Link Prediction across Heterogeneous Social Networks: A Survey

Presenter: Jiawei Zhang

March 18, 2014

Outline

- Background Knowledge
- Problem Formulation
- Link Prediction in Homogeneous Networks
- Link Prediction in Heterogeneous Networks
- Link Prediction across Aligned Heterogeneous Networks
- Future Works
- Summary

Outline

- **Background Knowledge**
- Problem Formulation
- Link Prediction in Homogeneous Networks
- Link Prediction in Heterogeneous Networks
- Link Prediction across Aligned Heterogeneous Networks
- Future Works
- Summary

Background Knowledge

- **Social Networks**
  - **Definition:** A social network is a social structure made up of a set of social actors and a set of ties between these actors.
  - **Example:**

Background Knowledge

- **Social Networks**
  - **Representation:**

$$G = (V, E)$$

where $\boldsymbol{G}$ is the social network, $\boldsymbol{V}$ is the set of actors and $\boldsymbol{E}$ is the set of ties among actors.

Background Knowledge

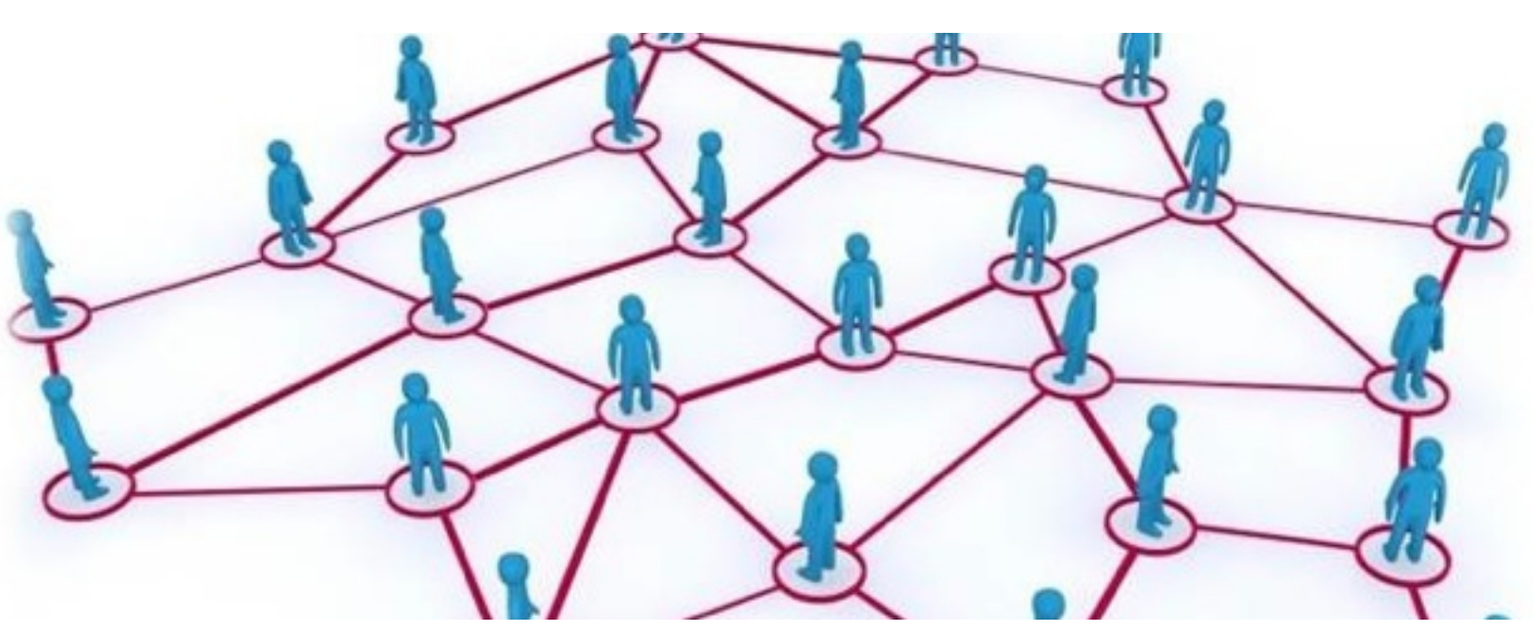
- **Homogeneous Social Networks**
  - **Definition:**

$$G = (V, E)$$

If $\boldsymbol{V}$ contains one single type nodes and $\boldsymbol{E}$ contains one single type of links, then $\boldsymbol{G}$ is a homogeneous social network.

Background Knowledge

- **Homogeneous Social Networks**
  - **Example:**



where $\boldsymbol{V}$ is the set of users and $\boldsymbol{E}$ is the set of social links among users.

Background Knowledge

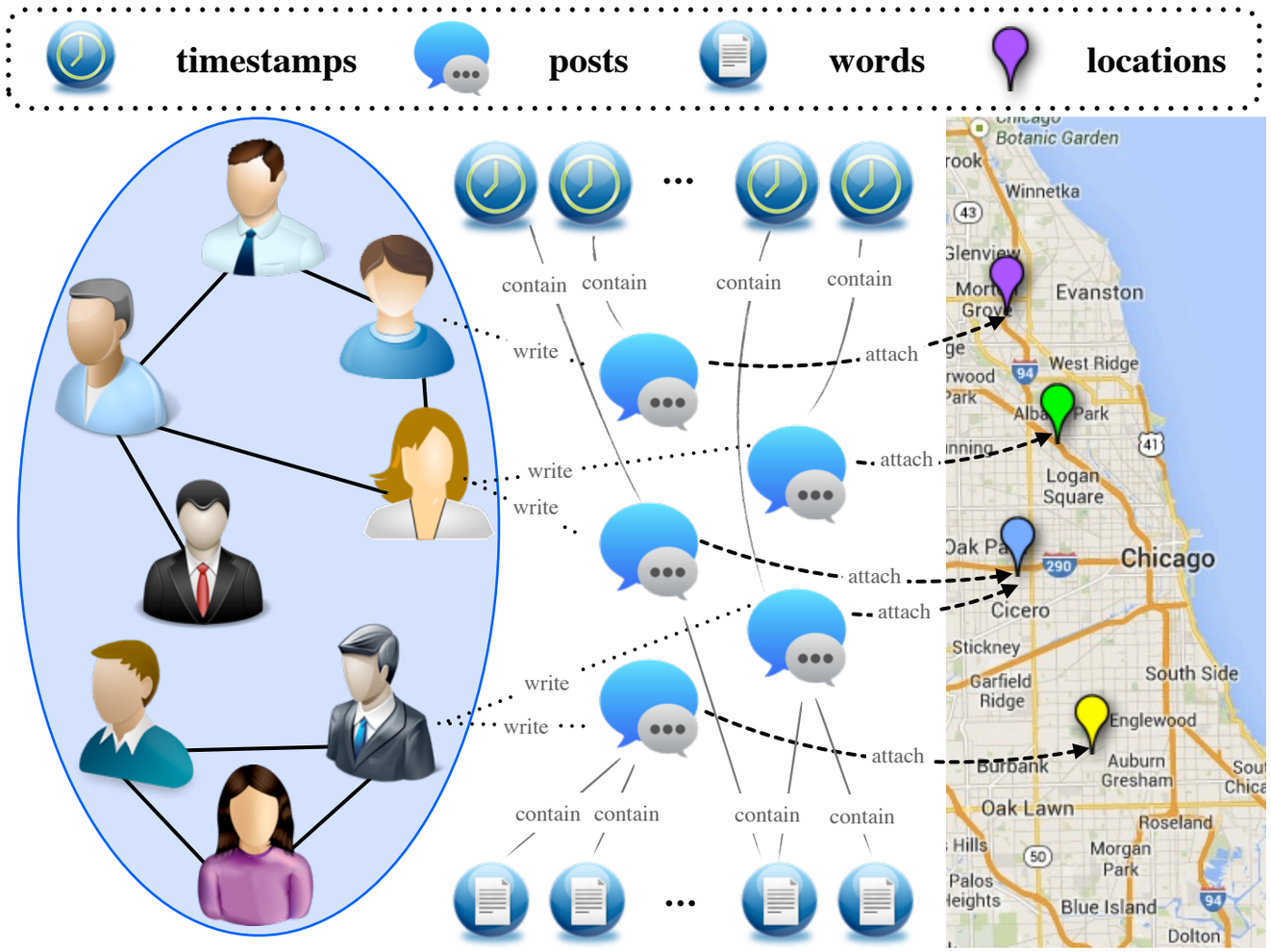
- **Heterogeneous Social Networks**
  - **Definition:**

$$G = (V, E)$$

where $V = \bigcup_i V_i$ is the sets of various kinds of nodes in the network and V_i is the i_{th} kind of nodes in $\boldsymbol{G}$; $E = \bigcup_j E_j$ is the sets of various types of links in the network and E_j is the j_{th} kind of links in $\boldsymbol{G}$.

Background Knowledge

- **Heterogeneous Social Networks**
  - **Example:**




SMS
SMS

foursquare

twitter

Temporal Activities

8 AM 12 PM 4 PM 8 PM 11 PM

User Accounts

Locations

Park West
Sheffield Neighbors
Lincoln Park
Near North Side
Near West Side
Loop
West Loop
Chicago
Near South Side
South Loop
W 31st St
E 31st St
W 35th St

Tips

Temporal Activities

8 AM 12 PM 4 PM 8 PM 11 PM

User Accounts

Locations

locate

Park West
Sheffield Neighbors
Lincoln Park
Near North Side
Near West Side
Loop
West Loop
Chicago
Near South Side
South Loop
W 31st St
E 31st St
W 35th St

locate

Tweets

Background Knowledge

- **Multi Aligned Heterogeneous Social Networks**
  - **Definition:**

$$\mathcal{G} = (G_{set}, A_{set})$$

where $G_{set} = \{G^{(1)}, G^{(2)}, \cdots, G^{(|G_{set}|)}\}$ is the set of $|G_{set}|$ different heterogeneous networks;

$A_{set} = \{A^{(1,2)}, A^{(1,3)}, \cdots, A^{(|G_{set}|,|G_{set}|-1)}\}$ is the set of ***anchor links*** among networks.

foursquare

Anchor Links

twitter

$A^{(1,2)}$

$G^{(1)}$

$G^{(2)}$

Temporal Activities

8 AM 12 PM 4 PM 8 PM 11 PM

User Accounts

Locations

Tips

Temporal Activities

8 AM 12 PM 4 PM 8 PM 11 PM

User Accounts

Locations

locate

locate

Tweets

Background Knowledge

- Homogeneous Social Networks
  - Definition & Example
- Heterogeneous Social Networks
  - Definition & Example
- Multiple Aligned Heterogeneous Social Networks
  - Definition & Example

Outline

- Background Knowledge
- **Problem Formulation**
- Link Prediction in Homogeneous Networks
- Link Prediction in Heterogeneous Networks
- Link Prediction across Aligned Heterogeneous Networks
- Future Works
- Summary

Problem Formulation

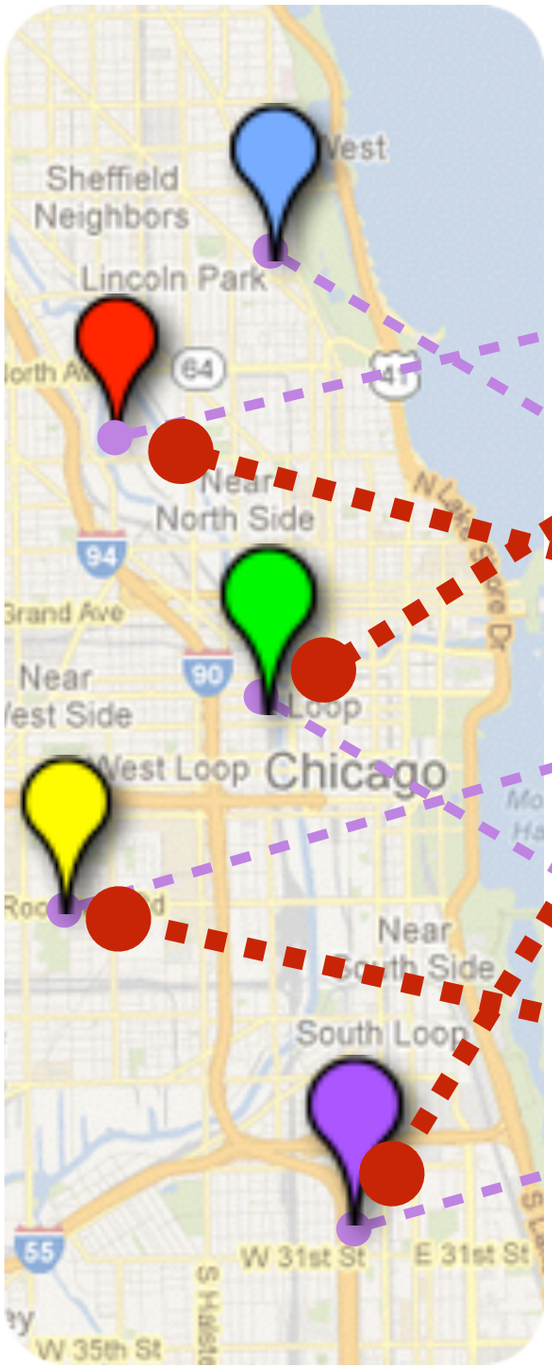
- **Link Prediction Problem**
  - **Definition:**

    Based on a snapshot of social network, e.g., $\boldsymbol{G}$, predicting the set of potential links to be formed in the future is formally defined as the **Link Prediction Problem.**

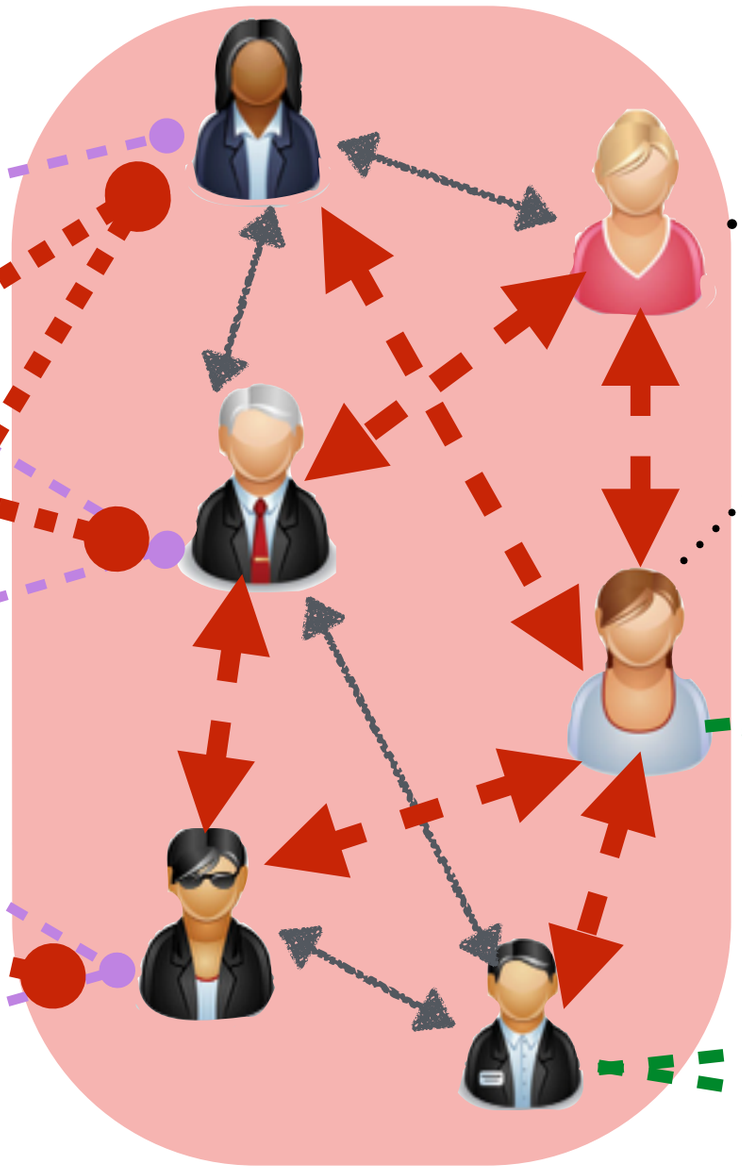
social links prediction

location links prediction

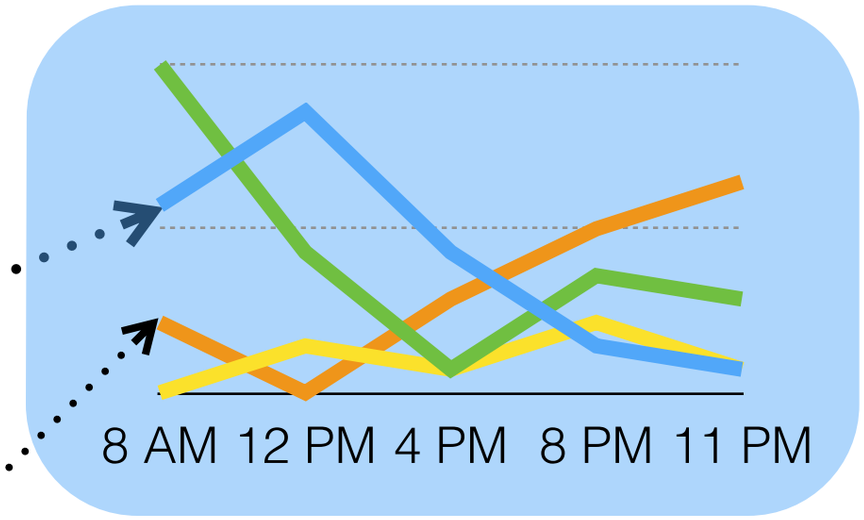
Locations



Social Links



Temporal Activities



Contents: Tweets

Problem Formulation

- Link Prediction Problem
  - Definition
  - Example

Outline

- Background Knowledge
- Problem Formulation
- **Link Prediction in Homogeneous Networks**
- Link Prediction in Heterogeneous Networks
- Link Prediction across Aligned Heterogeneous Networks
- Future Works
- Summary

Reminder

- **Homogeneous Social Networks**
  - **Definition:**

$$G = (V, E)$$

If $\boldsymbol{V}$ contains one single type nodes and $\boldsymbol{E}$ contains one single type of links, then $\boldsymbol{G}$ is a homogeneous social network.

Link Prediction for Homogeneous Social Networks

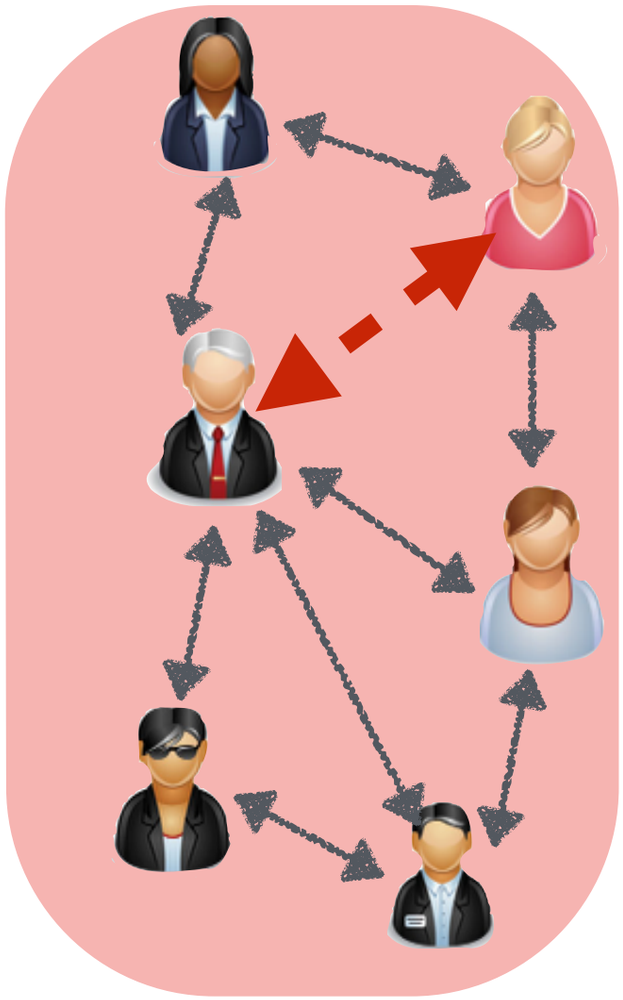
- **Unsupervised Link Predicators**
  - Local Neighbor based Link Predicators
  - Path based Link Predicators
- **Random Walk based Link Predicators**
- **…**

Link Prediction for Homogeneous Social Networks

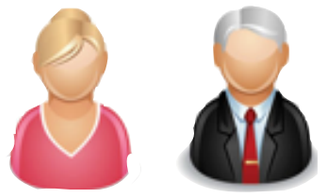
- **Unsupervised Link Predicators:**
  - measuring the closeness among nodes
  - assuming that close nodes are more likely to be connected

Link Prediction for Homogeneous Social Networks

- **Local Neighbor based Link Predicators:**
  - Preferential Attachment $PA(u,v) = |\Gamma(u)|\,|\Gamma(v)|$

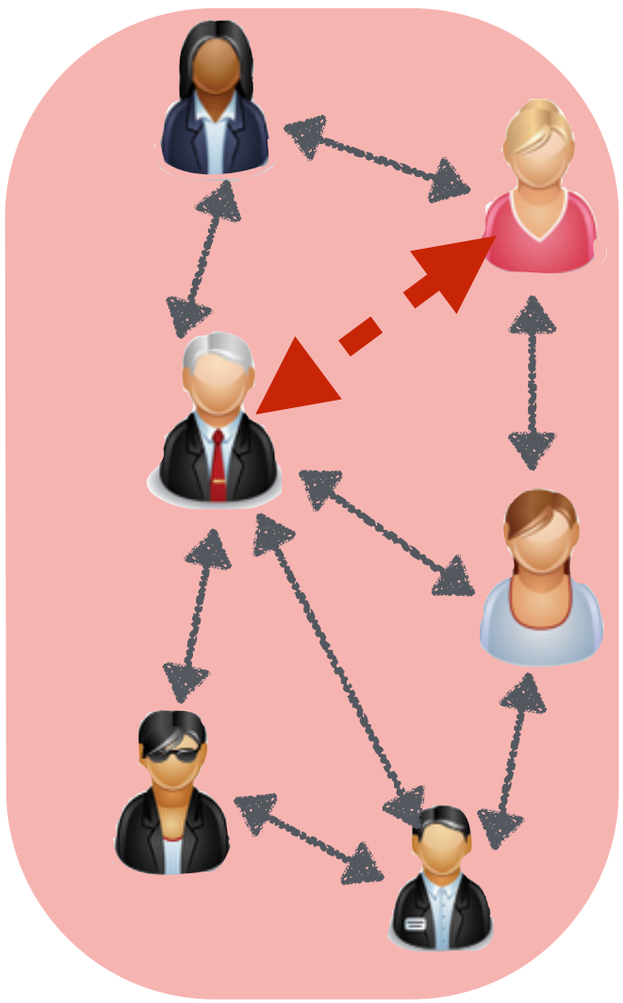


user u neighbors $\Gamma(u)$

$PA($  $) = 2 \times 4 = 8$

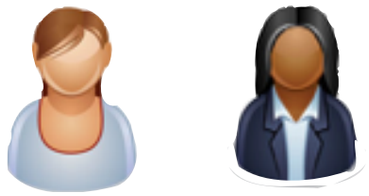
Link Prediction for Homogeneous Social Networks

- **Local Neighbor based Link Predicators:**
  - Common Neighbor $CN(u,v) = |\Gamma(u) \cap \Gamma(v)|$



user

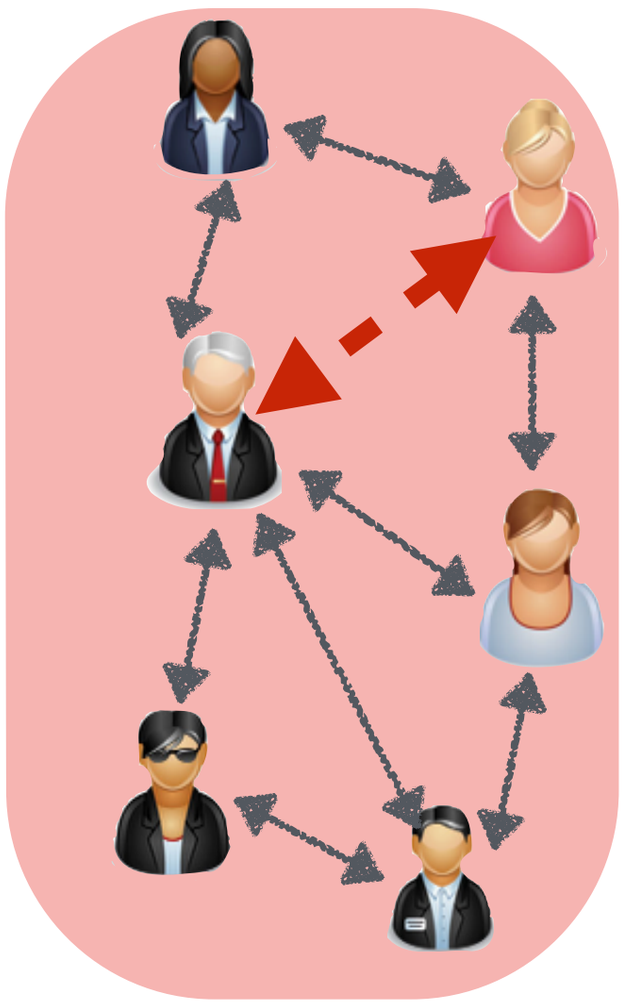
neighbors



$CN($  $) = |\{$  $\}| = 2$

Link Prediction for Homogeneous Social Networks

- **Local Neighbor based Link Predicators:**
  - Jaccard's Coefficient $JC(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$

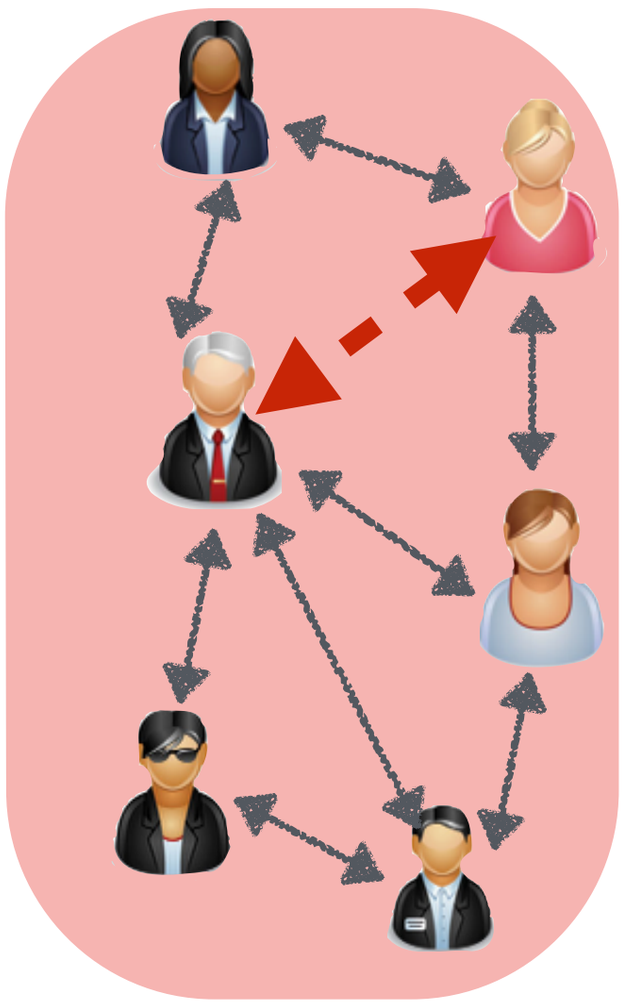


Link Prediction for Homogeneous Social Networks

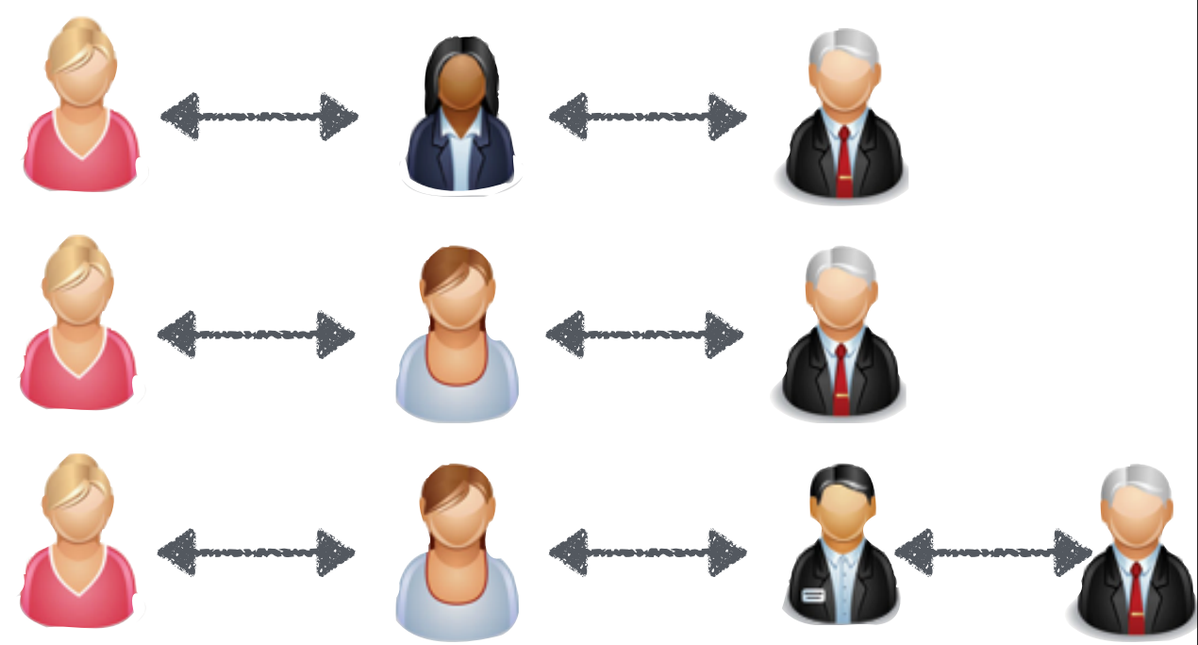
- **Local Neighbor based Link Predicators:**
  - Adamic/Adar Index $AA(u,v) = \sum_{w \in (\Gamma(u) \cap \Gamma(v))} \frac{1}{\log|\Gamma(w)|}$
  - Resource Allocation $RA(u,v) = \sum_{w \in (\Gamma(u) \cap \Gamma(v))} \frac{1}{|\Gamma(w)|}$
  - ...

Link Prediction for Homogeneous Social Networks

- **Path based Link Predicators:**
  - Shortest Path $SP(u,v) = \min\{|p_{u \rightsquigarrow v}|\}$



$p_{u \rightsquigarrow v}$



.....

$SP($  $) = 2$

Link Prediction for Homogeneous Social Networks

- **Path based Link Predicators:**
  - Katz $Katz(u,v) = \sum_{l=1}^{\infty} \beta^l \left| p_{u \rightsquigarrow v}^l \right|,$
  - …
- **Random Walk based Link Predicators:**
  - Hitting Time
  - Commute Time
  - ……

Link Prediction for Homogeneous Social Networks

- Unsupervised Link Predicators
  - Local Neighbor based Link Predicator
    - Preferential Attachment, Common Neighbor, Jaccard's Coefficient,
    - Adamic/Adar , Resource Allocation
  - Path based Link Predicator
    - Shortest Path, Katz
- Random Walk based Link Predicators
  - Hitting Time, Commute Time

Outline

- Background Knowledge
- Problem Formulation
- Link Prediction in Homogeneous Networks
- **Link Prediction in Heterogeneous Networks**
- Link Prediction across Aligned Heterogeneous Networks
- Future Works
- Summary

Reminder

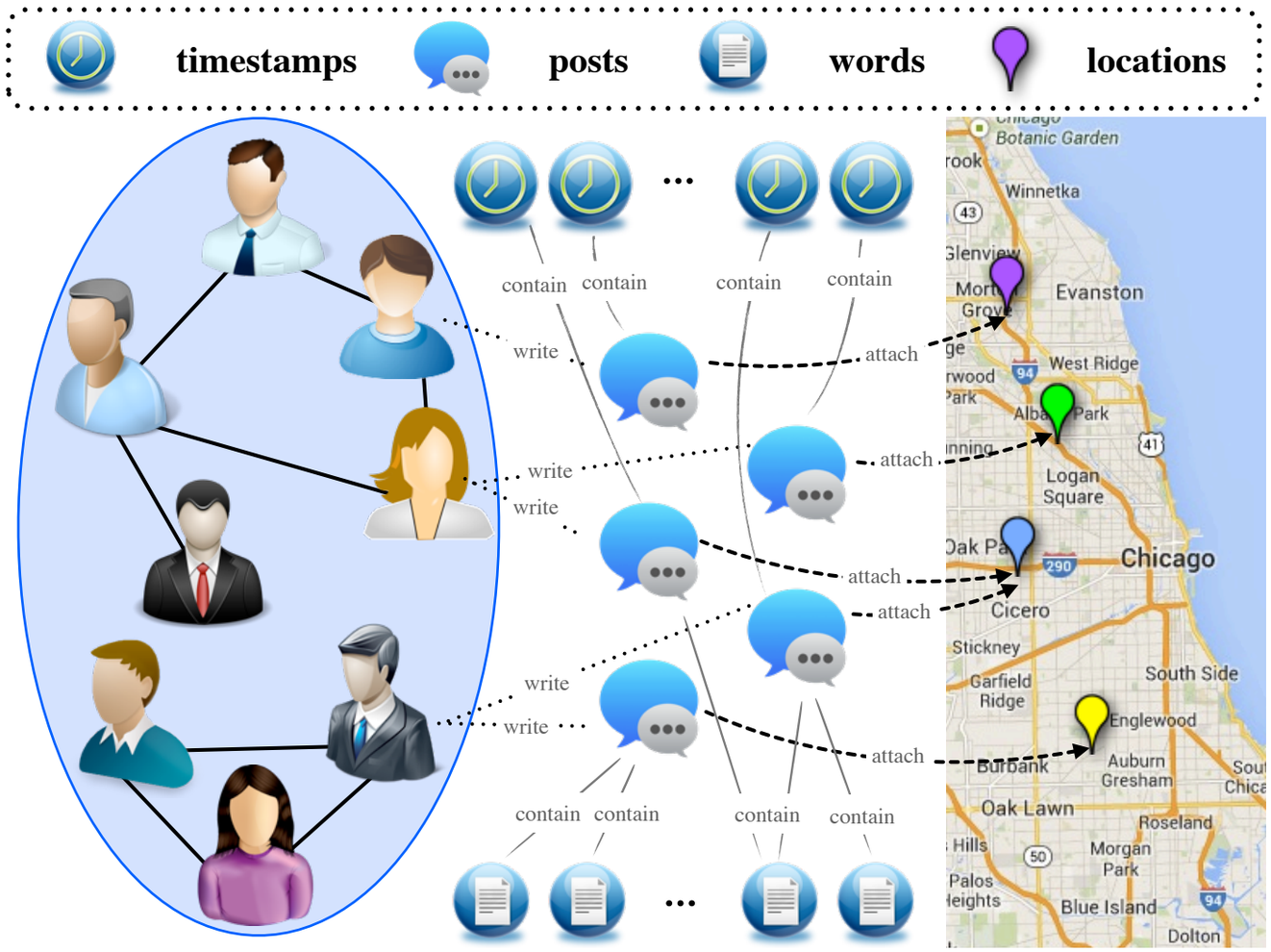
- **Heterogeneous Social Networks**
  - **Definition:**

$$G = (V, E)$$

where $V = \bigcup_i V_i$ is the sets of various kinds of nodes in the network and V_i is the i_{th} kind of nodes in $\boldsymbol{G}$; $E = \bigcup_j E_j$ is the sets of various types of links in the network and E_j is the j_{th} kind of nodes in $\boldsymbol{G}$.

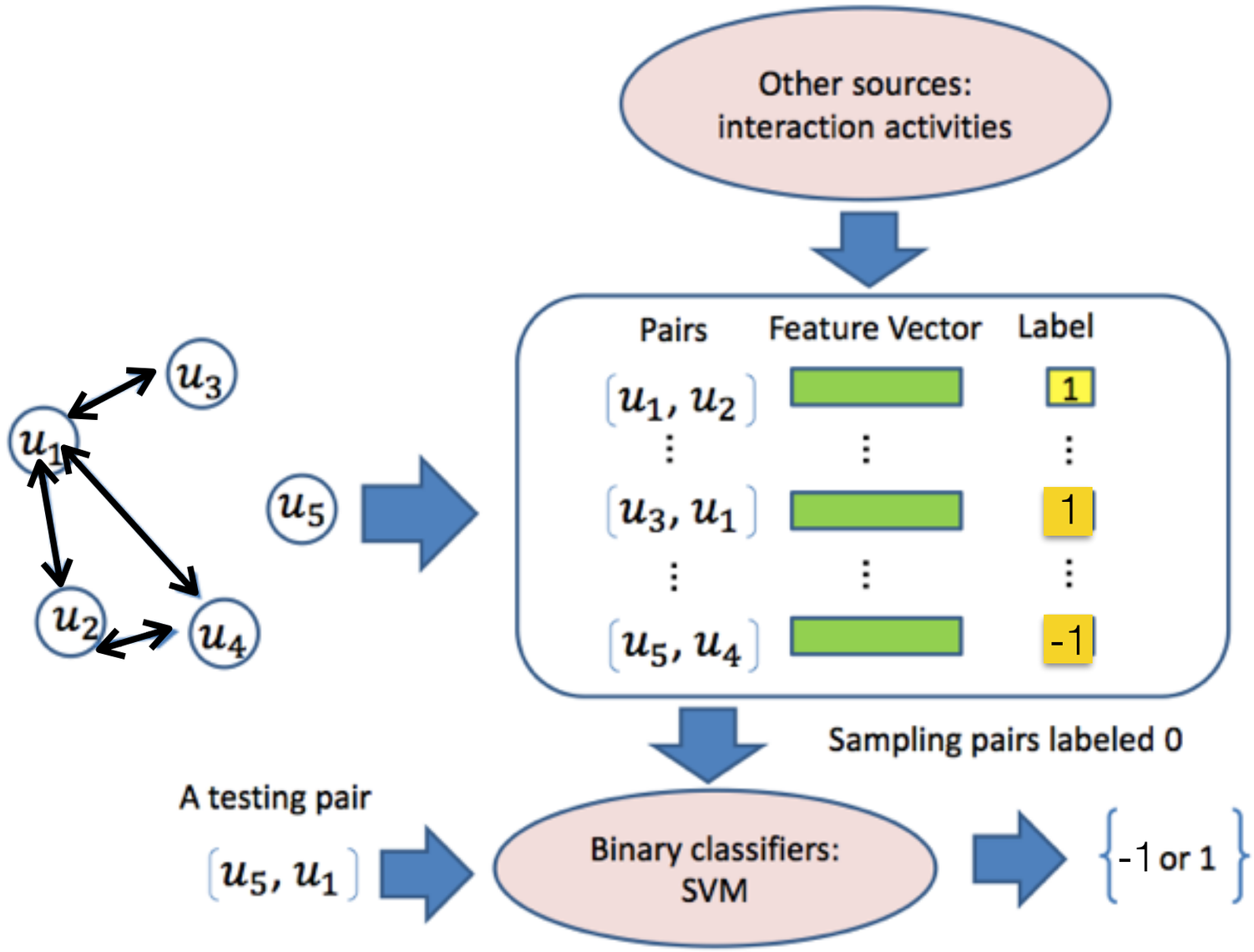
Reminder

- **Heterogeneous Social Networks**
  - **Example:**



Link Prediction in Heterogeneous Networks

- **Supervised Link Prediction**

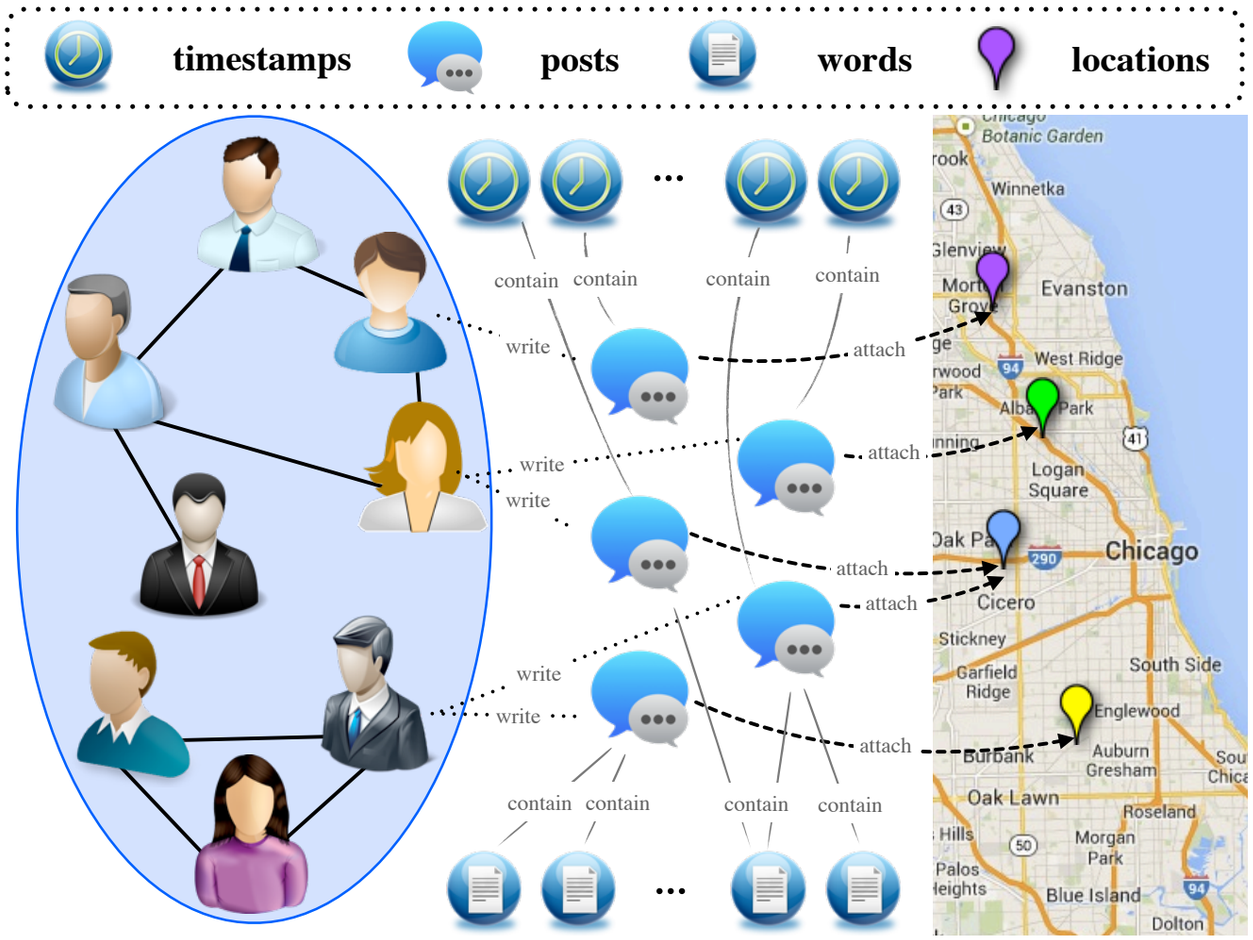


Link Prediction in Heterogeneous Networks

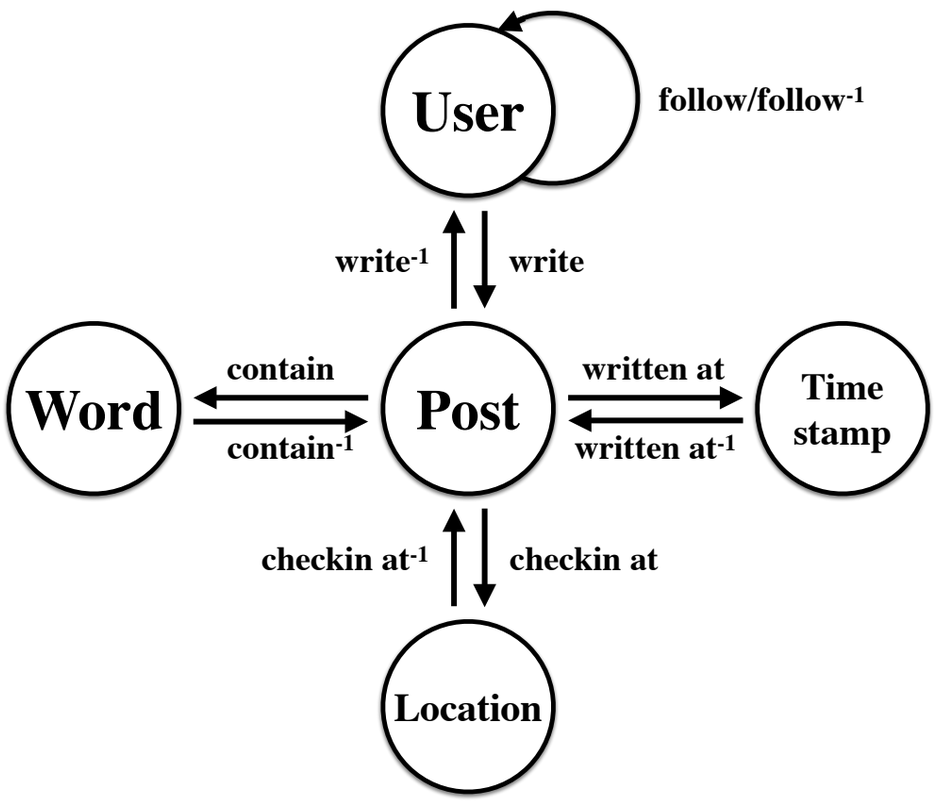
- **Supervised Link Prediction**
  - Feature Extraction
    - Local Neighbor, Path, Random Walk based Closeness Measures
    - Meta Path based Features
  - Classification Algorithm
    - Decision Tree, Naive Bayes, ….
    - SVM

Link Prediction in Heterogeneous Networks

- **Supervised Link Prediction**
  - Meta Path based Features for Heterogeneous Networks

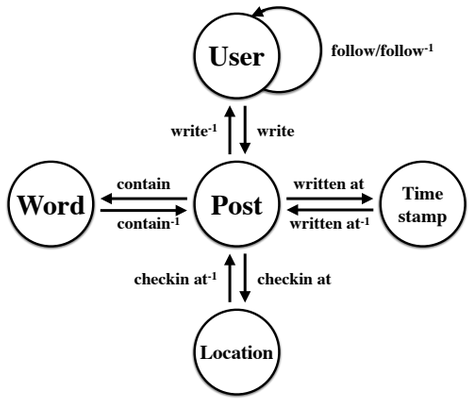


G



S_G

Link Prediction in Heterogeneous Networks



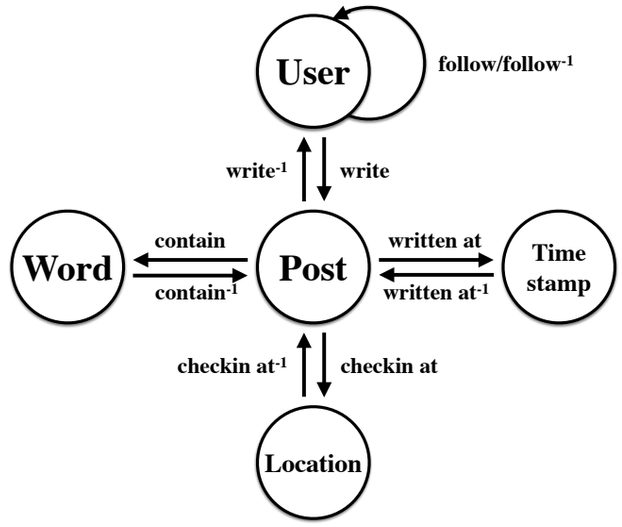
- **Supervised Link Prediction**
  - Meta Path based Features for Heterogeneous Networks
    - Schema **Definition**:

      The *schema* of network G is defined to be $S_G = (T, R)$, where T, R are the sets of node types and link types in G.

    - Meta Path **Definition**:

      Based on the given the network schema, $S_G = (T, R)$, $\Phi = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \cdots \xrightarrow{R_{k-1}} T_k$ is defined to be the *meta path* of network G, where $T_i \in T, i \in \{1, 2, \cdots, k\}$ and $R_i \in R, i \in \{1, 2, \cdots, k-1\}$

Link Prediction in Heterog Networks



- **Supervised Link Prediction**

  - *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 $\Phi_2(U, U)$.

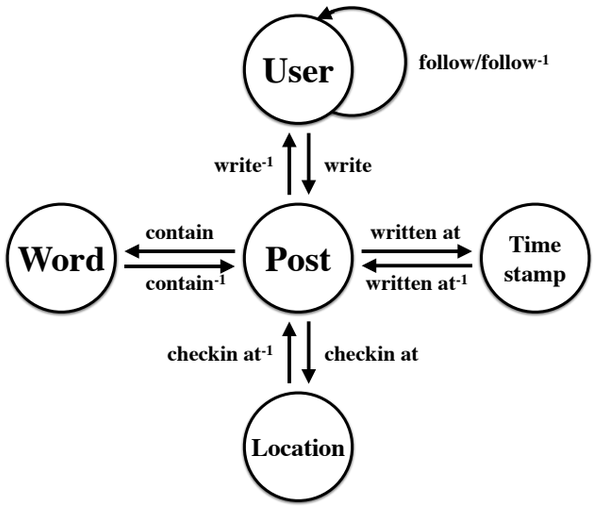
  - *ID 3. Common In Neighbor*: User $\xrightarrow{follow^{-1}}$ User $\xrightarrow{follow}$ User, whose notation is "$U \leftarrow U \to U$" or $\Phi_3(U, U)$.

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

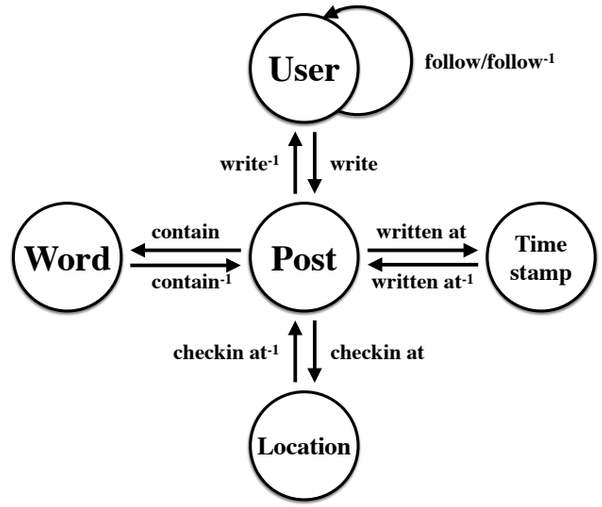
Link Prediction in Heterogeneous Networks



- **Supervised Link Prediction**
  - Meta Path based Features
    - **Definition:**

      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 meta path instances between u and v in the network.

Link Prediction in Heterogeneous Networks



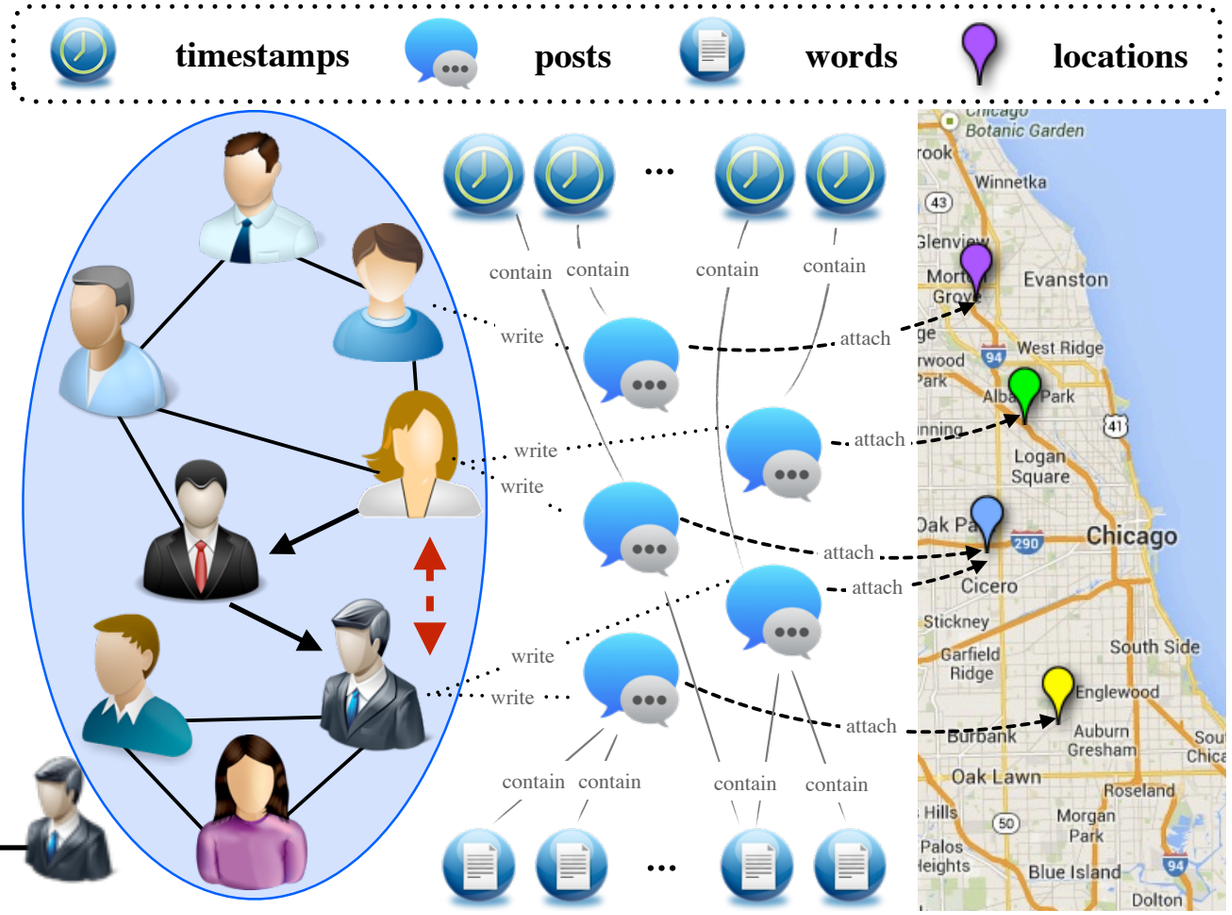
- **Supervised Link Prediction**
  - Meta Path based Features
    - Example:

meta path: $\text{“}U \rightarrow U \rightarrow U\text{”}$

path instance:

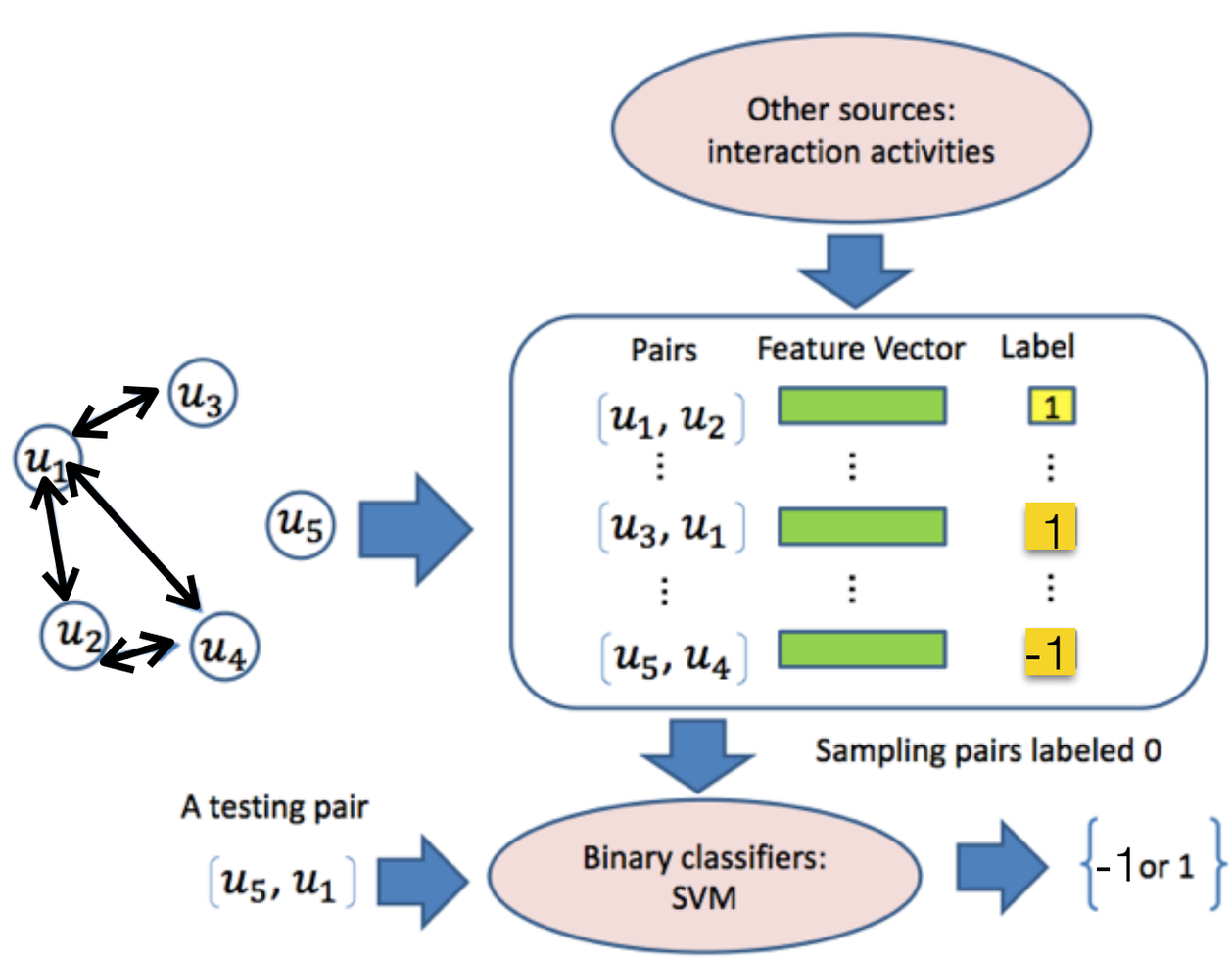
meta path: $\text{“}U \rightarrow P \rightarrow L \leftarrow P \leftarrow U\text{”}$

path instance:



Link Prediction in Heterogeneous Networks

- **Supervised Link Prediction**
  - Classification Algorithms
  - SVM



Link Prediction in Heterogeneous Networks

- Supervised Link Prediction Method
  - Feature Extraction
    - Meta Path based Features
  - Classification
    - SVM

Outline

- Background Knowledge
- Problem Formulation
- Link Prediction in Homogeneous Networks
- Link Prediction in Heterogeneous Networks
- **Link Prediction across Aligned Heterogeneous Networks**
- Future Works
- Summary

Reminder

- **Multi Aligned Heterogeneous Social Networks**
  - **Definition:**

$$\mathcal{G} = (G_{set}, A_{set})$$

where $G_{set} = \{G^{(1)}, G^{(2)}, \cdots, G^{(|G_{set}|)}\}$ is the set of $|G_{set}|$ different heterogeneous networks;

$A_{set} = \{A^{(1,2)}, A^{(1,3)}, \cdots, A^{(|G_{set}|,|G_{set}|-1)}\}$ is the set of ***anchor links*** among networks.

foursquare

Anchor Links

$A^{(1,2)}$

twitter

$G^{(1)}$

$G^{(2)}$

Temporal Activities

8 AM 12 PM 4 PM 8 PM 11 PM

User Accounts

Locations

Tips

Temporal Activities

8 AM 12 PM 4 PM 8 PM 11 PM

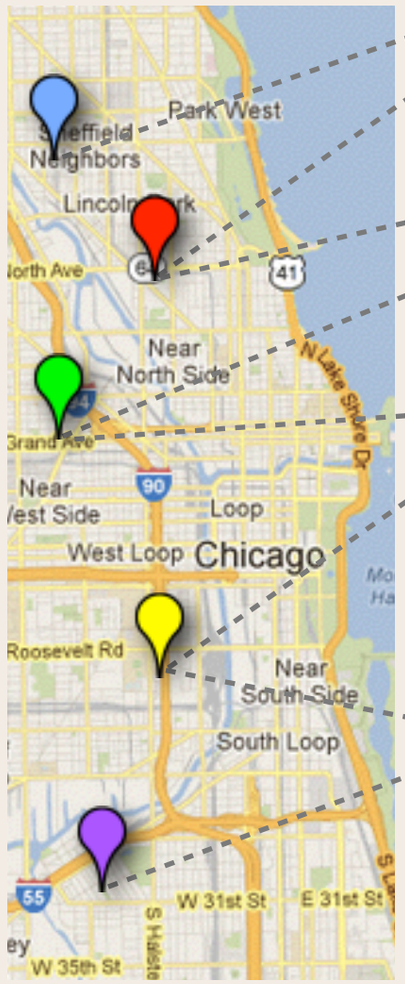
User Accounts

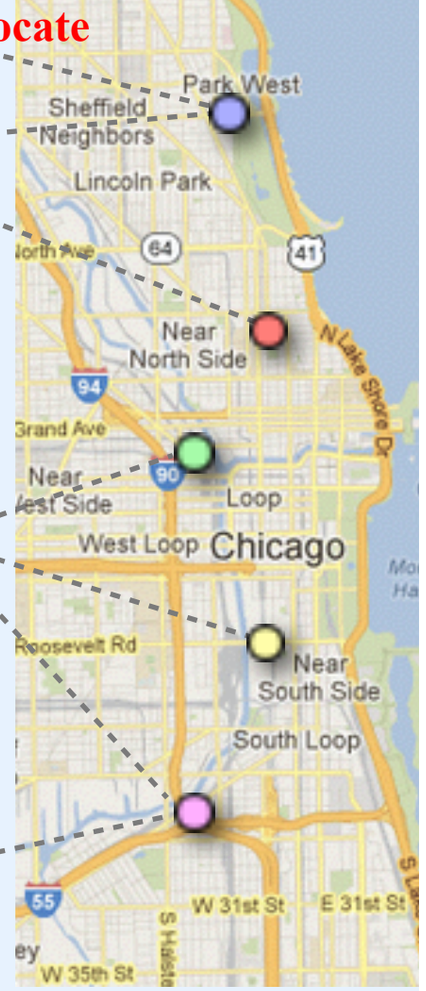
Locations

locate

locate

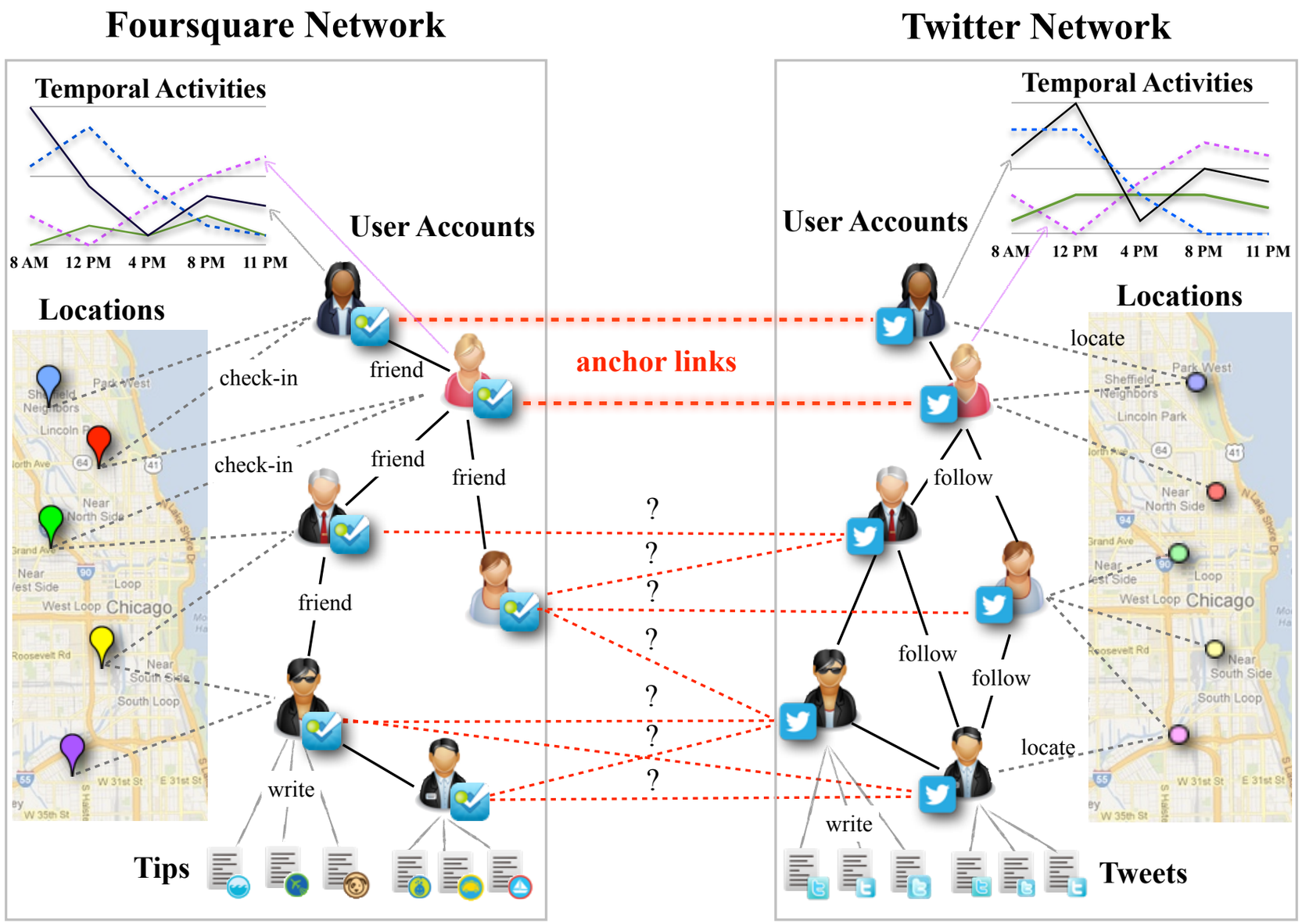
Tweets





Link Prediction across Aligned Heterogeneous Networks

- Anchor Link Prediction



Link Prediction across Aligned Heterogeneous Networks

- Anchor Link Prediction
  - Supervised Link Prediction Methods
    - Feature Extraction
    - Classification Algorithm

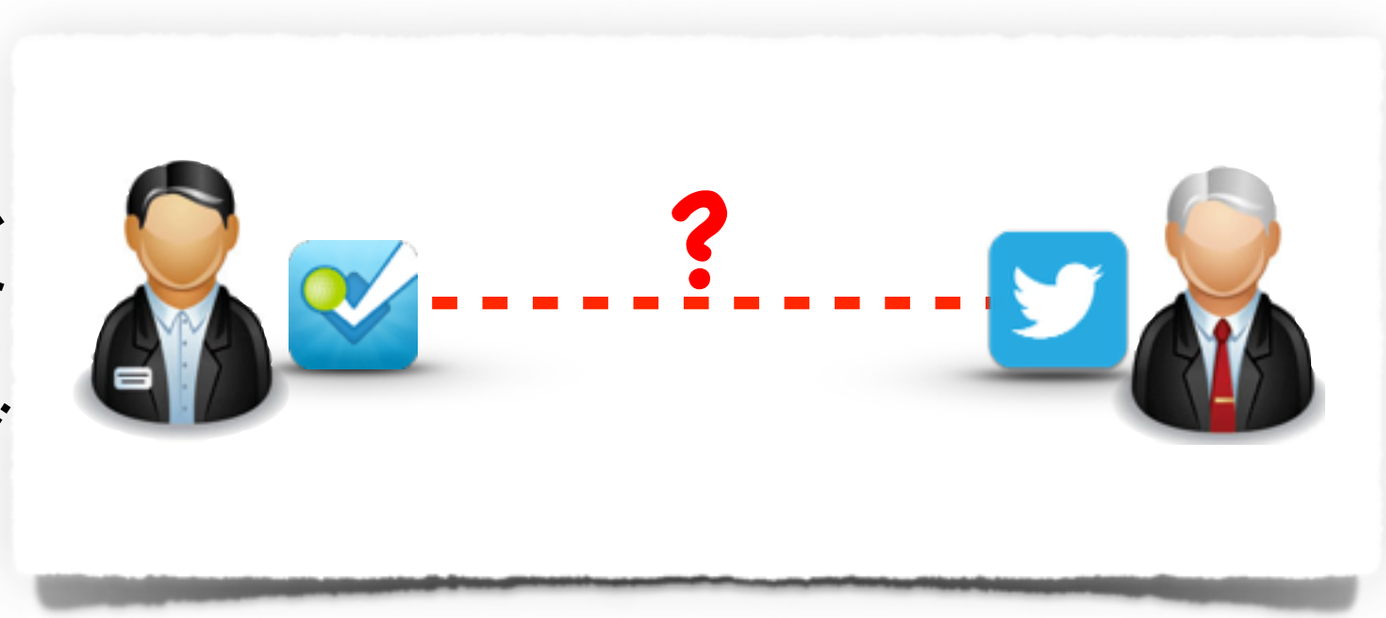
Link Prediction across Aligned Heterogeneous Networks

Social

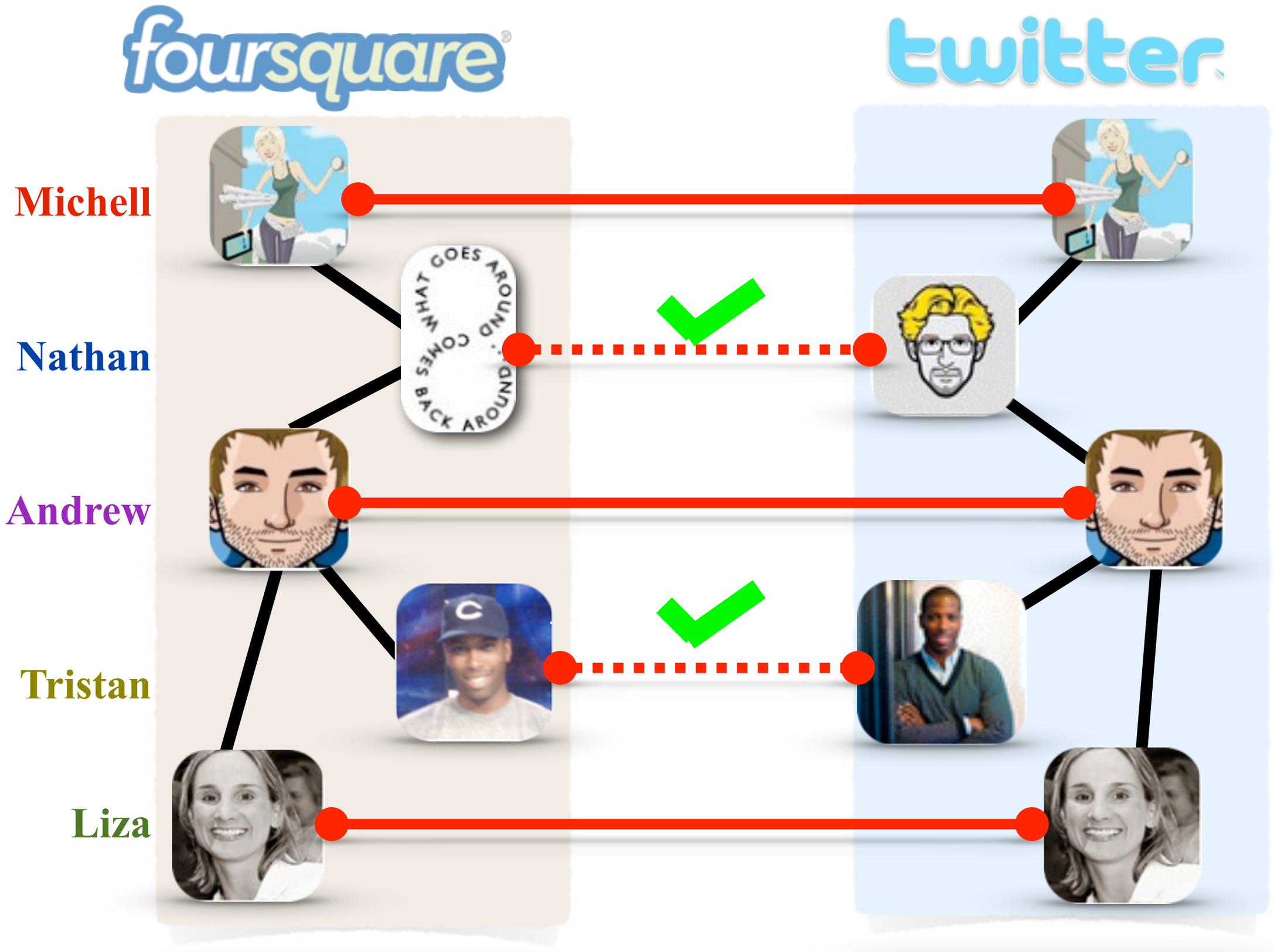
Spatial

Temporal

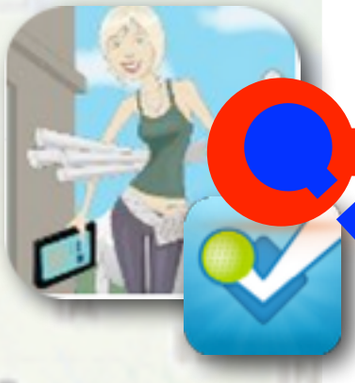
Content

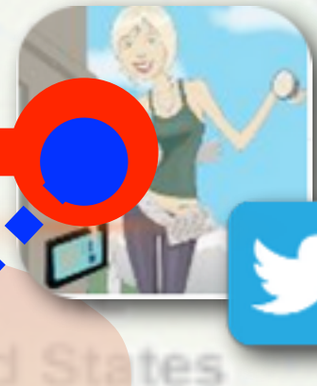


Social



foursquare





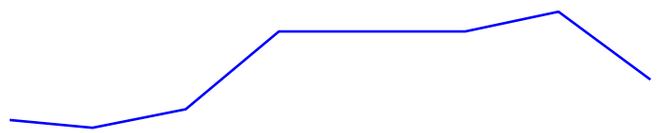
twitter

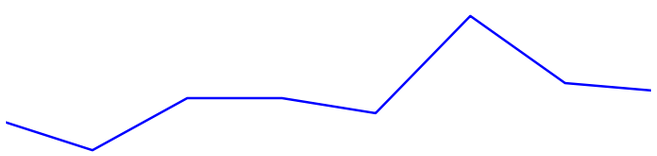
Temporal

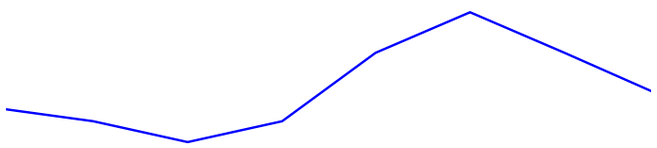
User Activities

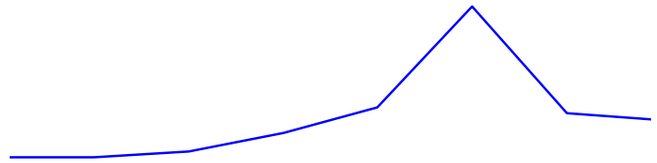
foursquare

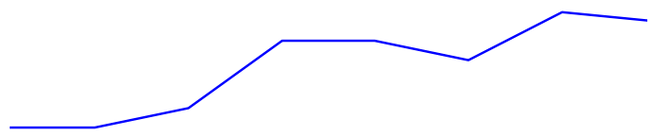
twitter

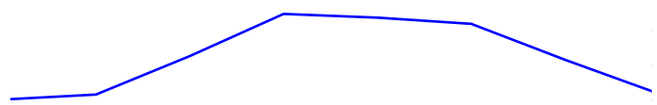


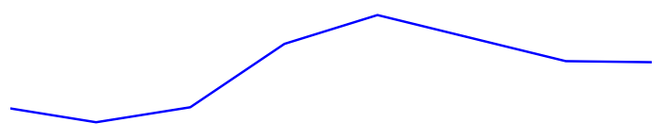


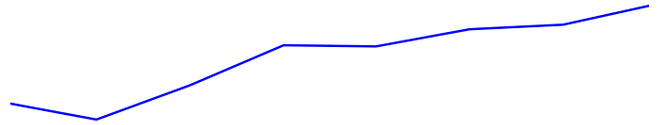


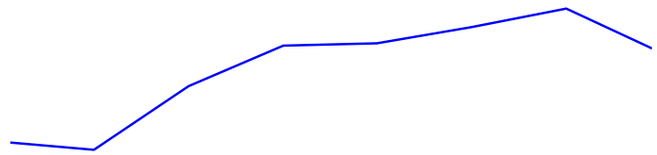


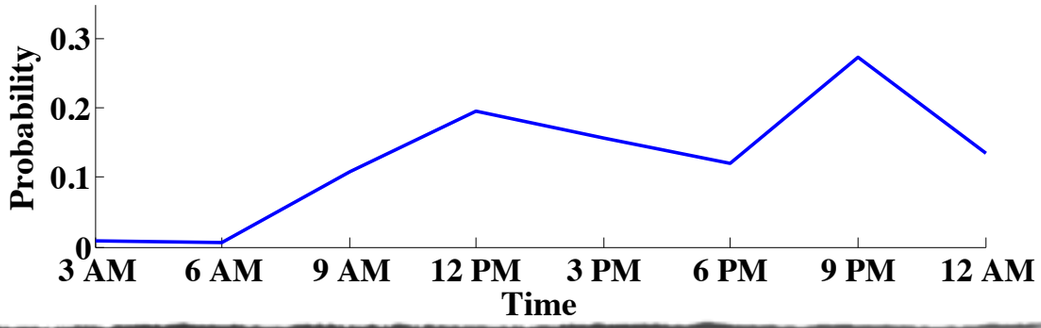












Content

Choice of Words

art (65,2), style (16,3)
audit (3,2), grill (19,2)

happy (27,5), enjoy (9,4)
week (18,4), shows (6,6)

awsm (2,3), kids (20,3)
red (61,3), open (11,4)

ask (6,5), coffee (8,3)
mochi (1,3), hangout (5,2)

win (19,4), amazing (55,5)
awesome (51,4), please (9,4)

awsm

=

awesome

Foursquare Network

Temporal Activities

8 AM 12 PM 4 PM 8 PM 11 PM

Locations

User Accounts

check-in

check-in

friend

friend

friend

friend

write

Tips

anchor links

0.6

0.7

0.8

0.1

0.9

0.3

0.2

0.8

Twitter Network

Temporal Activities

8 AM 12 PM 4 PM 8 PM 11 PM

User Accounts

Locations

locate

follow

follow

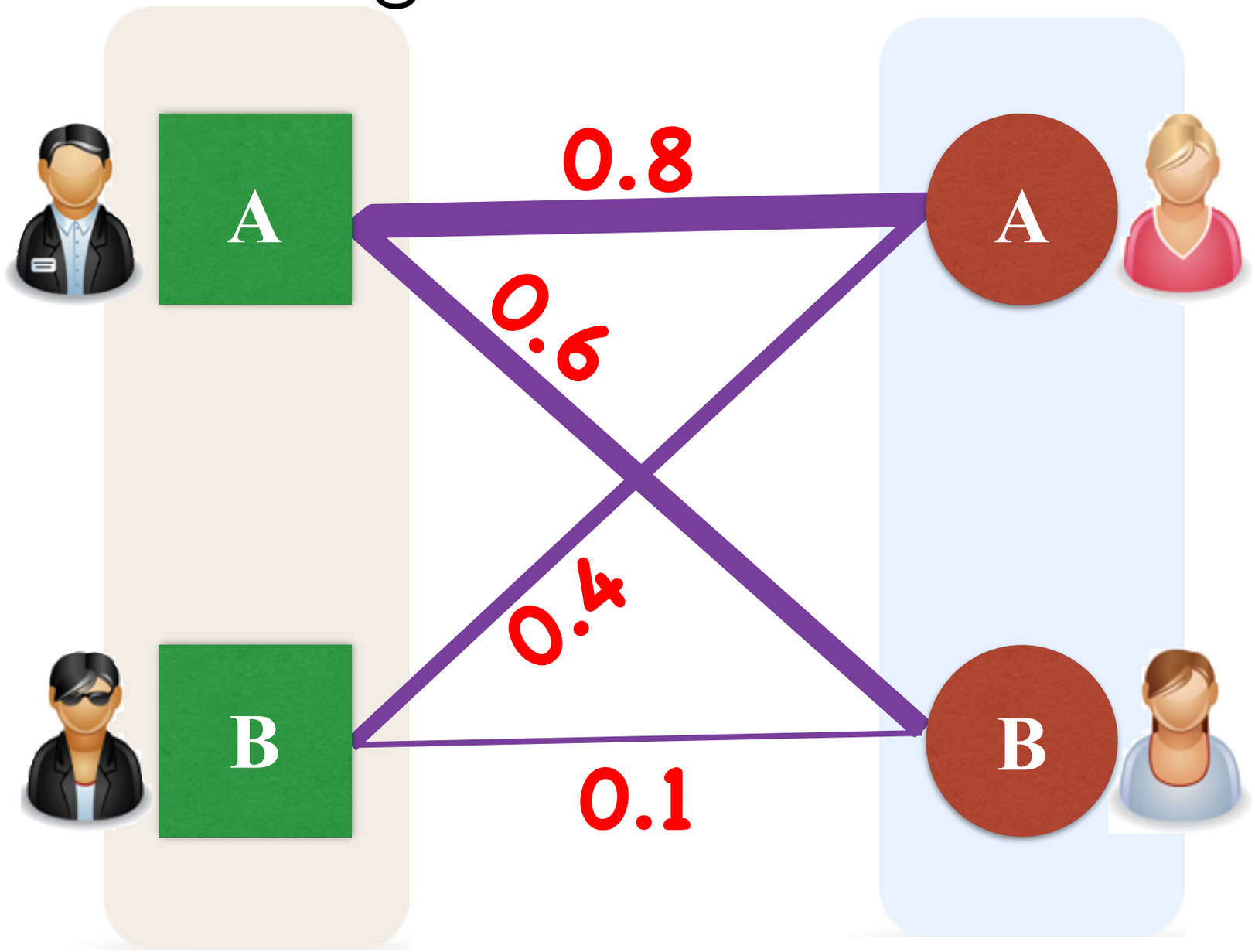
follow

locate

write

Tweets

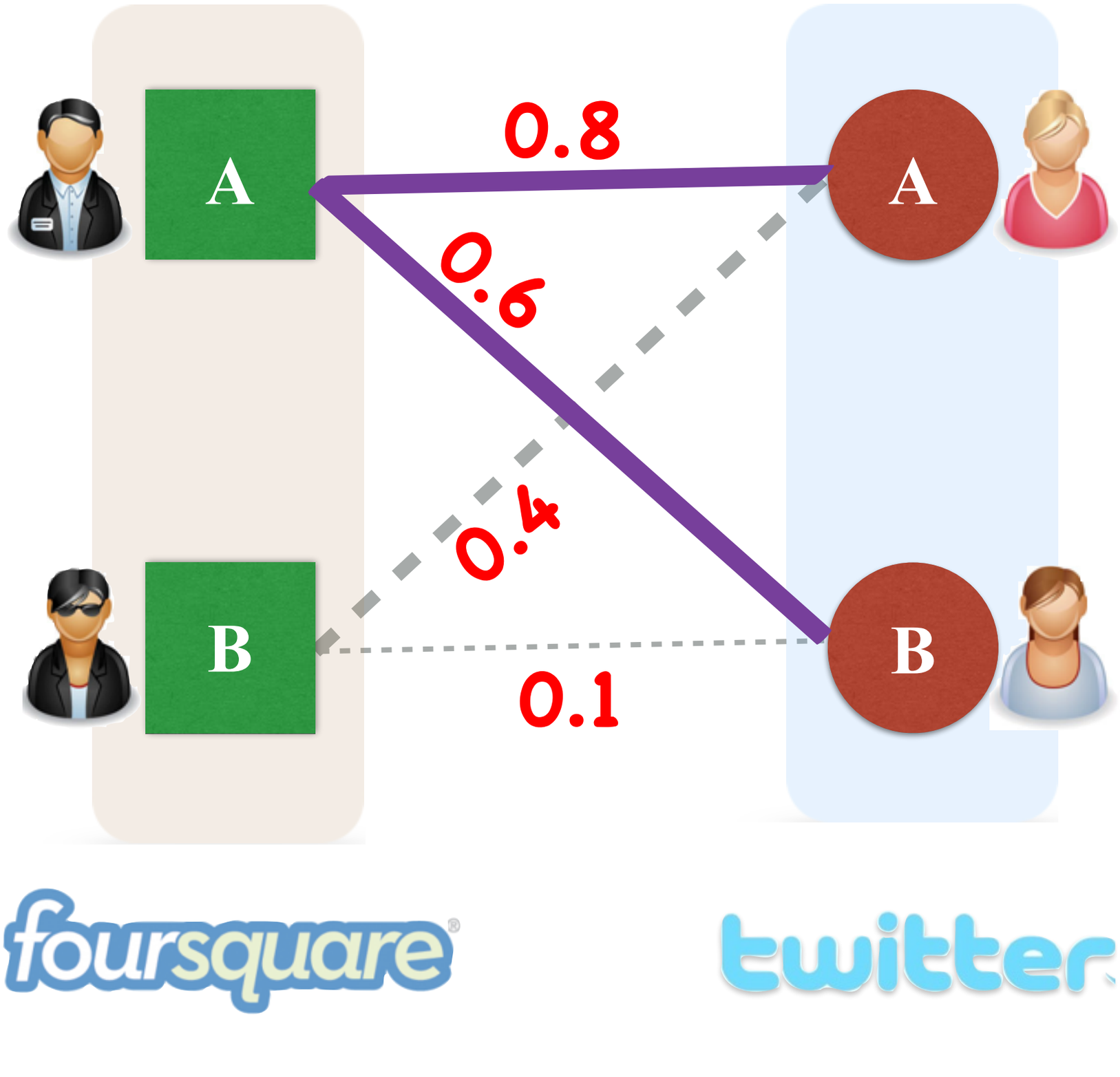
Link Prediction across Aligned Heterogeneous Networks



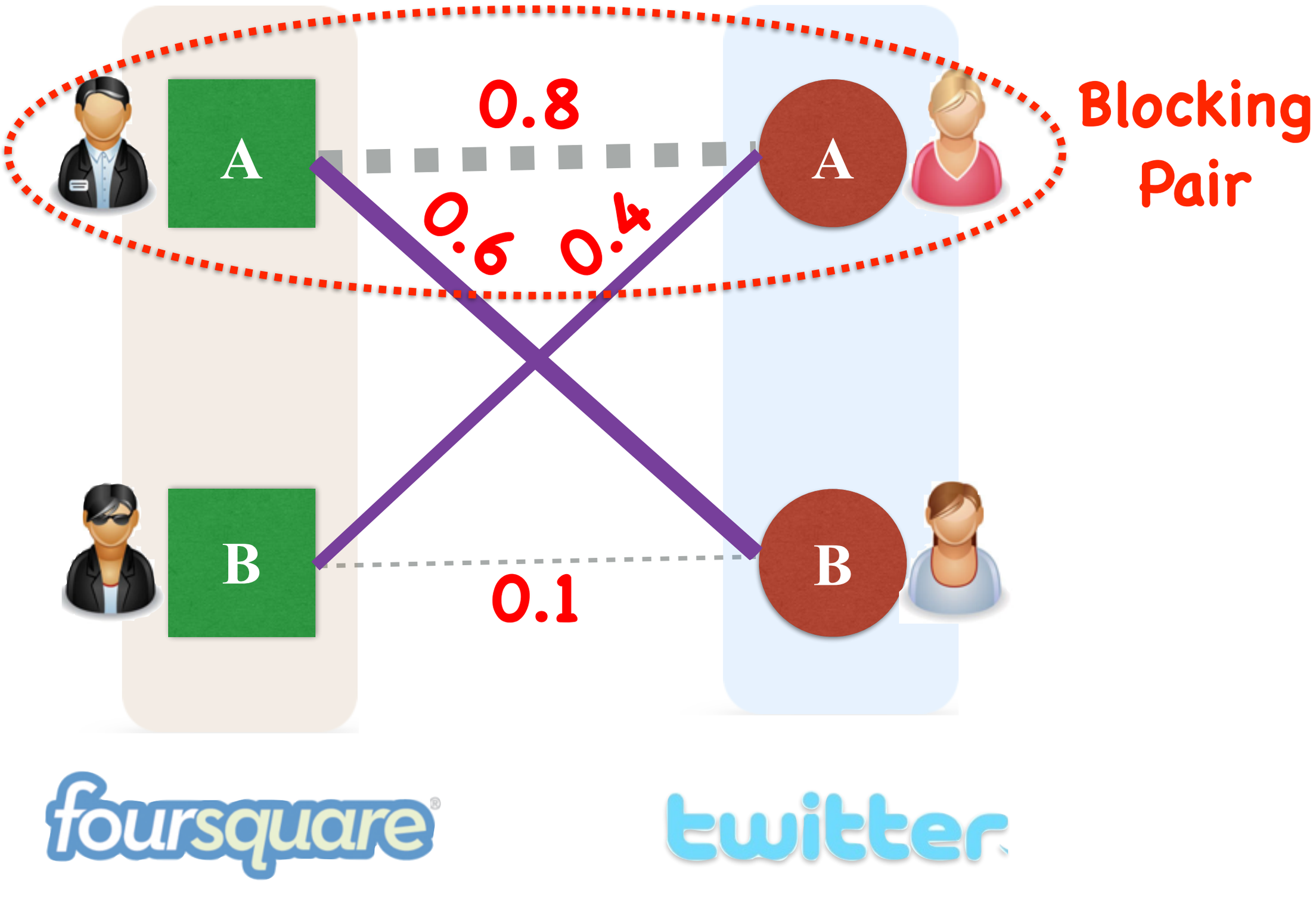
foursquare

twitter

Traditional Supervised Link Prediction



Max Sum of Scores w.r.t. 1-1 Constraint

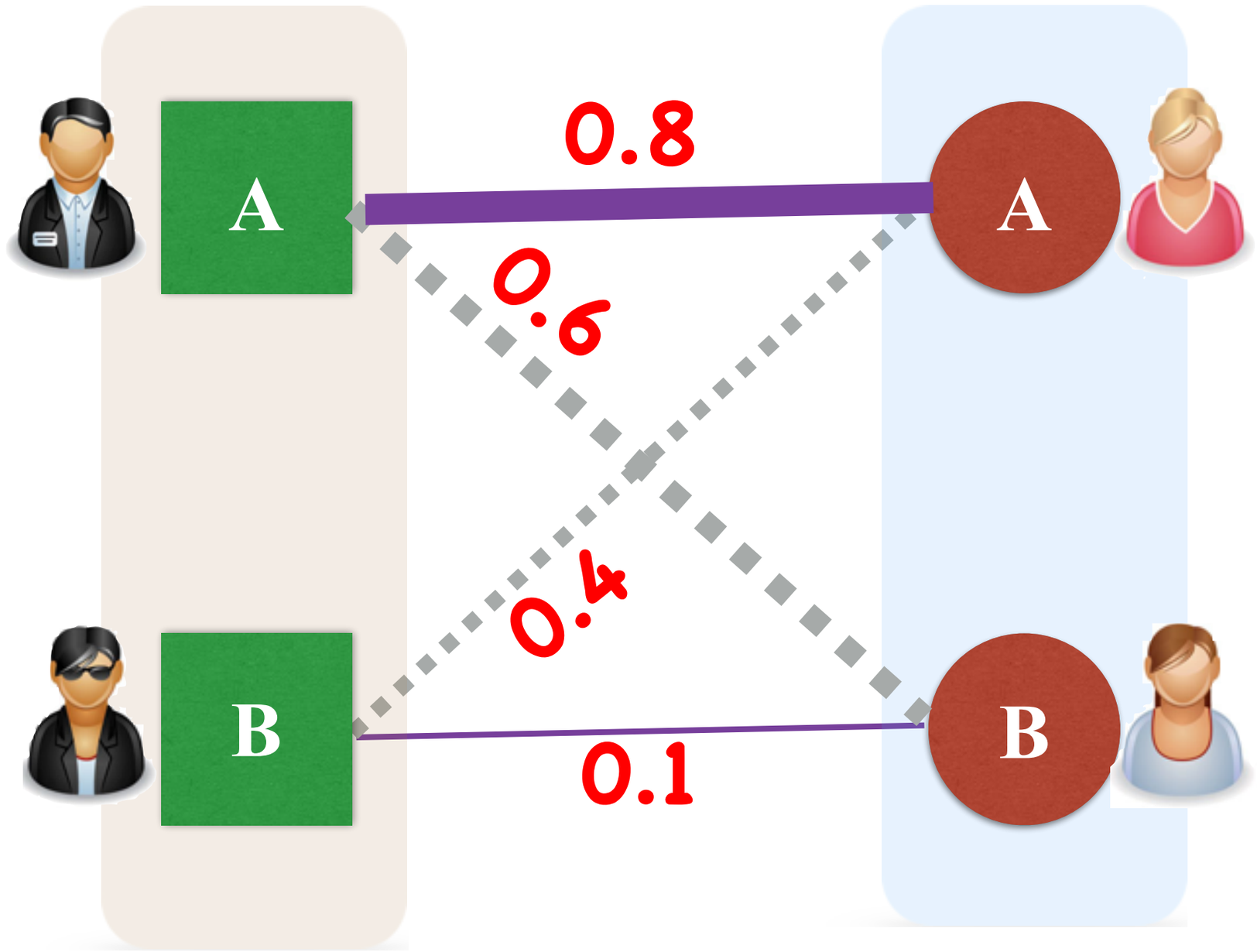


A

B

A

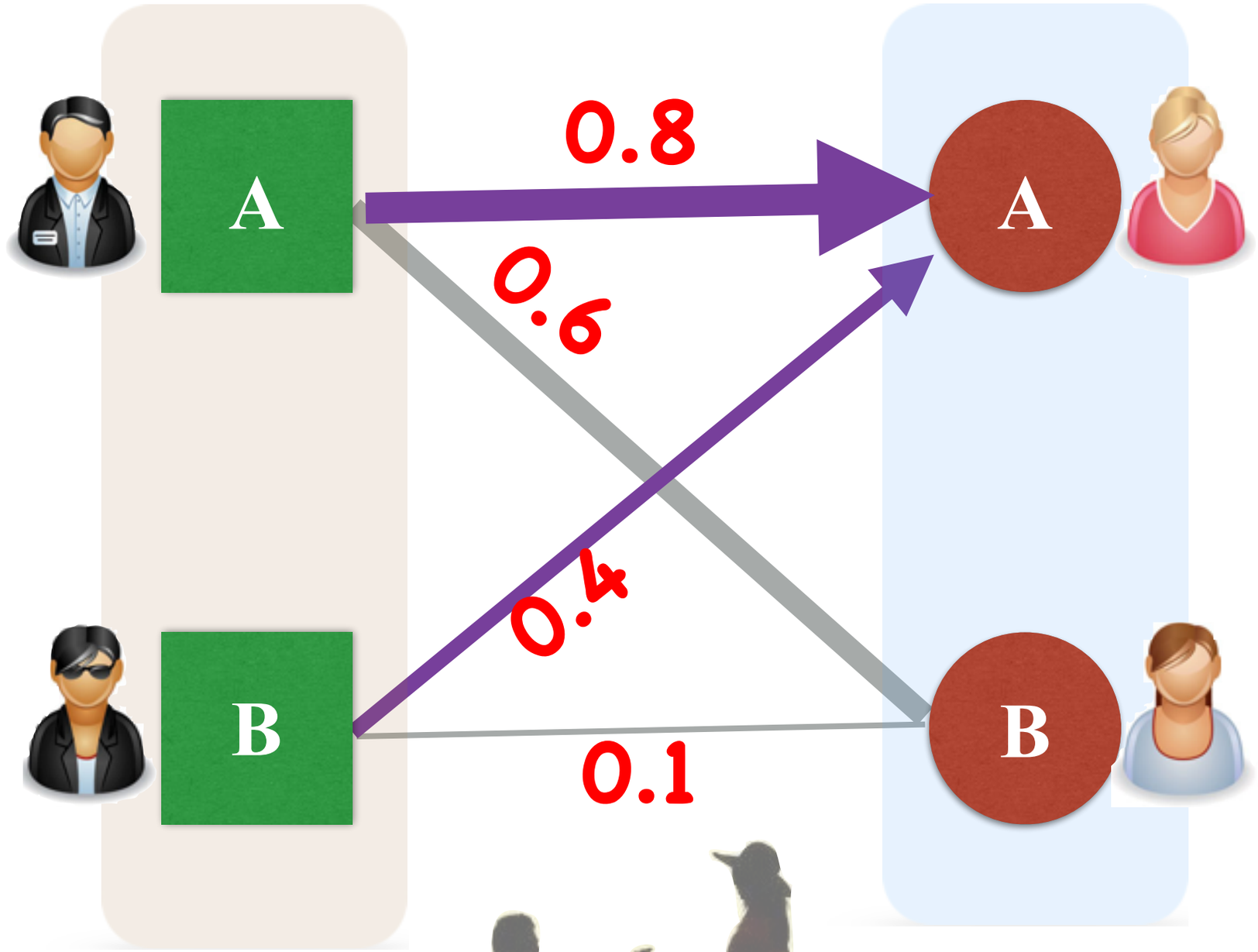
Stable Matching/Marriage



foursquare

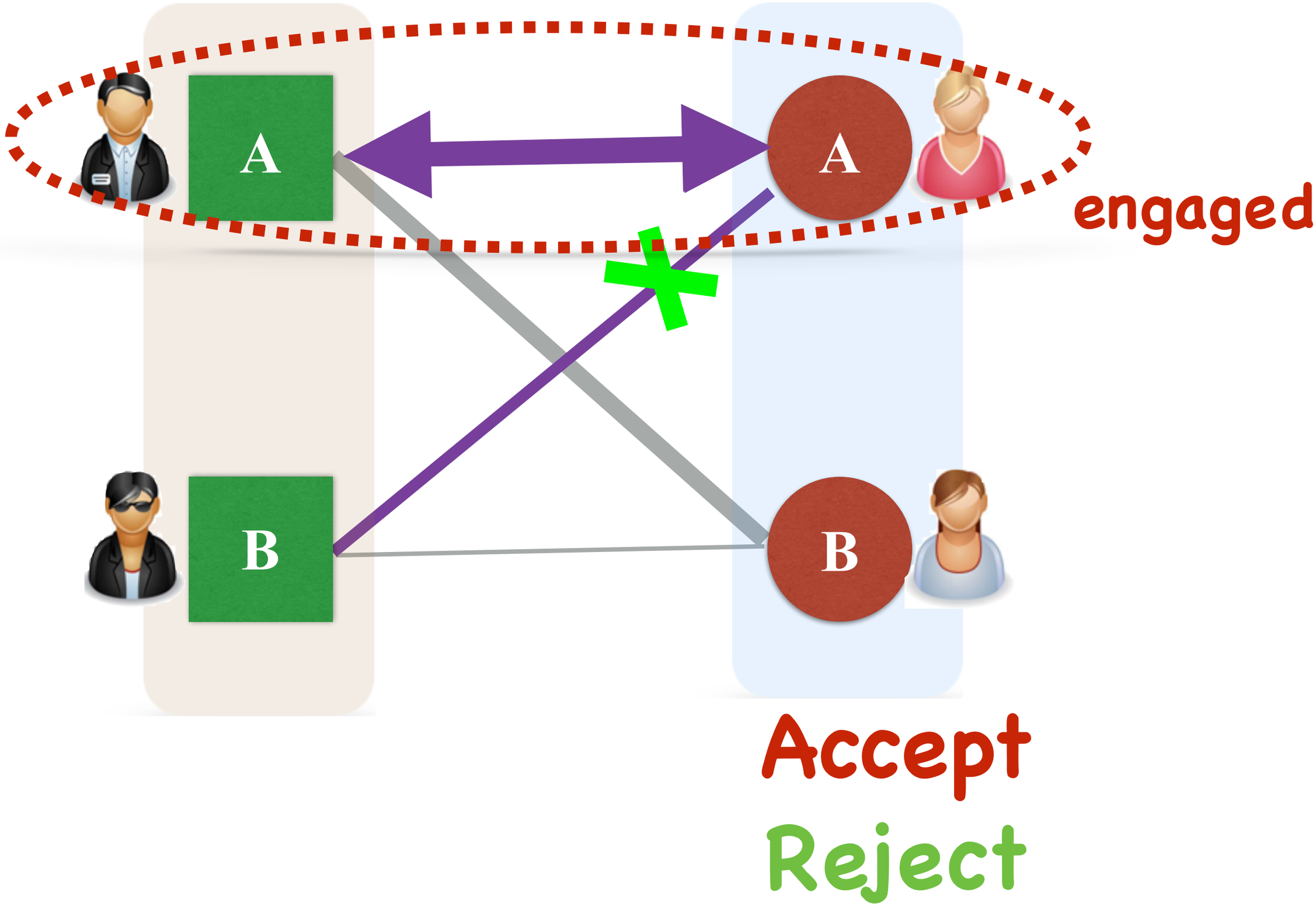
twitter

Stable Matching/Marriage

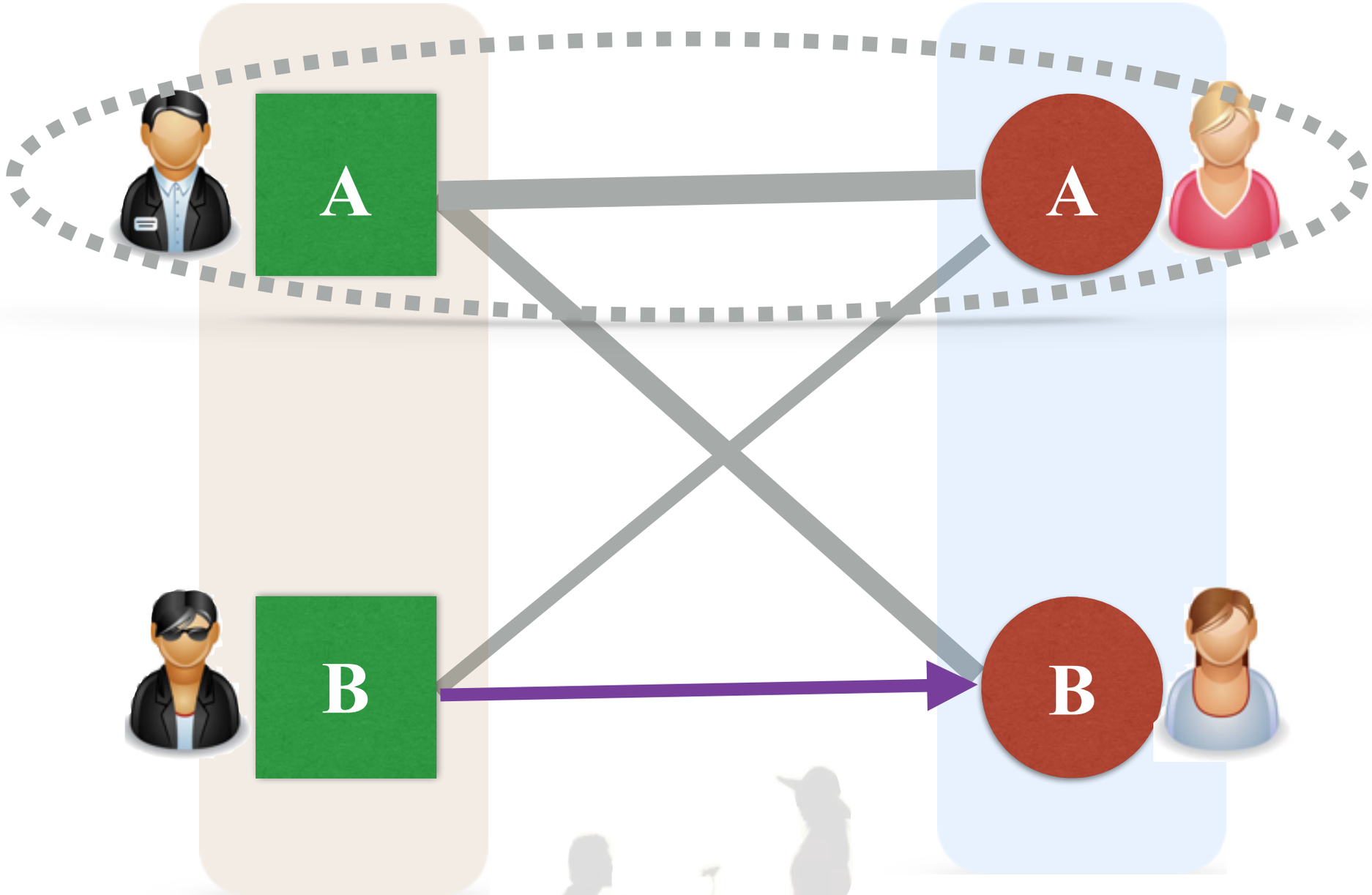


Propose

Stable Matching/Marriage

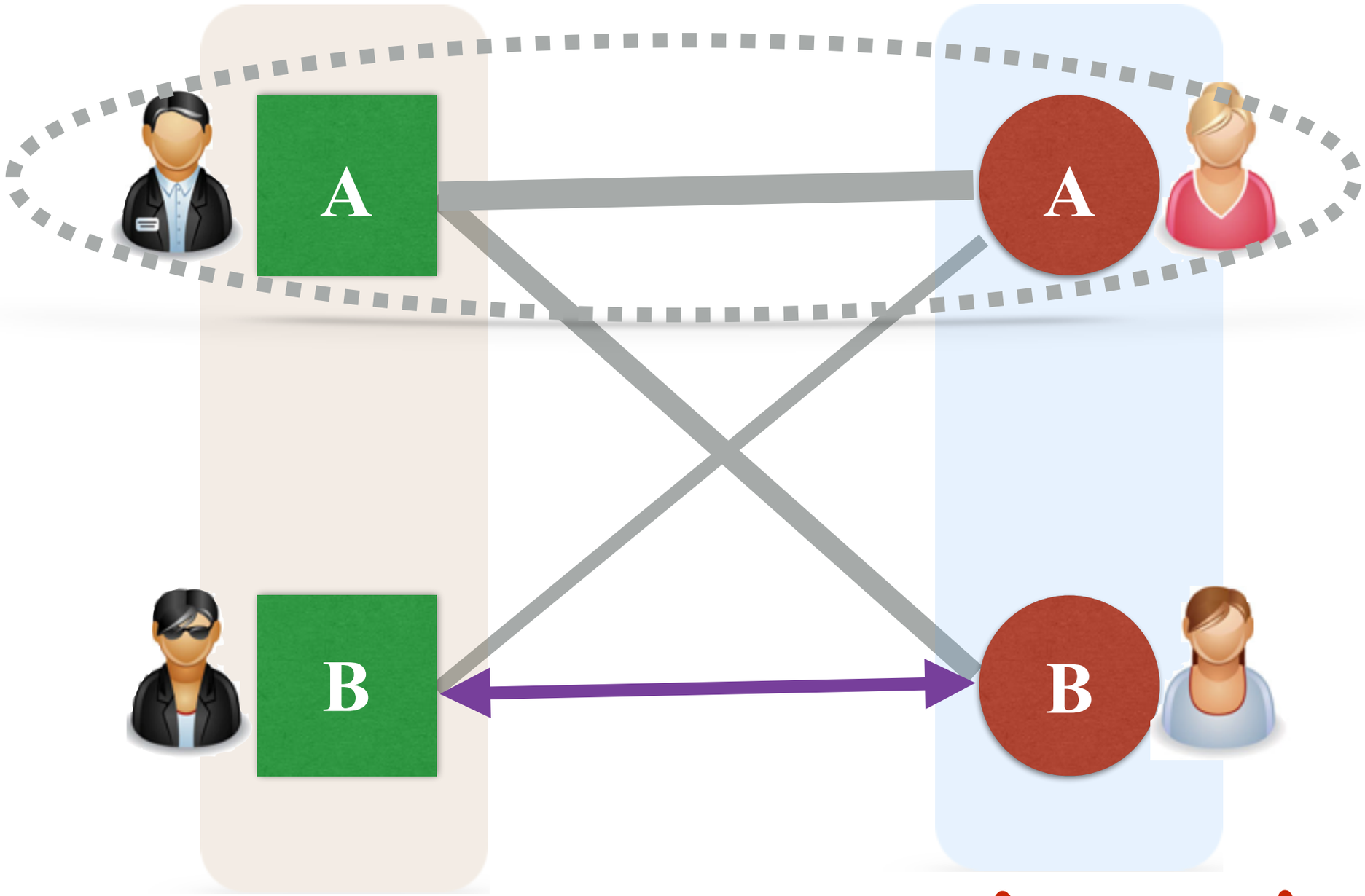


Stable Matching/Marriage



Propose

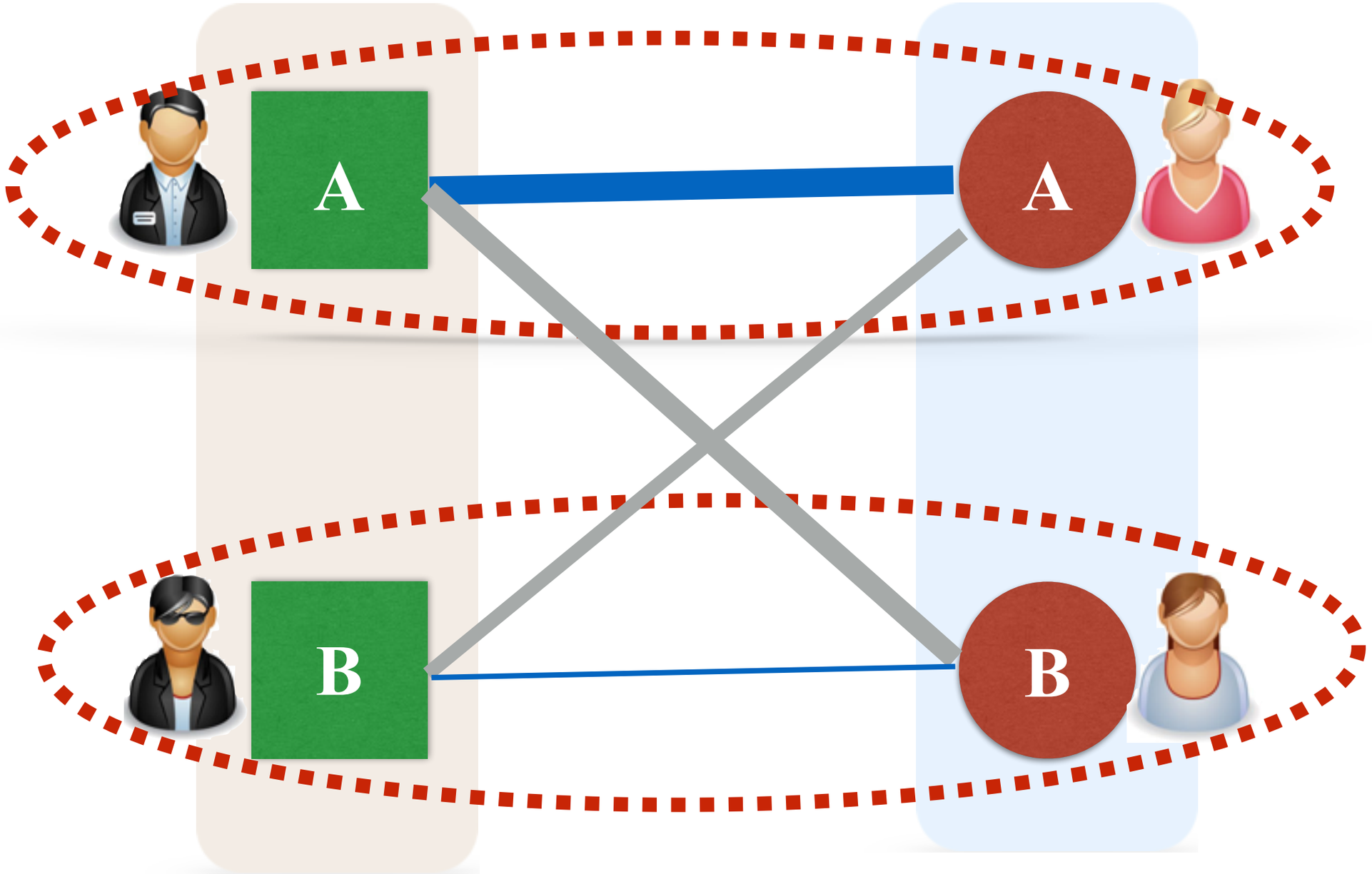
Stable Matching/Marriage



Accept

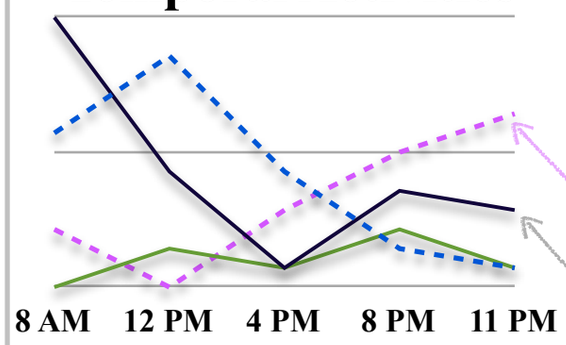
Reject

Stable Matching/Marriage



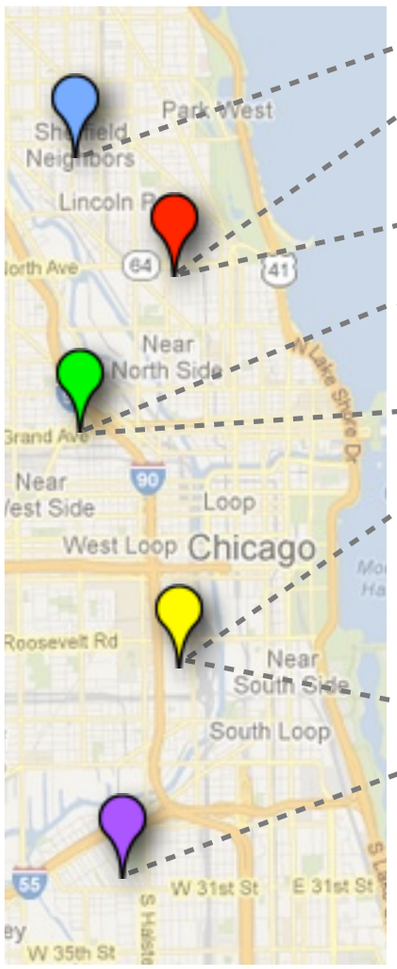
Foursquare Network

Temporal Activities



User Accounts

Locations



check-in

check-in

friend

friend

friend

friend

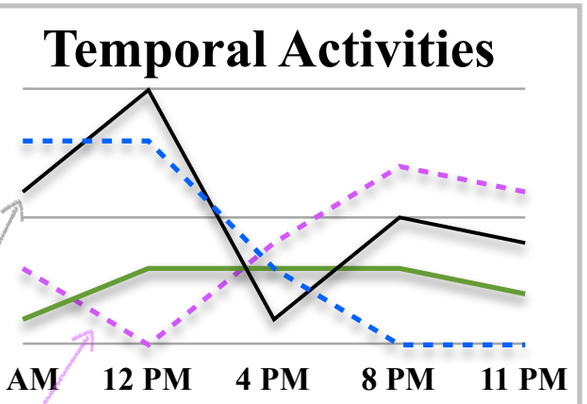
write

Tips

anchor links

Twitter Network

Temporal Activities



User Accounts

Locations

locate

follow

follow

follow

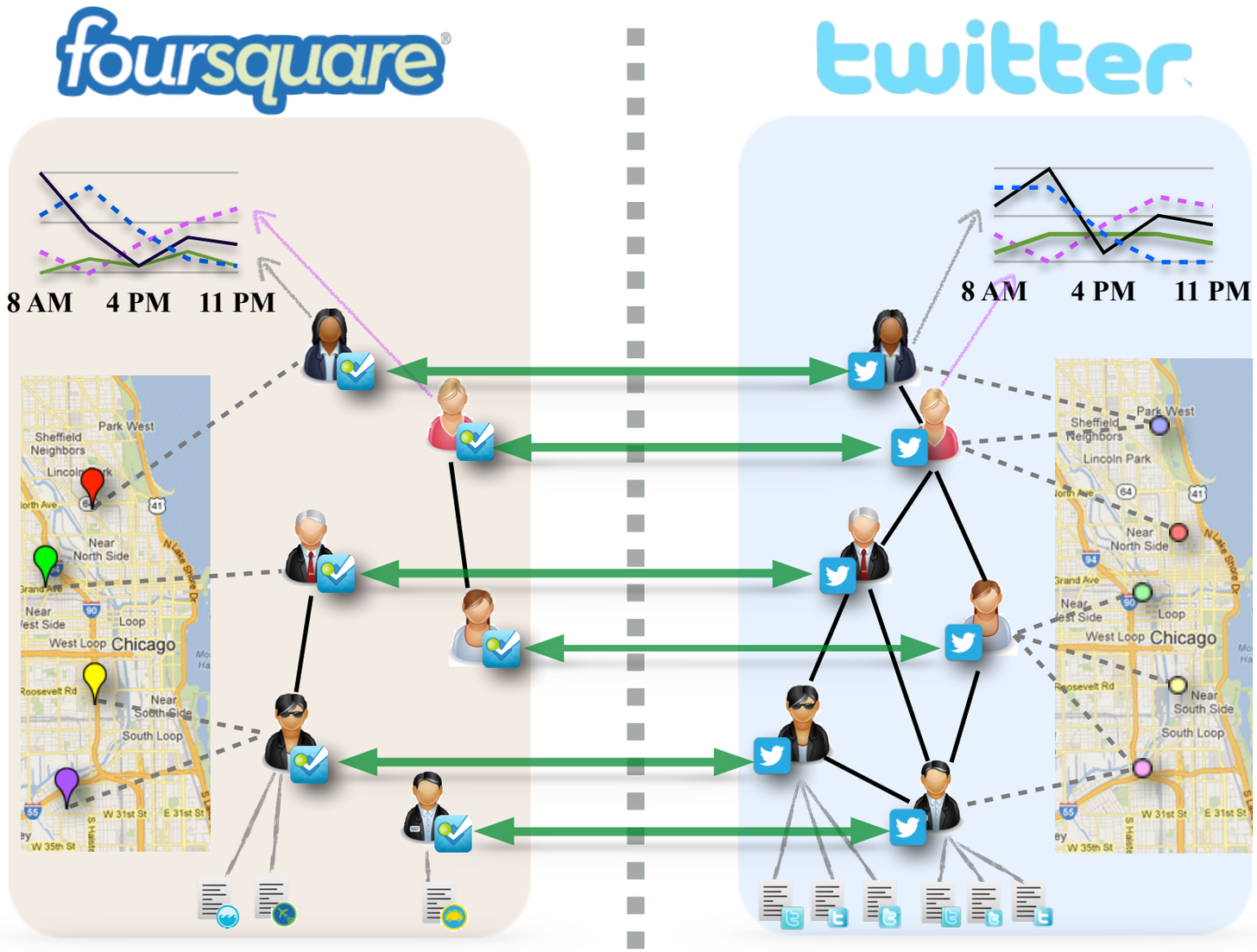
locate

write

Tweets

Link Prediction across Aligned Heterogeneous Networks

- Information Transfer across Aligned Networks for Link Prediction Tasks



Link Prediction across Aligned Heterogeneous Networks

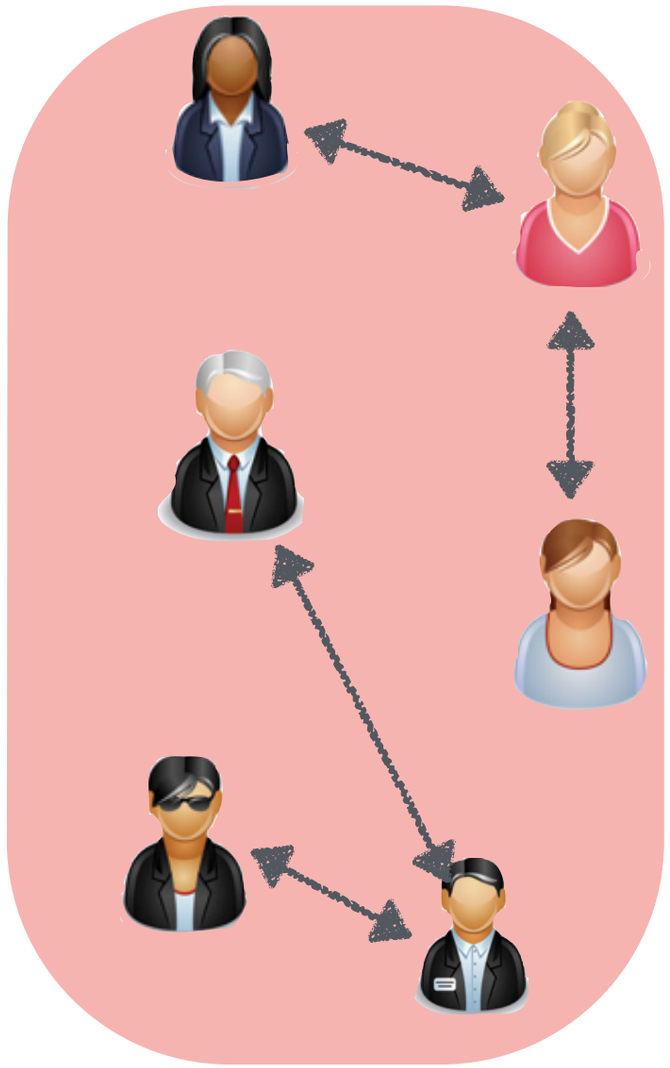
- Anchor Link Prediction
  - Supervised Link Prediction
  - Stable Matching
- Information Transfer across Aligned Networks

Outline

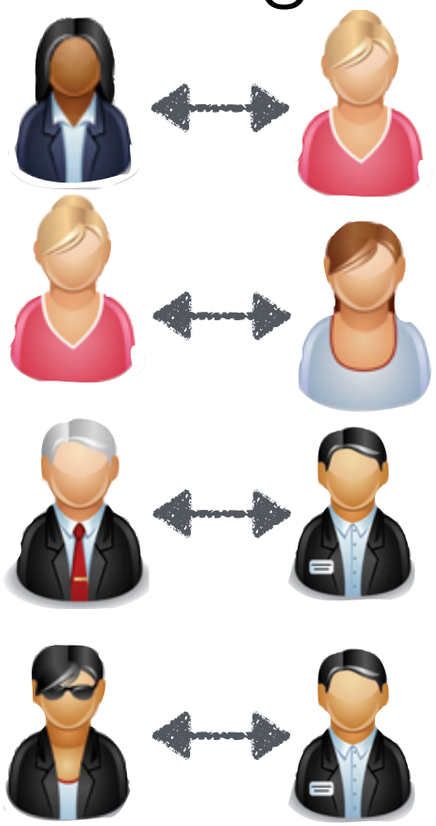
- Background Knowledge
- Problem Formulation
- Link Prediction in Homogeneous Networks
- Link Prediction in Heterogeneous Networks
- Link Prediction across Aligned Heterogeneous Networks
- **Future Works**
- Summary

Future Works

- Class Imbalance Problem in Supervised Link Prediction

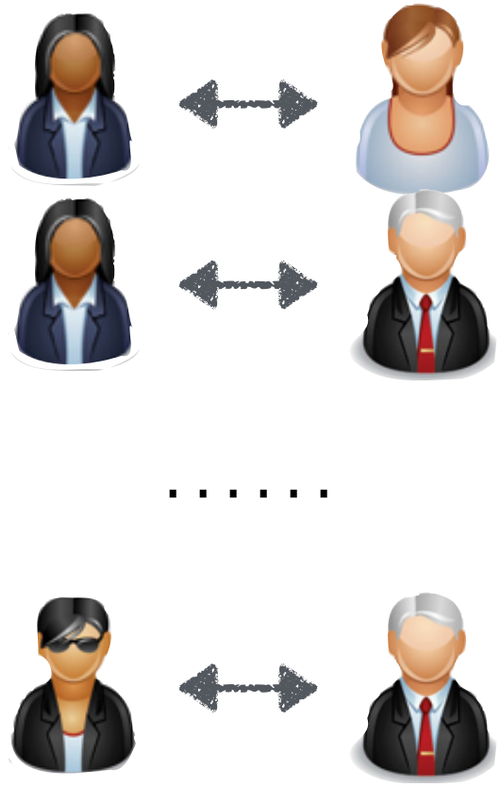


existing links



4

unconnected links

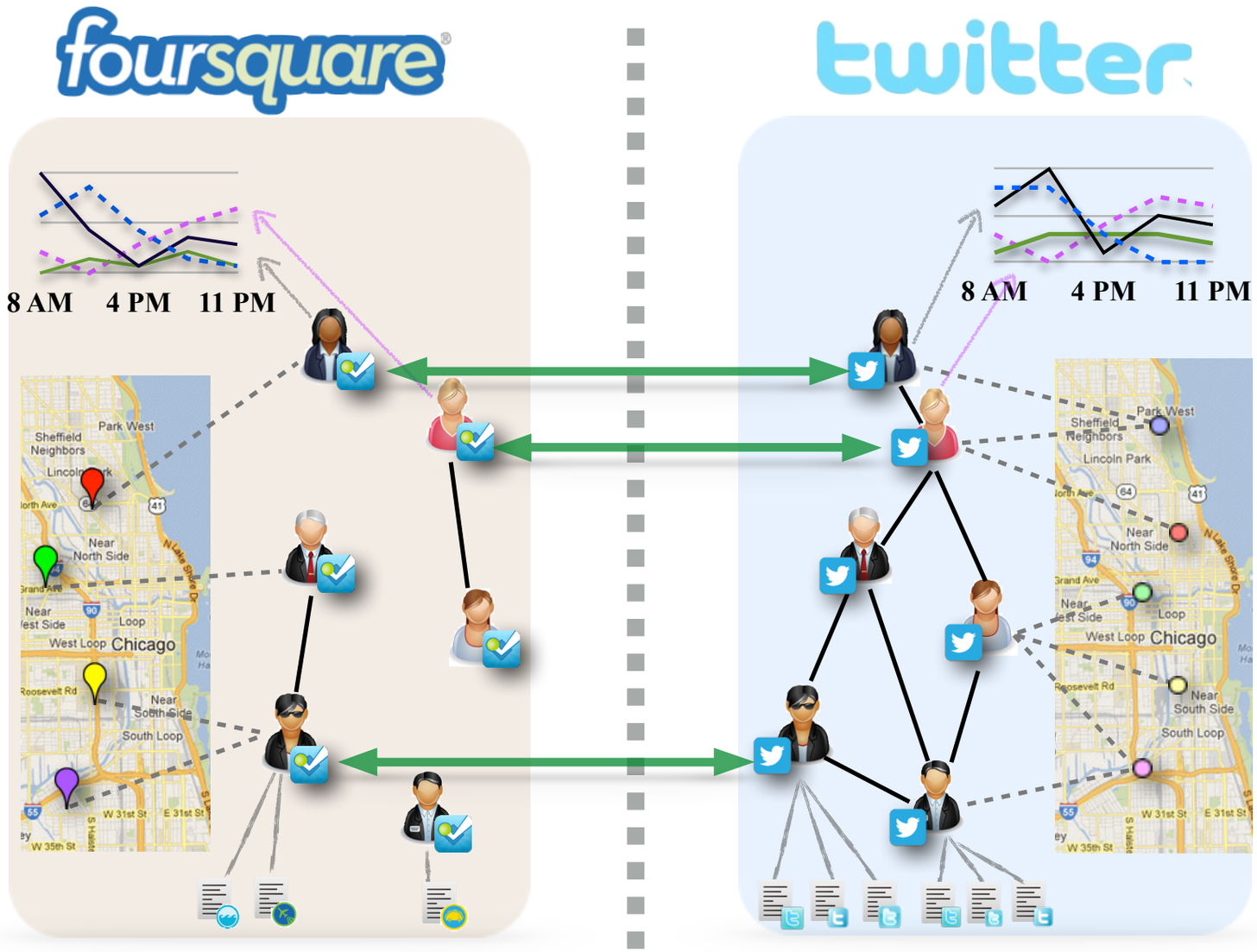


.....

6(6-1)/2 - 4 = 11

Future Works

- Information Transfer across Aligned Networks for both anchor and non-anchor users



Outline

- Background Knowledge
- Problem Formulation
- Link Prediction in Homogeneous Networks
- Link Prediction in Heterogeneous Networks
- Link Prediction across Aligned Heterogeneous Networks
- Future Works
- **Summary**

Summary

- Link Prediction Problem Definition
- Link Prediction for Homogeneous Networks
  - various unsupervised predicators
- Link Prediction for Heterogeneous Networks
  - Supervised Link Prediction Methods
- Link Prediction across Aligned Heterogeneous Networks
  - Anchor Link Prediction Problem: Supervised Method + Stable Matching
- Future Works
  - Class Imbalance Problem, Information Transfer across Aligned Networks

Thanks!

Q & A