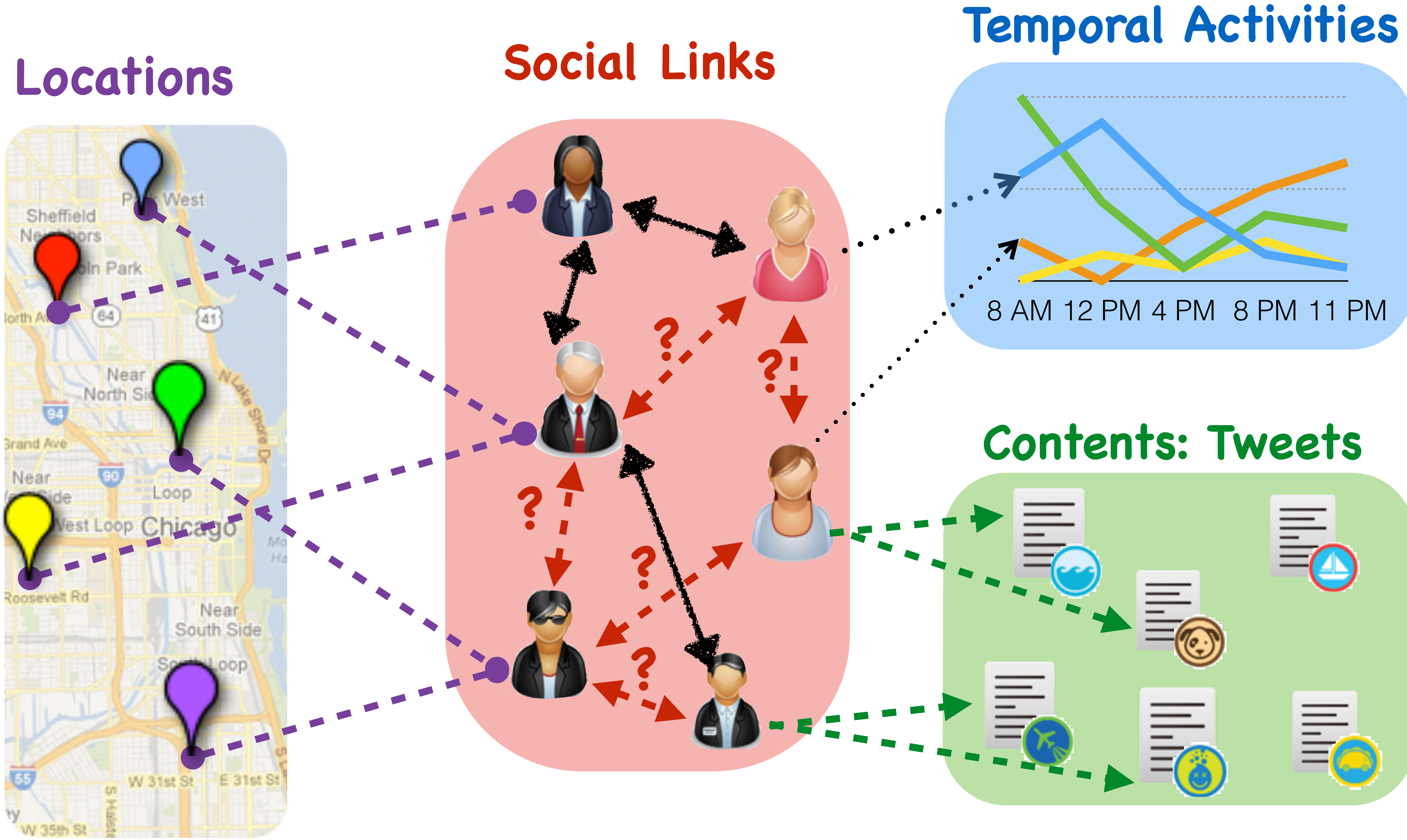
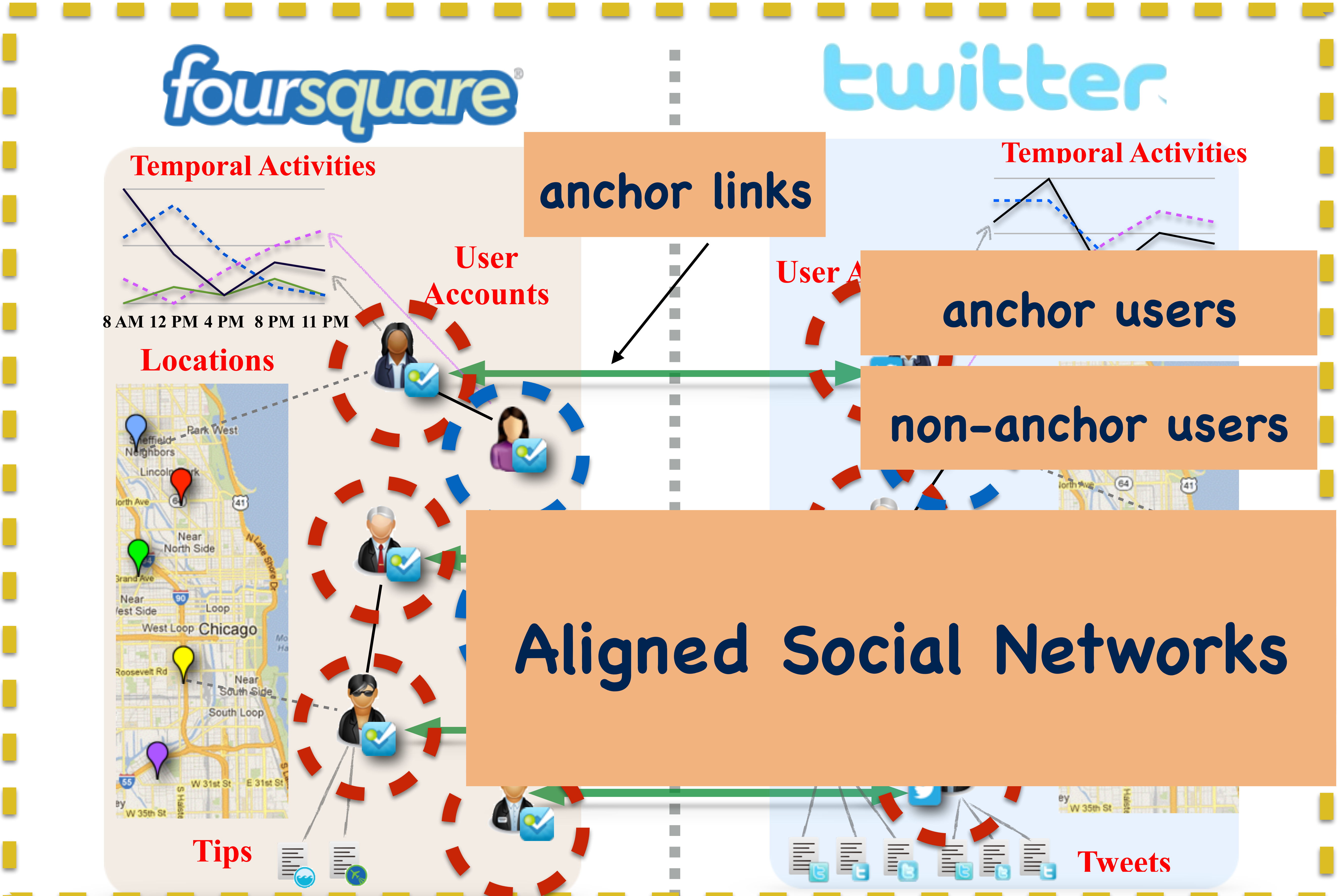
Meta-path based Multi-Network Collective Link Prediction

Jiawei Zhang[1,2], Philip S. Yu[1], Zhi-Hua Zhou[2]
University of Illinois at Chicago[2], Nanjing University[2]

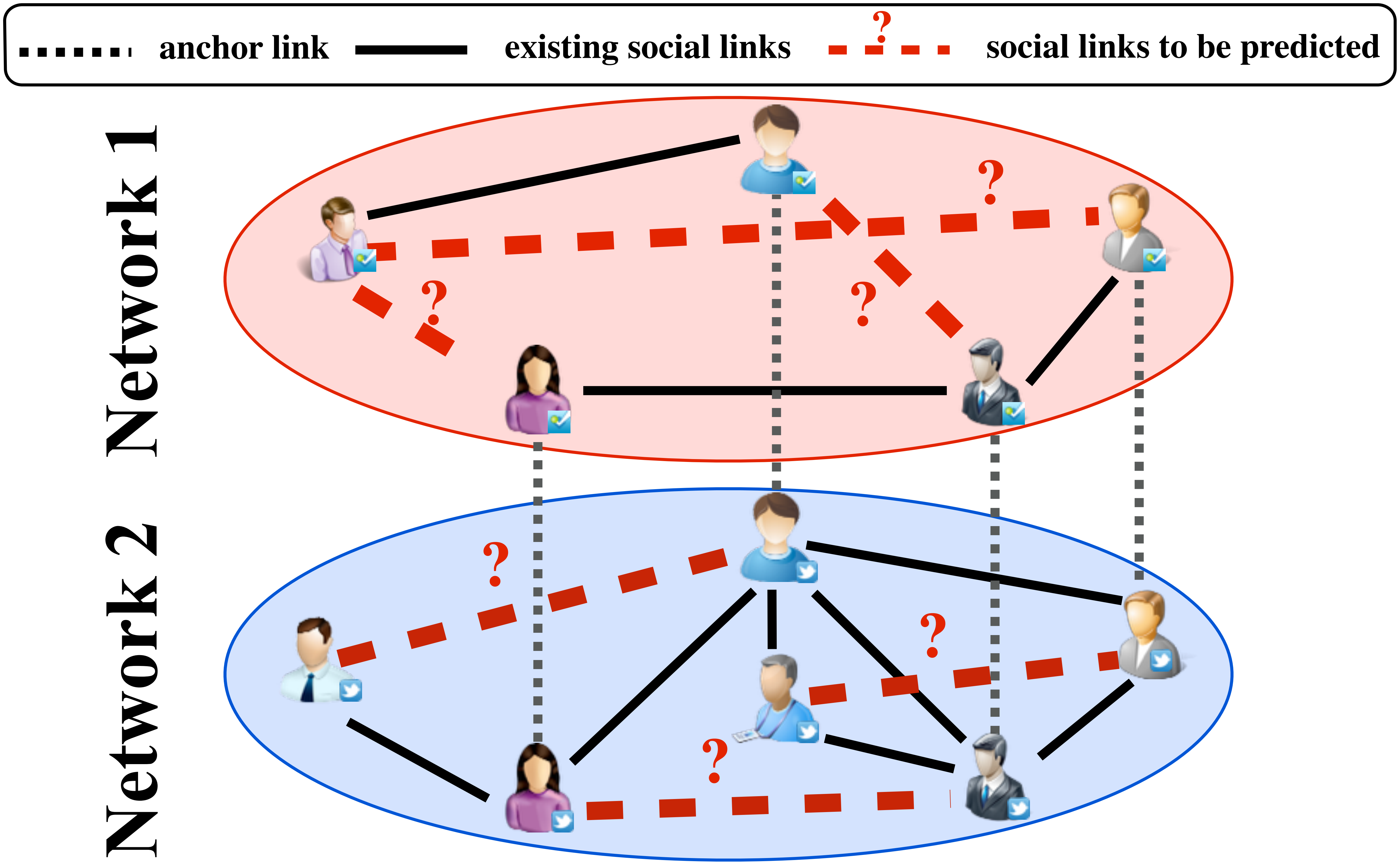
Traditional social link prediction in one single social network



Users use multiple social networks simultaneously



Predicting social links in multiple aligned networks simultaneously



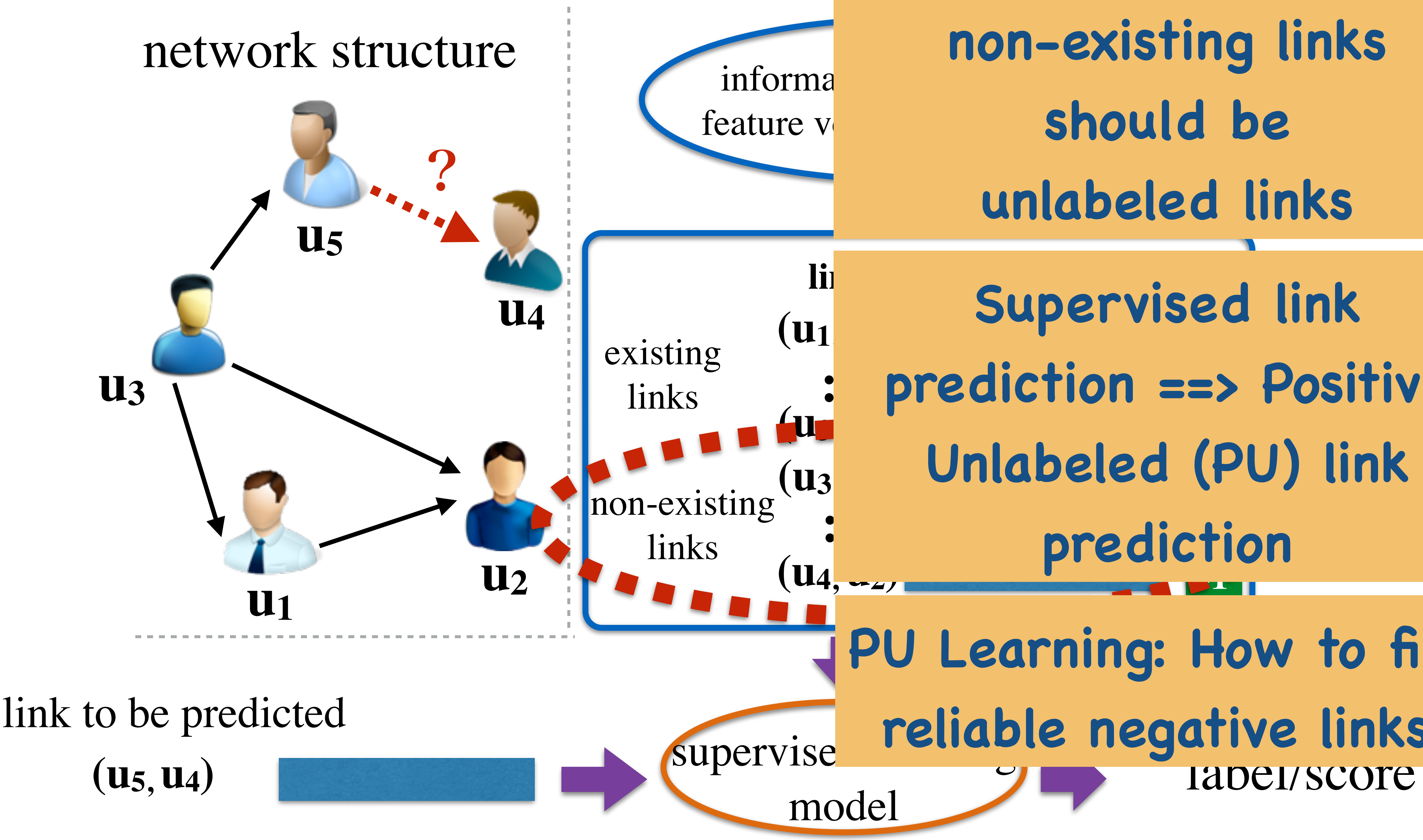
class imbalance problem
negative instances >>
positive instances

non-existing links
!=
negative links

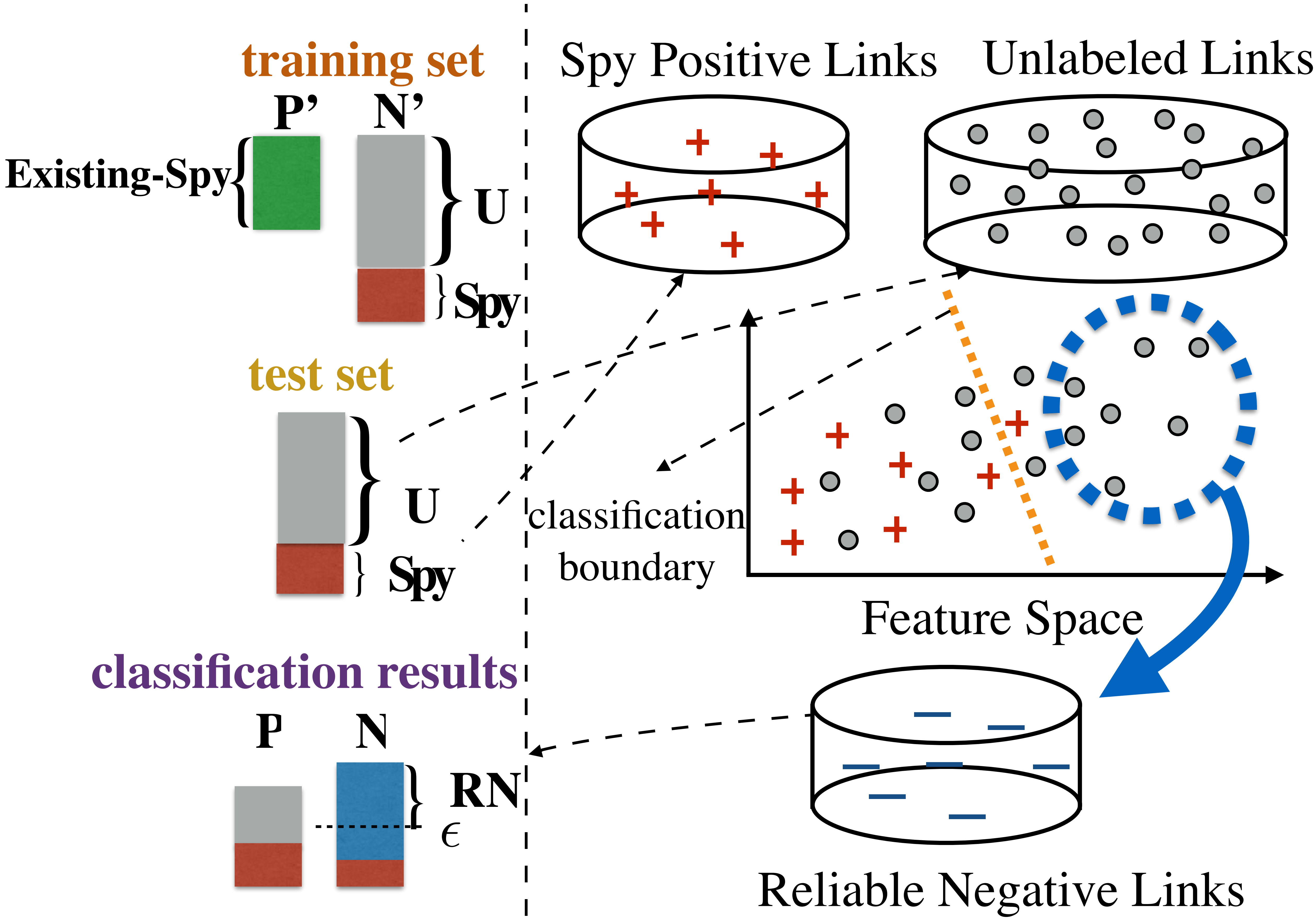
non-existing links
should be
unlabeled links

Supervised link prediction ==> Positive Unlabeled (PU) link prediction

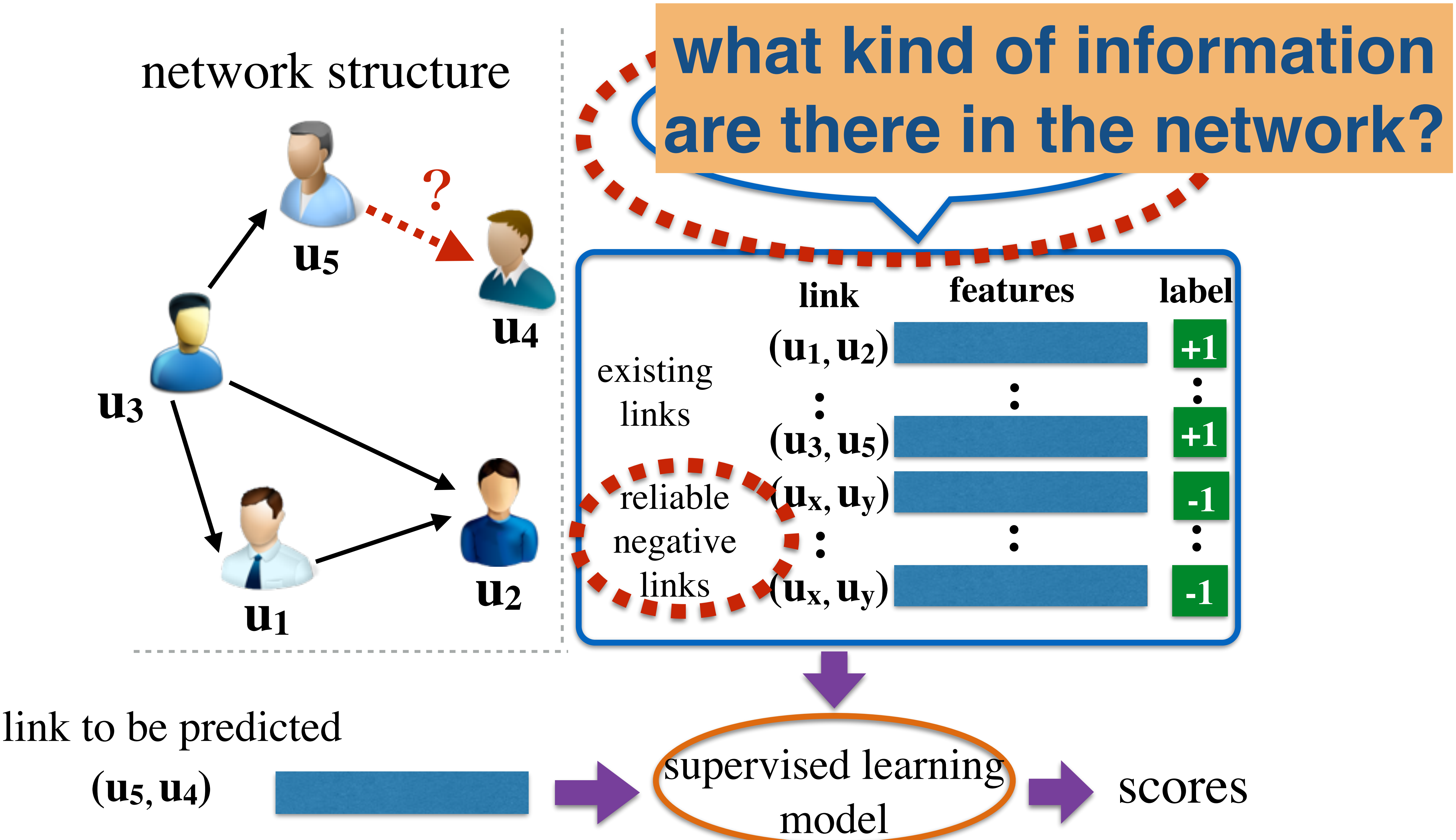
PU Learning: How to find reliable negative links?



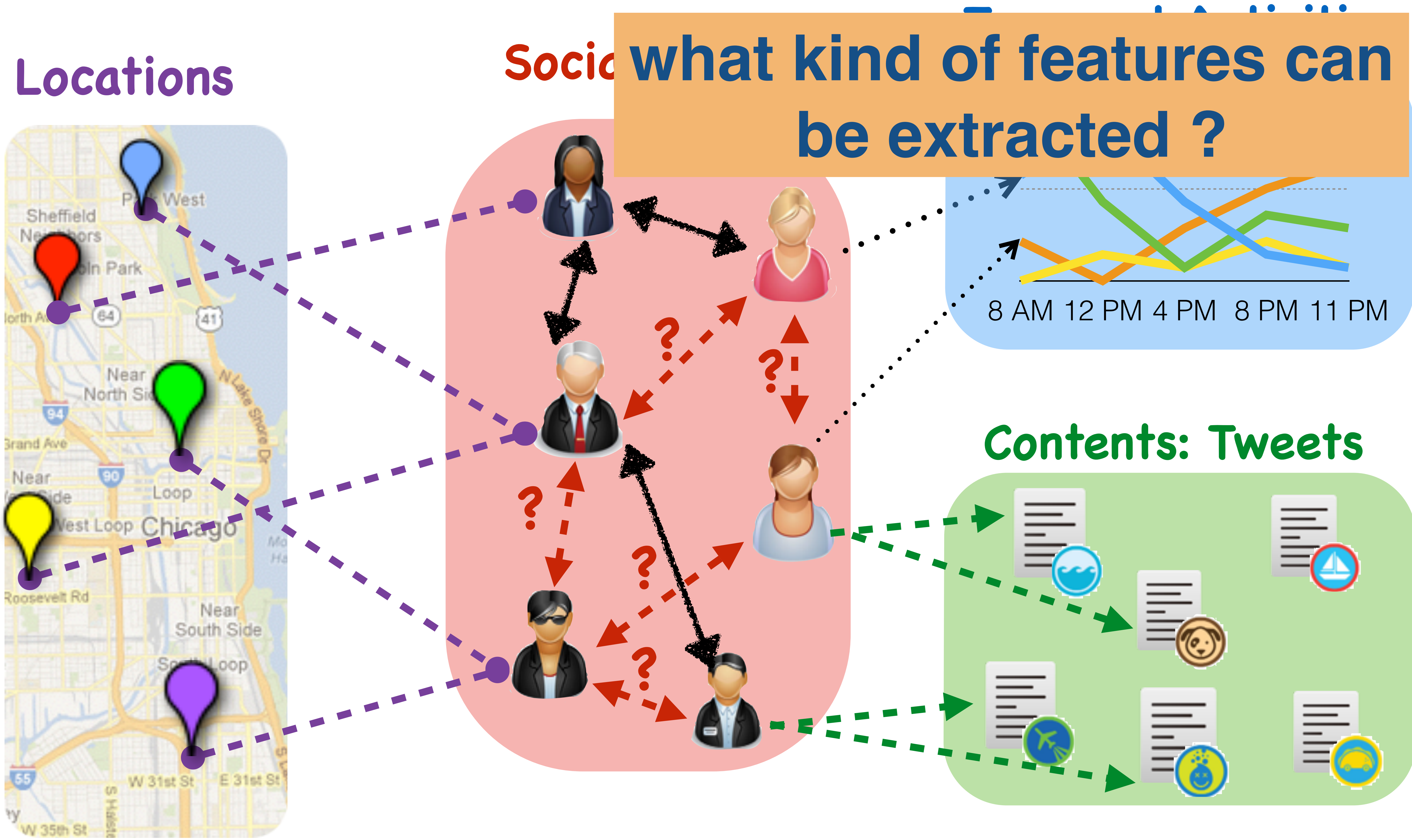
Reliable Negative Links Extraction



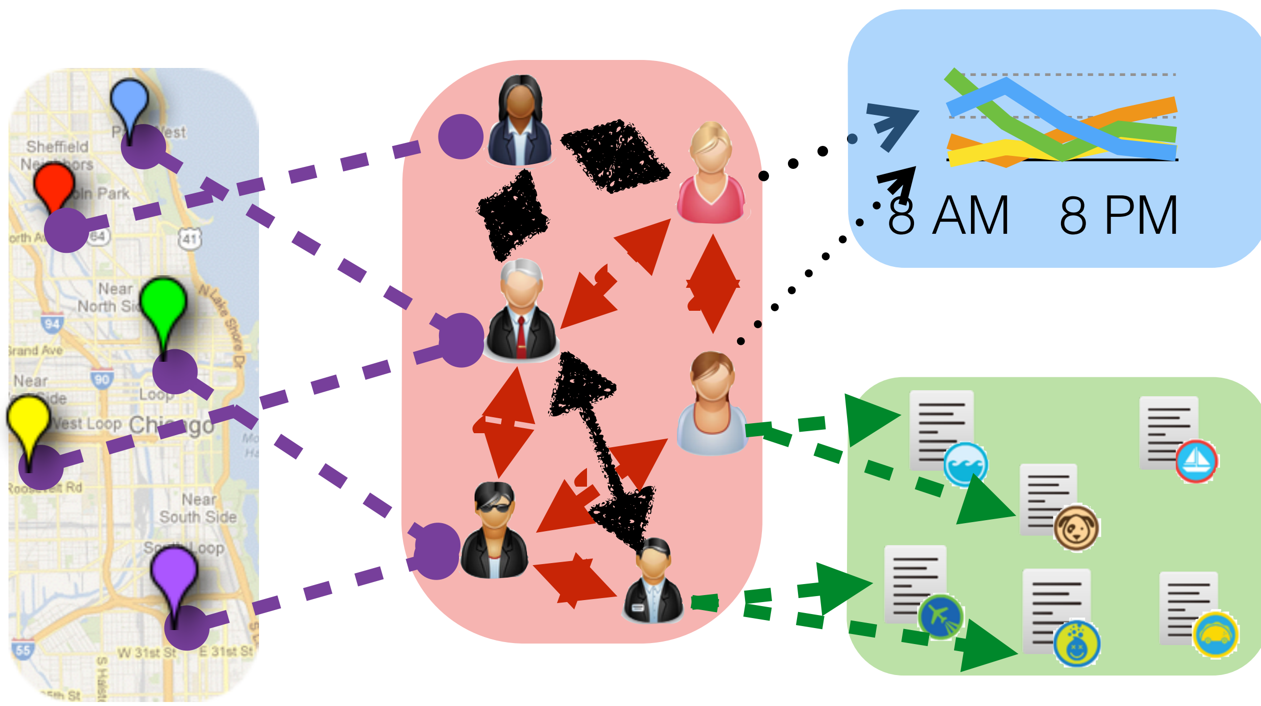
PU Link Prediction Setting



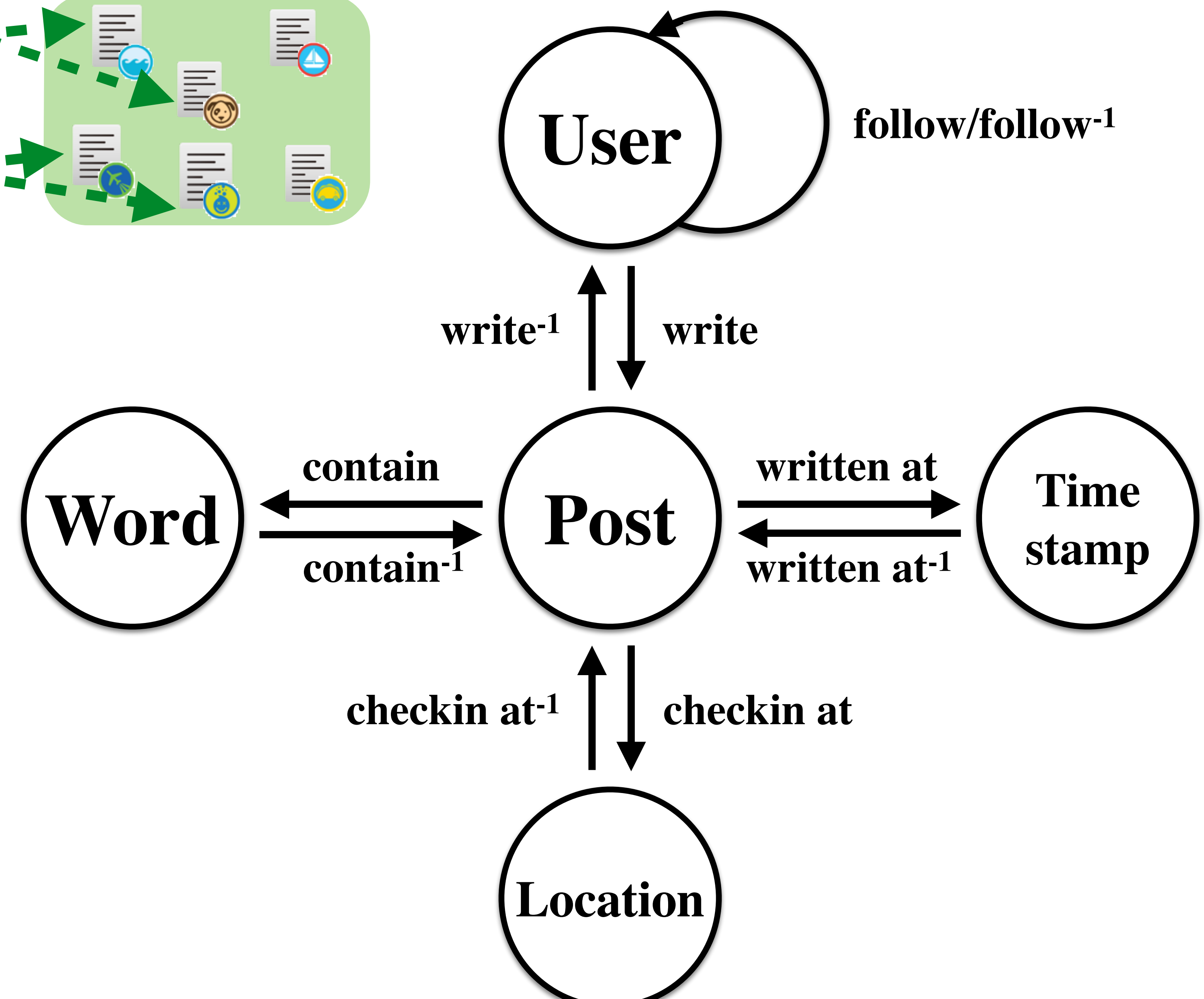
Heterogeneous Information



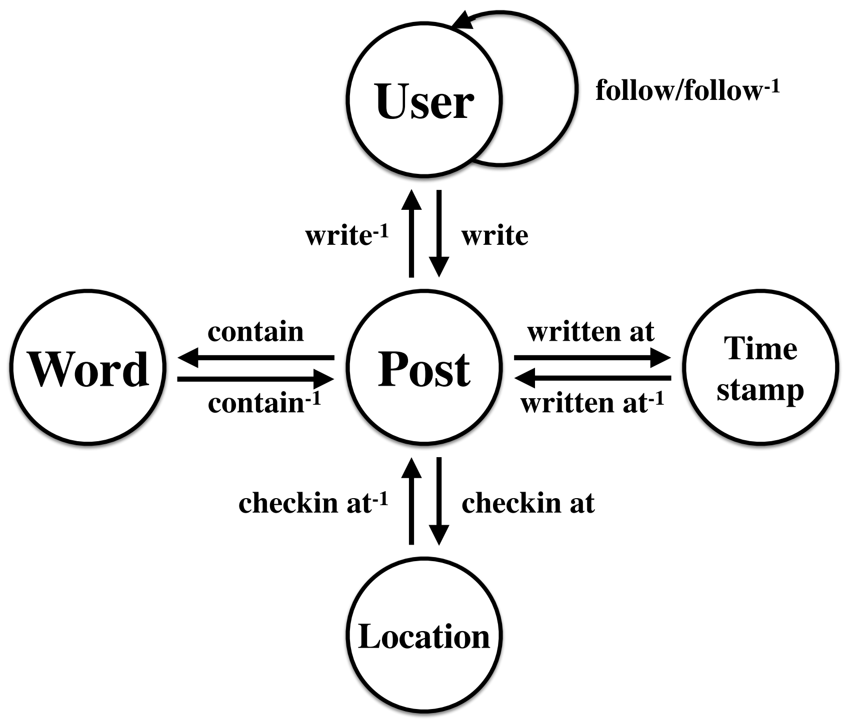
Network Schema



social network schema



Intra-network social 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*.

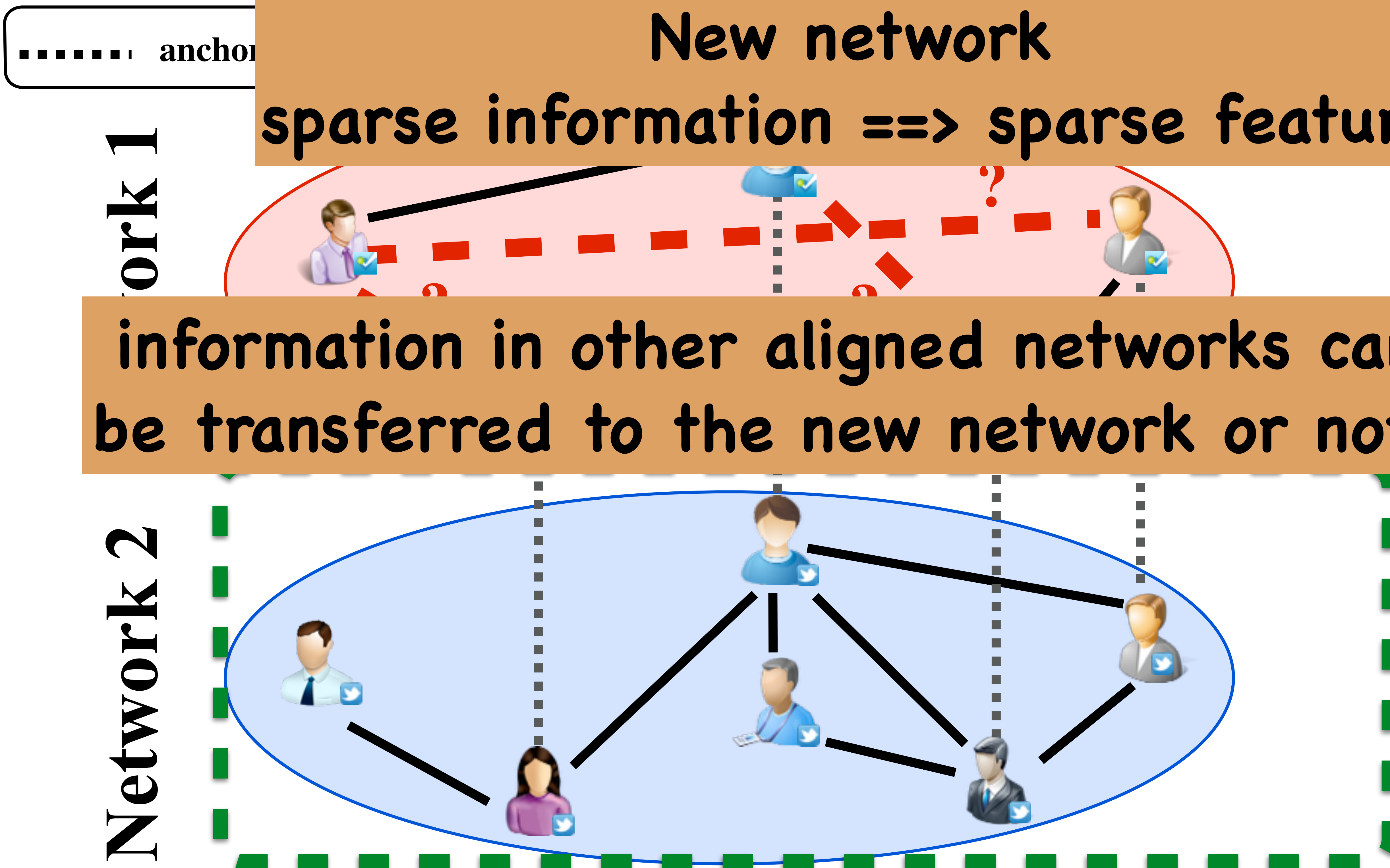
Homogeneous Intra-Network Social Meta Path

- *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 $\Phi_6(U, U)$.

New network problem



New network

sparse information ==> sparse feature

information in other aligned networks can be transferred to the new network or not?

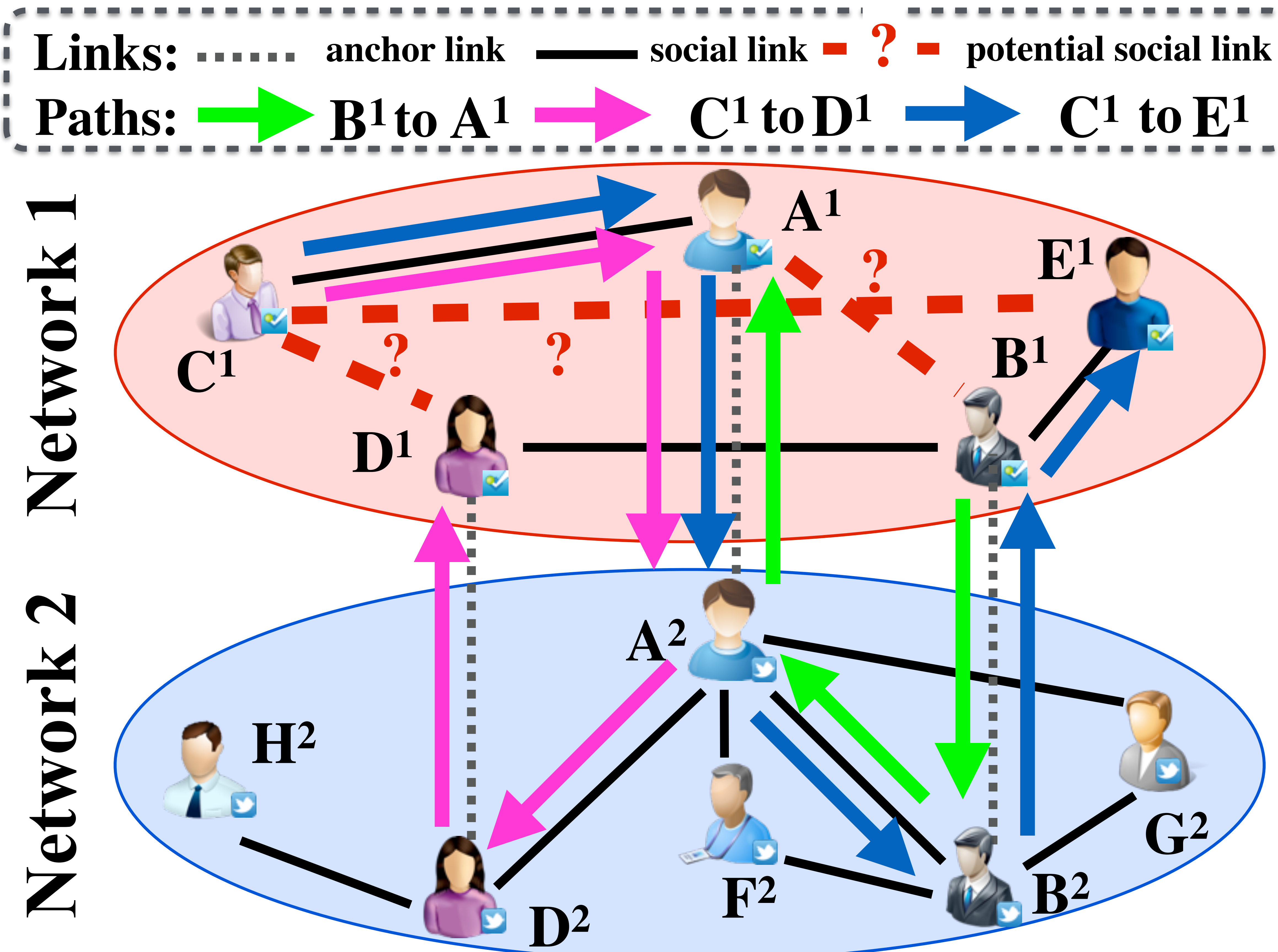
Anchor Meta path & Inter-network social meta paths

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.

Definition 13 (Inter-Network Meta Path): Meta path $\Psi = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \cdots \xrightarrow{R_{k-1}} T_k$ is an *inter-network meta path* across G^i and G^j iff $\exists m \in \{1, 2, \cdots, k-1\}$, $T_m \xleftrightarrow{R_m} T_{m+1} = \Upsilon(U^i, U^j)$.

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 $\Psi_4(U^i,U^i)$;

Inter-network social meta path instances

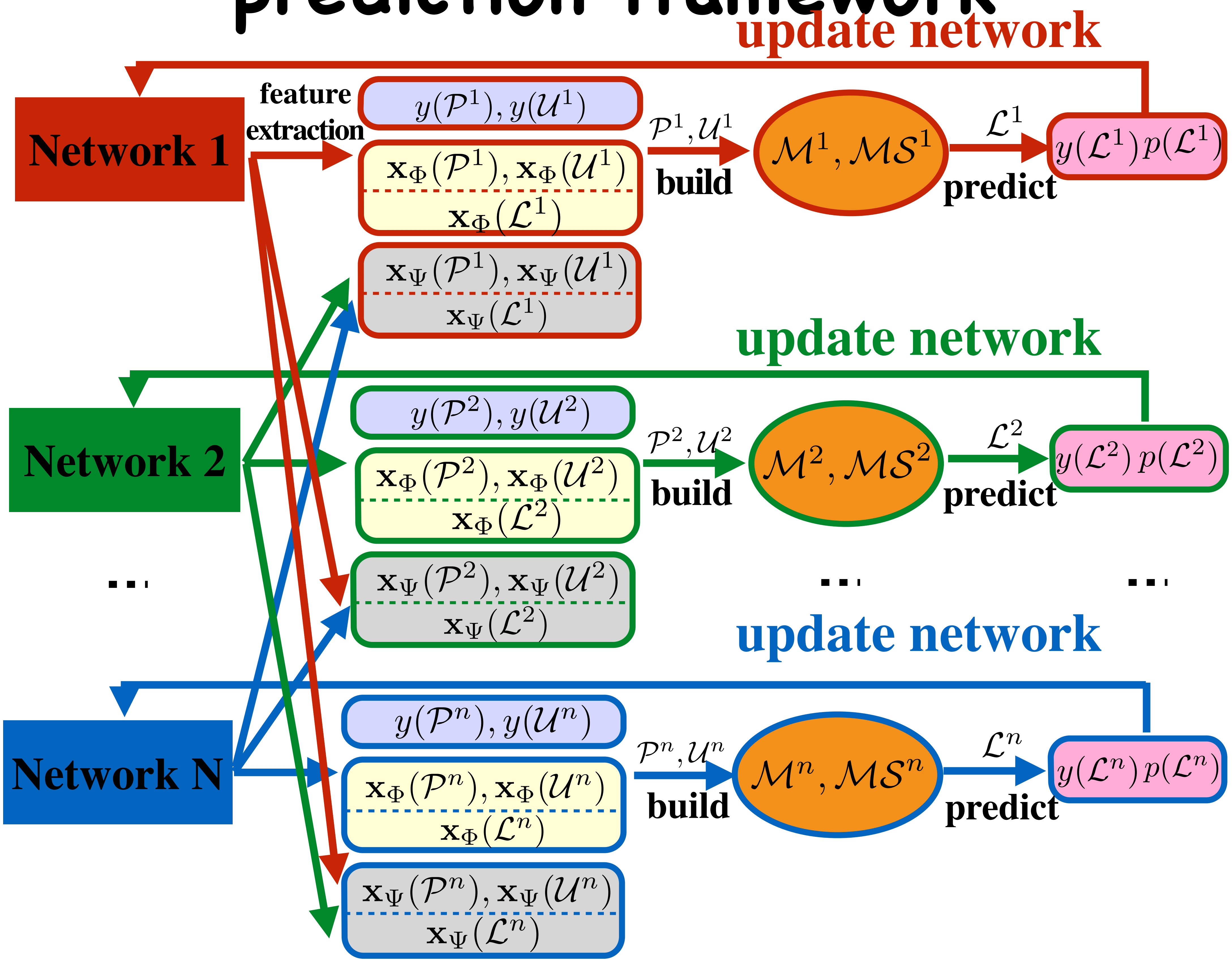


Meta path selection

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:

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

Multi-network collective link prediction framework



Dataset

- Foursquare and Twitter

Table 2: Properties of the Heterogeneous Networks

	property	network: Twitter	network: 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

Experiment Settings

- Ground truth: existing social link among users
  - hide part of the existing links in the test set
  - build model to discover these links

- Comparison Methods
  - MLI (Multi-network Link Identifier)
  - LI (Link Identifier): predict links in each network independently
  - SCAN(Supervised Cross-Aligned-Network link prediction): supervised link prediction, no meta path selection,
  - SCAN_s (SCAN with source network): features are extracted based on inter-network meta paths
  - SCAN_t (SCAN with target network): features are extracted based on intra-network meta paths

- Evaluation Metrics
  - AUC, Accuracy, F1

Experiment Results

collective link prediction is better than independent link prediction

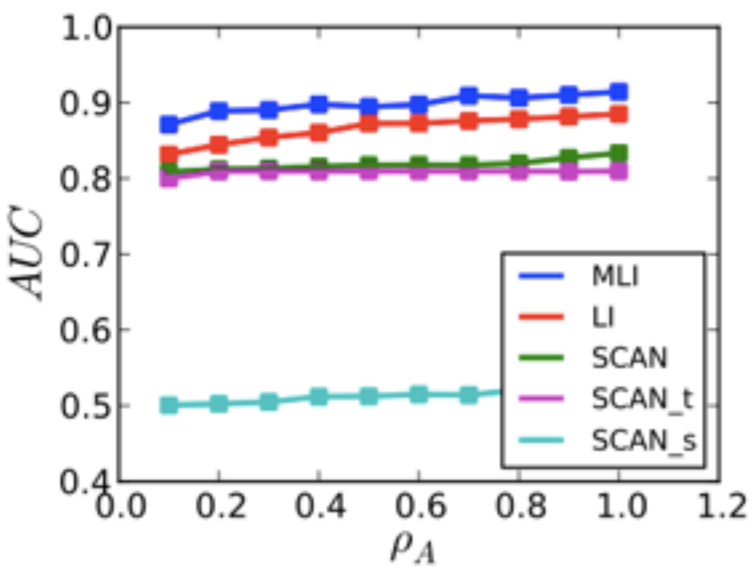
PU link prediction setting and meta path selection can improve the results

using features based on intra-network meta paths and inter-network meta paths simultaneously can achieve better results

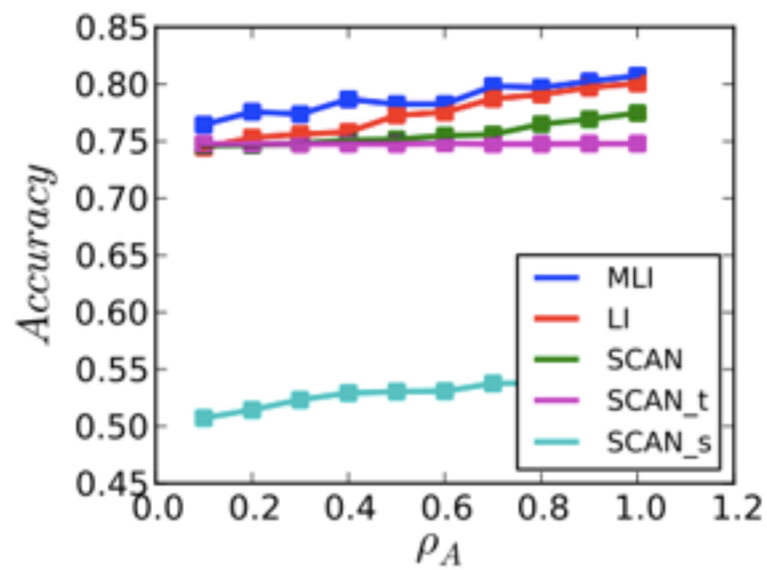
network	measure	methods					
Foursquare	AUC	MLI	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		LI	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		SCAN	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		SCANT	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		SCANs	0.524±0.013	0.524±0.017	0.524±0.012	0.524±0.005	0.524±0.002
	Accuracy	MLI	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		SCANs	0.548±0.011	0.548±0.055	0.548±0.007	0.548±0.008	0.548±0.007
	F1	MLI	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		SCANs	0.56±0.016	0.56±0.041	0.56±0.015	0.56±0.015	0.56±0.013

Parameter Analysis

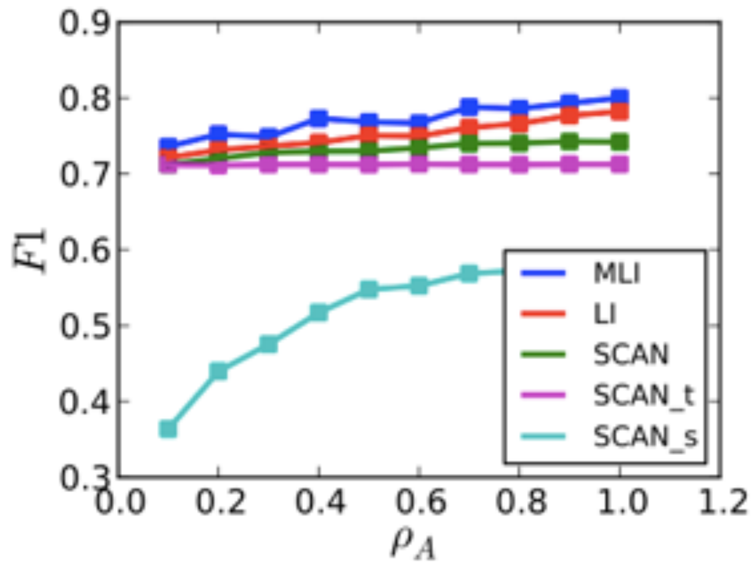
- ratio of anchor links



(a) Foursquare-AUC



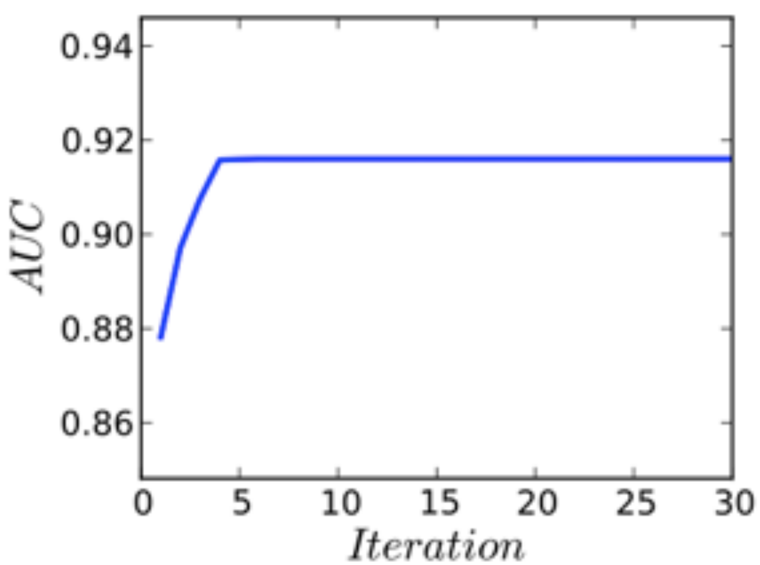
(b) Foursquare-Acc.



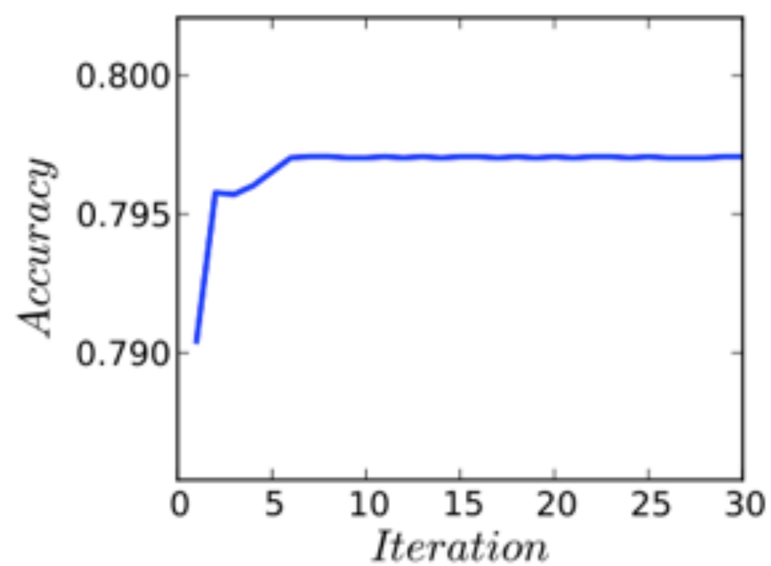
(c) Foursquare-F1

the more anchor links we have, the better performance we can achieve

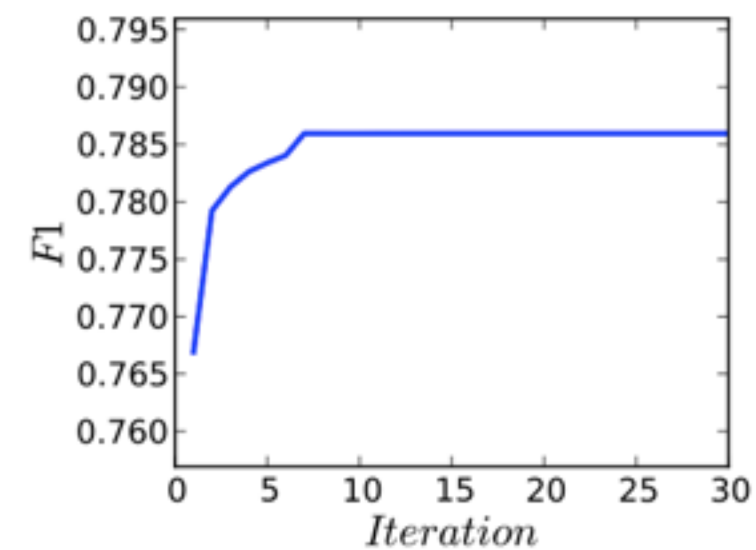
Convergence Analysis



(a) Foursquare-AUC



(b) Foursquare-Acc.



(c) Foursquare-F1

converge quickly, in less than 10 iterations

Conclusions

- Problem studied: collective link prediction across multiple aligned social networks

- Proposed Method:
  - PU Link Prediction Setting
  - Intra-network & Inter-network Meta Path based Feature Extraction
  - Meta path selection
  - Multi-network Collective PU Link Prediction Framework

- Experiment Results:
  - Collective Link Prediction is better than Independent Link Prediction
  - PU Link Prediction & Meta Path Selection can improve the results
  - Using information across networks can achieve better results
  - MLI can perform well consistently for different anchor link ratios & can converge quickly

Q&A