On Mining Big Data & Social Network Analysis

PHILIP S. YU

(PSYU@UIC.EDU)

UIC DISTINGUISHED PROFESSOR

WEXLER CHAIR PROFESSOR

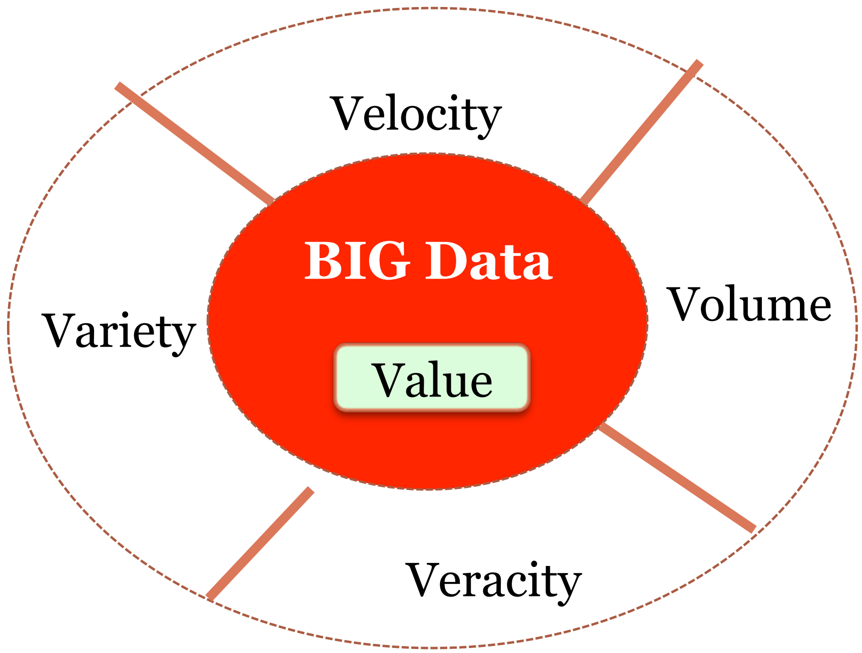
IN INFORMATION TECHNOLOGY

UNIVERSITY OF ILLINOIS AT CHICAGO

Big Data

Data Stream
with Concept Drift

- High Dimensional Data
- Heterogeneous Data Sources
- Unconventional Data Types
  - Graph/Network
  - Sequence
  - Text



Scalable
Mining
Algorithms

- Cleaness
- Trustworthiness
- Privacy

Outline

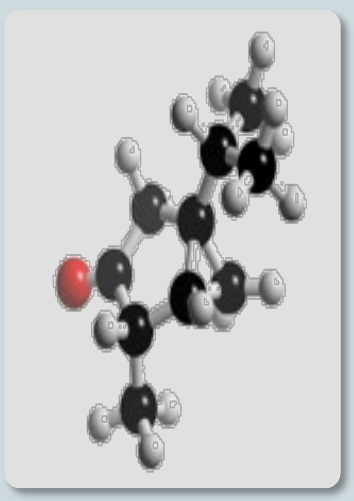
- Mining heterogeneous data sources
- Fusion knowledge across multiple social networks
- Using social networks to
  - understand customer purchase behavior
  - predict or promote real world activities
    - Inferring the impact of social media on crowdfunding

Information Fusion

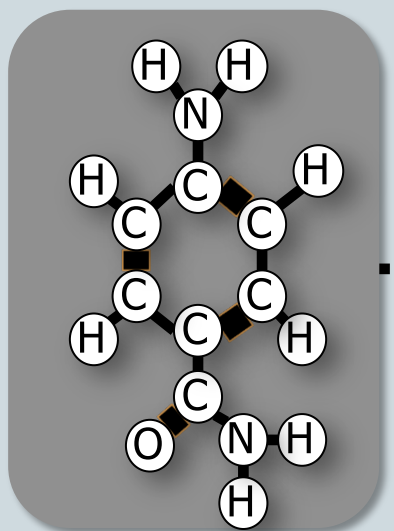
- Fusing information across multiple sources is the *Holy Grail* of big data research
- Many commercial companies have multiple sources of collecting customer information
  - Google has Google search, G-mail, Google Maps, Google+, YouTube, etc.
- Other examples
  - Detection of terrorist plots
  - Whereabouts on Malaysia MH370

Drug Discovery

Chemical Compound



Graph Object

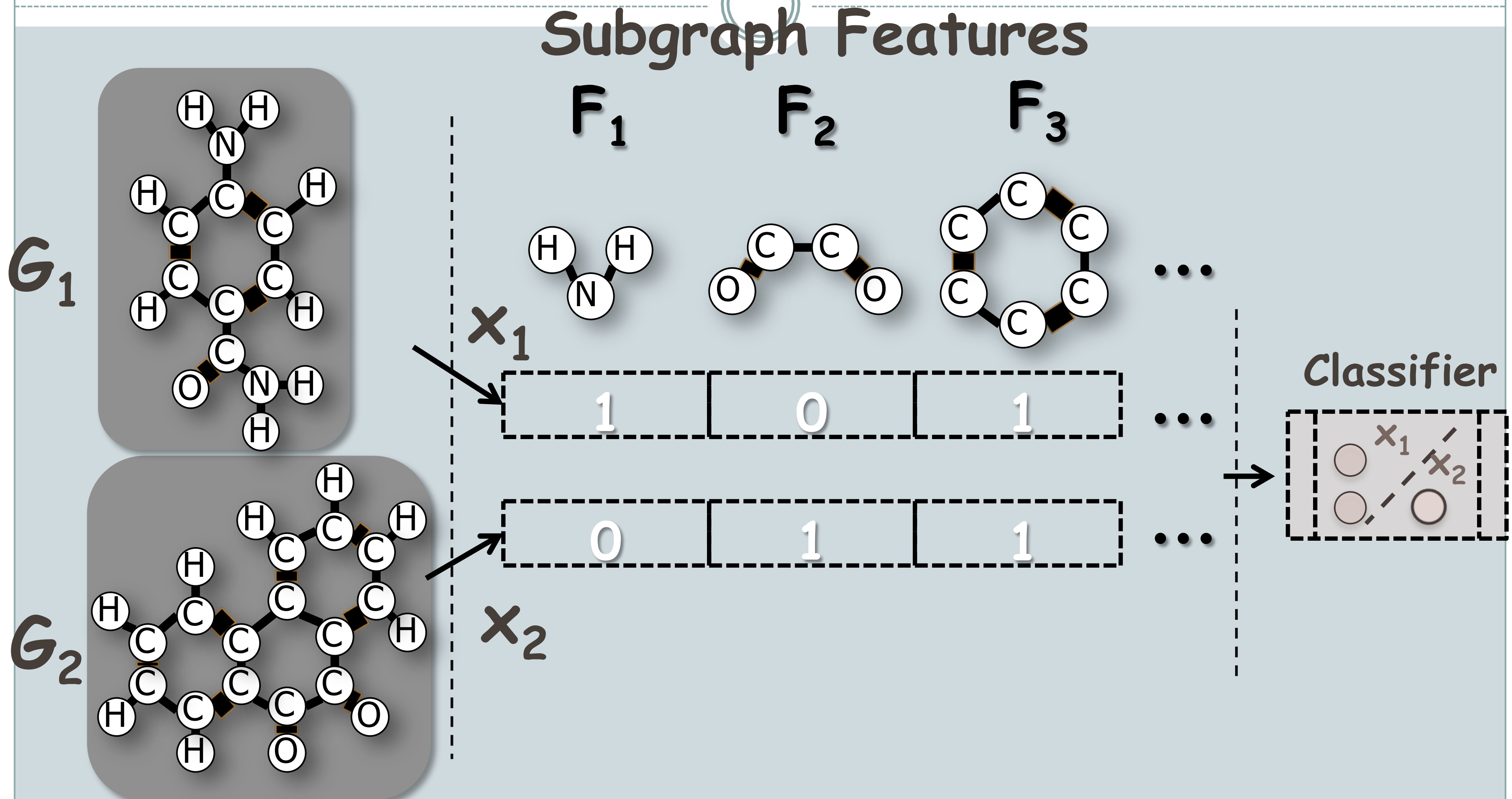


Anti-cancer activity

label

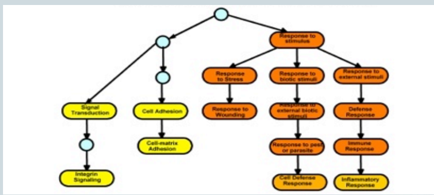
+/-

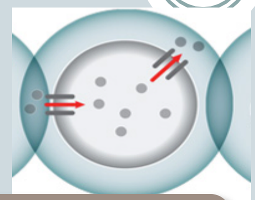
Subgraph Features

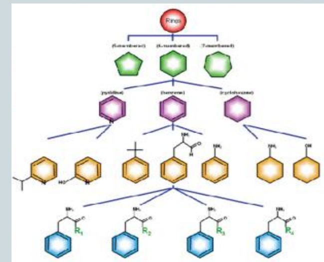


Graph Objects ⟶ Feature Vectors ⟶ Classifiers

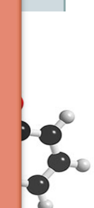
SLAP: HIN for Drugs



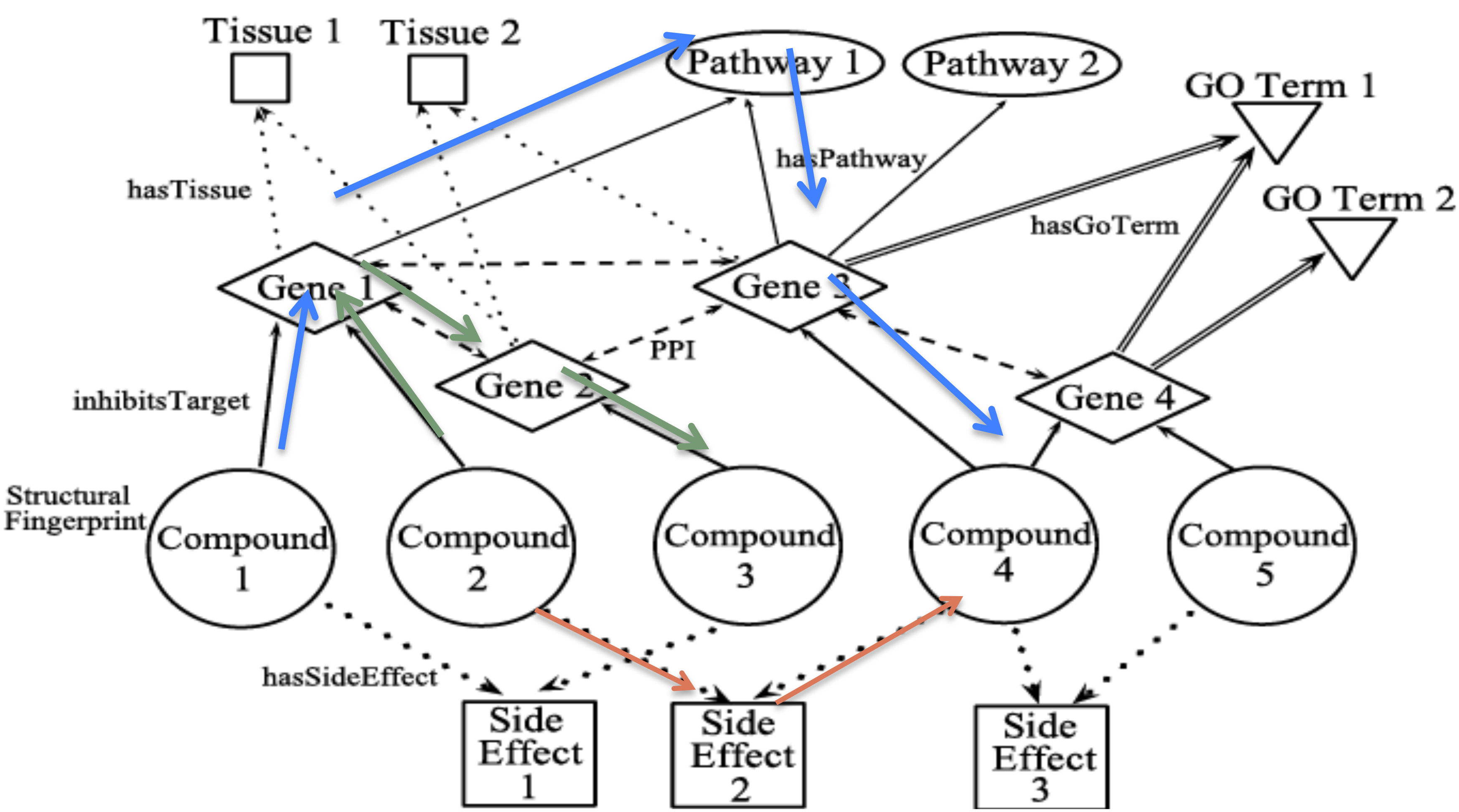




- SLAP is a subset of the Chem2Bio2RDF network
  - including 250,000 compounds with known bioactivities and the targets known to associated with these drugs
- Chem2Bio2RDF network semantically integrates ***42*** heterogeneous public datasets related to drug discovery
  - Major datasets include PubChem, ChEMBL, DrugBank, PharmGKB, BindingDB, STITCH, CTD, KEGG, SWISSPROT, PDB, SIDER, PubMed.



Path-based Collective Classification



$$P_1 : Compound \xrightarrow{hasSideEffect} Side\ Effect \xrightarrow{hasSideEffect^{-1}} Compound$$

$$P_2 : Compound \xrightarrow{inhibitsTarget} Gene \xrightarrow{hasPathway} Pathway \xrightarrow{hasPathway^{-1}} Gene \xrightarrow{inhibitsTarget^{-1}} Compound$$

$$P_3 : Compound \xrightarrow{inhibitsTarget} Gene \xrightarrow{PPI} Gene \xrightarrow{inhibitsTarget^{-1}} Compound$$

Mining Heterogeneous Information Networks

- Intuition
  - Two objects can be connected via different connectivity meta paths
    - E.g., two chemical compounds can be connected by

$$P_1 : Compound \xrightarrow{hasSideEffect} Side\ Effect \xrightarrow{hasSideEffect^{-1}} Compound$$

$$P_2 : Compound \xrightarrow{inhibitsTarget} Gene \xrightarrow{hasPathway} Pathway \xrightarrow{hasPathway^{-1}} Gene \xrightarrow{inhibitsTarget^{-1}} Compound$$

$$P_3 : Compound \xrightarrow{inhibitsTarget} Gene \xrightarrow{PPI} Gene \xrightarrow{inhibitsTarget^{-1}} Compound$$

  - Each connectivity meta path represents a different semantic meaning and implies different similarity semantics or relationships
- Challenges
  - How to assess the importance of a meta path?
  - How to identify, select and combine different meta paths together?

Multi-label Drug Target Prediction

Table 2: Classification performances "average score ± std (rank)" on Drug-Target Binding prediction task. "↓" indicates the smaller the value the better the performance; "↑" indicates the larger the value the better the performance.

criteria	#label	BSVM	ECC	PISL	PIML	PIPL
Micro-F1 ↑	10	0.532±0.046 (5)	0.576±0.053 (4)	0.608±0.046 (3)	0.611±0.040 (2)	0.625±0.042 (1)
	20	0.553±0.019 (5)	0.588±0.018 (4)	0.696±0.016 (3)	0.714±0.011 (2)	0.724±0.011 (1)
	30	0.536±0.052 (5)	0.585±0.054 (4)	0.674±0.032 (3)	0.695±0.025 (2)	0.706±0.026 (1)
	40	0.523±0.018 (5)	0.568±0.022 (4)	0.599±0.022 (3)	0.618±0.022 (2)	0.642±0.022 (1)
	50	0.521±0.028 (5)	0.571±0.036 (4)	0.603±0.031 (3)	0.635±0.028 (2)	0.653±0.026 (1)
Hamming Loss ↓	10	0.024±0.003 (5)	0.021±0.003 (4)	0.020±0.003 (2)	0.020±0.002 (3)	0.018±0.002 (1)
	20	0.019±0.001 (5)	0.017±0.000 (4)	0.012±0.001 (3)	0.012±0.001 (2)	0.011±0.001 (1)
	30	0.018±0.002 (5)	0.016±0.002 (4)	0.012±0.001 (3)	0.011±0.000 (2)	0.010±0.000 (1)
	40	0.017±0.001 (5)	0.015±0.001 (4)	0.014±0.001 (3)	0.013±0.001 (2)	0.012±0.001 (1)
	50	0.016±0.001 (5)	0.014±0.001 (4)	0.013±0.001 (3)	0.012±0.001 (2)	0.011±0.001 (1)
Subset 0/1 Loss ↓	10	0.147±0.012 (5)	0.128±0.017 (4)	0.123±0.011 (2)	0.124±0.010 (3)	0.113±0.010 (1)
	20	0.222±0.009 (5)	0.193±0.006 (4)	0.165±0.011 (3)	0.163±0.010 (2)	0.148±0.004 (1)
	30	0.265±0.019 (5)	0.223±0.029 (4)	0.214±0.007 (3)	0.207±0.004 (2)	0.182±0.003 (1)
	40	0.305±0.008 (5)	0.250±0.004 (2)	0.268±0.010 (4)	0.257±0.010 (3)	0.223±0.010 (1)
	50	0.351±0.009 (5)	0.288±0.018 (2)	0.306±0.013 (4)	0.288±0.020 (3)	0.261±0.017 (1)

Outline

- Mining heterogeneous data sources
- Fusing knowledge across multiple social networks
- Using social networks to
  - Understand customer purchase behavior
  - Predict or promote real world activities
    - Inferring the impact of social media on crowdfunding

Social Network

- Huge size
  - Facebook: more than a billion nodes
- High volume of new content generated
  - Rapidly and dynamically changing focus
- Rich information with many different types of data
- Noisy
- High aggregate value, but challenging to mine

Background

- Many social networks with different objectives
  - Facebook
  - Twitter
  - Foursquare
  - LinkedIn
  - YouTube
  - Instagram
  - WhatsApp
  - Google+
- Individuals often participate in multiple social networks

Fusion of Multiple Social Networks

- Each social network only capture a partial or biased view of an individual
- Newly formed social networks can be benefitted from information collected in more established networks
- Publicly available social network data can be rich and useful
- Fusing multiple social networks has the additional challenge on identity matching

Issues

- How to connect the multiple accounts of the same users in different social networks?
- How to transfer knowledge across different social networks?

Foursquare

Discover places that your friends love

Sign up with Facebook or Sign up with email

1. ## Millennium Park

9.6 201 E Randolph St (btwn Columbus Dr & Michigan Ave)
Park · 10

This spot is popular

"... are the **Crown Fountain**, a public art and video..." (8 tips)
"... Jay **Pritzker Pavilion** and the BP Pedestrian..." (6 tips)
"... photogenic **Cloud Gate sculpture** (nicknamed "The..." (5 tips)

Save

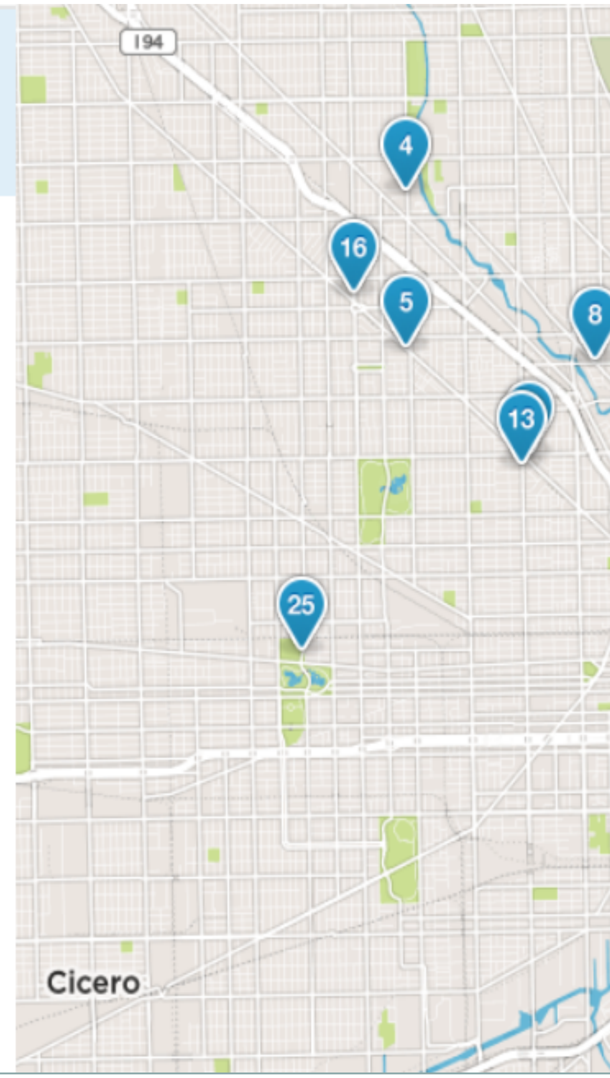
2. ## Intelligentsia Coffee

9.5 53 E Randolph St (btwn Wabash Ave & Garland Ct)
Coffee Shop · 3 · $ $ $ $ · View Menu

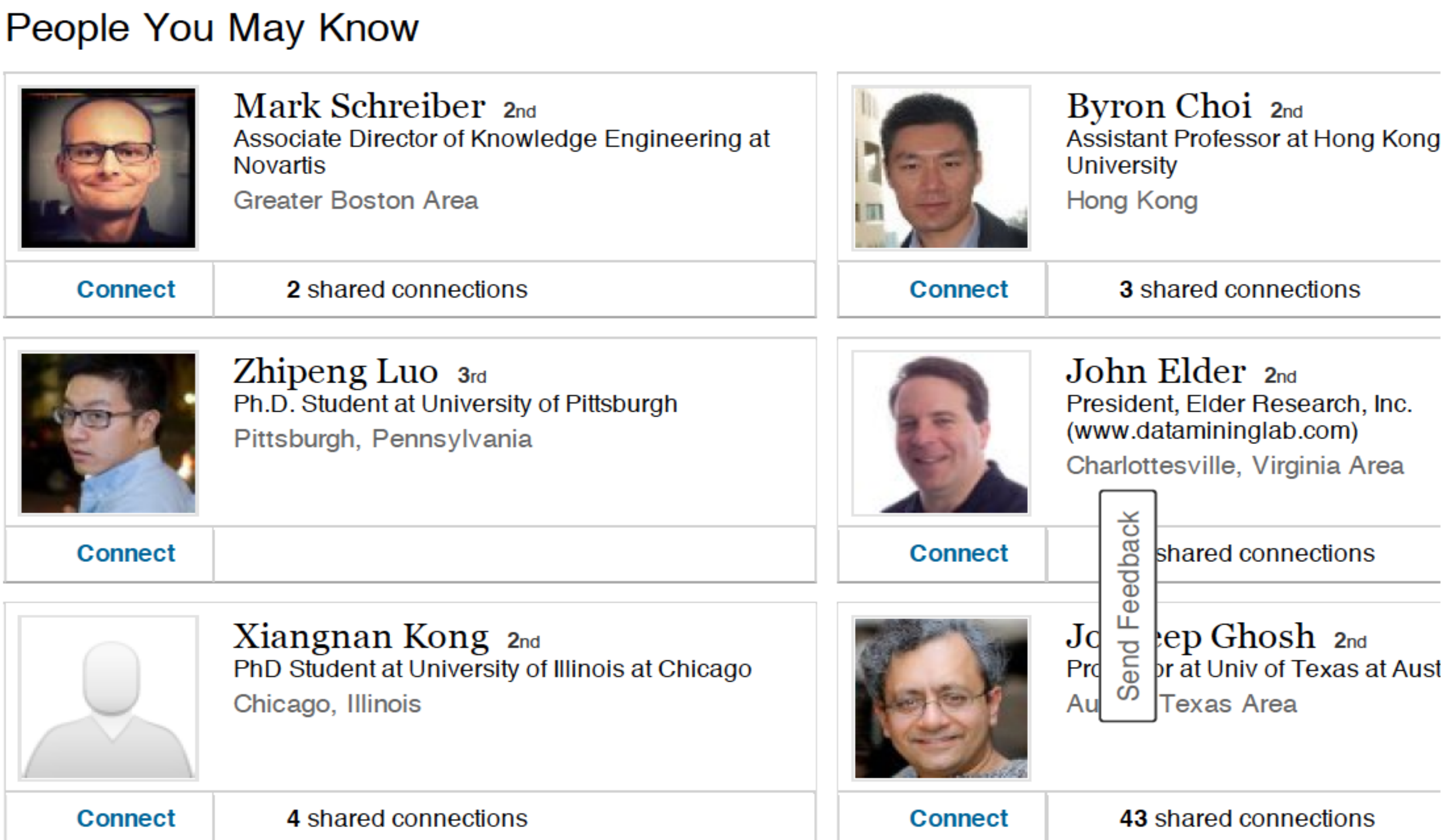
Lots of people like this place

"**Best coffee in Chicago**! Plus an outdoor..." (4 tips)
"... is good. The **espresso** brownies are extra good!!!" (13 tips)
"The **Pour Over** prepared on the chemex is..." (4 tips)

Save

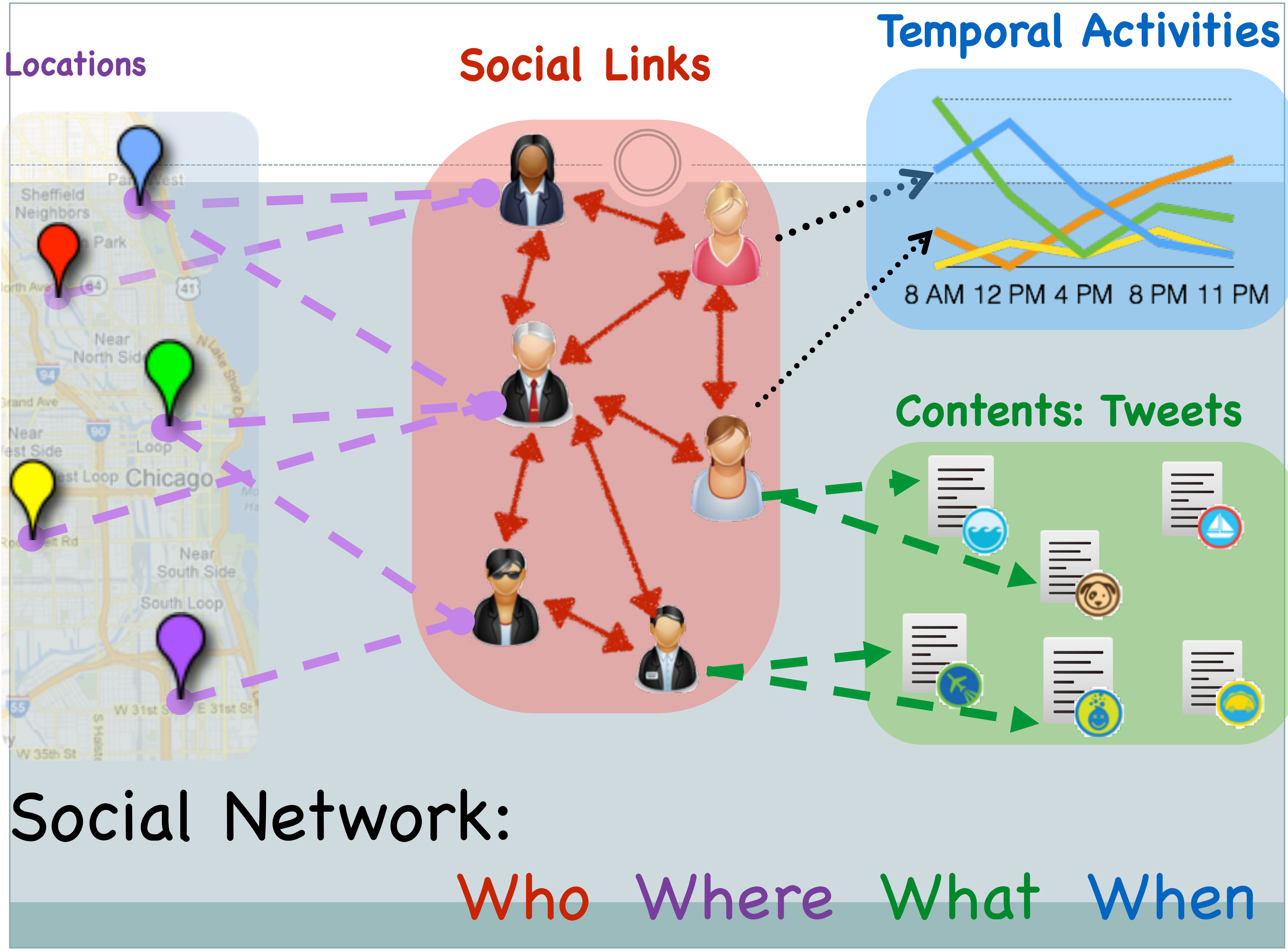


Friend Recommendation (Social Link Prediction)

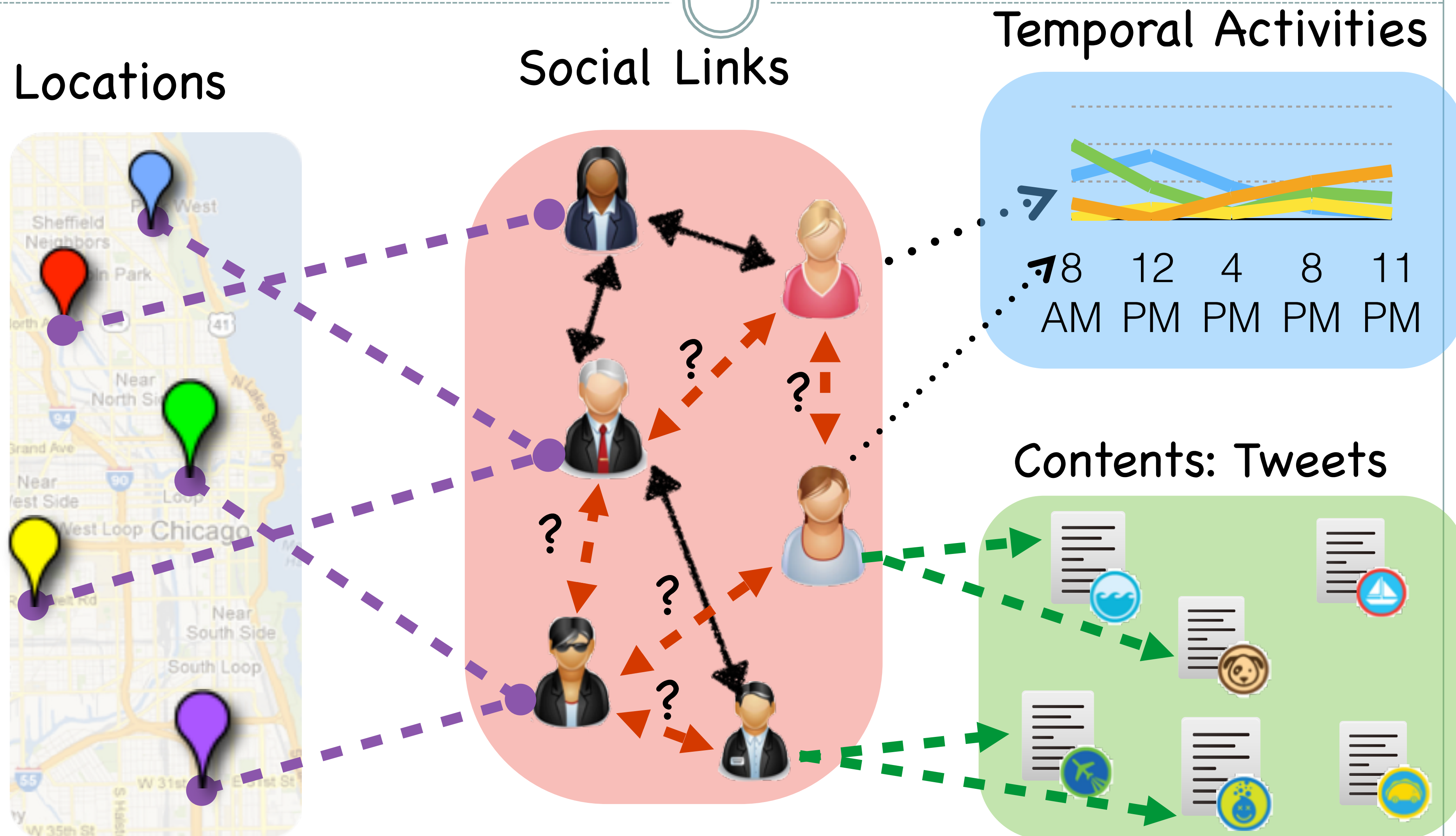


Challenges

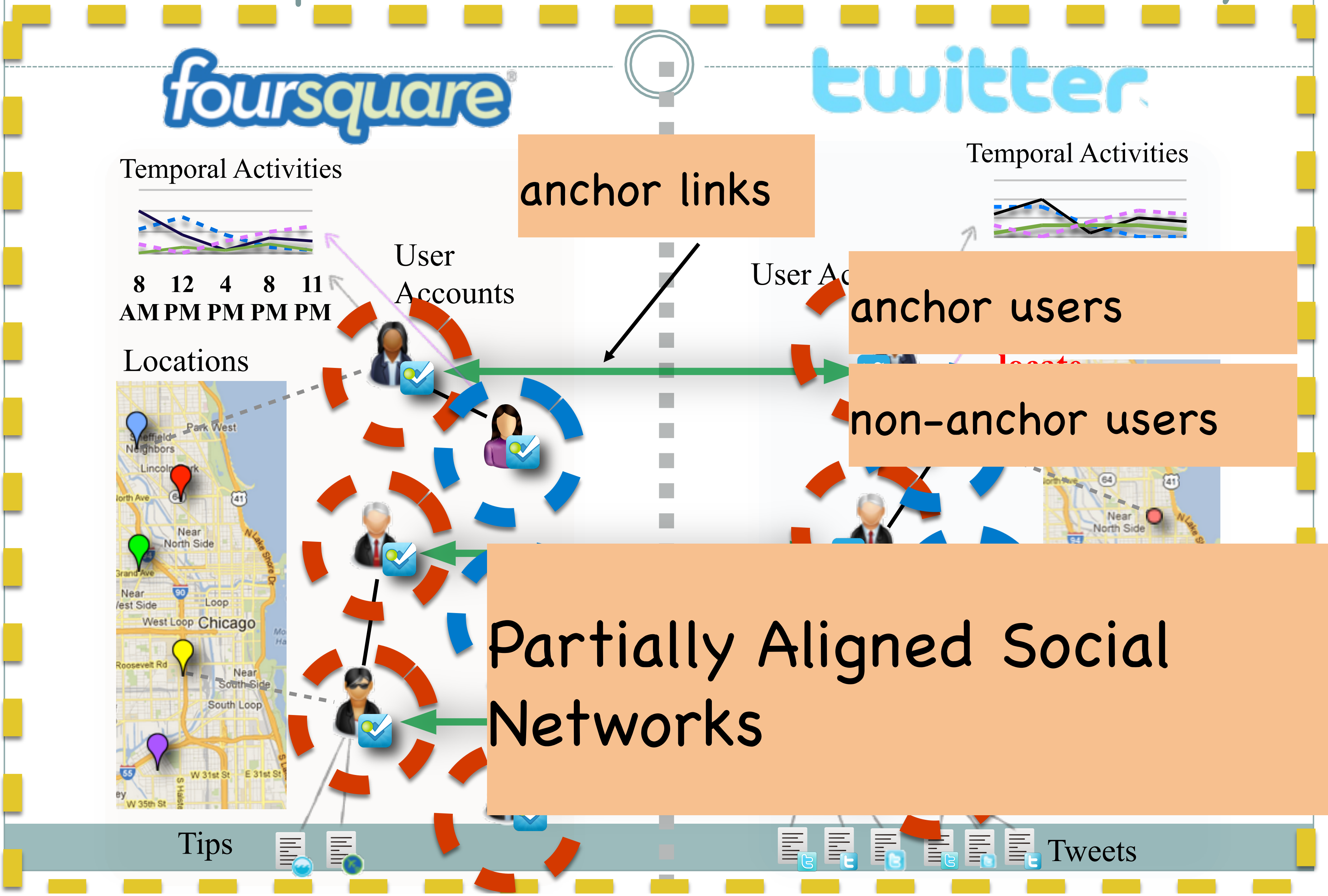
- How to improve the accuracy of friend recommendation (link prediction)?
  - Can we use information from other social networks, especially
    - Well established
    - Public available


Locations
Social Links
Temporal Activities
8 AM 12 PM 4 PM 8 PM 11 PM
Contents: Tweets
Chicago
Social Network:
Who Where What When

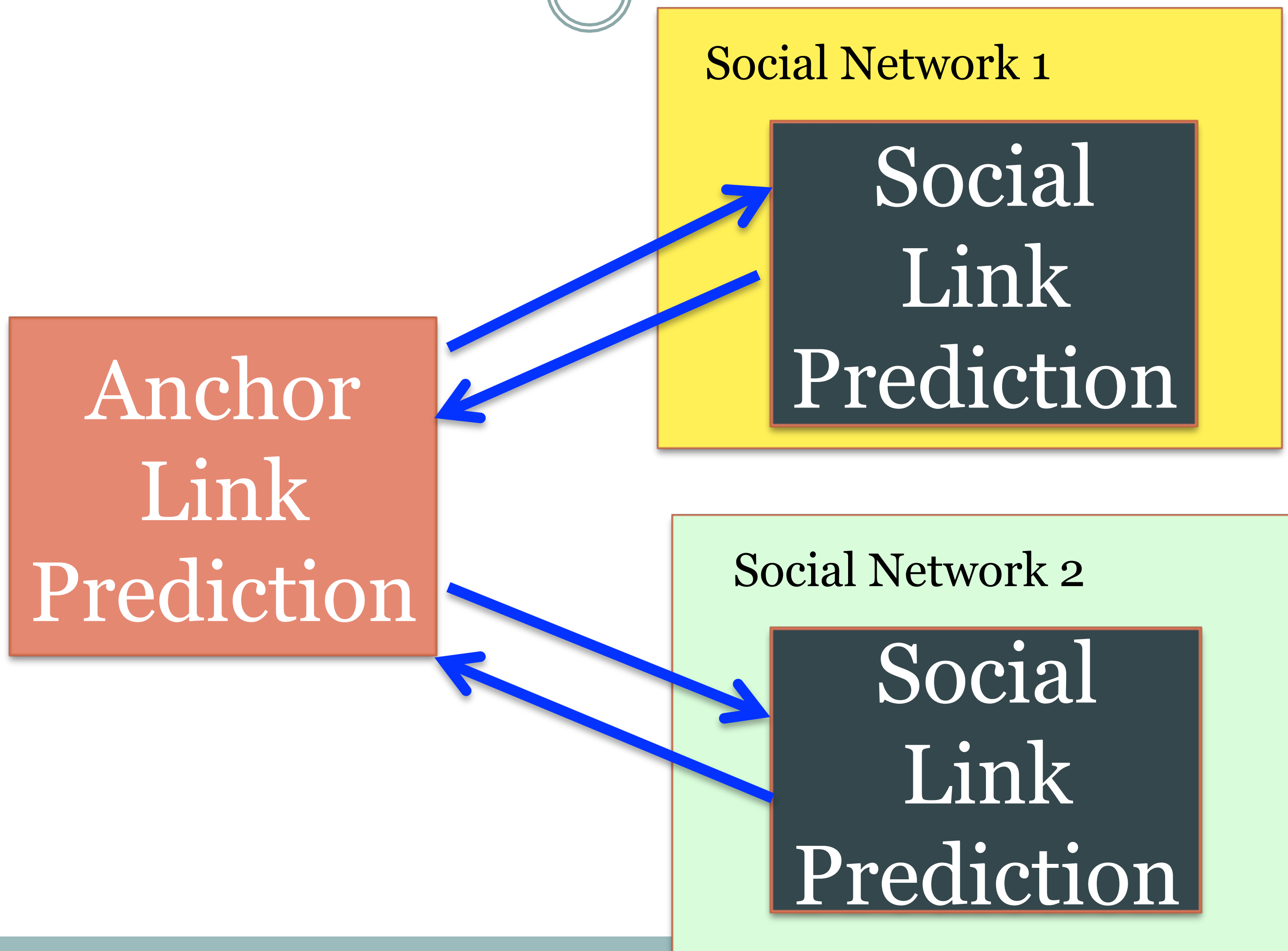
Traditional Social Link Prediction in One Single Social Network



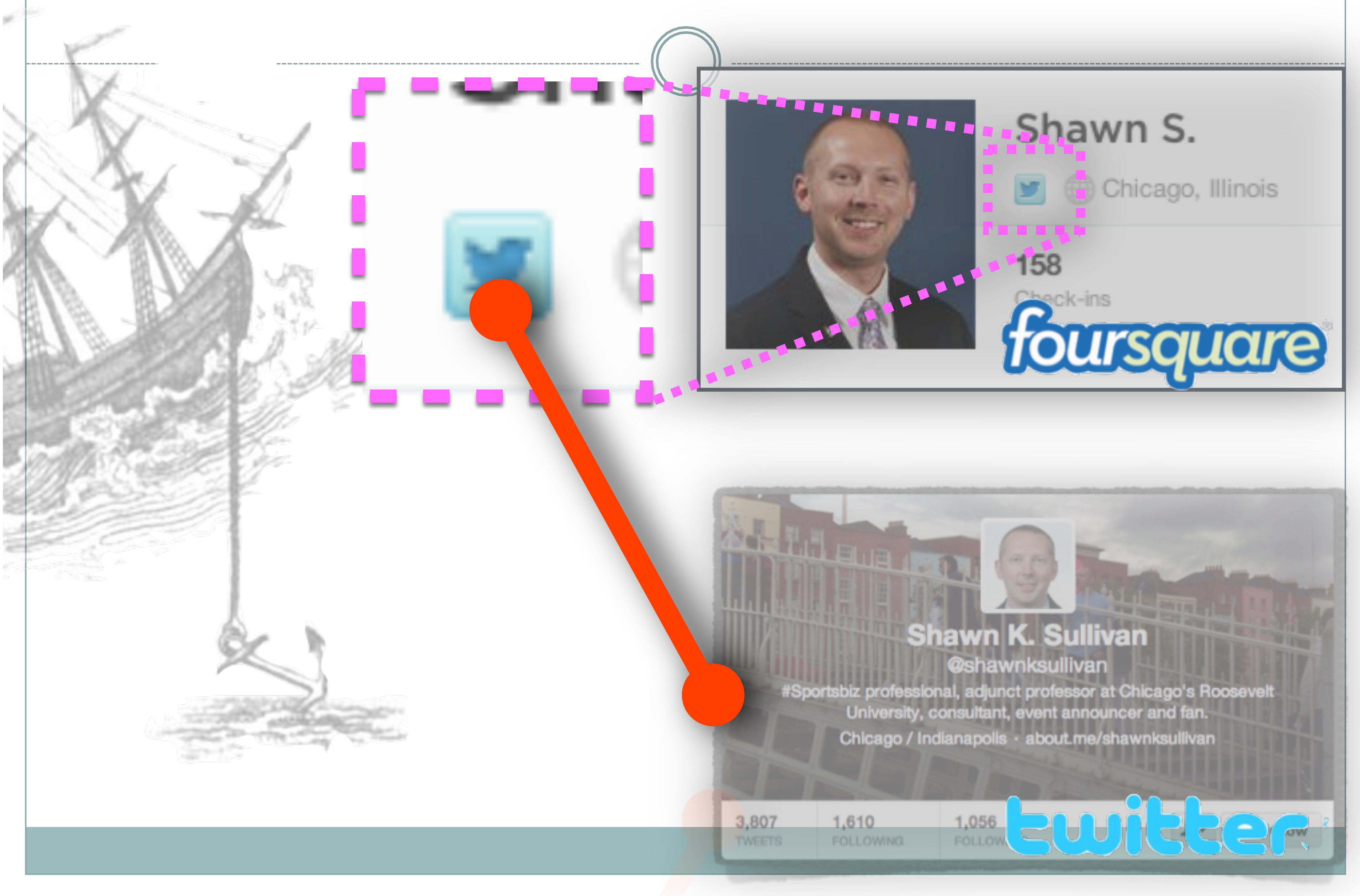
Use Multiple Social Networks Simultaneously



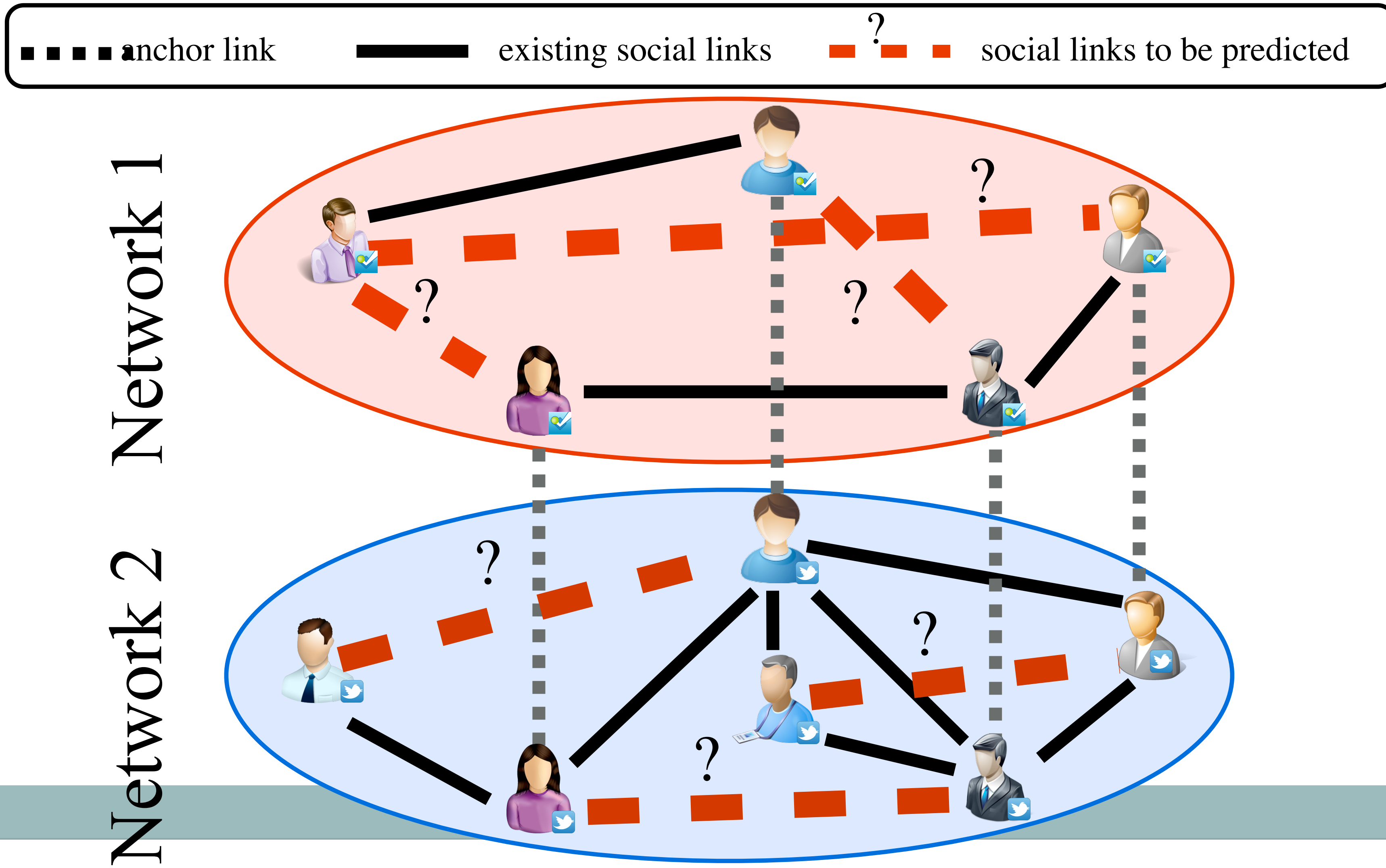
Basic Idea


Social Network 1
Social Link Prediction
Anchor Link Prediction
Social Network 2
Social Link Prediction

Anchor Links across Networks

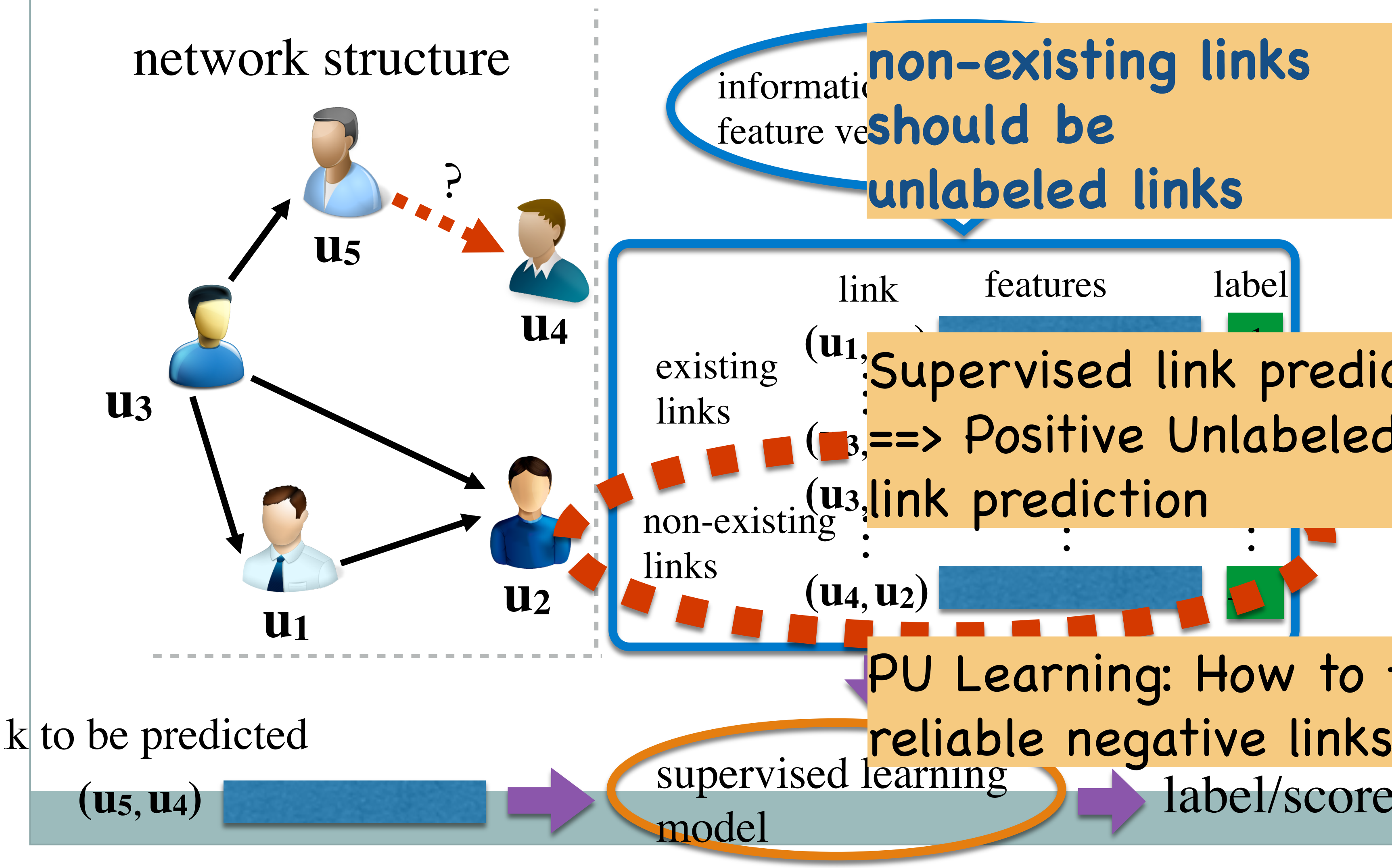


Predicting social links in multiple aligned networks simultaneously



Disadvantages of Su Prediction S

non-existing links
!=
negative links

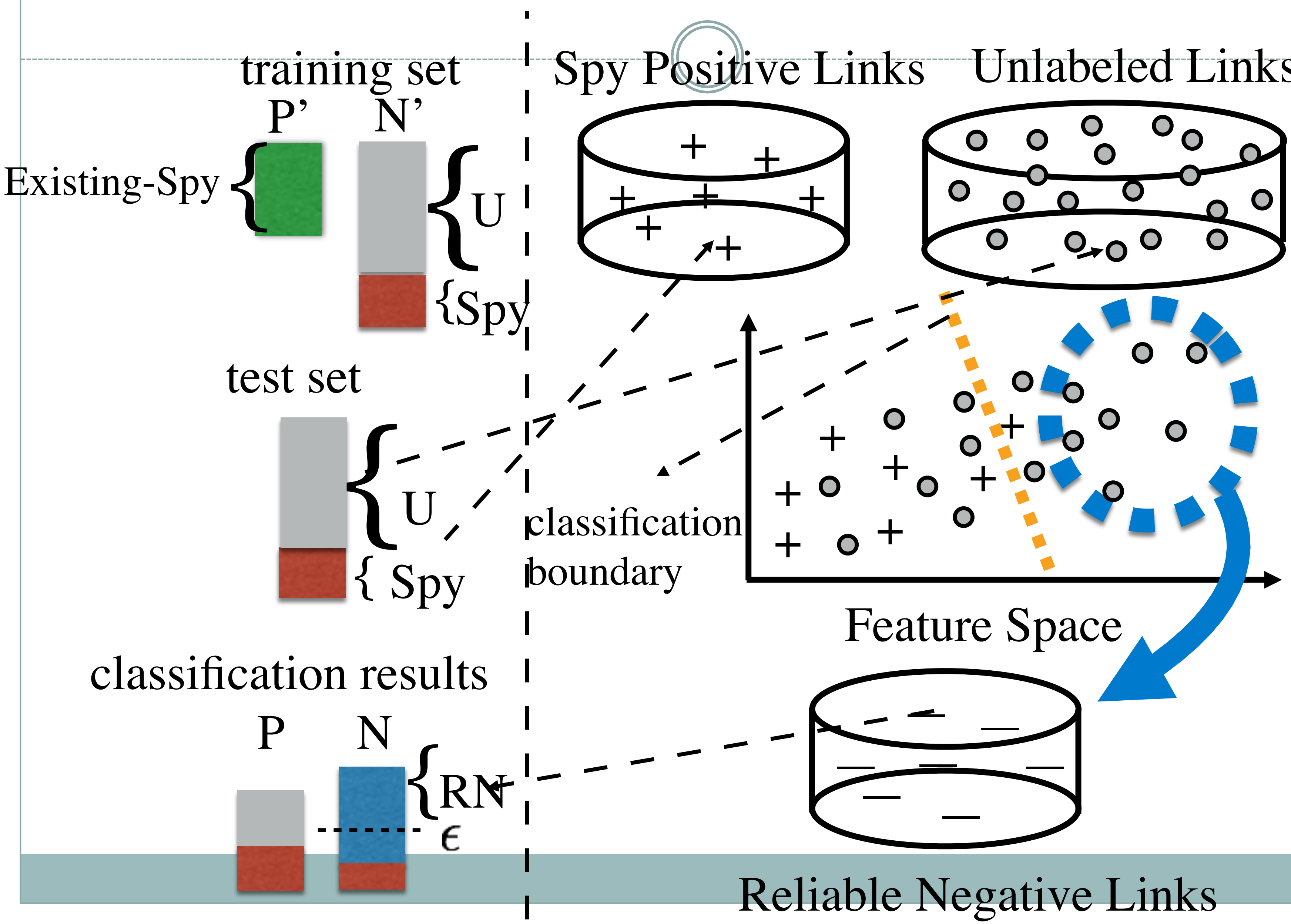


non-existing links
should be
unlabeled links

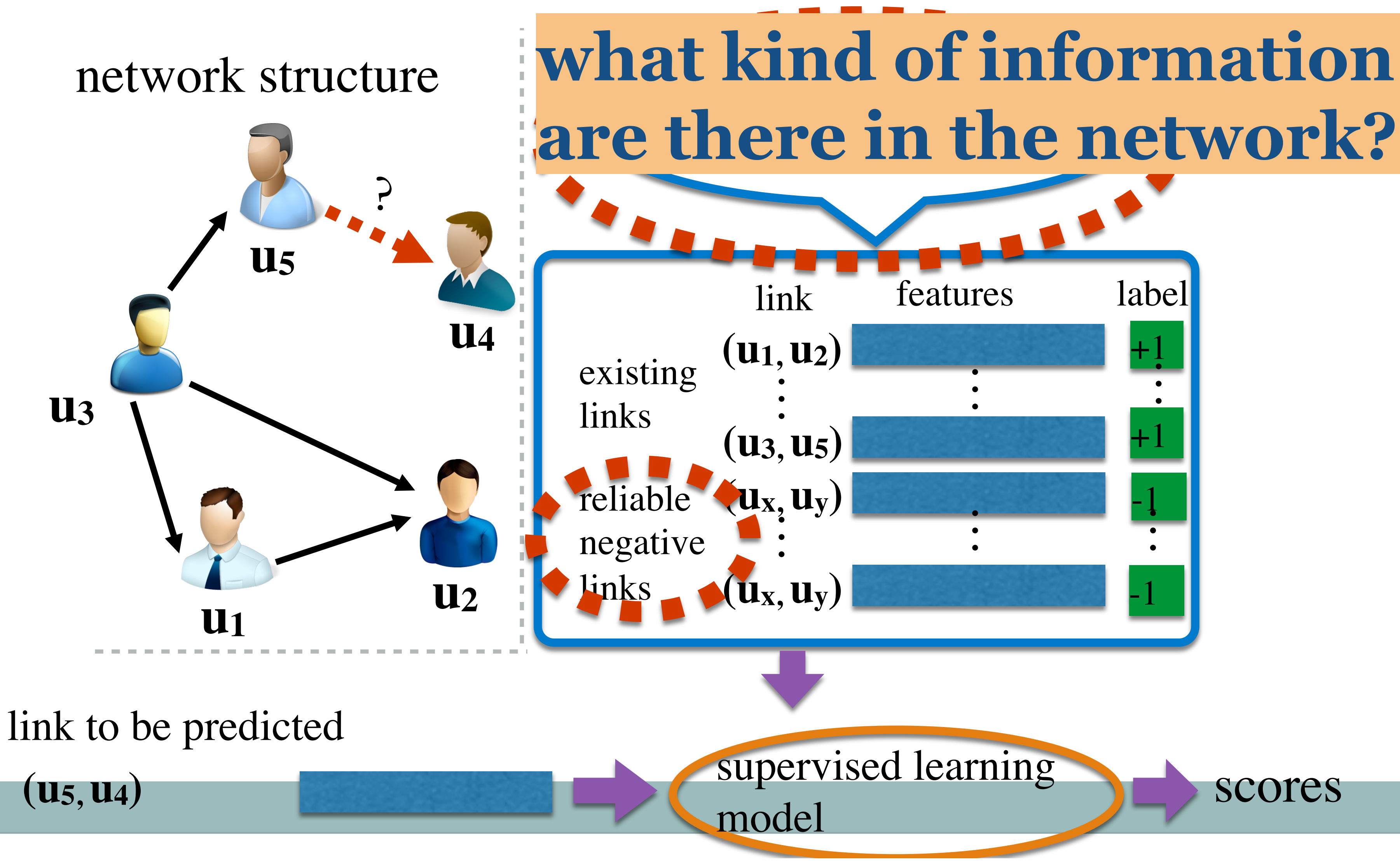
Supervised link prediction
==> Positive Unlabeled (PU)
link prediction

PU Learning: How to find
reliable negative links?

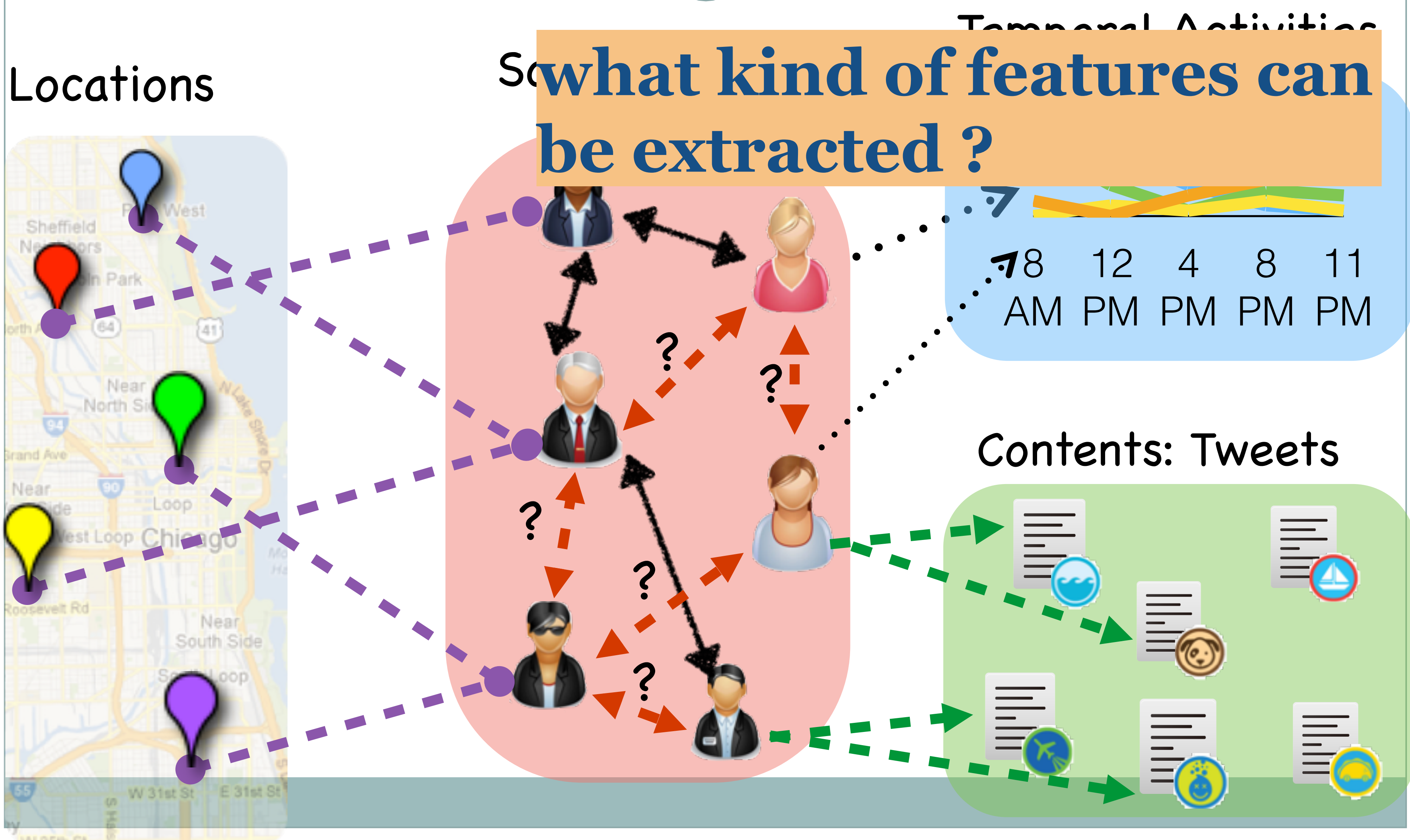
Reliable Negative Links Extraction



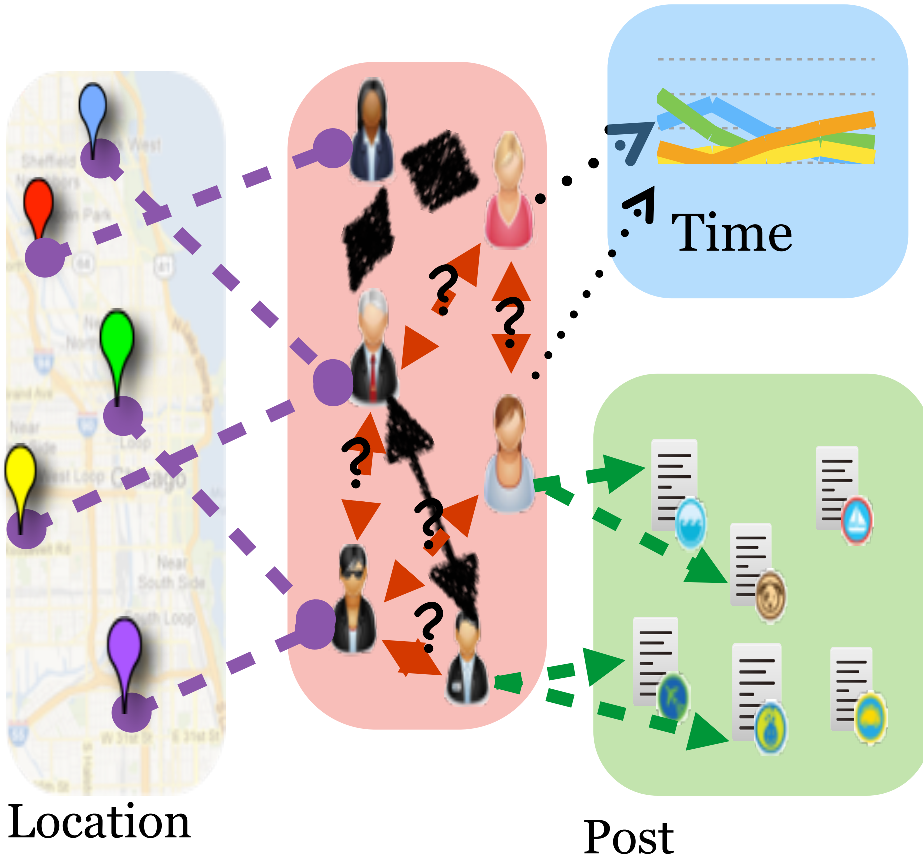
PU Link Prediction Setting



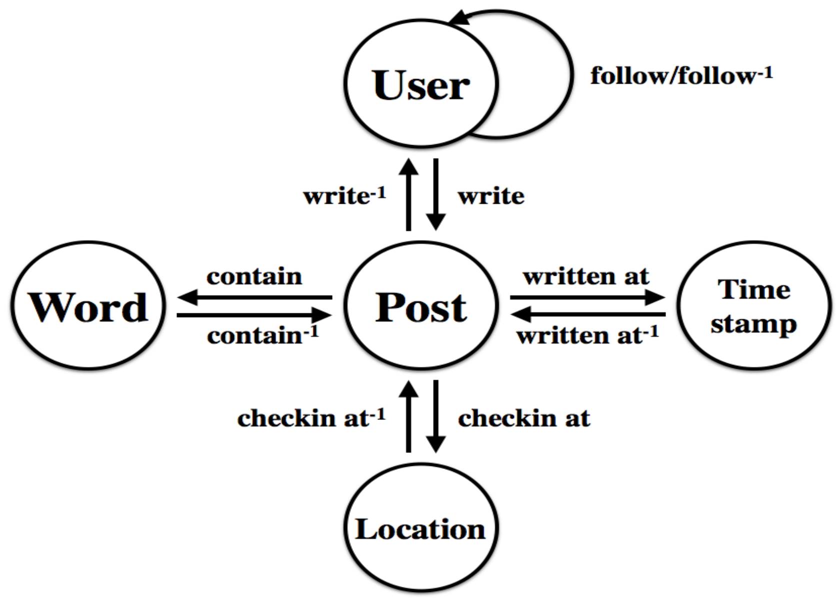
Heterogeneous Information



Network Schema

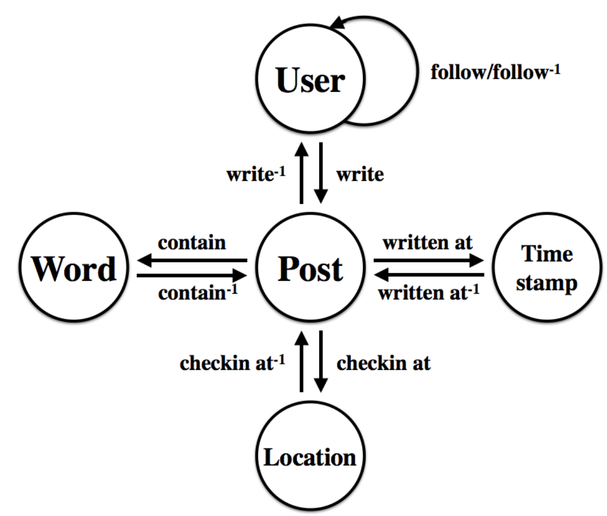


social network schema

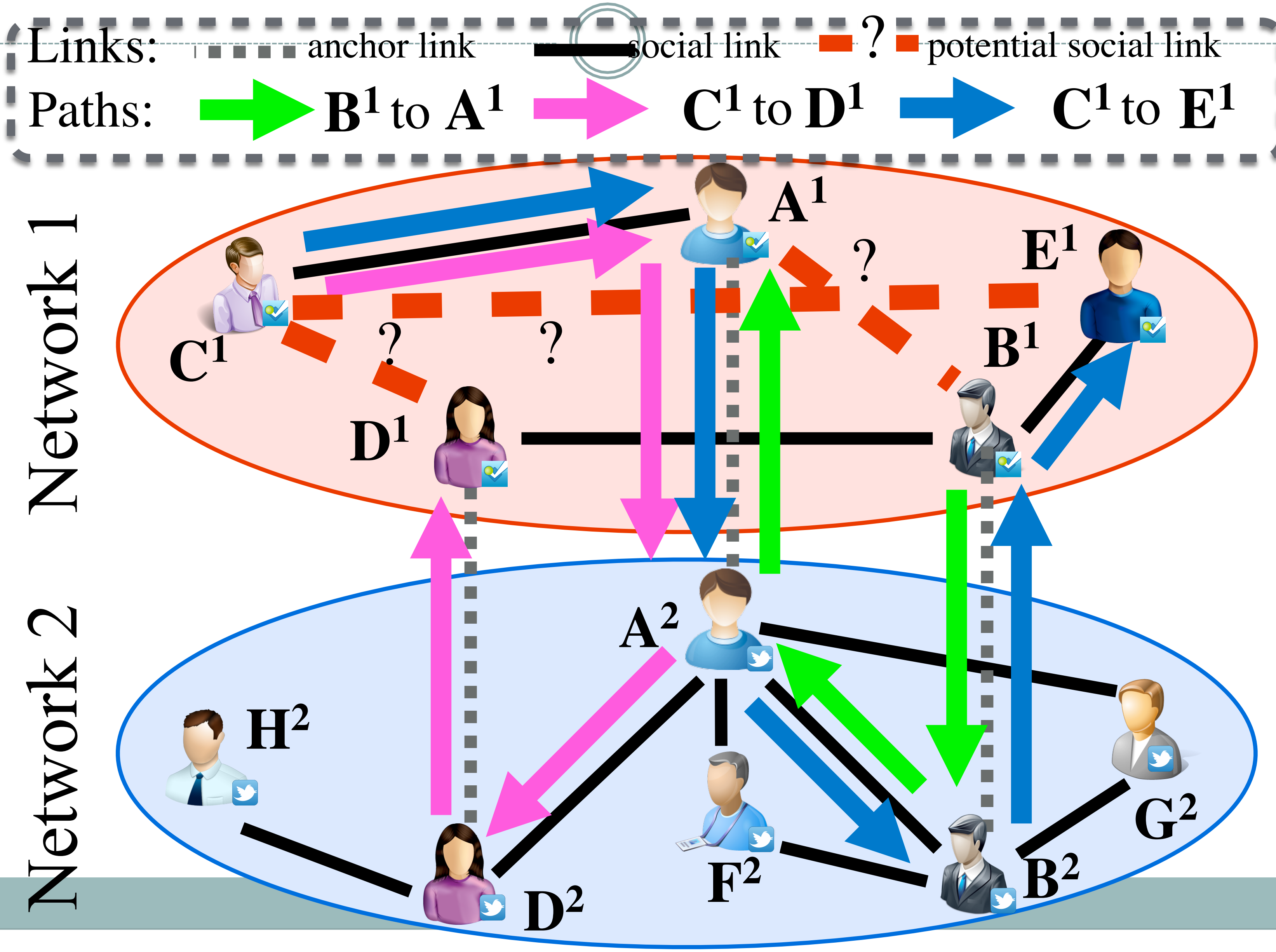


Intra-network social meta paths

- Two users U1 and U2 are considered to be similar
  - Connected through some homogeneous paths
    - U1 -> U3 <- U2 or U1 -> U3 <- U4 <-U2
  - Connect through some heterogeneous paths
    - U1 -> P1 -> **Word** <- P2 <- U2
    - U1 -> P1 -> **Location** <- P2 <- U2
    - U1 -> P1 -> **Time** <- P2 <- U2



Inter-network social meta path instances



Inter-network social meta path instances

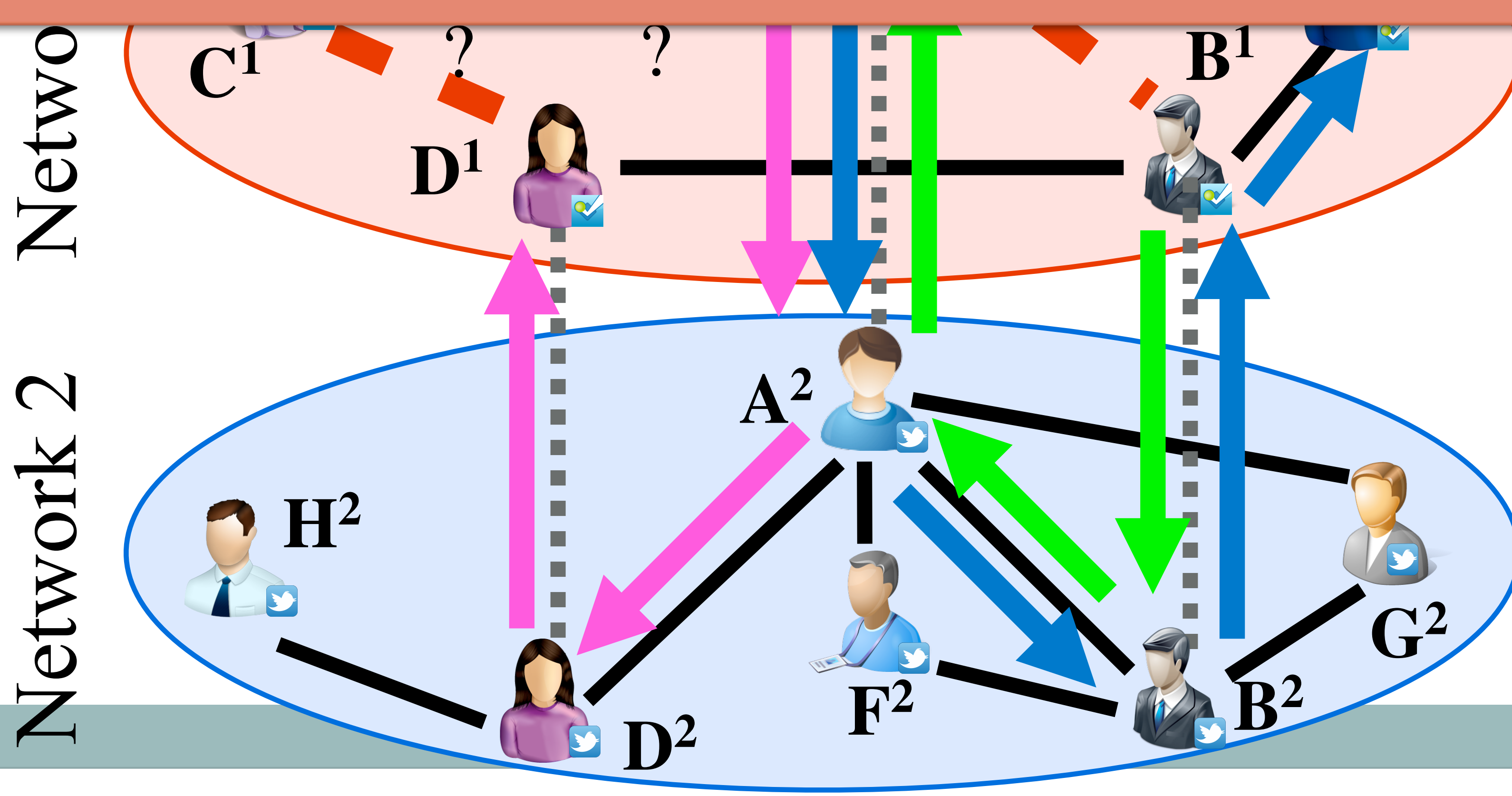
Not just social links

Also need to consider other hetrogeneous links:

U1 -> P1 -> **Word** <- P2 <- U2

U1 -> P1 -> **Location** <- P2 <- U2

U1 -> P1 -> **Time** <- P2 <- U2



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

collective link prediction is better than independent link prediction

network	measure	methods	Remaining information rates ρ_F of Foursquare. 0.1	0.2	0.3	0.4	0.5
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

Outline

- Mining heterogeneous data sources
- Fusing knowledge across multiple social networks
- Using social networks to
  - Understand customer purchase behavior
  - Predict or promote real world activities
    - Inferring the impact of social media on crowdfunding

Motivation

- Social networks can capture and contain rich information
- Most companies cannot afford to offer its own social networks to collect customer information
- Information available in public social networks may be crawled to
  - Gain better understanding of customer
  - Offer more targeted service

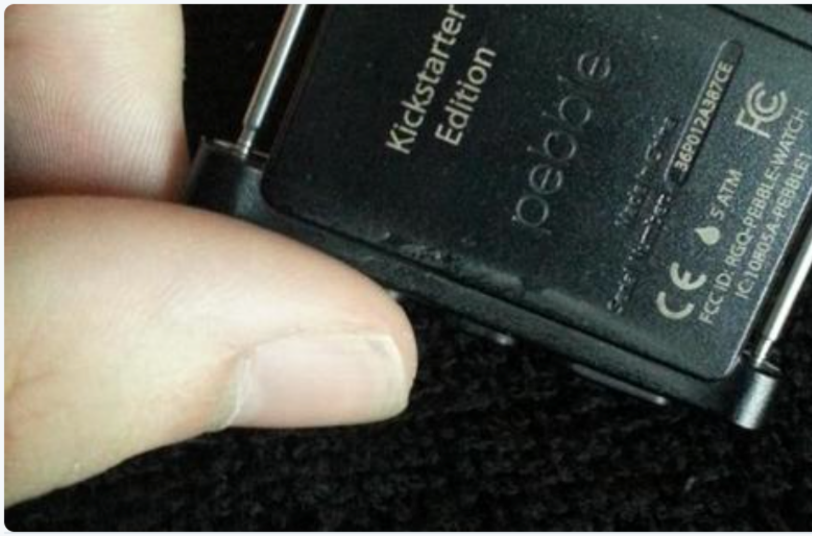
Examples

- Some real world examples of utilizing public available social network information
  - Insurance fraud detection
  - Job recruiting, Applicant screening
  - College Admission

Understanding Your Customers

- Most e-commerce companies, like Amazon, only have transaction data of their customers.
- These e-commerce companies do not own or operate social networks.
- Although the transaction data can provide the buying history, the e-commerce companies lack information on
  - The customer feedback on the product purchased
  - The friend of their customers which may show similar interests

Tyler T. @TuckertCTD · Aug 6
@**Pebble** I've had this white Kickstarter Edition since day one.
thing I've ever **bought**. #FreshHotFly



Marcus Wright @marcuswtech · Jul 3
Friend just **bought** one of these! **pebble** e-paper cherry red watch p-cr001
amzn.to/1qJMyPP #**pebble** #smartwatch

Jeremy Yancey @jeremy_yancey · May 1
Finally caved...curiosity got the better of me and I **bought** a #**Pebble**
#smartwatch Love it! ift.tt/1fDrVkm

3 1

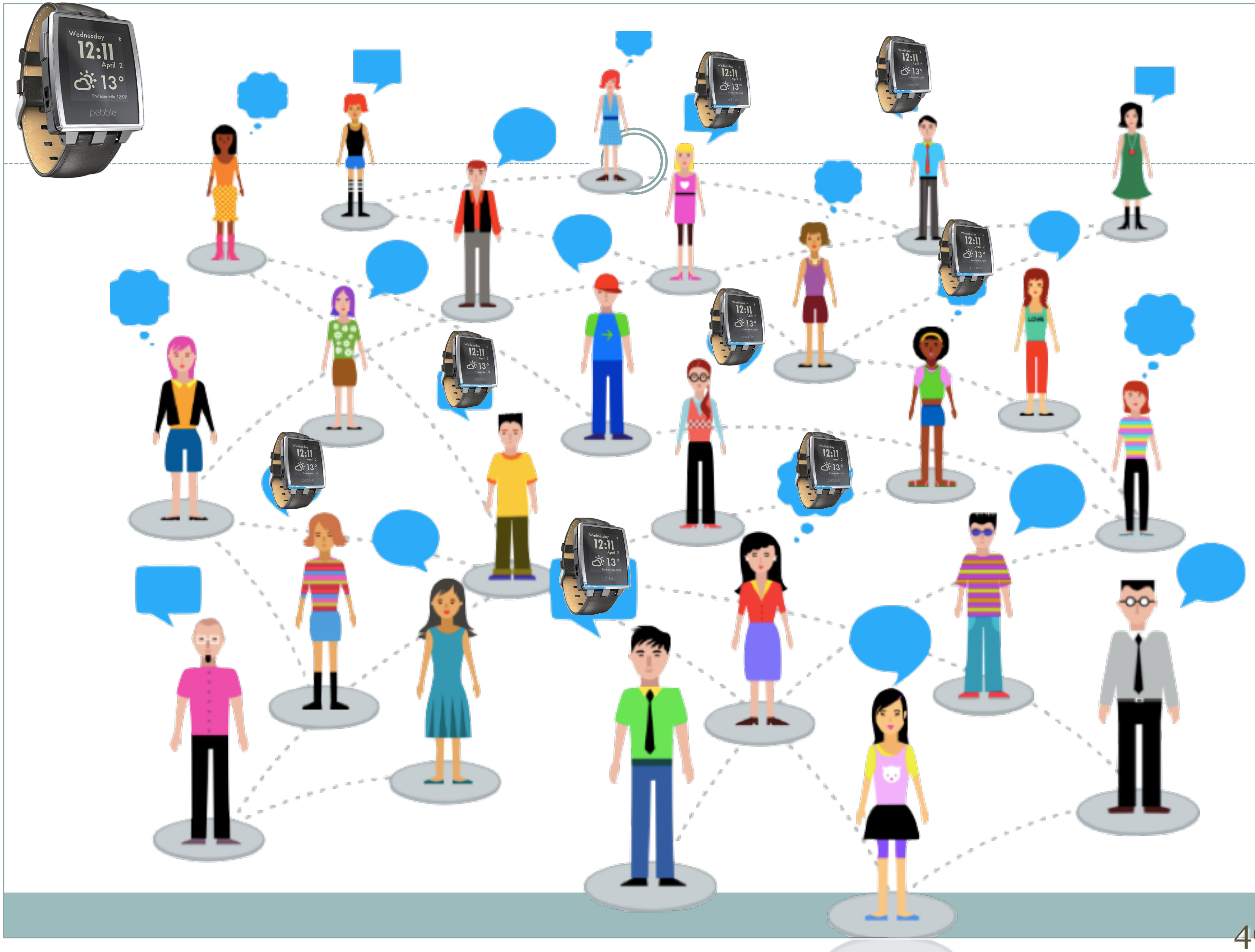
Joah Gerstenberg @therealjoahg · Aug 21
Just **bought** a drill gun on #pebbleminer :) @**pebble** #pebblesteel

3 14

Identifying Your Customers in Social Networks

Potential Applications:

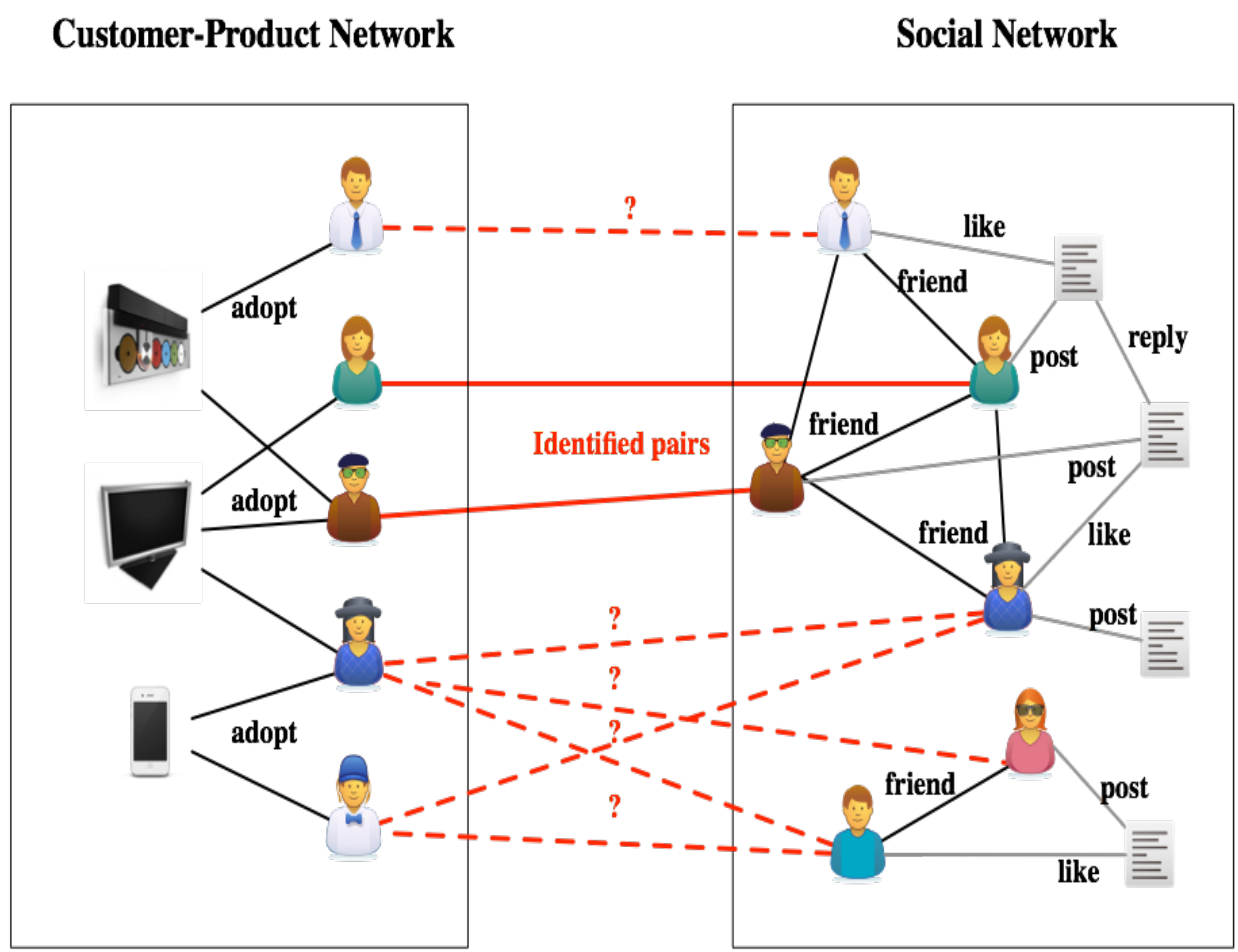
1. Analyze your customers' opinions


Wednesday
12:11
April 2
13°
pebble

Identifying Your Customers in Social Networks

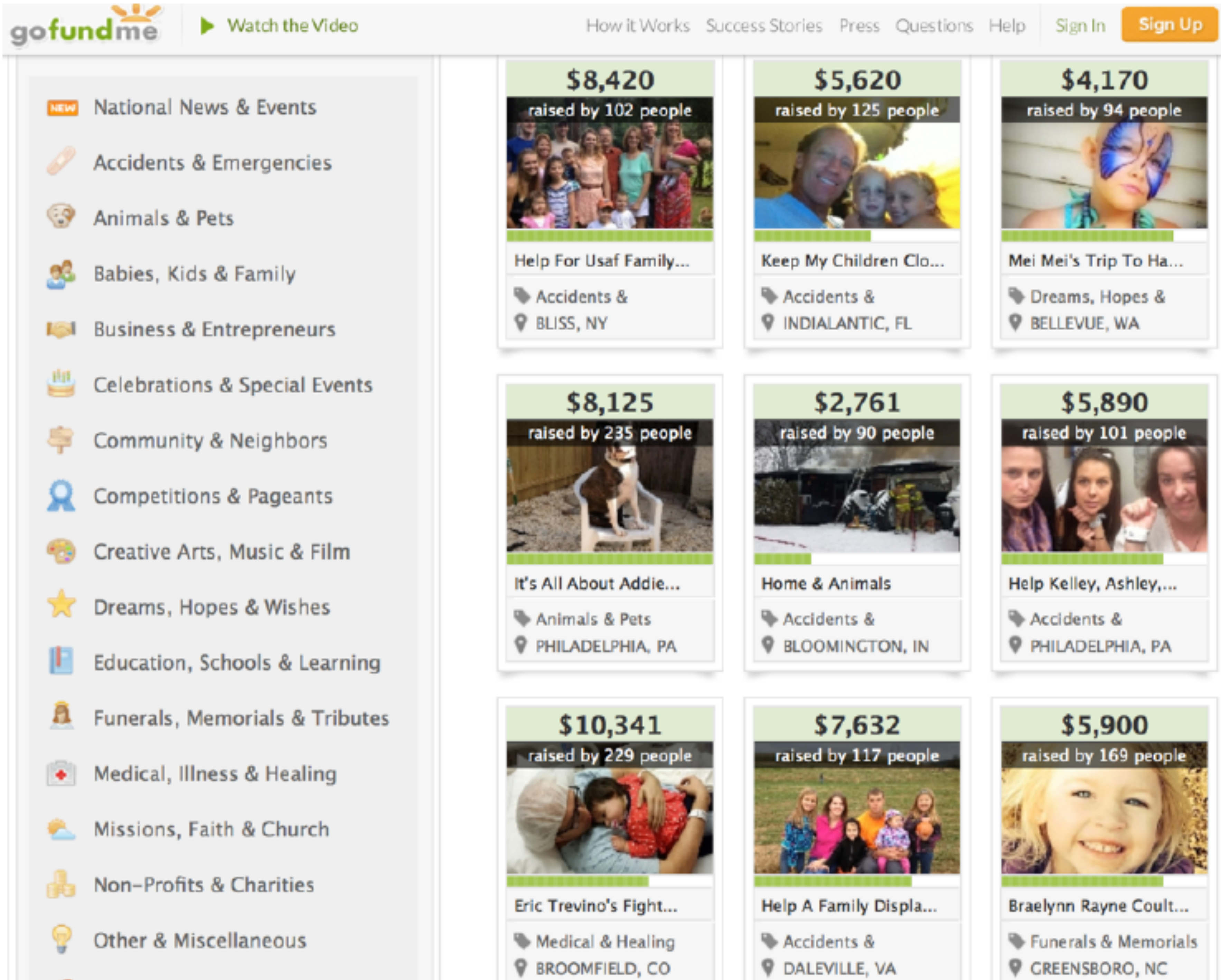
Potential Applications:

1. Analyze your customers' opinions
2. Personalized Product Recommendation
3. Discover the communities of your customers
4. Maximize product adoption


Customer-Product Network
Social Network
adopt
adopt
adopt
?
?
?
?
?
Identified pairs
like
friend
post
reply
friend
post
friend
like
post
friend
post
like

Outline

- Mining heterogeneous data sources
- Fusing knowledge across multiple social networks
- Using social networks to
  - Understand customer purchase behavior
  - Predict or promote real world activities
    - Inferring the impact of social media on crowdfunding


gofundme
Watch the Video
How it Works
Success Stories
Press
Questions
Help
Sign In
Sign Up
NEW
National News & Events
Accidents & Emergencies
Animals & Pets
Babies, Kids & Family
Business & Entrepreneurs
Celebrations & Special Events
Community & Neighbors
Competitions & Pageants
Creative Arts, Music & Film
Dreams, Hopes & Wishes
Education, Schools & Learning
Funerals, Memorials & Tributes
Medical, Illness & Healing
Missions, Faith & Church
Non-Profits & Charities
Other & Miscellaneous
$8,420
raised by 102 people
Help For Usaf Family...
Accidents &
BLISS, NY
$5,620
raised by 125 people
Keep My Children Clo...
Accidents &
INDIALANTIC, FL
$4,170
raised by 94 people
Mei Mei's Trip To Ha...
Dreams, Hopes &
BELLEVUE, WA
$8,125
raised by 235 people
It's All About Addie...
Animals & Pets
PHILADELPHIA, PA
$2,761
raised by 90 people
Home & Animals
Accidents &
BLOOMINGTON, IN
$5,890
raised by 101 people
Help Kelley, Ashley,...
Accidents &
PHILADELPHIA, PA
$10,341
raised by 229 people
Eric Trevino's Fight...
Medical & Healing
BROOMFIELD, CO
$7,632
raised by 117 people
Help A Family Displa...
Accidents &
DALEVILLE, VA
$5,900
raised by 169 people
Braelynn Rayne Coult...
Funerals & Memorials
GREENSBORO, NC

FORM 1: An affordable, professional 3D printer

by Formlabs

Home Updates 49 Backers 2,068 Comments 1,823 Cambridge, MA Technology

Funded! This project was successfully funded on Oct 26, 2012.

Share 2,793 Tweet <> Embed

2,068
backers

$2,945,885
pledged of $100,000 goal

0
seconds to go

Project by
Formlabs
Cambridge, MA
Contact me

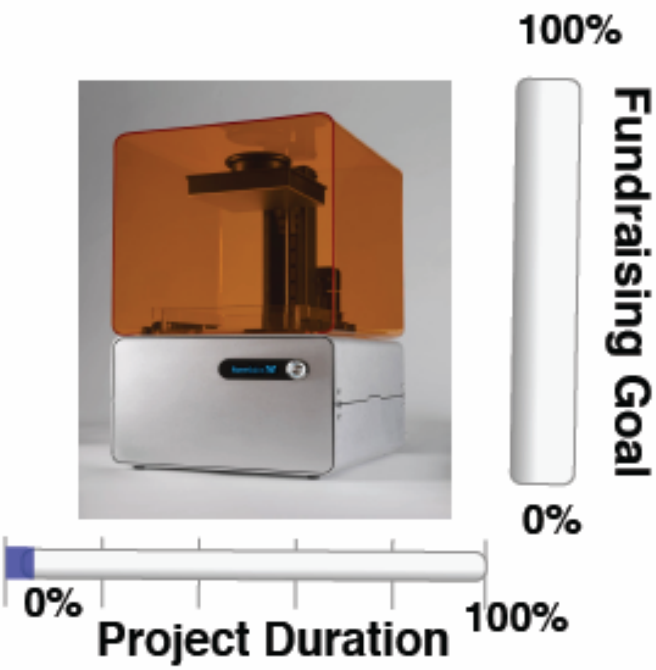
First created . 5 backed

Has not connected Facebook

Website: formlabs.com

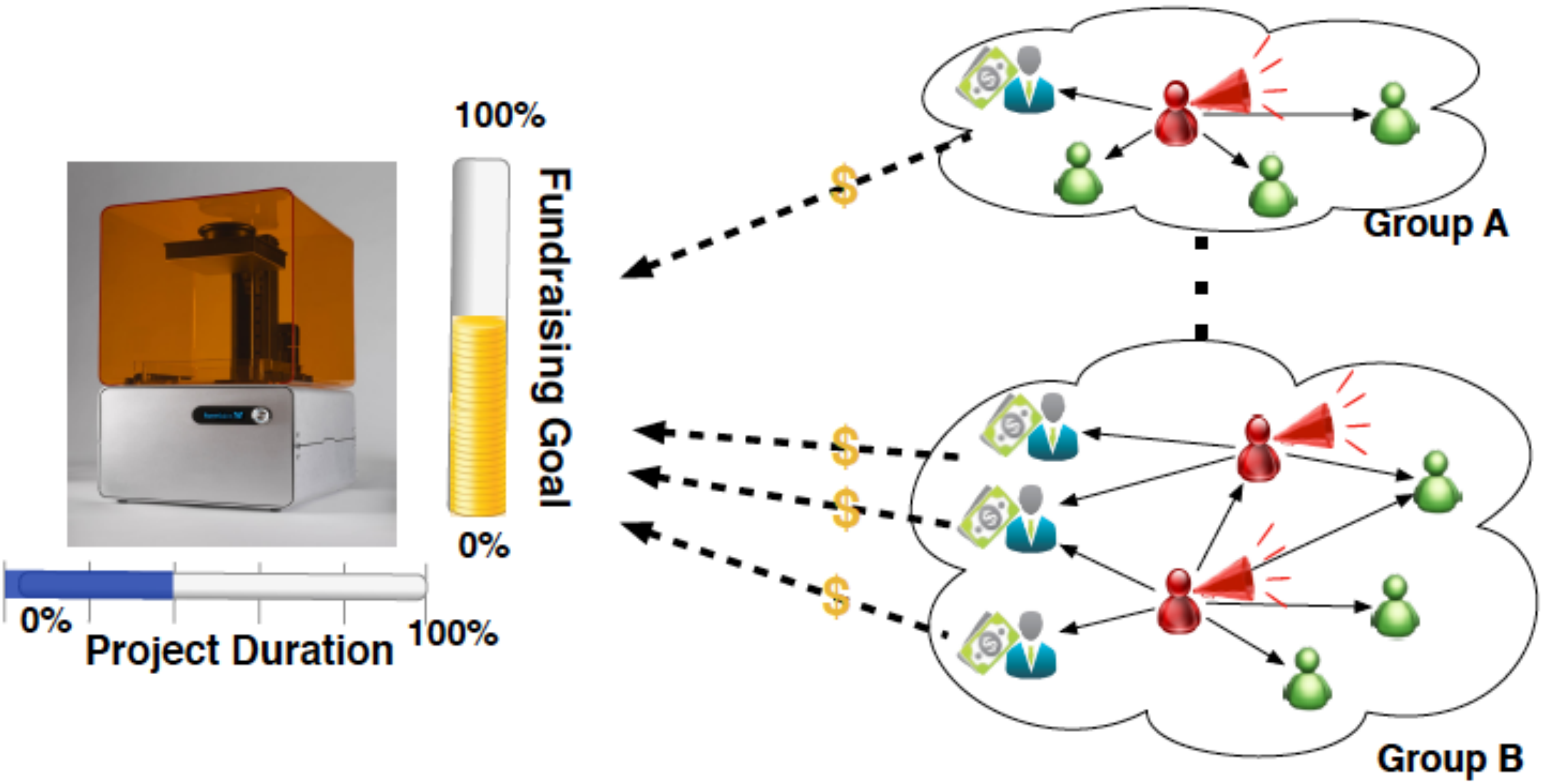
See full bio

Impact of Social Media on Crowdfunding



Unique properties: 1. fundraising goal 2. project duration

Impact of Social Media on Crowdfunding



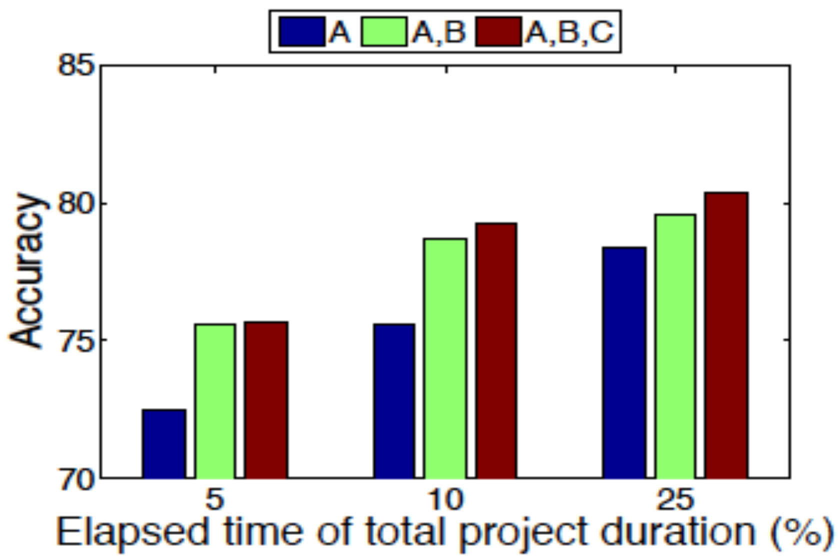
Unique properties: 1. fundraising goal 2. project duration

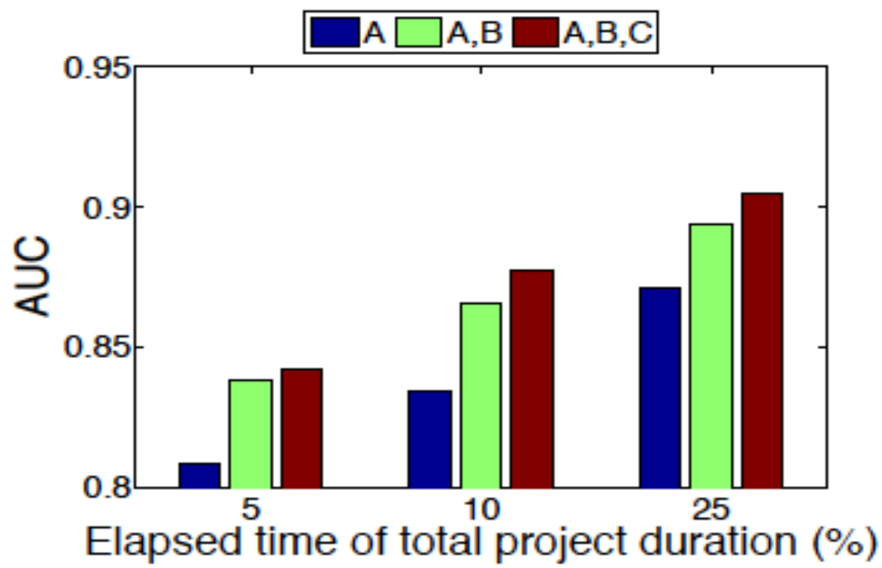
Features

- Project features:
  - funding goals, median amount of pledge options, number of backers, average amount of pledge per backer, elasped days since launched
- Social activity features
  - number of tweets, number of promoters, number of patrons, number of uniquely mentioned users, fraction of promoters from external sources
- Social structure features:
  - average number of followers of promoters, number of edges, diameter, number of connected components, number of triads, global clustering coefficient

Experiments (Early Prediction)

(2) Predict whether a project will succeed or fail (within 25% of project duration)

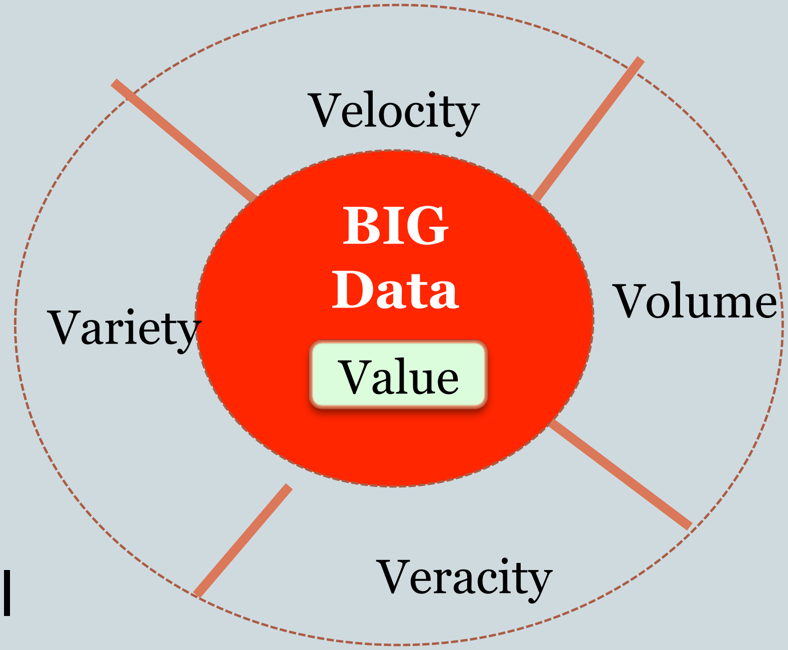




A: Project features; B:Social activity features; C: Social structure feature

Summary

- Mining heterogeneous data sources
- Fusing knowledge across multiple social networks
- Using social networks to
  - Understand customer purchase behavior
  - Predict or promote real world activities
    - Inferring the impact of social media on crowdfunding



References

- Y. Sun, J. Han, X. Yan, P.S. Yu, and T. Wu, "*PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks*", PVLDB, 2011.
- X. Kong, P.S. Yu, Y. Ding, and D. Wood "*Meta Path-based Collective Classification on Heterogeneous Information Networks*", ACM CIKM, 2012.
- X. Kong, B. Cao, and P.S. Yu, "*Multi-Label Classification by Mining Label and Instance Correlations from Heterogeneous Information Networks*", ACM KDD, 2013.
- Y. Sun, B. Norick, J. Han, X. Yan, P.S. Yu, and X. Yu, "*PathSelClus: Integrating Meta-Path Selection with User-Guided Object Clustering in Heterogeneous Information Networks*", ACM Trans. Knowledge Discovery from Data,, 2013.
- J. Zhang, and P.S. Yu, "*Meta-path based Multi-network Collective Link Prediction*", ACM KDD 2014.
- C. Shi, X. Kong, Y. Huang, P.S. Yu, and B. Wu "*HeteSim: A General Framework for Relevance Measure in Heterogeneous Networks*", IEEE Trans. on Knowledge and Data Engineering,2014.
- C. Shi, R. Wang, Y. Li, P.S. Yu, and B. Wu, "*Ranking-based Clustering on General Heterogeneous Information Networks by Network Projection*", ACM CIKM, 2014.
- B. Cao, X. Kong, and P.S. Yu, "*Collective Prediction of Multiple Types of Links in Heterogeneous Information Networks*", IEEE ICDM. 2014.

References

- C. Lu, S. Xie, X. Kong, and P.S. Yu, "*Inferring the Impacts of Social Media on Crowdfunding*", ACM WSDM 2014.
- X. Kong, J. Zhang, and P.S. Yu, "*Inferring Anchor Links across Multiple Heterogeneous Social Networks*", ACM CIKM 2013.
- J. Zhang, X. Kong, and P.S. Yu, "*Predicting Social Links for New Users across Aligned Heterogeneous Social Networks*" IEEE ICDM, 2013.
- J. Zhang, X. Kong, and P.S. Yu, "*Transferring Heterogeneous Links across Location-Based Social Networks*", ACM WSDM 2014.
- S. Wu, H. Chien, K. Lin, and P.S. Yu, "*Learning the Consistent Behavior of Common Users for Target Node Prediction across Social Networks*", ICML 2014.