UIC

Community Detection for Emerging Networks

Jiawei Zhang[1], Philip S. Yu[1,2]

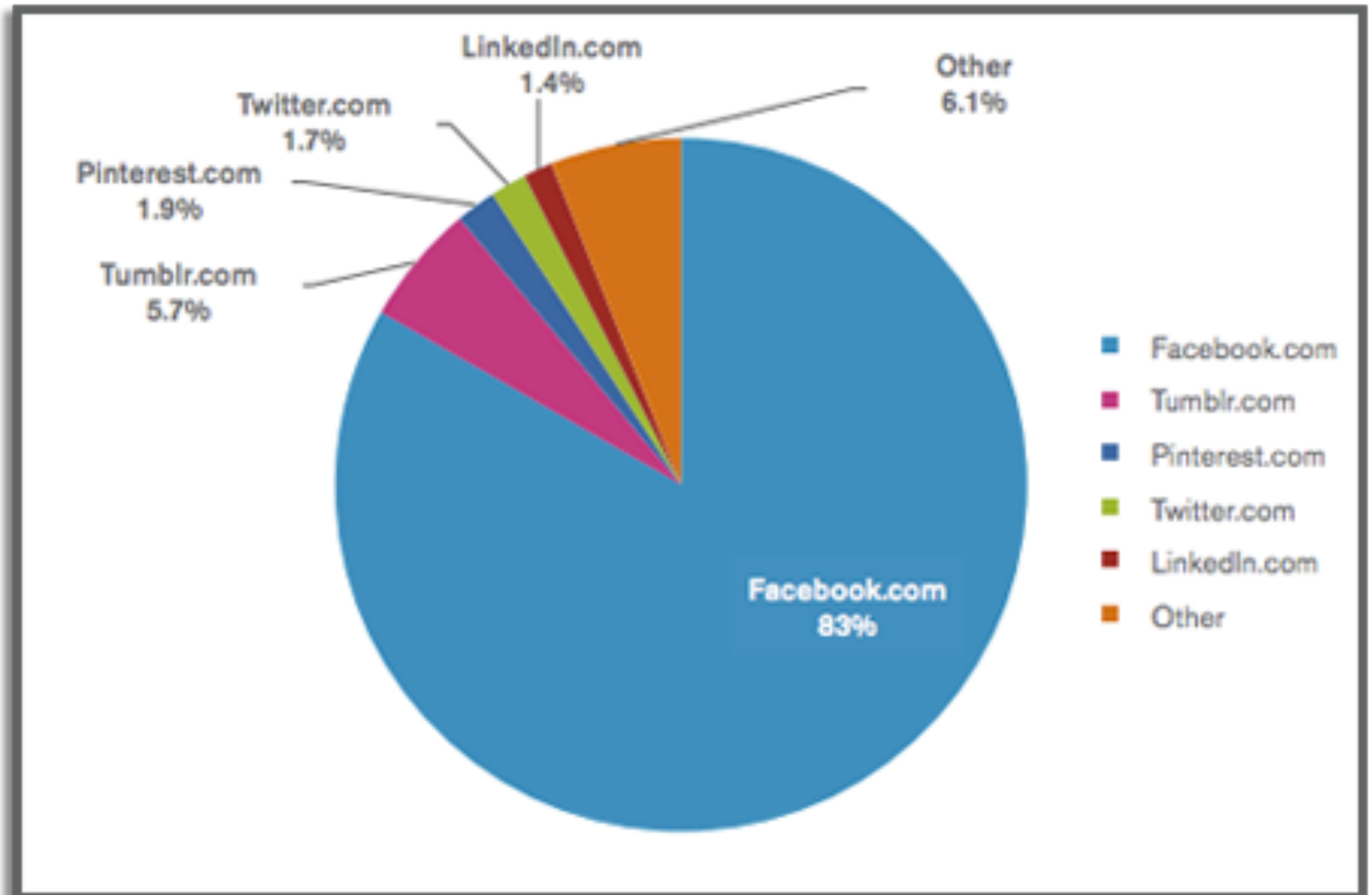
[1] University of Illinois at Chicago, USA
[2]Tsinghua University, China

New Social Networks Emerge Every Year

launch year

http://en.wikipedia.org/wiki/List_of_social_networking_websites

Emerging Networks Attract Limited Usages



Share of time spent

Emerging Networks Contains Sparse Information

Emerging Network Community Detection

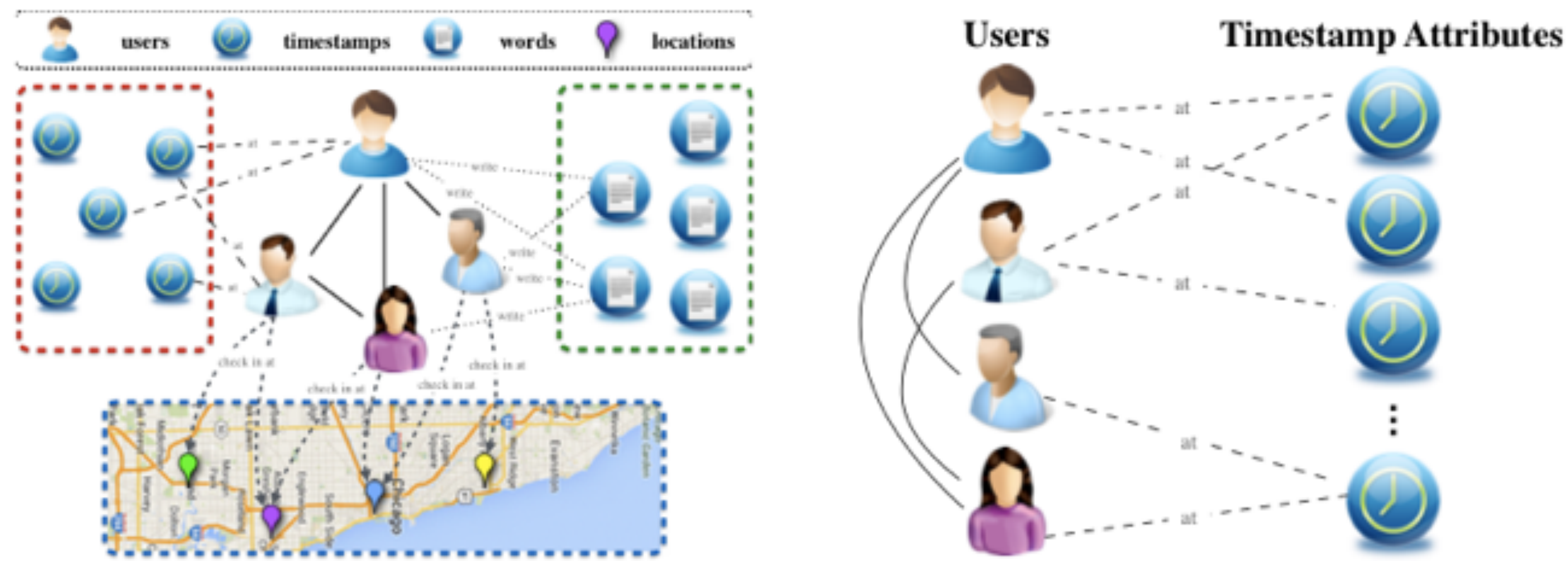
Hard to calculate effective closeness measures among users due to the sparse information

closeness measures among users:

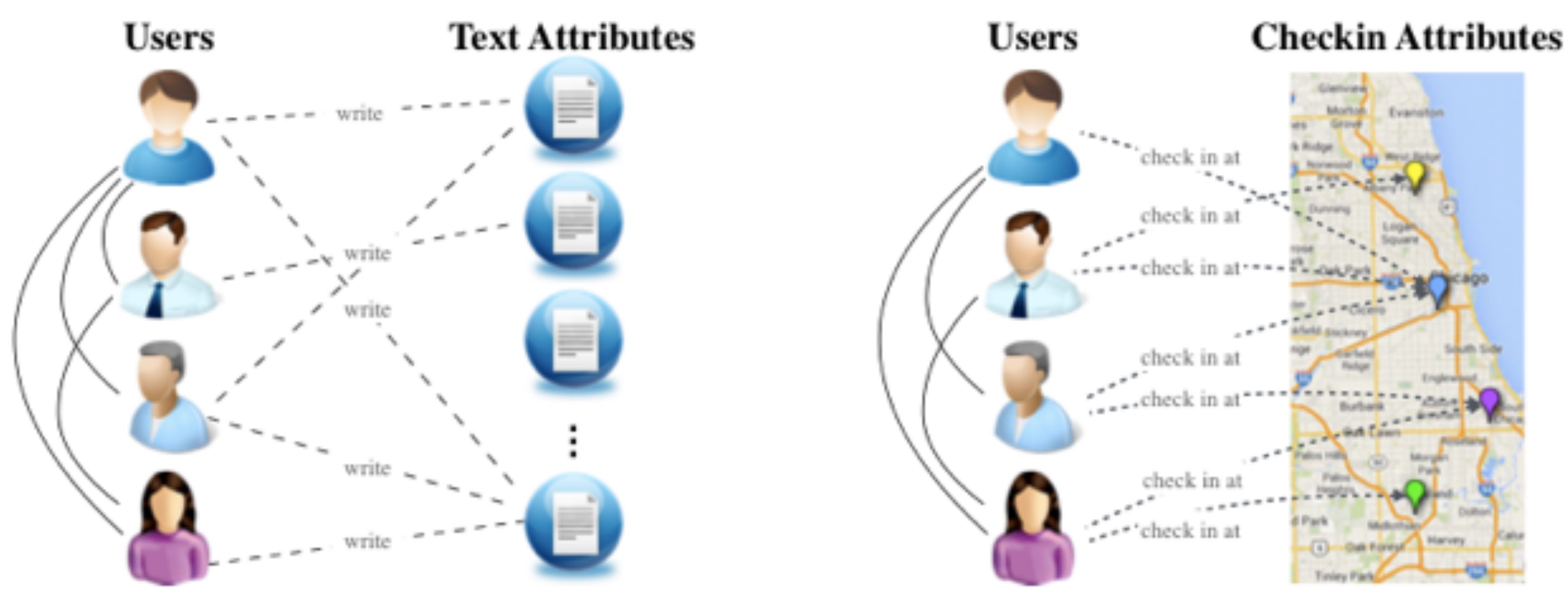
Intimacy

Challenge 1: Information Sparsity Problem

- Solution: use both Link and Attribute information

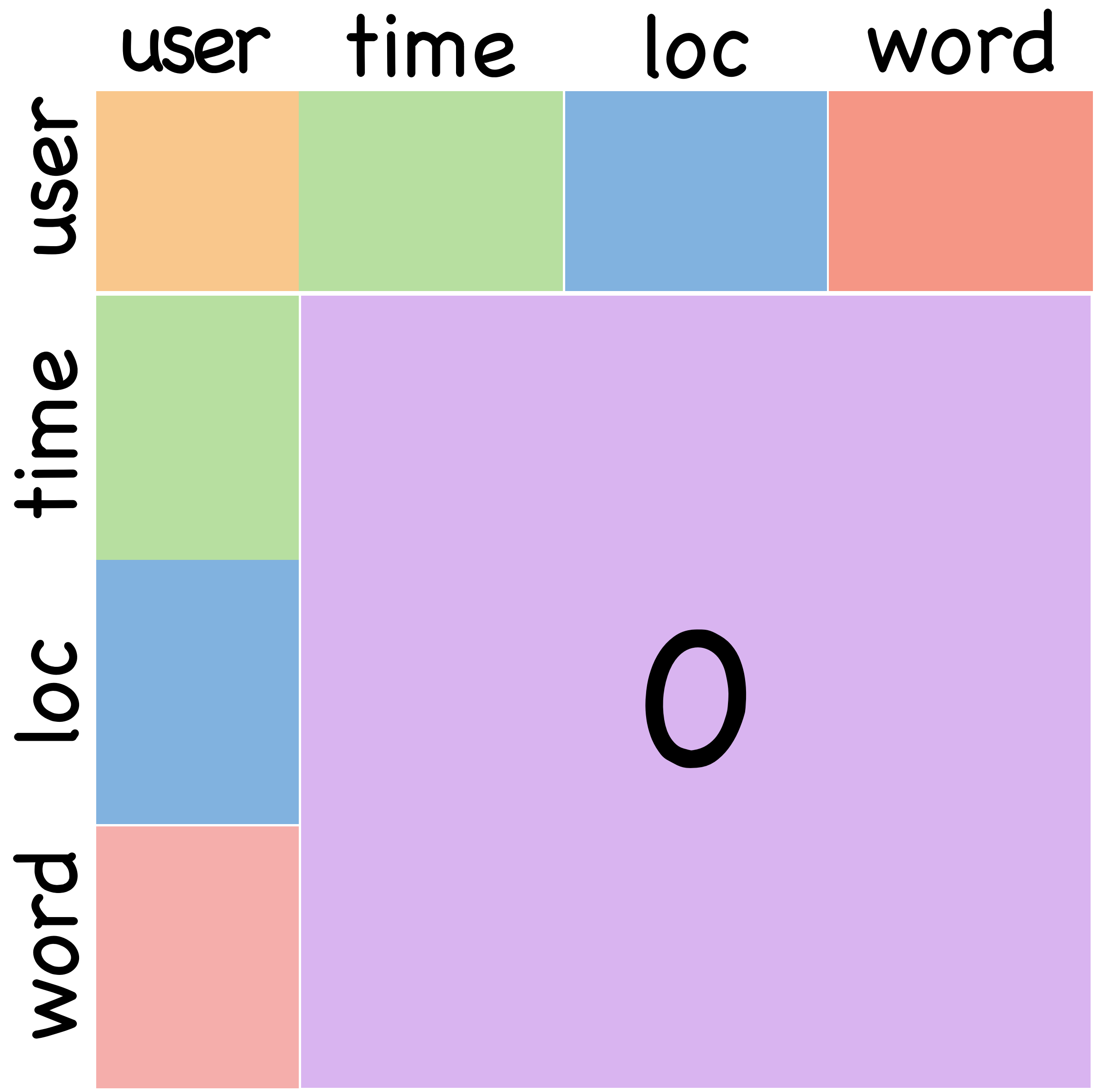


(a) augmented network (b) timestamp attribute



(c) text attribute (d) checkin attribute

Intimacy Calculation with both Connection and Attribute Information



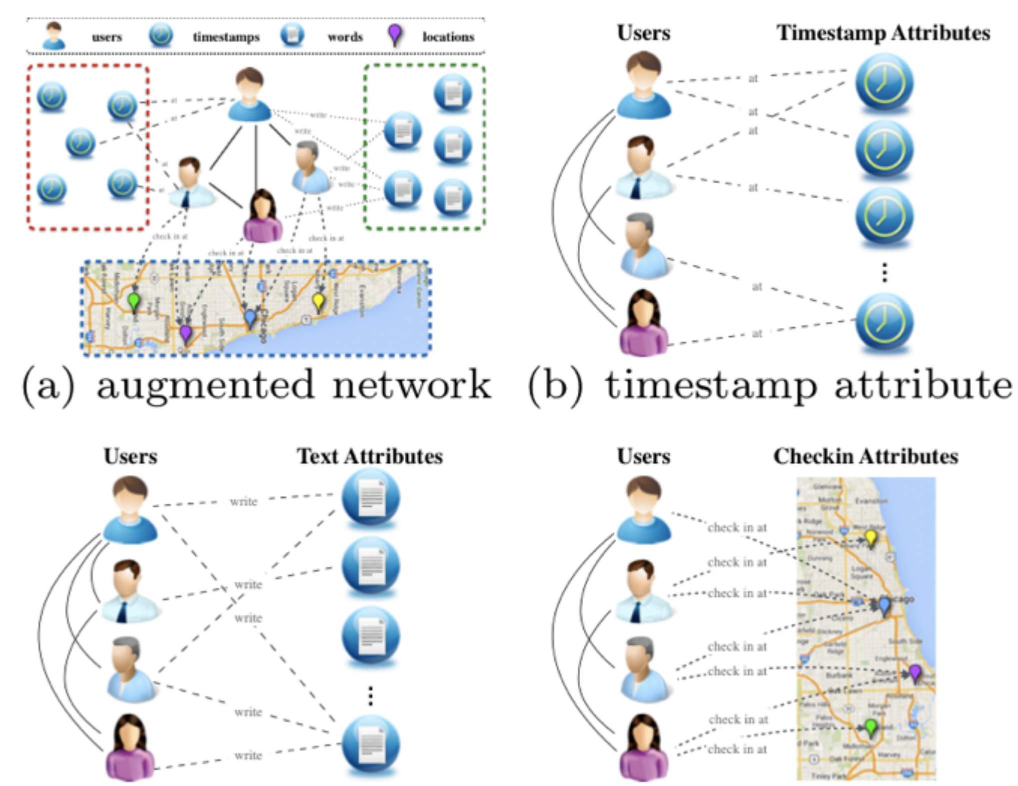
$$\tilde{\mathbf{Q}}_{aug} = \begin{bmatrix} \tilde{\mathbf{Q}} & \tilde{\mathbf{R}} \\ \tilde{\mathbf{S}} & \mathbf{0} \end{bmatrix}$$

network transitional matrix

weighted normalized adjacency matrices

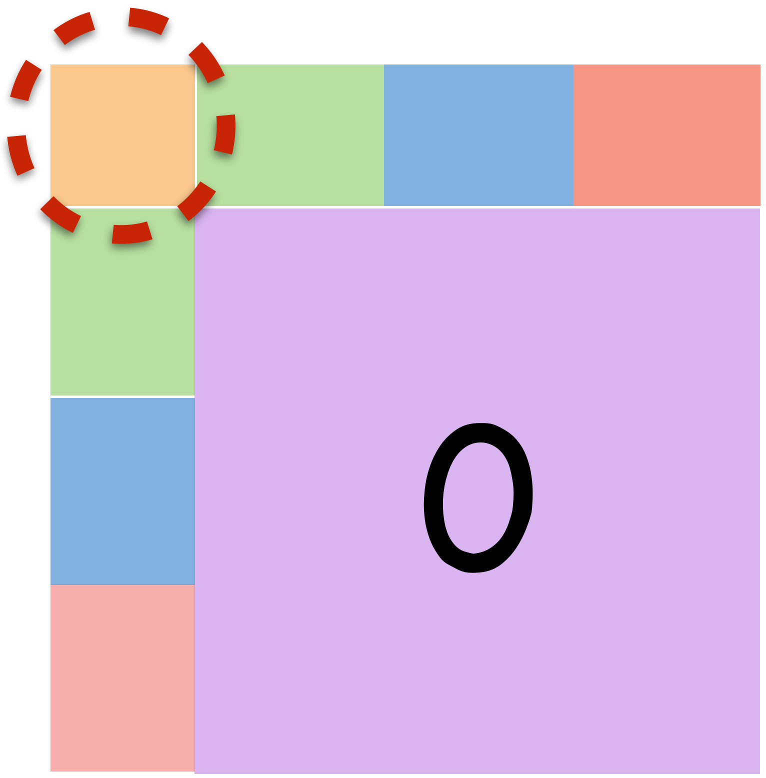
(1) among users

(2) between users and attributes



(a) augmented network (b) timestamp attribute (c) text attribute (d) checkin attribute

Intimacy Calculation with both Connection and Attribute Information



$$\left(\mathbf{I} + \alpha \tilde{\mathbf{Q}}_{aug}\right)^{\tau}$$

high-dimensional
stationary network transitional matrix

we only care about the intimacy matrix among users (lower dimension)

$$\tilde{\mathbf{H}}_{aug} = \left(\mathbf{I} + \alpha \tilde{\mathbf{Q}}_{aug}\right)^{\tau} (1 : |\mathcal{V}|, 1 : |\mathcal{V}|)$$

intimacy matrix
among users

sub-matrix
at the upper left corner

stationary network transitional matrix calculation

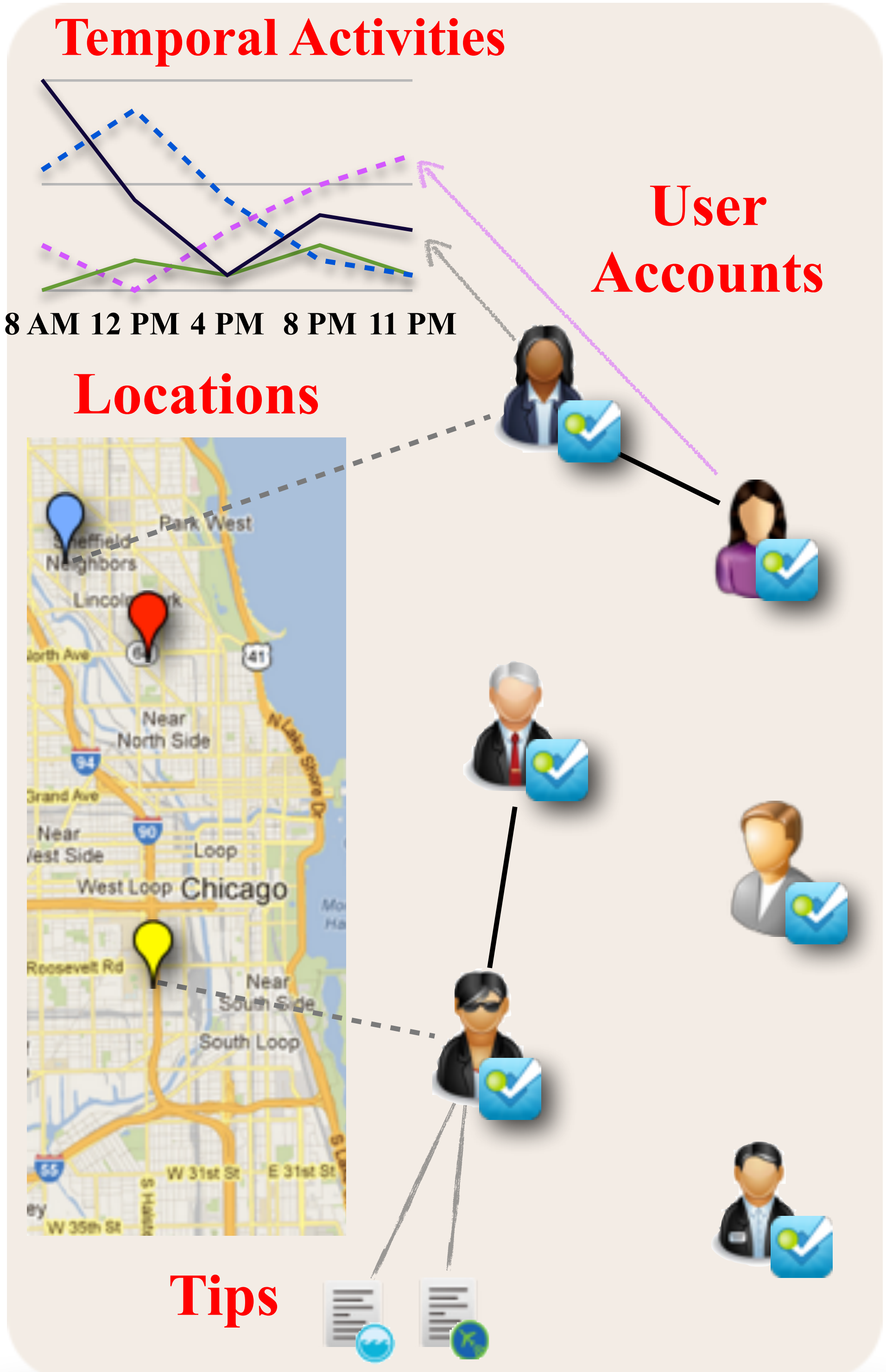
LEMMA 3.1. $(\tilde{\mathbf{Q}}_{aug})^k = \begin{bmatrix} \tilde{\mathbf{Q}}_k & \tilde{\mathbf{Q}}_{k-1}\tilde{\mathbf{R}} \\ \tilde{\mathbf{S}}\tilde{\mathbf{Q}}_{k-1} & \tilde{\mathbf{S}}\tilde{\mathbf{Q}}_{k-2}\tilde{\mathbf{R}} \end{bmatrix}$, $k \geq 2$, *where*

$$\tilde{\mathbf{Q}}_k = \begin{cases} \mathbf{I}, & \text{if } k = 0, \\ \tilde{\mathbf{Q}}, & \text{if } k = 1,, \\ \tilde{\mathbf{Q}}\tilde{\mathbf{Q}}_{k-1} + \tilde{\mathbf{R}}\tilde{\mathbf{S}}\tilde{\mathbf{Q}}_{k-2}, & \text{if } k \geq 2 \end{cases} \quad \tilde{\mathbf{Q}}_k \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|} \text{ and}$$

the heterogeneous network intimacy matrix is defined as

$$\begin{aligned} \tilde{\mathbf{H}}_{aug} &= \left(\mathbf{I} + \alpha\tilde{\mathbf{Q}}_{aug}\right)^\tau (1 : |\mathcal{V}|, 1 : |\mathcal{V}|) \\ &= \left(\sum_{t=0}^{\tau} \binom{\tau}{t} \alpha^t (\tilde{\mathbf{Q}}_{aug})^t\right) (1 : |\mathcal{V}|, 1 : |\mathcal{V}|) \\ &= \left(\sum_{t=0}^{\tau} \binom{\tau}{t} \alpha^t \left((\tilde{\mathbf{Q}}_{aug})^t (1 : |\mathcal{V}|, 1 : |\mathcal{V}|)\right)\right) \\ &= \left(\sum_{t=0}^{\tau} \binom{\tau}{t} \alpha^t \tilde{\mathbf{Q}}_t\right), \end{aligned}$$

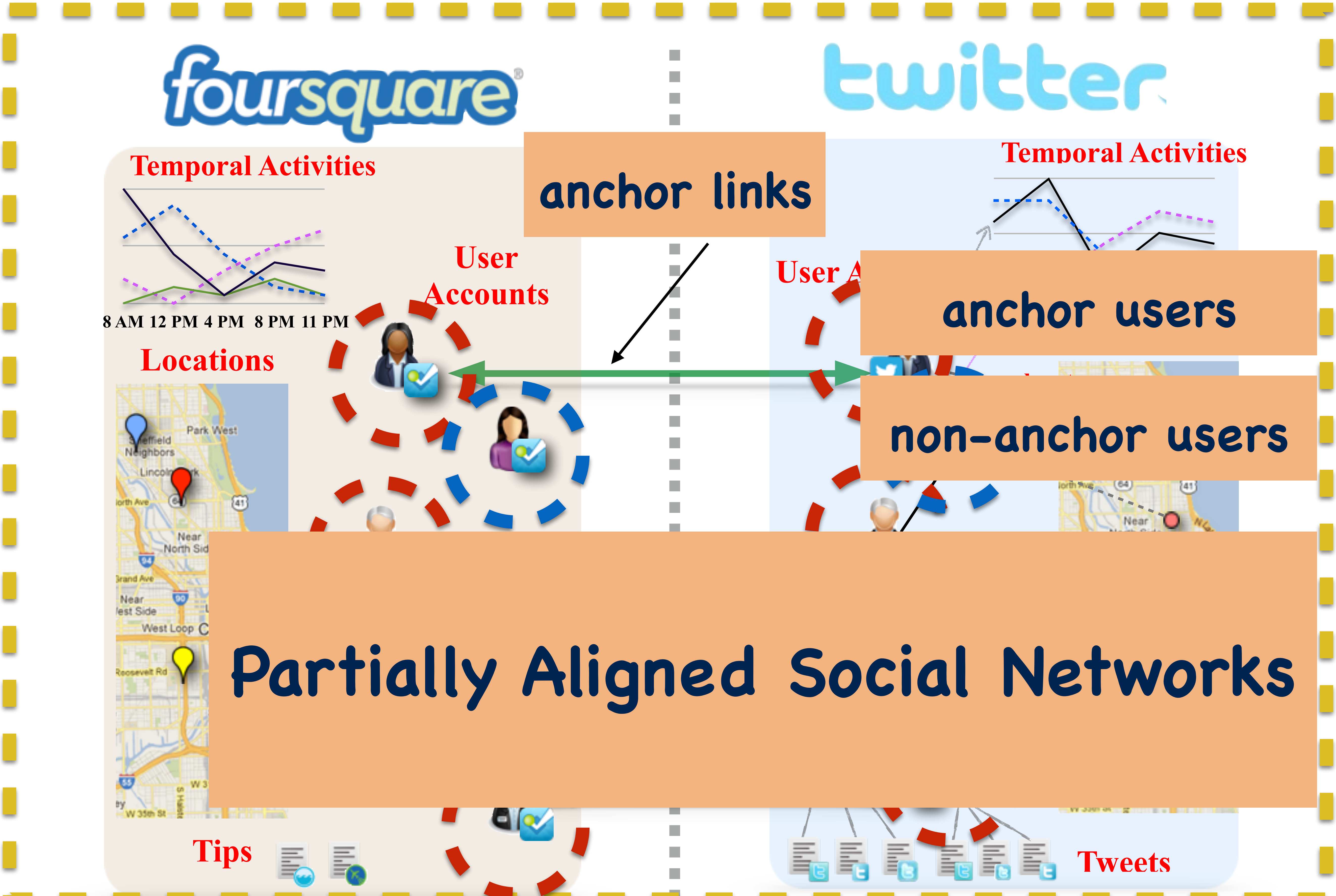
Challenge 2: Cold Start Community Detection



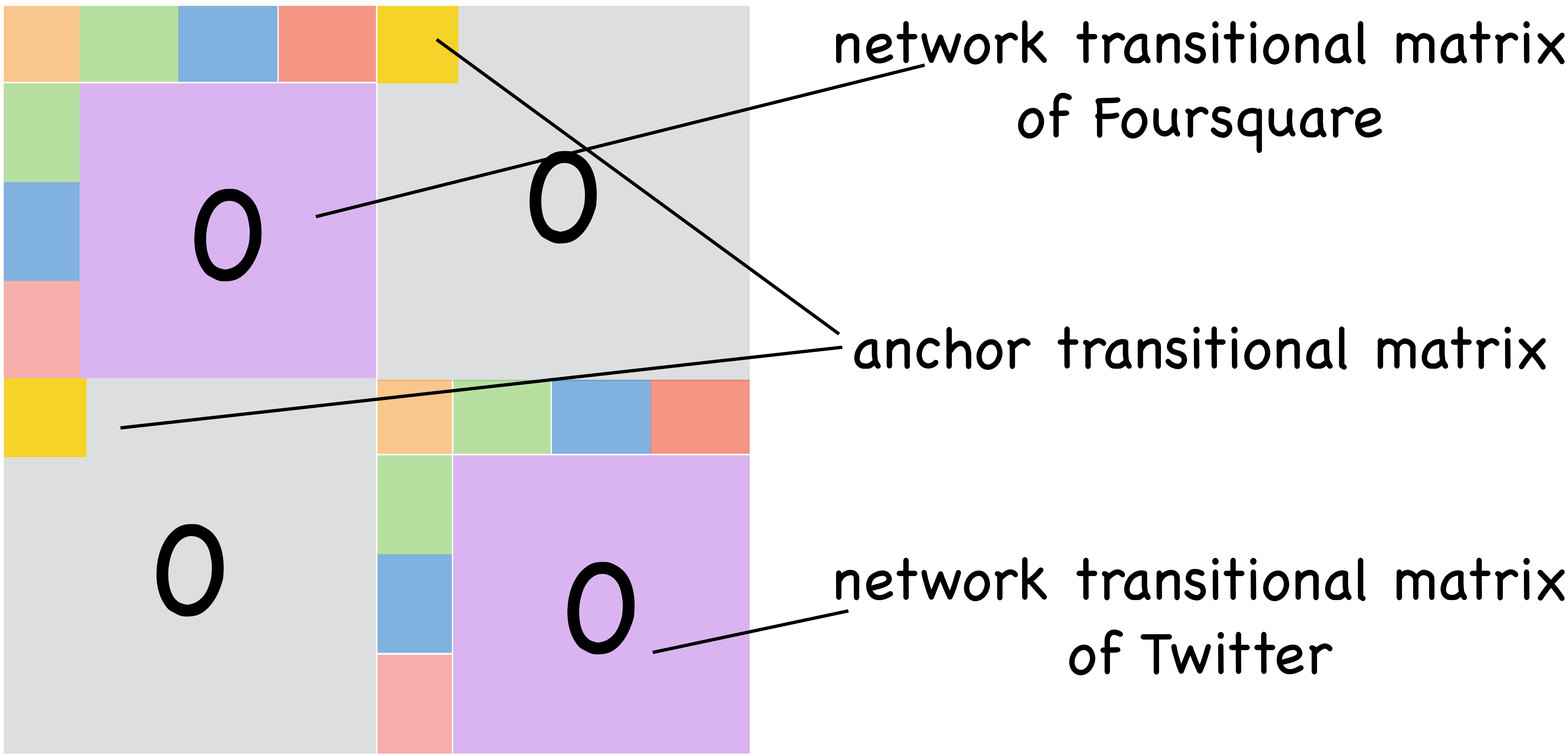
Emerging Network
Community Detection

A special case: Cold Start
Community Detection
(no social activities exist at all)

Users use multiple social networks simultaneously



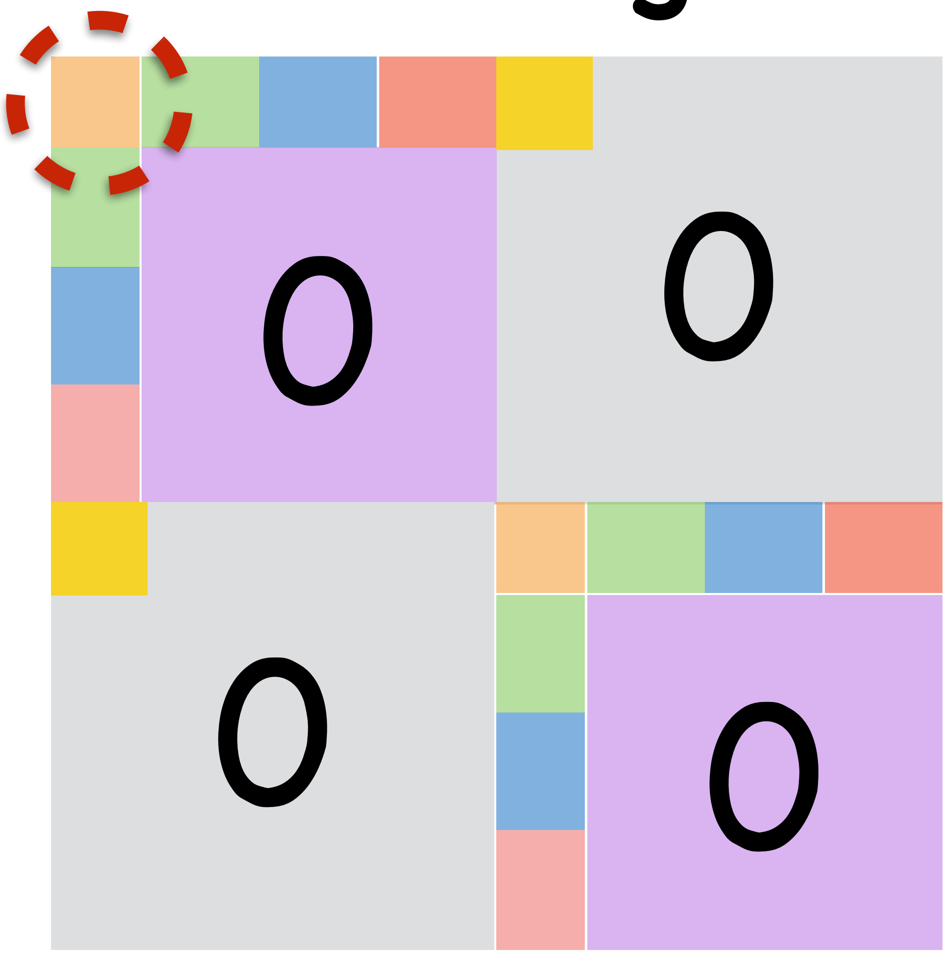
Intimacy Calculation with Information across Aligned Networks



$$\bar{\mathbf{Q}}_{align} = \begin{bmatrix} \bar{\mathbf{Q}}^{t}_{aug} & \bar{\mathbf{T}}^{t,s} \\ \bar{\mathbf{T}}^{s,t} & \bar{\mathbf{Q}}^{s}_{aug} \end{bmatrix}$$

weighted aligned network transitional matrix

Intimacy Calculation with Information across Aligned Networks



$$(\mathbf{I} + \alpha \bar{\mathbf{Q}}_{align})^{\tau}$$

high-dimensional
stationary aligned
network transitional matrix

we only care about the intimacy matrix among users (lower dimension)

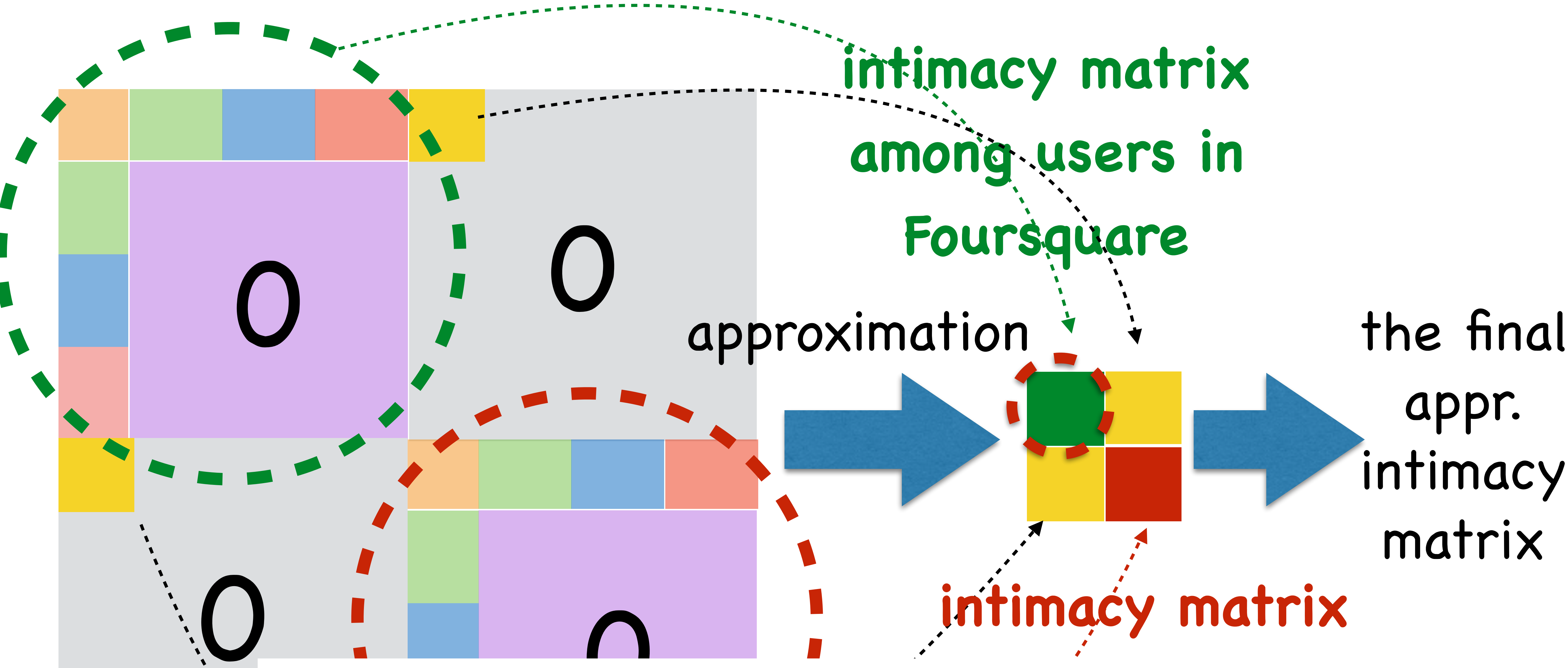
$$\bar{\mathbf{H}}_{align} = (\mathbf{I} + \alpha \bar{\mathbf{Q}}_{align})^{\tau}(1 : |\mathcal{V}^t|, 1 : |\mathcal{V}^t|)$$

intimacy matrix among
users in Foursquare

sub-matrix
at the upper left corner

Challenge 3: High Time and Space Costs

Solution: Approximated Intimacy Calculation



$$\bar{\mathbf{Q}}^{user}_{align} = \begin{bmatrix} (1-\rho^{t,s})\tilde{\mathbf{Q}}^{t}_{\mathcal{T}^t} & (\rho^{t,s})\mathbf{T}^{t,s} \\ (\rho^{s,t})\mathbf{T}^{s,t} & (1-\rho^{s,t})\tilde{\mathbf{Q}}^{s}_{\mathcal{T}^s} \end{bmatrix}$$

Approximated Intimacy Calculation

LEMMA 3.2. *For the given matrix* $(\mathbf{I} + \alpha\bar{\mathbf{Q}}_{align})$, *its* k_{th} *power meets* $(\mathbf{I} + \alpha\bar{\mathbf{Q}}_{align})^k\mathbf{P} = \mathbf{P}\mathbf{\Lambda}^k, k \geq 1$, *matrices* $\mathbf{P}$ *and* $\mathbf{\Lambda}$ *contain the eigenvector and eigenvalues of* $(\mathbf{I} + \alpha\bar{\mathbf{Q}}_{align})$. *The* i_{th} *column of matrix* $\mathbf{P}$ *is the eigenvector of* $(\mathbf{I}+\alpha\bar{\mathbf{Q}}_{align})$ *corresponding to its* i_{th} *eigenvalue* λ_i *and diagonal matrix* $\mathbf{\Lambda}$ *has value* $\Lambda(i,i) = \lambda_i$ *on its diagonal.*

$$\bar{\mathbf{H}}_{align}^{approx} = \left(\mathbf{P}^*(\mathbf{\Lambda}^*)^\tau(\mathbf{P}^*)^{-1}\right)(1:|\mathcal{V}^t|, 1:|\mathcal{V}^t|),$$

where $(\mathbf{I} + \alpha\bar{\mathbf{Q}}_{align}^{user}) = \mathbf{P}^*\mathbf{\Lambda}^*(\mathbf{P}^*)^{-1}$, τ *is the stop step.*

Clustering based on Intimacy Matrix

$$\min_{\mathbf{U},\mathbf{V}} \left\| \bar{\mathbf{H}}_{align} - \mathbf{U}\mathbf{V}\mathbf{U}^T \right\|_F^2 + \theta \left\| \mathbf{U} \right\|_F^2 + \beta \left\| \mathbf{V} \right\|_F^2 ,$$

$$s.t., \mathbf{U} \geq \mathbf{0}, \mathbf{V} \geq \mathbf{0},$$

where $\mathbf{U}$ is the latent feature vectors, $\mathbf{V}$ stores the correlation among rows of $\mathbf{U}$, θ and β are the weights of $\|\mathbf{U}\|_F^2$, $\|\mathbf{V}\|_F^2$ respectively.

The latent feature vectors in $\mathbf{U}$ can be used to detect communities in some traditional clustering methods, e.g., Kmeans [3].

Parameter Adjustment: weights of different information types and sources

Experiments

- Dataset

Table 1: Properties of the Heterogeneous 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

anchor links: 3,388

Experiments

- Comparison Methods

  - CADE-A (Exact intimacy matrix based CAD with parameter Adjustment)
  - CADA-A (Approximated intimacy matrix based CAD with parameter Adjustment)
  - CADE (Exact intimacy matrix based CAD)
  - CADA (Approximated intimacy matrix based CAD)
  - SINFL (Social Influence-based clustering)
  - NCUT (Normalized Cut)
  - KMEANS

Experiments

- Evaluation Metrics

  - *normalized Davies-Bouldin index*: $ndbi(\mathcal{C}) = \frac{1}{K}\sum_{i=1}^{K}\min_{j\neq i}\frac{d(c_i,c_j)+d(c_j,c_i)}{\sigma_i+\sigma_j+d(c_i,c_j)+d(c_j,c_i)}$, where c_i is the centroid of $U_i \in \mathcal{C}$, $d(c_i, c_j)$ is the distance between c_i and c_j, σ_i denotes the average distance between items in U_i and centroid c_i [23].
  - *Silhouette*: Let $a(u) = \frac{1}{|U_i|-1}\sum_{v\in U_i, v\neq u} d(u,v)$ and $b(u) = \min_{j,j\neq i}\left(\frac{1}{|U_j|}\sum_{v\in U_j} d(u,v)\right)$, the *Silhouette index* is defined to be $silhouette(\mathcal{C}) = \frac{1}{K}\sum_{i=1}^{K}(\frac{1}{|U_i|}\sum_{u\in U_i}\frac{b(u)-a(u)}{\max\{a(u),b(u)\}})$ [9].
  - *Entropy*: $E(\mathcal{C}) = -\sum_{i=1}^{K} P(i)\log P(i)$, where $P(i) = \frac{|U_i|}{|\mathcal{V}|}$ [23].

methods with approximated intimacy matrix can save lots of space and time

measure	methods	Information Sampling Rate 0.0	0.1	0.2	0.3	0.4	0.5
ndbi	CADE-A	**0.954**	**0.959**	**0.966**	**0.969**	**0.968**	**0.972**
	CADA-A	0.917	0.922	0.923	0.925	0.938	0.946
	CADE	0.938	0.944	0.949	0.949	0.954	0.957
	CADA	0.914	0.914	0.918	0.923	0.932	0.936
	SINFL	-	0.881	0.889	0.901	0.907	0.913
	NCUT	-	0.864	0.870	0.889	0.889	0.893
	KMEANS	-	0.842	0.859	0.881	0.886	0.887

Table 3: Space and time costs in calculating $\bar{\mathbf{H}}_{align}$.

emerging network	cost	method: **exact**	**approx.**
Foursquare	space cost(MB)	19526	1627
	time cost(s)	65996.17	6499.97

Summary

- Problem Studied: **Emerging Network Community Detection** & **Cold Start Community Detection**
- Calculate the **Intimacy** scores among users in the emerging network with both **Connection** and **Attribute** information across **Partially Aligned Networks**.
- To lower the time and space cost: **Approximated Intimacy Calculation**

Q & A

Anchor Links across Networks