UIC

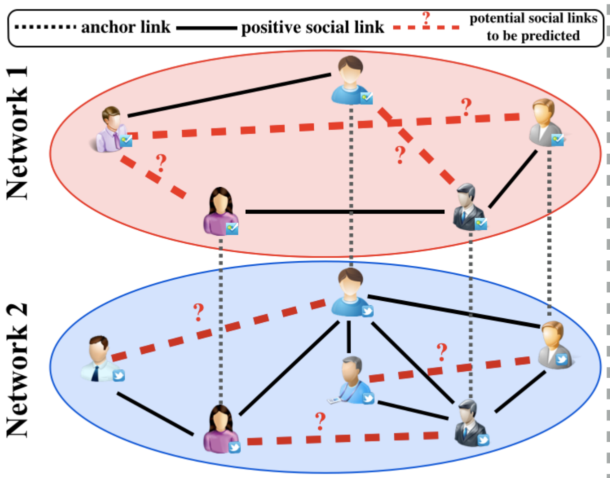
KDD 2014 THE 20TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING AUGUST 2014 | NEW YORK CITY

Meta-path based Multi-Network Collective Link Prediction

Jiawei Zhang[1,2] Philip S. Yu[1] Zhi-Hua Zhou[2]

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

1. Multi-Network Collective Link Prediction Problem



1. Problem Studied: Collective social link prediction across multiple aligned heterogeneous social networks.

2. Anchor User & Non-anchor Users: Users who join multiple social networks simultaneously are called anchor users. The remaining unshared users are called non-anchor users.

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

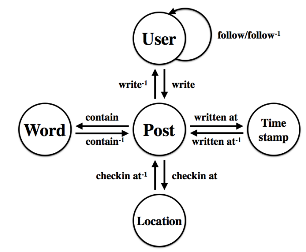
4. Collective Social Link Prediction: Predicting social links to be formed in the future based on a snapshot of multiple aligned online social network simultaneously is called collective social link prediction problem.

Challenge 1: What kind of features can be extracted from the network?

Challenge 2: How to formulate the social link prediction problem?

Challenge 3: How to predict social links in multiple aligned networks simultaneously?

2. Feature extraction based on intra & inter-network meta paths



Definition 10 (Intra-Network Social Meta Path): For a given meta path $\Phi = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \cdots \xrightarrow{R_{k-1}} T_k$ defined based on S_G, if T_1 and T_k are both the "User" node type, then P is defined as a *social meta path*. Depending on whether $T_1, \cdots, T_k$ and $R_1, \cdots, R_{k-1}$ are the same or not, P can be divided into two categories: *homogeneous intra-network social meta path* and *heterogeneous intra-network social meta path*.

Definition 11 (Intra-Network Social Meta Path based Features): For a given link (u, v), the feature extracted for it based on meta path $\Phi = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \cdots \xrightarrow{R_{k-1}} T_k$ from the network is defined to be the expected number of formed path instances between u and v in the network:

$$x(u,v) = I(u,T_1)I(v,T_k) \sum_{n_1\in\{u\},n_2\in T_2,\cdots,n_k\in\{v\}} \prod_{i=1}^{k-1} p(n_i,n_{i+1}) I((n_i,n_{i+1}),R_i),$$

where $p(n_i, n_{i+1}) = 1.0$ if $(n_i, n_{i+1}) \in E_{u,u}$ and otherwise, $p(n_i, n_{i+1})$ denotes the *formation probability* of link (n_i, n_{i+1}) to be introduced in Subsection 3.2.

intra-network meta paths

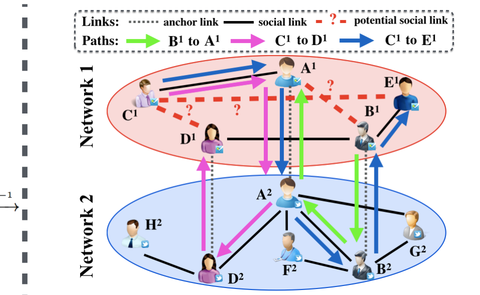
Homogeneous Intra-Network Social Meta Path

- *ID 0. Follow*: User $\xrightarrow{follow}$ User, whose notation is "$U \to U$" or $\Phi_0(U,U)$.
- *ID 1. Follower of Follower*: User $\xrightarrow{follow}$ User $\xrightarrow{follow}$ User, whose notation is "$U \to U \to U$" or $\Phi_1(U,U)$.
- *ID 2. Common Out Neighbor*: User $\xrightarrow{follow}$ User $\xrightarrow{follow^{-1}}$ User, whose notation is "$U \to U \leftarrow U$" or $\Phi_2(U,U)$.
- *ID 3. Common In Neighbor*: User $\xrightarrow{follow^{-1}}$ User $\xrightarrow{follow}$ User, whose notation is "$U \leftarrow U \to U$" or $\Phi_3(U,U)$.

Heterogeneous Intra-Network Social Meta Path

- *ID 4. Common Words*: User $\xrightarrow{write}$ Post $\xrightarrow{contain}$ Word $\xrightarrow{contain^{-1}}$ Post $\xrightarrow{write^{-1}}$ User, whose notation is "$U \to P \to W \leftarrow P \leftarrow U$" or $\Phi_4(U,U)$.
- *ID 5. Common Timestamps*: User $\xrightarrow{write}$ Post $\xrightarrow{contain}$ Time $\xrightarrow{contain^{-1}}$ Post $\xrightarrow{write^{-1}}$ User, whose notation is "$U \to P \to T \leftarrow P \leftarrow U$" or $\Phi_5(U,U)$.
- *ID 6. Common Location Checkins*: User $\xrightarrow{write}$ Post $\xrightarrow{attach}$ Location $\xrightarrow{attach^{-1}}$ Post $\xrightarrow{write^{-1}}$ User, whose notation is "$U \to P \to L \leftarrow P \leftarrow U$" or $\Phi_6(U,U)$.

Definition 12 (Anchor Meta Path): Let U^i, U^j be the user nodes of G^i and G^j respectively and $A^{i,j}$ be the anchor links between G^i and G^j. Meta path $\Upsilon = T_1 \xleftrightarrow{R_1} T_2$ is an *anchor meta path* between network G^i and G^j iff $T_1 = U^i$ and $T_2 = U^j$ and $R_1 = A^{i,j}$. The notation of *anchor meta path* from G^i to G^j is $\Upsilon(U^i, U^j)$ and the length of $\Upsilon(U^i, U^j)$ is 1.



inter-network meta paths

Category 1: $\Upsilon(U^i,U^j)\circ(\Phi(U^j,U^j)\cup\Phi_0(U^j,U^j))\circ\Upsilon(U^j,U^i)$, whose notation is $\Psi_1(U^i,U^i)$;
Category 2.: $(\Phi(U^i,U^i)\cup\Phi_0(U^i,U^i))\circ\Upsilon(U^i,U^j)\circ(\Phi(U^j,U^j)\cup\Phi_0(U^j,U^j))\circ\Upsilon(U^j,U^i)$, whose notation is $\Psi_2(U^i,U^i)$;
Category 3.: $\Upsilon(U^i,U^j)\circ(\Phi(U^j,U^j)\cup\Phi_0(U^j,U^j))\circ\Upsilon(U^j,U^i)\circ(\Phi(U^i,U^i)\cup\Phi_0(U^i,U^i))$, whose notation is $\Psi_3(U^i,U^i)$;
Category 4.: $(\Phi(U^i,U^i)\cup\Phi_0(U^i,U^i))\circ\Upsilon(U^i,U^j)\circ(\Phi(U^j,U^j)\cup\Phi_0(U^j,U^j))\circ\Upsilon(U^j,U^i)\circ(\Phi(U^i,U^i)\cup\Phi_0(U^i,U^i))$, whose notation is $\Psi_4(U^i,U^i)$;
where $\Phi(U^i,U^i)\cup\Phi_0(U^i,U^i) = \{\Phi_0(U^i,U^i),\cdots,\Phi_6(U^i,U^i)\}$ denote the 7 *intra-network social meta paths* of network G^i

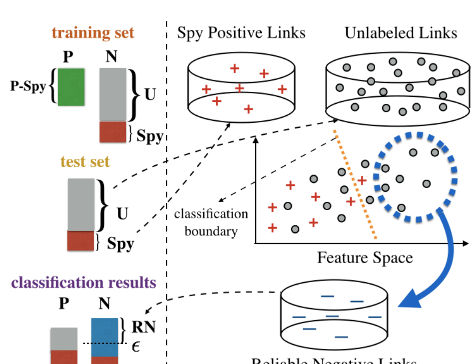
meta path selection

Let variable $X_i \in [\mathbf{x}_\Phi^T, \mathbf{x}_\Psi^T]^T$ be a feature extracted based on a meta path in $\{\Phi, \Psi\}$ and variable Y be the *label*. $P(Y = y)$ denotes the *prior probability* that links in the training set having label y and $P(X_i = x)$ represents the *frequency* that feature X_i has value x. Information theory related measure *mutual information* (mi) is used as the ranking criteria:

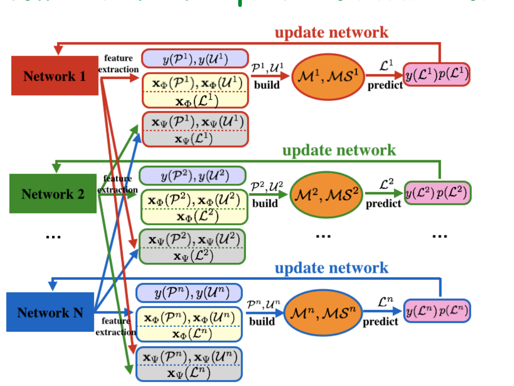
$$mi(X_i) = \sum_x \sum_y P(X_i = x, Y = y) \log \frac{P(X_i = x, Y = y)}{P(X_i = x)P(Y = y)}$$

3. Positive-Unlabeled (PU) Link Prediction Framework

PU link prediction



Collective PU link prediction framework

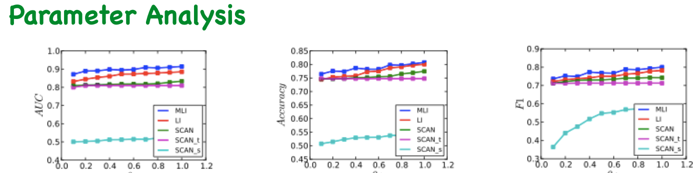


4. Experiment Results & Parameter Analysis & Convergence Analysis

network	measure	methods	0.1	0.2	0.3	0.4	0.5
			Remaining information rates ρ_T of Twitter				
Foursquare	AUC	MLI	**0.862±0.003**	**0.867±0.004**	**0.87±0.003**	**0.873±0.005**	**0.885±0.003**
		LI	0.831±0.005	0.834±0.004	0.846±0.004	0.853±0.005	0.855±0.005
		SCAN	0.81±0.007	0.81±0.008	0.812±0.005	0.817±0.007	0.816±0.01
		SCANT	0.81±0.007	0.81±0.007	0.81±0.007	0.81±0.007	0.809±0.007
		SCANS	0.504±0.007	0.51±0.003	0.511±0.005	0.516±0.005	0.522±0.004
	Accuracy	MLI	**0.78±0.003**	**0.786±0.005**	**0.789±0.004**	**0.794±0.005**	**0.793±0.004**
		LI	0.745±0.011	0.762±0.005	0.768±0.007	0.772±0.007	0.777±0.008
		SCAN	0.749±0.007	0.754±0.006	0.754±0.007	0.757±0.006	0.758±0.007
		SCANT	0.748±0.003	0.748±0.003	0.747±0.003	0.748±0.003	0.748±0.003
		SCANS	0.692±0.011	0.717±0.008	0.725±0.008	0.746±0.008	0.741±0.006
	F1	MLI	**0.768±0.004**	**0.774±0.005**	**0.778±0.006**	**0.784±0.006**	**0.785±0.005**
		LI	0.721±0.02	0.734±0.01	0.734±0.012	0.736±0.012	0.744±0.012
		SCAN	0.717±0.01	0.718±0.007	0.714±0.009	0.715±0.009	0.718±0.011
		SCANT	0.713±0.01	0.712±0.01	0.712±0.01	0.713±0.01	0.713±0.01
		SCANS	0.509±0.02	0.514±0.014	0.524±0.014	0.529±0.013	0.54±0.009
Twitter	AUC	MLI	**0.837±0.004**	**0.858±0.004**	**0.905±0.005**	**0.926±0.003**	**0.924±0.002**
		LI	0.772±0.009	0.829±0.008	0.871±0.009	0.887±0.002	0.887±0.002
		SCAN	0.706±0.008	0.771±0.012	0.799±0.009	0.817±0.002	0.819±0.002
		SCANT	0.555±0.133	0.678±0.006	0.753±0.044	0.754±0.019	0.764±0.014
		SCANS	0.687±0.008	0.687±0.002	0.687±0.005	0.687±0.002	0.687±0.002
	Accuracy	MLI	**0.821±0.005**	**0.864±0.001**	**0.892±0.008**	**0.914±0.004**	**0.925±0.002**
		LI	0.706±0.002	0.834±0.011	0.877±0.003	0.898±0.005	0.912±0.001
		SCAN	0.594±0.006	0.716±0.009	0.781±0.005	0.801±0.003	0.823±0.002
		SCANT	0.547±0.062	0.645±0.038	0.723±0.048	0.786±0.004	0.8±0.002
		SCANS	0.59±0.009	0.59±0.007	0.59±0.004	0.59±0.004	0.59±0.002
	F1	MLI	**0.713±0.009**	**0.762±0.005**	**0.791±0.006**	**0.81±0.004**	**0.81±0.002**
		LI	0.651±0.006	0.671±0.023	0.749±0.014	0.779±0.007	0.801±0.003
		SCAN	0.6±0.017	0.633±0.023	0.657±0.013	0.684±0.004	0.703±0.004
		SCANT	0.552±0.113	0.574±0.016	0.604±0.031	0.618±0.003	0.63±0.001
		SCANS	0.575±0.025	0.575±0.016	0.575±0.005	0.575±0.006	0.575±0.004

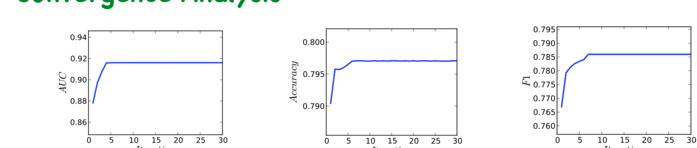
network	measure	methods	0.1	0.2	0.3	0.4	0.5
			Remaining information rates ρ_F of Foursquare.				
Foursquare	AUC	MLI	**0.677±0.023**	**0.776±0.011**	**0.844±0.008**	**0.887±0.005**	**0.906±0.003**
		LI	0.573±0.019	0.68±0.023	0.806±0.01	0.853±0.004	0.866±0.003
		SCAN	0.549±0.009	0.56±0.009	0.662±0.03	0.745±0.009	0.786±0.014
		SCANT	0.5±0.083	0.503±0.007	0.613±0.012	0.739±0.008	0.764±0.013
		SCANS	0.524±0.013	0.524±0.017	0.524±0.012	0.524±0.005	0.524±0.002
	Accuracy	MLI	**0.632±0.01**	**0.692±0.007**	**0.755±0.005**	**0.769±0.004**	**0.779±0.002**
		LI	0.568±0.013	0.624±0.053	0.699±0.004	0.722±0.006	0.761±0.01
		SCAN	0.558±0.007	0.6±0.006	0.683±0.071	0.714±0.009	0.721±0.007
		SCANT	0.491±0.019	0.568±0.004	0.65±0.008	0.685±0.007	0.714±0.007
		SCANS	0.548±0.011	0.548±0.055	0.548±0.007	0.548±0.008	0.548±0.007
	F1	MLI	**0.644±0.01**	**0.695±0.022**	**0.722±0.013**	**0.742±0.005**	**0.761±0.005**
		LI	0.63±0.017	0.635±0.015	0.66±0.007	0.684±0.01	0.715±0.016
		SCAN	0.6±0.02	0.609±0.006	0.614±0.031	0.632±0.018	0.645±0.018
		SCANT	0.534±0.196	0.559±0.004	0.565±0.016	0.584±0.011	0.645±0.011
		SCANS	0.56±0.016	0.56±0.041	0.56±0.015	0.56±0.015	0.56±0.013
Twitter	AUC	MLI	**0.884±0.004**	**0.891±0.003**	**0.915±0.003**	**0.917±0.003**	**0.923±0.002**
		LI	0.841±0.003	0.847±0.002	0.852±0.003	0.862±0.002	0.873±0.002
		SCAN	0.801±0.003	0.814±0.002	0.819±0.003	0.817±0.002	0.819±0.002
		SCANT	0.802±0.002	0.802±0.002	0.802±0.002	0.802±0.002	0.802±0.002
		SCANS	0.508±0.002	0.543±0.002	0.584±0.003	0.631±0.001	0.653±0.002
	Accuracy	MLI	**0.92±0.003**	**0.927±0.002**	**0.927±0.003**	**0.929±0.004**	**0.93±0.003**
		LI	0.899±0.004	0.904±0.004	0.908±0.004	0.913±0.002	0.916±0.003
		SCAN	0.831±0.005	0.835±0.003	0.837±0.006	0.842±0.001	0.844±0.002
		SCANT	0.827±0.003	0.827±0.003	0.827±0.003	0.827±0.003	0.827±0.003
		SCANS	0.568±0.004	0.577±0.003	0.585±0.002	0.587±0.002	0.591±0.003
	F1	MLI	**0.804±0.002**	**0.808±0.002**	**0.809±0.003**	**0.811±0.003**	**0.812±0.003**
		LI	0.776±0.005	0.785±0.005	0.792±0.005	0.8±0.003	0.804±0.003
		SCAN	0.682±0.006	0.686±0.004	0.69±0.006	0.699±0.001	0.703±0.003
		SCANT	0.683±0.003	0.683±0.003	0.683±0.003	0.683±0.003	0.683±0.003
		SCANS	0.53±0.006	0.546±0.006	0.559±0.004	0.564±0.004	0.571±0.004

Parameter Analysis



(a) Foursquare-AUC (b) Foursquare-Acc. (c) Foursquare-F1

Convergence Analysis



(a) Foursquare-AUC (b) Foursquare-Acc. (c) Foursquare-F1

5. Acknowledgement

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 the Pinnacle Lab at Singapore Management University, NSFC (61333014, 61321491) and 111 Program (B14020).