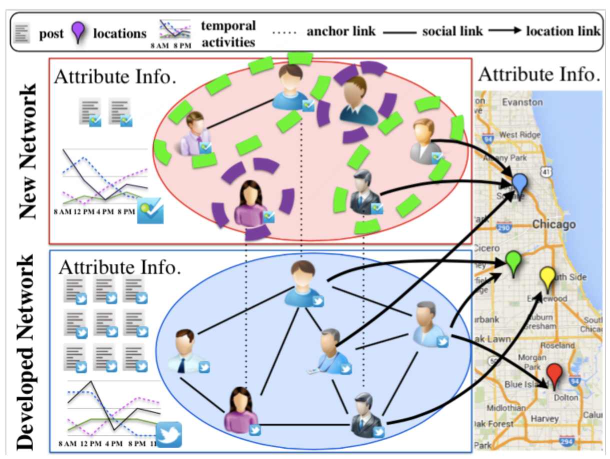
Community Detection for Emerging Networks

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

[1] University of Illinois at Chicago [2] Tsinghua University

2015 SIAM International Conference on DATA MINING

1. Emerging Network Community Detection Problem



1. Problem Studied: Community Detection for Emerging Networks.

2. Emerging Networks: Emerging Networks refers to Networks with Very Sparse Information (Connections and Attribute Information).

3. Aligned Social Networks: networks sharing common anchor users are called aligned social networks.

4. Cold Start Community Detection: A special case of "Emerging Network Community Detection" is "Cold Start Community Detection", where the target network is brand new containing no users' social activities at all.

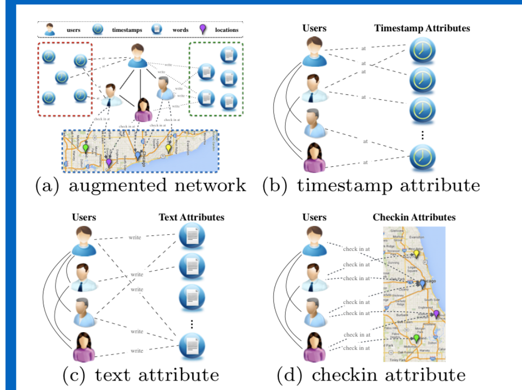
Challenge 0: How to measure the closeness (**Intimacy**) among users in the emerging network?

Challenge 1: What kind of information can we use in the network?

Challenge 2: How to transfer information across aligned networks?

Challenge 3: How to address the high space and time costs?

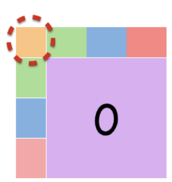
2. Intimacy Calculation with Heterogeneous Information



(a) augmented network (b) timestamp attribute

(c) text attribute (d) checkin attribute

Intimacy Calculation with both Connection and Attribute Information

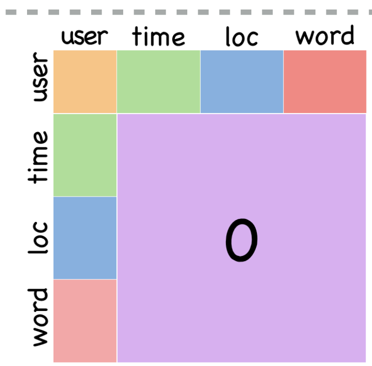


$$\left(\mathbf{I} + \alpha \tilde{\mathbf{Q}}_{aug}\right)^{\tau}$$

high-dimensional stationary network transitional matrix

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

$$\underbrace{\tilde{\mathbf{H}}_{aug}}_{\text{intimacy matrix among users}} = \left(\mathbf{I} + \alpha \tilde{\mathbf{Q}}_{aug}\right)^{\tau} \underbrace{(1:|\mathcal{V}|, 1:|\mathcal{V}|)}_{\text{sub-matrix at the upper left corner}}$$



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

network transitional matrix

normalized adjacency matrices
(1) among users
(2) between users and attributes

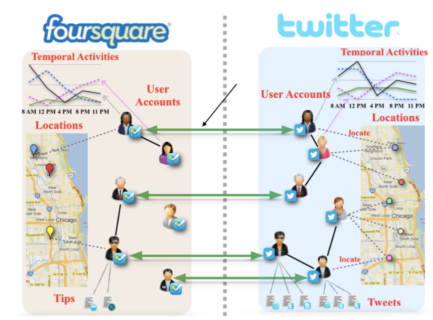
Lemma 3.1. $(\tilde{\mathbf{Q}}_{aug})^k = \begin{bmatrix} \tilde{\mathbf{Q}}_k & \tilde{\mathbf{Q}}_{k-1}\tilde{\mathbf{R}} \\ \tilde{\mathbf{S}}\tilde{\mathbf{Q}}_{k-1} & \tilde{\mathbf{S}}\tilde{\mathbf{Q}}_{k-2}\tilde{\mathbf{R}} \end{bmatrix}$, $k \geq 2$, *where*

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

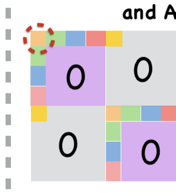
the heterogeneous network intimacy matrix is defined as

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

3. Intimacy Calculation across Aligned Networks



Intimacy Calculation with both Connection and Attribute Information



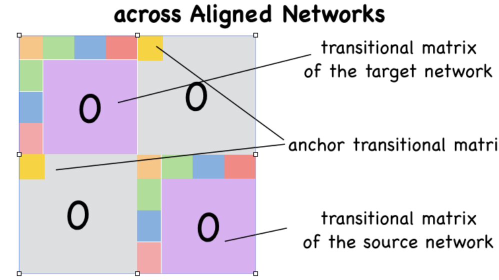
$$(\mathbf{I} + \alpha \bar{\mathbf{Q}}_{align})^{\tau}$$

high-dimensional stationary aligned network transitional matrix

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

$$\underbrace{\bar{\mathbf{H}}_{align}}_{\text{intimacy matrix among users in target network}} = (\mathbf{I} + \alpha \bar{\mathbf{Q}}_{align})^{\tau} \underbrace{(1:|\mathcal{V}^t|, 1:|\mathcal{V}^t|)}_{\text{sub-matrix at the upper left corner}}$$

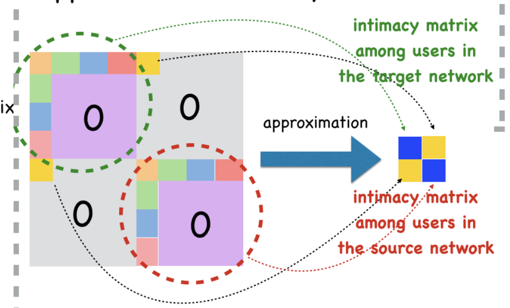
Intimacy Calculation with Information across Aligned Networks



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

aligned network transitional matrix

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.*

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

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

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

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

where $\mathbf{U}$ is the latent feature vectors, $\mathbf{V}$ stores the correlation among rows of $\mathbf{V}$, 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].

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

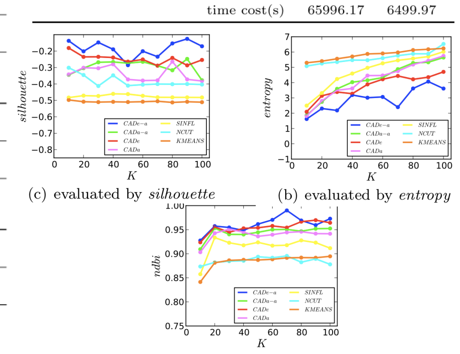
4. Experiment Results & Parameter Analysis & Time/Space Cost

Table 2: Community Detection Result of Foursquare.

		Information Sampling Rate σ_F										
measure	methods	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
ndbi	CADE-A	**0.954**	**0.959**	**0.966**	**0.969**	**0.968**	**0.972**	**0.974**	**0.979**	**0.984**	**0.989**	**0.991**
	CADA-A	0.917	0.922	0.923	0.925	0.938	0.946	0.946	0.946	0.947	0.949	0.950
	CADE	0.938	0.944	0.949	0.949	0.954	0.957	0.959	0.966	0.966	0.969	0.969
	CADA	0.914	0.914	0.918	0.923	0.932	0.936	0.939	0.940	0.942	0.942	0.946
	SINFL	-	0.881	0.889	0.901	0.907	0.913	0.913	0.916	0.916	0.917	0.917
	NCUT	-	0.864	0.870	0.889	0.889	0.893	0.894	0.894	0.894	0.897	0.897
	KMEANS	-	0.842	0.859	0.881	0.886	0.887	0.889	0.890	0.892	0.893	0.894
entropy	CADE-A	**3.001**	**2.859**	**2.753**	**2.482**	**2.361**	**2.342**	**2.167**	**2.25**	**2.140**	**1.994**	**1.932**
	CADA-A	4.150	4.137	4.133	4.108	4.084	4.025	4.013	3.856	3.506	3.70	3.68
	CADE	3.751	3.751	3.726	3.718	3.621	3.585	3.38	3.233	3.173	3.005	2.998
	CADA	4.360	4.237	4.213	4.211	4.102	4.061	4.021	4.015	3.97	3.851	3.823
	SINFL	-	5.147	5.105	5.063	4.981	4.968	4.934	4.892	4.856	4.768	4.668
	NCUT	-	5.823	5.691	5.618	5.517	5.512	5.494	5.485	5.473	5.467	5.459
	KMEANS	-	6.182	5.993	5.909	5.888	5.878	5.829	5.812	5.762	5.730	5.699
silhouette	CADE-A	**-0.396**	**-0.272**	**-0.28**	**-0.257**	**-0.251**	**-0.244**	**-0.224**	**-0.216**	**-0.150**	**-0.147**	**-0.132**
	CADA-A	-0.401	-0.384	-0.380	-0.377	-0.287	-0.279	-0.271	-0.270	-0.260	-0.237	-0.238
	CADE	-0.401	-0.302	**-0.275**	-0.270	-0.264	-0.262	-0.242	**-0.222**	**-0.196**	-0.186	**-0.129**
	CADA	-0.401	-0.381	-0.380	-0.372	-0.272	-0.260	-0.259	-0.251	-0.247	-0.246	-0.204
	SINFL	-	-0.482	-0.472	-0.469	-0.463	-0.462	-0.461	-0.459	-0.457	-0.428	-0.408
	NCUT	-	-0.415	-0.413	-0.413	-0.412	-0.410	-0.410	-0.408	-0.408	-0.345	-0.336
	KMEANS	-	-0.515	-0.515	-0.510	-0.508	-0.504	-0.498	-0.467	-0.464	-0.452	-0.434

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

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



(c) evaluated by *silhouette* (b) evaluated by *entropy* (a) evaluated by *ndbi*

5. Acknowledgement

This work is supported in part by NSF through grants CNS-1115234, and OISE-1129076, and the Pinnacle Lab at Singapore Management University.