

# 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



<sup>#</sup>Simon Fraser University

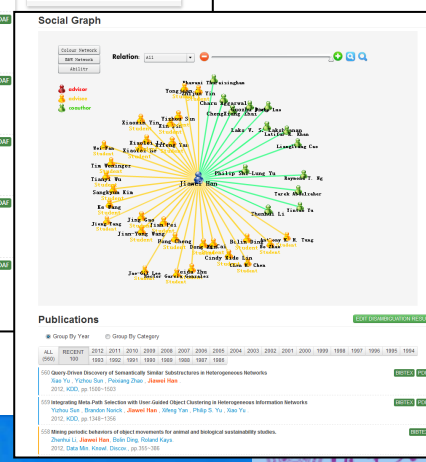
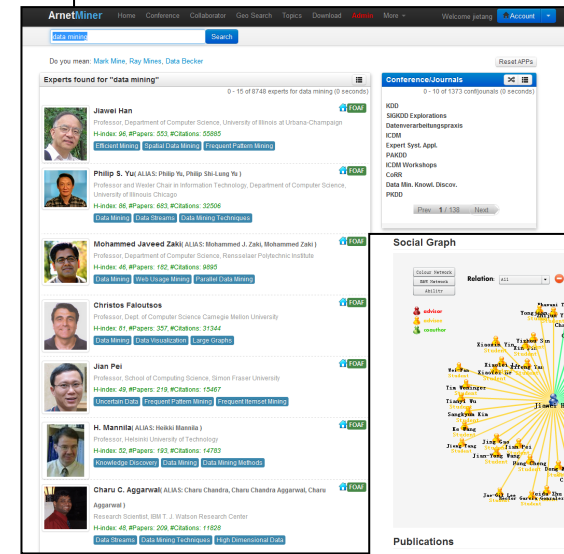
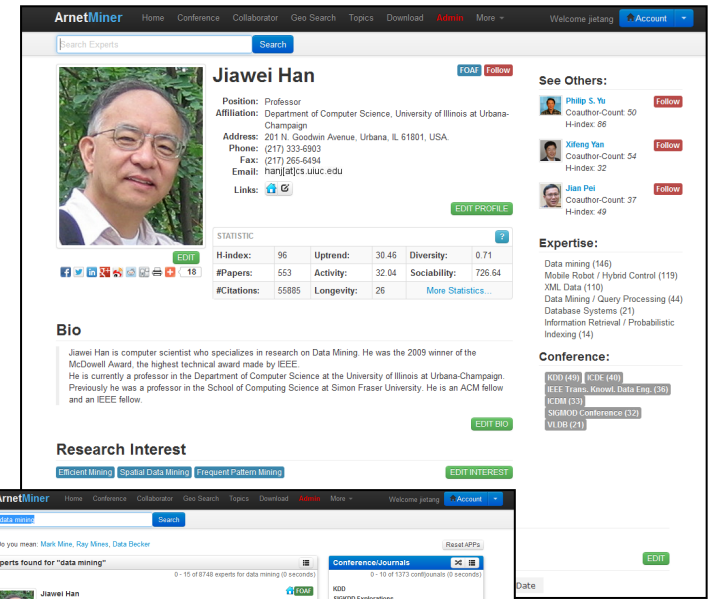


<sup>\*</sup>University of Illinois at Chicago



# 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



# 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

**Educational history**

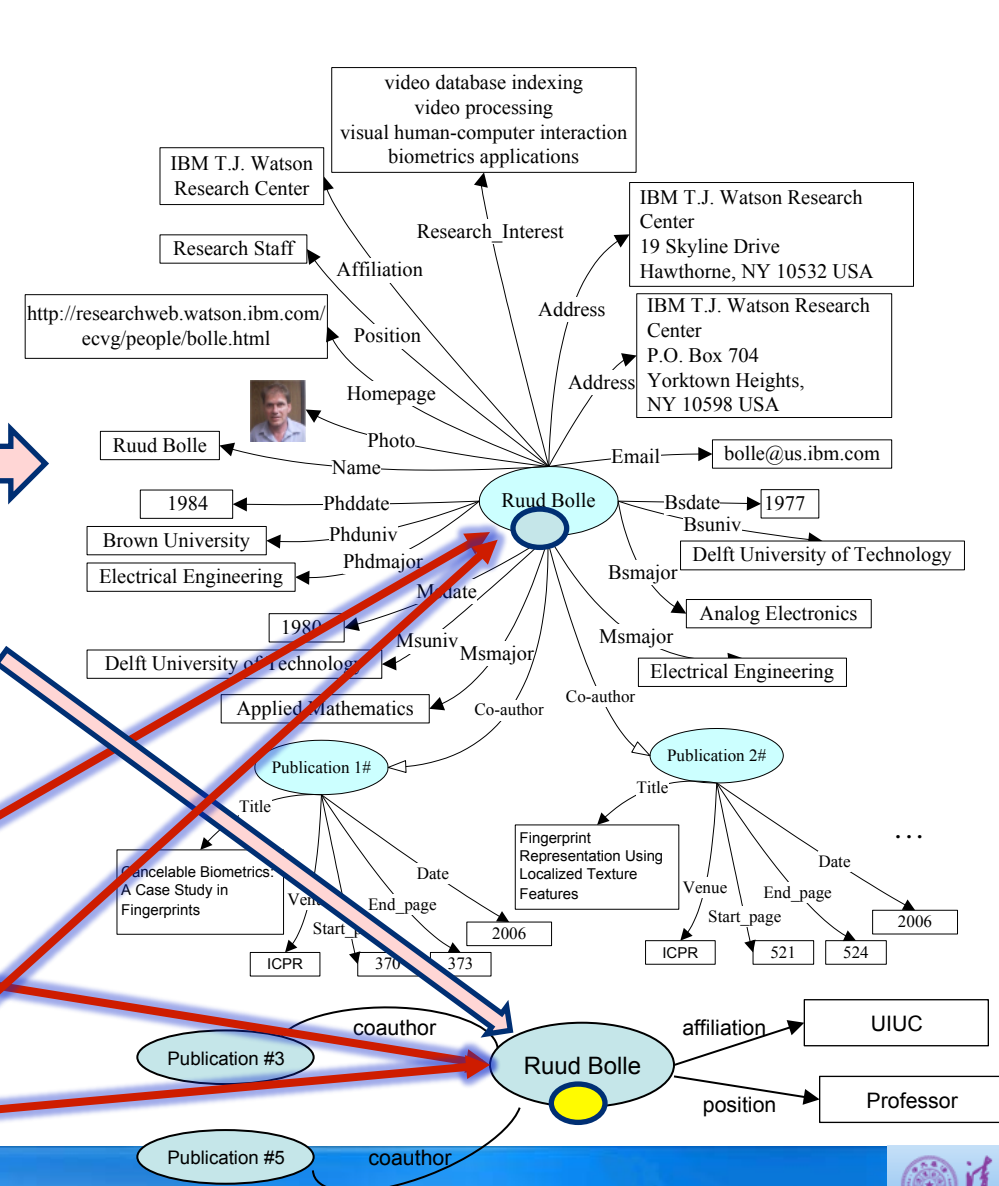
**Ruud M. Bolle** was born in Voorburg, The Netherlands. He received the Bachelor's Degree in Analog Electronics in 1977 and the Master's Degree in Electrical Engineering in 1980, both from Delft University of Technology, Delft, The Netherlands. In 1983 he received the Master's Degree in Applied Mathematics 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.

**Academic services**

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

**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.

DBLP: Ruud Bolle		Publications
2006		
50	EE	Nalini K. Ratha, Jonathan Connell, Ruud M. Bolle, Sharat Chikkerur: Cancelable Biometrics: A Case Study in Fingerprints. ICPR (4) 2006: 370-373
43	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	Andrew Senior, 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
45	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>



**Jiawei Han**  
Position: Professor  
Affiliation: Department of Computer Science, University of Illinois at Urbana-Champaign  
Address: 201 N. Goodwin Avenue, Urbana, IL 61801, USA.  
Phone: (217) 333-6903  
Fax: (217) 265-6494  
Email: hanj[at]cs.uiuc.edu

STATISTIC			
H-index:	96	Uptrend:	30.46
Diversity:	0.71	#Papers:	553
Activity:	32.04	Sociability:	726.64
#Citations:	55685	Longevity:	26

**See Others:**  
Philip S. Yu (Coauthor-Count: 50, H-index: 86)  
Xifeng Yan (Coauthor-Count: 54, H-index: 32)  
Jan Pei (Coauthor-Count: 37, H-index: 49)

**Expertise:**  
Data mining (146)  
Mobile Robot / Hybrid Control (119)  
XML Data (110)  
Data Mining / Query Processing (44)  
Database Systems (21)  
Information Retrieval / Probabilistic Indexing (14)

**Conference:**  
KDD (40)  
ICDE (40)  
IEEE Trans. Knowl. Data Eng. (35)  
ICDM (33)  
SIGMOD Conference (32)  
VLDB (21)

**M. I. 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

STATISTIC			
H-index:	75	Uptrend:	7.2
Diversity:	0.03	#Papers:	242
Activity:	11.12	Sociability:	331.69
#Citations:	44312	Longevity:	23

**See Others:**  
Philip S. Yu (Coauthor-Count: 50, H-index: 86)  
Xifeng Yan (Coauthor-Count: 54, H-index: 32)  
Jan Pei (Coauthor-Count: 37, H-index: 49)

**H. Garcia**  
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) 723-0685  
Fax: (650) 726-2588  
Email: hector@cs.stanford.edu

STATISTIC			
H-index:	80	Uptrend:	-4.04
Diversity:	0.22	#Papers:	195
Activity:	4.86	Sociability:	407.19
#Citations:	57908	Longevity:	25

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

**Research Interest:**  
Database Systems | Data Management | Data Warehousing

**Scott**  
Position:  
Affiliation:  
Address:  
Phone:  
Links:

STATISTIC			
H-index:	96	Uptrend:	-4.04
Diversity:	0.22	#Papers:	195
Activity:	4.86	Sociability:	407.19
#Citations:	57908	Longevity:	25


Extracted more than 1,000,000 researcher profiles from the Web

[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?

Miner

Home |



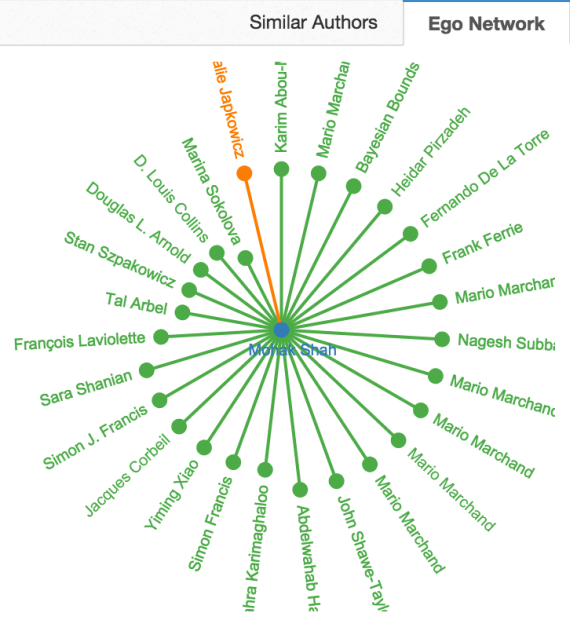
Upload

## Mohak Shah ✓

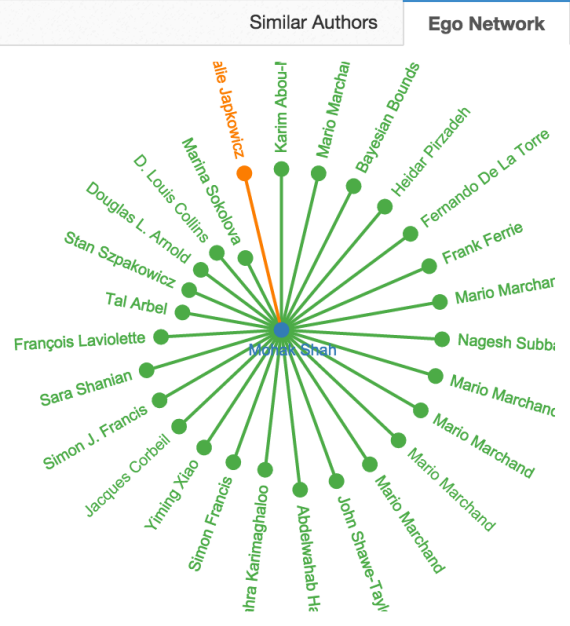
- Centre for Intelligent Machines McGill University
- Postdoctoral Fellow
- 1-514-398-8702
- mohak [at] cim [dot] mcgill [dot] ca
- <http://www.cim.mcgill.ca/~mohak/Site/Home.html>

[Follow](#) [Update](#)

Similar Authors

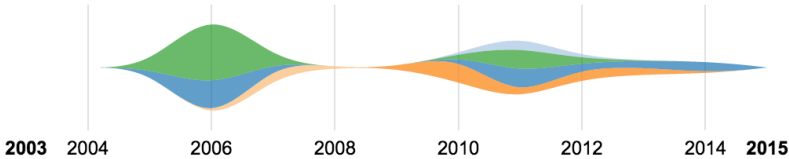


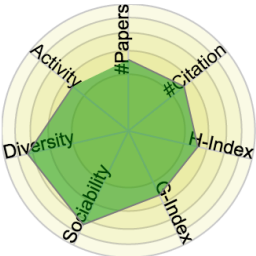
Ego Network



Research Interests

- Learning Algorithms
- Unsupervised Learning
- Conditional Random Fields
- Decision Trees
- Feature Selection





# Required semantics are distributed in multiple sources



LinkedIn

The screenshot shows the LinkedIn profile of Mohak Shah. At the top, there is a search bar and navigation links for Home, Profile, Connections, Education, Jobs, Interests, and Business. Below the navigation is a banner for 'Academic Executive Brief'. The profile header includes a photo of Mohak Shah, his name, and a '3rd' connection indicator. His current role is 'Sr Manager- Data Science at Bosch; General Chair, KDD 2016' in Palo Alto, California. His education includes 'McGill University'. A 'Send Mohak InMail' button and '500+ connections' are also visible. The 'Experience' section lists three roles: 'Sr Manager, Data Science' at Bosch (April 2014 - Present), 'General Chair' at ACM SIGKDD 2016, and 'Machine Learning Lab Manager' at GE Software (June 2012 - April 2014). The 'Research Manager, Data Mining and Machine Learning' role at Accenture (February 2011 - June 2012) is partially visible at the bottom.

Videolectures

videolectures.net  
exchange ideas & share knowledge



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

## Mohak Shah

homepage: <http://www.mohakshah.com/Site/Home.html>

search externally: Google Scholar, Springer, CiteSeer, Microsoft Academic

## 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**



Tom Dieterich  
Professor at Oregon State University and Director of Intelligent Systems Research


**Videolectures**



Thomas Dietterich  
School of Electrical Engineering and Computer Science, Oregon State University


**Same Person**

**Google Scholar**



Thomas Dietterich  
Professor of Computer Science, Oregon State University

**Arnetminer**



Thomas Dietterich  
Professor and Director of Intelligent Systems  
School of Electrical Engineering and Computer Science

# COSNET: Connecting Social Networks with Local and Global Consistency



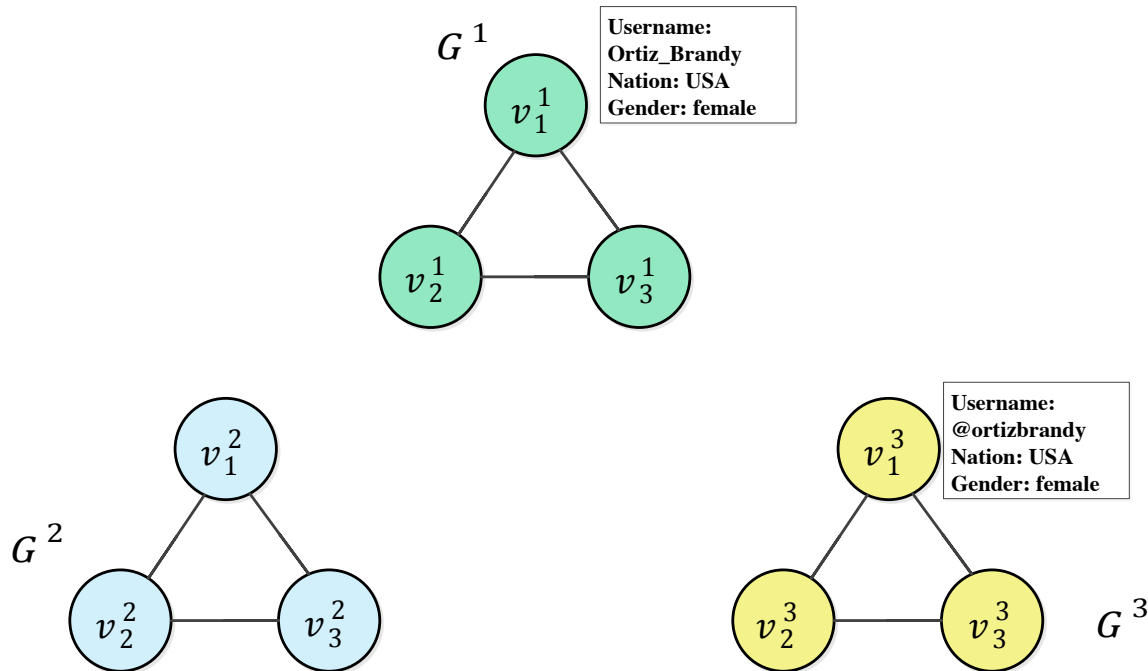
- **Input:**  $\mathbf{G} = \{G^1, G^2, \dots, G^m\}$ , with  $G^k = (V^k, E^k, R^k)$
- **Formalization:**  $\mathbf{X} = \{x_i\}$ , all possible pairwise matchings and each corresponds to  $y_i \in \{1, 0\}$
- **COSNET:** an energy-based model

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



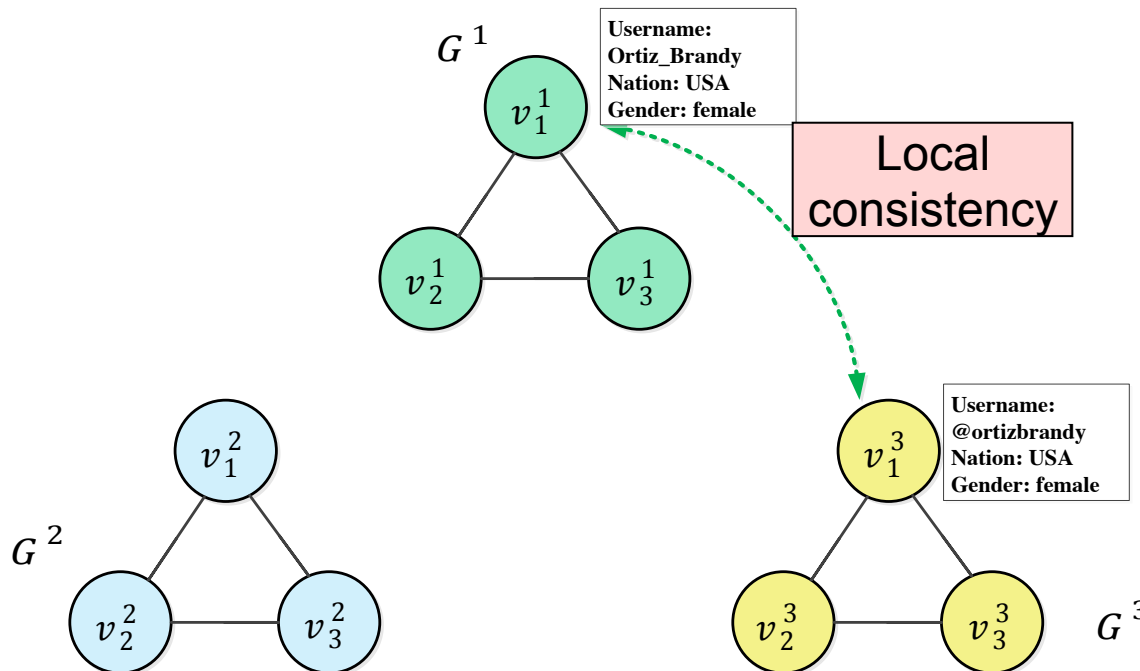
# Local vs. Global consistency

- Given three networks,



# Local vs. Global consistency

- Local matching: matching users by profiles



## Pairwise similarity features

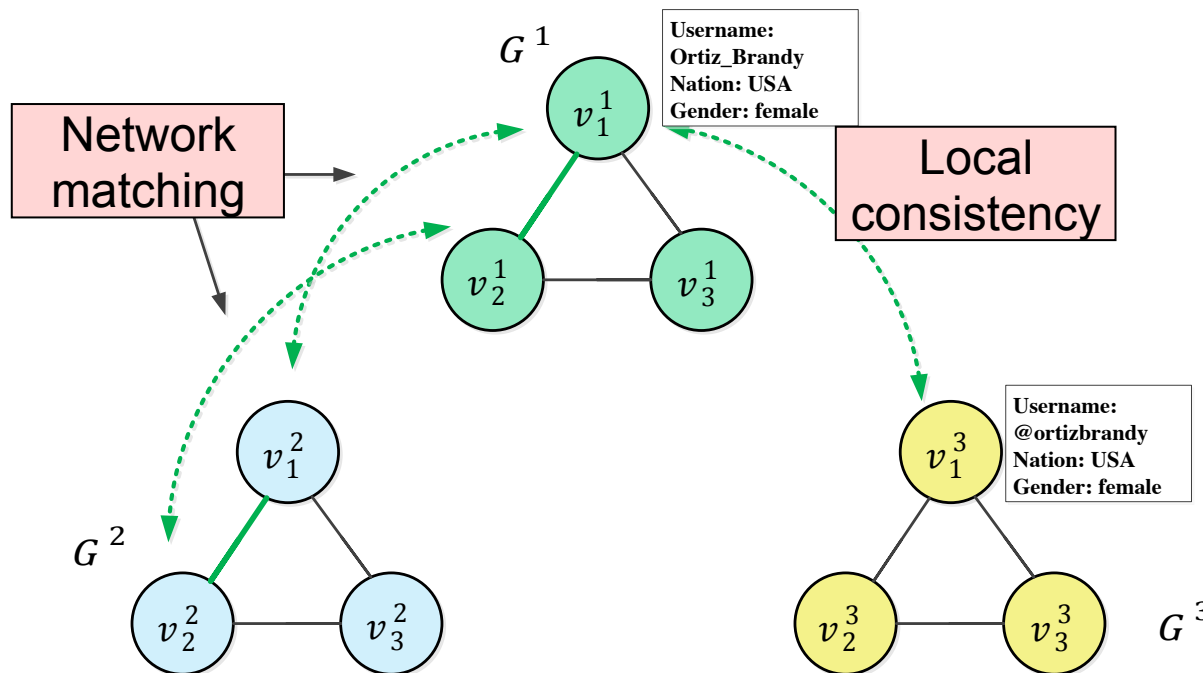
- Username similarity and uniqueness
- Profile content similarity
- Ego network similarity
- Social status

## Energy function

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

# Local vs. Global consistency

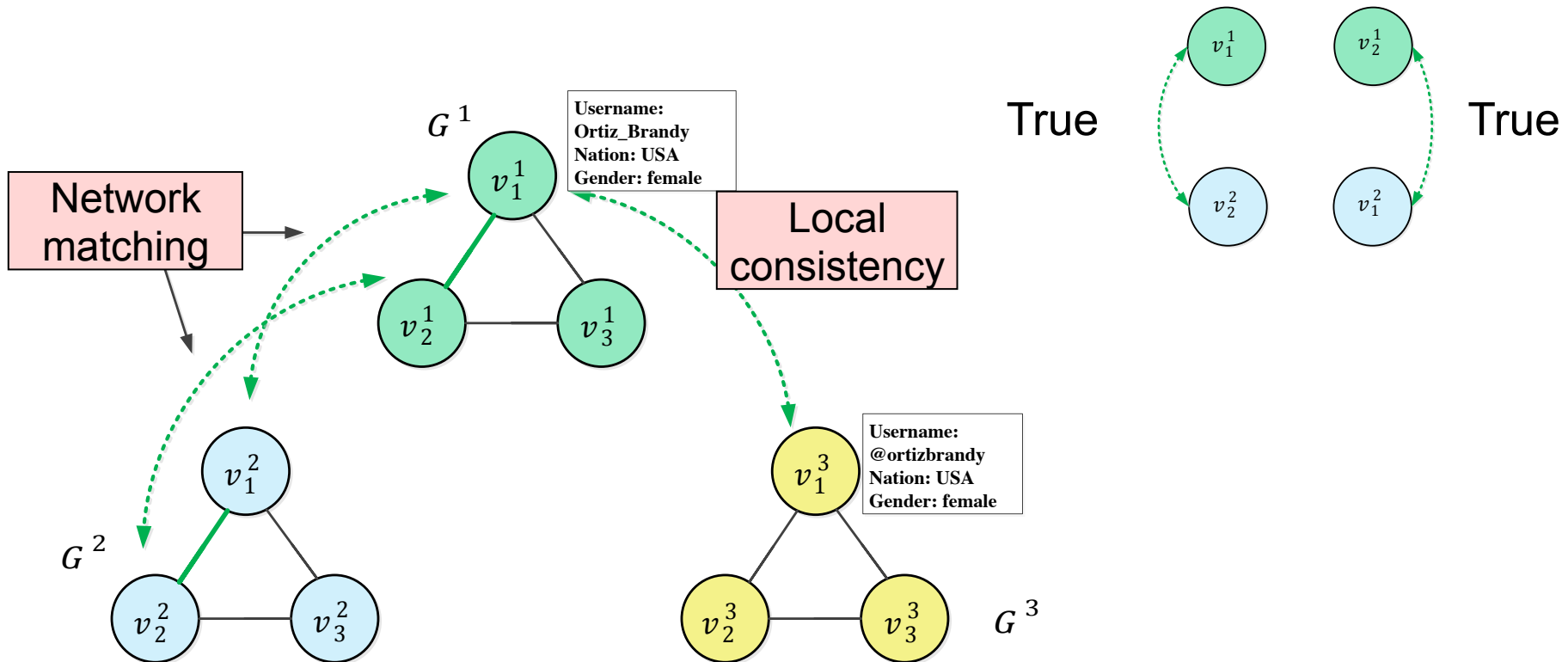
- Network matching: matching users' ego networks



Encourage “neighborhood-preserving matching”

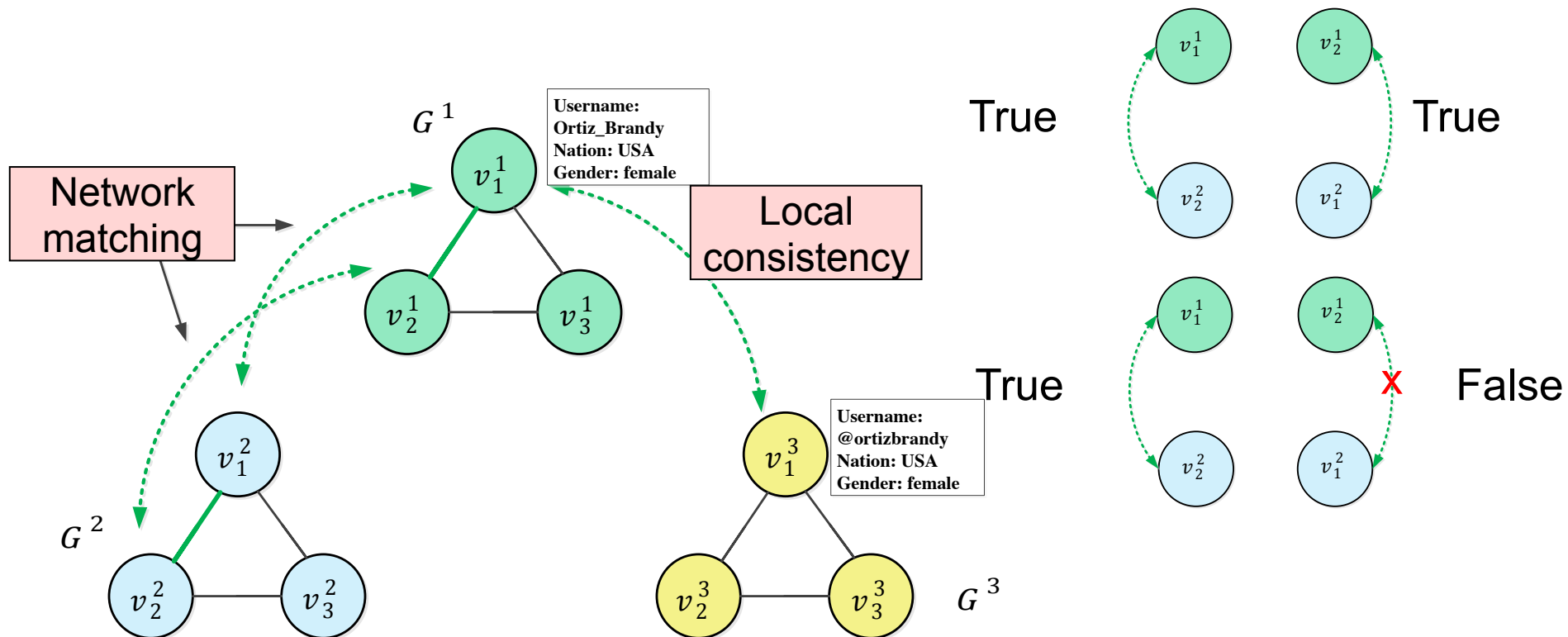
# Local vs. Global consistency

- Network matching: matching users' ego networks



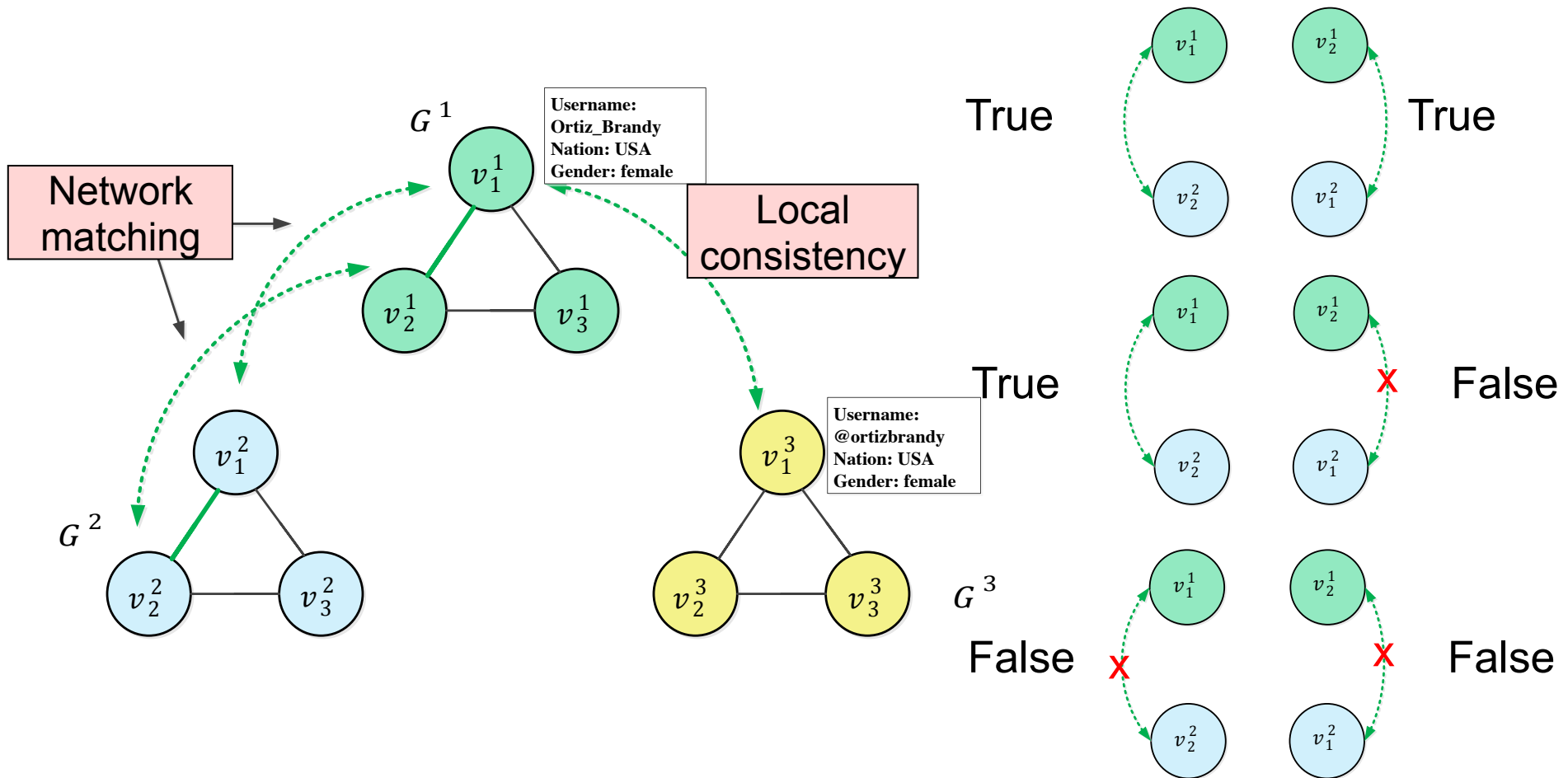
# Local vs. Global consistency

- Network matching: matching users' ego networks



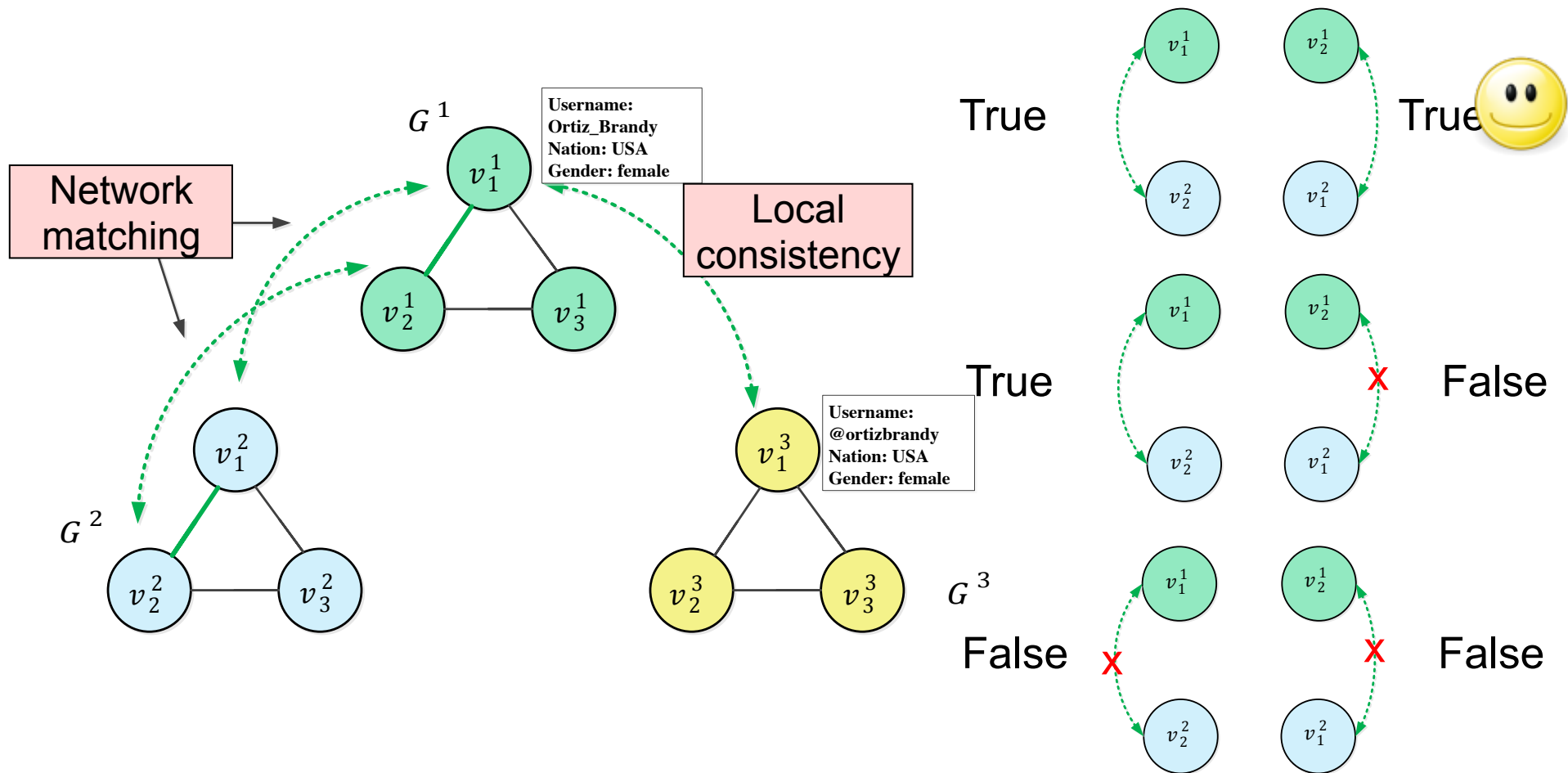
# Local vs. Global consistency

- Network matching: matching users' ego networks



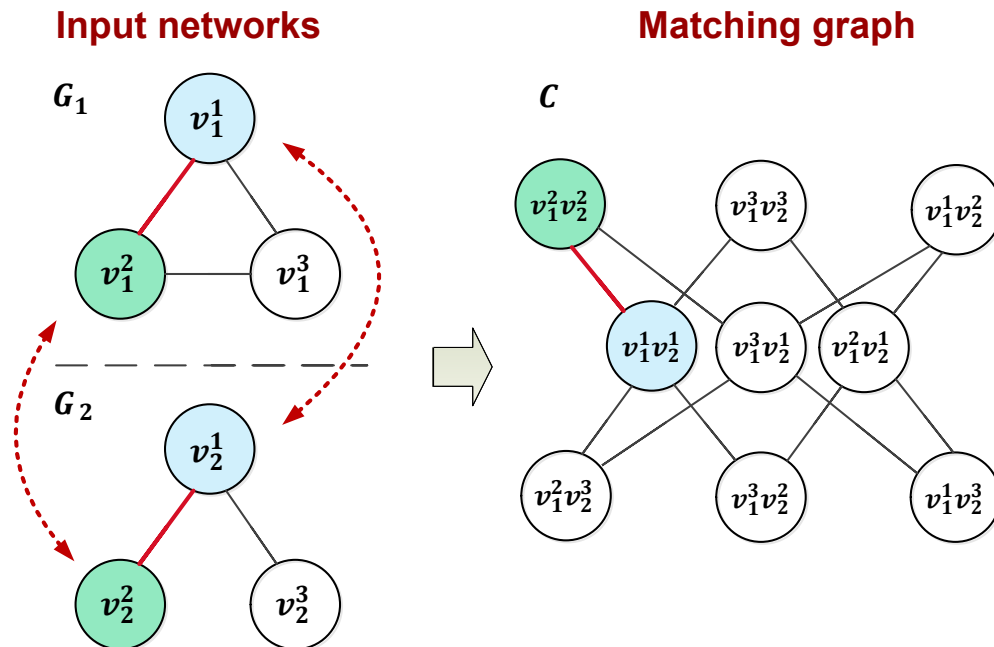
# Local vs. Global consistency

- Network matching: matching users' ego networks



# Network Matching

- Network matching: matching users' ego networks



Energy function

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

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



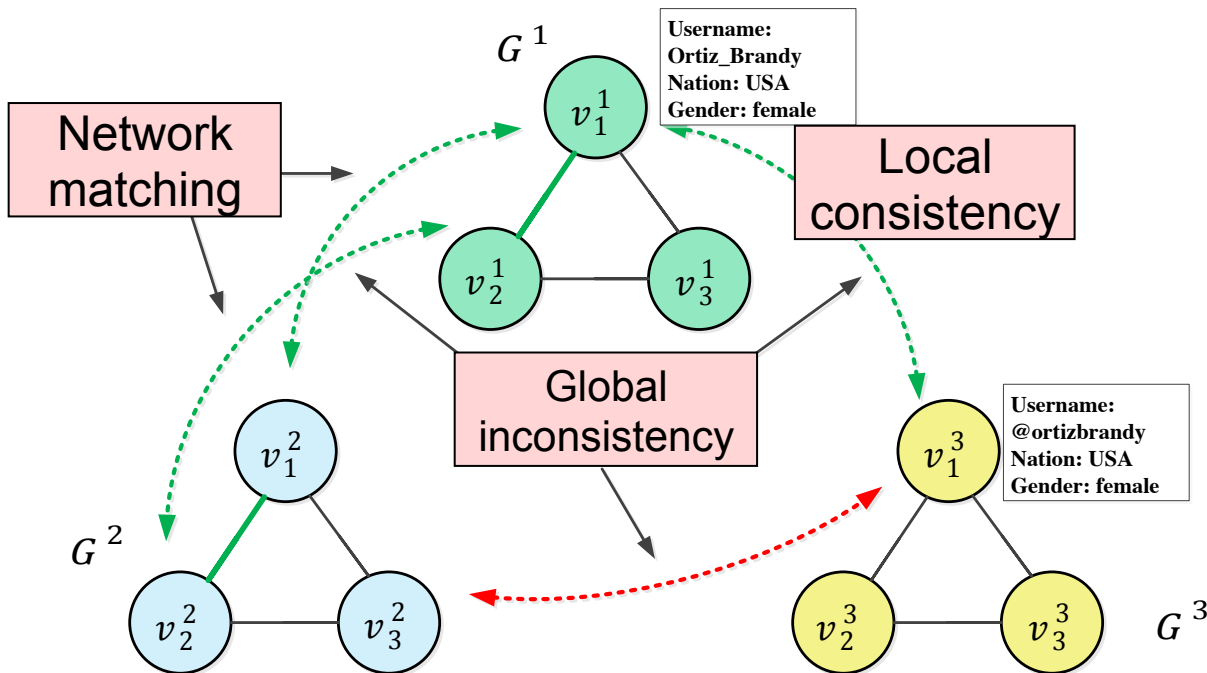


# Candidate Pruning

- 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

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

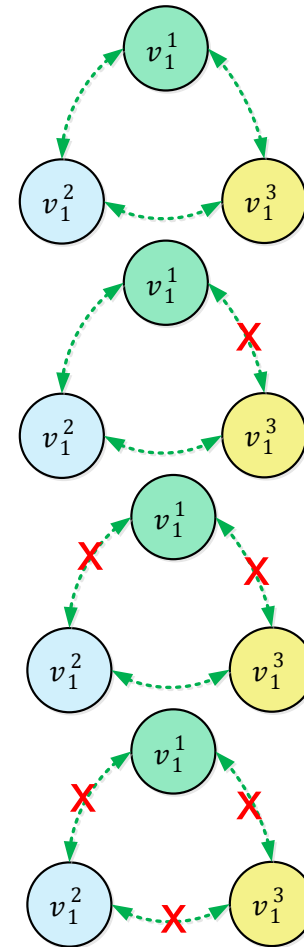
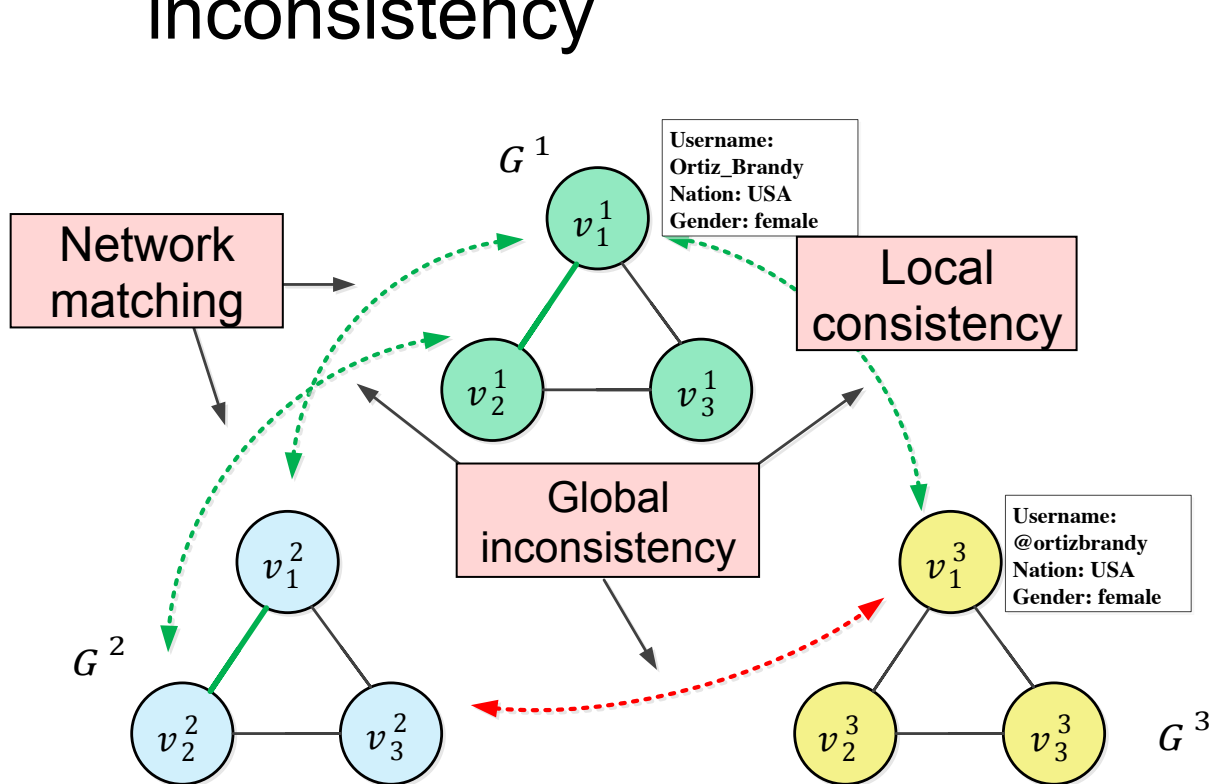
and

For the pair  $\langle \mathcal{V}_{i_n}^2, \mathcal{V}_{i_1}^1 \rangle$ , 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”

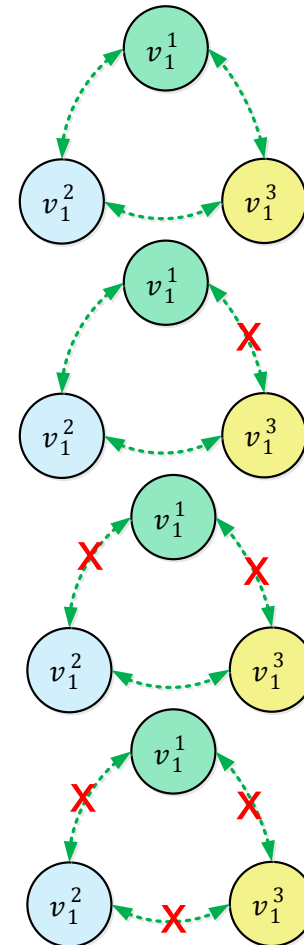
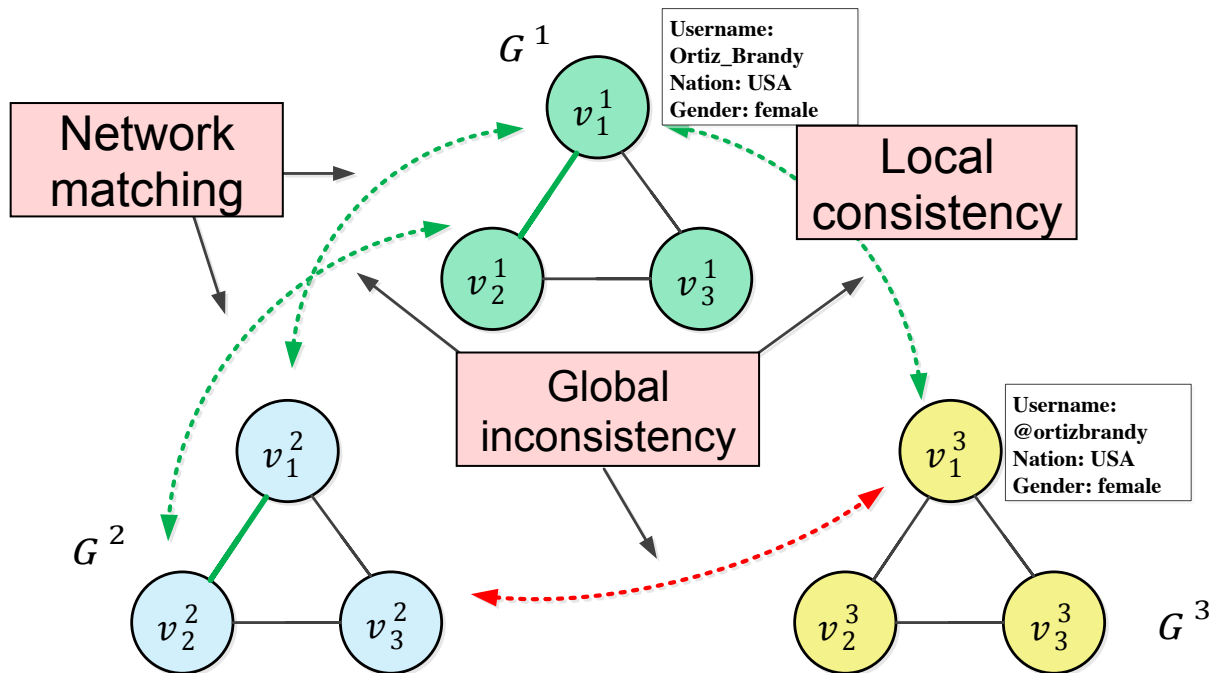
# Local vs. Global consistency

- Global consistency: matching users by avoiding global inconsistency



# Local vs. Global consistency

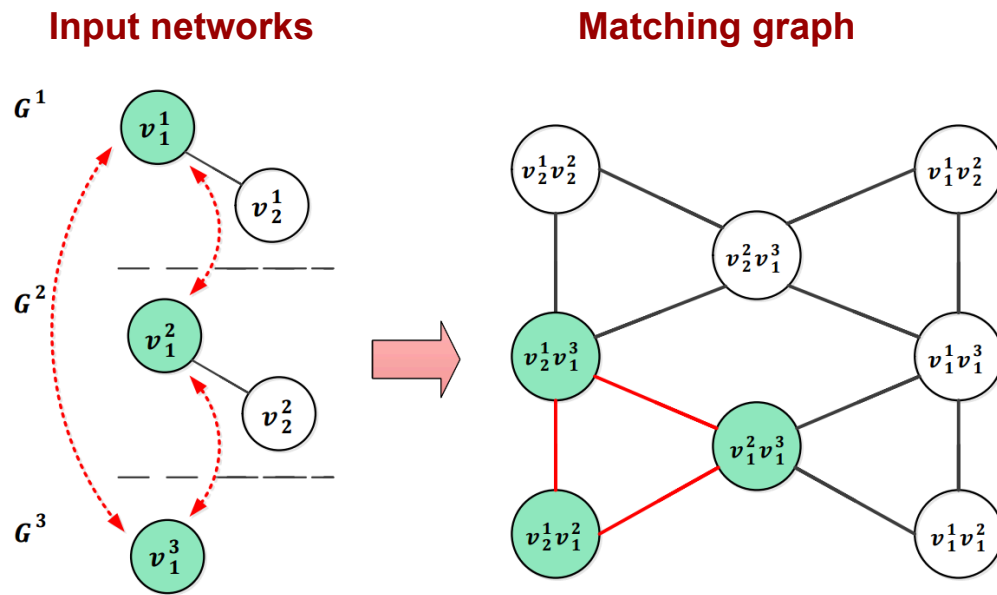
- Global consistency: matching users by avoiding global inconsistency



Inconsistent!



# Avoid Global Inconsistency

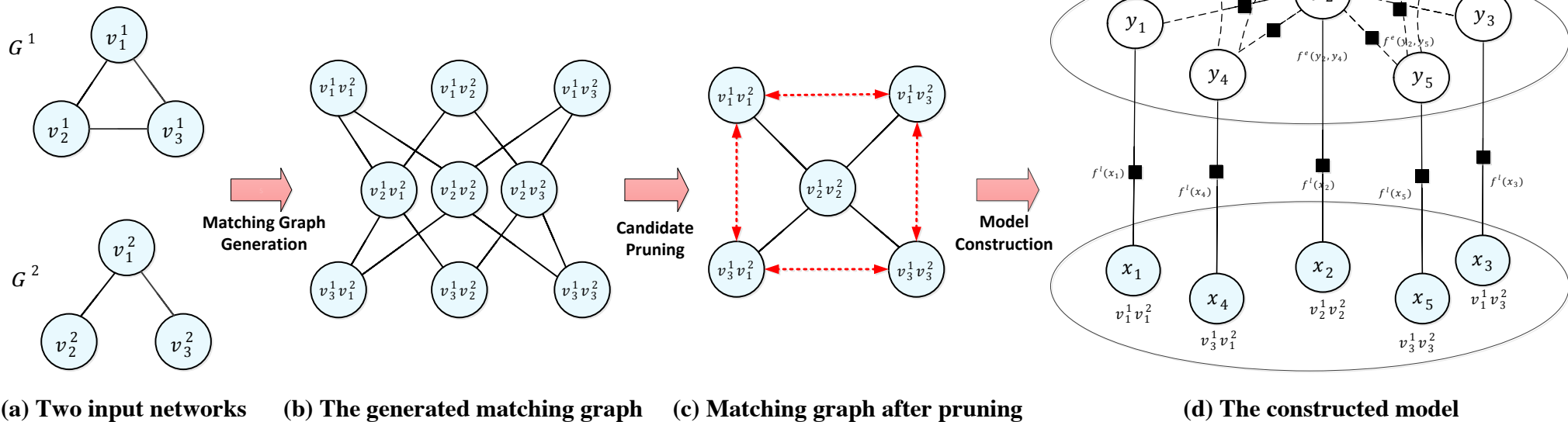


Energy function

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

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

# Model Construction



Objective function by combining all the energy functions

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

# Model Learning

- Max-margin learning

$$\begin{aligned} \min_W \quad & \frac{1}{2} \|W\|^2 + \mu\xi \\ \text{s.t.} \quad & E(\hat{Y}, X; W) \leq E(Y, X; W) - \Delta(Y, \hat{Y}) + \xi \end{aligned}$$

- 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

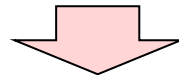
$$\begin{aligned} E(Y, X; W) &= \sum_{f \in \mathcal{F}} E_f(Y_f, X_f; W) \\ &= \sum_{f \in \mathcal{F}} \sum_{\mathbf{x}_i \in X_f} \left( \frac{1}{|\mathcal{F}_i|} \mathbf{w}_i^\top \mathbf{g}_l(\mathbf{x}_i, y_i^f) + \mathbf{w}_f^\top f(Y_f) \right) \\ \text{s.t.} \quad & y_i^f = y_i, \quad \forall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

# Model Learning (cont.)

- Dual decomposition

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

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



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

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

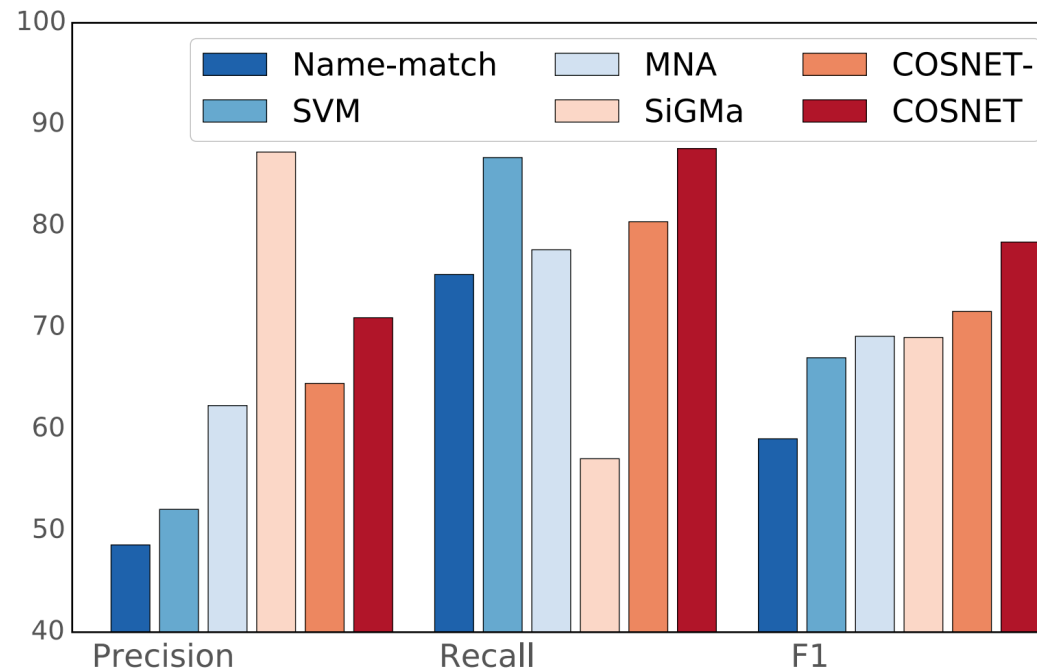
The resulting objective function is convex and non-differentiable, and can be solved by **projected sub-gradient** method



# Results

# Connecting AMiner with ...

- LinkedIn and VideoLectures

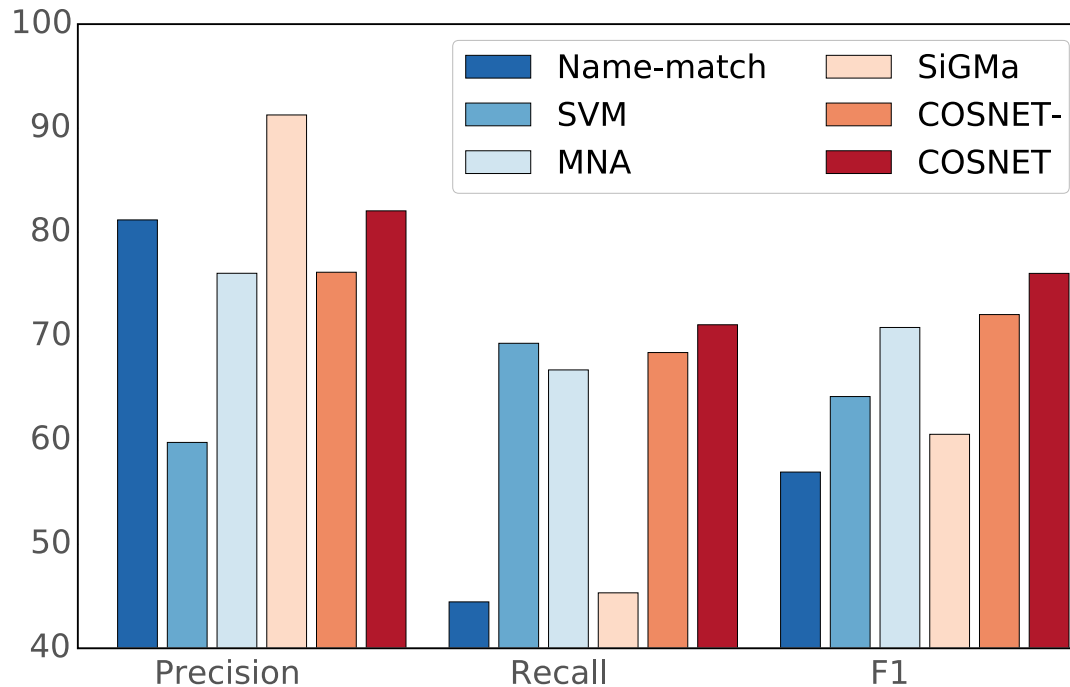


**Name-match:** match name only;  
**SVM:** use classifier to identify the same user;  
**MNA:** an optimization method;

**SiGMa:** local propagation;  
**COSNET:** our method;  
**COSNET-:** w/o global consistency.

# Connecting Social Media Sites

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



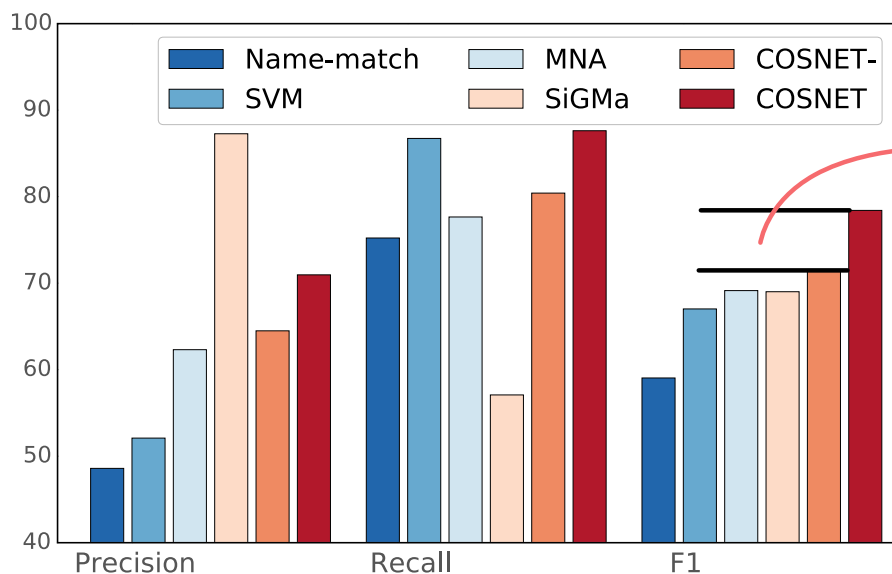
**Name-match:** match name only;  
**SVM:** use classifier to identify the same user;  
**MNA:** an optimization method;

**SiGMa:** local propagation;  
**COSNET:** our method;  
**COSNET-:** w/o global consistency.

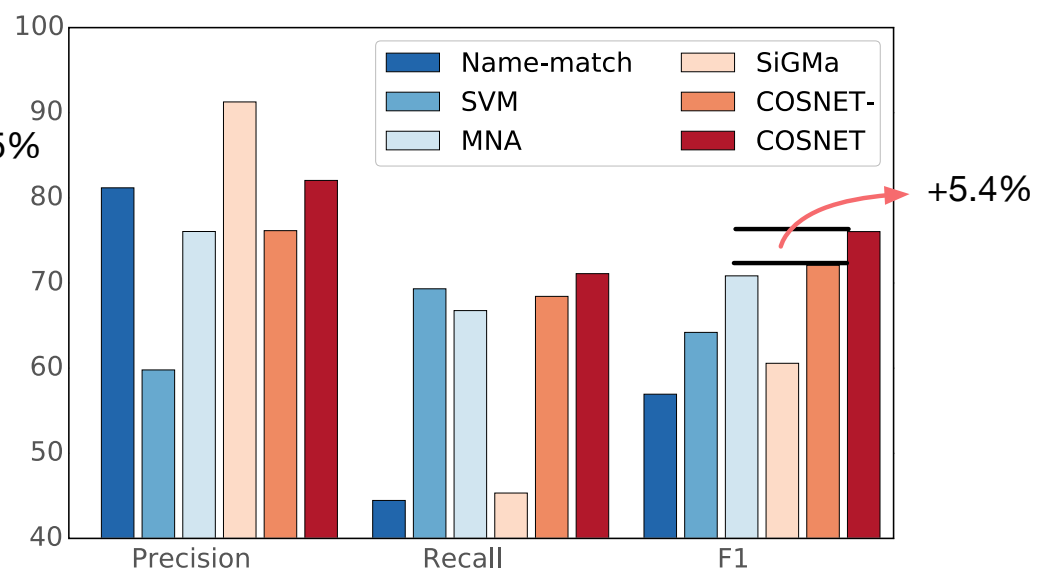
# Effects of Global Consistency



COSNET-: w/o global consistency.



Academia Collection



SNS Collection

# Application in AMiner

**videolectures.net** - Video contents  
exchange ideas & share knowledge

**in** - Personal profiles  
- Business connections  
- Skills and expertise

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

**uspto** - Patents data

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

**AMiner** Whatever comes to your mind

**Jiawei Han (韩家炜)** Following

Department of Computer Science, University of Illinois at Urbana-Champaign  
Professor  
(217) 333-6903  
hanj@cs.uiuc.edu  
http://www.cs.uiuc.edu/~hanj/

**Research Interests**

- Data Mining
- Information Extraction
- Data Analysis
- Machine Learning
- Text Mining

**Similar Authors**

**Ego Network**

**Timeline**

1985 1990 1995 2000 2005 2010 2014

**Papers** 790

**Lectures** 13

**Patents** 1

**Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions**  
Wei Shen, Jianyong Wang, **Jiawei Han**  
Knowledge and Data Engineering, IEEE Transactions (2015)  
Bibtex <http://dx.doi.org/10.1109/TKDE.2014.2327028>

# Thanks!



**Data & source code**

<http://aminer.org>

<http://aminer.org/cosnet>