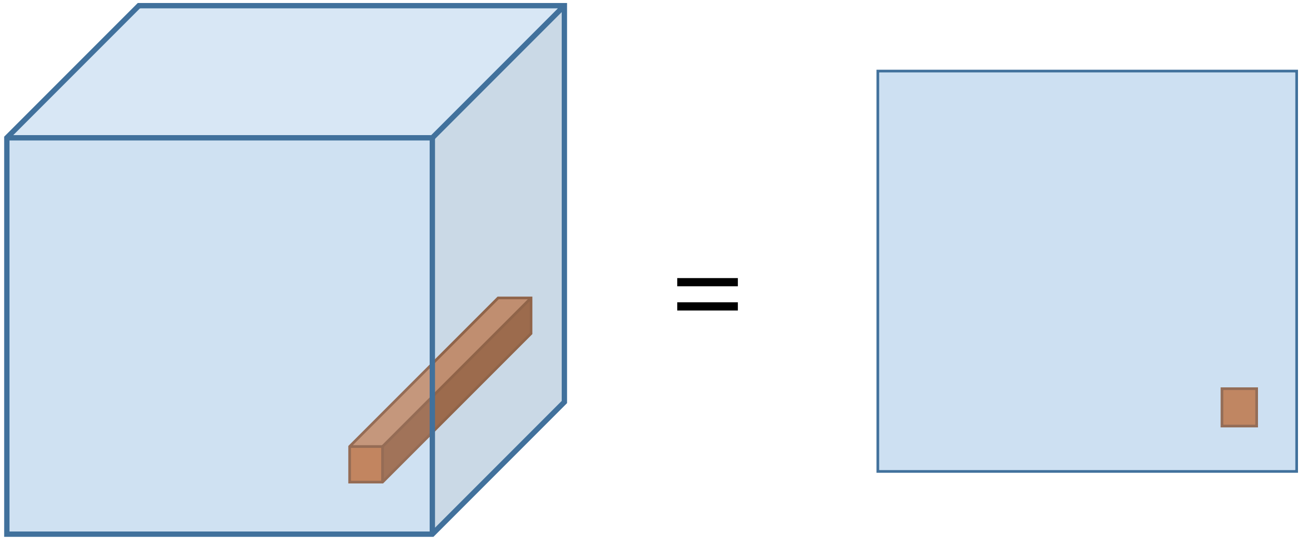


$$p_1 < p_2 \iff \hat{f}(\mathbf{n}, \rho, \mathbf{N}, p_1) > \hat{f}(\mathbf{n}, \rho, \mathbf{N}, p_2)$$

Axiom 5: Cross Dimensions

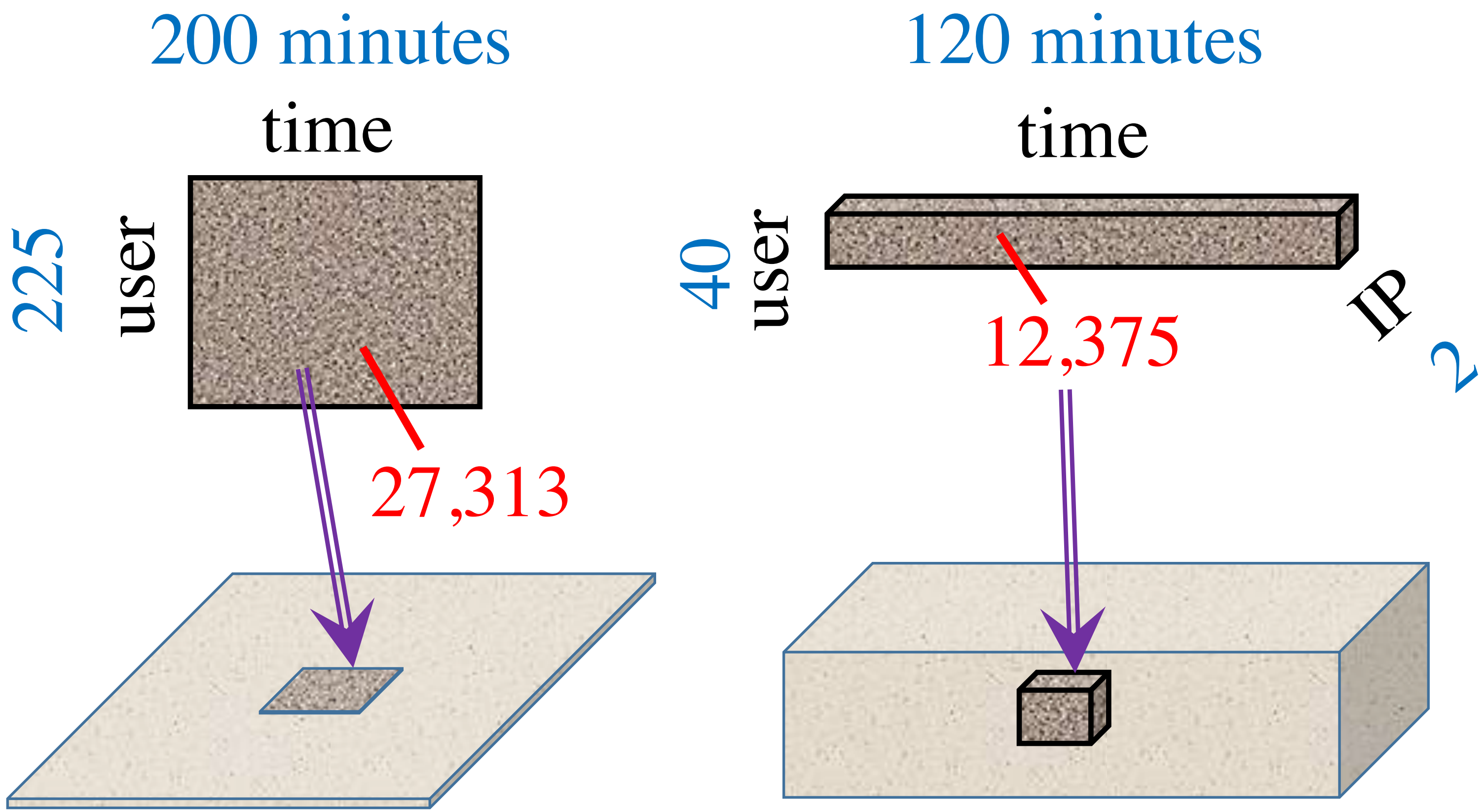
$$f_{K-1}\left([n_k]_{k=1}^{K-1}, c, [N_k]_{k=1}^{K-1}, C\right) = f_K\left(([n_k]_{k=1}^{K-1}, N_K), c, [N_k]_{k=1}^{K}, C\right)$$

Not including a mode is the same as including all values for that mode.



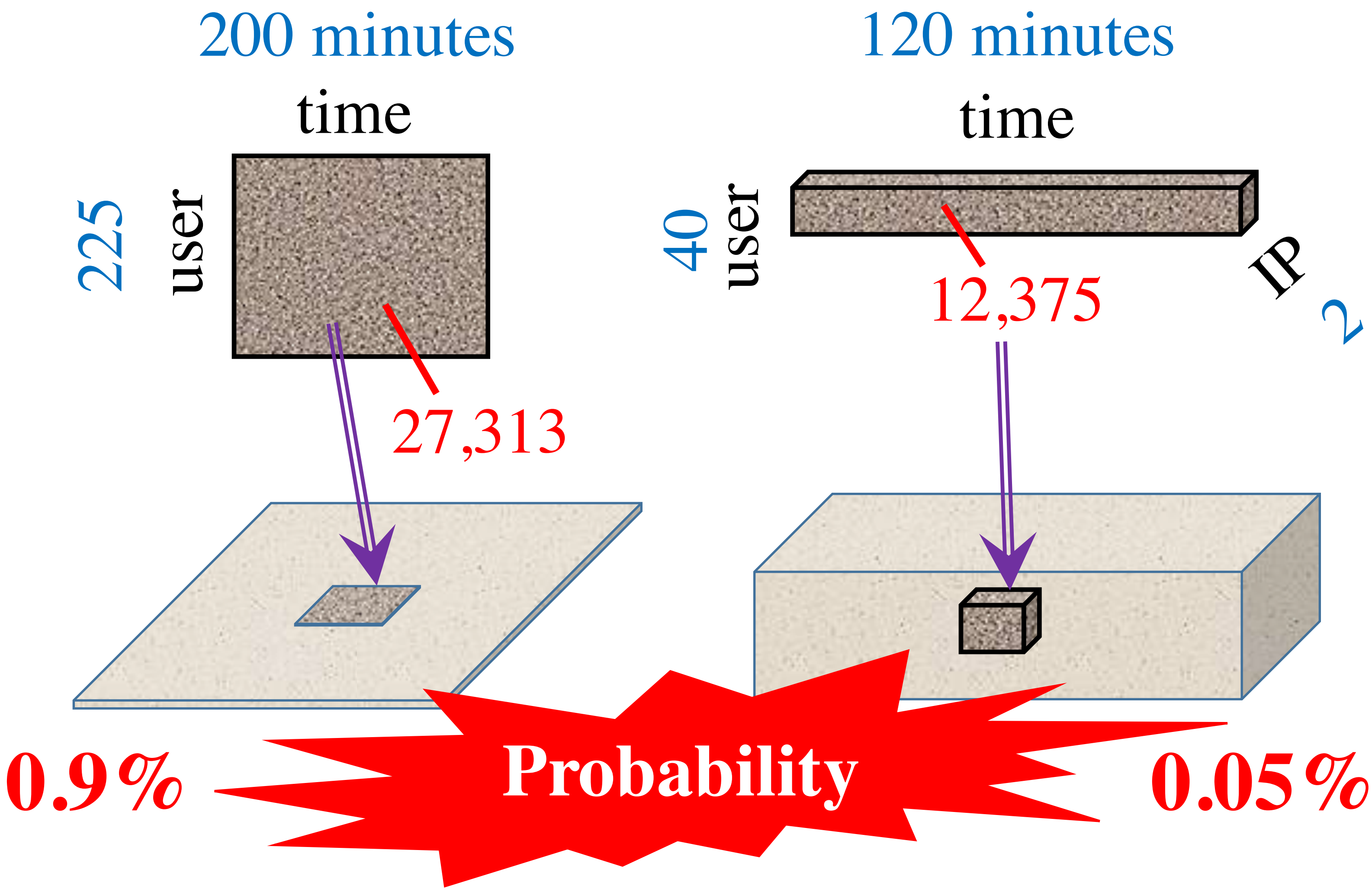
- New information (more modes) can only make our blocks more suspicious

Scoring the Suspiciousness



Q: Which is more suspicious?

Scoring the Suspiciousness



A General Suspiciousness Metric

- Negative log likelihood of block's probability

$$f(n,c,N,C) = -\log\left[Pr(Y_n = c)\right]$$

Lemma *Given an $n_1 \times \cdots \times n_K$ block of mass c in $N_1 \times \cdots \times N_K$ data of total mass C, the suspiciousness function is*

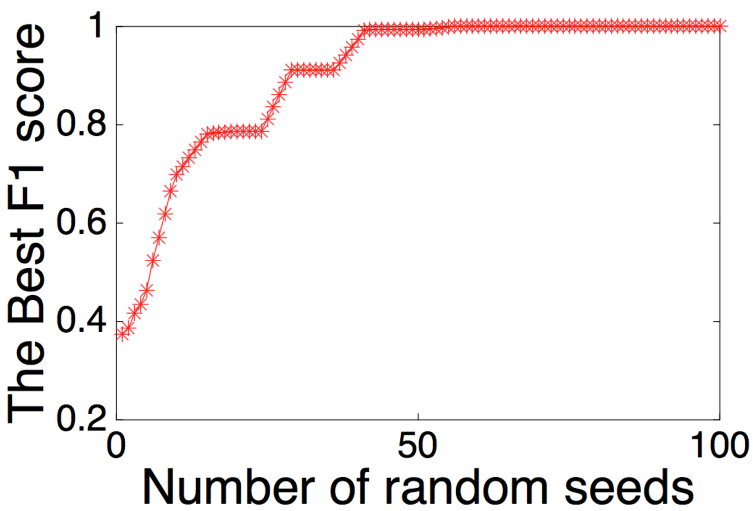
$$\boxed{f(\mathbf{n},c,\mathbf{N},C) = c(\log\frac{c}{C} - 1) + C\prod_{i=1}^{K}\frac{n_i}{N_i} - c\sum_{i=1}^{K}\log\frac{n_i}{N_i}}$$

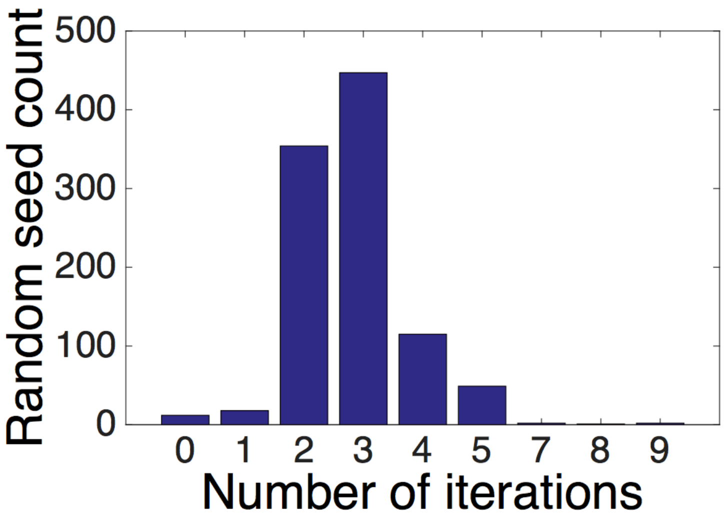
Using ρ as the block's density and p is the data's density, we have the simpler formulation

$$\boxed{\hat{f}(\mathbf{n},\rho,\mathbf{N},p) = \left(\prod_{i=1}^{K} n_i\right) D_{KL}(\rho||p)}$$

CrossSpot Algorithm

- Local search to maximize the metric
  - Start with seed blocks
  - Parameter-free: iteratively update the blocks
  - Scalable: parallelize to multiple machines





Advantages

	Method	Scores Blocks	Axioms: Density 1	Size 2	Concentration 3	Contrast 4	Multi-modal 5
Metrics	**SUSPICIOUSNESS**	✔	✔	✔	✔	✔	✔
	Mass	✓	✓	✗	✗	✗	✓
	Density	✓	✓	✗	✓	✗	✗
	Average Degree [9]	✓	✓	✗	✗	✗	N/A
	Singular Value [10]	✓	✓	✓	✓	✗	✗
Methods	**CROSSSPOT**	✔	✔	✔	✔	✔	✔
	Subgraph [30, 10, 36]	✓	✓	✓	✓	✗	N/A
	CopyCatch [6]	✓	✓	✓	✓	✗	N/A
	EigenSpokes [31]	✗	N/A				
	TrustRank [14, 8]	✗	N/A				
	BP [28, 1]	✗	N/A				

Results: Dense Block Detection

- Synthetic data
  - 1,000×1,000×1,000 of 10,000 random data
  - Block#1: 30×30×30 of 512 3 modes
  - Block#2: 30×30×1,000 of 512 2 modes
  - Block#3: 30×1,000×30 of 512 2 modes
  - Block#4: 1,000×30×30 of 512 2 modes

	Recall				Overall Evaluation		
	Block #1	Block #2	Block #3	Block #4	Precision	Recall	F1 score
HOSVD (r=20)	93.7%	29.5%	23.7%	21.3%	**0.983**	0.407	0.576
HOSVD (r=10)	91.3%	24.4%	18.5%	19.2%	0.972	0.317	0.478
HOSVD (r=5)	85.7%	10.0%	9.5%	11.4%	0.952	0.195	0.324
CROSSSPOT	**100%**	**99.9%**	**94.9%**	**95.4%**	0.978	**0.967**	**0.972**

Results: Tweeting Hashtags

User$\times$hashtag$\times$IP$\times$minute	Mass c	Suspiciousness
$582\times3\times294\times$**56,940**	5,941,821	111,799,948
$188\times1\times313\times$**56,943**	2,344,614	47,013,868
$75\times1\times2\times2{,}061$	689,179	19,378,403

User ID	Time	IP address (city, province)	Tweet text with hashtag
USER-D	11-18 12:12:51	IP-1 (Deyang, Shandong)	**#Snow#** the Samsung GALAXY SII QQ Service customized version...
USER-E	11-18 12:12:53	IP-1 (Deyang, Shandong)	**#Snow#** the Samsung GALAXY SII QQ Service customized version...
USER-F	11-18 12:12:54	IP-2 (Zaozhuang, Shandong)	**#Snow#** the Samsung GALAXY SII QQ Service customized version...
USER-E	11-18 12:17:55	IP-1 (Deyang, Shandong)	**#Li Ning - a weapon with a hero#** good support activities!
USER-F	11-18 12:17:56	IP-2 (Zaozhuang, Shandong)	**#Li Ning - a weapon with a hero#** good support activities!
USER-D	11-18 12:18:40	IP-1 (Deyang, Shandong)	**#Toshiba Bright Daren#** color personality test to find out your sense...
USER-E	11-18 17:00:31	IP-2 (Zaozhuang, Shandong)	**#Snow#** the Samsung GALAXY SII QQ Service customized version...
USER-D	11-18 17:00:49	IP-2 (Zaozhuang, Shandong)	**#Toshiba Bright Daren#** color personality test to find out your sense...
USER-F	11-18 17:00:56	IP-2 (Zaozhuang, Shandong)	**#Li Ning - a weapon with a hero#** good support activities!

Results: Network Attacks

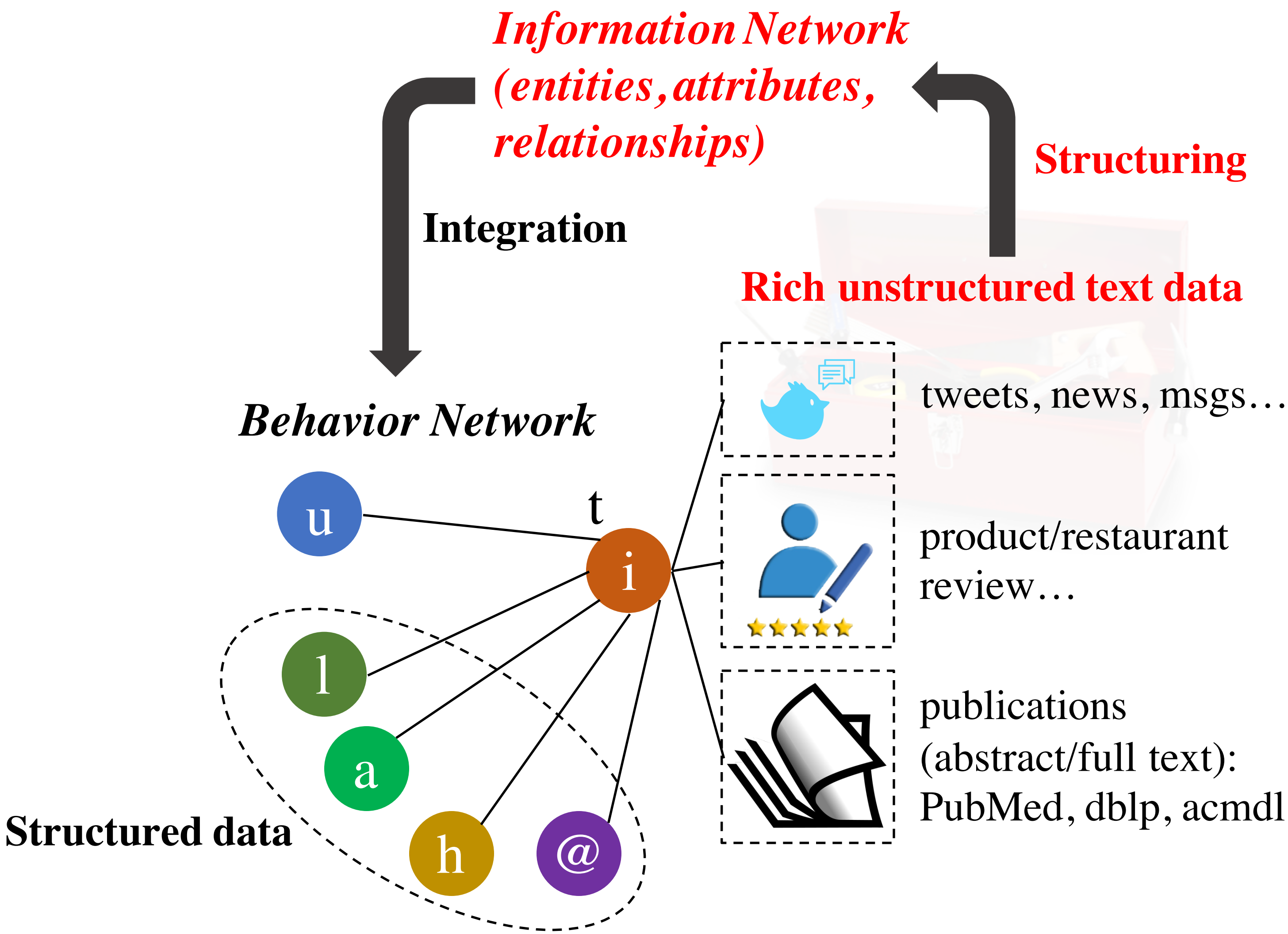
	#	Src-IP×dst-IP×port×second	Mass c	Suspiciousness
CrossSpot	1	411×9×6×**3,610**	47,449	552,465
	2	533×6×1×**3,610**	30,476	400,391
	3	5×5×2×**3,610**	18,881	317,529
	4	11×7×7×**3,610**	20,382	295,869
HOSVD	1	15×1×1×1,336	4,579	80,585
	2	1×2×2×1,035	1,035	18,308
	3	1×1×1×1,825	1,825	34,812
	4	1×13×6×181	1,722	29,224

Summary

- Ill-gotten Facebook Likes, Zombie Followers
- **Observations, Representations, Models**
  - **CopyCatch:** Catching ill-gotten Likes by core search
  - **LockInfer:** Adding seed selection before search
  - **CatchSync:** Catching smart zombie followers with high recall (recovering power-law distributions)
  - **CrossSpot:** Defining suspiciousness across dimensions

II. Structuring behavioral content and integrating behavioral analysis with information networks

Data to Network to Knowledge

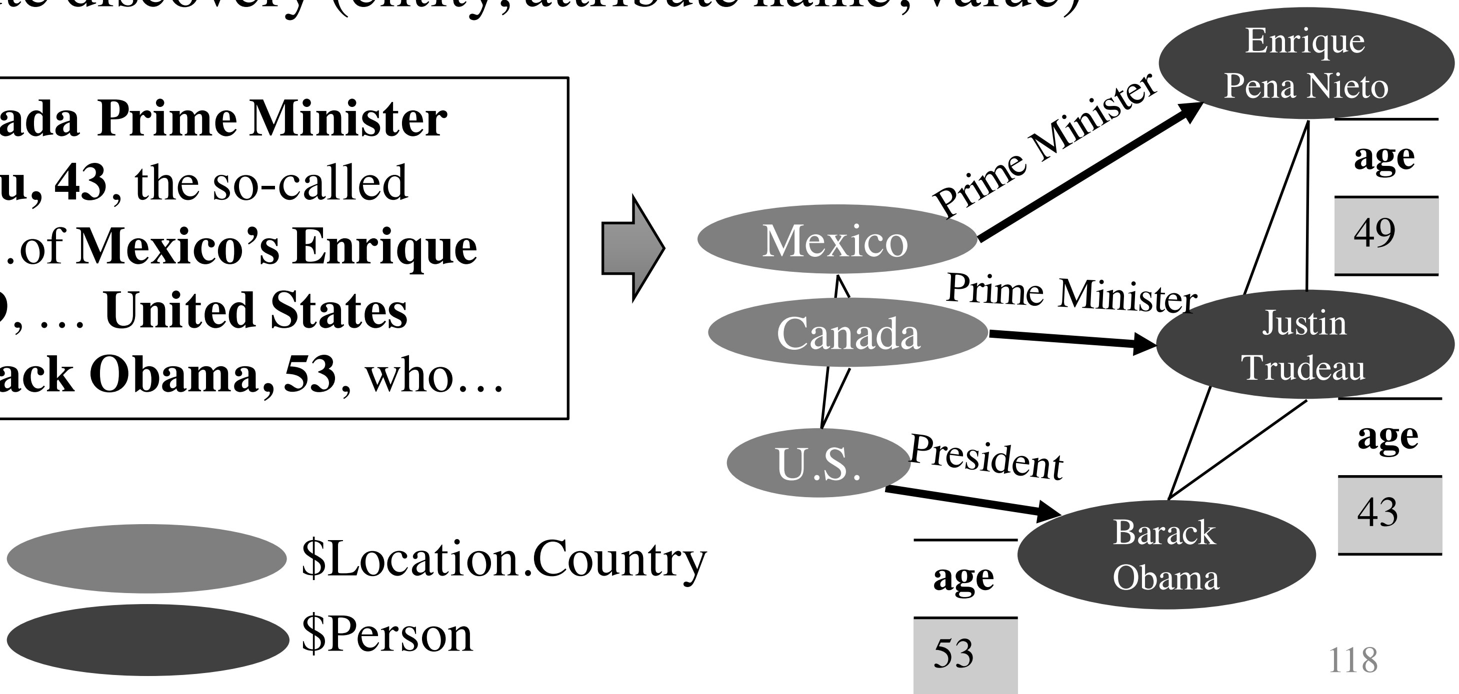


Construction of Heterogeneous Information Networks from Text

ὢPhilosophy: Not extensive "labeling" but exploring the power of massive text corpora!

- Mining phrases (the minimal semantic units)
- Entity recognition and typing
- Attribute discovery (entity, attribute name, value)

…here by **Canada Prime Minister Justin Trudeau, 43**, the so-called #APEChottie…of **Mexico's Enrique Pena Nieto, 49**, … **United States President Barack Obama, 53**, who…

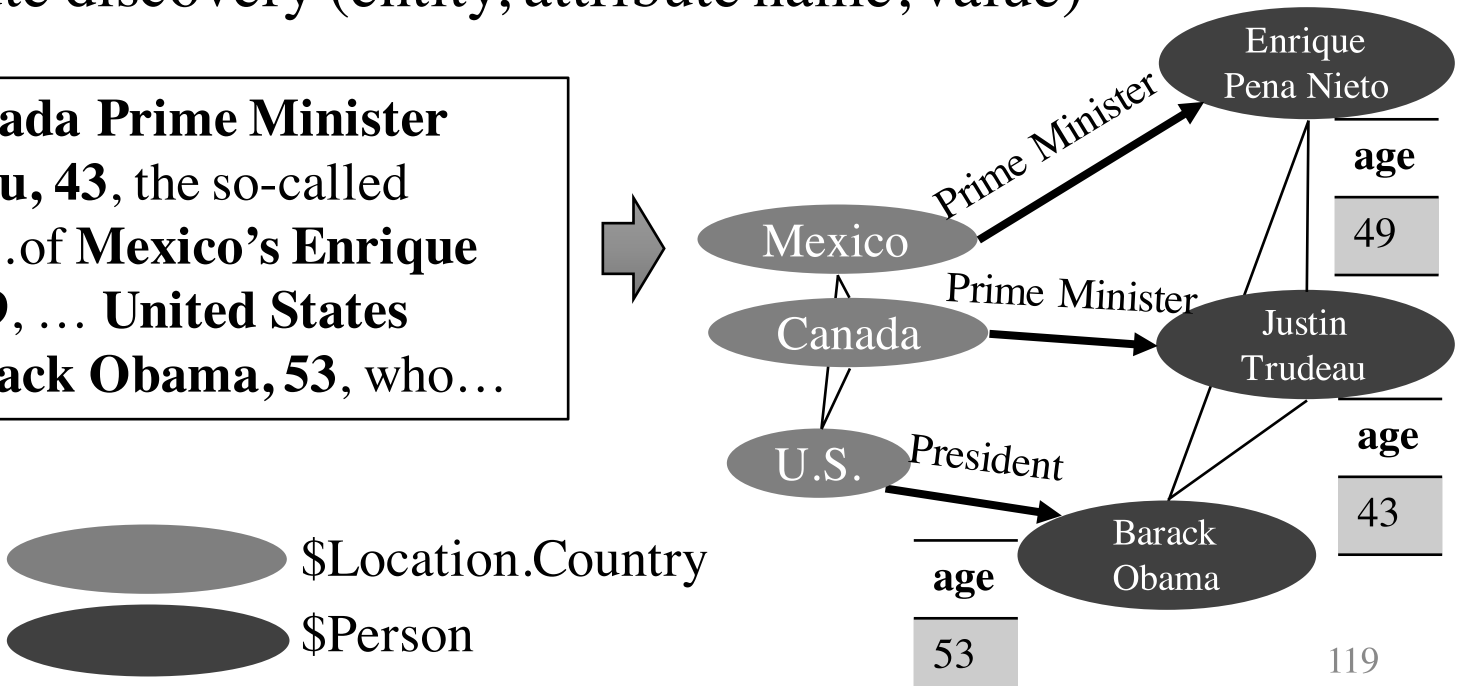


Construction of Heterogeneous Information Networks from Text

ὢPhilosophy: Not extensive "labeling" but exploring the power of massive text corpora!

- ❑**Mining phrases (the minimal semantic units)**
- ❑Entity recognition and typing
- ❑Attribute discovery (entity, attribute name, value)

…here by **Canada Prime Minister Justin Trudeau, 43**, the so-called #APEChottie…of **Mexico's Enrique Pena Nieto, 49**, … **United States President Barack Obama, 53**, who…



Why Mining Phrases?

- **Unigrams** are *ambiguous but* **phrases** are natural, *unambiguous* semantic units
  - Ex.: "United" vs. United States, United Airline, United Parcel Service
- Mining semantically meaningful phrases
  - Transform text data from *word granularity* to *phrase granularity*
  - Enhance the power at manipulating unstructured data using information networks
- Phrase mining: Most NLP methods may need annotation and training
  - Annotate hundreds of documents as training data
  - Train a supervised model based on part-of-speech features
  - Limitations: High annotation cost
  - May not be scalable to domain-specific, dynamic, emerging applications
    - Scientific domains, query logs, or social media, e.g., Yelp, Twitter
- Minimal/no training but making good use of massing corpora

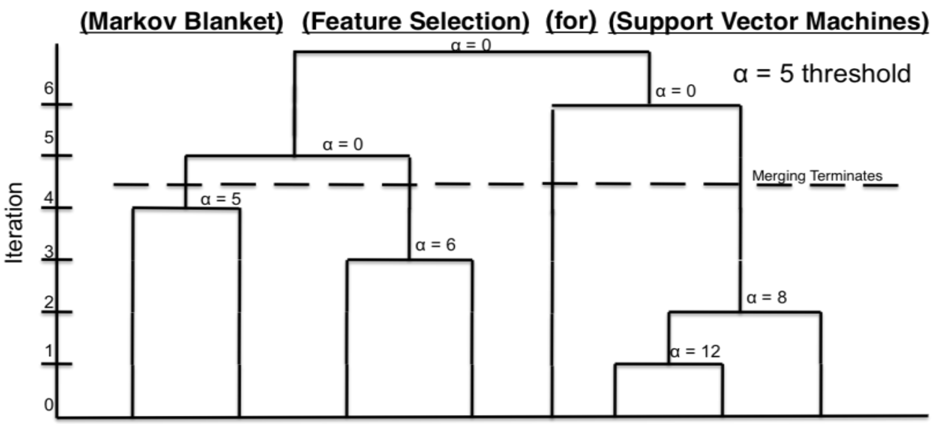
Strategies for Phrase Mining

- Strategy 1: Simultaneously inferring phrases and topics
  - Bigram topical model [Wallach'06], topical n-gram model [Wang, et al.'07], phrase discovering topic model [Lindsey, et al.'12]
  - High model complexity: Tends to overfitting; High inference cost: Slow
- Strategy 2: Post topic modeling phrase construction
  - Label topic [Mei et al.'07], TurboTopic [Blei & Lafferty'09], KERT [Danilevsky, et al.'14]
  - Words in the same phrase may be assigned to different topics
    - Ex. … knowledge discovery using least squares support vector machine …
- **Our solution 1: ToPMine** [El-kishky, et al., VLDB'15]
  - First Phrase Mining then Topic Modeling (No training data at all)
- **Our solution 2: SegPhrase+** [Liu, et al., SIGMOD'15]
  - Integrating phrase mining and document segmentation (with minimal training data)

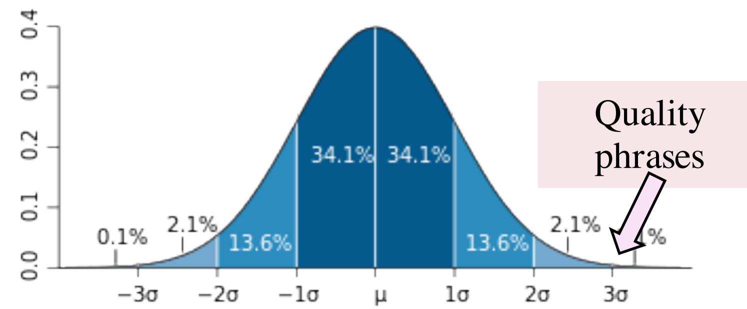
ToPMine: The Overall Phrase Mining Framework

- **ToPMine** [El-Kishky et al. VLDB'15]
  - First phrase construction, then topic mining
  - Contrast with KERT: First topic modeling, then phrase mining
- **The ToPMine Framework:**
  - Perform frequent ***contiguous pattern*** mining to extract candidate phrases and their counts
  - Perform agglomerative merging of adjacent unigrams as guided by a significance score—This segments each document into a ***"bag-of-phrases"***
  - The newly formed bag-of-phrases are passed as input to PhraseLDA, an extension of LDA, that constrains all words in a phrase to each sharing the same latent topic

Phrase Mining: Frequent Pattern Mining + Statistical Analysis



Markov Blanket Feature Selection for Support Vector Machines.



Based on significance score [Church et al.'91]:

$$\alpha(P_1, P_2) \approx (f(P_1 \bullet P_2) - \mu_0(P_1,P_2)) \,/\, f(P_1 \bullet P_2)^{1/2}$$

[Markov blanket] [feature selection] for [support vector machines]
[knowledge discovery] using [least squares] [support vector machine] [classifiers]
…[support vector] for [machine learning]…

Phrase	Raw freq.	True freq.
[support vector machine]	90	80
[vector machine]	95	0
[support vector]	100	20

What Kind of Phrases are of "High Quality"?

- Judging the quality of phrases
  - **Popularity**
    - "information retrieval" vs. "cross-language information retrieval"
  - **Concordance**
    - "powerful tea" vs. "strong tea"
    - "active learning" vs. "learning classification"
  - **Informativeness**
    - "this paper" (frequent but not discriminative, not informative)
  - **Completeness**
    - "vector machine" vs. "support vector machine"

ToPMine: Experiments on Yelp Reviews

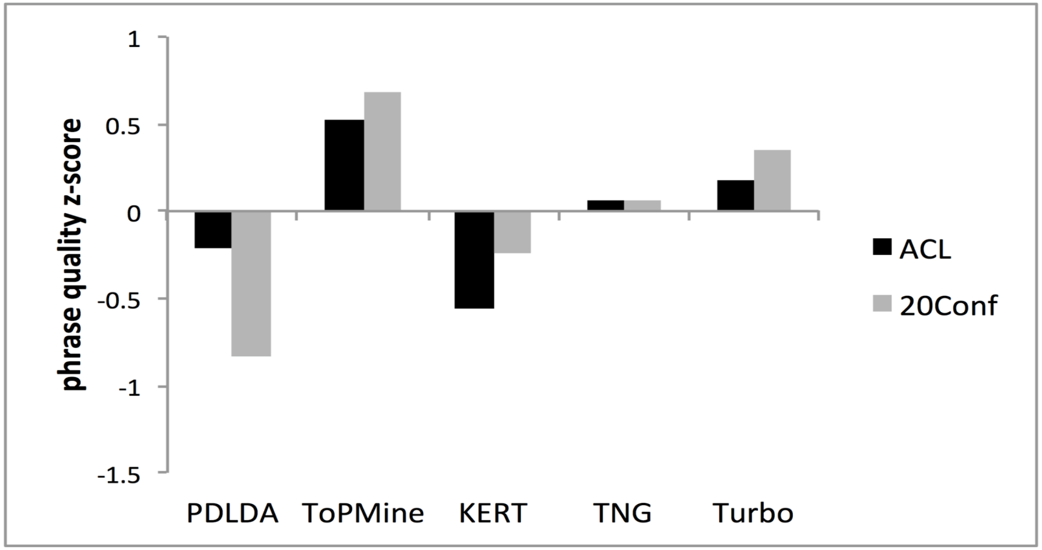
	Topic 1	*Topic 2*	*Topic 3*	*Topic 4*	*Topic 5*
unigrams	coffee	food	room	store	good
	ice	good	parking	shop	food
	cream	place	hotel	prices	place
	flavor	ordered	stay	find	burger
	egg	chicken	time	place	ordered
	chocolate	roll	nice	buy	fries
	breakfast	sushi	place	selection	chicken
	tea	restaurant	great	items	tacos
	cake	dish	area	love	cheese
	sweet	rice	pool	great	time
n-grams	ice cream	spring rolls	parking lot	grocery store	mexican food
	iced tea	food was good	front desk	great selection	chips and salsa
	french toast	fried rice	spring training	farmer's market	food was good
	hash browns	egg rolls	staying at the hotel	great prices	hot dog
	frozen yogurt	chinese food	dog park	parking lot	rice and beans
	eggs benedict	pad thai	room was clean	wal mart	sweet potato fries
	peanut butter	dim sum	pool area	shopping center	pretty good
	cup of coffee	thai food	great place	great place	carne asada
	iced coffee	pretty good	staff is friendly	prices are reasonable	mac and cheese
	scrambled eggs	lunch specials	free wifi	love this place	fish tacos

ToPMine: Faster and Generating Better Quality Phrases

Running time of different algorithms

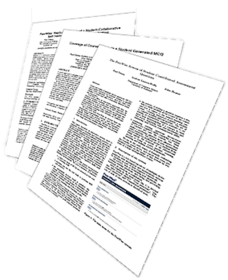
Method	*sampled dblp titles (k=5)*	*dblp titles (k=30)*	*sampled dblp abstracts*	*dblp abstracts*
PDLDA	3.72(hrs)	~20.44(days)	1.12(days)	~95.9(days)
Turbo Topics	6.68(hrs)	>30(days)*	>10(days)*	>50(days)*
TNG	146(s)	5.57 (hrs)	853(s)	NA†
LDA	**65(s)**	3.04 (hrs)	353(s)	13.84(hours)
KERT	68(s)	3.08(hrs)	1215(s)	NA†
ToP-Mine	67(s)	**2.45(hrs)**	**340(s)**	**10.88(hrs)**

Phrase quality measured by z-score



SegPhrase: From Raw Corpus to Quality Phrases and Segmented Corpus

Raw Corpus | **Quality Phrases** | **Segmented Corpus**



Document 1
Citation recommendation is an interesting but challenging research problem in data mining

Document 2
In this study, we investigate the problem in the context of heterogeneous information networks using data mining technique.

Document 3
Principal Component Analysis is a linear dimensionality reduction technique commonly used in machine learning applications.

+ A small set of labels or a general KB

Input Raw Corpus → **Quality Phrases** ↔ **Segmented Corpus**

Phrase Mining | **Phrasal Segmentation**

Integrating phrase mining with phrasal segmentation

J. Liu et al. **Mining Quality Phrases from Massive Text Corpora.** SIGMOD, 2015 (**won Grand Prize in Yelp Dataset Challenge**).

Experiments: Interesting Phrases Generated
(From the Titles and Abstracts of SIGMOD)

Query	SIGMOD	
Method	**SegPhrase+**	**Chunking (TF-IDF & C-Value)**
1	data base	data base
2	database system	database system
3	relational database	query processing
4	query optimization	query optimization
5	query processing	relational database
…	…	…
51	sql server	database technology
52	**relational data**	database server
53	data structure	**large volume**
54	**join query**	performance study
55	web service	web service
…	…	…
201	**high dimensional data**	**efficient implementation**
202	location based service	sensor network
203	xml schema	**large collection**
204	**two phase locking**	**important issue**
205	deep web	frequent itemset
…	…	…

Only in SegPhrase+

Only in Chunking

Mining Quality Phrases in Multiple Languages

- Both ToPMine and SegPhrase+ are extensible to mining quality phrases in multiple languages
  - SegPhrase+ on Chinese (From Chinese Wikipedia)
  - ToPMine on Arabic (From Quran Fus7a Arabic)(no preprocessing)
    - Experimental results of Arabic phrases:

      اورفك → Those who disbelieve

      مسب الله نمحرلا ميحرلا → In the name of God the Gracious and Merciful

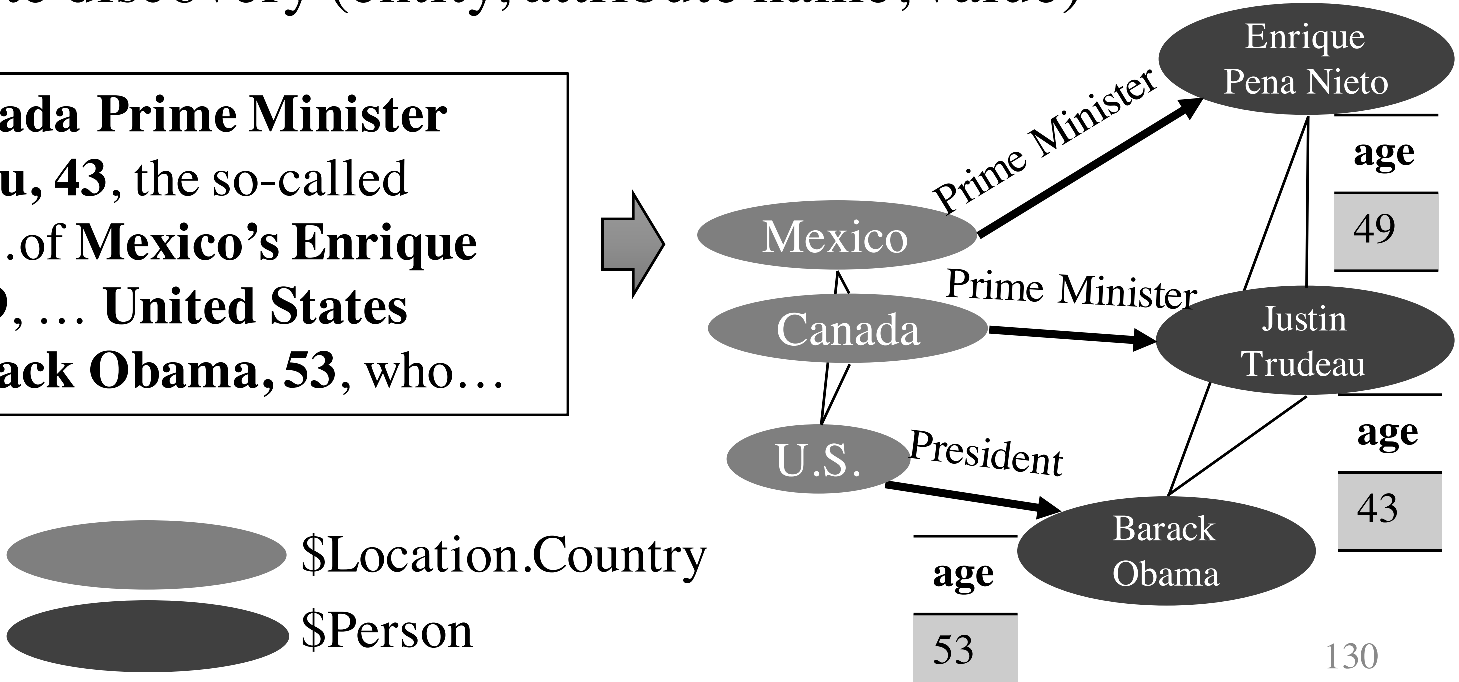
Rank	Phrase	In English
…	…	…
62	首席_执行官	CEO
63	中间_偏右	Middle-right
…	…	…
84	百度_百科	Baidu Pedia
85	热带_气旋	Tropical cyclone
86	中国科学院_院士	Fellow of Chinese Academy of Sciences
…	…	…
1001	十大_中文_金曲	Top-10 Chinese Songs
1002	全球_资讯网	Global Info Website
1003	天一阁_藏_明代_科举_录_选刊	A Chinese book name
…	…	…
9934	国家_戏剧_院	National Theater
9935	谢谢_你	Thank you
…	…	…

Construction of Heterogeneous Information Networks from Text

Philosophy: Not extensive "labeling" but exploring the power of massive text corpora!

- Mining phrases (the minimal semantic units)
- **Entity recognition and typing**
- Attribute discovery (entity, attribute name, value)

…here by **Canada Prime Minister Justin Trudeau, 43**, the so-called #APEChottie…of **Mexico's Enrique Pena Nieto, 49**, … **United States President Barack Obama, 53**, who…



Why Entity Recognition and Typing from Massive Corpora?

- Traditional named entity recognition systems are designed for major types (e.g., PER, LOC, ORG) and general domains (e.g., news)
  - Require additional steps to adapt to new domains/types
  - Expensive human labor on annotation
    - 500 documents for entity extraction; 20,000 queries for entity linking
  - Unsatisfying agreement due to various granularity levels and scopes of types
- Entities obtained by entity linking techniques have *limited* coverage and freshness
  - > 50% unlinkable entity mentions in Web corpus [Lin et al., EMNLP'12]
  - > 90% in our experiment corpora: tweets, Yelp reviews, …
- A new approach: ClusType: Entity Recognition and Typing by Relation Phrase-Based Clustering [Ren, et al., KDD 2015]
  - Recognizing entity mentions of target types with minimal/no human supervision and with no requirement that entities can be found in a KB (distant supervision)

Recognizing Typed Entities

Identifying token span as entity mentions in documents and labeling their types

Target Types

- **FOOD**
- **LOCATION**
- **JOB_TITLE**
- **EVENT**
- **ORGANIZATION**
- …

The best BBQ I've tasted in Phoenix! I had the pulled pork sandwich with coleslaw and baked beans for lunch. ... The owner is very nice. …

Plain text

The best BBQ:Food I've tasted in Phoenix:LOC ! I had the [pulled pork sandwich]:Food with coleslaw:Food and [baked beans]:Food for lunch. … The owner:JOB_TITLE is very nice. …

Text with typed entities

FOOD

LOCATION

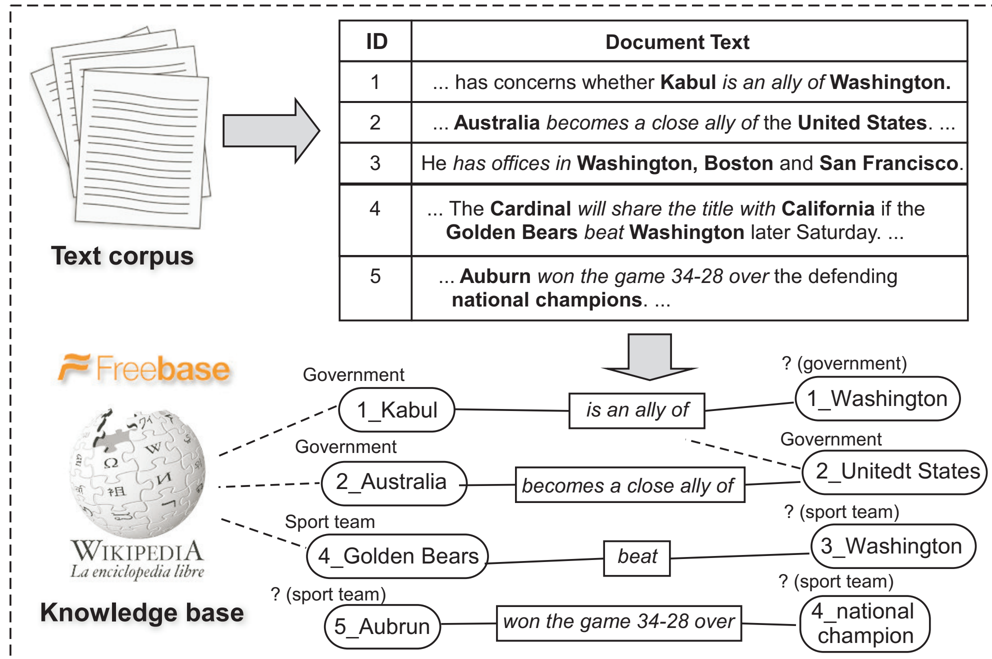
EVENT

Enabling structured analysis of unstructured text corpus

ClusType: A Distant Supervision Framework

Problem: *Distantly-supervised entity recognition in a domain-specific corpus*

- Given: (1) a domain-specific corpus *D*, (2) a knowledge base (e.g., Freebase), (3) a set of target types (*T*) from a KB
- Detect candidate entity mentions in D, and categorize each candidate mention by target types or Not-Of-Interest (NOI)



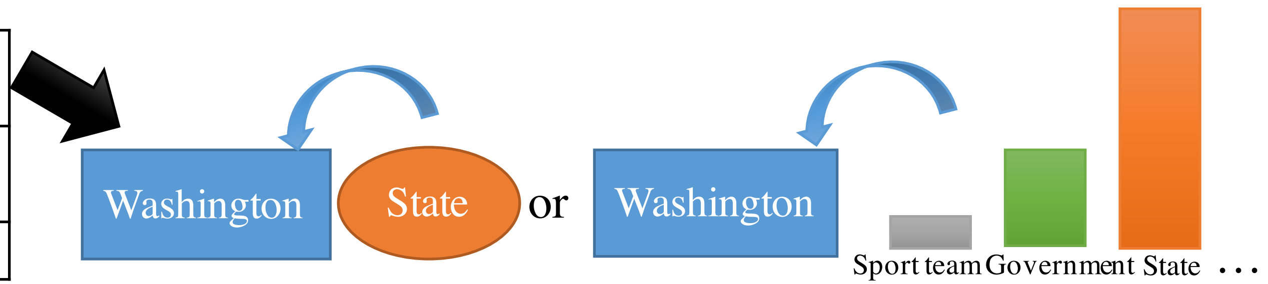
Solution:

- Detect entity mentions from text
- Map candidate mentions to KB entities of target types
- Use confidently mapped {mention, type} to infer types of remaining candidate mentions

Entity Recognition and Typing: Challenges and Solutions

- Challenge 1: Domain Restriction: Extensive training, use general-domain corpora, not work well on specific, dynamic or emerging domains (*e.g.*, tweets, Yelp reviews)
  - Solution: Domain-agnostic phrase mining: Extracts candidate entity mentions with minimal linguistic assumption (e.g., only use POS tagging)
- Challenge 2: Name ambiguity: Multiple entities may share the same surface name
  - Solution: Model each mention based on its surface name and context

While Griffin is not the part of Washington's plan on Sunday's game, …
… news from Washington indicates that the congress is going to…
It is one of the best state parks in Washington.



- Challenge 3: Context Sparsity: There are many ways to describe the same relation
  - Solution: cluster *relation phrase*, infer synonymous relation phrases

Sentence	Freq.
The magnitude 9.0 quake **caused widespread devastation in [Kesennuma city]**	12
… tsunami that **ravaged [northeastern Japan]** last Friday	31
The resulting tsunami **devastate [Japan]**'s northeast	244

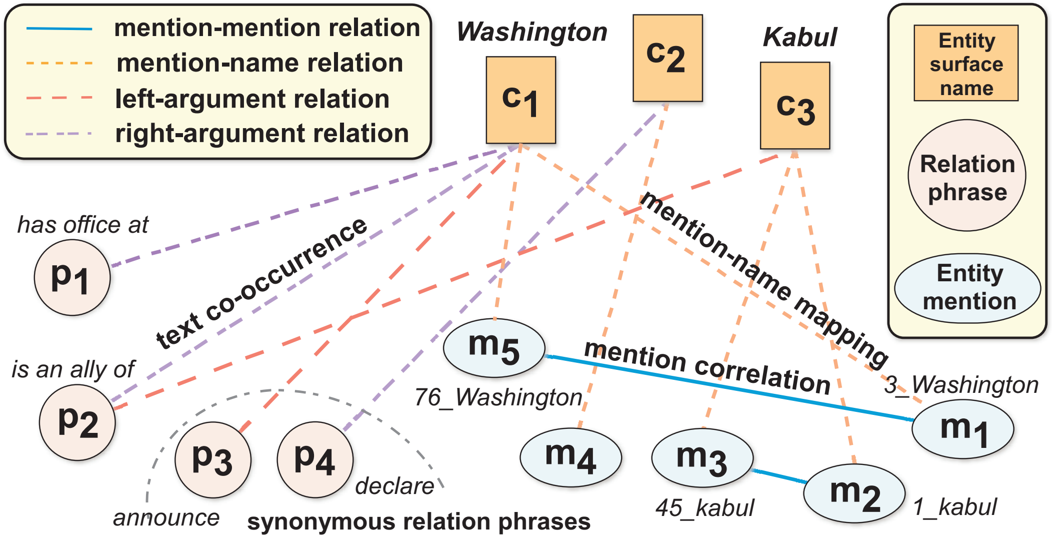
The ClusType Framework: Phrase Segmentation and Heterogeneous Graph Construction

- POS-constrained phrase segmentation for mining candidate entity mentions and relation phrases, simultaneously
- Construct a heterogeneous graph to represent available information in a unified form

Entity mentions are kept as individual objects **to be disambiguated**

Linked to entity surface names & relation phrases

Weight assignment: The more two objects are likely to share the same label, the larger the weight will be associated with their connecting edge



The ClusType Framework: Mutual Enhancement of Type Propagation and Relation Phrase Clustering

- With the constructed graph, formulate a graph-based semi-supervised learning of two tasks jointly:

Type propagation on heterogeneous graph

Derived entity argument types serve as **good feature** for clustering relation phrases

Multi-view relation phrase clustering

Propagate type information among entities bridges via synonymous relation phrases

Mutually enhancing each other; leads to quality recognition of unlinkable entity mentions

ClusType: A General Framework Overview

- **Candidate Generation**
  - Perform phrase mining on a POS-tagged corpus to extract candidate entity mentions and relation phrases
- **Construction of Heterogeneous Graphs**
  - Construct a heterogeneous graph to encode our insights on modeling the type for each entity mention
  - Collect seed entity mentions as labels by linking extracted mentions to the KB
- **Relation Phrase Clustering**
  - Estimate type indicator for unlinkable candidate mentions with the proposed type propagation integrated with relation phrase clustering on the constructed graph

Candidate Generation

- Phrase mining incorporating both *corpus-level statistics* and *syntactic constraints*
  - **Global significance score**: Filter low-quality candidates; **generic POS tag patterns**: remove phrases with improper syntactic structure
  - Extend ToPMine to partition corpus into segments which meet both significance threshold and POS patterns → candidate entity mentions & relation phrases

Relation phrase: Phrase that denotes a unary or binary relation in a sentence

Pattern	Example
V	disperse; hit; struck; knock;
P	in; at; of; from; to;
V P	locate in; come from; talk to;
VW*(P)	caused major damage on; come lately

V-verb; P-prep; W-{adv | adj | noun | det | pron}
W* denotes multiple W; (P) denotes optional.

Experiment: Entity detection: Performance comparison between our method and an NP chunker

Method	NYT		Yelp		Tweet	
	Prec	Recall	Prec	Recall	Prec	Recall
Our method	**0.469**	**0.956**	**0.306**	**0.849**	0.226	**0.751**
NP chunker	0.220	0.609	0.296	0.247	**0.287**	0.181

Recall is most critical for this step, since later we cannot detect the misses (i.e., false negatives)

Type Inference: A Joint Optimization Problem

$$\mathcal{O}_{\alpha,\gamma,\mu} = \mathcal{F}(\mathbf{C}, \mathbf{P}_L, \mathbf{P}_R) + \mathcal{L}_\alpha\big(\mathbf{P}_L, \mathbf{P}_R, \{\mathbf{U}^{(v)}, \mathbf{V}^{(v)}\}, \mathbf{U}^*\big) + \Omega_{\gamma,\mu}(\mathbf{Y}, \mathbf{C}, \mathbf{P}_L, \mathbf{P}_R). \quad (2)$$

Mention modeling & mention correlation

$$\mathcal{F}(\mathbf{C}, \mathbf{P}_L, \mathbf{P}_R) = \sum_{i=1}^{n}\sum_{j=1}^{l} W_{L,ij}\left\| \frac{\mathbf{C}_i}{\sqrt{D^{(\mathcal{C})}_{L,ii}}} - \frac{\mathbf{P}_{L,j}}{\sqrt{D^{(\mathcal{P})}_{L,jj}}} \right\|_2^2 + \sum_{i=1}^{n}\sum_{j=1}^{l} W_{R,ij}\left\| \frac{\mathbf{C}_i}{\sqrt{D^{(\mathcal{C})}_{R,ii}}} - \frac{\mathbf{P}_{R,j}}{\sqrt{D^{(\mathcal{P})}_{R,jj}}} \right\|_2^2$$

$$\Omega_{\gamma,\mu}(\mathbf{Y}, \mathbf{C}, \mathbf{P}_L, \mathbf{P}_R) = \big\|\mathbf{Y} - f(\Pi_{\mathcal{C}}\mathbf{C}, \Pi_L\mathbf{P}_L, \Pi_R\mathbf{P}_R)\big\|_F^2 + \frac{\gamma}{2}\sum_{c\in\mathcal{C}}\sum_{i,j=1}^{M_c} W_{ij}^{(c)}\left\| \frac{\mathbf{Y}_i}{\sqrt{D^{(c)}_{ii}}} - \frac{\mathbf{Y}_j}{\sqrt{D^{(c)}_{jj}}} \right\|_2^2 + \mu\|\mathbf{Y} - \mathbf{Y}_0\|_F^2$$

Type propagation between entity surface names and relation phrases

$$\mathcal{L}_\alpha\big(\mathbf{P}_L, \mathbf{P}_R, \{\mathbf{U}^{(v)}, \mathbf{V}^{(v)}\}, \mathbf{U}^*\big) = \sum_{v=0}^{d} \beta^{(v)}\big(\|\mathbf{F}^{(v)} - \mathbf{U}^{(v)}\mathbf{V}^{(v)T}\|_F^2 + \alpha\|\mathbf{U}^{(v)}\mathbf{Q}^{(v)} - \mathbf{U}^*\|_F^2\big). \quad (3)$$

Multi-view relation phrases clustering)

ClusType: Experiment Setting

- Datasets: 2013 New York Times news (~110k docs) [event, PER, LOC, ORG]; Yelp Reviews (~230k) [Food, Job, …]; 2011 Tweets (~300k) [event, product, PER, LOC, …]
- Seed mention sets: < 7% extracted mentions are mapped to Freebase entities
- Evaluation sets: manually annotate mentions of target types for subsets of the corpora
- Evaluation metrics: Follows named entity recognition evaluation (Precision, Recall, F1)
- Compared methods
  - **Pattern**: Stanford pattern-based learning; **SemTagger**:bootstrapping method which trains contextual classifier based on seed mentions; **FIGER**: distantly-supervised sequence labeling method trained on Wiki corpus; **NNPLB**: label propagation using ReVerb assertion and seed mention; **APOLLO**: mention-level label propagation using Wiki concepts and KB entities;
  - **ClusType-NoWm**: ignore mention correlation; **ClusType-NoClus**: conducts only type propagation; **ClusType-TwpStep**: first performs hard clustering then type propagation

Comparing ClusType with Other Methods and Its Variants

Performance comparison on three datasets in terms of Precision, Recall and F1 score

Table 5: Performance comparisons on three datasets in terms of Precision, Recall and F1 score.

Data sets	NYT			Yelp			Tweet		
Method	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Pattern [9]	0.4576	0.2247	0.3014	0.3790	0.1354	0.1996	0.2107	0.2368	0.2230
FIGER [16]	0.8668	0.8964	0.8814	0.5010	0.1237	0.1983	**0.7354**	0.1951	0.3084
SemTagger [12]	0.8667	0.2658	0.4069	0.3769	0.2440	0.2963	0.4225	0.1632	0.2355
APOLLO [29]	0.9257	0.6972	0.7954	0.3534	0.2366	0.2834	0.1471	0.2635	0.1883
NNPLB [15]	0.7487	0.5538	0.6367	0.4248	0.6397	0.5106	0.3327	0.1951	0.2459
ClusType-NoClus	0.9130	0.8685	0.8902	0.7629	0.7581	0.7605	0.3466	0.4920	0.4067
ClusType-NoWm	0.9244	0.9015	0.9128	0.7812	0.7634	0.7722	0.3539	**0.5434**	0.4286
ClusType-TwoStep	0.9257	0.9033	0.9143	0.8025	0.7629	0.7821	0.3748	0.5230	0.4367
ClusType	**0.9550**	**0.9243**	**0.9394**	**0.8333**	**0.7849**	**0.8084**	0.3956	0.5230	**0.4505**

- **vs. FIGER**: Effectiveness of our candidate generation and type propagation
- vs. **NNPLB** and **APOLLO**: ClusType utilizes not only semantic-rich relation phrase as type cues, but also cluster synonymous relation phrases to tackle context sparsity
- vs. our **variants**: (i) models mention correlation for name disambiguation; and (ii) integrates clustering in a mutually enhancing way

Comparing on Trained NER System

- Compare with Stanford NER,which is trained on general-domain corpora including ACE corpus and MUC corpus, on three types: PER, LOC, ORG

F1 score comparison with trained NER

Table 6: F1 score comparison with trained NER.

Method	NYT	Yelp	Tweet
Stanford NER [6]	0.6819	0.2403	0.4383
ClusType-NoClus	0.9031	0.4522	0.4167
ClusType	**0.9419**	**0.5943**	**0.4717**

[6] J. R. Finkel, T. Grenager and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In ACL'05.

- ClusType and its variants outperform Stanford NER on both dynamic corpus (NYT) and domain-specific corpus (Yelp)
- ClusType has lower precision but higher Recall and F1 score on Tweet → Superior recall of ClusType mainly come from domain-independent candidate generation

Example Output and Relation Phrase Clusters

Example output of ClusType and the compared methods on the Yelp dataset

ClusType	SemTagger	NNPLB
The best **BBQ:Food** I've tasted in **Phoenix:LOC** ! I had the [**pulled pork sandwich**]:**Food** with **coleslaw:Food** and [**baked beans**]:**Food** for lunch. ...	The best BBQ I've tasted in **Phoenix:LOC** ! I had the pulled [**pork sandwich**]:**LOC** with **coleslaw:Food** and [**baked beans**]:**LOC** for lunch. ...	The best **BBQ:Loc** I've tasted in **Phoenix:LOC** ! I had the pulled pork **sandwich:Food** with coleslaw and baked **beans:Food** for **lunch:Food**. ...
I only go to **ihop:LOC** for **pancakes:Food** because I don't really like anything else on the menu. Ordered [**chocolate chip pancakes**]:**Food** and a [**hot chocolate**]:**Food**.	I only go to ihop for pancakes because I don't really like anything else on the menu. Ordered [**chocolate chip pancakes**]:**LOC** and a [**hot chocolate**]:**LOC**.	I only go to ihop for pancakes because I don't really like anything else on the menu. Ordered chocolate chip pancakes and a hot chocolate.

❑ Extracts more mentions and predicts types with higher

Example relation phrase clusters and corpus-wide frequency from the NYT dataset

ID	Relation phrase
1	recruited by (5.1k); employed by (3.4k); want hire by (264)
2	go against (2.4k); struggling so much against (54); run for re-election against (112); campaigned against (1.3k)
3	looking at ways around (105); pitched around (1.9k); echo around (844); present at (5.5k);

- Not only synonymous relation phrases, but also both sparse and frequent relation phrase can be clustered together
- → boosts sparse relation phrases with type information of frequent relation phrases

Fine-grained Entity Typing

- **Fine-grained Entity Typing**: Type labels for a mention forms a "*type-path*" (not necessarily ending in a leaf node) in a given (tree-structured) type hierarchy

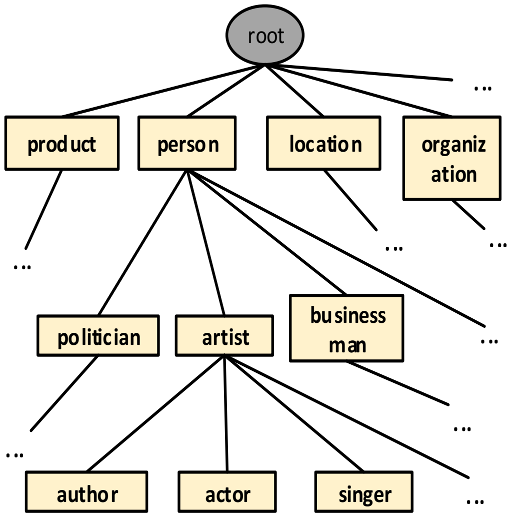
ID	Sentence
S1	Republican presidential candidate ***Donald Trump*** spoke during a campaign event in Rock Hill.
S2	***Donald Trump***'s company has threatened to withhold up to $1 billion of investment if the U.K. government decides to ban his entry into the country.
S3	In ***Trump***'s TV reality show, "The Apprentice", 16 people competed for a job.
...	...

Type-path

Person → politician

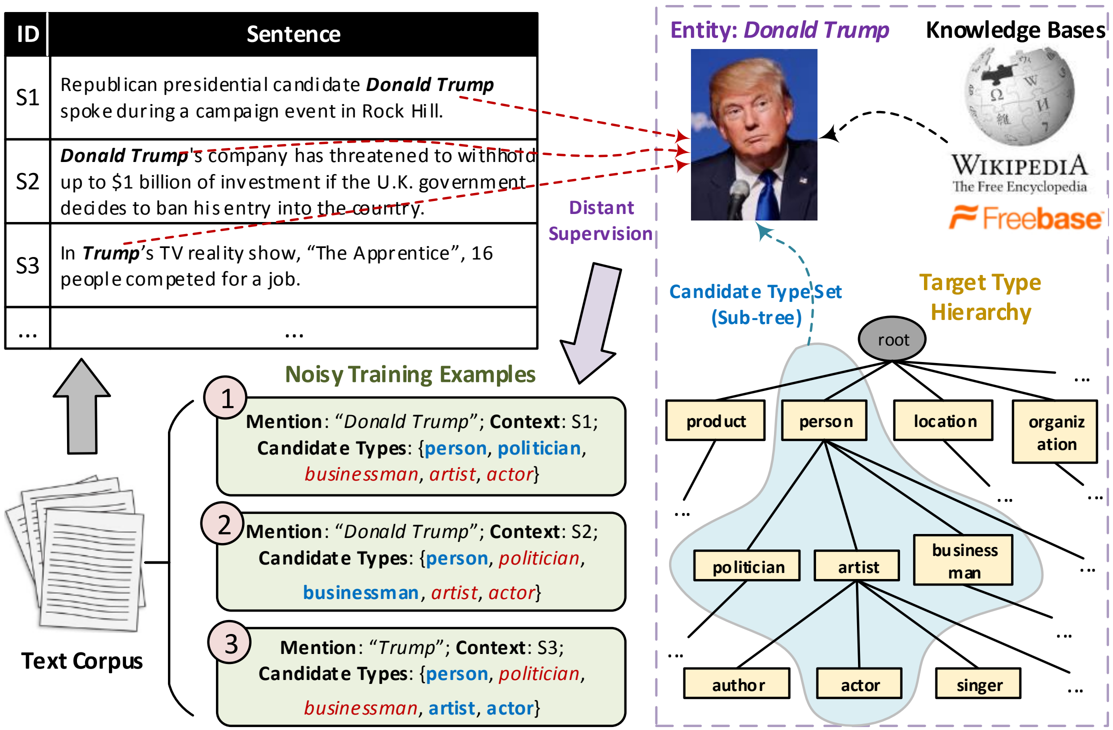
Person → businessman

Person → artist → actor



- Manually annotating training corpora with **100+** entity types
  - Expensive & Error-prone
- **Current practice**: use distant supervision to ***automatically labeled training corpora***

Label Noise in Entity Typing



Donald Trump is mentioned is sentences S1-S3.

- Distant supervision
  - Assign *same* types (blue region) to *all* the mentions
  - Does not consider ***local contexts*** when assigning type labels
  - Introduce ***label noise*** to the mentions

The types assigned to entity Trump include person, artist, actor, politician, businessman, while only {person, politician} are correct types for the **mention "*Trump*" in S1**

Label Noise in Entity Typing (cont.)

- Current typing systems either **ignore this issue**
  - assume all candidate labels obtained by supervision are "true" labels

Dataset	Wiki	OntoNotes	BBN	NYT
# of target types	113	89	47	446
(1) noisy mentions (%)	27.99	25.94	22.32	51.81
(2a) sibling pruning (%)	23.92	16.09	22.32	39.26
(2b) min. pruning (%)	28.22	8.09	3.27	32.75
(2c) all pruning (%)	45.99	23.45	25.33	61.12

- Or use **simple pruning heuristics** to ***delete*** mentions with conflicting types
  - aggressive deletion of mentions → significant loss of training data

The larger the target type set, the more severe the loss!

Label Noise Reduction: Task Description

- Define a *new* task, called **Label Noise Reduction in Entity Typing,** to identify the correct type-path for *each mention in training set*, from its *noisy candidate type set*
  - **VS. typical typing systems**: they focus on designing models for typing *unlabeled mentions*
  - The first systematic study of type label noise in distant supervision
  - A fundamental task for entity typing systems (the bottleneck of their performance)
- **Problem Definition**
  - **Input**:
    - (1) Automatically labeled training corpus: set of (mention, context, candidate type labels) triples
    - (2) Knowledge base, along with its entity-type facts (i.e., set of (entity, type) tuples)
    - (3) Target type hierarchy **T**
  - **Output**: Estimate ***a single type-path*** (not required to end in a leaf node) in the hierarchy **T**, based on the mention itself as well as its context in the sentence
- **Non-goals**: Entity mention detection; Entity linking; Type hierarchy creation

Label Noise Reduction: Challenges

Presence of incorrect type labels in a mention's candidate type set

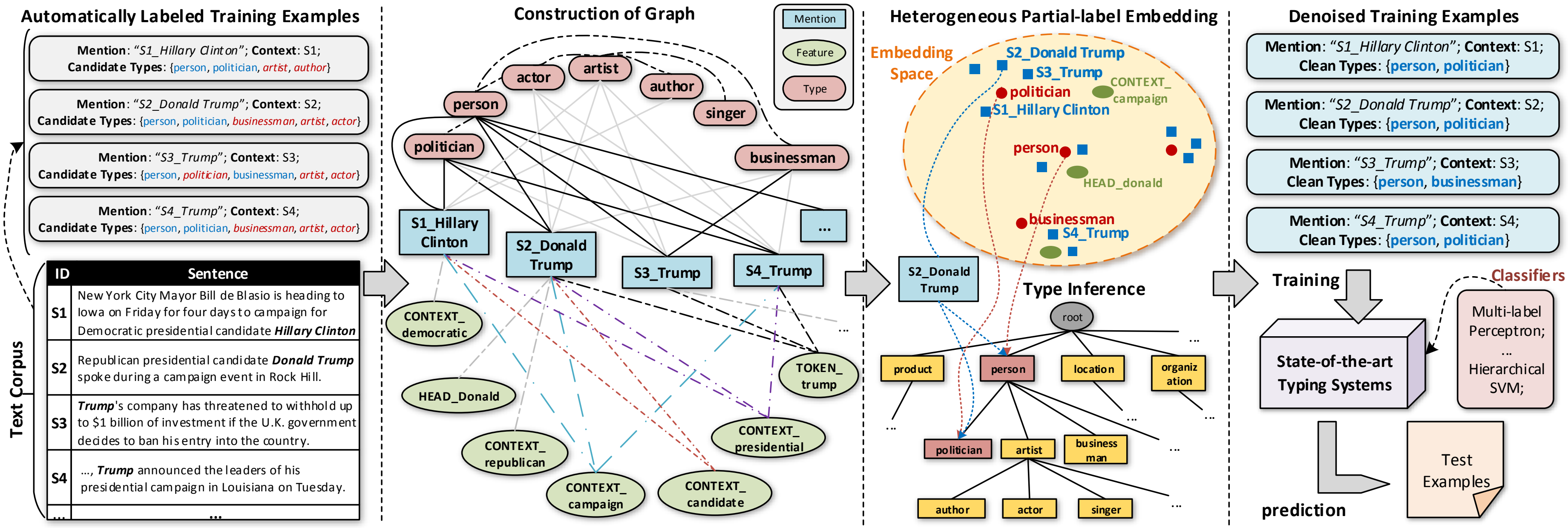
- Supervised/semi-supervised techniques both assume "*all labels are correct/reliable labels*"
- How to accurately estimate the relatedness between mentions and types?
  - **Aspect I**: How to model the *noisy associations between mention and its candidate labels*, to indicate the "truth status" of the candidate labels
  - **Aspect II**: How to incorporate the *semantic similarity between types*, as we are estimating the type-path holistically for a mention
    - vs. estimating individual labels independently

Label Noise Reduction: Solution Ideas

- Propose a weakly-supervised (unsupervised) approach, where the end goal is to estimate the *relatedness between mentions and types*
    1. sim(mention, **true** candidate label) > sim(mention, **false** candidate label)
    2. sim(mention, **fine-grained** true label) > sim(mention, **coarse-grained** true label)

1. Model the "truth status" of candidate labels as "latent values" using a novel ***partial-label loss*** → progressively estimate them by incorporating multiple signals:
    - *Co-occurrences between text features and mentions* in the large corpus
    - *Collective associations between type labels and mentions* in the large corpus
2. Model *semantic similarity between types* (*i.e.*, type correlation) derived from KB, to ensure holistic type-path estimation

Label Noise Reduction: Framework Overview



1. Generate text features and construct a heterogeneous graph
2. Perform joint embedding of the constructed graph G into the same low-dimensional space
3. For each mention, search its candidate type sub-tree in a top-down manner and estimate the true type-path from learned embedding

Text Features for Fine-grained Typing

- Features are extracted from:
  - (1) mention's name string: *e.g., head token, POS tags, Brown Cluster of head token*
  - (2) mention's context in the sentence: *e.g., n-grams, dependency roles*

Feature	Description	Example
Head	Syntactic head token of the mention	"HEAD_Turing"
Token	Tokens in the mention	"Turing", "Machine"
POS	Part-of-Speech tag of tokens in the mention	"NN"
Character	All character trigrams in the head of the mention	":tu", "tur", ..., "ng:"
Word Shape	Word shape of the tokens in the mention	"Aa" for "Turing"
Length	Number of tokens in the mention	"2"
Context	Unigrams/bigrams before and after the mention	"CXT_B:Maserati ,", "CXT_A:and the"
Brown Cluster	Brown cluster ID for the head token (learned using $\mathcal{D}$)	"4_1100", "8_1101111", "12_111011111111"
Dependency	Stanford syntactic dependency [16] associated with the head token	"GOV:nn", "GOV:turing"

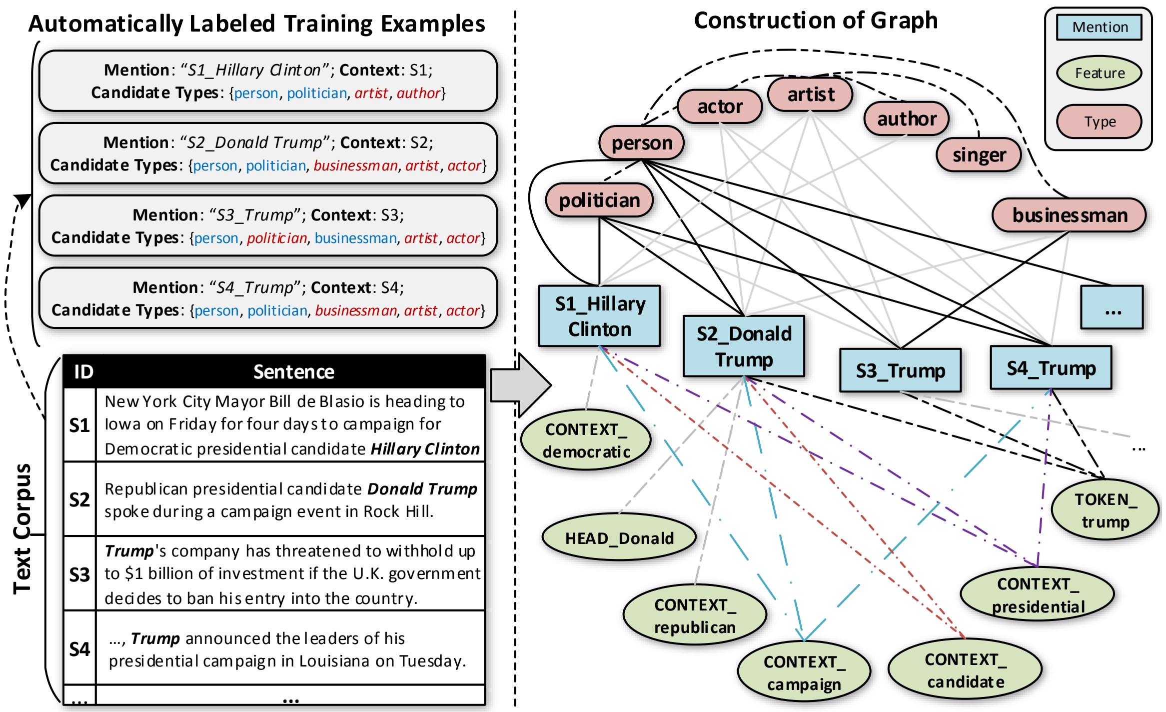
- "*Turing Machine*" is used as an example mention from the sentence:
  - "The band's former drummer Jerry Fuchs—who was also a member of Maserati, Turing Machine and The Juan MacLean—died after falling down an elevator shaft.".

Construction of Heterogeneous Graphs

- With three types of objects extracted from corpus: entity mentions, target types, and text features

Three types of links:

1. **Mention-type link**: represents each mention's candidate type assignment
2. **Mention-feature link**: captures *corpus-level co-occurrences* between mentions and text features
3. **Type correlation link**: encodes the type correlation derived from KB or target type hierarchy



Mention-Type Association Subgraph

- Forms a bipartite graph between entity mentions and target types
  - Each mention is linked to its candidate types with binary weight
  - Some links are "false" links in the constructed mention-type subgraph
  - The likelihood of a mention-type link is measured by the relevance between the corresponding mention and type

Example: In sentence S1, context words *democratic* and *presidential* infer that type **politician** is more relevant than type **actor** for mention "Hillary Clinton"

Hypothesis 1 (Partial Label Association):
A mention should be embedded closer to its most relevant candidate type than to any other non-candidate type, yielding higher similarity between the corresponding embedding vectors.

ID	Sentence
S1	New York City Mayor Bill de Blasio is heading to Iowa on Friday for four days to campaign for Democratic presidential candidate ***Hillary Clinton***
S2	Republican presidential candidate ***Donald Trump*** spoke during a campaign event in Rock Hill.
S3	***Trump***'s company has threatened to withhold up to $1 billion of investment if the U.K. government decides to ban his entry into the country.
S4	..., ***Trump*** announced the leaders of his presidential campaign in Louisiana on Tuesday.

Mention-Feature Co-occurrence Subgraph

- Intuition
  - Mentions sharing many text features tend to have close type semantics
  - Text features which co-occur with many entity mentions in the corpus likely represent similar entity types.

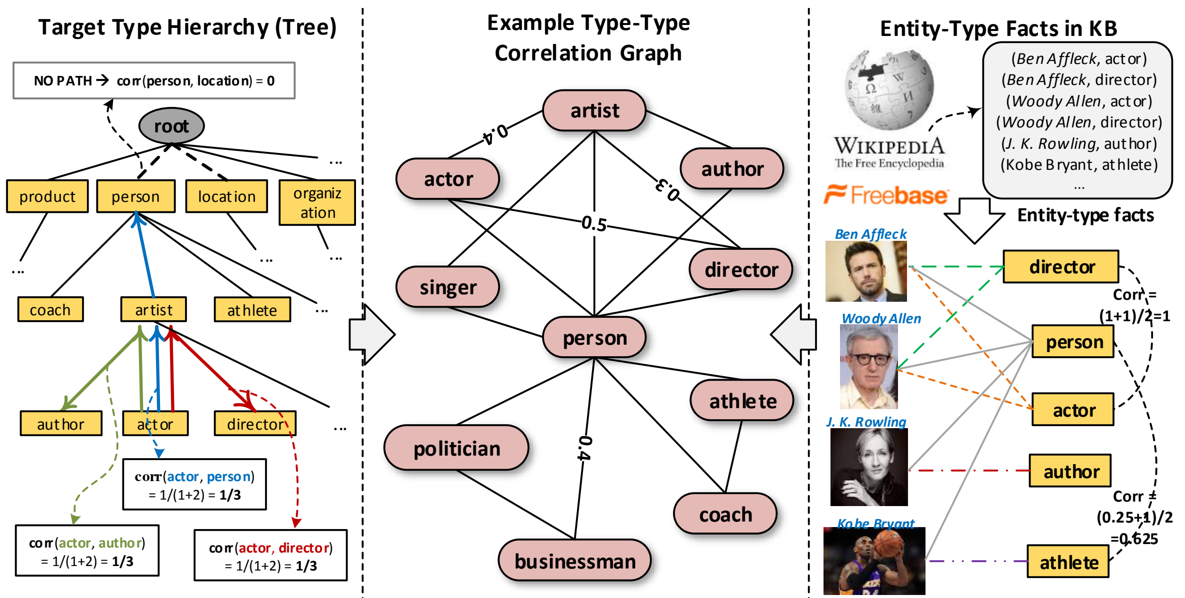
Example: mentions "Donald Trump" in S2 and "Trump" in S4 share multiple features (e.g., *Trump*, *presidential* and *campaign*), and thus are likely of the same type **politician**. Conversely, features *campaign* and *presidential* likely represent the same type politician since they co-occur with similar sets of mentions in the corpus.

ID	Sentence
S1	New York City Mayor Bill de Blasio is heading to Iowa on Friday for four days to campaign for Democratic presidential candidate ***Hillary Clinton***
S2	Republican presidential candidate ***Donald Trump*** spoke during a campaign event in Rock Hill.
S3	***Trump***'s company has threatened to withhold up to $1 billion of investment if the U.K. government decides to ban his entry into the country.
S4	..., ***Trump*** announced the leaders of his presidential campaign in Louisiana on Tuesday.

Hypothesis 2 (Mention-Feature Co-occurrences): *If two entity mentions share similar features, they should be close to each other in the embedding space (i.e., high similarity score). If two features co-occur with a similar set of mentions, their embedding vectors tend to be similar.*

Type Correlation Subgraph

- Build a homogeneous graph to represent the semantic similarity between types
  - **Simple way**: Use distance in the target type hierarchy
    - In target type hierarchy, types closer to each other tend to be more related
    - **Example:** actor is more related to artist than to person in the left column
  - **Advanced way**: Exploit entity-type facts in KB
    - Given two target types, the correlation between them is proportional to the number of entities they share in the KB



Hypothesis 3 (Type Correlation):
If high correlation exists between two target types based on either type hierarchy or KB, they should be embedded close to each other.

Heterogeneous Partial-Label Embedding (PLE): The Joint Optimization Problem

$$\min_{\{\mathbf{u}_i\}_{i=1}^N,\{\mathbf{c}_j\}_{j=1}^M,\{\mathbf{v}_k,\mathbf{v}'_k\}_{k=1}^K} \mathcal{O} = \mathcal{O}_{MY} + \mathcal{O}_{MF} + \mathcal{O}_{YY}$$

$$\mathcal{O}_{MY} = \sum_{i=1}^N \ell_i + \frac{\lambda}{2}\sum_{i=1}^N \|\mathbf{u}_i\|_2^2 + \frac{\lambda}{2}\sum_{k=1}^K \|\mathbf{v}_k\|_2^2.$$

$$\ell_i = \max\left\{0, 1 - \left[\max_{y\in\mathcal{Y}_i} s(m_i,y) - \max_{y'\in\overline{\mathcal{Y}}_i} s(m_i,y')\right]\right\}$$

Partial label loss between mentions and types (Hypo 1)

$$\mathcal{O}_{MF} = -\sum_{(m_i,f_j)\in G_{MF}} w_{ij}\cdot\log p(f_j|m_i)$$

Model mention-feature links using second-order skip-gram objective (Hypo 2)

$$\mathcal{O}_{YY} = -\sum_{(y_k,y_{k'})\in G_{YY}} w_{kk'}\left[\log p(y_{k'}|y_k) + \log p(y_k|y_{k'})\right]$$

Type correlation based on KB (Hypo 3)

PLE: Partial-Label Loss

$$\ell_i = \max\left\{0, 1 - \left[\max_{y \in \mathcal{Y}_i} s(m_i, y) - \max_{y' \in \overline{\mathcal{Y}}_i} s(m_i, y')\right]\right\}$$

- **Intuition**
  - For mention m_i, the maximum score associated with its candidate types $\mathcal{Y}_i$ is greater than the maximum score associated with any other non-candidate types $\overline{\mathcal{Y}_i}$, where the scores are measured using current embedding vectors.

- **vs. multi-label learning**
  - A large margin is enforced between *all* candidate types and non-candidate types without considering noisy types.

PLE: Second-Order Proximity Model

- Intuition
  - Nodes with similar distributions over neighbors are similar to each other
- Define the probability of feature f_j generated by mention m_i for each link (m_i, f_j) in the mention-feature subgraph as follows

$$p(f_j|m_i) = \exp(\mathbf{c}_j^T \mathbf{u}_i) / \sum_{f_{j'} \in \mathcal{F}} \exp(\mathbf{c}_{j'}^T \mathbf{u}_i)$$

- Enforce the conditional distribution specified by embeddings, i.e., $p(\cdot\,|m_i)$, to be close to the empirical distribution (i.e., link distribution of m_i over all features in the mention-feature subgraph)

Learning Algorithm for PLE

$$\min_{\{\mathbf{u}_i\}_{i=1}^N, \{\mathbf{c}_j\}_{j=1}^M, \{\mathbf{v}_k, \mathbf{v}'_k\}_{k=1}^K} \mathcal{O} = \mathcal{O}_{MY} + \mathcal{O}_{MF} + \mathcal{O}_{YY}$$

- Can be efficiently solved by alternative minimization algorithm based on block coordinate descent schema
- Algorithm complexity is linear to #links in the heterogeneous graph
- Mini-batch stochastic sub-gradient descent can also be applied for our problem

Algorithm 1: Model Learning of PLE

Input: $G = \{G_{MY}, G_{MF}, G_{YY}\}$, regularization parameter λ, learning rate α, number of negative samples Z

Output: entity mention embeddings $\{\mathbf{u}_i\}_{i=1}^N$, feature embeddings $\{\mathbf{c}_j\}_{j=1}^M$, type embeddings $\{\mathbf{v}_k\}_{k=1}^K$

Initialize: $\{\mathbf{u}_i\}$, $\{\mathbf{c}_j\}$, and $\{\mathbf{v}_k\}$ as random vectors
while *$\mathcal{O}$ in Eq. (7) not converge* **do**
  for *each link in G_{MF} and G_{YY}* **do**
    Draw Z negative links from noise distribution $P_n(\cdot)$
  end
  for $m_i \in \mathcal{M}$ **do**
    $\mathbf{u}_i \leftarrow \mathbf{u}_i - \alpha \cdot \partial\mathcal{O}/\partial\mathbf{u}_i$ with $\partial\mathcal{O}/\partial\mathbf{u}_i$ defined in Eq. (9)
  end
  for $f_j \in \mathcal{F}$ **do**
    $\mathbf{c}_j \leftarrow \mathbf{c}_j - \alpha \cdot \partial\mathcal{O}/\partial\mathbf{c}_j$ using $\partial\mathcal{O}/\partial\mathbf{c}_j$ defined in Eq. (10)
  end
  for $y_k \in \mathcal{Y}$ **do**
    $\mathbf{v}_k \leftarrow \mathbf{v}_k - \alpha \cdot \partial\mathcal{O}/\partial\mathbf{v}_k$ based on $\partial\mathcal{O}/\partial\mathbf{v}_k$ in Eq. (11)
    $\mathbf{v}'_k \leftarrow \mathbf{v}'_k - \alpha \cdot \partial\mathcal{O}/\partial\mathbf{v}'_k$ using $\partial\mathcal{O}/\partial\mathbf{v}'_k$ in Eq. (12)
  end
end

Top-Down Type Inference

- Perform top-down search in the candidate type sub-tree to estimate the correct type-path

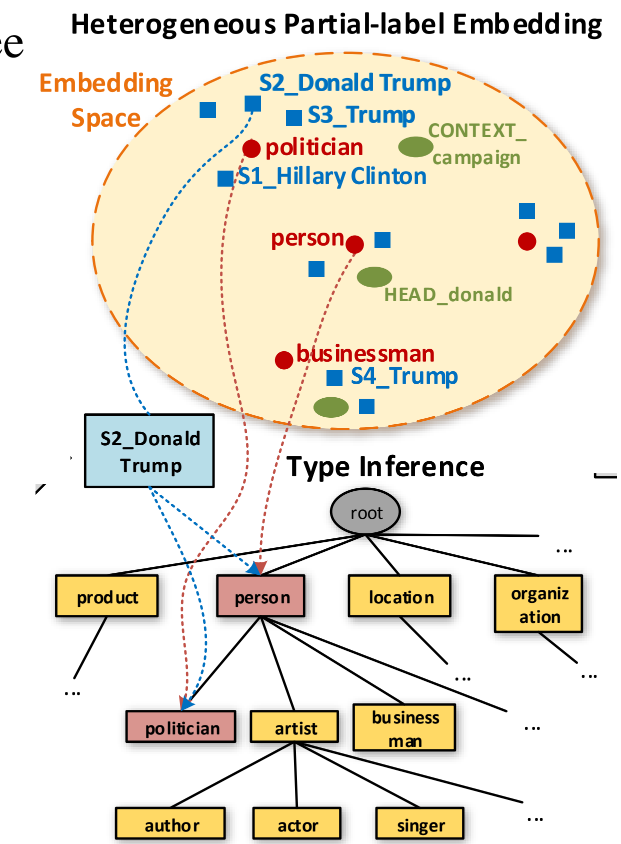
---

Algorithm 2: Type Inference

Input: candidate type sub-tree $\{\mathcal{Y}_i\}$, mention embeddings $\{\mathbf{u}_i\}$, type embeddings $\{\mathbf{v}_k\}$, threshold η

Output: estimated type-path $\{\mathcal{Y}_i^*\}$ for $m_i \in \mathcal{M}$

for $m_i \in \mathcal{M}$ **do**
    Initialize: $\mathcal{Y}_i^*$ as $\emptyset$, r as the root of $\mathcal{Y}$
    while $\mathcal{C}_i(r) \neq \emptyset$ **do**
        $r \leftarrow \mathrm{argmax}_{y_k \in \mathcal{C}_i(r)}\, s(\mathbf{u}_i, \mathbf{v}_k)$
        if $s(\mathbf{u}_i, \mathbf{v}_r) > \eta$ **then**
            Update the type-path: $\mathcal{Y}_i^* \leftarrow \mathcal{Y}_i^* \bigcup \{r\}$
        else
            return $\mathcal{Y}_i^*$ as the estimated type-path for m_i
        end
    end
end



Experiment Setting

- Datasets:
  - (1) **Wiki**: 1.5M sentences sampled from ~780k Wikipedia articles
  - (2) **OntoNotes**: 13,109 news
  - (3) **BBN**: 2,311 Wall Street Journal articles

Data sets	Wiki	OntoNotes	BBN
#Types	113	89	47
#Documents	780,549	13,109	2,311
#Sentences	1.51M	143,709	48,899
#Training mentions	2.69M	223,342	109,090
#Ground-truth mentions	563	9,604	121,001
#Features	644,860	215,642	125,637
#Edges in graph	87M	5.9M	2.9M

Experiment Setting

- Compared Methods
  - **(1) Sib**: removes siblings types; **(2) Min**: removes types that appear only once in the document; **(3) All**: first performs Sib pruning then Min pruning; (4)**DeepWalk**: embedding a homogeneous graph with binary edges**; (5) LINE**: second-order LINE; **(5) WSABIE**: adopts WARP loss with kernel extension; **(6) PTE**: applied PTE joint training algorithm on subgraphs G_{MF} and G_{MY}. **(7) PL-SVM**: uses a margin-based loss to handle label noise. **(8) CLPL**: uses a linear model to encourage large average scores for candidate types.
  - For PLE, we compare **(1)PLE**: adopts KB-based type correlation subgraph; **(2)PLE-CoH**: adopts type hierarchy-based correlation subgraph; **(3)PLE-NoCo**: does not consider type correlation.

Intrinsic Experiments: Effectiveness of Label Noise Reduction

- Goal: compare how accurately PLE and the other methods can estimate the true types of mentions from its noisy candidate type set

Method	Wiki							OntoNotes						
	Acc	Ma-P	Ma-R	Ma-F1	Mi-P	Mi-R	Mi-F1	Acc	Ma-P	Ma-R	Ma-F1	Mi-P	Mi-R	Mi-F1
Raw	0.373	0.558	**0.681**	0.614	0.521	**0.719**	0.605	0.480	0.671	**0.793**	0.727	0.576	**0.786**	0.665
Sib [7]	0.373	0.583	0.636	0.608	0.578	0.653	0.613	0.487	0.710	0.732	0.721	0.675	0.702	0.688
Min [7]	0.373	0.561	0.679	0.615	0.524	0.717	0.606	0.481	0.680	0.777	0.725	0.592	0.763	0.667
All [7]	0.373	0.585	0.634	0.608	0.581	0.651	0.614	0.487	0.716	0.724	0.720	0.686	0.691	0.689
DeepWalk-Raw [21]	0.328	0.598	0.459	0.519	0.595	0.367	0.454	0.441	0.625	0.708	0.664	0.598	0.683	0.638
LINE-Raw [29]	0.349	0.600	0.596	0.598	0.590	0.610	0.600	0.549	0.699	0.770	0.733	0.677	0.754	0.714
WSABIE-Raw [34]	0.332	0.554	0.609	0.580	0.557	0.633	0.592	0.482	0.686	0.743	0.713	0.667	0.721	0.693
PTE-Raw [28]	0.419	0.678	0.597	0.635	0.686	0.607	0.644	0.529	0.687	0.754	0.719	0.657	0.733	0.693
PLE-NoCo	0.556	0.795	0.678	0.732	0.804	0.668	0.730	0.593	0.768	0.773	0.770	0.751	0.762	0.756
PLE-CoH	0.568	0.805	0.671	0.732	0.808	0.704	0.752	0.620	0.789	0.785	0.787	0.778	0.769	0.773
PLE	**0.589**	**0.840**	0.675	**0.749**	**0.833**	0.705	**0.763**	**0.639**	**0.814**	0.782	**0.798**	**0.791**	0.766	**0.778**

40.57% improvement in Accuracy and 23.89% improvement in Macro-Precision compared to the best baseline on Wiki dataset

- **vs. pruning strategies**: LNR *identifies true types* from the candidate type sets instead of *aggressively deleting instances* with noisy type labels
- **vs. other embedding methods**: PLE obtains superior performance because it effectively *models the noisy type labels*
- **vs. PLE variants**: (i) PLE captures *type semantic similarity*; (ii) modeling type correlation with entity-type facts in KB yields more accurate and complete type correlation statistics than type hierarchy-based approach

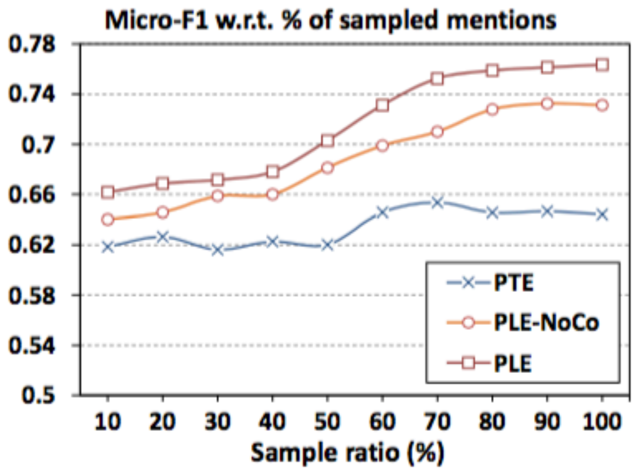
Intrinsic Experiments: Effectiveness of Label Noise Reduction

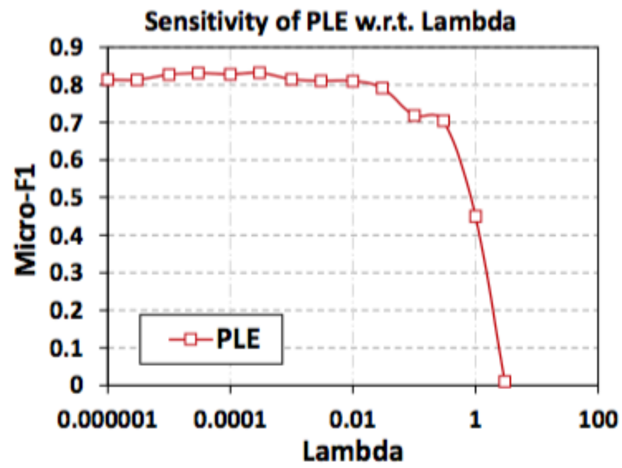
- Example output on news articles

Text	*NASA* says it may decide by tomorrow whether another space walk will be needed ...	... the board of *directors* which are composed of twelve members directly appointed by the *Queen*.
Wiki Page	https://en.wikipedia.org/wiki/NASA	https://en.wikipedia.org/wiki/Elizabeth_II
Cand. type set	person, artist, location, structure, organization, company, news_company	person, artist, actor, author, person_title, politician
WSABIE	person, artist	person, artist
PTE	organization, company, news_company	person, artist
PLE	organization, company	person, person_title

- PLE predicts fine-grained types with better accuracy (e.g., person_title)
- and avoids from overly-specific predictions (e.g., news_company)

Intrinsic Experiments: Effectiveness of Label Noise Reduction





- Testing the effect of training set size
  - Performance of all methods improves as the ratio increases, and becomes *insensitive* as the sampling ratio > 0.7
- Testing the effect of training set size
  - Performance of PLE becomes insensitive as becomes small enough (i.e., 0.01)

Extrinsic Experiments: Fine-Grained Entity Typing

- Compare performance gain of two state-of-the-art typing systems, when using denoised training data output by different compared methods

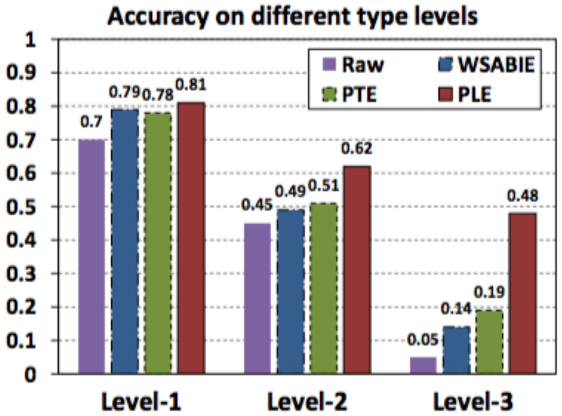
Typing System	Noise Reduction Method	Wiki Acc	Wiki Ma-F1	Wiki Mi-F1	OntoNotes Acc	OntoNotes Ma-F1	OntoNotes Mi-F1	BBN Acc	BBN Ma-F1	BBN Mi-F1
N/A	PL-SVM [20]	0.428	0.613	0.571	0.465	0.648	0.582	0.497	0.679	0.677
N/A	CLPL [2]	0.162	0.431	0.411	0.438	0.603	0.536	0.486	0.561	0.582
HYENA [35]	Raw	0.288	0.528	0.506	0.249	0.497	0.446	0.523	0.576	0.587
	Min [7]	0.325	0.566	0.536	0.295	0.523	0.470	0.524	0.582	0.595
	All [7]	0.417	0.591	0.545	0.305	0.552	0.495	0.495	0.563	0.568
	WSABIE-Min [34]	0.199	0.462	0.459	0.400	0.565	0.521	0.524	0.610	0.621
	PTE-Min [28]	0.238	0.542	0.522	0.452	0.626	0.572	0.545	0.639	0.650
	PLE-NoCo	0.517	0.672	0.634	0.496	0.658	0.603	0.650	0.709	0.703
	PLE	**0.543**	**0.695**	**0.681**	**0.546**	**0.692**	**0.625**	**0.692**	**0.731**	**0.732**
FIGER [14]	Raw	0.474	0.692	0.655	0.369	0.578	0.516	0.467	0.672	0.612
	Min	0.453	0.691	0.631	0.373	0.570	0.509	0.444	0.671	0.613
	All	0.453	0.648	0.582	0.400	0.618	0.548	0.461	0.636	0.583
	WSABIE-Min	0.455	0.646	0.601	0.425	0.603	0.546	0.481	0.671	0.618
	PTE-Min	0.476	0.670	0.635	0.494	0.675	0.618	0.513	0.674	0.657
	PLE-NoCo	0.543	0.726	0.705	0.547	0.699	0.639	0.643	0.753	0.721
	PLE	**0.599**	**0.763**	**0.749**	**0.572**	**0.715**	**0.661**	**0.685**	**0.777**	**0.750**

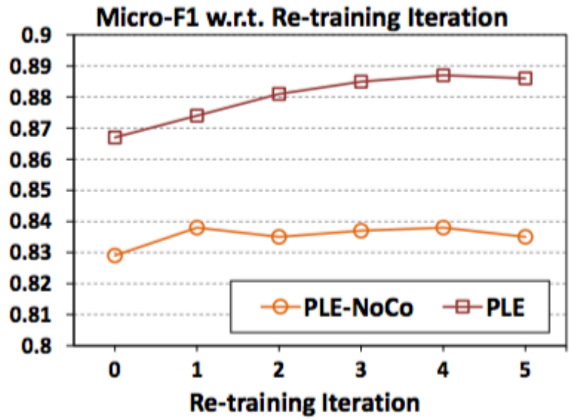
Table 9: Study of performance improvement on fine-grained typing systems **FIGER** [14] and **HYENA** [35] on the three datasets.

- **vs. other noise reduction methods**: the effectiveness of the proposed margin-based loss in modeling noisy candidate types
- **vs. partial-label learning methods**: PLE obtains superior performance because it jointly models type correlation derived from KB and feature-mention co-occurrences in the corpus

Case Analyses

- Testing at different type levels
  - It is more difficult to distinguish among deeper (more fine-grained) types.
  - PLE always outperforms the other two method, and achieves a 153% improvement in Accuracy.





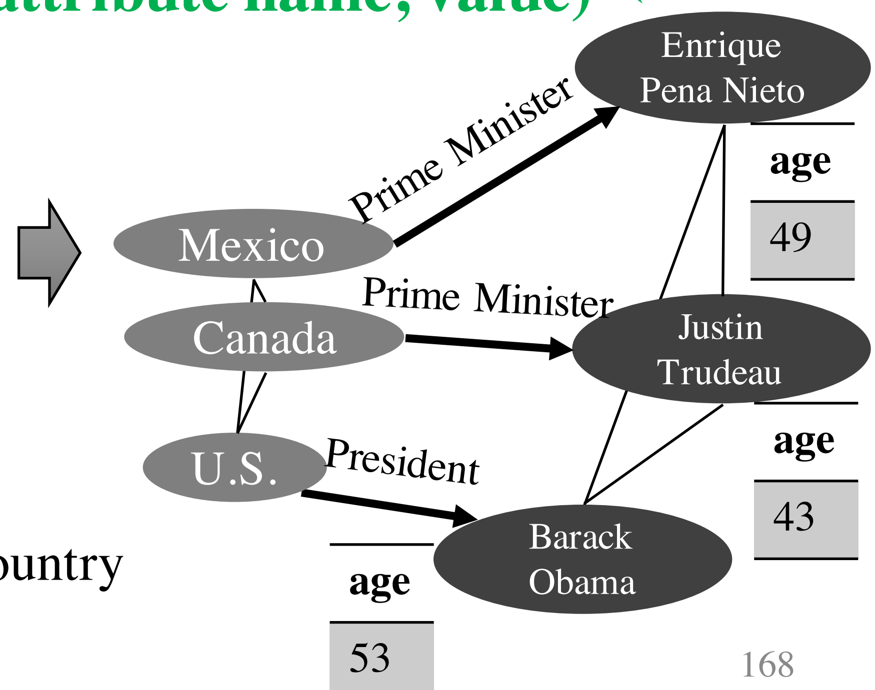
- Iterative re-training of PLE
  - Analyze the effect of boostrapping PLE
  - The performance gain becomes marginal after 3 iterations of re-training

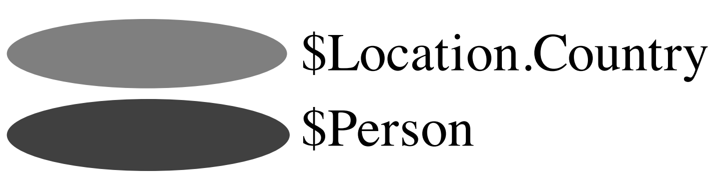
Construction of Heterogeneous Information Networks from Text

Philosophy: Not extensive "labeling" but exploring the power of massive text corpora!

- Mining phrases (the minimal semantic units)
- Entity recognition and typing
- **Attribute discovery (entity, attribute name, value)**

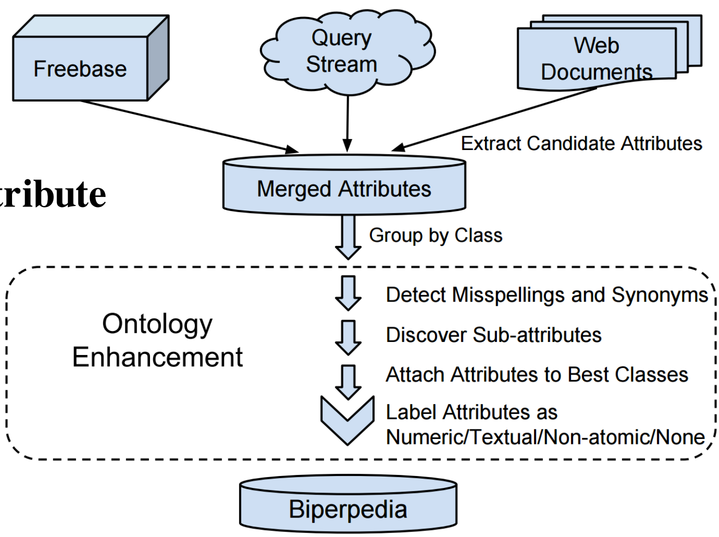
…here by **Canada Prime Minister Justin Trudeau, 43**, the so-called #APEChottie…of **Mexico's Enrique Pena Nieto, 49**, … **United States President Barack Obama, 53**, who…

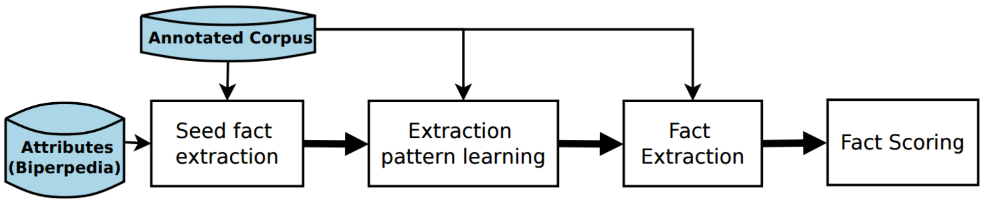




Google's Approaches on Attribute Extraction

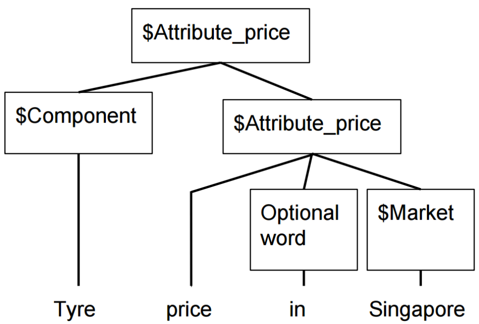
- Given Google's **query log**, web text and knowledge bases
  - "Obama wife name"… "Japan asian population", "Brazil female latino population", "Princeton economist"…
  - "Obama's wife, Michelle Obama, is a lawyer...", "Princeton economist Paul Krugman was awarded..."…
  - Obama: *$Person, $President*; Japan, Brazil: *$Location, $Country*; Princeton: *$Organization, $University*…
- Biperpedia (VLDB'14): **Attribute Name Extraction** from query log
  - $Person: wife name, daughter name
  - $Country: asian population, female latino population
  - $University: economist
- ReNoun (EMNLP'14): **Fact Extraction for Noun Phrase Attribute**
  - (Obama, wife, Michelle Obama)
  - (Princeton, economist, Paul Krugman)



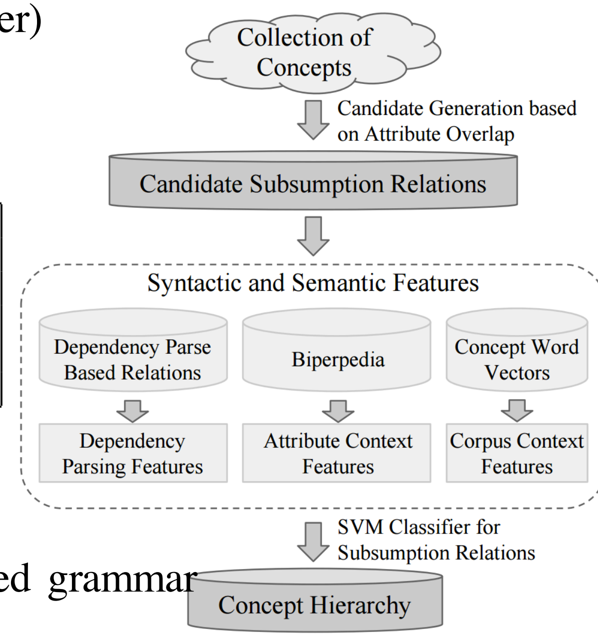


Google's Approaches on Attribute Extraction

- Latte (WebDB'15 Best Paper): **Concept (Type) Hierarchy Extraction** with attribute features
  - {country, address, zip code}: $University (sub) - $Location (super)
  - {online payment, non profit, tax return}: $University (sub) - $Organization (super)
  - {daughter name, wife name, age}: $President (sub) - $Person (super)

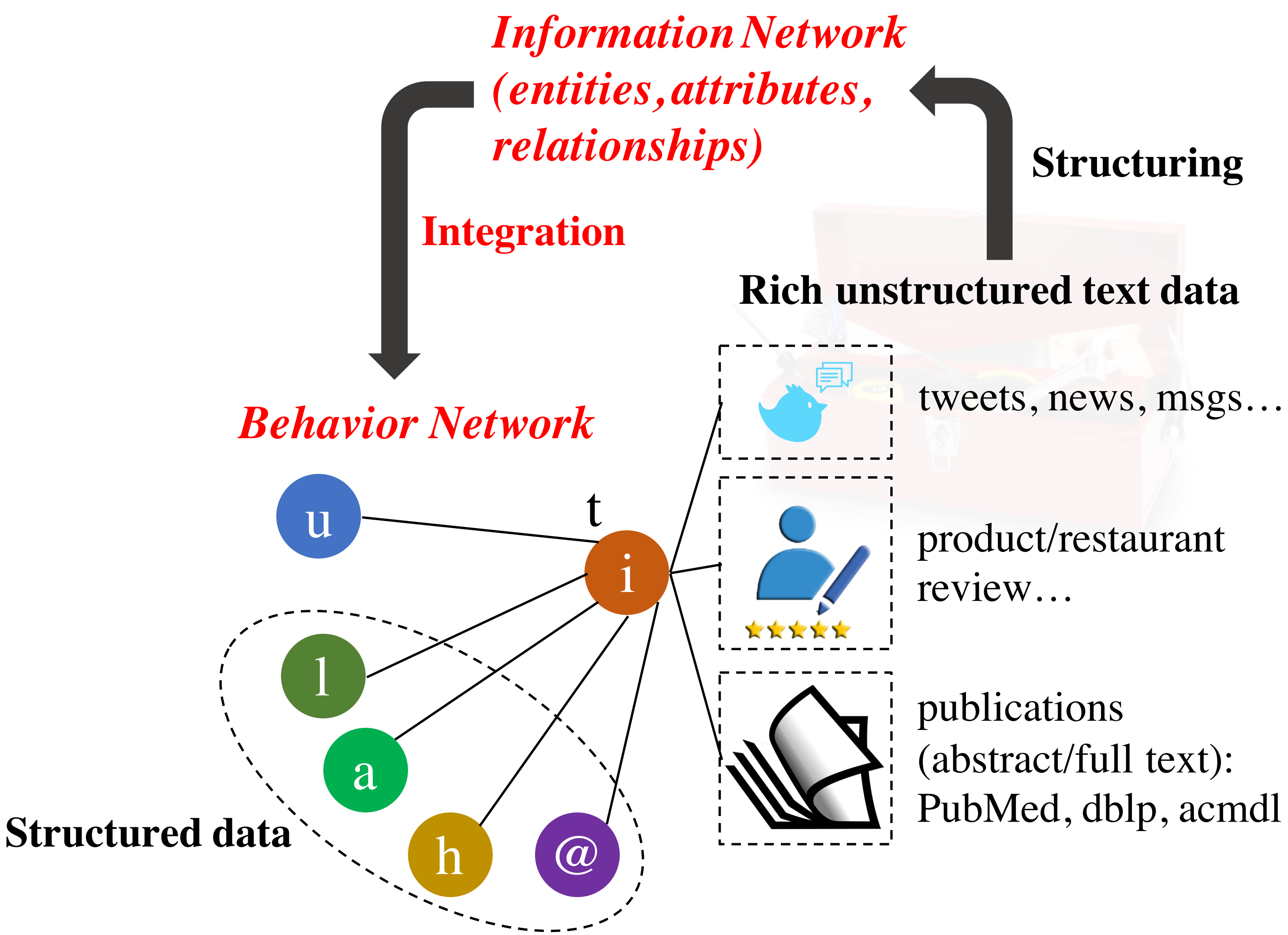



```
$Attribute_price ::= price
$Attribute_price ::= $Market $Attribute_price
$Attribute_price ::= $Attribute_price in $Market
$Attribute_price ::= $Component $Attribute_price
$Market          ::= Singapore | USA | Dubai | London | . . .
$Component       ::= battery | bumper | tyre | door | . . .
```



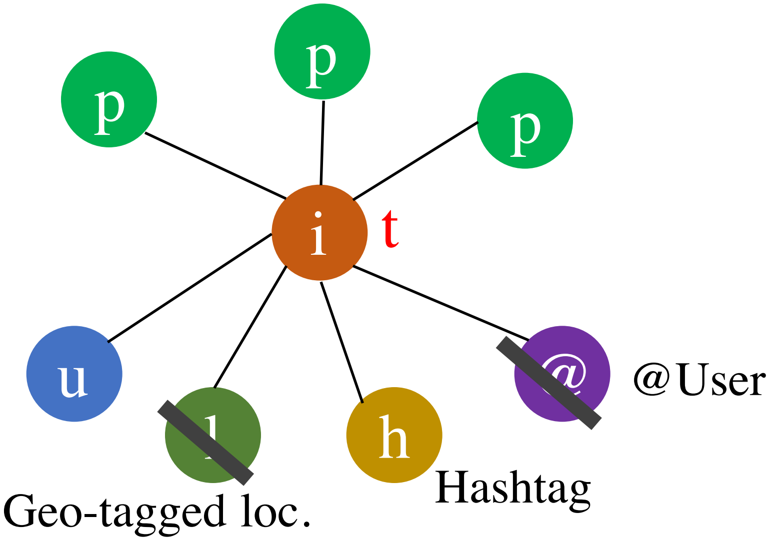
- ARI (WWW'16): **Attribute Name Structure Extraction** with rule-based grammar
  - Long-tail distribution of attribute names
  - $Person: $FamilyMember (name) - daughter, wife, mother, daughter name, wife name
  - $Country: ($Gender) ($Ethnicity) population - asian population, female latino population

Data to Network to Knowledge

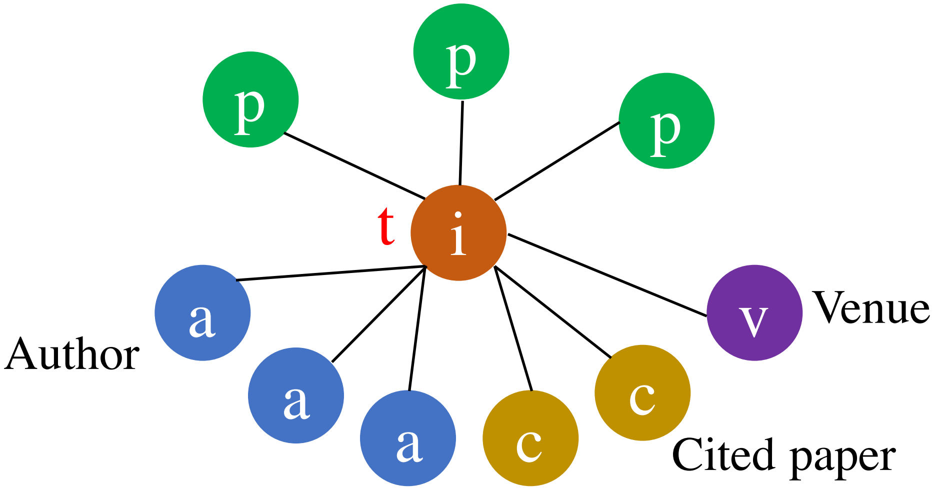


Bring Phrases to Behavior Modeling

- Tweeting behavior
  - Event **summary**



- Paper-publishing behavior
  - Research trend **summary**

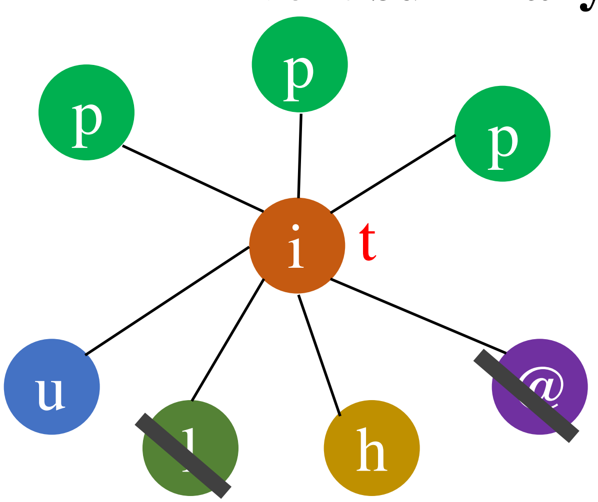


> 20:03:09 @ebekahwsm
> this better be the best halftime show ever in the history of halftimes shows. ever. #SuperBowl

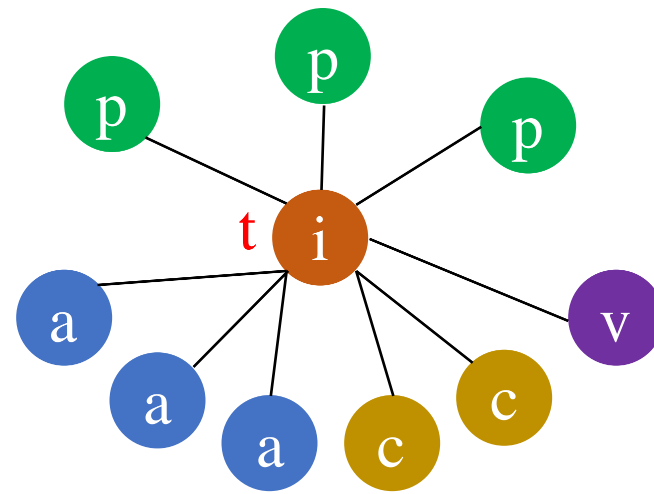
> 2009 P. Melville, W. Gryc, R. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification", KDD'09. Refs: p81623, p84395…

Tensor Fails

- Tweeting behavior
  - Event **summary**



- Paper-publishing behavior
  - Research trend **summary**

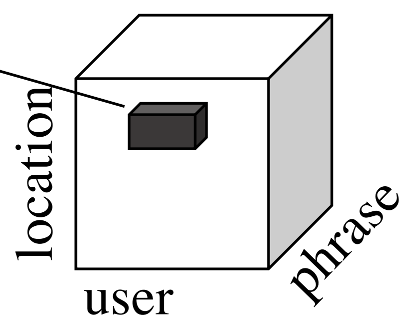


Q: How to represent and summarize **dynamic multi-contextual** behaviors?

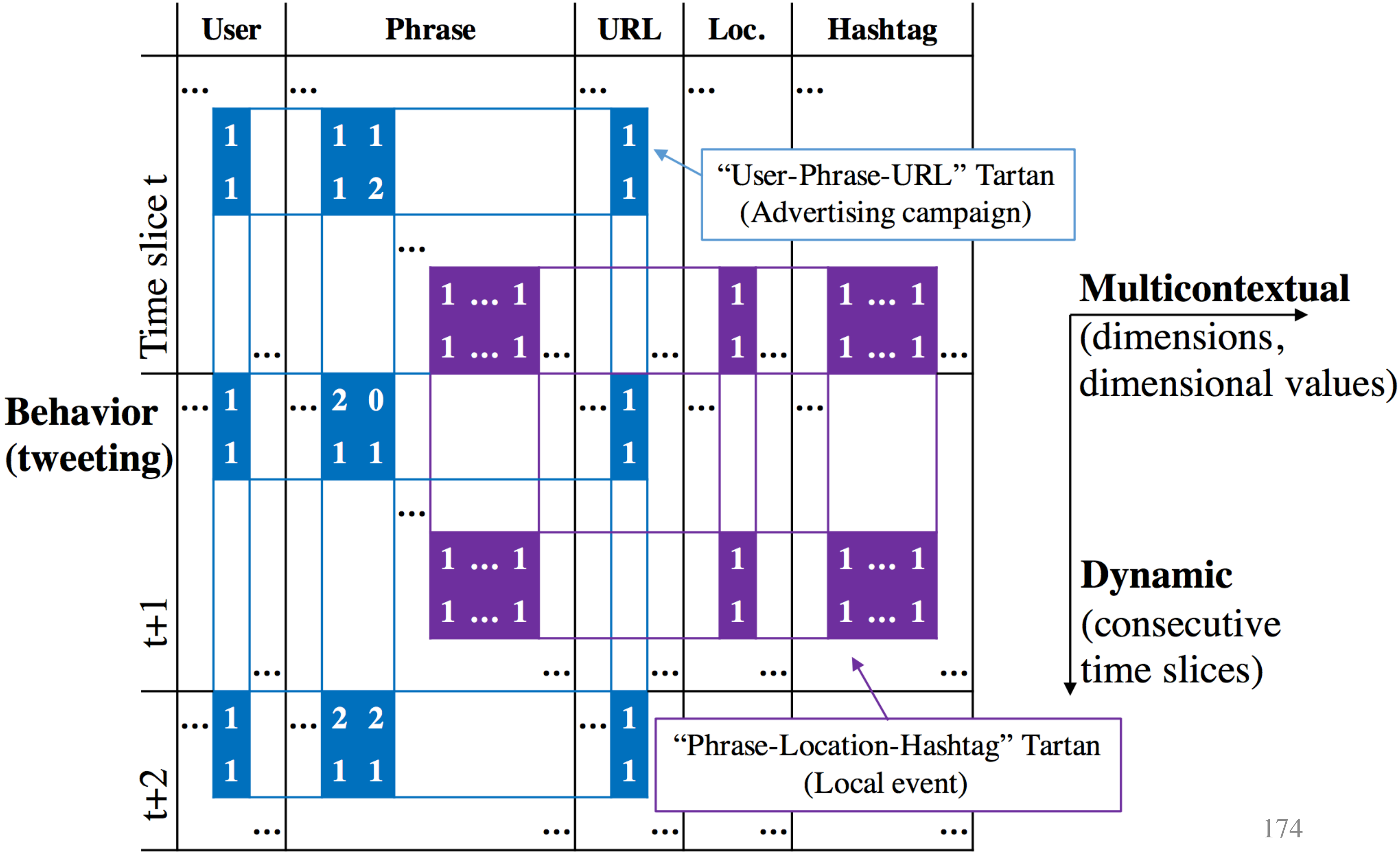
A set of values in dimensions (*one-guaranteed value*, *empty value*, *multi-values*)

✔ ☐ ☐

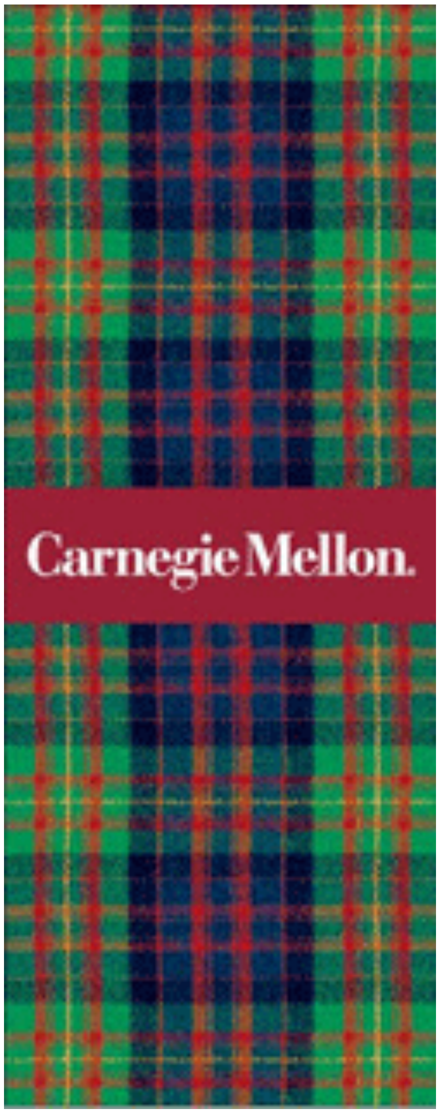
summary
(dense block)



Two-Level Matrix and "Tartan"



CMU Tartans

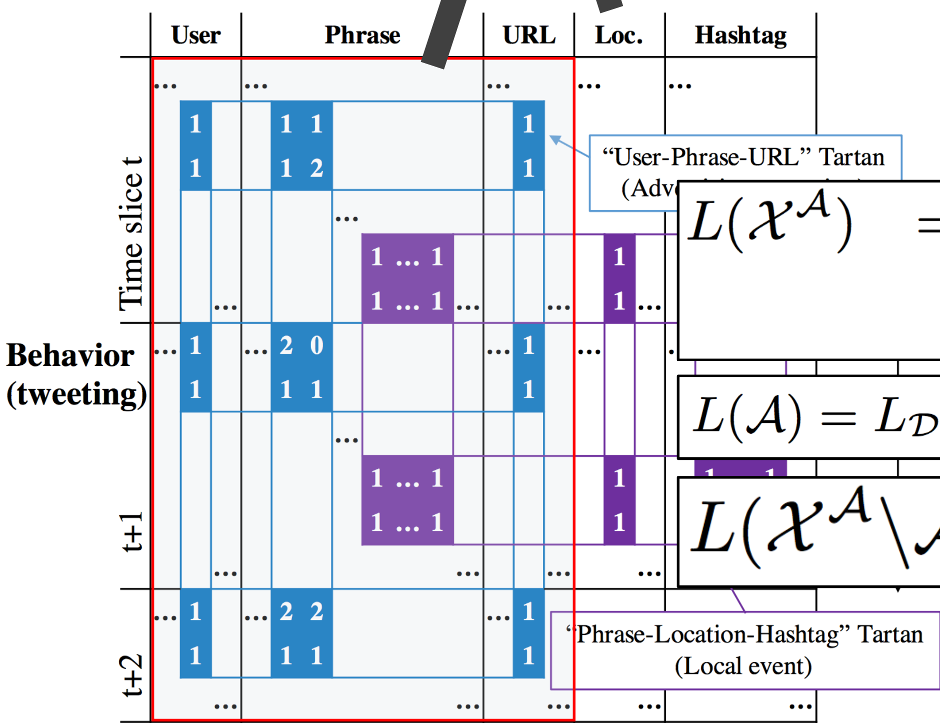


CARNEGIE MELLON UNIVERSITY
PITTSBURGH PENNSYLVANIA 1900

Optimize with MDL Principle

- *Maximize* the number of bits by encoding the Tartan

$$\boxed{f(\mathcal{A},\mathcal{X}) = L(\mathcal{X}^{\mathcal{A}}) - L(\mathcal{A}) - L(\mathcal{X}^{\mathcal{A}}\backslash\mathcal{A}).}$$

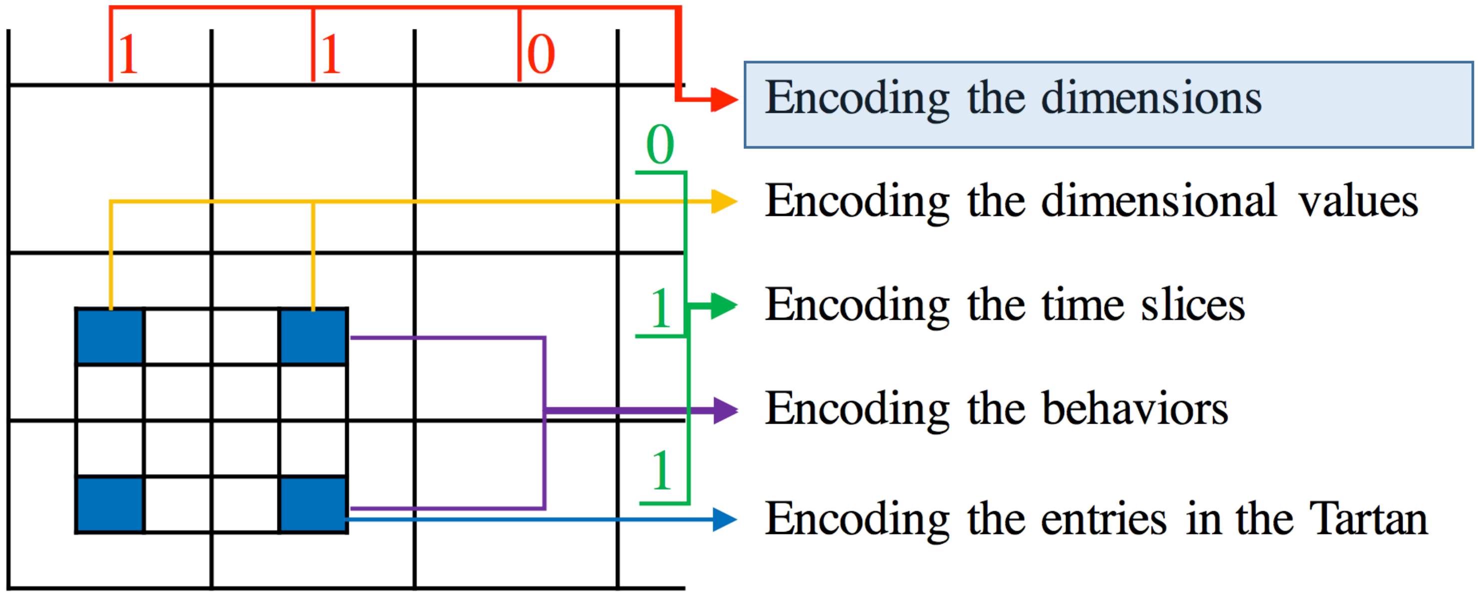


$$L(\mathcal{X}^{\mathcal{A}}) = g(V+C,C) + L_{\mathcal{D}}(\mathcal{A}) + L_{\mathcal{T}}(\mathcal{A}) + \sum\nolimits_{d\in\mathcal{D}} \log^* N_d + \sum\nolimits_{t\in\mathcal{T}} \log^* E^{(t)}.$$

$$L(\mathcal{A}) = L_{\mathcal{D}}(\mathcal{A}) + L_{\mathcal{V}}(\mathcal{A}) + L_{\mathcal{T}}(\mathcal{A}) + L_{\mathcal{B}}(\mathcal{A}) + L_{\mathcal{A}}(\mathcal{A}).$$

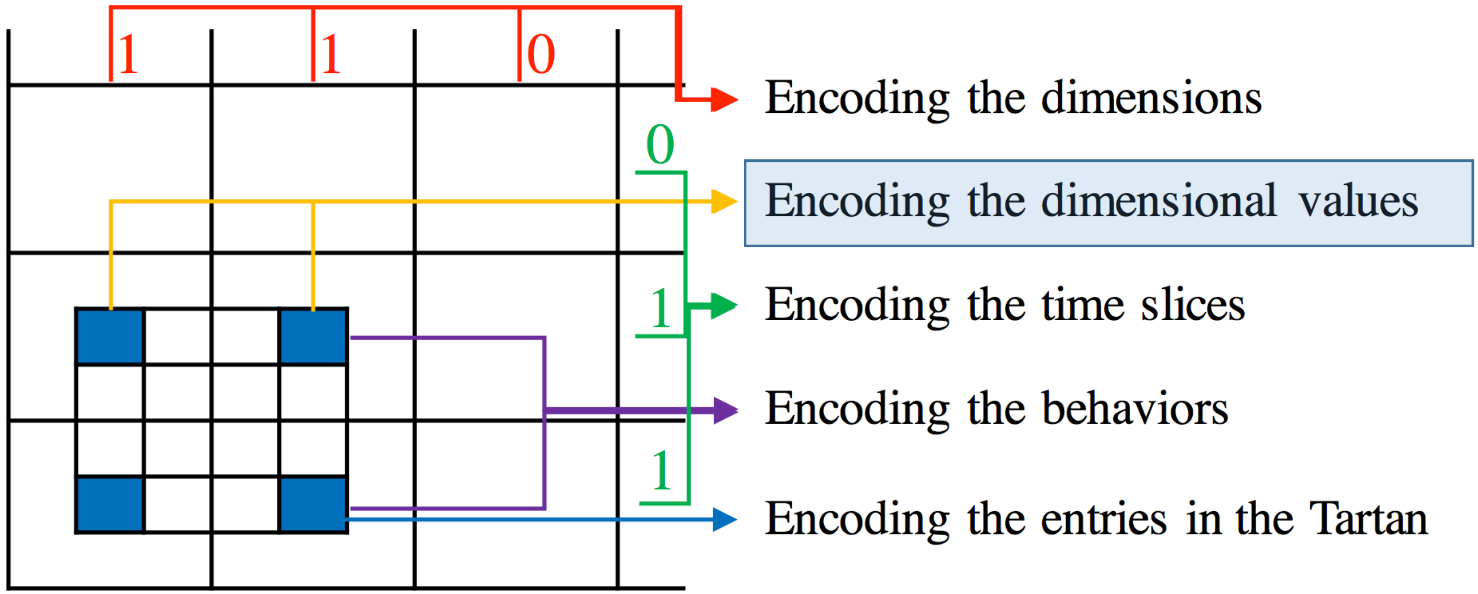
$$L(\mathcal{X}^{\mathcal{A}}\backslash\mathcal{A}) = g(V+C-v-c, C-c);$$

Encoding Tartan: Dimensions



$$
\begin{aligned}
H_{\mathcal{D}}(X) &= -\sum_{x\in\{0,1\}} P(X=x)\log P(X=x) \\
&= -\left(\frac{D^{\mathcal{A}}}{D}\log\frac{D^{\mathcal{A}}}{D} + \frac{D-D^{\mathcal{A}}}{D}\log\frac{D-D^{\mathcal{A}}}{D}\right). \\
L_{\mathcal{D}}(\mathcal{A}) &= \log^* D + \log^* D^{\mathcal{A}} + D\cdot H_{\mathcal{D}}(X) \\
&= \log^* D + \log^* D^{\mathcal{A}} + g(D, D^{\mathcal{A}}),
\end{aligned}
$$

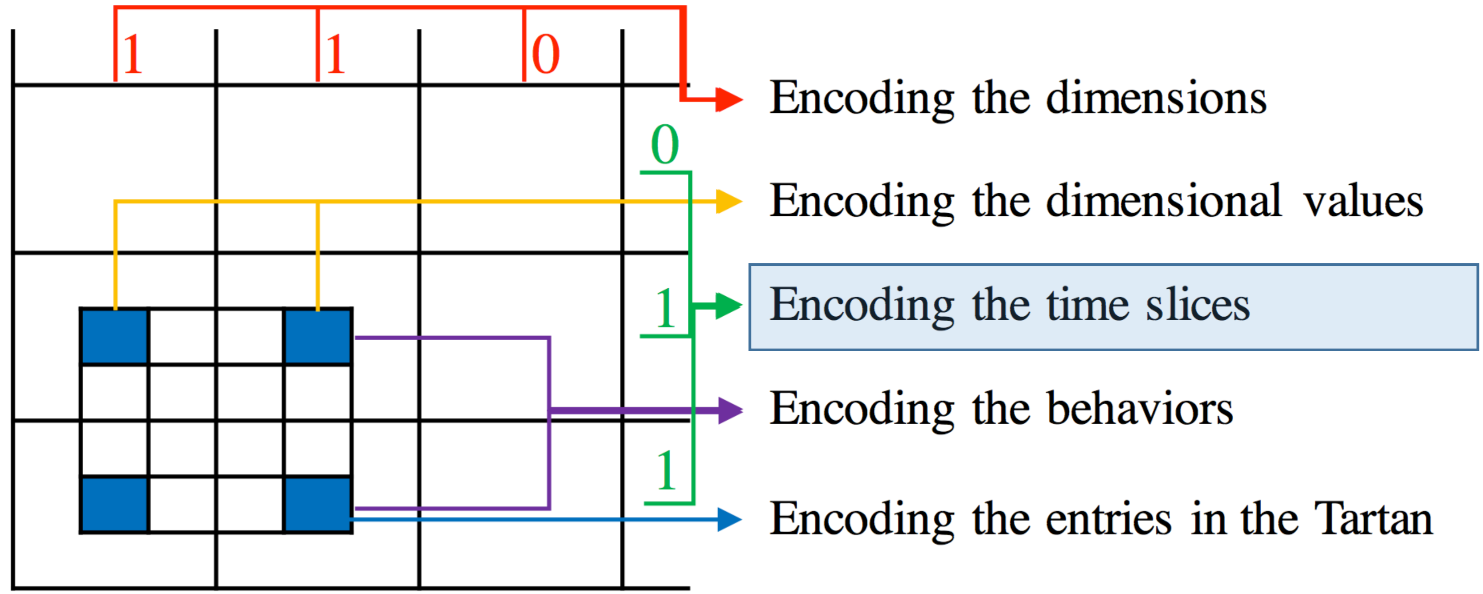
Encoding Tartan: Dimensional Values



$$H_{\mathcal{V}_d}(X) = -\left(\frac{n_d}{N_d}\log\frac{n_d}{N_d} + \frac{N_d-n_d}{N_d}\log\frac{N_d-n_d}{N_d}\right).$$

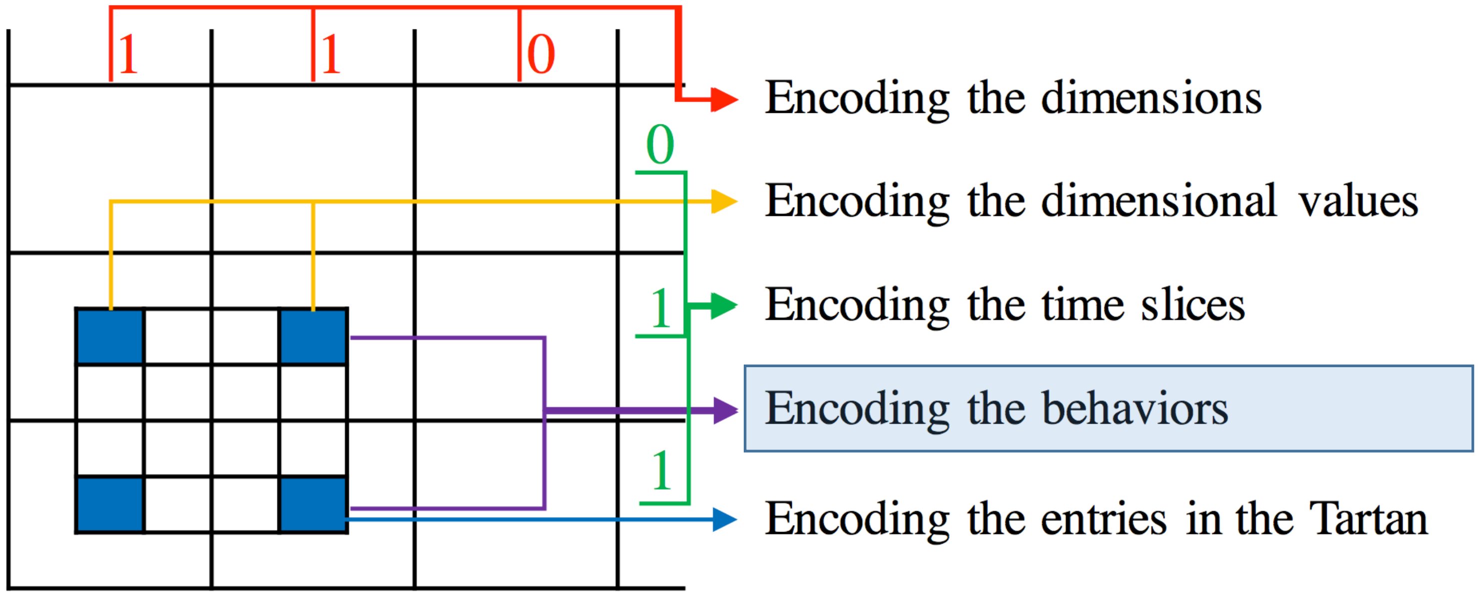
$$L_{\mathcal{V}}(\mathcal{A}) = \sum_{d\in\mathcal{D}}\left(\log^* N_d + \log^* n_d + g(N_d, n_d)\right).$$

Encoding Tartan: Time Slices



$$L_{\mathcal{T}}(\mathcal{A}) = \log^* T + \log^* T^{\mathcal{A}} + \log^* t_{start}$$

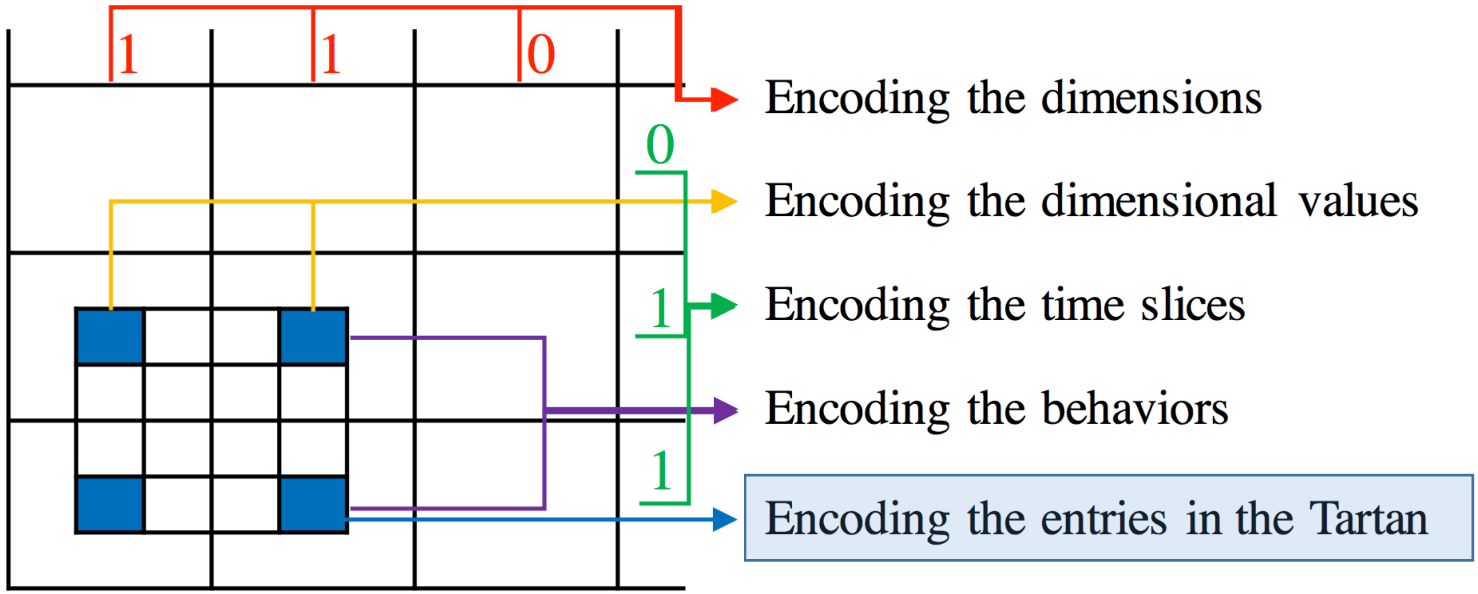
Encoding Tartan: Behaviors



$$H_{\mathcal{B}^{(t)}}(X) = -\left(\frac{e^{(t)}}{E^{(t)}} \log \frac{e^{(t)}}{E^{(t)}} + \frac{E^{(t)}-e^{(t)}}{E^{(t)}} \log \frac{E^{(t)}-e^{(t)}}{E^{(t)}}\right).$$

$$L_{\mathcal{B}}(\mathcal{A}) = \sum_{t\in\mathcal{T}} \left(\log^* E^{(t)} + \log^* e^{(t)} + g(E^{(t)}, e^{(t)})\right).$$

Encoding Tartan: Entries



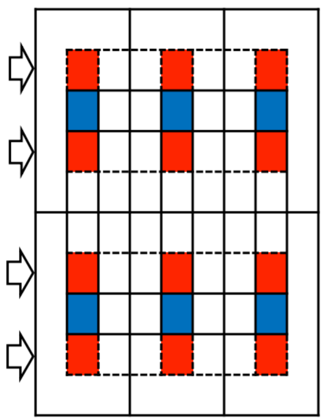
$$v = \left(\sum_{d\in\mathcal{D}} n_d\right)\left(\sum_{t\in\mathcal{T}} e^{(t)}\right).$$

$$c = \sum_{d\in\mathcal{D},t\in\mathcal{T}} \sum_{b\in\mathcal{B}^{(t)},i\in\mathcal{V}_d} \mathcal{X}_d^{(t)}(b,i).$$

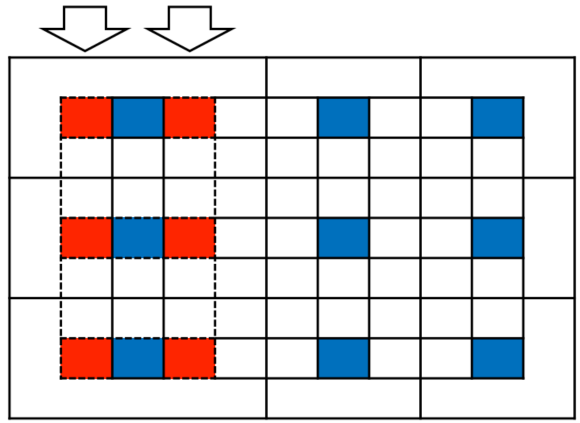
$$H_{\mathcal{A}}(X) = -\left(\frac{c}{v+c}\log\frac{c}{v+c} + \frac{v}{v+c}\log\frac{v}{v+c}\right).$$

$$L_{\mathcal{A}}(\mathcal{A}) = (v+c)H_{\mathcal{A}}(X) = g(v+c,c).$$

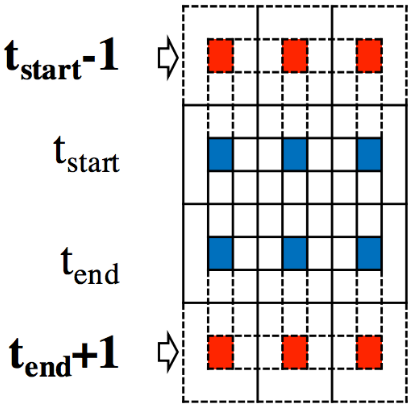
Greedy Search for the Local Optimum



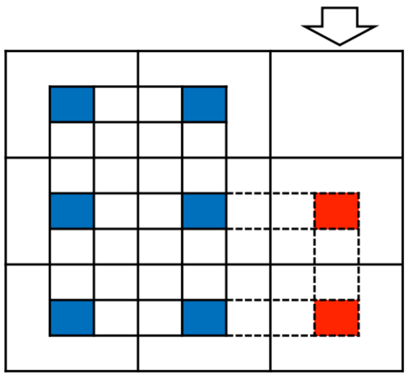
(a) Update the set of behaviors.



(b) Update the set of values.



(c) Update the consecutive time slices.



(d) Update the set of dimensions.

Time complexity:

$$\mathcal{O}(\sum_d N_d \log N_d + \sum_t E^{(t)} \log E^{(t)})$$

Experimental Results

- DM/ML research trend summaries with DBLP data

1997 2000 2003 2006 2009 2012

Author	Venue	Keyword	Cited	#Paper
76 Cheng-xiang Zhai Hui Fang S. Kambhampati	**7** SIGIR VLDB TKDE	**7** "information retrieval" "data integration" "text classification"	**68** p56743[1] p62995 p76869	**32** 2003- 2007

[1] "A language modeling approach to information retrieval"

Venue	Keyword	#Paper
5 ICML NIPS …	**6** "reinforcement learning" "machine learning"	**40** 1997- 2002

Author	Venue	Cited	#Paper
6 Jiawei Han Xifeng Yan	**1** SIG- MOD	**1** p76095[2]	**22** 2004- 2010

[2] "Frequent subgraph discovery"

Venue	Keyword	#Paper
3 ICDM AAAI TKDE	**1** "anomaly detection"	**25** 2005- 2013

Author	Venue	Keyword	#Paper
27 C. Faloutsos J. Pei P. S. Yu X. Lin C. Aggarwal…	**6** KDD ICDM ICDE TKDE …	**12** "large graphs" "data streams" "evolving data" "evolving graphs" …	**70** 2006- 2013

Author	Venue	Keyword	Cited	#Paper
12 Ryen White Hang Li Tie-Yan Liu Zhaohui Zheng…	**5** SIGIR WWW WSDM CIKM…	**3** "web search" "click-through data" "sponsored search"	**12** p82630[3] p116290 p103899 p106191…	**32** 2006- 2013

[3] "Optimizing search engines using clickthrough data"

Author	Venue	Keyword	#Paper
8 Qiang Yang Dou Shen Sinno Pan…	**3** KDD PAKDD AAAI	**6** "transfer learning" "data mining" "localization models"	**17** 2007- 2010

Experimental Results

- Event summaries with Super Bowl 2013 tweets

Time axis: 16:30, 17:00, 17:30, 18:00, 18:30, 19:00, 19:30, 20:00, 20:30, 21:00, 21:30, 22:00

Tweets	Phrase	Dimensions	Count	Tartan
16:30:31 My prediction Ravens 34 Niners 31 16:30:57 Ready for the big game :D, my prediction 24-20 SF #SuperBowl 16:31:14 My prediction for superbowl.. 48.. Jets over Bears 17-13 Mark Sanchez MVP 16:32:24 I predict Baltimore Ravens will win 27 to 24 or 25 or 26. Basically it will be a close game.	"my prediction"	user (3,325); **phrase 226**; hashtag (0); URL (0)	3,397 tweets	Tartan #1: (1 dim) 16:30-17:30
17:30:51 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist 17:31:01 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist 17:31:16 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist 17:31:19 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist	"make your prediction"	user (196); **phrase 4**; **RT @user 1**; **URL 1**	196 tweets	Tartan #2: (3 dims) 17:00-18:00
18:55:03 RT @49ers: Kaepernick is sacked on 3rd and goal. #49ers K David Akers makes 36-yard FG. Baltimore leads 7-3 with 3:58 left in 1st Qtr. #SB47 18:55:04 RT @49ers: Kaepernick is sacked on 3rd and goal. #49ers K David Akers makes 36-yard FG. Baltimore leads 7-3 with 3:58 left in 1st Qtr. #SB47 18:55:44 RT @Ravens: David Akers is good from 36 yards to make the score 7-3 Ravens. Nice job by the defense to tighten up in the red zone.	"7-3", "1st Qtr"	user (213); **phrase 21**; **RT @user 3**; URL (0)	215 tweets	Tartan #3: (2 dims) 18:30-19:30
20:20:01 RT @ExtraGrumpyCat: No Superbowl halftime show will ever surpass this. http://t.co/0VSy7Cv6 20:20:02 RT @WolfpackAlan: No Superbowl halftime show will ever surpass this. http://t.co/6BlloPXs 20:20:04 RT @ExtraGrumpyCat: No Superbowl halftime show will ever surpass this. http://t.co/0VSy7Cv6 20:20:05 RT @WolfpackAlan: No Superbowl halftime show will ever surpass this. http://t.co/6BlloPXs	halftime show"	user (617); **phrase 11**; **RT @user 4**; **URL 4**	617 tweets	Tartan #4: (3 dims) 20:00-21:00
20:20:47 (Manhattan, NY)...and every one of those girls took #ballet #Beyonce #superbowl 20:22:01 (New York, NY) I have the biggest lady boner for Beyonce #BeyonceBowl #DestinyBowl #DestinysChild #SuperBowl 20:24:32 (Manhattan, NY) No one can ever top that performance by Beyonce. EVER. #Beyonce #superbowl #halftimeshow	"beyonce", #beyonce, #superbowl, #DestinysChild	**location 2**; **phrase 55**; **hashtag 17**; URL (0)	166 tweets	Tartan #5: (3 dims) 20:00-21:00
21:44:42 Ahora si pff #49ers 23-28 #Ravens 21:44:44 Baltimore #Ravens 28-23 San Francisco #49ers 21:44:50 FG Akers #49ers 23-28 #Ravens 3Q 3:10 #SuperBolwXLVII #SuperBowl #NFL	"28-23", #49ers, #Ravens	user (650); **phrase 69**; **hashtag 11**; URL (0)	653 tweets	Tartan #6: (2 dims) 21:00-22:00
22:42:27 Congratulations Ravens!!!! 22:42:43 Congratulations Ray Lewis and the Ravens. 22:42:43 Game over.! Ravens won ray got his retirement ring now all y'all boys and girls go to sleep .! 22:42:52 "@LetThatBoyTweet: Game over. Ravens win the Super Bowl."	"congratulations", "game over"	user (1942); **phrase 248**; hashtag (0); URL (0)	1,950 tweets	Tartan #7: (1 dim) 22:00-23:30

Summary

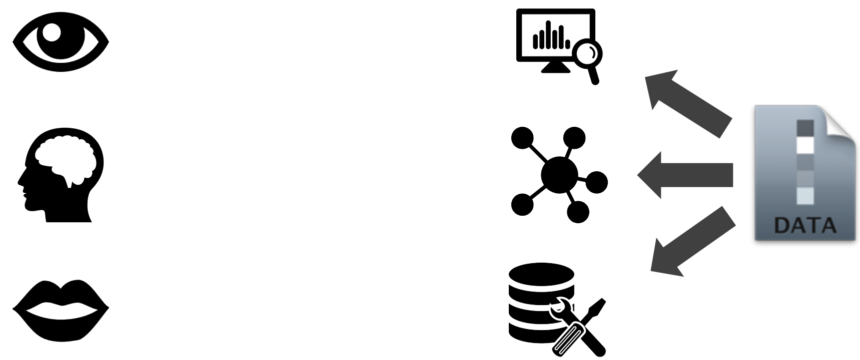
- Structuring text into heterogeneous information networks
- **Observations, Representations, Models**
  - **ToPMine/SegPhrase:** Quality phrase mining
  - **ClusType:** Entity recognition and typing
  - **MetaPAD:** Data-driven automatic attribute discovery for attributed network construction
    - Integrating text mining techniques
    - **Meta Pattern** Mining
- Integrating phrases into behavioral analysis
- **Observations, Representations, Models**
  - **CatchTartan:** Dynamic multicontextual. Tensor fails.

Experience-Driven **Data-Driven**

Observation

Representation

Models/Algorithms

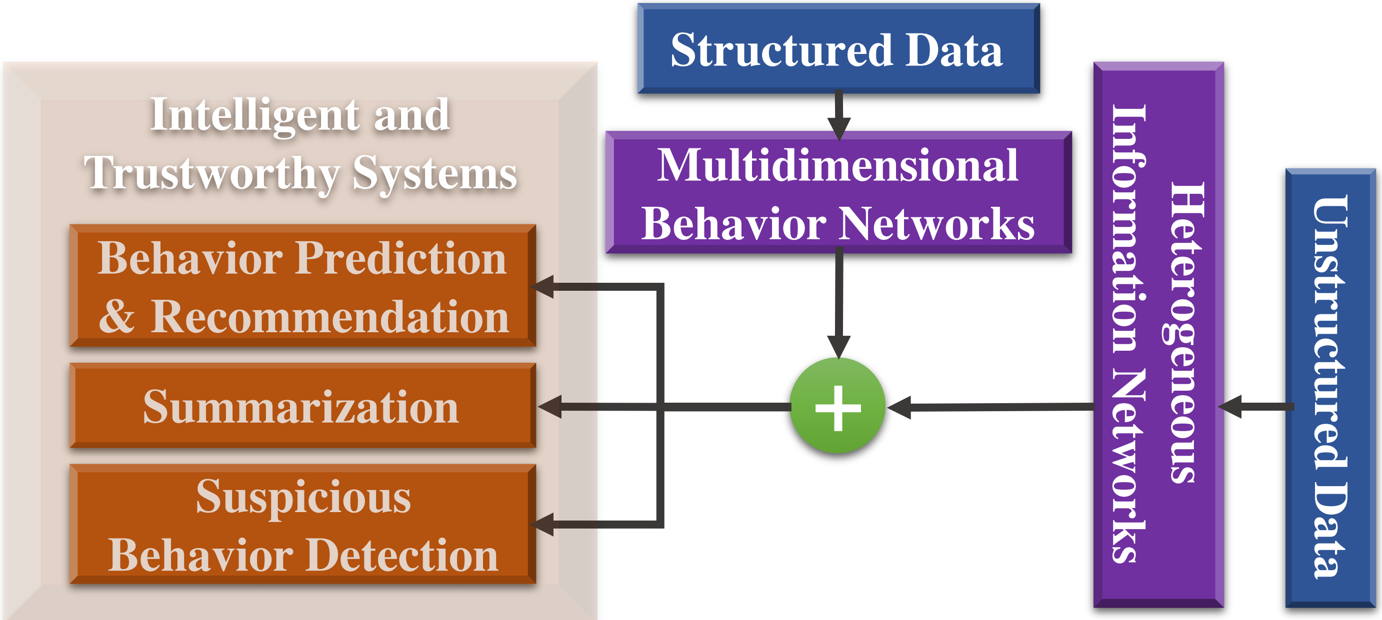


Conclusion

Data-Driven Behavioral Analytics

Data-Driven Behavioral Analytics

- Mining **behavior networks** with **social and spatiotemporal contexts** to support intelligent and trustworthy systems
  - Mining for **behavior prediction and recommendation**
  - Mining for **suspicious behavior detection**
- Structuring behavioral content and integrating behavioral analysis with **information networks**



Acknowledgement

NSF
NSFC National Natural Science Foundation of China
Tsinghua University 1911
Carnegie Mellon University
University of Maryland 1856
Microsoft Research 微软亚洲研究院
Tencent 腾讯

References

D. Blei, A. Ng, and M. Jordan. "Latent dirichlet allocation." JMLR, 2003.

J. Herlocker, J. Konstan, L. Terveen, J. Riedl. "Evaluating collaborative filtering recommender systems." ACM TOIS, 2004.

Y. Koren, R. Bell, C. Volinsky. "Matrix factorization techniques for recommender systems." Computer, 2009.

Y. Koren. "Factorization meets the neighborhood: A multifaceted collaborative filtering model." KDD, 2008.

Y. Koren. "Collaborative filtering with temporal dynamics." CACM, 2010.

M. Balabanovic and Y. Shoham. "FAB: Content-based, collaborative recommendation." CACM, 1997.

N. Liu and Q. Yang. "Eigenrank: A ranking-oriented approach to collaborative filtering." SIGIR, 2008.

N. Liu, M. Zhao, and Q. Yang. "Probabilistic latent preference analysis for collaborative filtering." CIKM, 2009.

References

H. Ma, H. Yang, M. Lyu, and I. King. "Sorec: Social recommendation using probabilistic matrix factorization." CIKM, 2008.

H. Ma, T. Zhou, M. Lyu, and I. King. "Improving recommender systems by incorporating social contextual information." ACM TOIS, 2011.

H. Ma, D. Zhou, C. Liu, M. Lyu, and I. King. "Recommender systems with social regularization." WSDM, 2011.

J. Leskovec, A. Singh, and J. Kleinberg. "Patterns of influence in a recommendation network." PAKDD, 2006.

P. Massa and A. Paolo. "Trust-aware recommender systems." RecSys, 2007.

M. Jamali and E. Martin. "TrustWalker: A random walk model for combining trust-based and item-based recommendation." KDD, 2009.

H. Ma, I. King, and M. Lyu. "Learning to recommend with social trust ensemble." SIGIR, 2009.

H. Ma. I. King, and M. Lyu. "Learning to recommend with explicit and implicit social relations." ACM TIST, 2011.

References

M. Faloutsos, P. Faloutsos, and C. Faloutsos. "On power-law relationships of the internet topology." SIGCOMM, 1999.

A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Weiner. "Graph structure in the web." Computer Networks, 2000.

F. Chung and L. Lu. "The average distances in random graphs with given expected degrees." PNAS, 2002.

J. Kleinberg. "Authoritative sources in a hyperlinked environment." JACM, 1999.

H. Kwak, C. Lee, H. Park, and S. Moon. "What is Twitter, a social network or a news media?" WWW, 2010.

B. Hooi, H.A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos. "Fraudar: Bounding graph fraud in the face of camouflage." KDD, 2016.

C. Aggarwal and J. Han. "Frequent pattern mining." Springer, 2014.

J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu. "FreeSpan: Frequent pattern-projected sequential pattern mining." KDD, 2000.

References

X. Yan and J. Han. "gspan: Graph-based substructure pattern mining." ICDM, 2003.

X. Yan and J. Han. "CloseGraph: Mining closed frequent graph patterns." KDD, 2003.

Y. Sun, J. Han, X. Yan, P.S. Yu, and T. Wu. "PathSim: Meta path-based top-k similarity search in heterogeneous information networks." VLDB, 2011.

Y. Sun, Y. Yu, and J. Han. "Ranking-based clustering of heterogeneous information networks with star network schema." KDD, 2009.

Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. "RankClus: Integrating clustering with ranking for heterogeneous information network analysis." EDBT, 2009.

Y. Sun, R. Barber, M. Gupta, C. Aggarwar, and J. Han. "Co-author relationship prediction in heterogeneous bibliographic networks." ASONAM, 2011.

A. El-Kishky, Y. Song, C. Wang, C.R. Voss, and J. Han. "Scalable topical phrase mining from text corpora." VLDB, 2014.

J. Liu, J. Shang, C. Wang, X. Ren, and J. Han. "Mining quality phrases from massive text corpora." SIGMOD, 2015.

References

X. Ren, A. El-Kishky, C. Wang, F. Tao, C.R. Voss, and J. Han. "Effective entity recognition and typing by relation phrase-based clustering." KDD, 2015.

X. Ren, W. He, M. Qu, C.R. Voss, H. Ji, and J. Han. "Label noise reduction in entity typing by heterogeneous partial-label embedding." KDD, 2016.

C. Wang, M. Danilevsky, N. Desai, Y. Zhang, P. Nguyen, T. Taula, and J. Han. "A phrase mining framework for recursive construction of a topical hierarchy." KDD, 2013.

E.E. Papalexakis, C. Faloutsos, N.D. Sidiropoulos. "ParCube: Sparse parallelizable tensor decompositions." PKDD, 2012.

D. Koutra, U. Kang, J. Vreeken, and C. Faloutsos. "VOG: Summarizing and understanding large graphs." SDM, 2014.

R. Gupta, A. Halevy, X. Wang, S.E. Whang, and F. Wu. "Biperpedia: An ontology for search applications." VLDB, 2014.

M. Yahya, S. Whang, R. Gupta, and A. Halevy. "ReNoun: Fact extraction for nominal attributes." EMNLP, 2014.

A. Halevy, N. Noy, S. Sarawagi, S.E. Whang, and X. Yu. "Discovering structure in the universe of attribute names." WWW, 2016.

References

Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. "Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation." SIGMOD, 2014.

Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han. "A confidence-aware approach for truth discovery on long-tail data." VLDB, 2014.

F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han. "Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation." KDD, 2015.

Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. "A survey on truth discovery." KDD Explorations Newsletter, 2016.

S. Zhi, B. Zhao, W. Tong, J. Gao, D. Yu, H. Ji, and J. Han. "Modeling truth existence in truth discovery." KDD, 2015.

S. Kumar, R. West, and J. Leskovec. "Disinformation on the Web: Impact, characteristics, and detection of Wikipedia hoaxes." WWW, 2016.

S. Kumar, F. Spezzano, and V.S. Subrahmanian. "Identifying malicious actors on social media." ASONAM, 2016. (tutorial)

Thank you!

Data-Driven Behavioral Analytics: Observations, Representations and Models