MCD: Mutual Community Detection across Multiple Social Networks

Philip S. Yu[1,2] and Jiawei Zhang[1]
[1]University of Illinois at Chicago
[2]Tsinghua University

Users use multiple social networks simultaneously

foursquare

twitter

Temporal Activities

8 AM 12 PM 4 PM 8 PM 11 PM

Locations

Tips

User Accounts

anchor links

Temporal Activities

User

anchor users

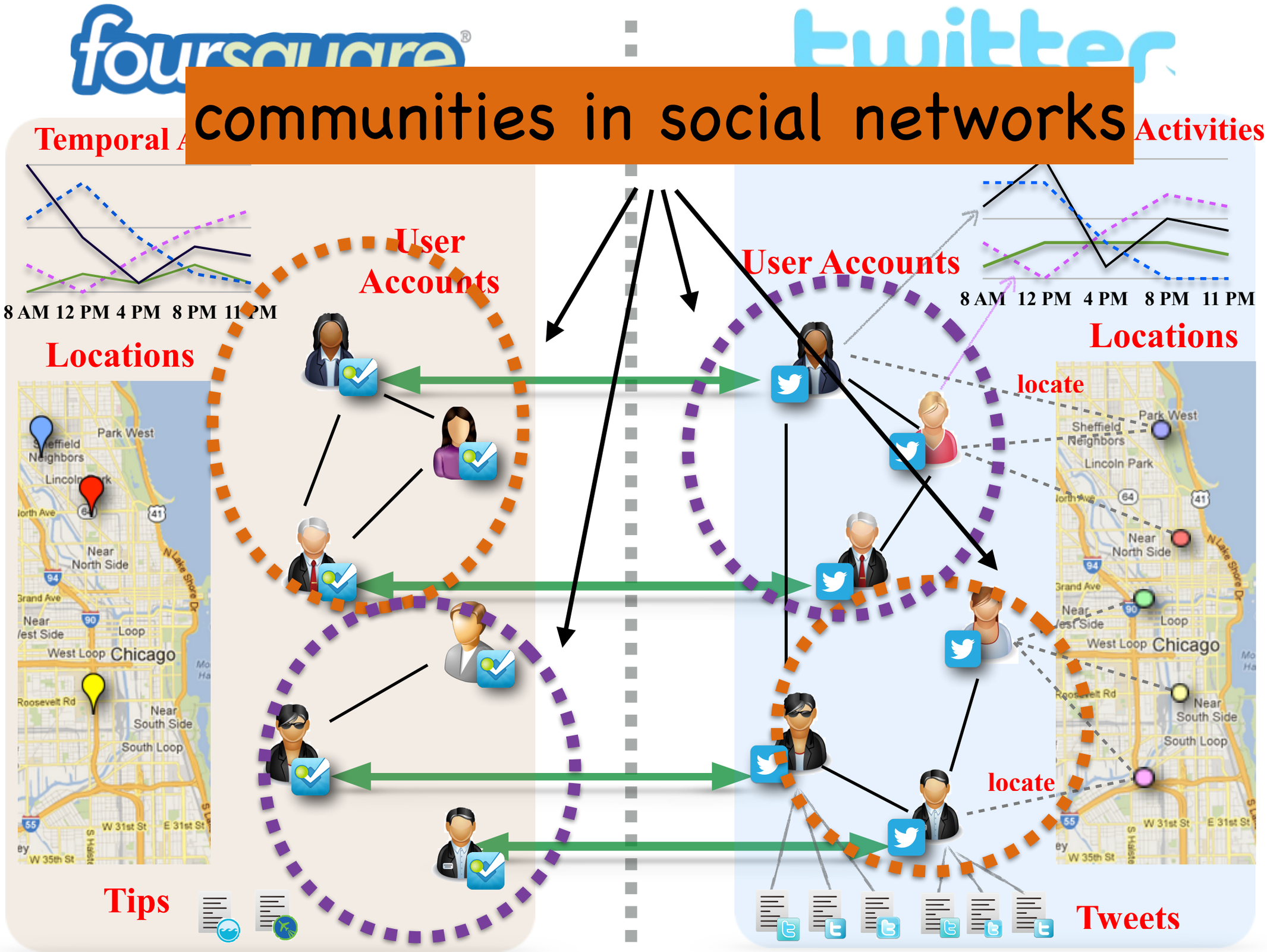
locate

non-anchor users

Partially Aligned Social Networks

Problem Studied: Simultaneous Community Detection of Multiple Partially Aligned Networks

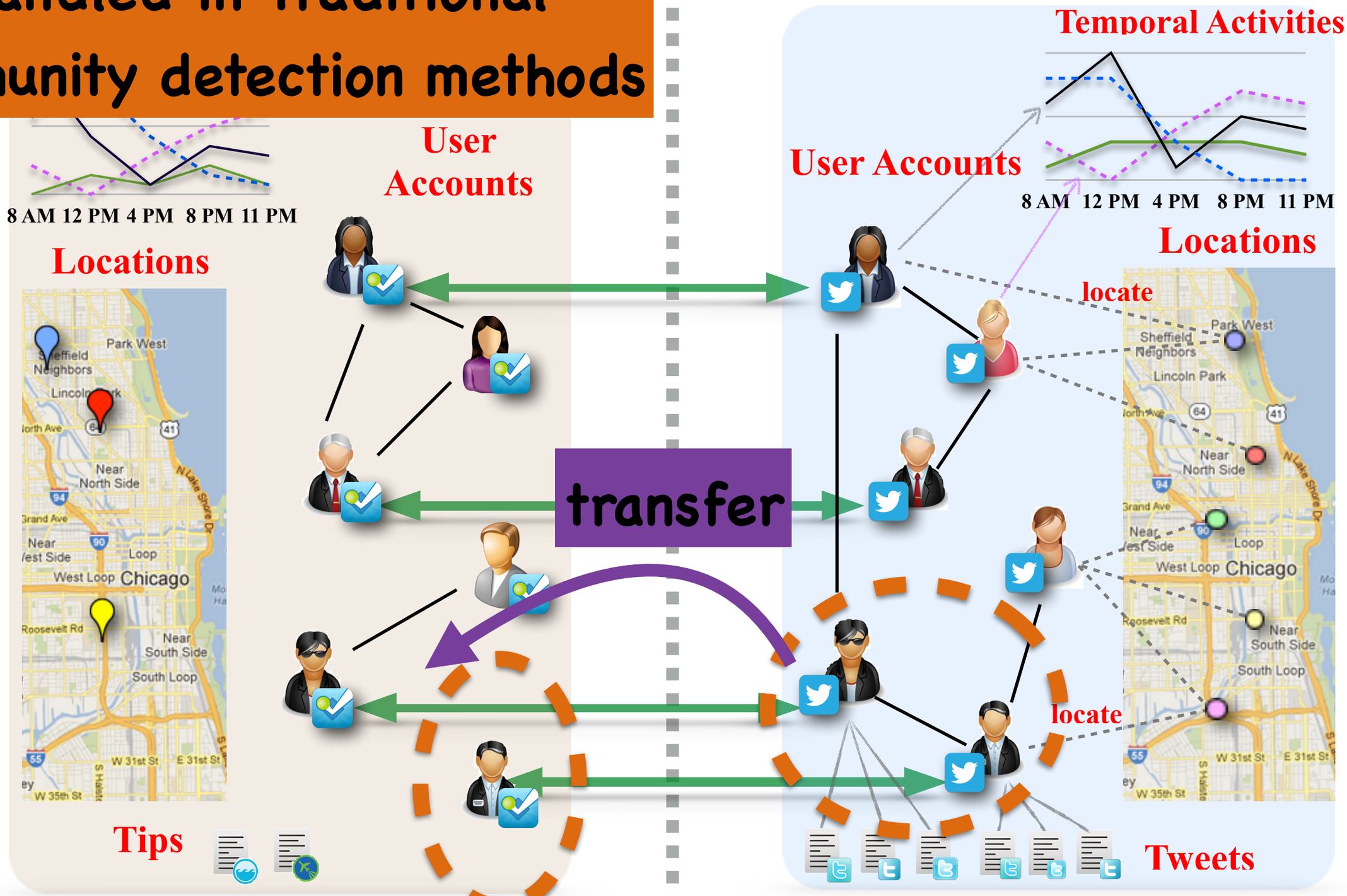
communities in social networks



Motivations: Mutual community detection

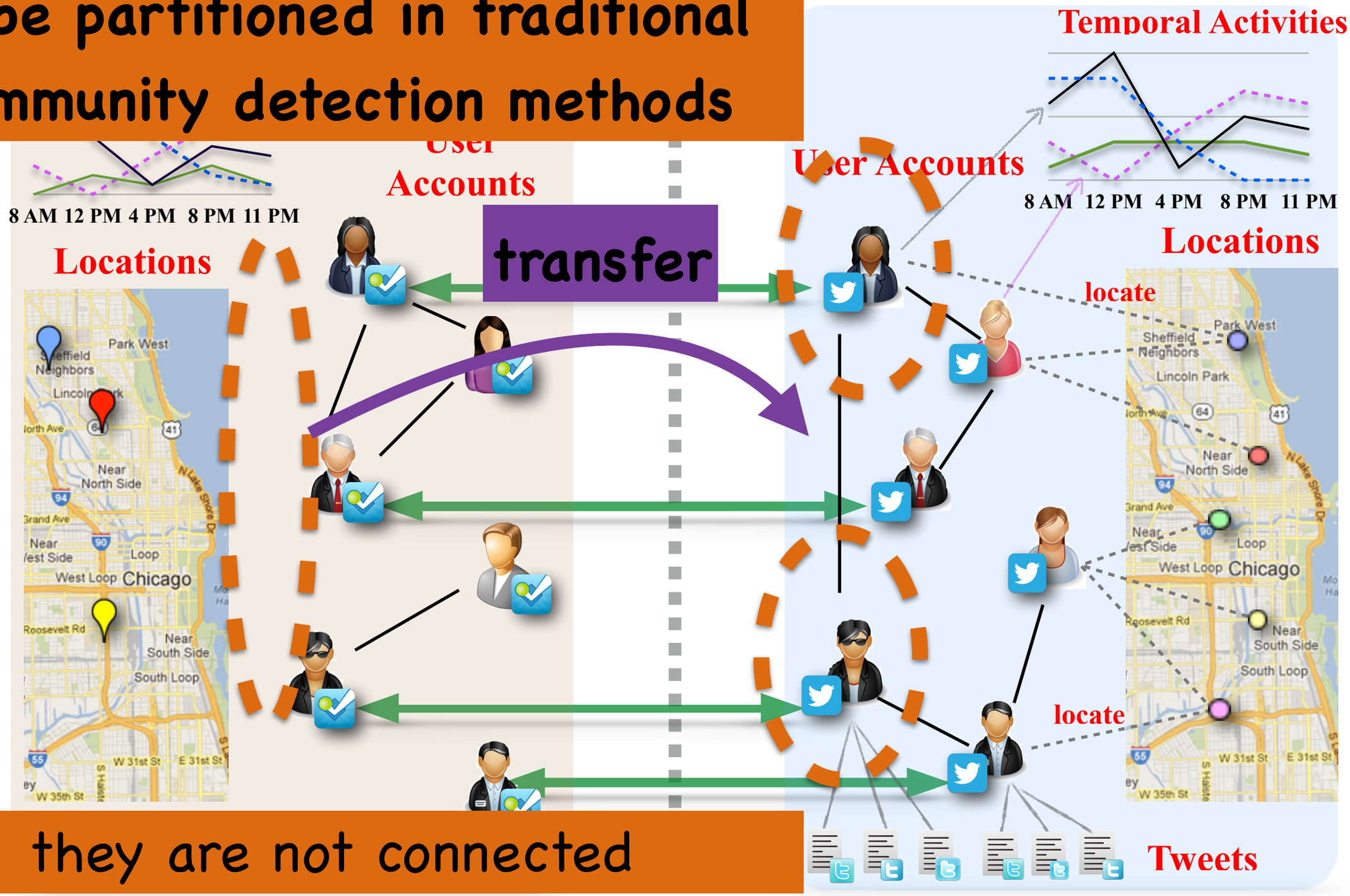
isolated user: cannot be handled in traditional community detection methods

twitter



Motivations: Mutual community detection

community boundary users: hard to be partitioned in traditional community detection methods



they are not connected

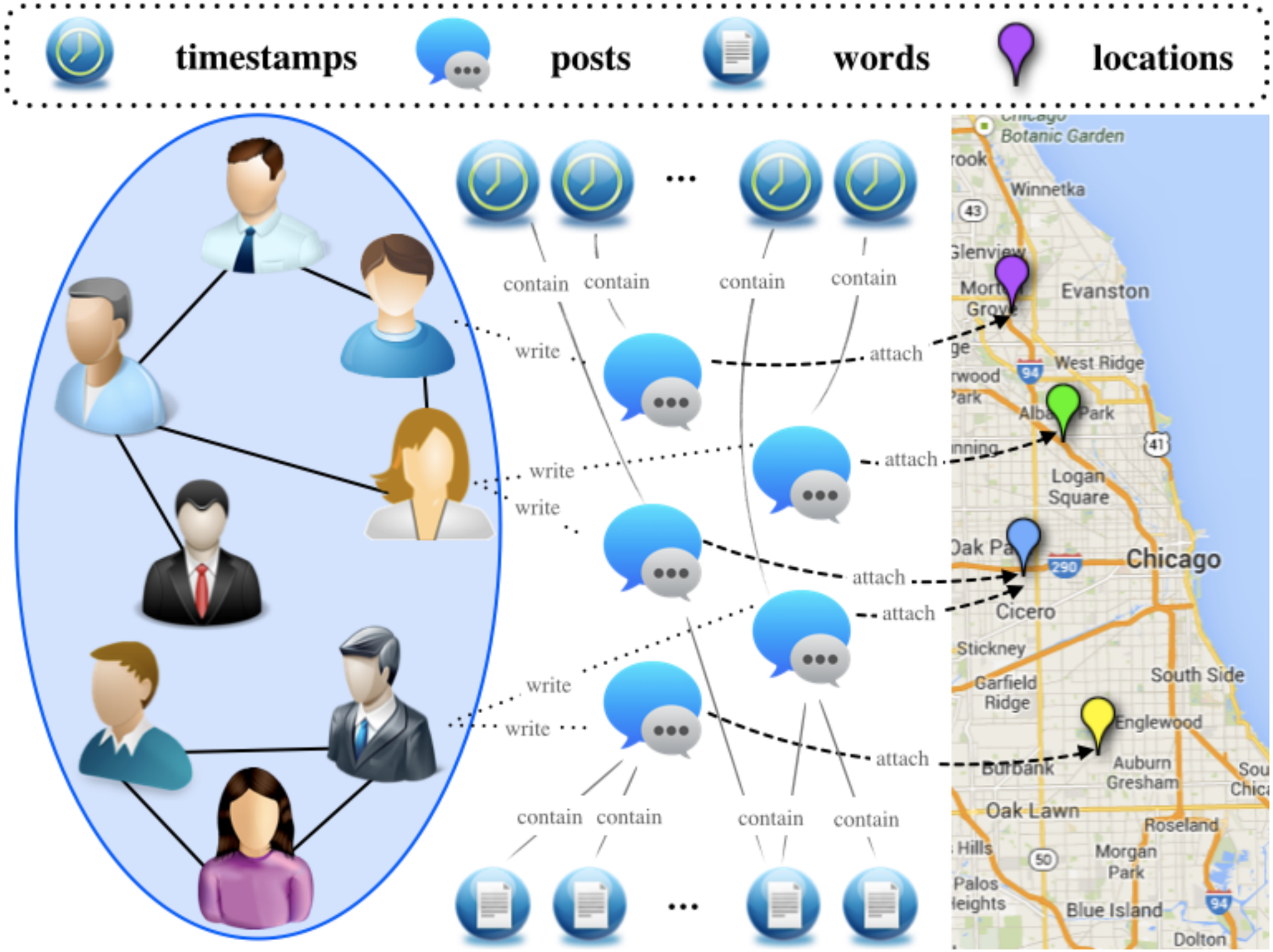
Challenges

- **Challenge 1:** Similarity measure among users with heterogeneous information in social networks

- **Challenge 2:** Community detection in each network

- **Challenge 3:** Mutual community detection

Challenge 1: Similarity Measure among Users with Heterogeneous Information



Solution: Social Meta Path based Similarity Measure

TABLE I
SUMMARY OF HNMPS.

ID	Notation	Heterogeneous Network Meta Path	Semantics
1	U $\rightarrow$ U	User $\xrightarrow{follow}$ User	Follow
2	U $\rightarrow$ U $\rightarrow$ U	User $\xrightarrow{follow}$ User $\xrightarrow{follow}$ User	Follower of Follower
3	U $\rightarrow$ U $\leftarrow$ U	User $\xrightarrow{follow}$ User $\xrightarrow{follow^{-1}}$ User	Common Out Neighbor
4	U $\leftarrow$ U $\rightarrow$ U	User $\xrightarrow{follow^{-1}}$ User $\xrightarrow{follow}$ User	Common In Neighbor
5	U $\rightarrow$ P $\rightarrow$ W $\leftarrow$ P $\leftarrow$ U	User $\xrightarrow{write}$ Post $\xrightarrow{contain}$ Word $\xrightarrow{contain^{-1}}$ Post $\xrightarrow{write^{-1}}$ User	Posts Containing Common Words
6	U $\rightarrow$ P $\rightarrow$ T $\leftarrow$ P $\leftarrow$ U	User $\xrightarrow{write}$ Post $\xrightarrow{contain}$ Time $\xrightarrow{contain^{-1}}$ Post $\xrightarrow{write^{-1}}$ User	Posts Containing Common Timestamps
7	U $\rightarrow$ P $\rightarrow$ L $\leftarrow$ P $\leftarrow$ U	User $\xrightarrow{write}$ Post $\xrightarrow{attach}$ Location $\xrightarrow{attach^{-1}}$ Post $\xrightarrow{write^{-1}}$ User	Posts Attaching Common Location Check-ins

Similarity score between users x and y based on meta paths 1-7

$$\text{Sim}(x, y) = \sum_i \omega_i \left(\frac{|\mathcal{P}_i(x \rightsquigarrow y)| + |\mathcal{P}_i(y \rightsquigarrow x)|}{|\mathcal{P}_i(x \rightsquigarrow \cdot)| + |\mathcal{P}_i(y \rightsquigarrow \cdot)|} \right)$$

$\mathcal{P}_i(x \rightsquigarrow y)$ the set of path instances of meta path i going from x to y

Challenge 2: Community detection in each heterogeneous network

- **Solution**: Partition users into different clusters based on certain cost function, e.g., normalized-cut

let $\mathcal{C} = \{U_1, U_2, \ldots, U_k\}$ be the community structures detected from G.

$$cut(\mathcal{C}) = \frac{1}{2}\sum_{i=1}^{k} S(U_i, \overline{U_i}) = \frac{1}{2}\sum_{i=1}^{k} \sum_{u \in U_i, v \in \overline{U_i}} S(u, v),$$

$$Ncut(\mathcal{C}) = \frac{1}{2}\sum_{i=1}^{k} \frac{S(U_i, \overline{U_i})}{S(U_i, \cdot)} = \sum_{i=1}^{k} \frac{cut(U_i, \overline{U}_i)}{S(U_i, \cdot)},$$

S(u, v) is the meta path based similarity between u and v

Challenge 3: Mutual community detection

- **Solution**: transfer information about the anchor users across networks based on **Clustering Discrepancy**

Let u_i and u_j be two anchor users in the network,

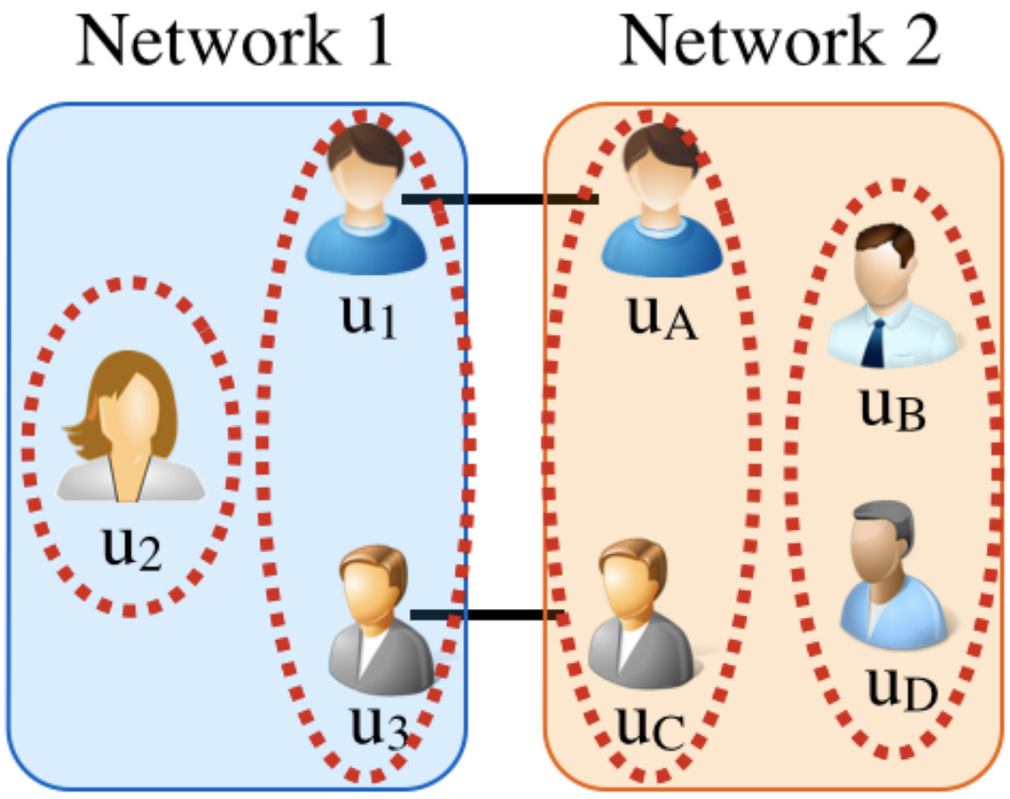
Definition 2 (Discrepancy): The discrepancy between the clustering results of u_i and u_j across aligned networks $G^{(1)}$ and $G^{(2)}$ is defined as the difference of confidence scores of u_i and u_j being partitioned in the same cluster across aligned networks.

$$d_{ij}(\mathcal{C}^{(1)}, \mathcal{C}^{(2)}) = \left(\mathbf{h}_i^{(1)}(\mathbf{h}_j^{(1)})^T - \mathbf{h}_i^{(2)}(\mathbf{h}_j^{(2)})^T\right)^2$$

$$d(\mathcal{C}^{(1)}, \mathcal{C}^{(2)}) = \sum_{i}^{n^{(1)}} \sum_{j}^{n^{(2)}} d_{ij}(\mathcal{C}^{(1)}, \mathcal{C}^{(2)}),$$

Normalized Discrepancy is used, Definition refer to the paper

Example: Discrepancy



Clustering Confidence Matrices

	cluster1	cluster2
u_1	1	0
u_2	0	1
u_3	1	0

$\mathbf{H}^{(1)}$

	cluster1	cluster2
u_A	1	0
u_B	0	1
u_C	1	0
u_D	0	1

$\mathbf{H}^{(2)}$

Anchor Transition Matrices

	u_A	u_B	u_C	u_D
u_1	1	0	0	0
u_2	0	0	0	0
u_3	0	0	1	0

$\mathbf{T}^{(1,2)}$

	u_1	u_2	u_3
u_A	1	0	0
u_B	0	0	0
u_C	0	0	1
u_D	0	0	0

$\mathbf{T}^{(2,1)}$

community detection discrepancy

$$\left\| \bar{\mathbf{H}}^{(1)} \left(\bar{\mathbf{H}}^{(1)} \right)^T - \bar{\mathbf{H}}^{(2)} \left(\bar{\mathbf{H}}^{(2)} \right)^T \right\|_F^2 = 0,$$

where

$$\bar{\mathbf{H}}^{(1)} = (\mathbf{T}^{(1,2)})^T \mathbf{H}^{(1)}$$

$$\bar{\mathbf{H}}^{(2)} = (\mathbf{T}^{(1,2)})^T (\mathbf{T}^{(2,1)})^T \mathbf{H}^{(2)}$$

Mutual Community Detection Objective Function

- The optimal Community Detection of networks $G^{(1)}$ and $G^{(2)}$ can be represented as

$$\arg \min_{\mathcal{C}^{(1)},\mathcal{C}^{(2)}} \alpha \cdot Ncut(\mathcal{C}^{(1)}) + \beta \cdot Ncut(\mathcal{C}^{(2)}) + \theta \cdot Nd(\mathcal{C}^{(1)},\mathcal{C}^{(2)})$$

where α, β and θ represents the weights of these terms and, for simplicity, α, β are both set as 1 in this paper.

sensitivity analysis of θ is available in the experiments

Mutual Community Detection Objective Function

$$\min_{\mathbf{H}^{(1)},\mathbf{H}^{(2)}} \alpha\cdot\mathrm{Tr}((\mathbf{H}^{(1)})^T\mathbf{L}^{(1)}\mathbf{H}^{(1)}) + \beta\cdot\mathrm{Tr}((\mathbf{H}^{(2)})^T\mathbf{L}^{(2)}\mathbf{H}^{(2)}) + \theta\cdot\frac{\left\|\bar{\mathbf{H}}^{(1)}\left(\bar{\mathbf{H}}^{(1)}\right)^T-\bar{\mathbf{H}}^{(2)}\left(\bar{\mathbf{H}}^{(2)}\right)^T\right\|_F^2}{\left\|\mathbf{T}^{(1,2)}\right\|_F^2\left(\left\|\mathbf{T}^{(1,2)}\right\|_F^2-1\right)},$$

$$s.t.\quad (\mathbf{H}^{(1)})^T\mathbf{D}^{(1)}\mathbf{H}^{(1)}=\mathbf{I},\ (\mathbf{H}^{(2)})^T\mathbf{D}^{(2)}\mathbf{H}^{(2)}=\mathbf{I},$$

The objective function i
with orthogonality constra
to solve because the con
but also numerically exper

Objective Equation of Normalized-Discrepancy between Networks G(1) and G(2)

solution: fix one variable, update the other variable until convergence

Convergence analysis is available in the experiments

Experiments

- **Dataset**: Foursquare &Twitter (Anchor Links: 3,388)

TABLE II
PROPERTIES OF THE HETEROGENEOUS SOCIAL NETWORKS

		network	
	property	**Twitter**	**Foursquare**
# node	user	5,223	5,392
	tweet/tip	9,490,707	48,756
	location	297,182	38,921
# link	friend/follow	164,920	76,972
	write	9,490,707	48,756
	locate	615,515	48,756

Experiments

- **Comparison Methods**
  - **MCD:** the method proposed in this paper
  - **SI-Clus**: Social Influence based multi-network clustering methods [34][37]
  - **NCUT**: partition users by minimizing the normalized-cut cost calculated with social links[23]
  - **KMeans**: extend traditional KMeans to partition users based on the social link information only[22]

[34] J. Zhang and P. Yu. Community detection for emerging networks. In *SDM*, 2015.

[37] Y. Zhou and L. Liu. Social influence based clustering of heterogeneous information networks. In *KDD*, 2013.

[22] G. Qi, C. Aggarwal, and T. Huang. Community detection with edge content in social media networks. In *ICDE*, 2012.

[23] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 2000.

Experiments

- **Evaluation Metrics**
  - **Clustering Quality Metrics**
    - (1) normalized-dbi, (2) entropy,
    - (3) density, (4) Silhouette index
- **Clustering Consensus Metrics**
  - (1) rand, (2) variation of information,
  - (3)mutual information, (4) normalized mutual information

Community Detection C

Partial Alignment of Foursquare and Twitter

network	measure	methods	remaining anchor link rates σ 0.1	0.2	0.3	0.4	0.5	0.6	0.7
Foursquare	ndbi	MCD	**0.927**	**0.924**	**0.95**	**0.969**	**0.966**	**0.961**	**0.958**
		SICLUS	0.891	0.889	0.88	0.877	0.894	0.883	0.89
		NCUT	0.863	0.863	0.863	0.863	0.863	0.863	0.863
		KMEANS	0.835	0.835	0.835	0.835	0.835	0.835	0.835
	entropy	MCD	**1.551**	**1.607**	**1.379**	**1.382**	**1.396**	**1.382**	**1.283**
		SICLUS	4.332	4.356	4.798	4.339	4.474	4.799	4.446
		NCUT	2.768	2.768	2.768	2.768	2.768	2.768	2.768
		KMEANS	2.369	2.369	2.369	2.369	2.369	2.369	2.369
	density	MCD	**0.216**	**0.205**	**0.196**	0.163	**0.239**	**0.192**	**0.303**
		SICLUS	0.116	0.121	0.13	0.095	0.143	0.11	0.13
		NCUT	0.154	0.154	0.154	0.154	0.154	0.154	0.154
		KMEANS	0.182	0.182	0.182	**0.182**	0.182	0.182	0.182
	silhouette	MCD	**-0.137**	**-0.114**	**-0.148**	**-0.156**	**-0.117**	**-0.11**	**-0.035**
		SICLUS	-0.168	-0.198	-0.173	-0.189	-0.178	-0.181	-0.21
		NCUT	-0.34	-0.34	-0.34	-0.34	-0.34	-0.34	-0.34
		KMEANS	-0.297	-0.297	-0.297	-0.297	-0.297	-0.297	-0.297

Community Detection Quality in Twitter

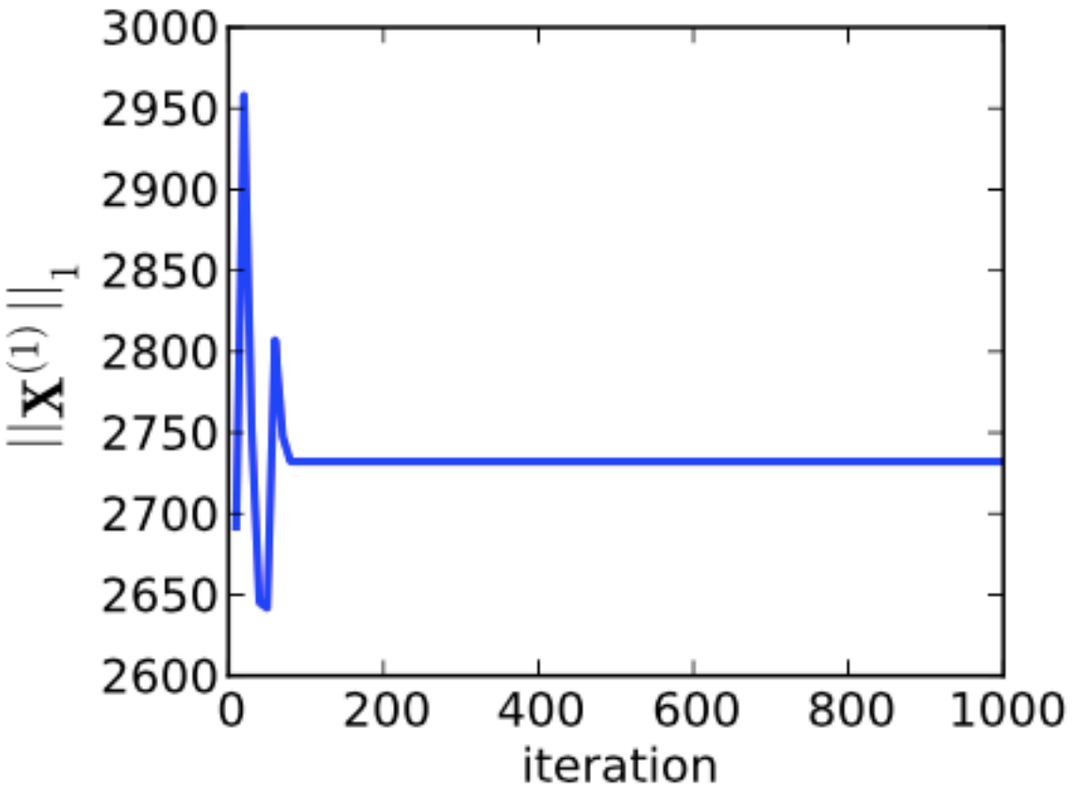
network	measure	methods	remaining anchor link rates σ 0.1	0.2	0.3	0.4	0.5	0.6	0.7
Twitter	ndbi	MCD	**0.962**	**0.969**	**0.955**	**0.969**	**0.97**	**0.958**	**0.952**
		SICLUS	0.815	0.843	0.807	0.83	0.826	0.832	0.835
		NCUT	0.759	0.759	0.759	0.759	0.759	0.759	0.759
		KMEANS	0.761	0.761	0.761	0.761	0.761	0.761	0.761
	entropy	MCD	**2.27**	**2.667**	**2.48**	**2.381**	**2.43**	**2.372**	**2.452**
		SICLUS	4.780	5.114	5.066	4.961	4.904	4.866	5.121
		NCUT	3.099	3.099	3.099	3.099	3.099	3.099	3.099
		KMEANS	3.245	3.245	3.245	3.245	3.245	3.245	3.245
	density	MCD	**0.14**	0.097	**0.142**	0.109	**0.15**	**0.158**	**0.126**
		SICLUS	0.055	0.017	0.044	0.026	0.04	0.062	0.016
		NCUT	0.107	0.107	0.107	0.107	0.107	0.107	0.107
		KMEANS	0.119	**0.119**	0.119	**0.119**	0.119	0.119	0.119
	silhouette	MCD	**-0.137**	**-0.179**	**-0.282**	**-0.175**	**-0.275**	**-0.273**	**-0.248**
		SICLUS	-0.356	-0.322	-0.311	-0.347	-0.346	-0.349	-0.323
		NCUT	-0.424	-0.424	-0.424	-0.424	-0.424	-0.424	-0.424
		KMEANS	-0.406	-0.406	-0.406	-0.406	-0.406	-0.406	-0.406

Consensus of the Community Detection Results between Foursquare and Twitter

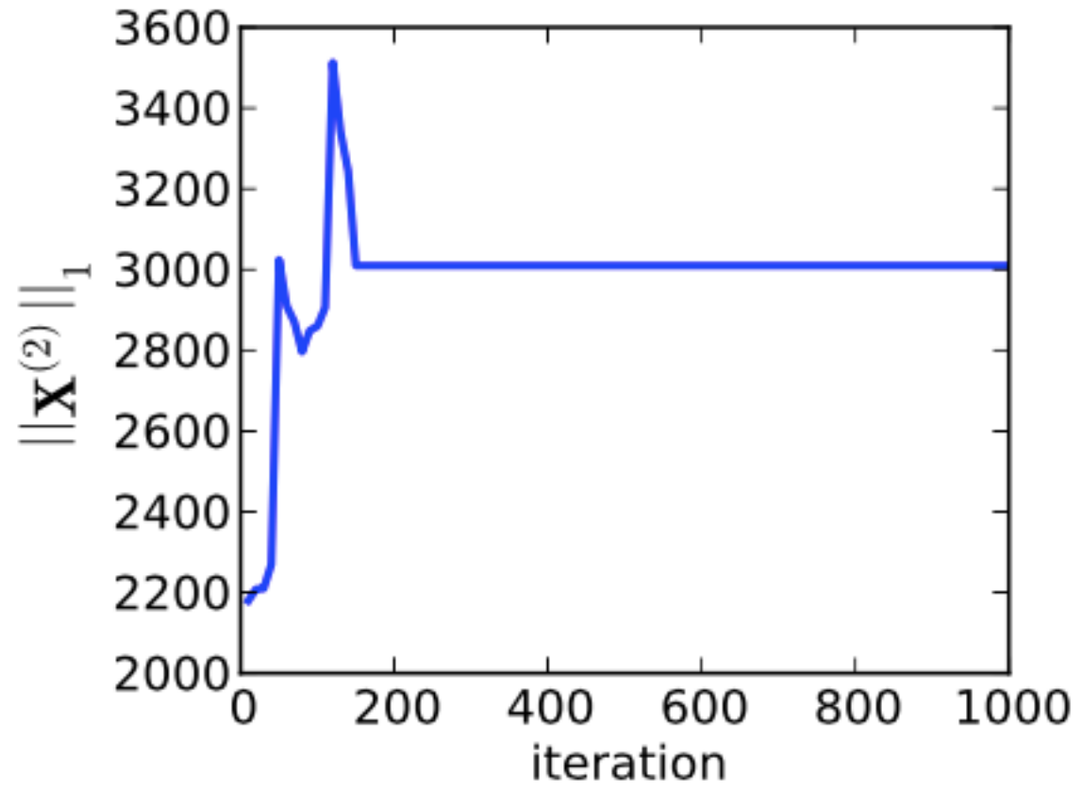
measure	methods	remaining anchor link rates σ						
		0.1	0.2	0.3	0.4	0.5	0.6	0.7
rand	MCD	**0.095**	**0.099**	**0.107**	**0.138**	**0.116**	**0.121**	**0.132**
	SICLUS	0.135	0.139	0.144	0.148	0.142	0.14	**0.132**
	NCUT	0.399	0.377	0.372	0.4	0.416	0.423	0.362
	KMEANS	0.436	0.387	0.4	0.358	0.403	0.363	0.408
vi	MCD	**3.309**	**4.052**	**4.058**	**3.902**	**4.038**	**4.348**	**3.973**
	SICLUS	7.56	8.324	8.414	8.713	8.756	8.836	8.832
	NCUT	5.384	5.268	5.221	4.855	5.145	5.541	5.909
	KMEANS	5.427	5.117	5.355	5.326	5.679	5.944	5.452
nmi	MCD	0.152	**0.152**	**0.149**	**0.141**	**0.149**	**0.156**	**0.142**
	SICLUS	**0.172**	0.097	0.081	0.06	0.056	0.069	0.078
	NCUT	0.075	0.074	0.111	0.108	0.109	0.099	0.05
	KMEANS	0.008	0.047	0.048	0.054	0.048	0.028	0.047
mi	MCD	0.756	**0.611**	**0.4**	0.258	**0.394**	**0.431**	**0.381**
	SICLUS	**0.780**	0.446	0.367	**0.277**	0.258	0.325	0.374
	NCUT	0.188	0.181	0.261	0.232	0.252	0.243	0.138
	KMEANS	0.02	0.112	0.119	0.135	0.127	0.078	0.119

Convergence Analysis

To address the objective function: iterative updating variables



(a) $\left\|\mathbf{X}^{(1)}\right\|_1$



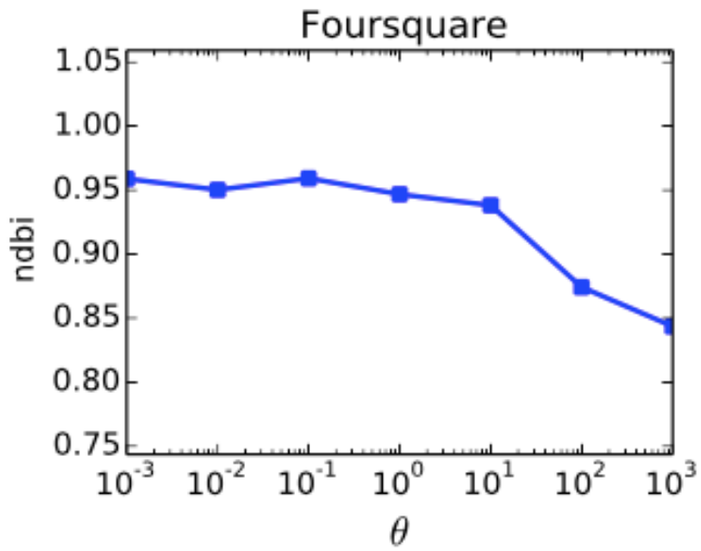
(b) $\left\|\mathbf{X}^{(2)}\right\|_1$

Fig. 3. $\|\mathbf{X}$

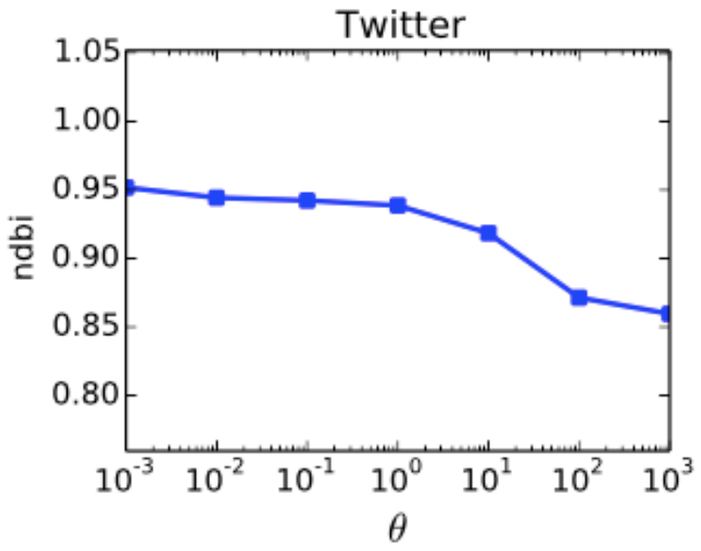
both X(1) and X(2) can converge within 200 iterations

Parameter Analysis

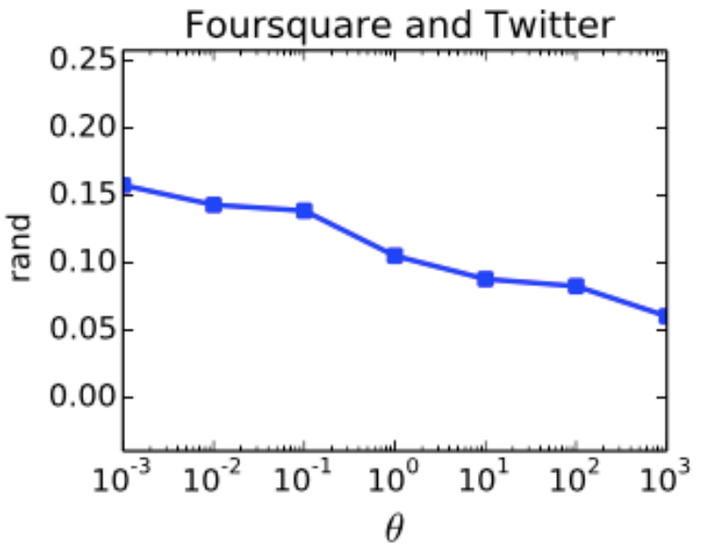
In the objective function: θ is the weight of the discrepancy term



(a) θ-ndbi (Foursquare)



(b) θ-ndbi (Twitter)



(c) θ-rand

higher ndbi: better quality; lower rand: more consensus

smaller θ: favors high quality results in each network

larger θ favors consensus results between networks

Summary

- Problem Studied: **Mutual Community Detection** across **Partially Aligned Social Networks**

- Proposed Method:
  - **Social Meta Path** based **Similarity Measure** among users
  - **Normalized-Cut** based **Isolated Community Detection**
  - **Normalized-Discrepancy** based **Mutual Community Detection**

Q&A