Transferring Heterogeneous Links across Location-Based Social Networks

WSDM 2014 New York City February 24-28

Jiawei Zhang Xiangnan Kong Philip S. Yu

UIC Department of Computer Science, University of Illinois at Chicago, College of Engineering

1. Multiple Aligned LBSNs & Link Prediction Problem

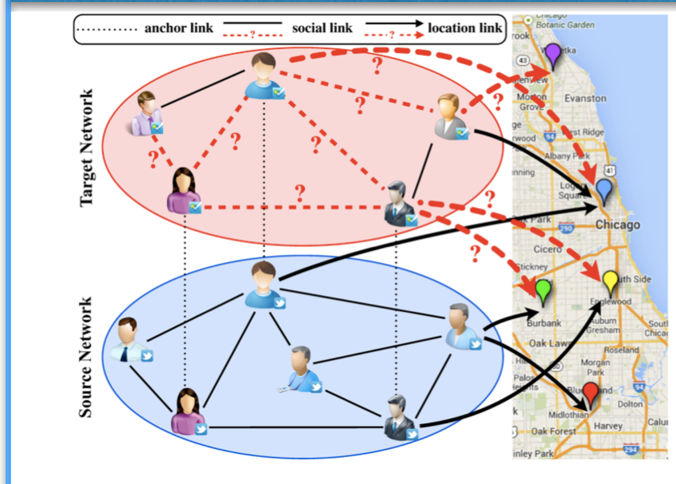


Figure 1: Example of collective link transferring across two aligned location-based social networks.

1. Location-based social networks (LBSNs) are one kind of online social networks offering geographic services.

2. Multiple Aligned Networks: users are usually involved in multiple networks simultaneously.

3. Link Prediction: predict links to be formed in the future based on a snapshot of online social networks.

4. Social Links & Location Links

Problem Studied:

Predict Social Links and Location Links for a New Location-based social network.

Challenges:

1. Collective Link Prediction: Multiple link prediction tasks in one network can be correlated.

2. New Network Problem: Information in new networks is very sparse.

2. Solve Challenge 1: Collective Link Prediction

(b) social link prediction

$$\hat{\mathcal{Y}}_s^t = arg \max_{\mathcal{Y}_s^t} P(\mathbf{y}(\boldsymbol{L}_s^t) = \mathcal{Y}_s^t | \mathbf{x}(\boldsymbol{L}_s^t))$$

(c) location link prediction

$$\hat{\mathcal{Y}}_l^t = arg \max_{\mathcal{Y}_l^t} P(\mathbf{y}(\boldsymbol{L}_l^t) = \mathcal{Y}_l^t | \mathbf{x}(\boldsymbol{L}_l^t))$$

(d) collective link prediction

$$\hat{\mathcal{Y}}_s^t, \hat{\mathcal{Y}}_l^t = arg \max_{\mathcal{Y}_s^t, \mathcal{Y}_l^t} P(\mathbf{y}(\boldsymbol{L}_s^t) = \mathcal{Y}_s^t | \mathbf{y}(\boldsymbol{L}_l^t) = \mathcal{Y}_l^t, \mathbf{x}(\boldsymbol{L}_s^t)) \times P(\mathbf{y}(\boldsymbol{L}_l^t) = \mathcal{Y}_l^t | \mathbf{y}(\boldsymbol{L}_s^t) = \mathcal{Y}_s^t, \mathbf{x}(\boldsymbol{L}_l^t))$$

x(L): feature vectors of links in L; y(L): labels of links in L, Ls: social links, Ll: location links.

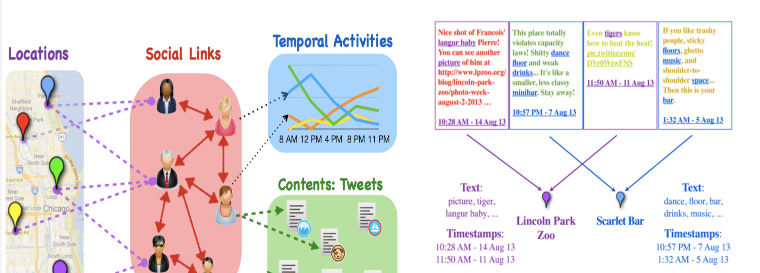


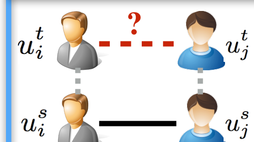
Figure 2: Example of information accumulation for locations from online posts.

Extracted Features

Table 2: Features extracted from vector x and y

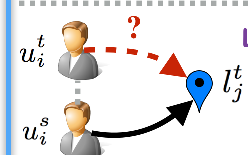
Features	Descriptions
Extended Degree Count (EDC)	$\|\|\boldsymbol{x}\|\|_1, \|\|\boldsymbol{y}\|\|_1$
Extended Degree Ratio (EDR)	$\|\|\boldsymbol{x}\|\|_1 / \|\|\boldsymbol{y}\|\|_1$
Extended Common Neighbour (ECN)	$\boldsymbol{x} \cdot \boldsymbol{y}$
Extended Jaccard's Coefficient (EJC)	$\frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\|\boldsymbol{x}\|\|_1 \cdot \|\|\boldsymbol{y}\|\|_1}$
Extended Preferential Attachment (EPA)	$\|\|\boldsymbol{x}\|\|_1 \cdot \|\|\boldsymbol{y}\|\|_1$
Euclidean Distance (ED)	$(\sum_k (x_k - y_k)^2)^{1/2}$
Cosine Similarity (CS)	$\frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\|\boldsymbol{x}\|\|_2 + \|\|\boldsymbol{y}\|\|_2}$

3. Solve Challenge 2: New Network Problem



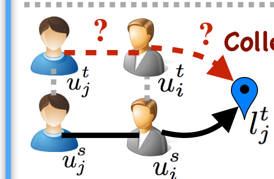
Social Link Prediction across Aligned LBSN

$$P(y(u_i^t, u_j^t) = 1 | G^t, G^s) = P\left(y(u_i^t, u_j^t) = 1 | \left[\boldsymbol{x}(u_i^t, u_j^t)^T, \boldsymbol{x}(u_i^s, u_j^s)^T, y(u_i^s, u_j^s)\right]^T\right)$$



Location Link Prediction across Aligned LBSN

$$P(y(u_i^t, l_j^t) = 1 | G^t, G^s) = P\left(y(u_i^t, l_j^t) = 1 | \left[\boldsymbol{x}(u_i^t, l_j^t)^T, \boldsymbol{x}(u_i^s, l_j^s)^T, y(u_i^s, l_j^s)\right]^T\right)$$



Collective Link Prediction across Aligned LBSN

$$\hat{\mathcal{Y}}_s^t, \hat{\mathcal{Y}}_l^t = arg \max_{\mathcal{Y}_s^t, \mathcal{Y}_l^t} P(\mathbf{y}(\boldsymbol{L}_s^t) = \mathcal{Y}_s^t | G^t, G^s, \mathbf{y}(\boldsymbol{L}_l^t) = \mathcal{Y}_l^t) \times P(\mathbf{y}(\boldsymbol{L}_l^t) = \mathcal{Y}_l^t | G^t, G^s, \mathbf{y}(\boldsymbol{L}_s^t) = \mathcal{Y}_s^t)$$

Alternatively Update to get the optimal labels

$$\hat{\mathcal{Y}}_s^{t(\tau)} = arg \max_{\mathcal{Y}_s^t} P(\mathbf{y}(\boldsymbol{L}_s^t) = \mathcal{Y}_s^t | G^t, G^s, \mathbf{y}(\boldsymbol{L}_s^t) = \hat{\mathcal{Y}}_s^{t(\tau-1)}, \mathbf{y}(\boldsymbol{L}_l^t) = \hat{\mathcal{Y}}_l^{t(\tau-1)})$$

$$\hat{\mathcal{Y}}_l^{t(\tau)} = arg \max_{\mathcal{Y}_l^t} P(\mathbf{y}(\boldsymbol{L}_l^t) = \mathcal{Y}_s^t | G^t, G^s, \mathbf{y}(\boldsymbol{L}_s^t) = \hat{\mathcal{Y}}_s^{t(\tau)}, \mathbf{y}(\boldsymbol{L}_l^t) = \hat{\mathcal{Y}}_l^{t(\tau-1)})$$

Algorithm 1 TRAIL

Input: two aligned heterogeneous LBSNs, G^s, G^t.
existing social links and location links: E_s^t, E_l^t
anchor links between G^t and G^s: $A^{t,s}$, $A^{s,t}$
potential social links and location links: L_s^t, L_l^t
Output: the inferred labels and existence probabilities of links in L_s^t and L_l^t: $\hat{\mathcal{Y}}_s^t, \hat{\mathcal{P}}_s^t, \hat{\mathcal{Y}}_l^t, \hat{\mathcal{P}}_l^t$

```
1: construct training sets, test sets with E_s^t, E_l^t, L_s^t and L_l^t.
2: converge ← False
3: while converge is False do
4:    extract features x^t(E_s^t) and x^t(L_s^t) for social links in E_s^t and L_s^t from G^t.
5:    extract features x^s(E_s^t) and x^s(L_s^t) for social links in E_s^t and L_s^t from G^s by utilizing anchor links in A^{t,s}.
6:    C_s ← train([x^t(E_s^t)^T, x^s(E_s^t)^T, y^s(E_s^t)]^T, y^t(E_s^t))
7:    Ŷ_s^t, P̂_s^t ← C_s.classify([x^t(L_s^t)^T, x^s(L_s^t)^T, y^s(L_s^t)]^T)
8:    update G^t with Ŷ_s^t, P̂_s^t
9:    Accumulate information for locations
10:   extract features x^t(E_l^t) and x^t(L_l^t) for location links in E_l^t and L_l^t from G^t.
11:   extract features x^s(E_l^t) and x^s(L_l^t) for location links in E_l^t and L_l^t from G^s by utilizing anchor links in A^{t,s}.
12:   C_l ← train([x^t(E_l^t)^T, x^s(E_l^t)^T, y^s(E_l^t)]^T, y^t(E_l^t))
13:   Ŷ_l^t, P̂_l^t ← C_l.classify([x^t(L_l^t)^T, x^s(L_l^t)^T, y^s(L_l^t)]^T)
14:   update G^t with Ŷ_l^t, P̂_l^t
15:   if Ŷ_s^t, P̂_s^t, Ŷ_l^t, P̂_l^t all converge then
16:      converge ← True
17:   end if
18: end while
19: Return Ŷ_s^t, P̂_s^t, Ŷ_l^t, P̂_l^t
```

4. Experiments

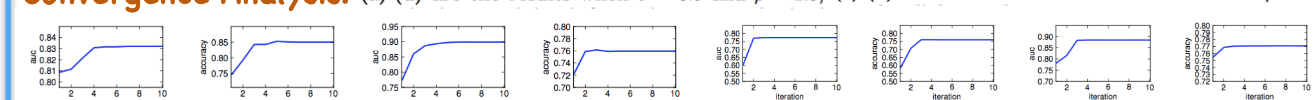
Table 4: Performance comparison of different methods for inferring social and location links for Foursquare of different remaining information rates. The anchor link sample rate ρ is set as 1.0.

link	measure	methods	remaining information rates σ 0.1	0.2	0.3	0.4	0.5	0.6	0.7
social	AUC	TRAIL	**0.810±0.012**	**0.824±0.009**	**0.837±0.008**	**0.844±0.009**	**0.832±0.003**	**0.852±0.009**	**0.847±0.009**
		TRAILT	0.691±0.040	0.684±0.039	0.704±0.033	0.729±0.006	0.718±0.020	0.732±0.005	0.730±0.008
		TRAILS	0.572±0.007	0.578±0.007	0.580±0.004	0.575±0.012	0.580±0.011	0.583±0.009	0.578±0.009
		SCAN	0.772±0.050	0.788±0.004	0.811±0.009	0.830±0.005	0.809±0.004	0.825±0.008	0.824±0.012
		SCANT	0.524±0.023	0.559±0.008	0.559±0.017	0.554±0.044	0.630±0.008	0.599±0.007	0.627±0.004
		SCANS	0.583±0.005	0.579±0.003	0.583±0.010	0.562±0.005	0.579±0.004	0.585±0.003	0.584±0.003
		CN	0.494±0.002	0.500±0.015	0.504±0.006	0.496±0.012	0.495±0.018	0.491±0.015	0.489±0.018
		JC	0.497±0.003	0.503±0.004	0.501±0.002	0.502±0.010	0.496±0.008	0.496±0.019	0.492±0.008
		AA	0.494±0.002	0.499±0.014	0.501±0.006	0.494±0.012	0.492±0.018	0.489±0.015	0.493±0.022
	Accuracy	TRAIL	**0.855±0.002**	**0.849±0.004**	**0.850±0.008**	**0.854±0.005**	**0.850±0.003**	**0.851±0.001**	**0.852±0.004**
		TRAILT	0.622±0.046	0.627±0.036	0.655±0.022	0.676±0.009	0.674±0.019	0.677±0.004	0.679±0.008
		TRAILS	0.548±0.004	0.551±0.006	0.552±0.004	0.549±0.000	0.551±0.002	0.553±0.003	0.544±0.001
		SCAN	0.747±0.003	0.752±0.007	0.748±0.000	0.754±0.008	0.746±0.005	0.745±0.007	0.747±0.003
		SCANT	0.512±0.009	0.522±0.002	0.520±0.001	0.537±0.006	0.554±0.008	0.542±0.003	0.567±0.007
		SCANS	0.557±0.002	0.547±0.006	0.553±0.002	0.545±0.006	0.552±0.007	0.551±0.002	0.551±0.004
		NAIVE	0.525±0.014	0.526±0.006	0.525±0.008	0.526±0.007	0.525±0.013	0.525±0.009	0.525±0.013
location	AUC	TRAIL	**0.848±0.005**	**0.856±0.010**	**0.870±0.010**	**0.878±0.007**	**0.899±0.007**	**0.886±0.022**	**0.887±0.009**
		TRAILT	0.839±0.006	0.850±0.003	0.857±0.009	0.866±0.008	0.862±0.005	0.871±0.005	0.869±0.003
		TRAILS	0.631±0.003	0.632±0.002	0.631±0.001	0.634±0.001	0.634±0.002	0.634±0.002	0.635±0.001
		SCAN	0.712±0.010	0.757±0.002	0.758±0.009	0.770±0.005	0.775±0.005	0.784±0.004	0.792±0.003
		SCANT	0.676±0.009	0.711±0.005	0.730±0.005	0.749±0.003	0.756±0.001	0.763±0.005	0.769±0.003
		SCANS	0.633±0.003	0.633±0.003	0.633±0.001	0.636±0.001	0.637±0.000	0.633±0.001	0.634±0.001
		FCF	0.598±0.008	0.638±0.015	0.638±0.005	0.654±0.012	0.664±0.007	0.661±0.007	0.664±0.010
	Accuracy	TRAIL	**0.719±0.004**	**0.736±0.001**	**0.749±0.006**	**0.754±0.003**	**0.753±0.002**	**0.760±0.002**	**0.761±0.002**
		TRAILT	0.674±0.009	0.697±0.004	0.706±0.005	0.709±0.001	0.717±0.006	0.716±0.007	0.717±0.002
		TRAILS	0.536±0.003	0.527±0.001	0.537±0.005	0.553±0.003	0.560±0.002	0.565±0.000	0.566±0.001
		SCAN	0.658±0.000	0.670±0.002	0.682±0.001	0.697±0.003	0.699±0.003	0.723±0.003	0.723±0.007
		SCANT	0.610±0.001	0.623±0.001	0.631±0.001	0.647±0.001	0.653±0.002	0.671±0.003	0.676±0.002
		SCANS	0.536±0.025	0.531±0.008	0.535±0.002	0.547±0.004	0.557±0.004	0.565±0.001	0.566±0.001
		NAIVE	0.536±0.014	0.536±0.002	0.536±0.001	0.537±0.008	0.536±0.012	0.536±0.009	0.537±0.019

Table 5: Performance comparison of different methods for inferring social and location links for Foursquare of different anchor link sample rates. The remaining informaiton rate σ is set as 1.0.

link	measure	methods	anchor link sample rates ρ 0.0	0.2	0.4	0.6	0.8	1.0
social	AUC	TRAIL	**0.712±0.004**	**0.733±0.019**	**0.761±0.017**	**0.782±0.007**	**0.821±0.012**	**0.855±0.008**
		TRAILT	0.712±0.012	0.711±0.007	0.711±0.012	0.711±0.010	0.712±0.014	0.712±0.005
		TRAILS	0.500±0.000	0.507±0.005	0.524±0.005	0.555±0.036	0.577±0.028	0.583±0.015
		SCAN	0.603±0.020	0.621±0.036	0.539±0.022	0.664±0.026	0.748±0.027	0.827±0.002
		SCANT	0.603±0.009	0.603±0.014	0.603±0.016	0.603±0.027	0.603±0.006	0.604±0.011
		SCANS	0.500±0.000	0.496±0.001	0.513±0.013	0.515±0.015	0.570±0.060	0.572±0.007
		CN	0.525±0.000	0.525±0.008	0.524±0.013	0.525±0.005	0.525±0.013	0.525±0.007
		JC	0.527±0.008	0.527±0.011	0.527±0.010	0.528±0.002	0.527±0.016	0.528±0.009
		AA	0.493±0.006	0.490±0.006	0.490±0.012	0.490±0.009	0.493±0.013	0.490±0.006
	Accuracy	TRAIL	**0.654±0.014**	**0.746±0.009**	**0.756±0.009**	**0.764±0.008**	**0.768±0.012**	**0.839±0.002**
		TRAILT	0.655±0.004	0.653±0.008	0.655±0.014	0.655±0.008	0.655±0.008	0.655±0.005
		TRAILS	0.500±0.000	0.501±0.003	0.535±0.009	0.529±0.006	0.535±0.004	0.545±0.014
		SCAN	0.554±0.028	0.567±0.009	0.563±0.007	0.605±0.014	0.656±0.011	0.748±0.012
		SCANT	0.553±0.002	0.553±0.004	0.553±0.003	0.554±0.002	0.553±0.001	0.553±0.003
		SCANS	0.500±0.000	0.498±0.003	0.515±0.008	0.529±0.003	0.536±0.003	0.541±0.005
		NAIVE	0.500±0.000	0.508±0.001	0.514±0.006	0.517±0.002	0.519±0.003	0.526±0.000
location	AUC	TRAIL	**0.871±0.020**	**0.876±0.011**	**0.891±0.006**	**0.881±0.028**	**0.916±0.016**	**0.925±0.007**
		TRAILT	0.871±0.015	0.872±0.004	0.872±0.013	0.872±0.003	0.872±0.017	0.872±0.014
		TRAILS	0.500±0.000	0.492±0.002	0.479±0.004	0.504±0.002	0.580±0.001	0.652±0.003
		SCAN	0.745±0.005	0.746±0.011	0.773±0.010	0.788±0.012	0.796±0.016	0.797±0.009
		SCANT	0.745±0.021	0.744±0.011	0.745±0.025	0.744±0.020	0.743±0.011	0.744±0.010
		SCANS	0.500±0.000	0.490±0.002	0.481±0.002	0.504±0.001	0.578±0.005	0.651±0.005
		FCF	0.682±0.006	0.683±0.002	0.682±0.007	0.683±0.002	0.683±0.006	0.682±0.003
	Accuracy	TRAIL	**0.734±0.008**	**0.754±0.005**	**0.765±0.006**	**0.775±0.003**	**0.789±0.008**	**0.797±0.010**
		TRAILT	0.735±0.002	0.734±0.007	0.734±0.007	0.734±0.006	0.735±0.004	0.735±0.004
		TRAILS	0.500±0.000	0.509±0.003	0.514±0.006	0.511±0.001	0.533±0.000	0.569±0.001
		SCAN	0.731±0.002	0.753±0.001	0.754±0.002	0.755±0.002	0.767±0.002	0.777±0.003
		SCANT	0.732±0.013	0.732±0.010	0.732±0.016	0.732±0.009	0.732±0.004	0.732±0.004
		SCANS	0.500±0.000	0.511±0.002	0.516±0.006	0.517±0.005	0.534±0.001	0.568±0.002
		NAIVE	0.500±0.000	0.509±0.001	0.517±0.001	0.517±0.005	0.525±0.010	0.536±0.004

Convergence Analysis: (a)-(d) are the results when $\sigma = 0.5$ and $\rho = 1.0$; (e)-(h) are the same results when $\sigma = 1.0$ and $\rho = 0.5$



(a) AUC of social links (b) *acc.* of social links (c) AUC of location links (d) *acc.* of location links (e) AUC of social links (f) *acc.* of social links (g) AUC of location links (h) *acc.* of location links

5. Acknowledgements

This work is supported in part by NSF through grants CNS-1115234, DBI-0960443, and OISE-1129076, US Department of Army through grant W911NF-12-1-0066, and Huawei Grant.