



# COSNET: Connecting Heterogeneous Social Networks with Local and Global Consistency

Yutao Zhang<sup>†</sup>, Jie Tang<sup>†</sup>, Zhilin Yang<sup>†</sup>, Jian Pei<sup>#</sup>, and Philip S. Yu<sup>\*</sup>

<sup>†</sup>Tsinghua University



清华大学

Tsinghua University



<sup>\*</sup>University of Illinois at Chicago





# AMiner II (ArnetMiner)

□ Academic Social Network Analysis and Mining system—AMiner (<http://aminer.org>)

- Online since 2006
- >38 million researcher profiles
- >100 million publications
- >241 million requests
- >12.35 Terabyte data
- 100K IP access from 170 countries per month
- 10% increase of visits per month

□ Deep analysis, mining, and search

The screenshot displays the ArnetMiner interface. At the top, a researcher profile for Jiawei Han is shown, including a photo, basic information (Position: Professor, Affiliation: Department of Computer Science, University of Illinois at Urbana-Champaign), and statistics (H-index: 96, #Papers: 553, #Citations: 55885). Below this is a 'Bio' section and a 'Research Interest' section listing 'Efficient Mining', 'Spatial Data Mining', and 'Frequent Pattern Mining'. To the right, sections for 'See Others', 'Expertise', and 'Conference' are visible. The middle section shows search results for 'data mining', listing profiles for Jiawei Han, Philip S. Yu, Mohammed Javed Zaki, Christos Faloutsos, Jian Pei, H. Mannila, and Charu Aggarwal. The bottom section features a 'Social Graph' visualization where nodes represent researchers and edges represent relationships like 'coauthor', 'collaborator', and 'mentee'. A 'Publications' section shows a timeline of papers from 1990 to 1994.

# Knowledge Acquisition from the Web

(ACM TKDD, WWW'12, ISWC'06, ICDM'07, ACL'07)



**Ruud Bolle**

Office: 1S-D58

**Contact Information**

Office: 1S-D58  
Letters: IBM T.J. Watson Research Center  
P.O. Box 704  
Yorktown Heights, NY 10598 USA  
Packages: IBM T.J. Watson Research Center  
19 Skyline Drive  
Hawthorne, NY 10532 USA  
Email: [bolle@us.ibm.com](mailto:bolle@us.ibm.com)

**Educational history**

Ruud M. Bolle was born in Voorburg, The Netherlands in 1977. He received a Degree in Analog Electronics in 1977 and the Master's Degree in Electrical Engineering in 1980, both from Delft University of Technology, Delft, The Netherlands. In 1983 he received the Master's Degree in Applied Mathematics and in 1984 the Ph.D. in Electrical Engineering from Brown University, Providence, Rhode Island. In 1984 he became a Research Staff Member at the IBM Thomas J. Watson Research Center in the Artificial Intelligence Department of the Computer Science Department. In 1988 he became manager of the newly formed Exploratory Computer Vision Group which is part of the Math Sciences Department.

Currently, his research interests are focused on video database indexing, video processing, visual human-computer interaction and computer vision.

**Academic services**

Ruud M. Bolle is a Fellow of the IEEE and the AIPR. He is Area Editor of Computer Vision and Image Understanding and Associate Editor of Pattern Recognition. Ruud M. Bolle is a Member of the IBM Academy of Technology.

Author's  
and in  
code  
tson  
ce  
np

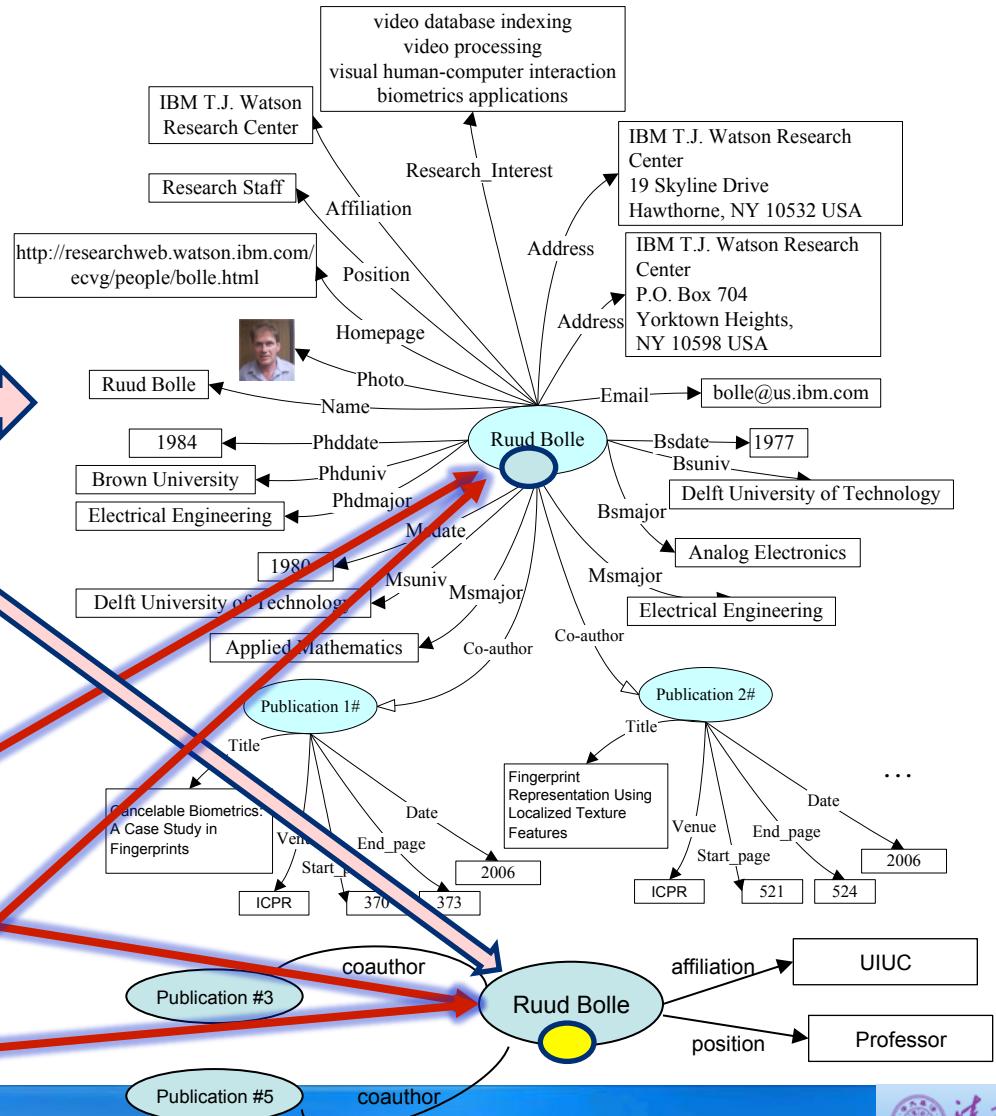
Ruud

puter  
Ruud

DBLP: Ruud Bolle

**Publications**

		2006
50	EE	Nalini K. Ratha, Jonathan Connell, Ruud M. Bolle, Sharat Chikkerur: Cancellable Biometrics: A Case Study in Fingerprints. ICPR (4) 2006: 370-373
49	EE	Sharat Chikkerur, Sharath Pankanti, Alan Jea, Nalini K. Ratha, Ruud M. Bolle: Fingerprint Representation Using Localized Texture Features. ICPR (4) 2006: 521-524
43	EE	Arun Hampapur, Ying-li Tian, Lisa Brown, Sharath Pankanti, Ruud M. Bolle: Appearance models for occlusion handling. Image Vision Comput. 24(11): 1233-1243 (2006)
2005		
47	EE	Ruud M. Bolle, Jonathan H. Connell, Sharath Pankanti, Nalini K. Ratha, Andrew W. Senior: The Relation between the ROC Curve and the CMC. AutoID 2005: 15-20
46	EE	Sharat Chikkerur, Venu Govindaraju, Sharath Pankanti, Ruud M. Bolle, Nalini K. Ratha: Novel Approaches for Minutiae Verification in Fingerprint Images. WACV. 2005: 111-116
...		





# Researcher Profile Database<sup>[1]</sup>

**ArnetMiner** Home Conference Collaborator Geo Search Topics Download Admin More Account Welcome jietang FOAF Follow

**Search Experts** Search

**Jiawei Han** FOAF Follow

Position: Professor  
Affiliation: Department of Computer Science, University of Illinois at Urbana-Champaign  
Address: 140 N Goodwin Avenue, Urbana, IL 61801, USA  
Phone: (217) 333-6932  
Fax: (217) 265-6494  
Email: han@cs.uiuc.edu  
Links: [ORCID](#) [Google Scholar](#)

**STATISTIC** ?  
H-index: 96 Uptrend: 30.46 Diversity: 0.71  
#Papers: 553 Activity: 32.04 Sociability: 726.64  
#Citations: 55885 Longevity: 26 [More Statistics...](#)

**Bio**  
Jiawei Han is computer scientist who specializes in research on Data Mining. He was the 2009 winner of the McDowell Award, the highest technical award made by IEEE. He is currently a professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign. Previously he was a professor in the School of Computing Science at Simon Fraser University. He is an ACM fellow and an IEEE fellow.

**Research Interest**  
[Efficient Mining](#) [Spatial Data Mining](#) [Frequent Pattern Mining](#) [EDIT INTEREST](#)

**Education**  
Phd University [EDIT](#)

**M. I. Jordan** FOAF Follow

ALIAS: Michael I. Jordan, Michael Jordan, Michael Irwin Jordan

Position: Professor  
Affiliation: Department of EECS Department of Statistics University of California, Berkeley  
Address: University of California, Berkeley EECS Department 731 Soda Hall #1776 Berkeley, CA 94720-1776  
Phone: (510) 642-3806  
Fax: (510) 642-5775  
Email: jordan@stat.berkeley.edu  
Links: [ORCID](#) [Google Scholar](#) [EDIT PROFILE](#)

**STATISTIC** ?  
H-index: 75 Uptrend: 7.2 Diversity: 0.03  
#Papers: 242 Activity: 11.12 Sociability: 331.69  
#Citations: 44312 Longevity: 23 [More Statistics...](#)

**H. Garcia** FOAF Follow

ALIAS: H. Garcia Molina, H. Garcia-Molina, Hector Garcia Molina, Hector Garcia Molina

Position: Professor  
Affiliation: Departments of Computer Science and Electrical Engineering, Stanford University  
Address: Department of Computer Science Stanford University Gates Hall 4A, Room 434 Stanford, CA 94305-9040 USA  
Phone: (650) 723-0885  
Fax: (650) 725-2588  
Email: hector@cs.stanford.edu

**See Others:**  
Andreas Paepcke Cauthor-Count: 32 H-index: 37  
Jennifer Widom Cauthor-Count: 24 H-index: 79  
D. Barbara Cauthor-Count: 26 H-index: 0

**Expertise:**  
Data (115) Database Systems (60) Time Systems / Automated Ware Test Data (30) Mining (23) Mobile Robot / Hybrid Control (22) General library / Information Access [EDIT PROFILE](#)

**Conference:**  
KDD (49) ICDE (40) IEEE Trans. Knowl. Data Eng. (36) VLDB (32) SIGMOD Conference (32) VLDB (21)

**ArnetMiner** Home Conference Collaborator

**Search Experts** Search

**Scott** FOAF Follow

Position:   
Affiliation:   
Address:   
Phone:   
Links: [ORCID](#) [Google Scholar](#)

**STATISTIC** ?  
H-index: 96 Uptrend: -4.04 Diversity: 0.22  
#Papers: 195 Activity: 4.86 Sociability: 407.19  
#Citations: 57908 Longevity: 25 [More Statistics...](#)

**Expertise:**  
Wireless network / End-to-end Routing Behavior (80) ATM Networks (21)

**Research Interest**  
[Database Systems](#) [Data Management](#) [Data Warehousing](#) [EDIT INTEREST](#)

**Monterrey, Mexico, in 1974. From Stanford University, Stanford, California, he received in 1975 a MS in electrical engineering and a PhD in computer science in 1979. He holds an honorary PhD from ETH Zurich (2007). Garcia-Molina is a Fellow of the Association for Computing Machinery and of the American Academy of Arts and Sciences; is a member of the National Academy of Engineering; received the 1999 ACM SIGMOD Innovations Award; is a Venture Advisor for Onset Ventures, and is a member of the Board of Directors of Oracle.**

[1] J. Tang, L. Yao, D. Zhang, and J. Zhang. A Combination Approach to Web User Profiling. ACM Transactions on Knowledge Discovery from Data (TKDD), (vol. 5 no. 1), Article 2 (December 2010), 44 pages.

# Is this Enough?

AMiner

Whatever comes to your mind
Search

Home
|
 Profile

Mohak Shah

Centre for Intelligent Machines McGill University

Postdoctoral Fellow

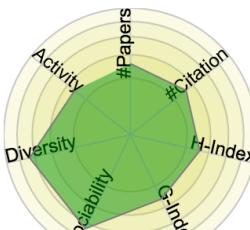
1-514-398-8702

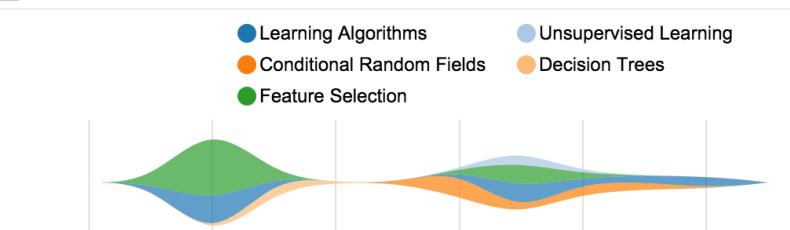
mohak [at] cim [dot] mcgill [dot] ca

<http://www.cim.mcgill.ca/~mohak/Site/Home.html>

Upload
 Update

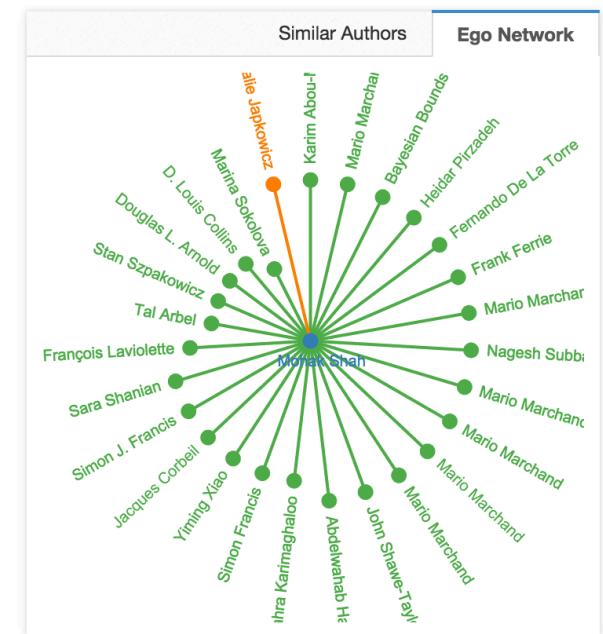
in
 Research Interests





Similar Authors

Ego Network



5


 清华大学  
Tsinghua University

# Required semantics are distributed in multiple sources



## LinkedIn

LinkedIn profile of Mohak Shah:

**Mohak Shah**  
Sr Manager- Data Science at Bosch; General Chair, KDD 2016  
Palo Alto, California | Research

Current: Bosch, ACM SIGKDD 2016  
Previous: GE Software, Accenture, McGill University  
Education: McGill University

Send Mohak InMail

500+ connections

Contact Info

Experience

Sr Manager, Data Science  
Bosch  
April 2014 – Present (1 year 5 months) | Palo Alto, California

General Chair  
ACM SIGKDD 2016

Machine Learning Lab Manager  
GE Software  
June 2012 – April 2014 (1 year 11 months) | San Ramon, California

Research Manager, Data Mining and Machine Learning  
Accenture  
February 2011 – June 2012 (1 year 5 months) | Greater Chicago Area

BOSCH  
Invented for life

GE

accenture

## Videolectures

videolectures.net  
exchange ideas & share knowledge

World Summit Award



HOME • BROWSE LECTURES • PEOPLE • CONFERENCES • ACADEMIC ORGANISATIONS •

## Mohak Shah

homepage: <http://www.mohakshah.com/Site/Home.html>  
search externally: Google Scholar, Springer, CiteSeer, Microsoft Academ

## Lecture:



lecture  
Generalized Agreement Statistics over Fixed Group of Experts  
as author at Sessions, together with: Data & Web Mining Lab (produced by), 59 views



# Identity Linking

- Identifying users from multiple heterogeneous networks and integrating semantics from the different networks together.

LinkedIn

LinkedIn profile of Tom Dietterich:

**Tom Dietterich**  
Professor at Oregon State University and Director of Intelligent Systems Research  
Covallis, Oregon Area | Higher Education  
Current: Bigt!, Inc. Oregon State University  
Previous: Dado Corp. an EMC Company, Smart Desktop Division of Pi Corporation, MusicBrands, Inc.  
Education: Stanford University

284 connections  
414 contacts  
[www.linkedin.com/pub/tom-dietterich/0/97/828/](http://www.linkedin.com/pub/tom-dietterich/0/97/828/)

Videolectures

Videolectures.net profile of Thomas Dietterich:

**Thomas Dietterich**  
organization: School of Electrical Engineering and Computer Science, Oregon State University, <http://eecs.oregonstate.edu>  
Homepage: <http://www.ece.oregonstate.edu/~tgda>  
Search externally: Google Scholar, Springer, Google Scholar, Microsoft Academic Search, Scopus, DBLP  
Description:  
The focus of my research is machine learning: How can we make computer systems that adapt and learn from their experience? Is data sets to expand scientific knowledge and build more useful computer applications? My laboratory combines research on machine problems in science and engineering.

Same Person

Google Scholar profile of Thomas Dietterich:

**Thomas Dietterich**  
Professor of Computer Science, Oregon State University  
Machine Learning - Computational Sustainability - Artificial Intelligence - Reinforcement Learning  
Verified email at cs.oregonstate.edu  
Homepage

Citation indices  
Citations: 6919 Since 2008: 1400  
Cited by: 13  
H-index: 34  
i10-index: 937  
Citations to my articles  
Cited by: 21 Since: 2008 Rev: 1.26 Revs / Year  
Select All Name Export  
Title / Author  
Ensemble methods in machine learning  
Multiple classifier systems, 1–11  
Solving multiclass learning problems via error-correcting output codes  
Approximate statistical tests for comparing supervised classification learning algorithms  
An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization  
Machine Learning 32(1), 165–199  
Machine Learning 32(1), 165–199

Google Scholar

Arnetminer profile of Thomas Dietterich:

**Thomas Dietterich** [HOME](#) [FOLLOW](#)  
ALIAS: Thomas G. Dietterich  
Professor and Director of Intelligent Systems  
School of Electrical Engineering and Computer Science  
School of Electrical Engineering and Computer Science  
1148 Kelley Engineering Center Oregon State University  
Covallis, Oregon 97331-5501  
P (541) 737-5559  
F (541) 737-1380  
tgda@cs.oregonstate.edu

EDIT AVATAR  
Share to: [Facebook](#) [Twitter](#) [LinkedIn](#) [Google+](#) [Tumblr](#) [StumbleUpon](#) [Digg](#) [Reddit](#) [Email](#) [Print](#) 0

Profile Activity Citation  
Scopus H-index 22

Arnetminer

# COSNET: Connecting Social Networks with Local and Global Consistency

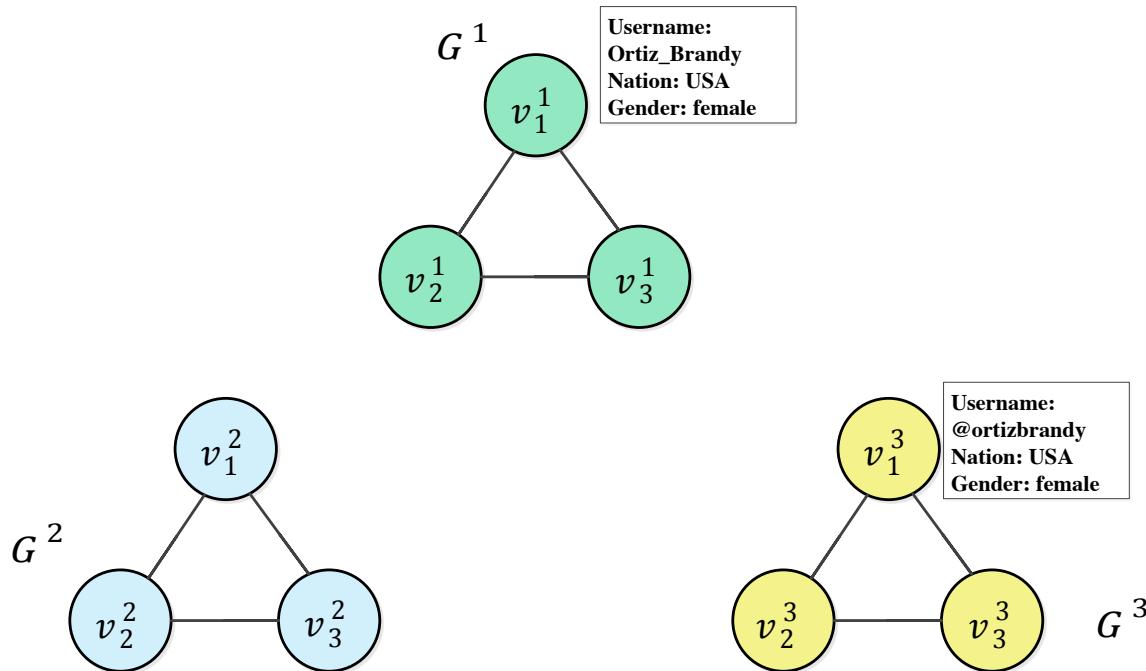


- **Input:**  $\mathbf{G}=\{G^1, G^2, \dots, G^m\}$ , with  $G^k=(V^k, E^k, R^k)$
- **Formalization:**  $\mathbf{X}=\{x_i\}$ , all possible pairwise matchings and each corresponds to  $y_i \in \{1,0\}$
- **COSNET:** an energy-based model

$$Y^* = \arg \min E(Y, X)$$

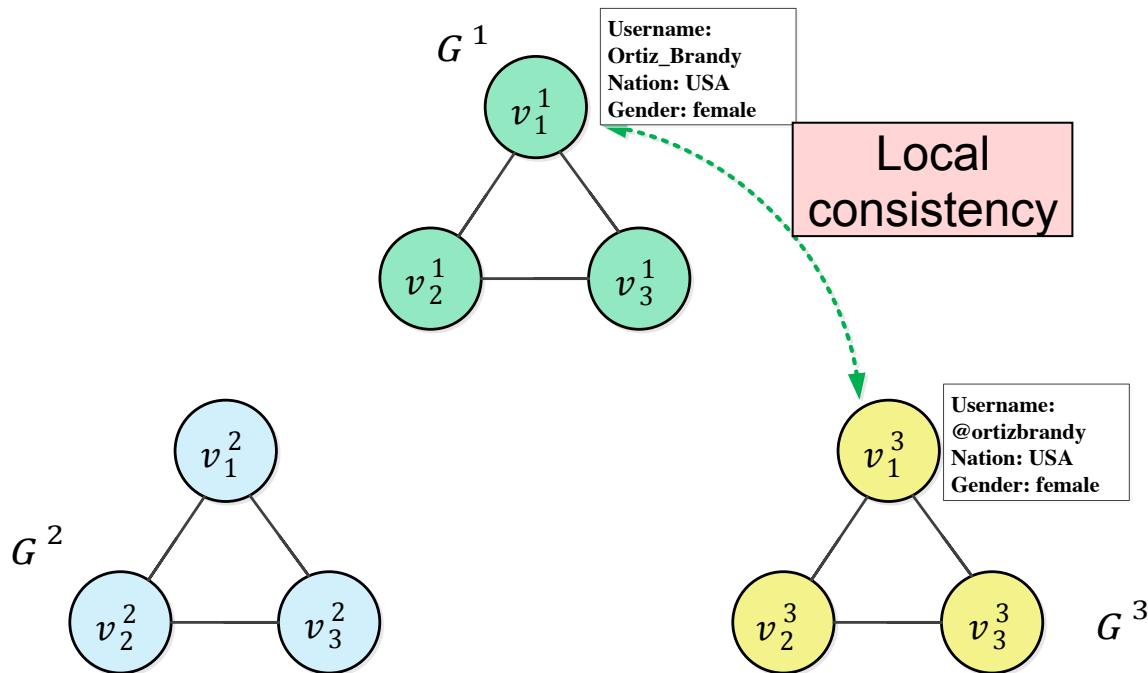
# Local vs. Global consistency

- Given three networks,



# Local vs. Global consistency

- Local matching: matching users by profiles



Pairwise similarity features

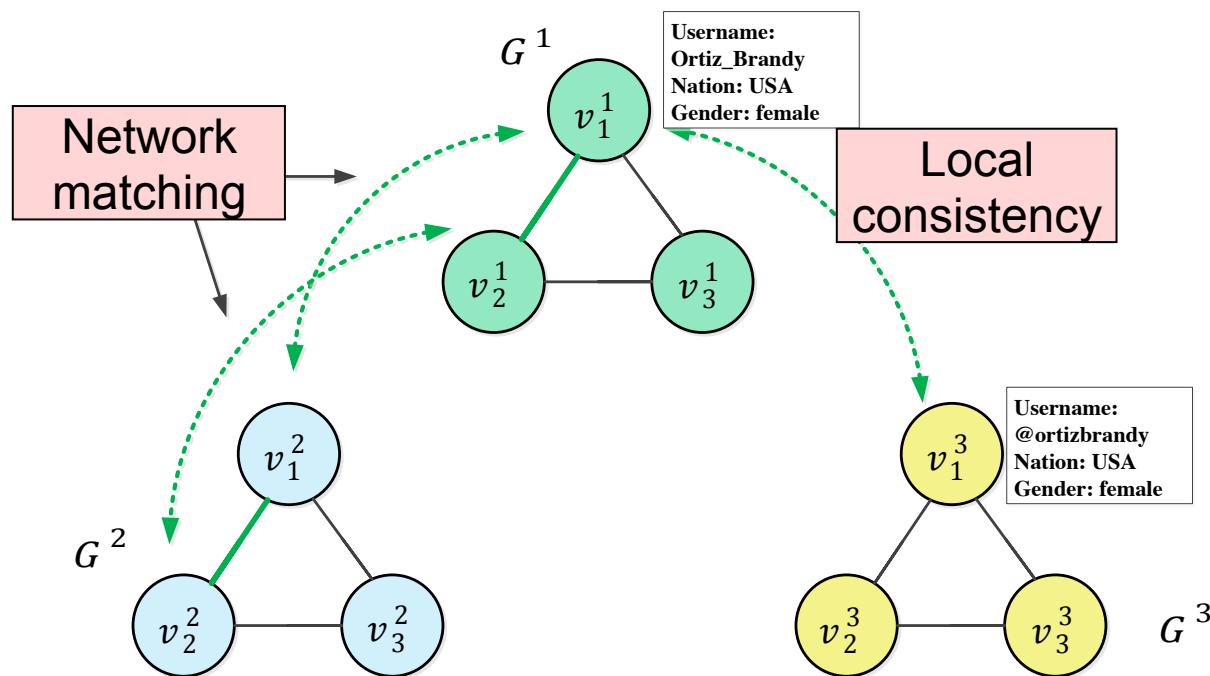
- Username similarity and uniqueness
- Profile content similarity
- Ego network similarity
- Social status

Energy function

$$E_l(Y, X) = \sum_i \mathbf{w}_l^\top \mathbf{g}_l(\mathbf{x}_i, y_i)$$

# Local vs. Global consistency

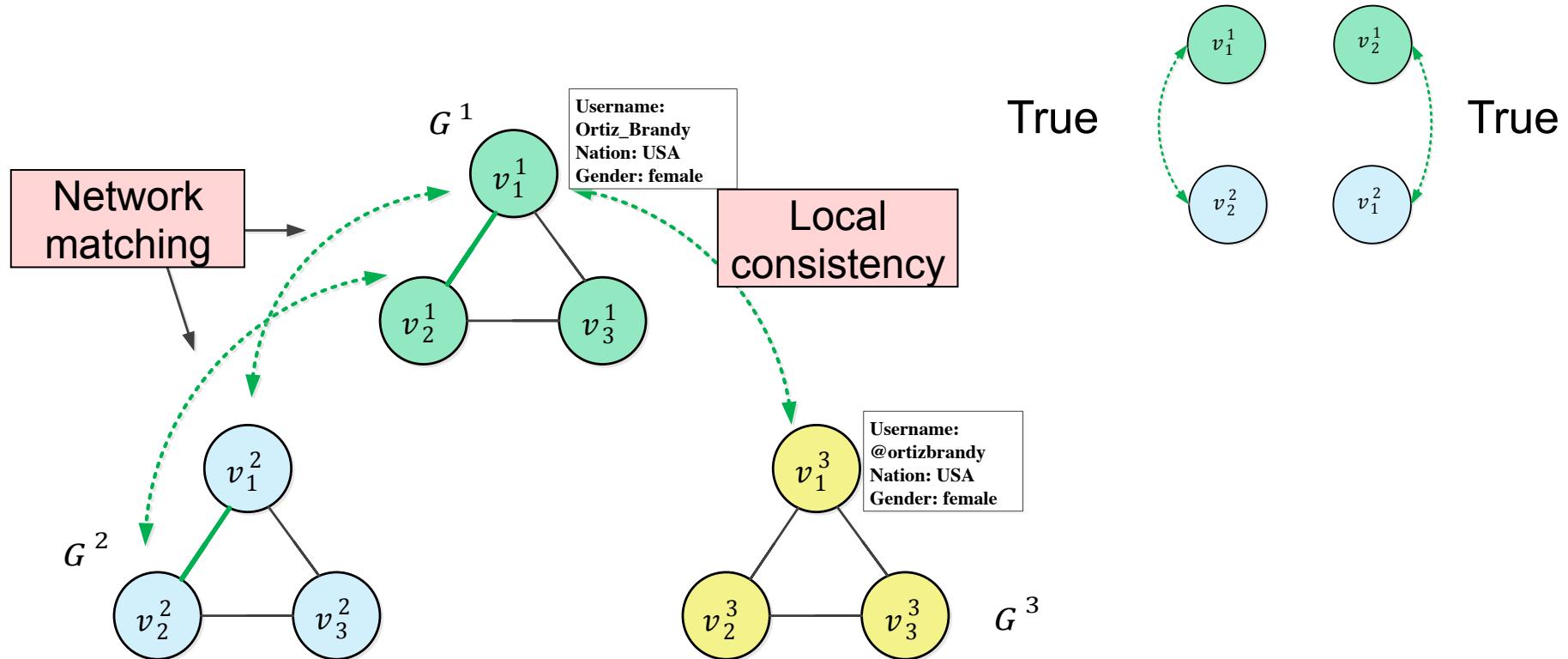
- Network matching: matching users' ego networks



Encourage “neighborhood-preserving matching”

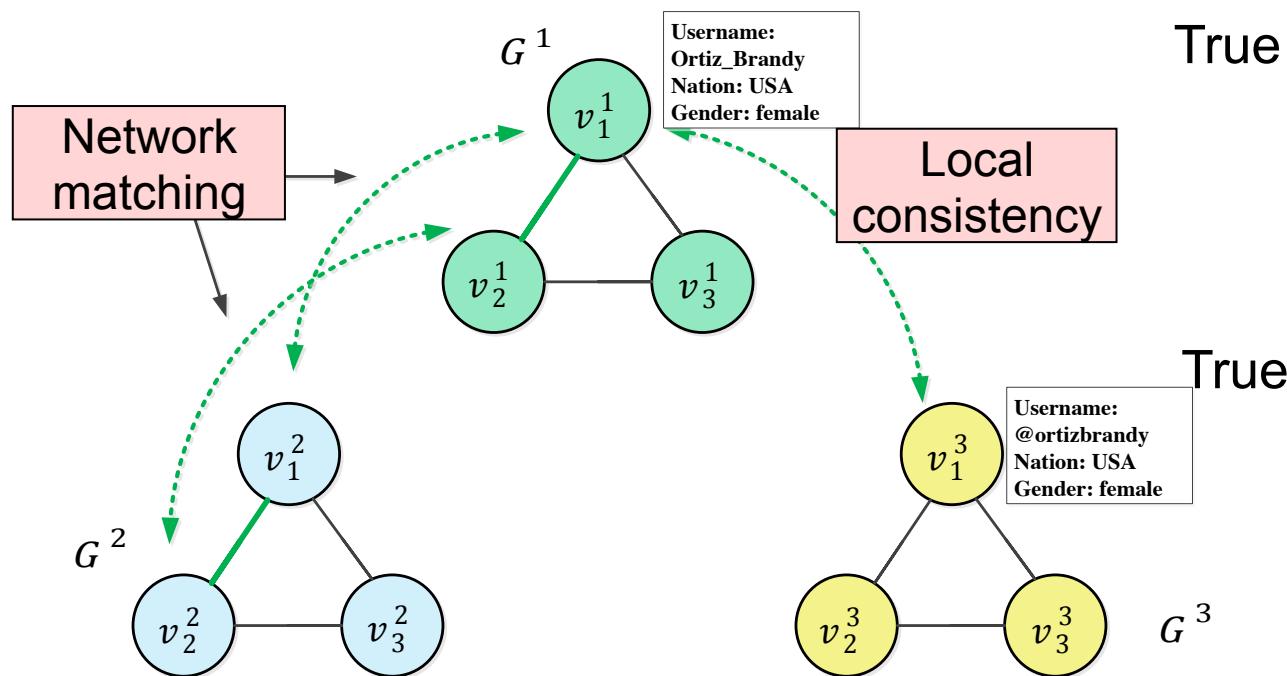
# Local vs. Global consistency

- Network matching: matching users' ego networks



# Local vs. Global consistency

- Network matching: matching users' ego networks



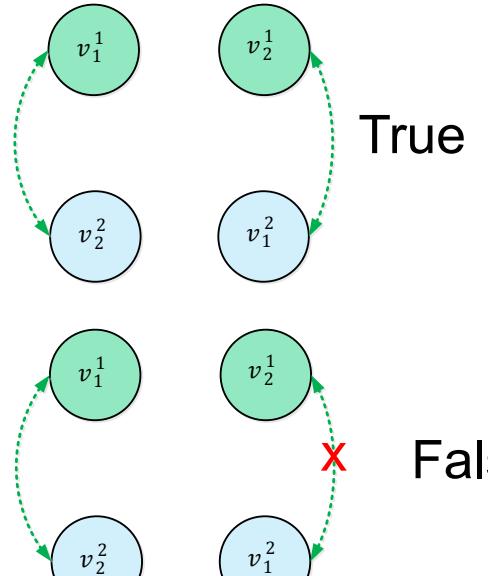
True

Local consistency

True

Username:  
@ortizbrandy  
Nation: USA  
Gender: female

$G^3$

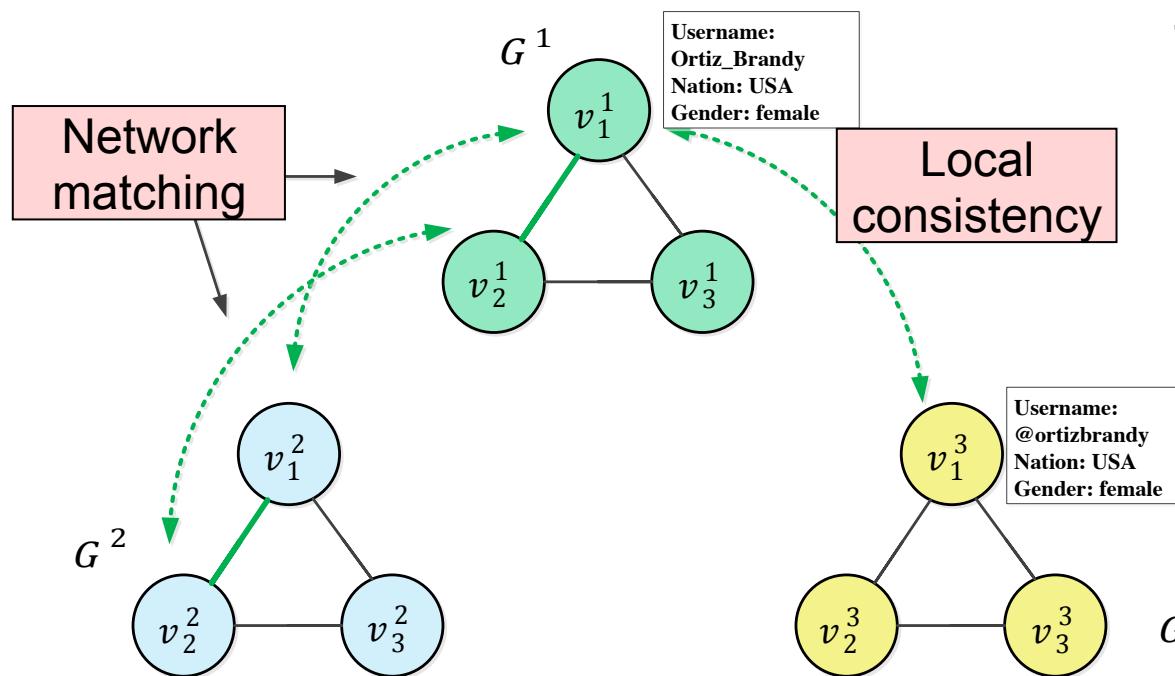


True

False

# Local vs. Global consistency

- Network matching: matching users' ego networks

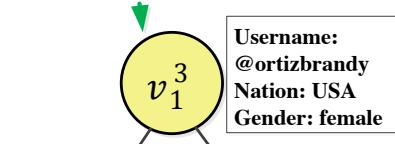


True

True

False

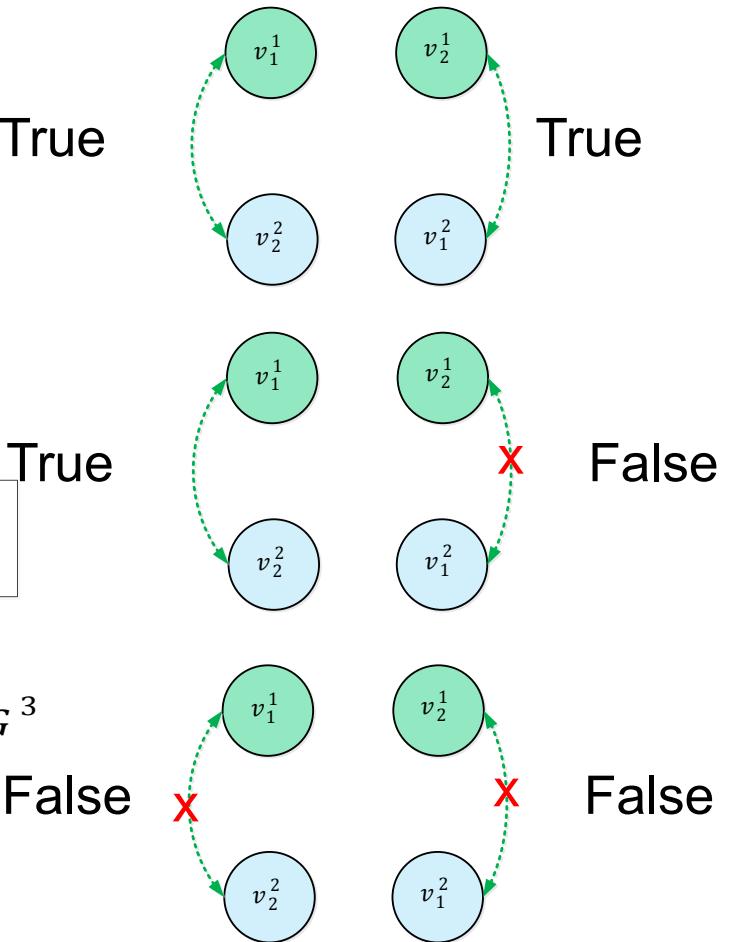
Local consistency



True

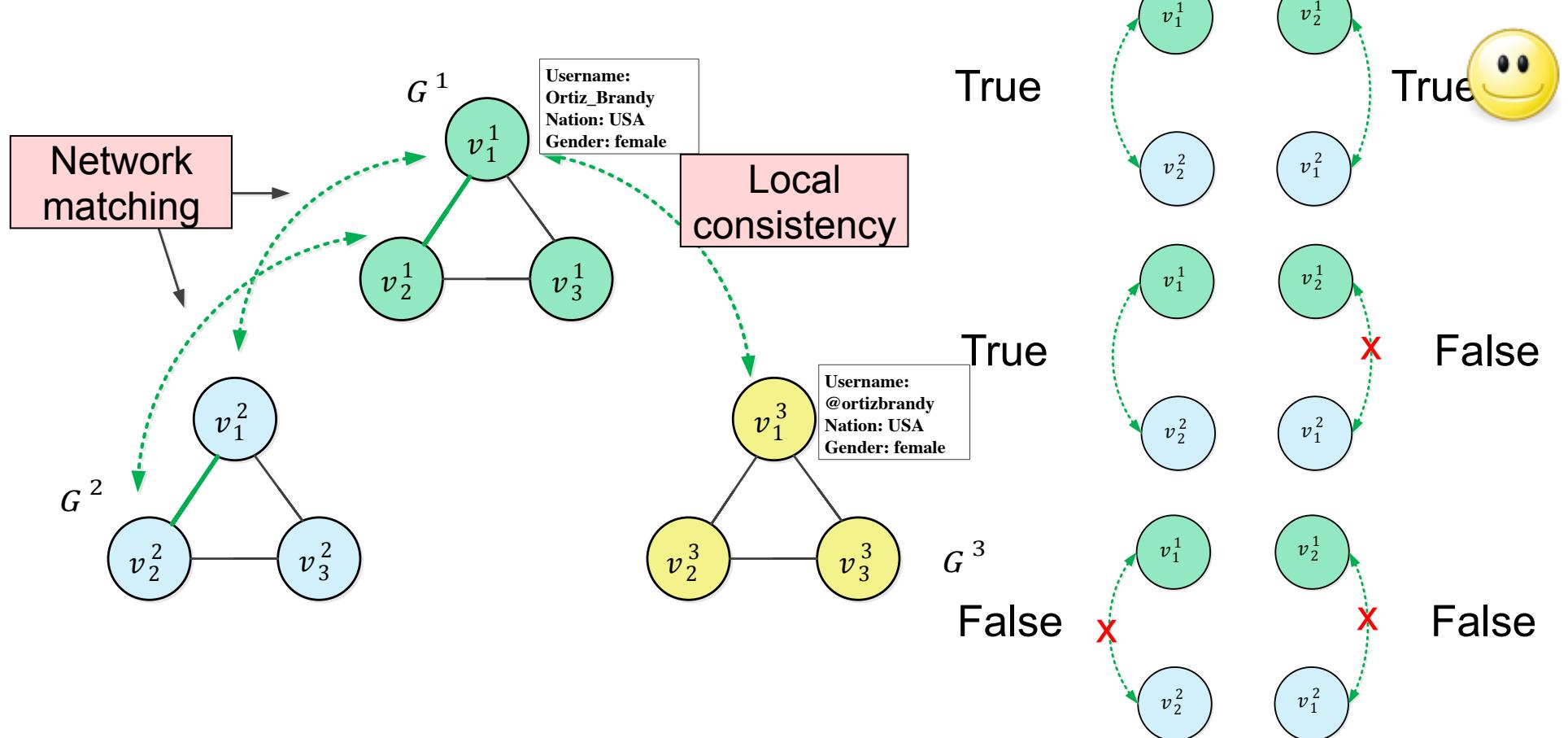
False

False



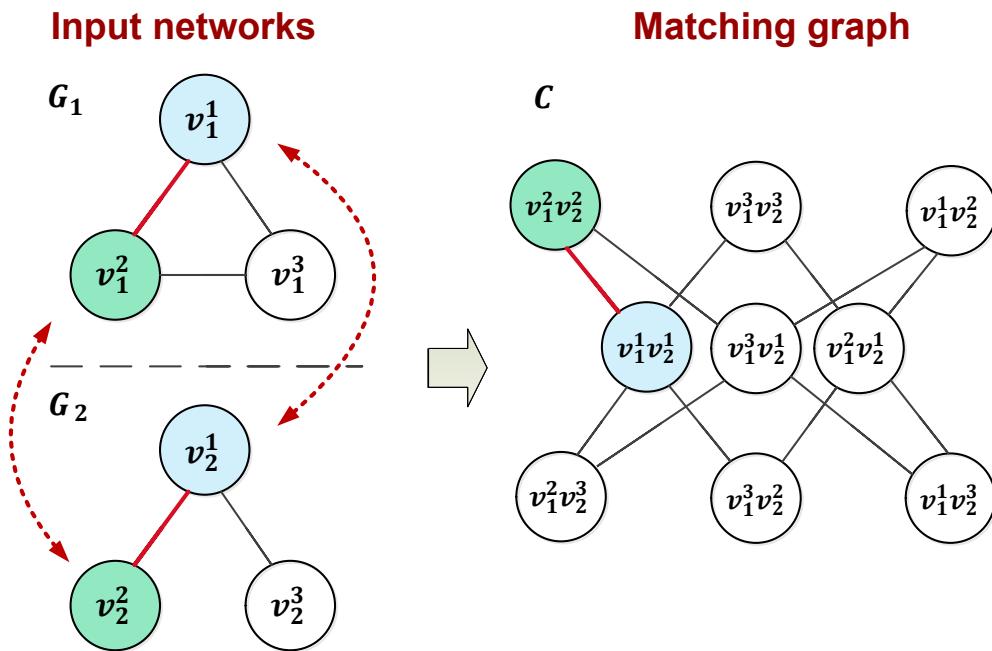
# Local vs. Global consistency

- Network matching: matching users' ego networks



# Network Matching

- Network matching: matching users' ego networks



**Energy function**

$$E_e(Y, X) = \sum_{\langle \mathbf{x}_i, \mathbf{x}_j \rangle \in E_{MG}} \mathbf{w}_e^\top \mathbf{f}_e(y_i, y_j)$$

$$\mathbf{f}_e(y_i, y_j) = \begin{cases} (1, 0, 0)^\top & \text{if } y_i = y_j = 0 \\ (0, 1, 0)^\top & \text{if } y_i + y_j = 1 \\ (0, 0, 1)^\top & \text{if } y_i = y_j = 1 \end{cases}$$

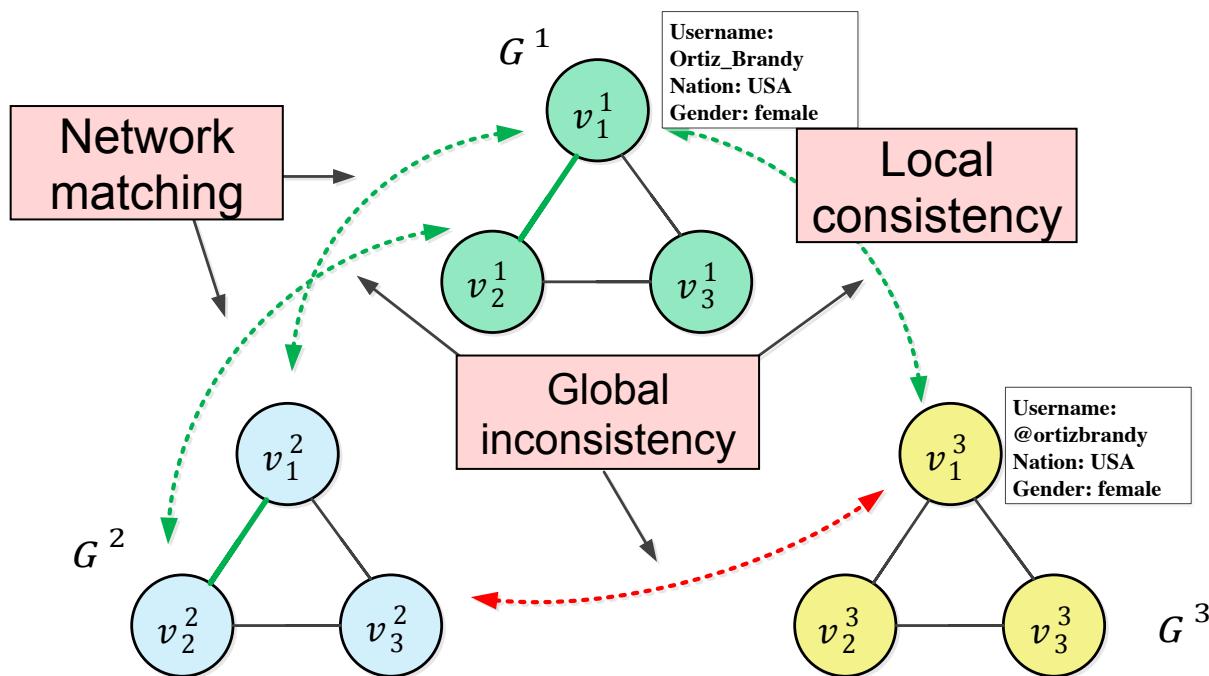


# Candidate Pruning

- Content-based method
  - Username similarity above a threshold
- Structure-based similarity
  - Starting from a seed mapping set and iteratively propagate the m

# Local vs. Global consistency

- Global consistency: matching users by avoiding global inconsistency



**DEFINITION 2 (GLOBAL INCONSISTENCY).** Given a set of social networks  $\mathbf{G}$ , a set of user pairs  $X$  and the corresponding labels  $Y$ , if there exists a sequence of user pairs  $\langle \mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \dots, \mathbf{x}_{i_n} \rangle$ , such that

$$\forall i = i_1, i_2, \dots, i_n, y_i = 1$$

and

$$\forall k = 1, 2, \dots, n-1, \mathcal{V}_{i_k}^2 = \mathcal{V}_{i_{k+1}}^1$$

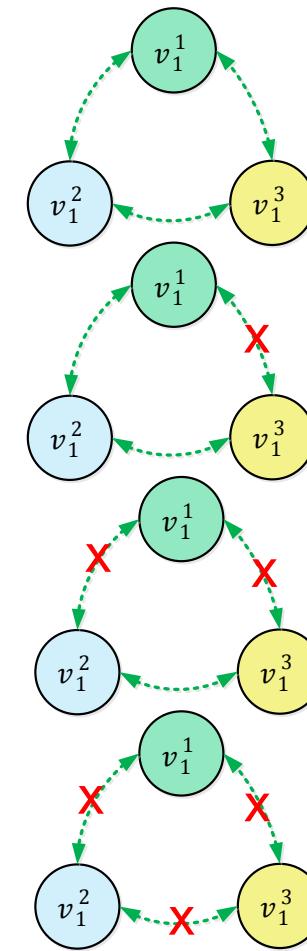
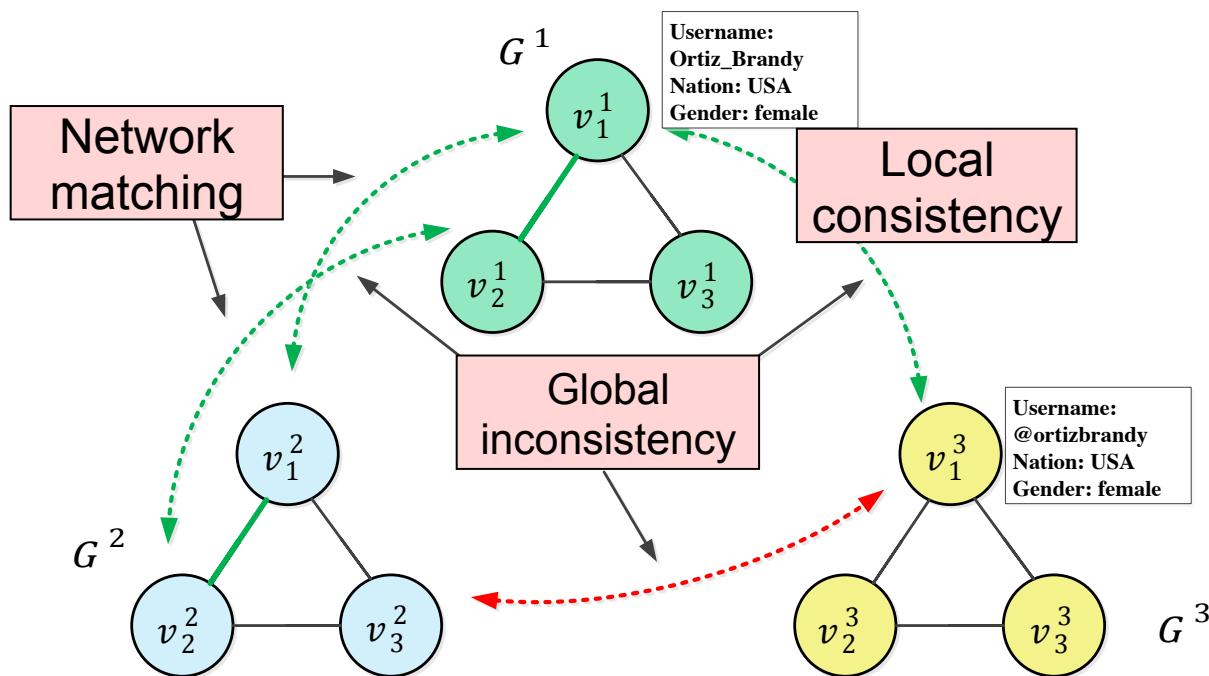
and

For the pair  $\langle \mathcal{V}_{i_n}^2, \mathcal{V}_1^1 \rangle$ , if the corresponding label  $y_j = 0$  then we say that the assigned labels  $Y$  causes global inconsistency given  $\mathbf{G}$  and  $X$ .

Avoid “global inconsistency”

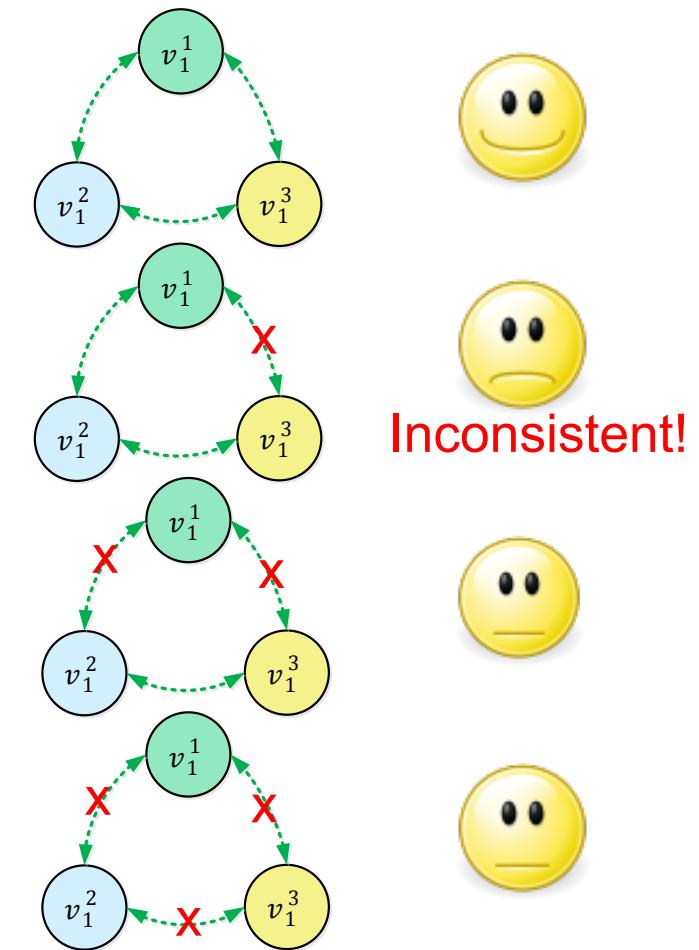
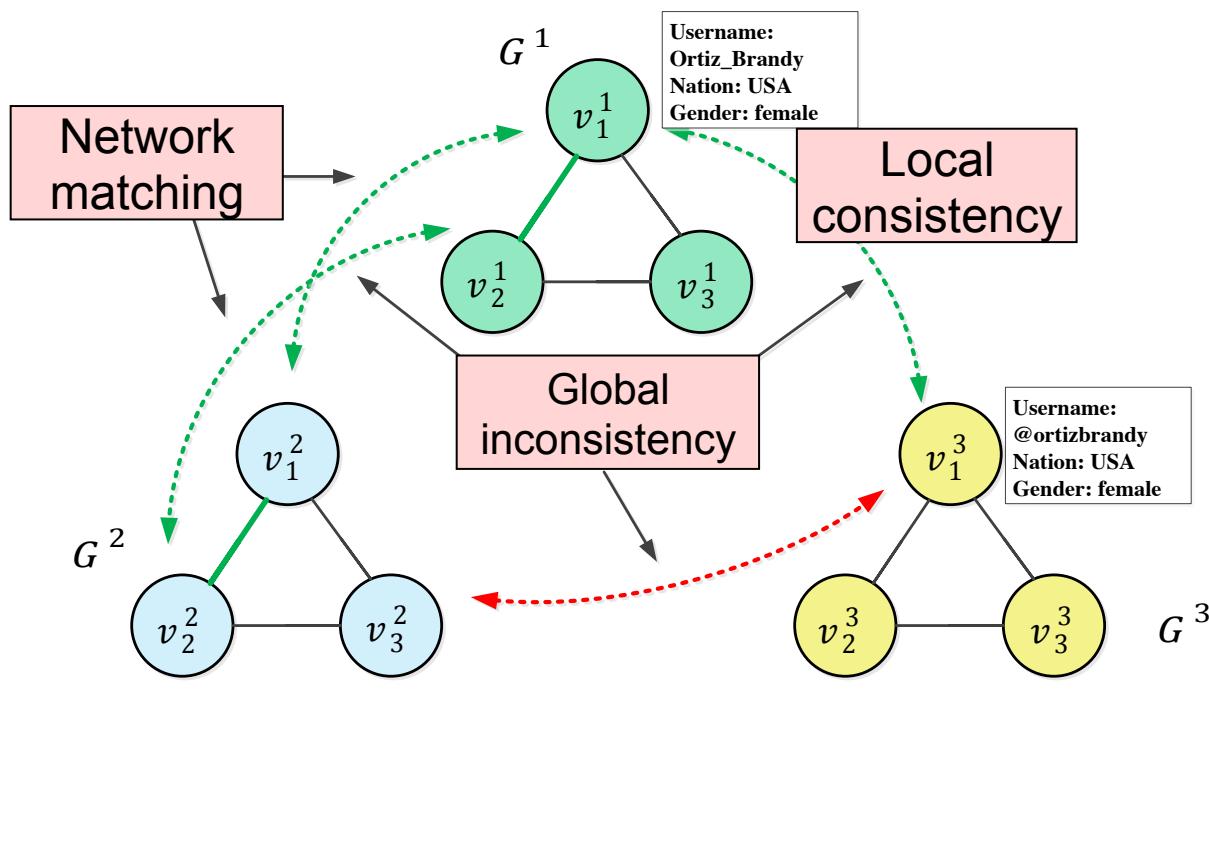
# Local vs. Global consistency

- Global consistency: matching users by avoiding global inconsistency



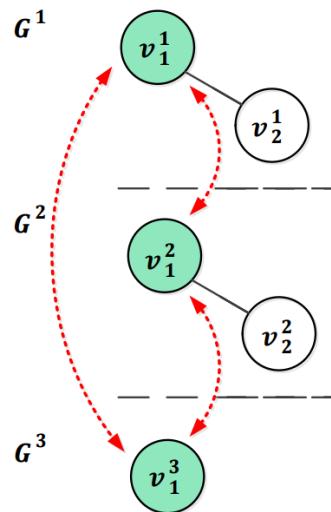
# Local vs. Global consistency

- Global consistency: matching users by avoiding global inconsistency

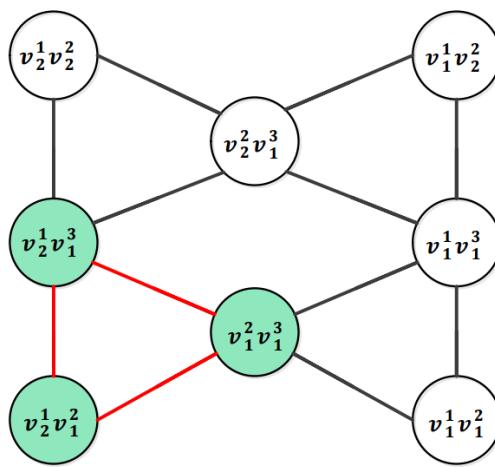


# Avoid Global Inconsistency

**Input networks**



**Matching graph**

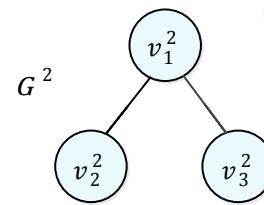
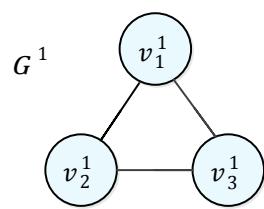


**Energy function**

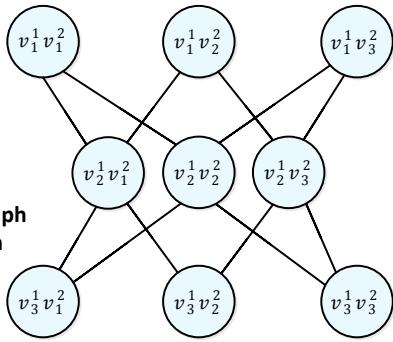
$$E_t(Y, X) = \sum_{c \in T_{MG}} \mathbf{w}_t^\top \mathbf{f}_t(Y_c)$$

$$\mathbf{f}_t(y_i, y_j) = \begin{cases} (1, 0, 0, 0)^\top & \text{if } |Y_c| = 0 \\ (0, 1, 0, 0)^\top & \text{if } |Y_c| = 1 \\ (0, 0, 1, 0)^\top & \text{if } |Y_c| = 2 \\ (0, 0, 0, 1)^\top & \text{if } |Y_c| = 3 \end{cases}$$

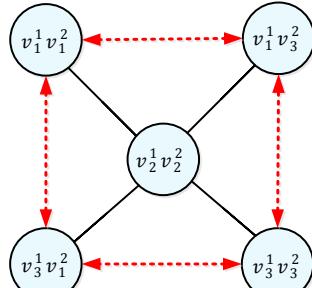
# Model Construction



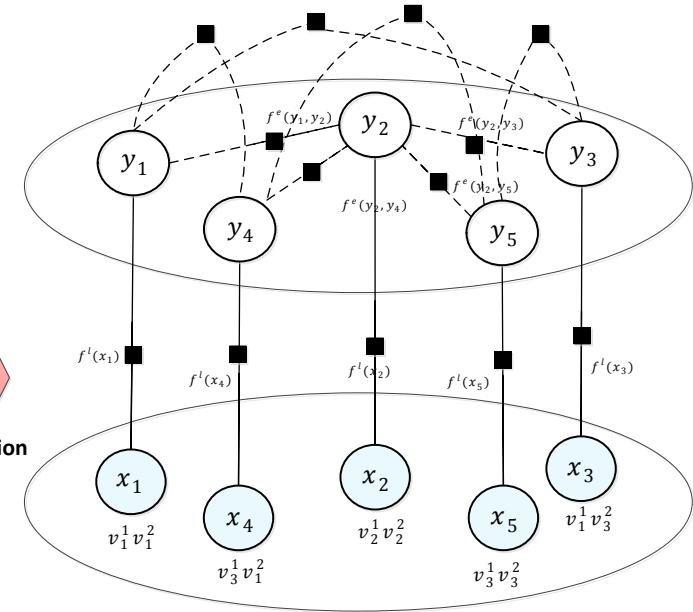
**Matching Graph Generation**



**Candidate Pruning**



**Model Construction**



(a) Two input networks

(b) The generated matching graph

(c) Matching graph after pruning

(d) The constructed model

Objective function by combining all the energy functions

$$\begin{aligned}
 E(Y, X) = & \sum_{\mathbf{x}_i \in V_{MG}} \mathbf{w}_l^\top \mathbf{g}_l(\mathbf{x}_i, y_i) + \sum_{\langle \mathbf{x}_i, \mathbf{x}_j \rangle \in E_{MG}} \mathbf{w}_e^\top \mathbf{f}_e(y_i, y_j) \\
 & + \sum_{c \in T_{MG}} \mathbf{w}_t^\top \mathbf{f}_t(Y_c)
 \end{aligned} \tag{2}$$



# Model Learning

- Max-margin learning

$$\min_W \frac{1}{2} \|W\|^2 + \mu \xi$$

$$\text{s.t. } E(\hat{Y}, X; W) \leq E(Y, X; W) - \Delta(Y, \hat{Y}) + \xi$$

- As the original problem is intractable, we use Lagrangian relaxation to decompose the original objective function into a set of easy-to-solve sub-problems

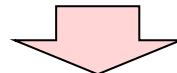
$$\begin{aligned} E(Y, X; W) &= \sum_{f \in \mathcal{F}} E_f(Y_f, X_f; W) \\ &= \sum_{f \in \mathcal{F}} \sum_{x_i \in X_f} \left( \frac{1}{|\mathcal{F}_i|} \mathbf{w}_l^\top \mathbf{g}_l(x_i, y_i^f) + \mathbf{w}_f^\top f(Y_f) \right) \\ \text{s.t. } y_i^f &= y_i, \quad \forall f, y_i \in Y_f \end{aligned}$$



# Model Learning (cont.)

- Dual decomposition

$$L(Y, X, \boldsymbol{\lambda}; W) = \min_W \sum_{f \in \mathcal{F}} \left( \sum_{y_i \in Y_f} \frac{1}{|\mathcal{F}_i|} \mathbf{w}_l^\top \mathbf{g}_l(\mathbf{x}_i, y_i^f) + \mathbf{w}_f^\top f(Y_f) \right) \\ + \sum_{f \in \mathcal{F}} \sum_{y_i \in Y_f} \lambda_i^f (y_i - y_i^f)$$



This provides a **lower bound** to the original function

$$\min_{W, \boldsymbol{\lambda}} \frac{1}{2} \|W\|^2 + \mu(E(\hat{Y}, X; W) - \max_{\boldsymbol{\lambda}} L(Y, X, \boldsymbol{\lambda}; W))$$

$$\text{s.t. } \sum_{y_i \in Y_i} \lambda_i^f = 0, \quad \forall f \in \mathcal{F}$$

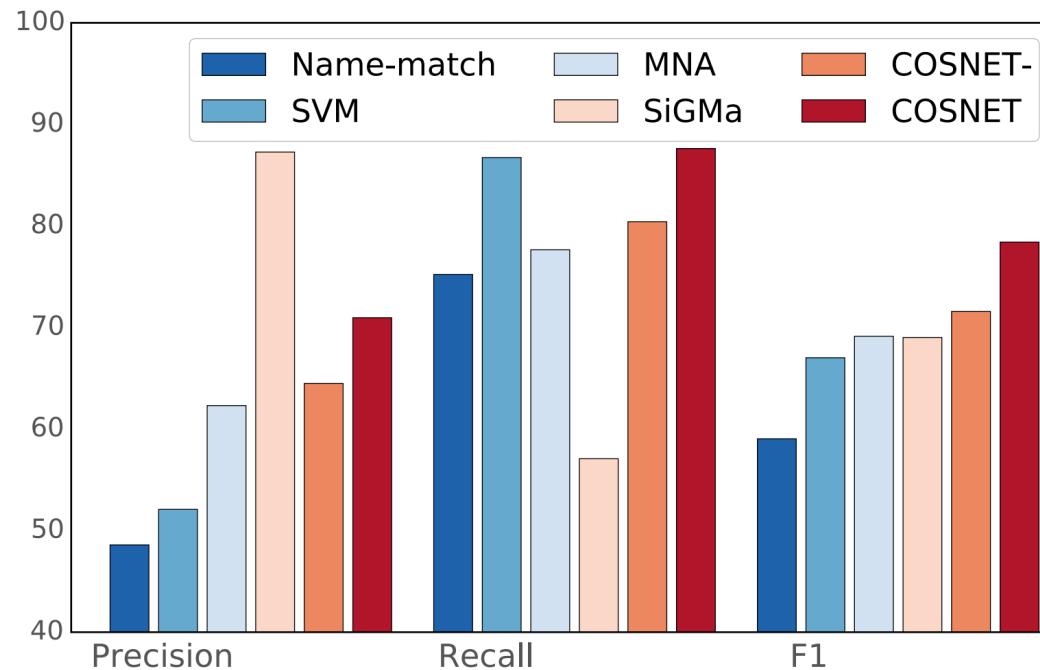
The resulting objective function is convex and non-differentiable, and can be solved by **projected sub-gradient** method



# Results

# Connecting AMiner with ...

- LinkedIn and VideoLectures



Name-match: match name only;

SVM: use classifier to identify the same user;

MNA: an optimization method;

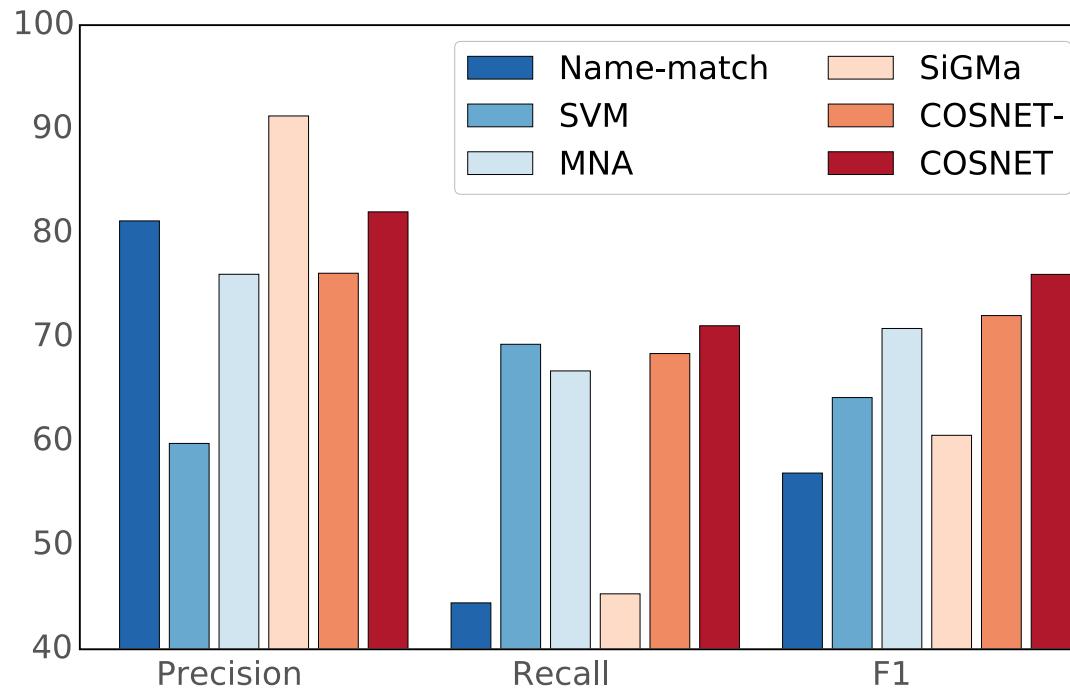
SiGMA: local propagation;

COSNET: our method;

COSNET-: w/o global consistency.

# Connecting Social Media Sites

- Twitter, LiveJournal, Last.fm, Flickr, MySpace



Name-match: match name only;

SVM: use classifier to identify the same user;

MNA: an optimization method;

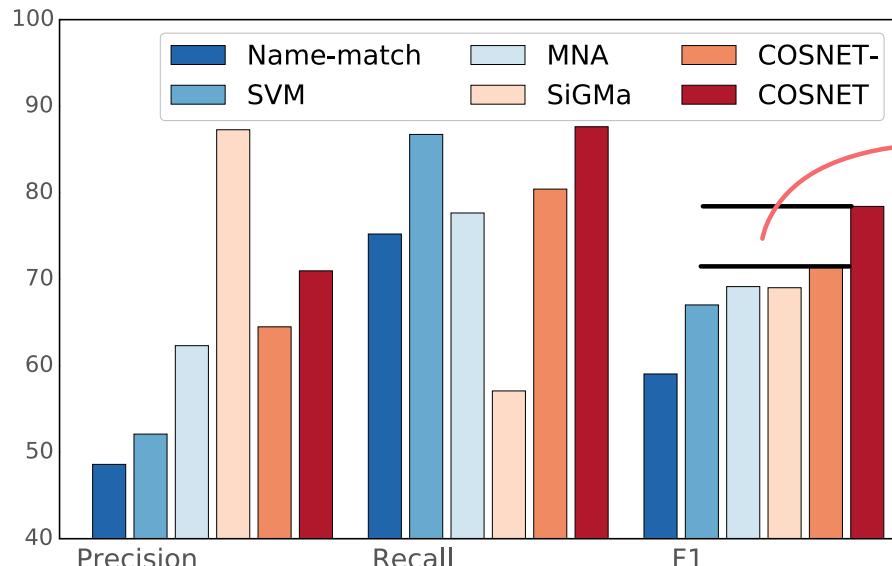
SiGMA: local propagation;

COSNET: our method;

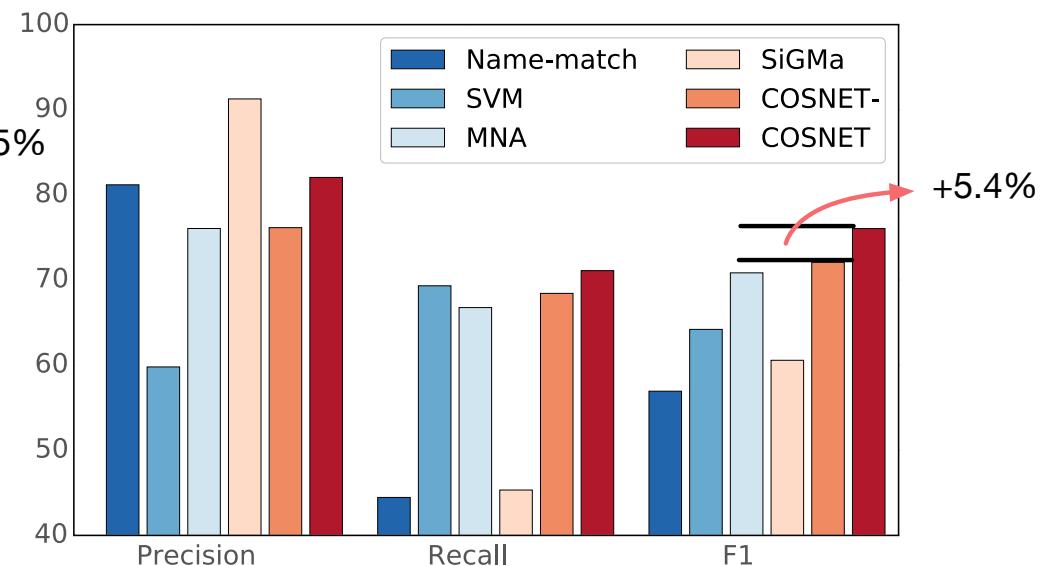
COSNET-: w/o global consistency.

# Effects of Global Consistency

COSNET-: w/o global consistency.



Academia Collection



SNS Collection

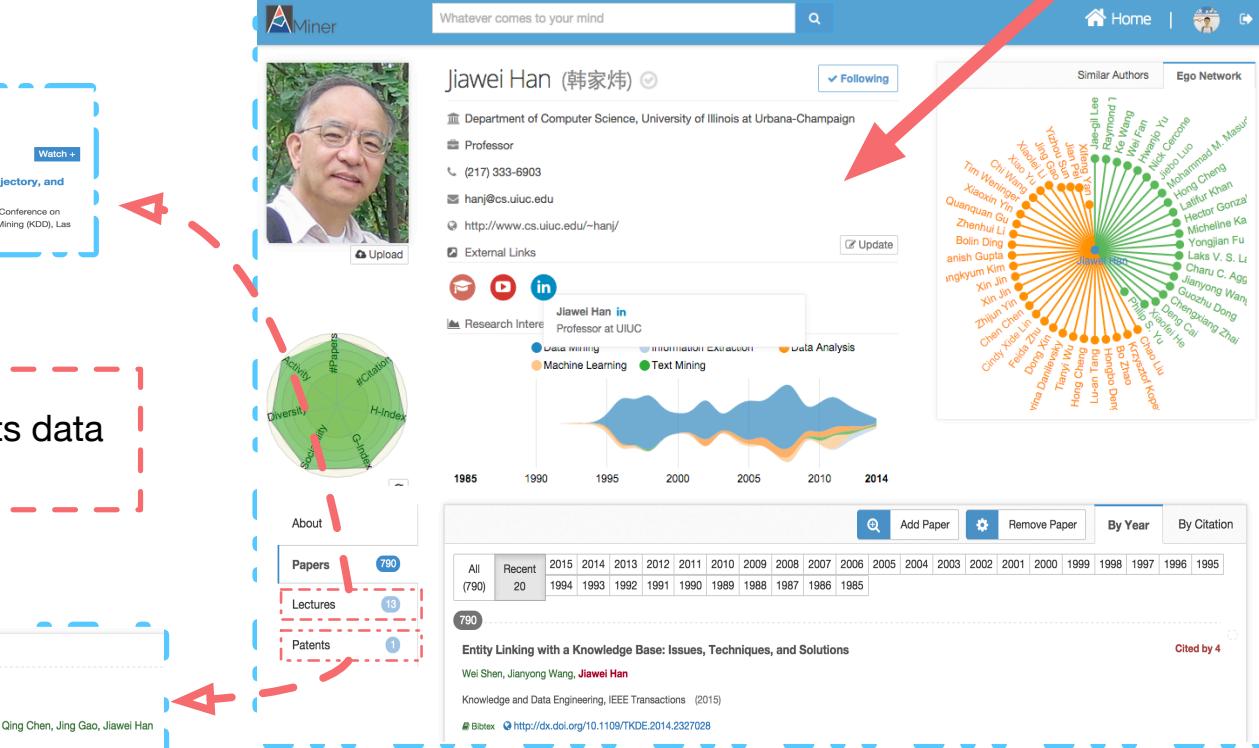
# Application in AMiner



- Video contents

**Bringing Structure to Text: Mining Phrases, Entity Concepts, Topics, and Hierarchies**  
20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), New York 2014

**Mining Massive RFID, Trajectory, and Traffic Data Sets**  
14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Las Vegas 2008



- Personal profiles
- Business connections
- Skills and expertise



- Patents data

1 Systems and Methods for Detecting a Novel Data Class  
Mohammad Mehedy Masud, Latifur Rahman Khan, Bhavani Marlenne Thuraisingham, Qing Chen, Jing Gao, Jiawei Han  
Publication-date: 2012-03-01 Application-date: 2011-08-22



# Thanks!



**Data & source code**

<http://aminer.org>

<http://aminer.org/cosnet>