"Meta-path based Multi-Network Collective Link Prediction"

Jiawei Zhang (Univ. of Illinois), Philip S. Yu (Univ. of Illinois), Zhi-Hua Zhou (Nanjing Univ.)

KDD2014読み会 @京都大学 2014/09/23

横井 祥

京都大学 情報学科 鹿島研究室

紹介する論文

Meta-path based Multi-Network Collective Link Prediction

Jiawei Zhang
Big Data and Social Computing (BDSC) Lab
University of Illinois at Chicago
Chicago, IL, USA
jzhan9@uic.edu

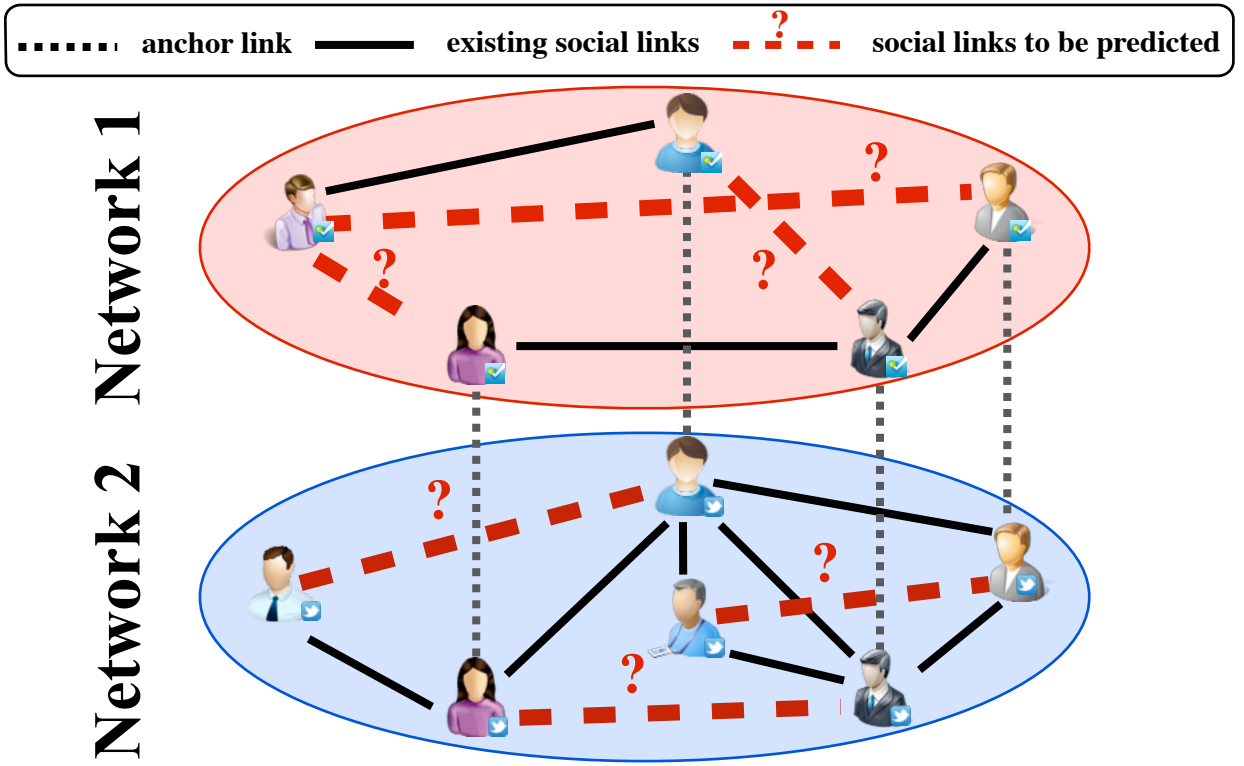
Philip S. Yu
Big Data and Social Computing (BDSC) Lab
University of Illinois at Chicago
Chicago, IL, USA
psyu@cs.uic.edu

Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology,
Nanjing University
Nanjing 210023, China
zhouzh@lamda.nju.edu.cn

- 図表はすべて，主著者 Jiawei Zhang が公開している論文PDF，発表スライドPDF，ポスターPDFのものを利用しています.
  - http://www.cs.uic.edu/~jzhang2/publication.html

解く問題：multiple partial aligned networks の同時リンク予測

- **リンク予測** (Link Prediction)
  - e.g. 「友達かも」
  - 「未来にリンクが張られるかどうか」を予測する
- ＋ 複数ネットワーク
- ＋ 一部接続されている



複数のネットワークを連携させる理由

- **New network problem** [25, 26]
  - リンク予測で用いられてきた従来の特徴量 (e.g. common neighborhoods) は、既存のネットワークから十分な情報を取り出せる (既存ネットワークが十分密) であることを仮定
  - 新しい地理的領域やコミュニティでは, social network は疎

- → multiple partial aligned networks でのリンク予測 [25, 26]
  - 多くの場合, ユーザは他のSNSを利用している
  - **情報が潤沢な (かもしれない) 他のSNS (social network) の情報を利用 (transfer) したい**

multiple (partial) aligned networks のリンク予測の既存手法 [25, 26] の問題点

- anchor users (両方のSNSを利用しているユーザ) にしか適用できない
- **Network difference problem** [15, 23]
  - 異なるネットワークは異なる特性を持つため, 情報を転移させると性能が落ちることがある
  - …という問題を解決しきれていない

- **問題設定**：
  multiple partial aligned networks の同時リンク予測
- 提案手法：
  - PU (Positive Unlabeled) Link Prediction
  - Meta Path ベースの特徴量
  - 特徴量の選別
  - → Multi-Network Link Prediction Framework
- 実験

- 問題設定：
  multiple partial aligned networks の同時リンク予測
- **提案手法**：
  - PU (Positive Unlabeled) Link Prediction
  - Meta Path ベースの特徴量
  - 特徴量の選別
  - → Multi-Network Link Prediction Framework
- 実験

- 問題設定：
  multiple partial aligned networks の同時リンク予測
- **提案手法**：
  - **PU (Positive Unlabeled) Link Prediction**
  - Meta Path ベースの特徴量
  - 特徴量の選別
  - → Multi-Network Link Prediction Framework
- 実験

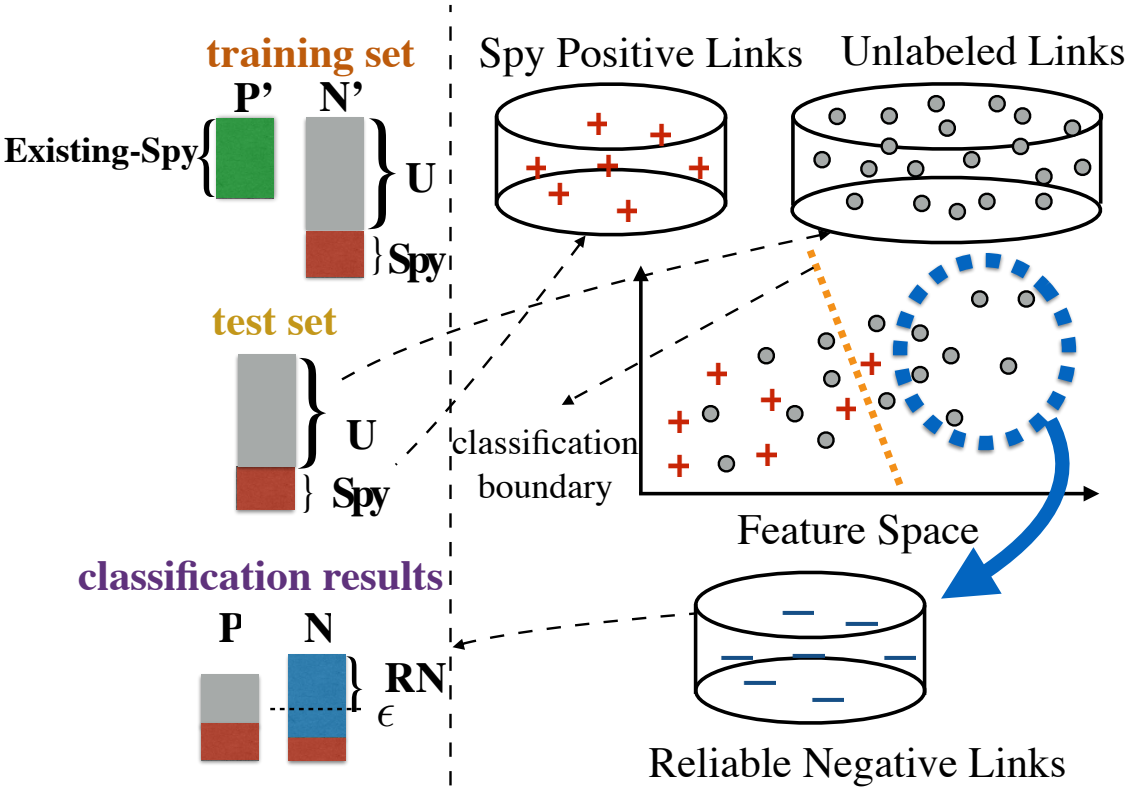
教師データの扱い：PU Link Prediction

- Supervised link prediction ?
  - 与えられた social networks のエッジ (Positive link) は疎
  - エッジが張られていない頂点の組を negative とすると
    - negative links >> positive links

教師データの扱い：PU Link Prediction

- Supervised link prediction ?
  - 与えられた social networks のエッジ (positive link) は疎
  - エッジが張られていない頂点の組を negative とすると
    - negative links >> positive links
- **PU (Positive Unlabeled) link prediction**
  - エッジが張られていない頂点の組は **unlabeled** と考えるべき
  - positive links と unlabeled links が与えられた半教師あり学習と考える
  - → unlabeled の中から **reliable negative** を探したい

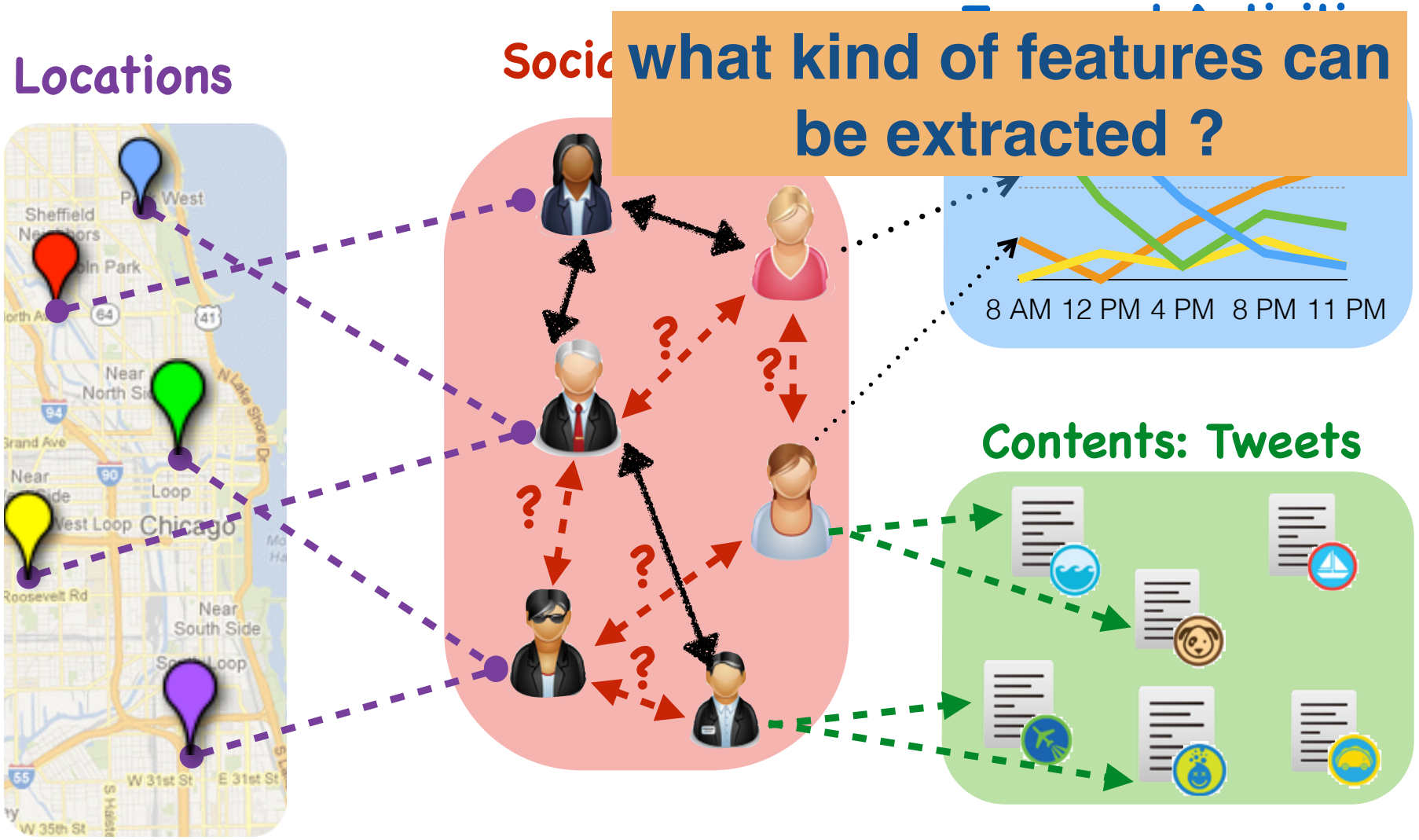
Reliable Negative リンクの抽出



- Spy の中で最も Negative っぽいもの
  …よりも更に Negative なものを, **Reliable Negative** とする.
- Spy Technique has been proposed: [13](CIKM 2013).
- [28] (ICDM 2011) apply this technique to graph mining area.

- 問題設定：
  multiple partial aligned networks の同時リンク予測
- **提案手法**：
  - PU (Positive Unlabeled) Link Prediction [13][28]
  - **Meta Path ベースの特徴量**
  - 特徴量の選別
  - → Multi-Network Link Prediction Framework
- 実験

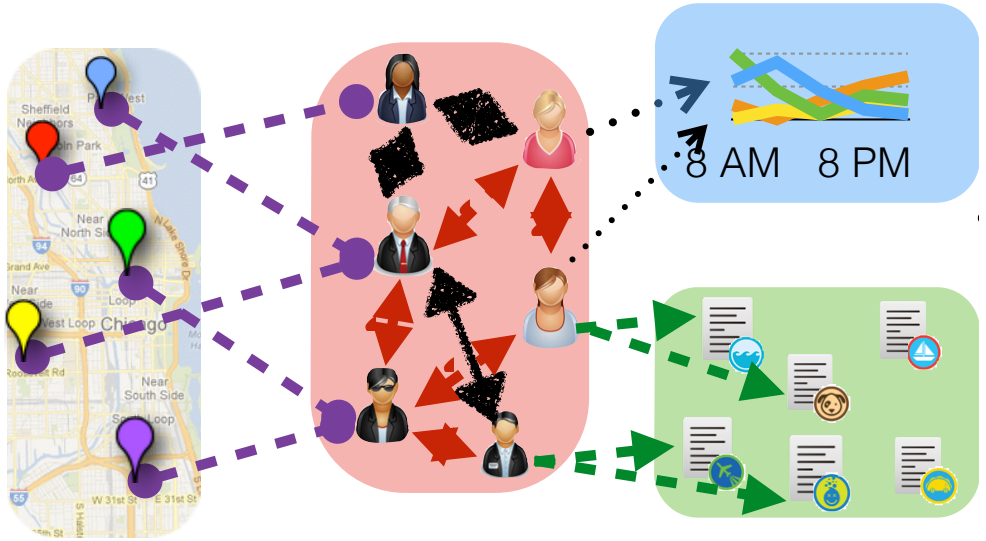
特徴量をどう取り出すか

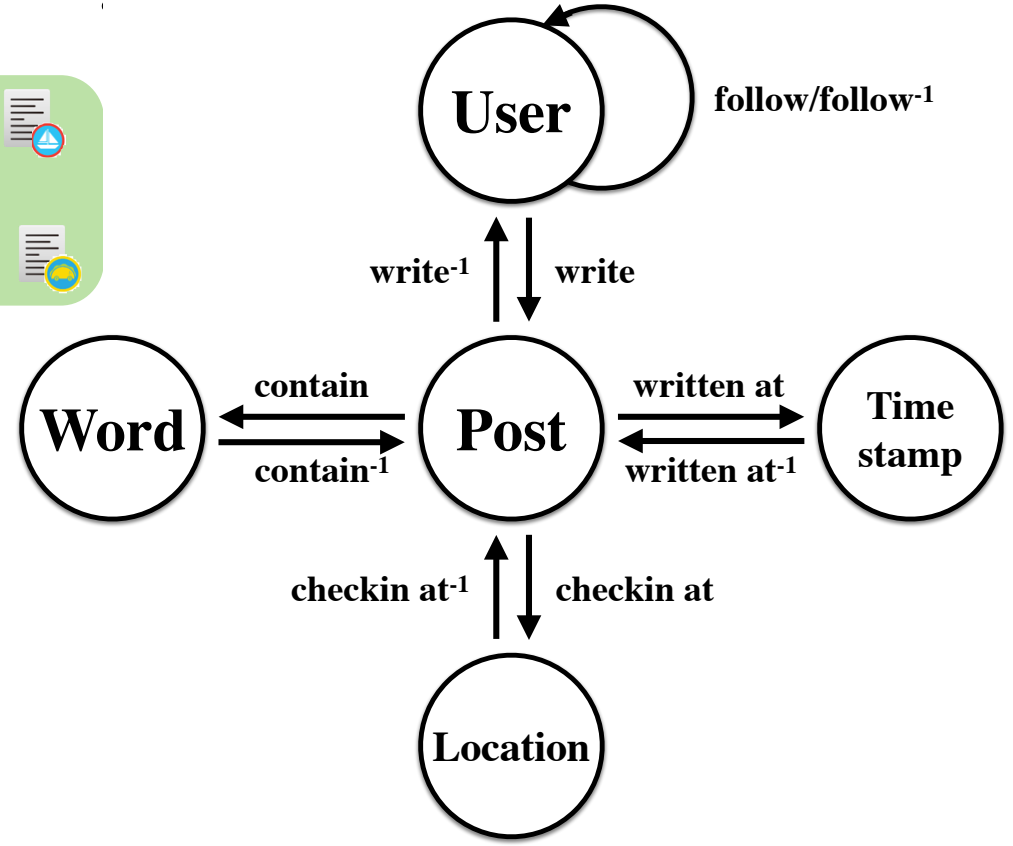


Meta Path：グラフの "パス" の拡張
ノードやエッジの種類を考える

- "**Meta Path**: A Key to Mining **Heterogeneous** Information Networks" (Jiawei Han, Univ. of Illinois)
- いままでの path: 「あるかないか」「パス長」
  - "Alice --(follow)--> Bob --(follow)--> Eve"
  - **Path**: * → * → *
- 色々な種類のノード、色々な種類のエッジがあれば、パスにも色々な種類が考えられる
  - "Alice --(checkin at)--> Kyoto Univ. --(checkin at $^{-1}$)--> Bob"
  - **Meta Path**: User → Location → User

SNSのNetwork Schema





Intra-network Social Meta Path

- **Homogeneous** Intra-network Social Meta Path
- **同種**のノードを繋ぐパス

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

Intra-network Social Meta Path

- **Heterogeneous** Intra-network Social Meta Path
- **別種**のノードを繋ぐパス

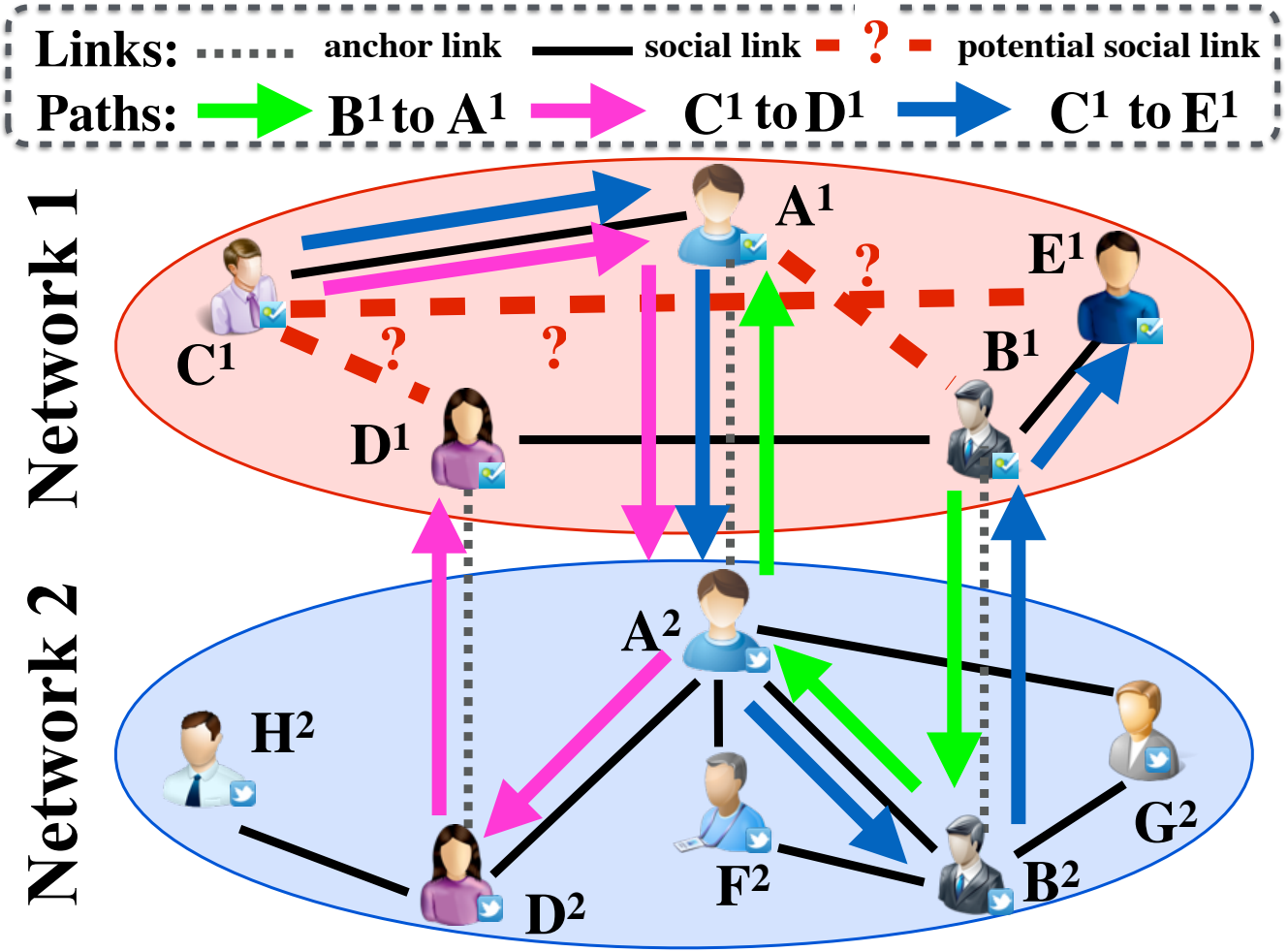
  - *ID 4. Common Words*: User $\xrightarrow{write}$ Post $\xrightarrow{contain}$ Word $\xrightarrow{contain^{-1}}$ Post $\xrightarrow{write^{-1}}$ User, whose notation is "$U \to P \to W \leftarrow P \leftarrow U$" or $\Phi_4(U, U)$.

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

Inter-network Social Meta Path

- anchor link を含むパス＝異種ネットワークを行き来するパス



特徴量の計算

- x_Φ(u, v): (u, v) 間の Meta Path Φ による特徴量
  - $\Phi = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \cdots \xrightarrow{R_{k-1}} T_k$
  - $$x(u,v) = I(u,T_1)I(v,T_k) \sum_{n_1 \in \{u\}, n_2 \in T_2, \cdots, n_k \in \{v\}} \prod_{i=1}^{k-1} p(n_i, n_{i+1}) I((n_i, n_{i+1}), R_i)$$
- (u, v) 間の Φ の形の **Meta Path の本数**
- p (**formation probability**; 将来リンクが張られる見込み確率) **で重み付け**

- 問題設定：
  multiple partial aligned networks の同時リンク予測
- **提案手法**：
  - PU (Positive Unlabeled) Link Prediction [13][28]
  - Meta Path ベースの特徴量 (inspired by [17] work)
    - Meta Path
    - Intra-network & Inter-network Social Meta Path
  - **特徴量の選別**
  - → Multi-Network Link Prediction Framework
- 実験

Meta Path Selection：有用な特徴量の選別

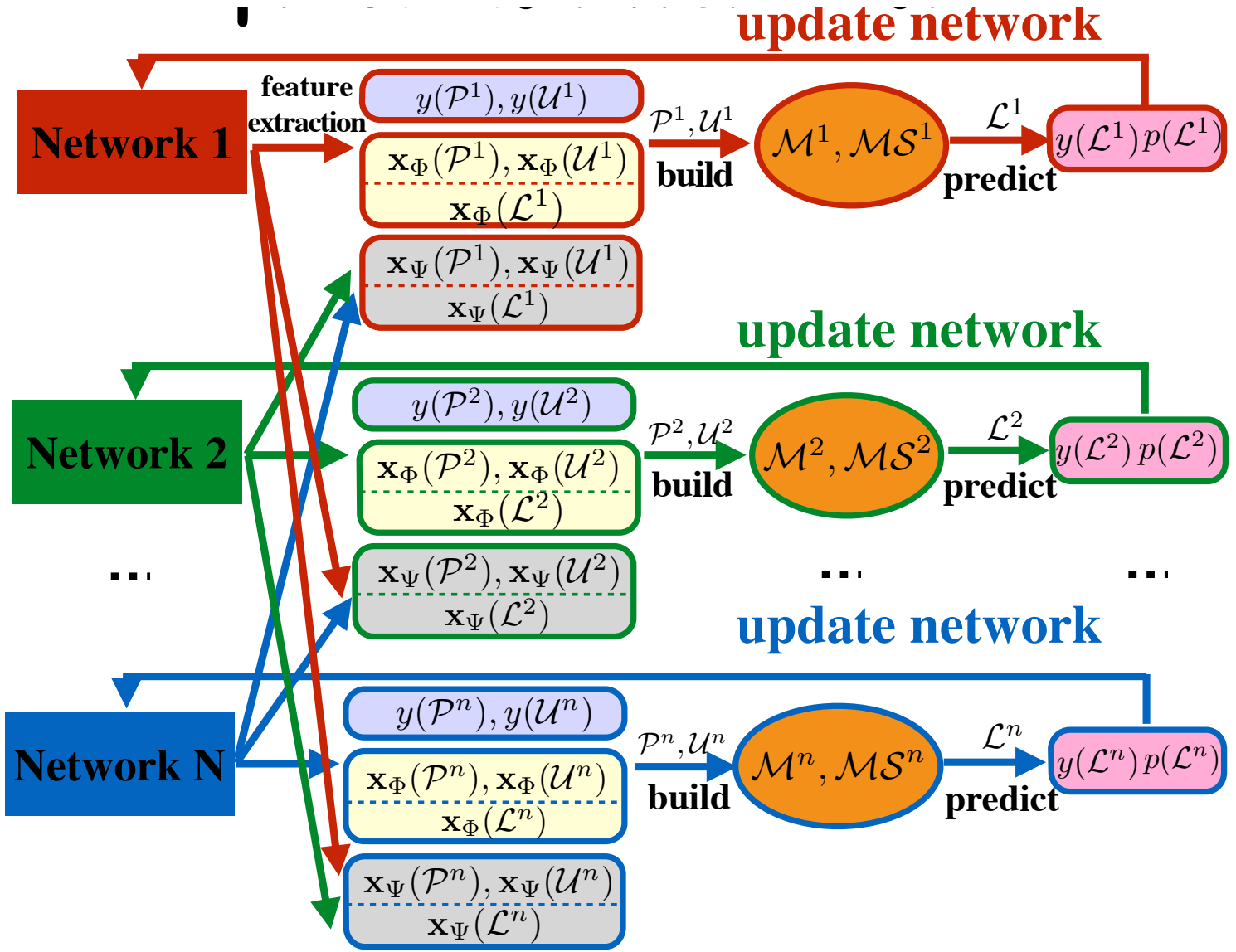
- $$mi(X_i) = \sum_x \sum_y P(X_i = x, Y = y) \log \frac{P(X_i = x, Y = y)}{P(X_i = x)P(Y = y)}$$

- X_i: Meta Path Φ_i に対応する特徴量を実現値とする確率変数
- Y: ラベル (将来リンクが生成されるかどうかを表す)
- **将来リンクが生成されるかどうかを表す確率変数**とのmi(**相互情報量**)が大きな特徴量のみ残す

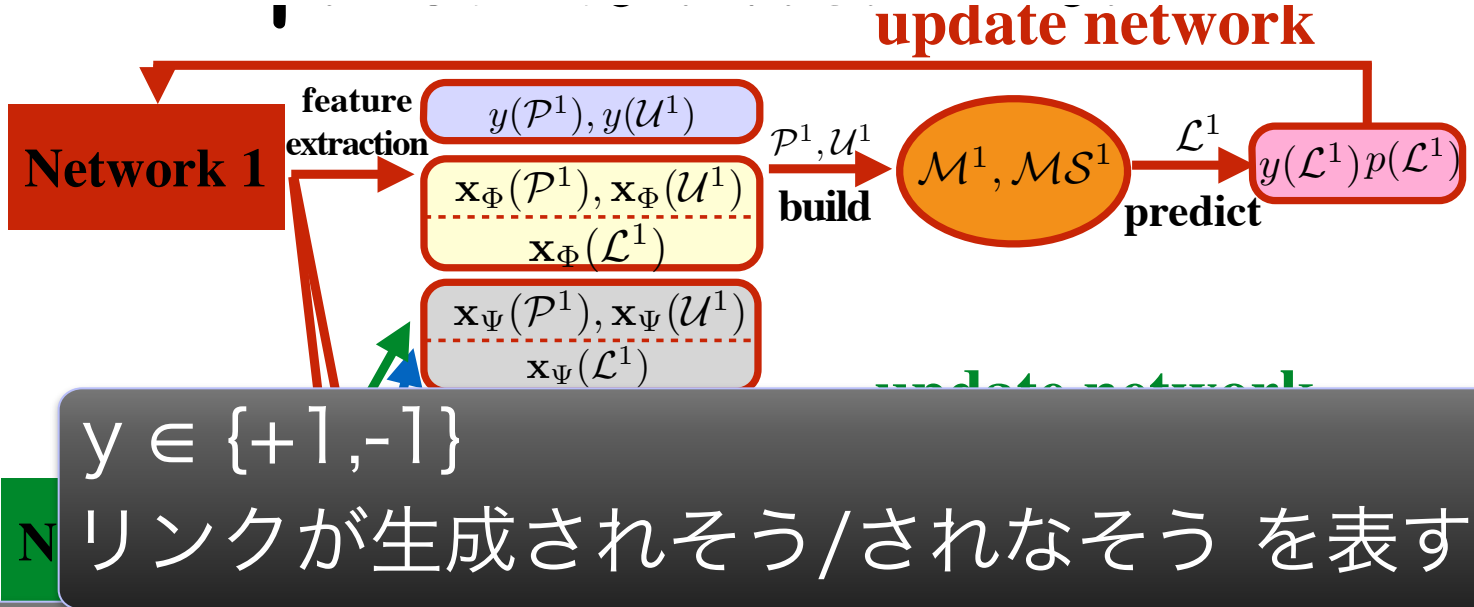
- → **network difference problem** (違う種類のネットワークから情報を転移すると性能が落ちることがあるよ問題) の解消

- 問題設定：
  multiple partial aligned networks の同時リンク予測
- **提案手法**：
  - PU (Positive Unlabeled) Link Prediction [13][28:2011]
  - Meta Path ベースの特徴量 (inspired by [17:2011] work)
    - Meta Path
    - Intra-network & Inter-network Social Meta Path
  - 特徴量の選別 (Meta Path Selection)
    - Network difference problem の解決
  - → **Multi-Network Link Prediction Framework**
- 実験

複数ネットワークの同時リンク予測
Multi-PU Link Prediction Framework



複数ネットワークの同時リンク予測
Multi-PU Link Prediction Framework



y ∈ {+1,-1}
リンクが生成されそう/されなそう を表すラベル

1. リンクにMeta Pathベースの特徴ベクトルを付与 (y利用)
2. PU でモデル p(y=+1|**x**) を作成
   / 利用する特徴量集合の更新(?)
3. モデルで y を更新 (argmax の計算の便宜上, 単独のネットワークで計算)

- 問題設定：
  multiple partial aligned networks の同時リンク予測
- 提案手法：
  - PU (Positive Unlabeled) Link Prediction [13][28]
  - Meta Path ベースの特徴量 (inspired by [17] work)
    - Meta Path
    - Intra-network & Inter-network Social Meta Path
  - 特徴量の選別 (Meta Path Selection)
    - Network difference problem の解決
  - → MLI (複数ネットワークの同時リンク予測)
- **実験**

実験：Dataset

- Foursquare and Twitter

Table 2: 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

実験：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

proposed

MLI (Multi-network Link Identifier)

LI (Link Identifier): predict links in each network independently

[25, 26]

- SCAN(Supervised Cross-Aligned-Network link prediction): supervised link prediction, no meta path selection,

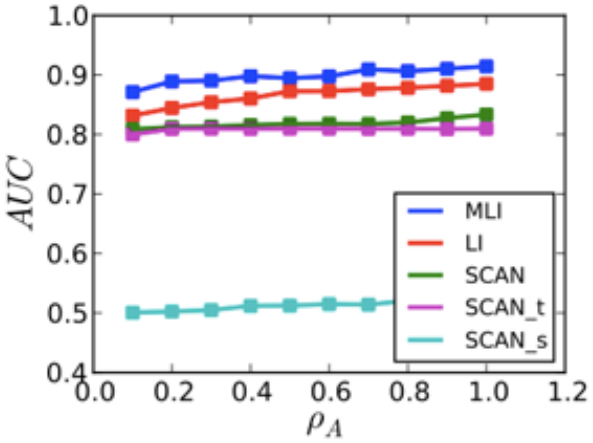
SCAN_s (SCAN with source network): features are extracted based on inter-network meta paths

SCAN_t (SCAN with target network): features are extracted based on intra-network meta paths

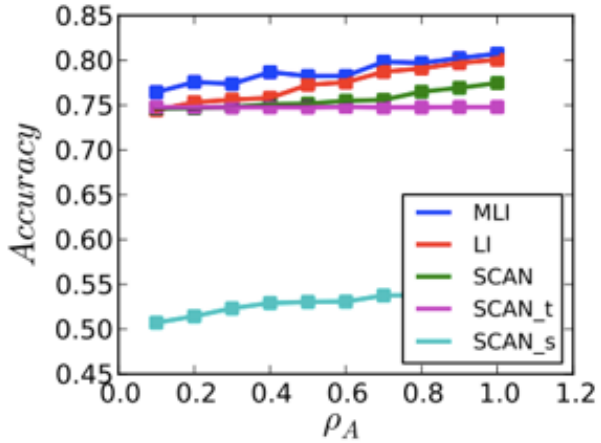
- Evaluation Metrics
  - AUC, Accuracy, F1

実験：結果

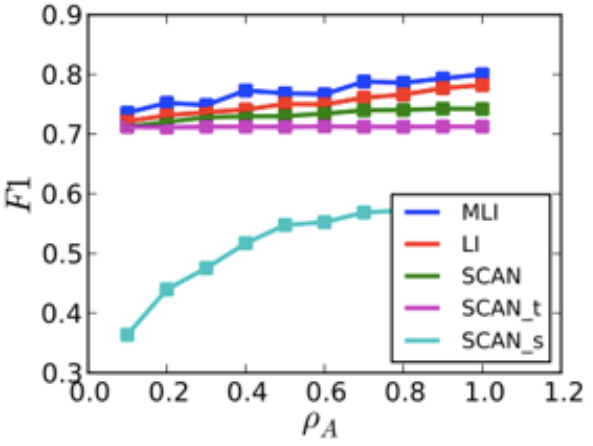
- PU link prediction & meta path & meta path selection (3種類の方法の合わせ技) で,
  既存手法 [25, 26] よりスコア向上
- 複数ネットワークを連携させて予測にするとスコアは更に向上
- anchor link (複数SNSを同時に利用しているユーザ) の割合 (横軸) が大きいとスコアは向上



(a) Foursquare-AUC



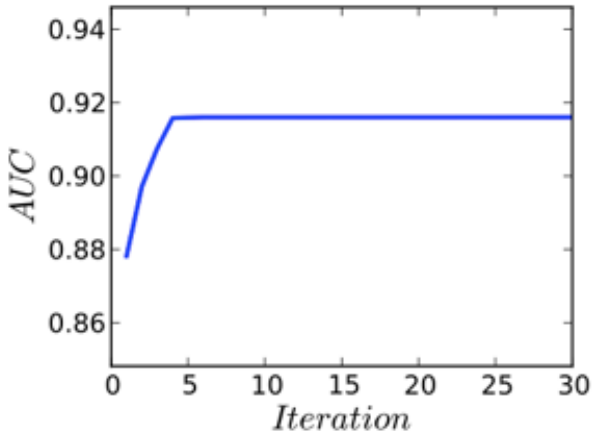
(b) Foursquare-Acc.



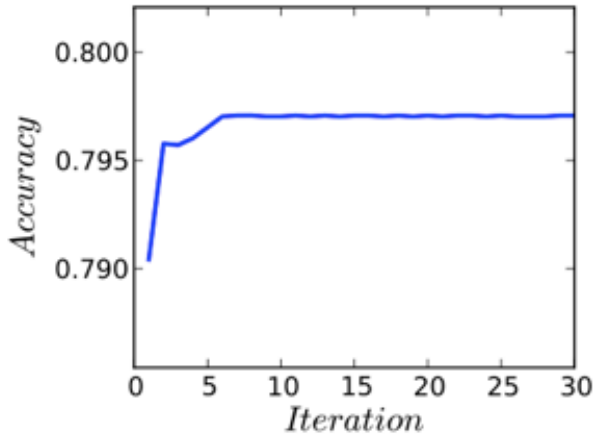
(c) Foursquare-F1

実験：結果

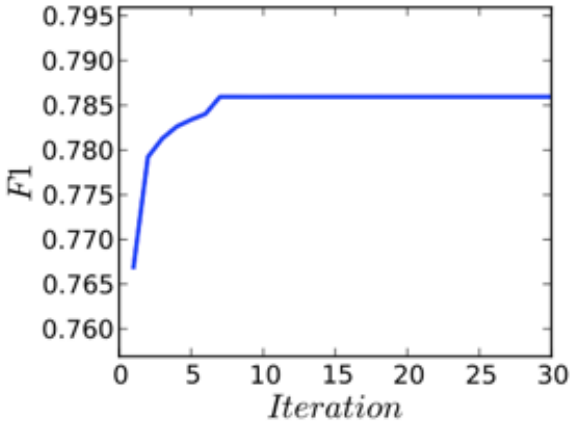
- Multi-PU Link Prediction Framework のループは10回程度で収束



(a) Foursquare-AUC



(b) Foursquare-Acc.



(c) Foursquare-F1

まとめ

- 問題設定：
  multiple partial aligned networks の同時リンク予測
- 提案手法：
  - PU (**Positive Unlabeled**) Link Prediction [13][28]
  - **Meta Path** ベースの特徴量 (inspired by [17] work)
    - Meta Path
    - Intra-network & Inter-network Social Meta Path
  - 特徴量の選別 (**Meta Path Selection**)
    - Network difference problem の解決
  - → **MLI** (複数ネットワークの同時リンク予測)
- 実験結果：
  - 3種類のテクニックの合わせ技でスコア向上, 複数ネットワークの連携で更に向上
  - anchor link の割合が大きいとスコア向上
  - ループは10回程度で収束