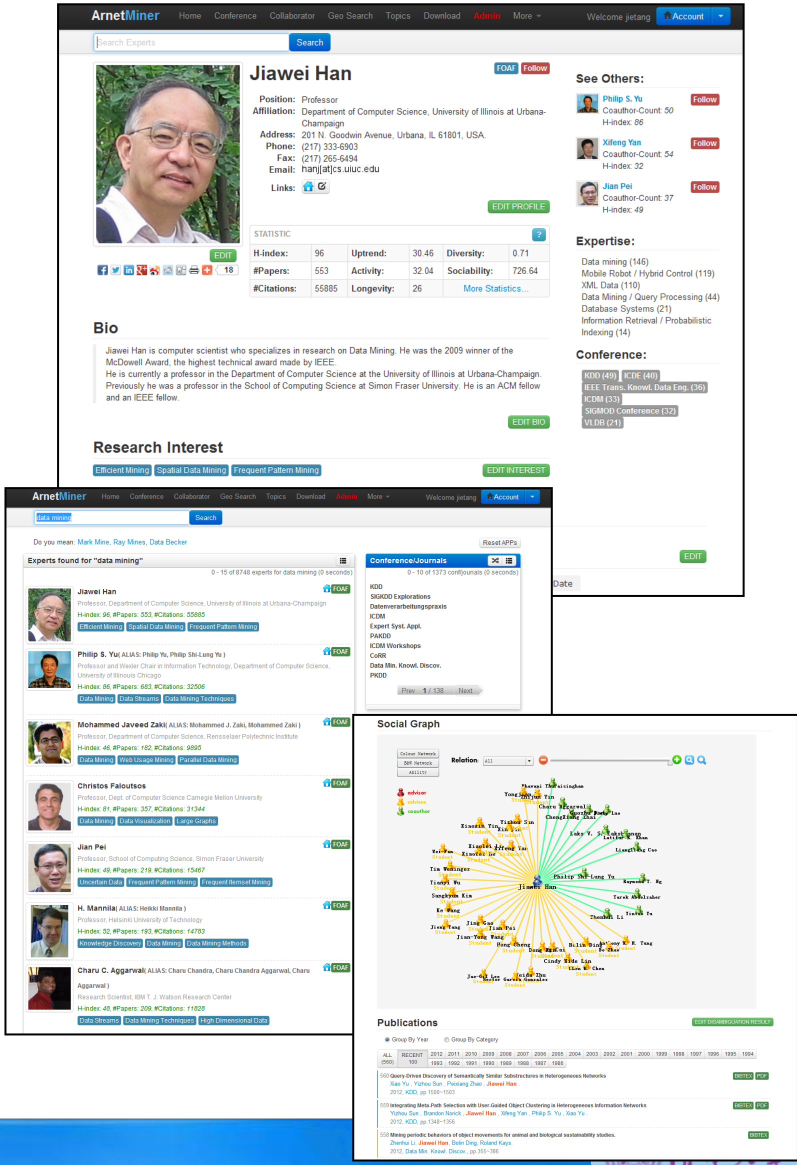
COSNET: Connecting Heterogeneous Social Networks with Local and Global Consistency

Yutao Zhang+, Jie Tang+, Zhilin Yang+, Jian Pei#, and Philip S. Yu*

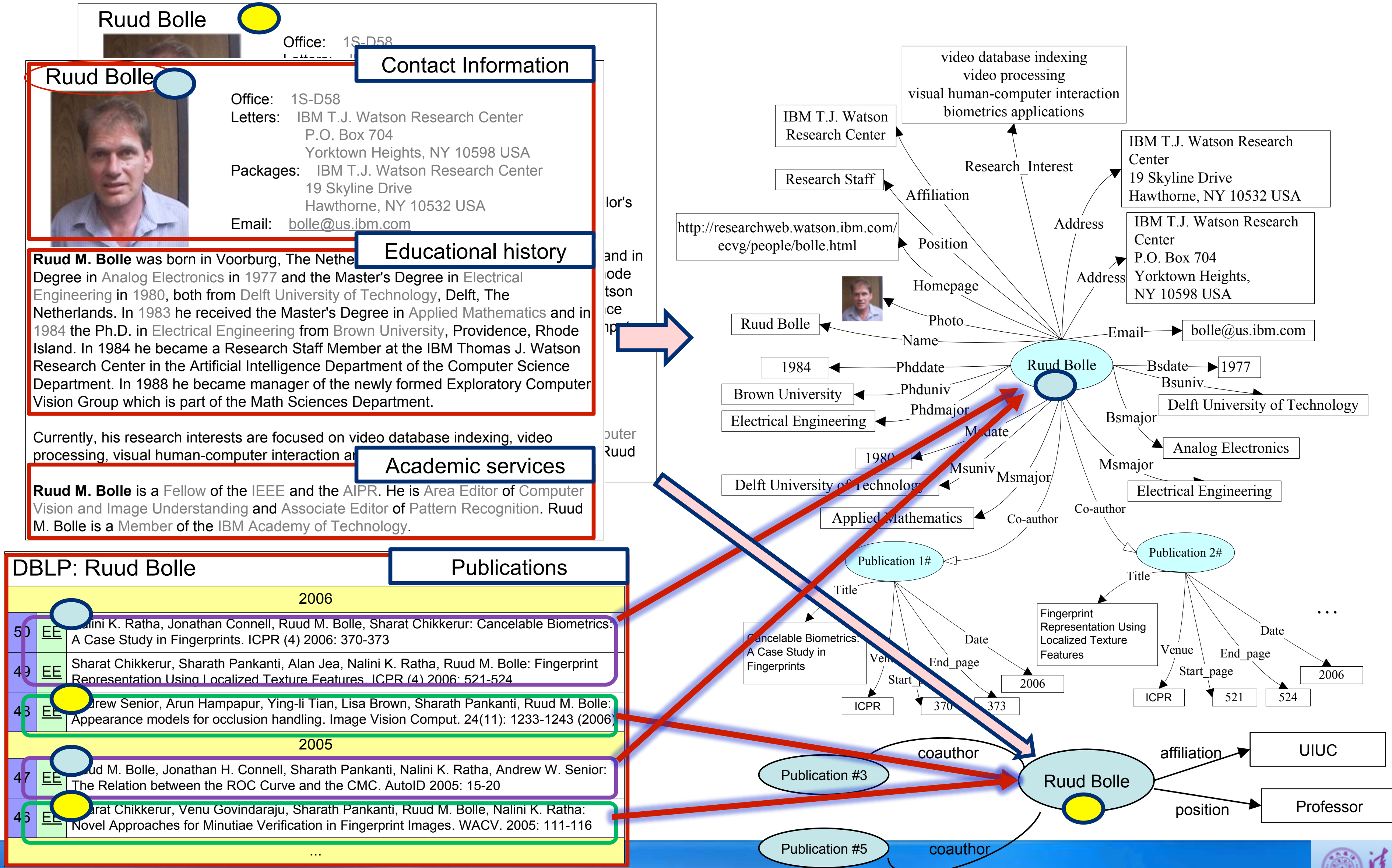
+Tsinghua University #Simon Fraser University *University of Illinois at Chicago

AMiner II (ArnetMiner)

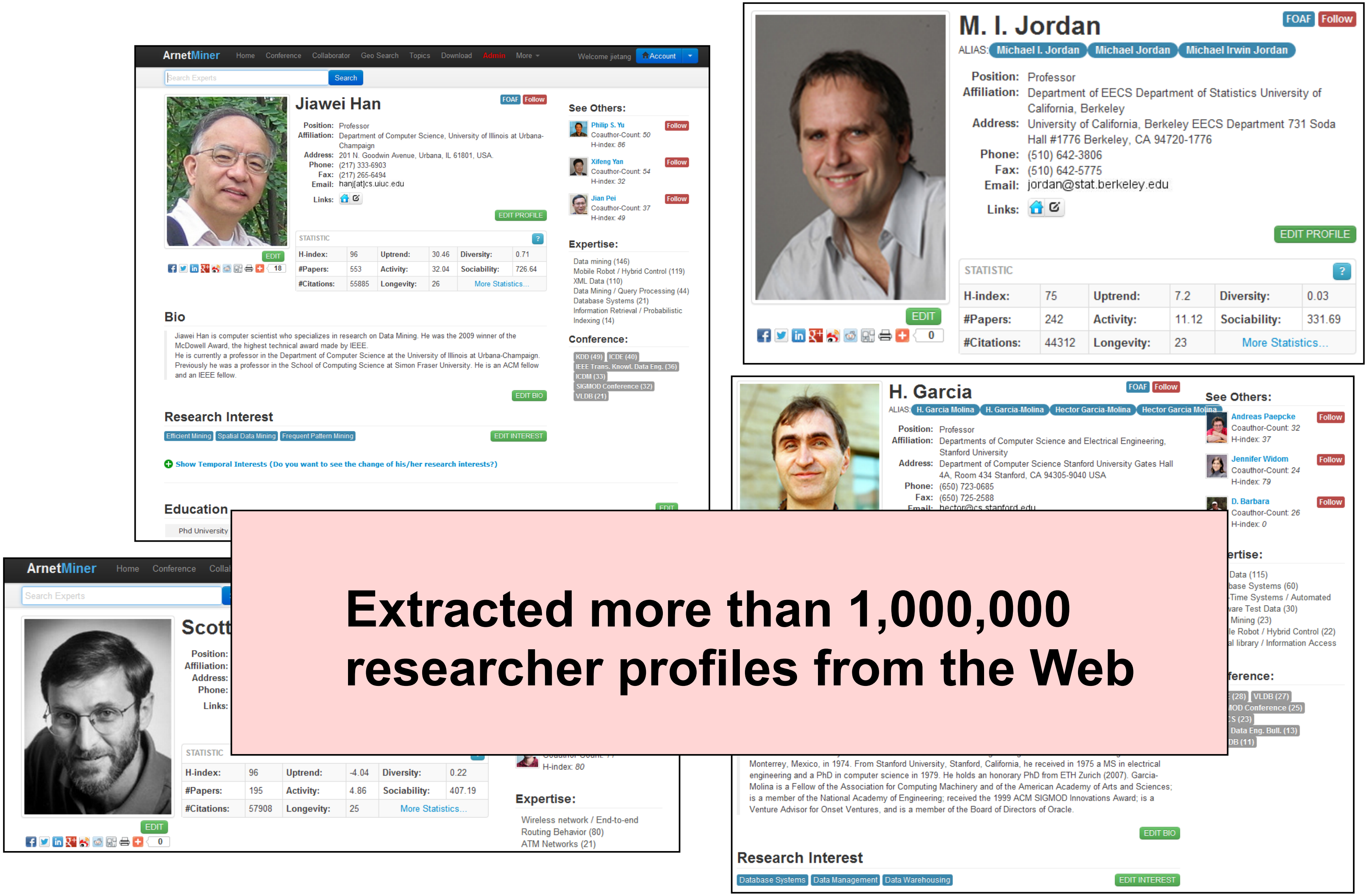
- Academic Social Network Analysis and Mining system—AMiner (http://aminer.org)
  - Online since 2006
  - **>38 million** researcher profiles
  - **>100 million** publications
  - **>241 million** requests
  - **>12.35 Terabyte** data
  - 100K IP access from 170 countries per month
  - 10% increase of visits per month
- Deep analysis, mining, and search



Knowledge Acquisition from the Web
(ACM TKDD, WWW'12, ISWC'06, ICDM'07, ACL'07)



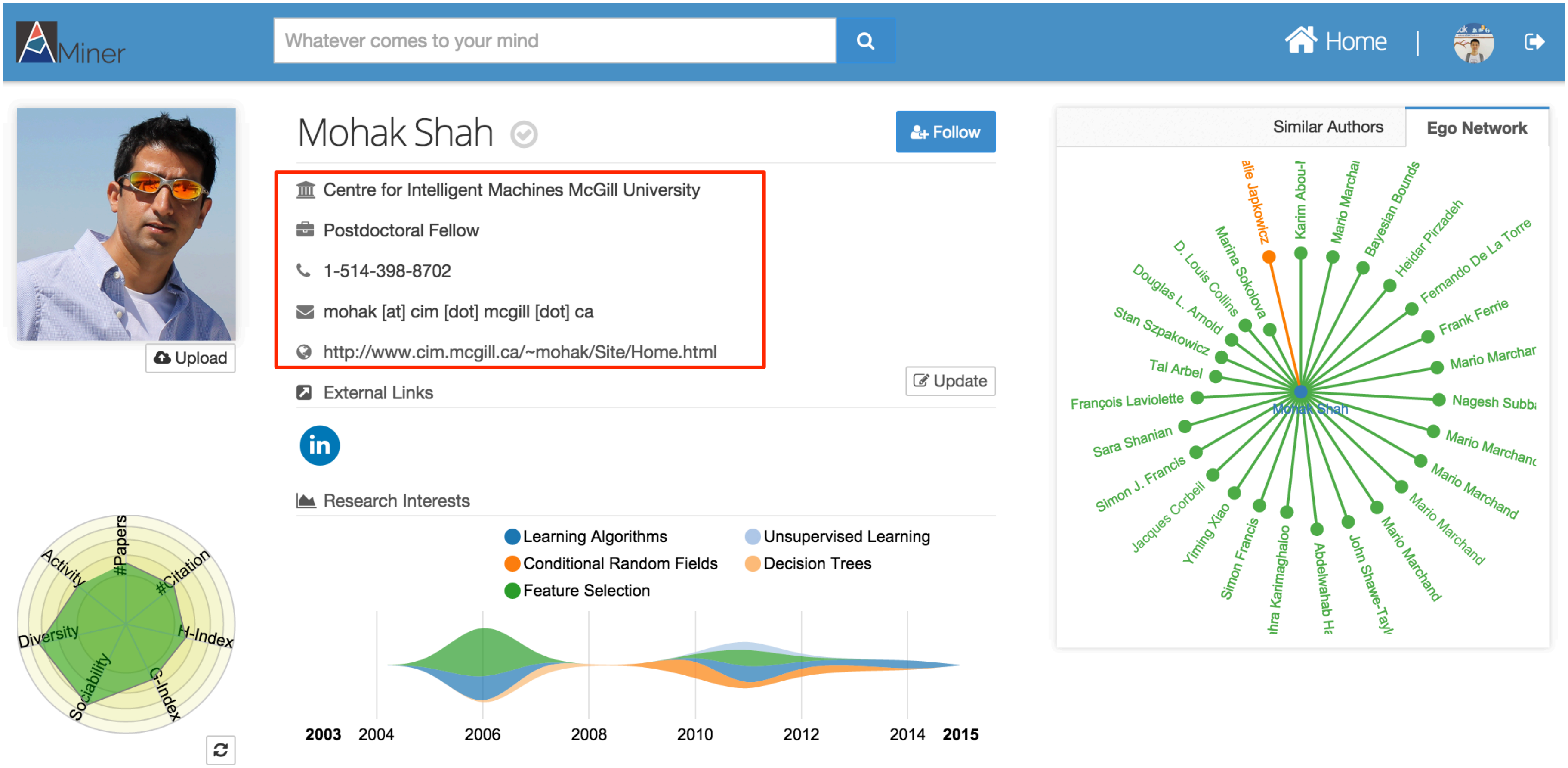
Researcher Profile Database[1]



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

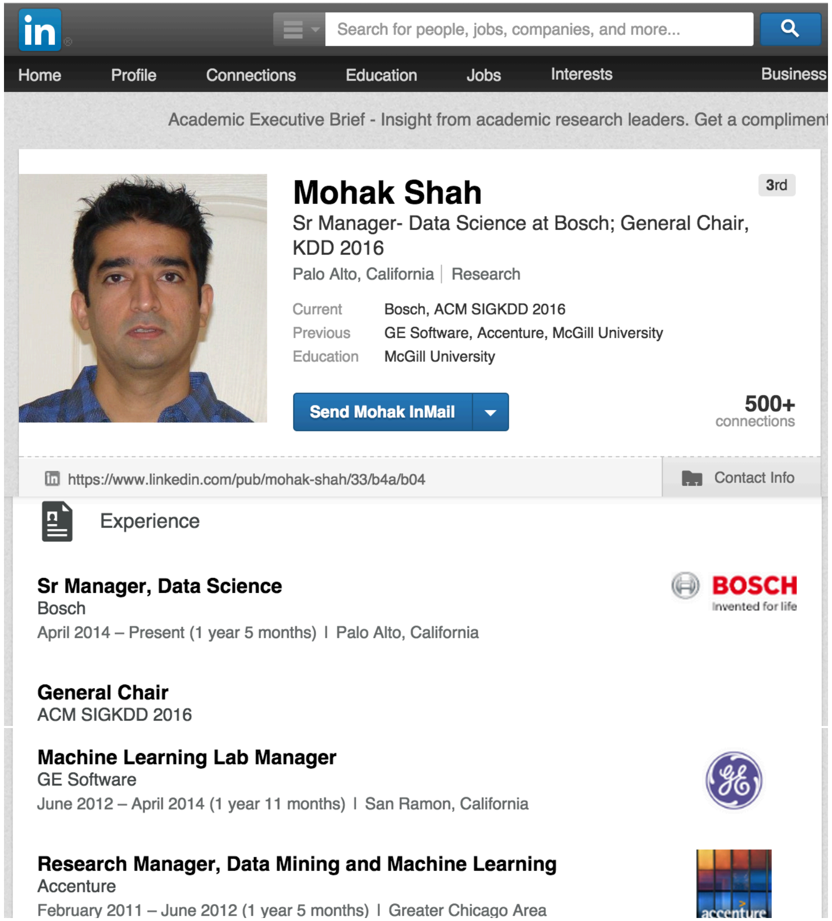
[1] J. Tang, L. Yao, D. Zhang, and J. Zhang. A Combination Approach to Web User Profiling. ACM Transactions on Knowledge Discovery from Data (TKDD), (vol. 5 no. 1), Article 2 (December 2010), 44 pages.

Is this Enough?



Required semantics are distributed in multiple sources

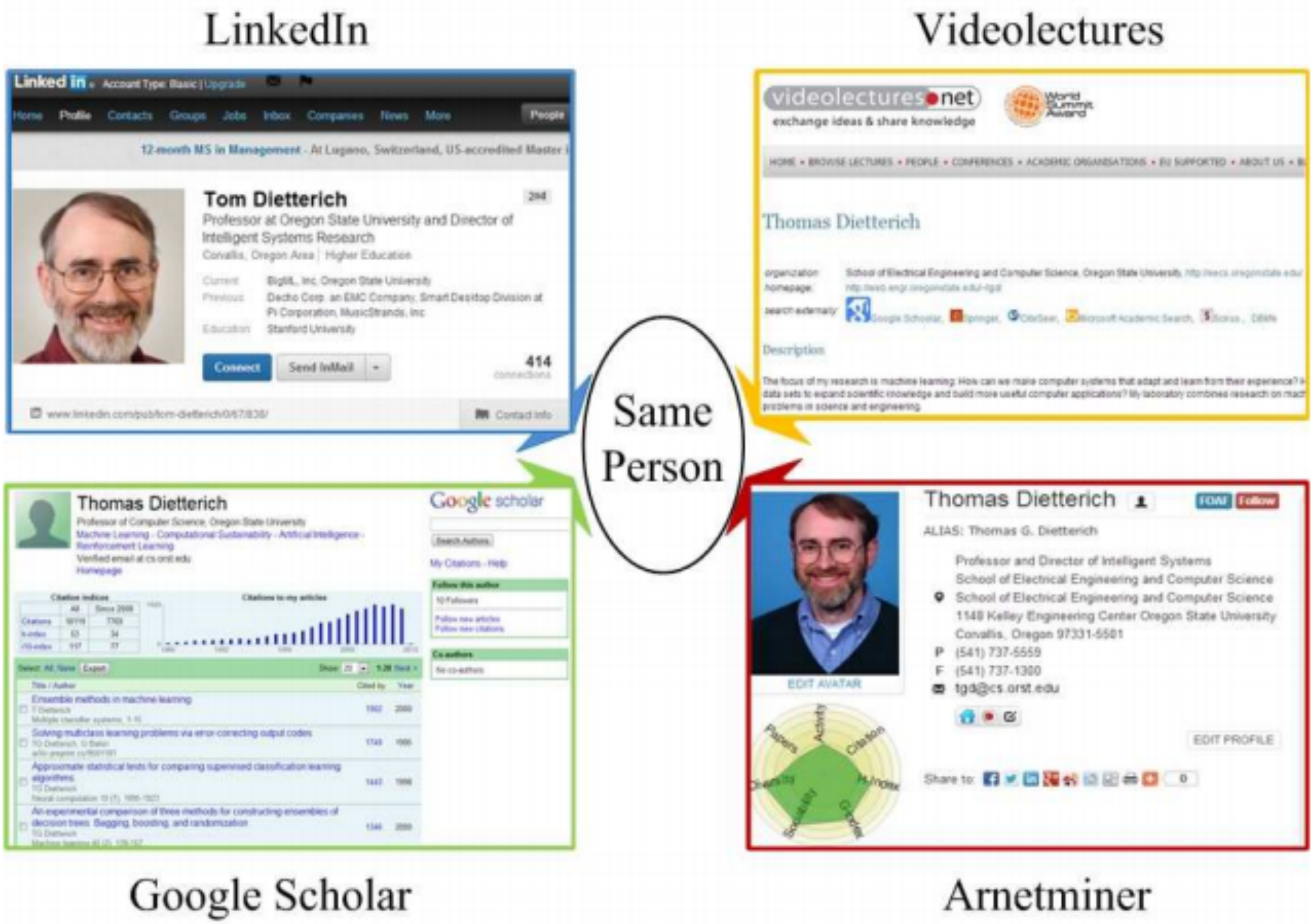
LinkedIn



Videolectures

Identity Linking

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



COSNET: Connecting Social Networks with Local and Global Consistency

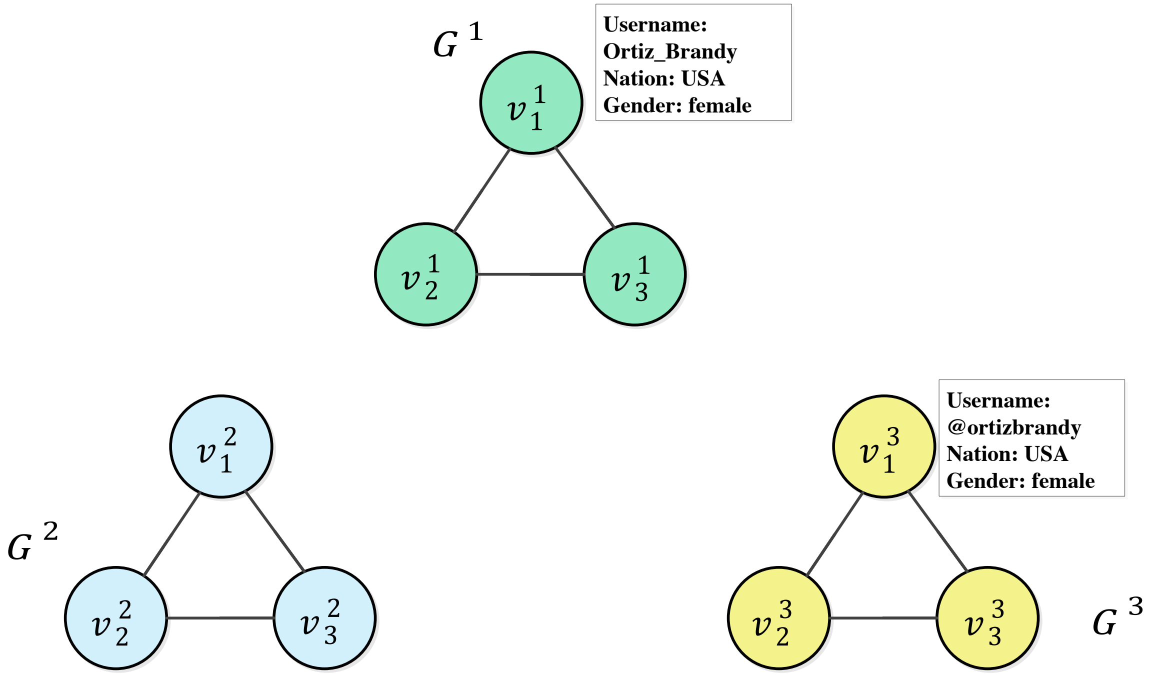
- Input: $\mathbf{G}=\{G^1, G^2, \ldots, G^m\}$, with $G^k=(V^k, E^k, R^k)$
- Formalization: $\mathbf{X}=\{x_i\}$, all possible pairwise matchings and each corresponds to $y_i \in \{1,0\}$
- COSNET: an energy-based model

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

[1] Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, and Philip Yu. COSNET: Connecting Heterogeneous Social Networks with Local and Global Consistency. KDD'15.

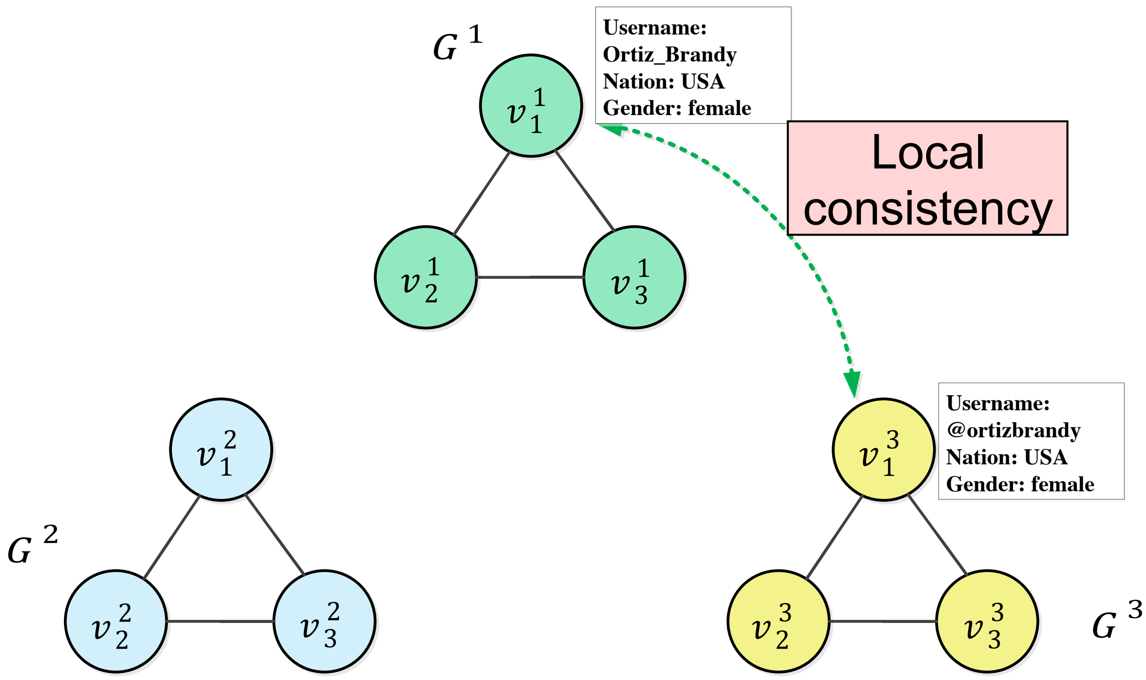
Local vs. Global consistency

- Given three networks,



Local vs. Global consistency

- Local matching: matching users by profiles



Pairwise similarity features

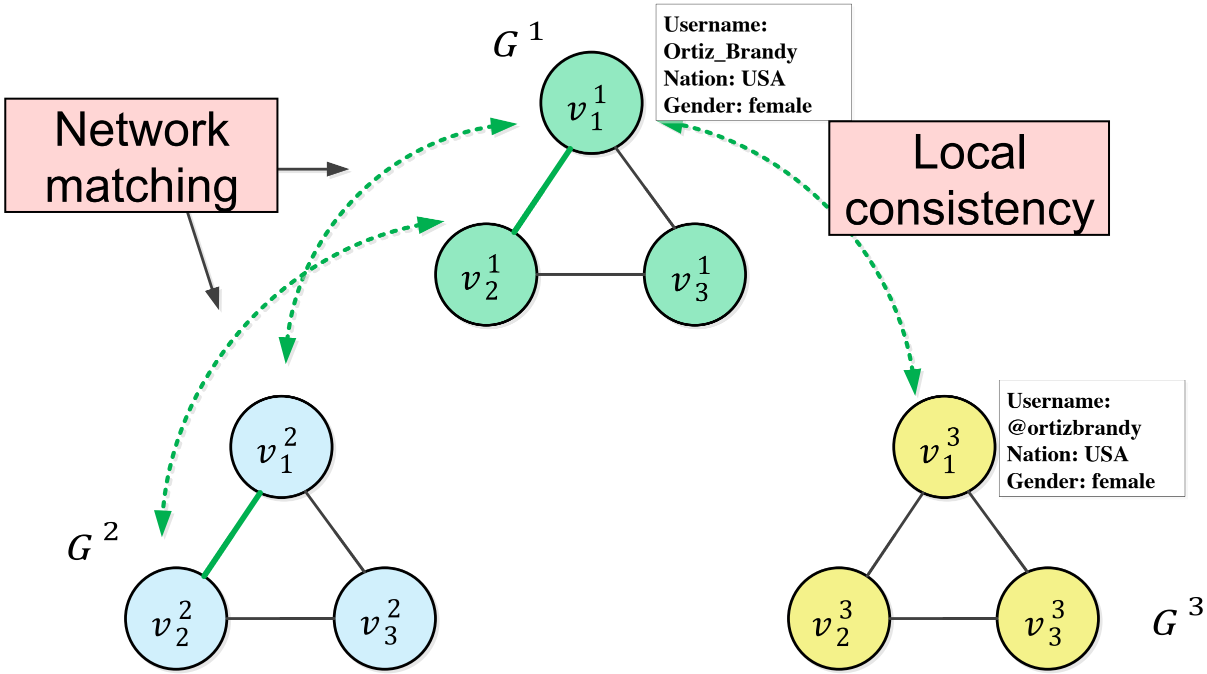
- Username similarity and uniqueness
- Profile content similarity
- Ego network similarity
- Social status

Energy function

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

Local vs. Global consistency

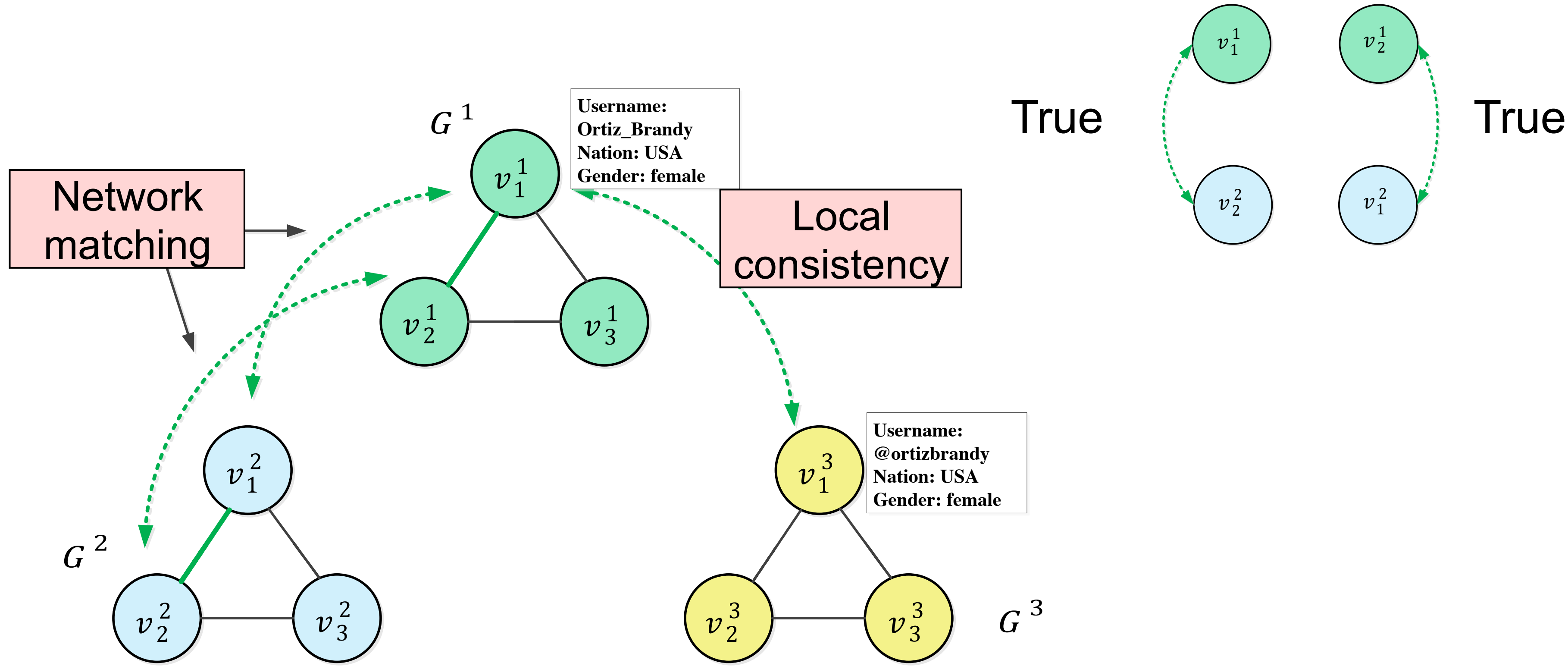
- Network matching: matching users' ego networks



Encourage "neighborhood -preserving matching"

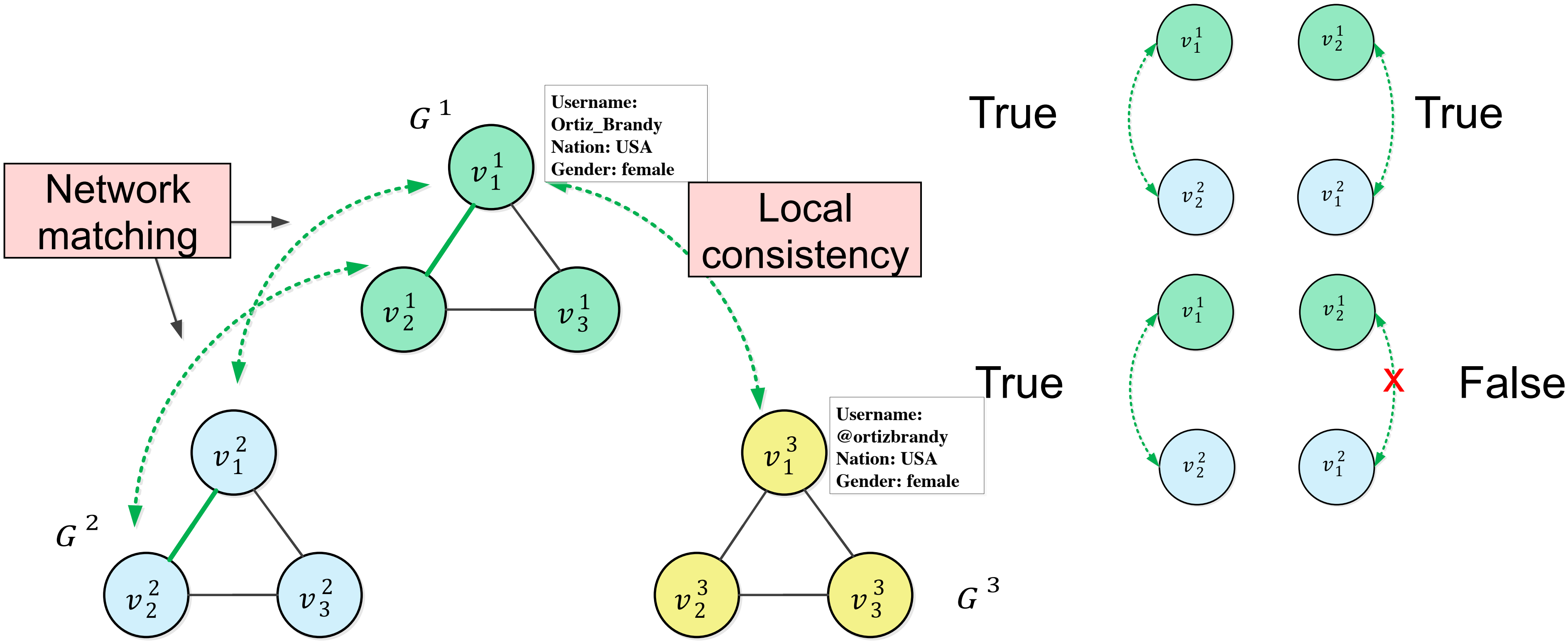
Local vs. Global consistency

- Network matching: matching users' ego networks



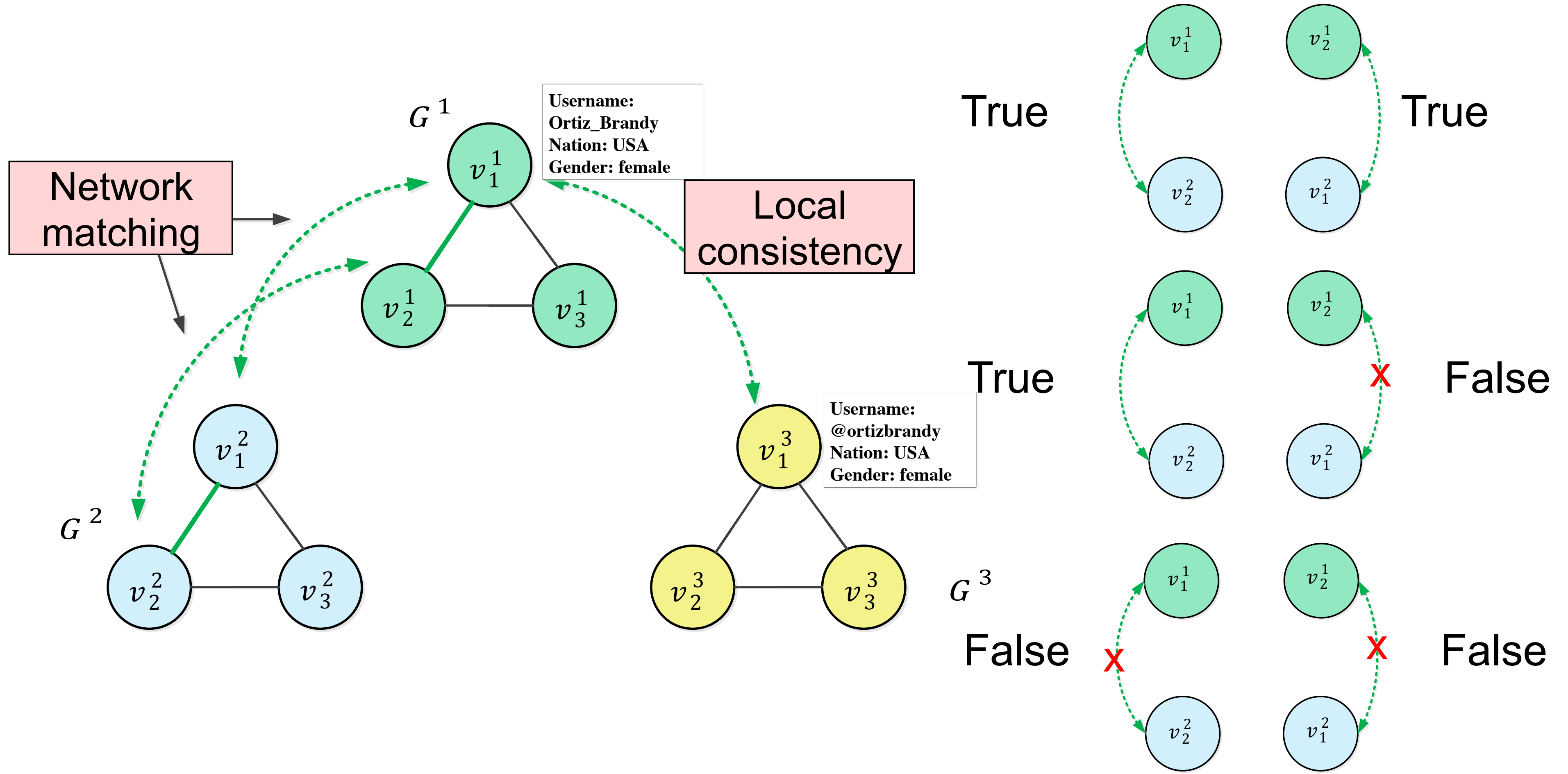
Local vs. Global consistency

- Network matching: matching users' ego networks



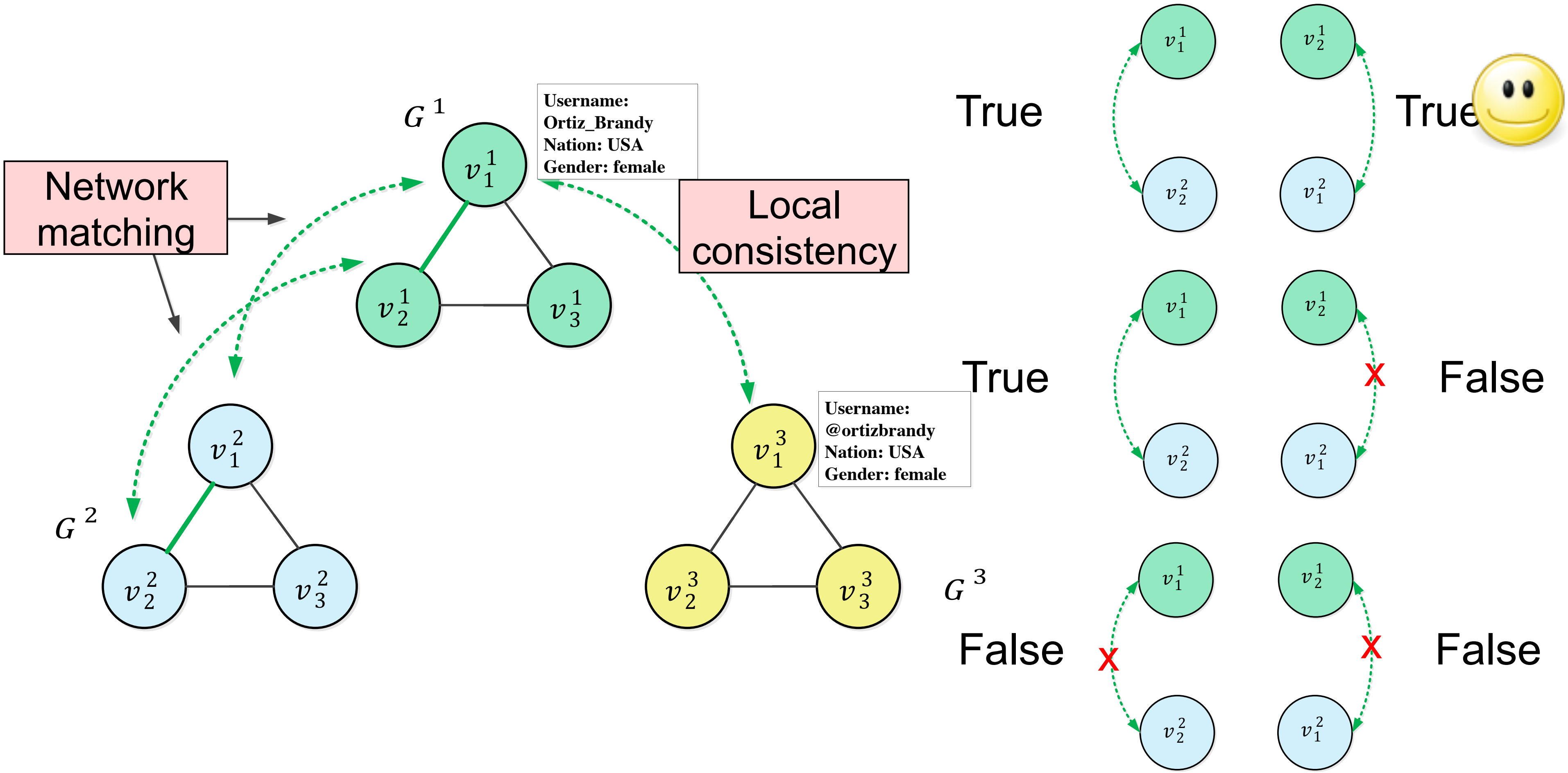
Local vs. Global consistency

- Network matching: matching users' ego networks



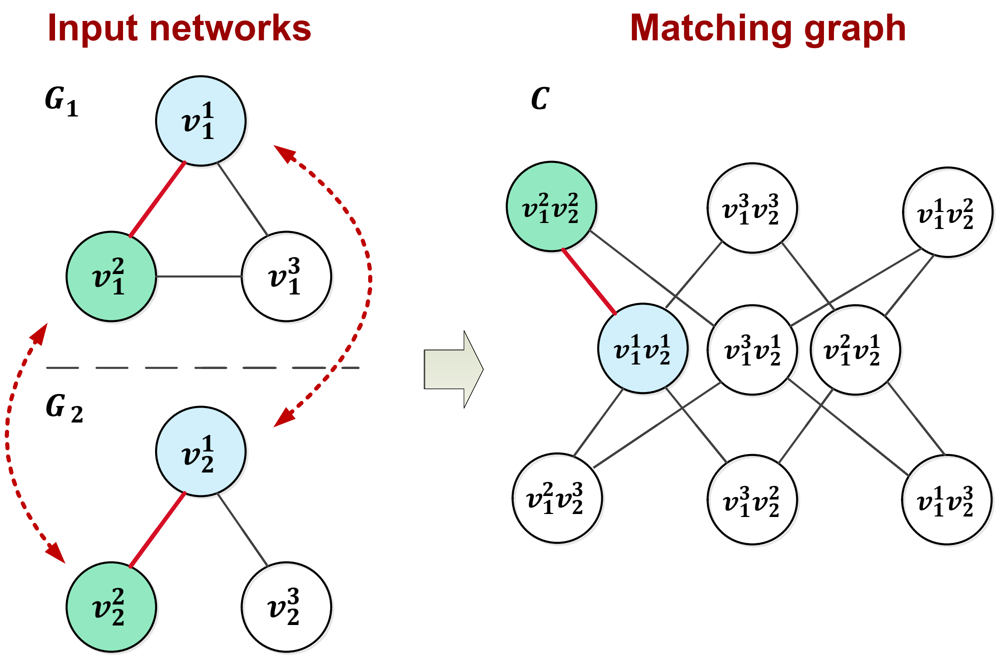
Local vs. Global consistency

- Network matching: matching users' ego networks



Network Matching

- Network matching: matching users' ego networks



Energy function

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

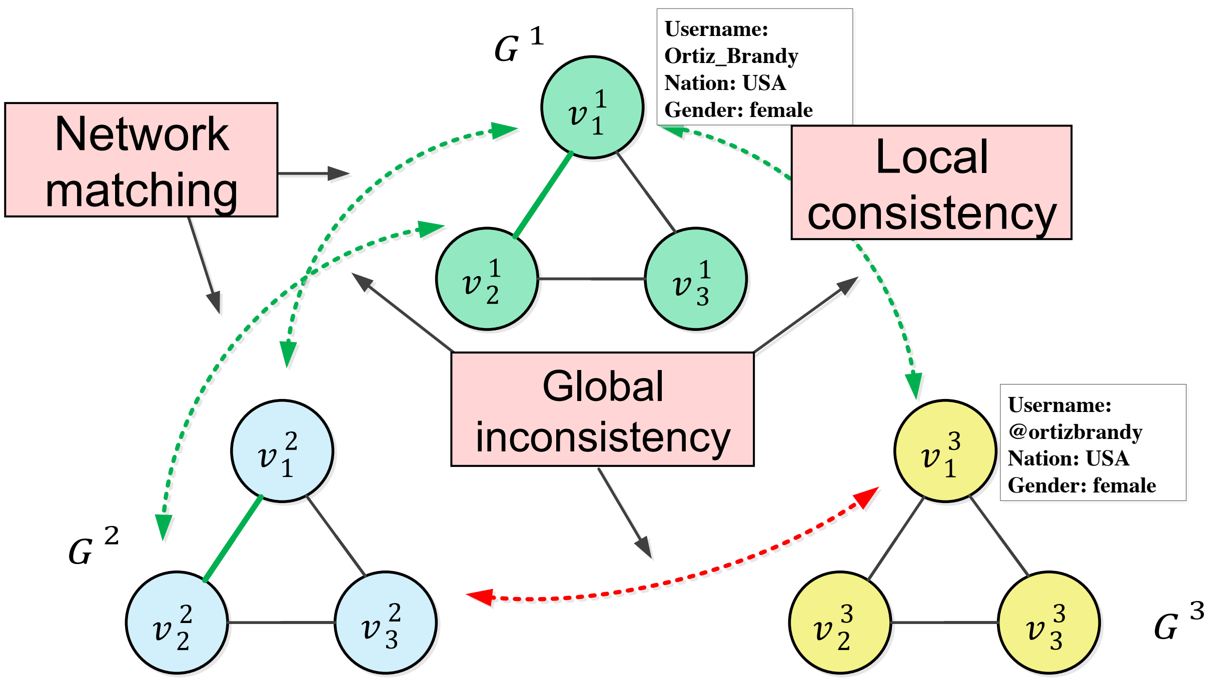
$$\mathbf{f}_e(y_i, y_j) = \begin{cases} (1,0,0)^{\intercal} & \text{if } y_i = y_j = 0 \\ (0,1,0)^{\intercal} & \text{if } y_i + y_j = 1 \\ (0,0,1)^{\intercal} & \text{if } y_i = y_j = 1 \end{cases}$$

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* $\langle \mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \cdots, \mathbf{x}_{i_n} \rangle$*, such that*

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

and

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

and

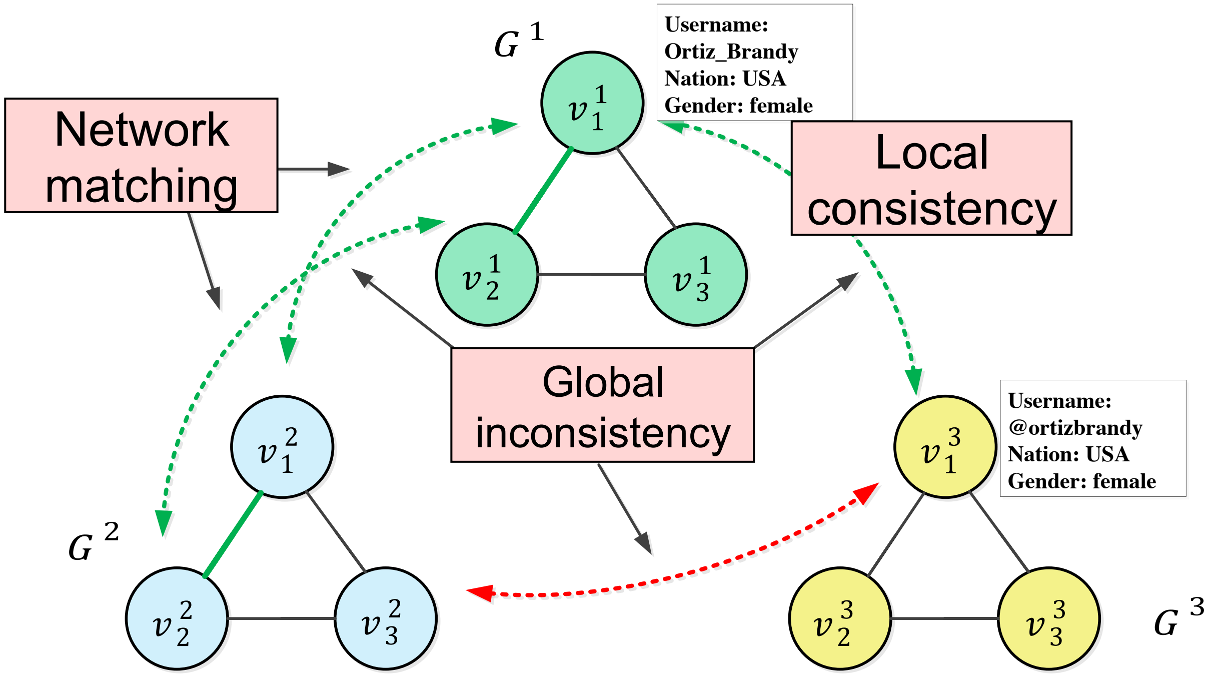
For the pair $\langle \mathcal{V}_{i_n}^2, \mathcal{V}_1^1 \rangle$*, if the corresponding label* $y_j = 0$

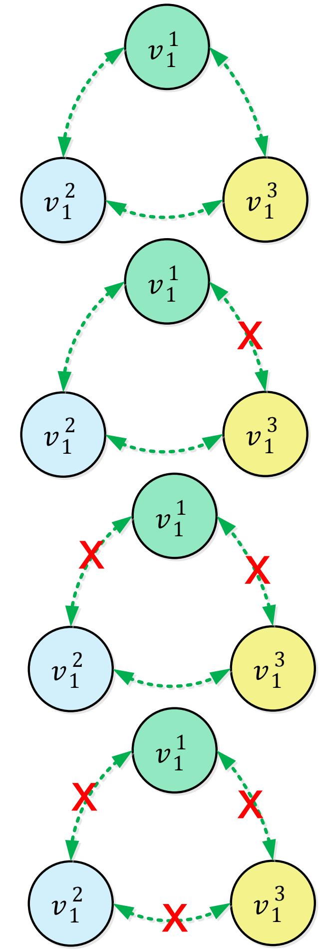
then we say that the assigned labels Y *causes global inconsistency given* $\mathbf{G}$ *and* X.

Avoid "global inconsistency"

Local vs. Global consistency

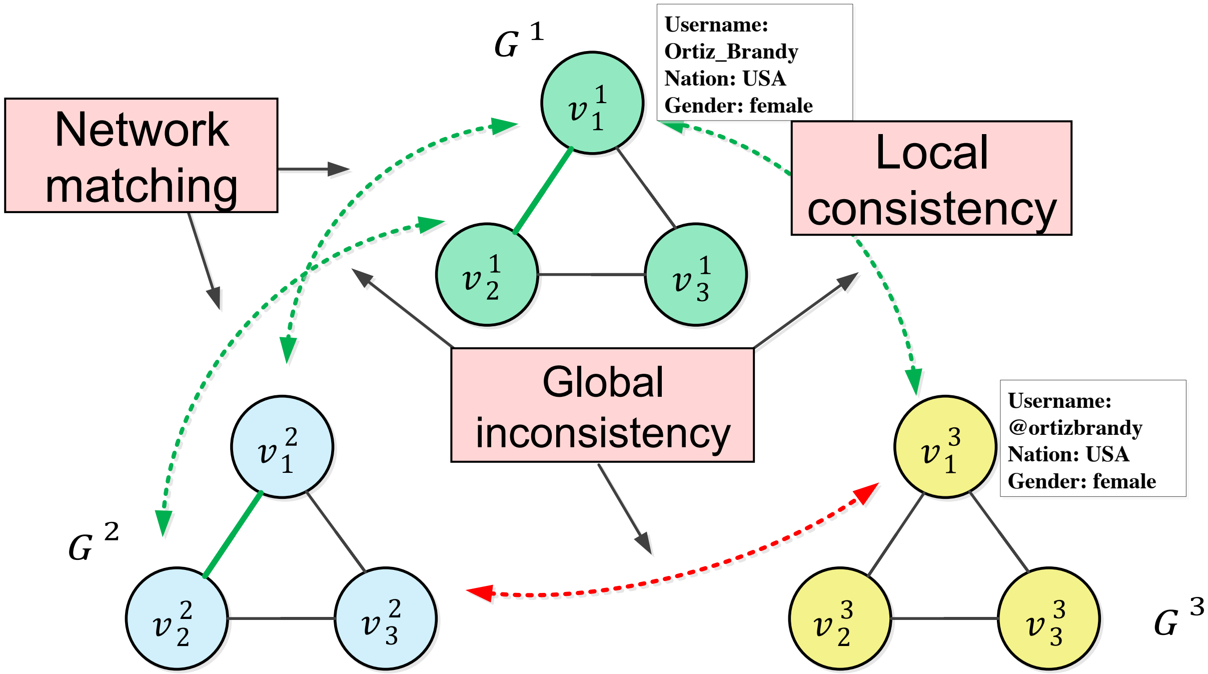
- Global consistency: matching users by avoiding global inconsistency

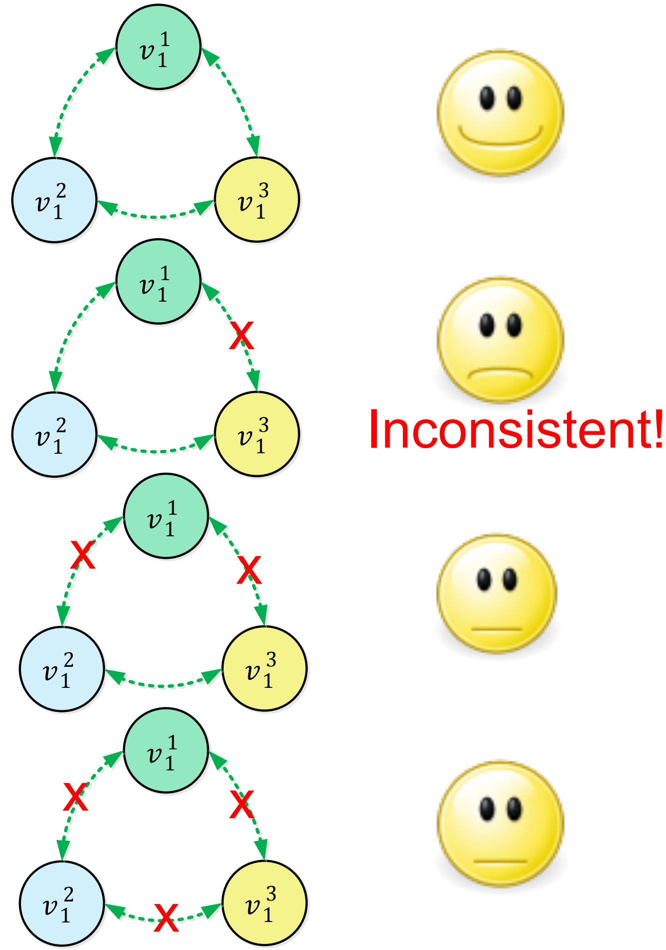




Local vs. Global consistency

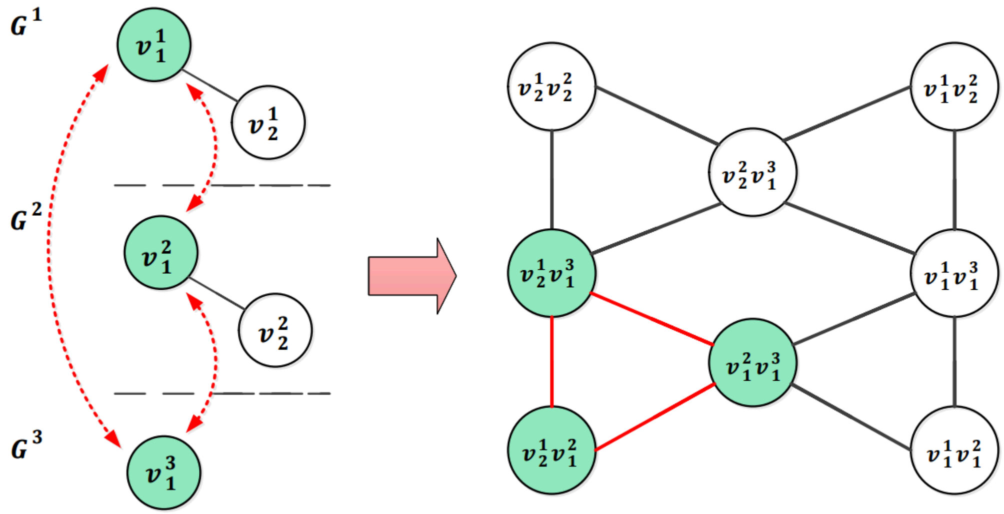
- Global consistency: matching users by avoiding global inconsistency





Avoid Global Inconsistency

Input networks **Matching graph**

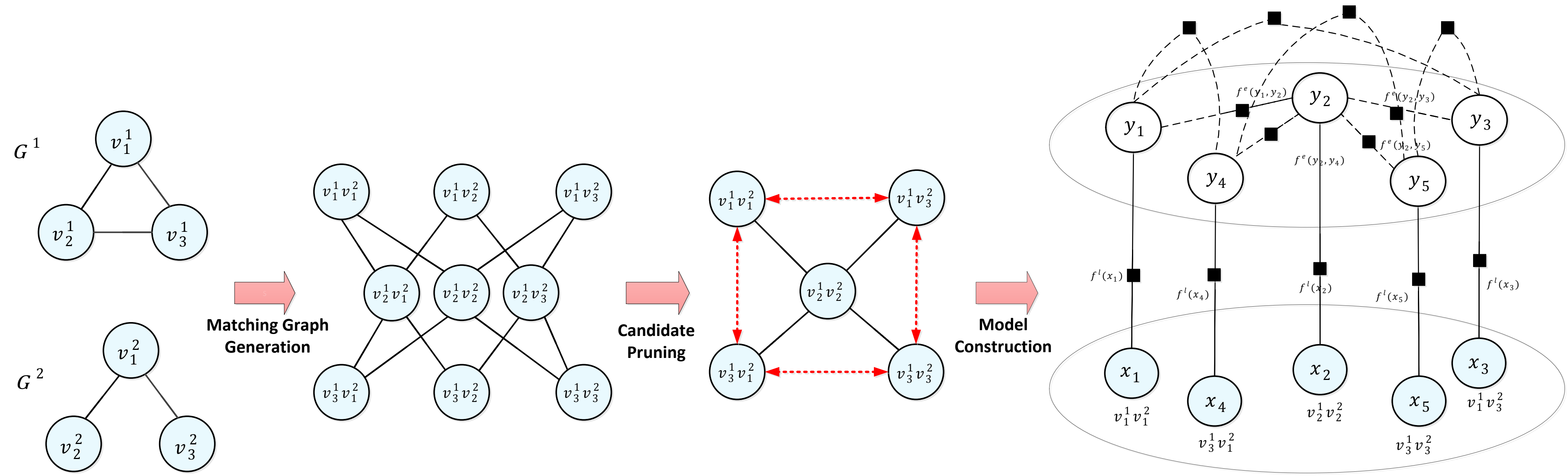


Energy function

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

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

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

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

Model Learning

- Max-margin learning

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

- 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

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

Model Learning (cont.)

- Dual decomposition

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

This provides a lower bound to the original function

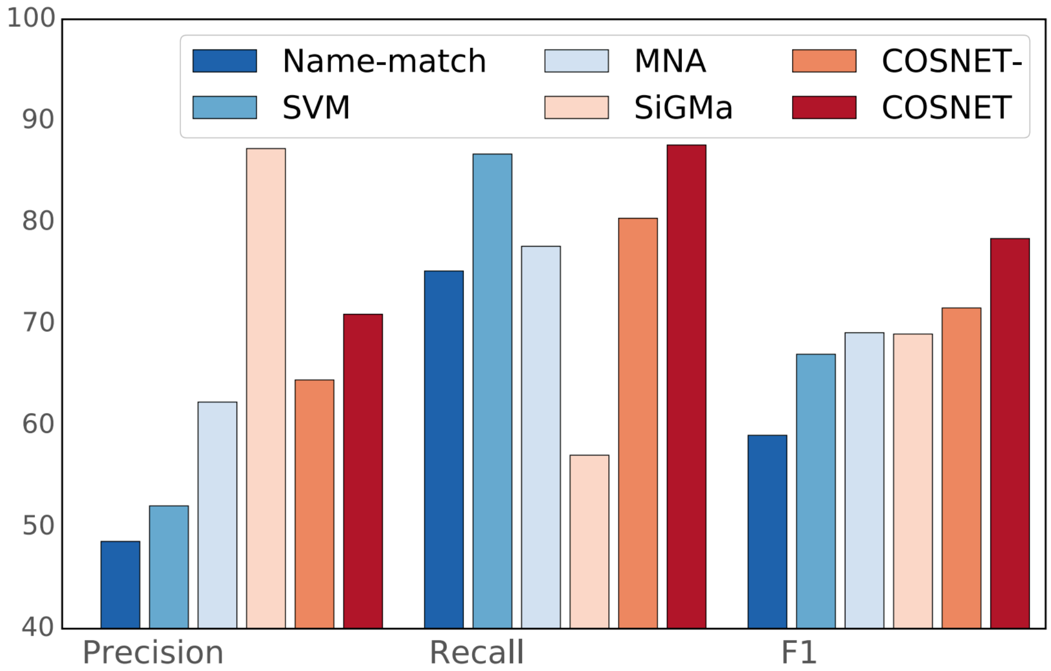
$$\min_{W,\boldsymbol{\lambda}} \frac{1}{2}||W||^2 + \mu(E(\hat{Y},X;W) - \max_{\boldsymbol{\lambda}} L(Y,X,\boldsymbol{\lambda};W))$$
$$\text{s.t.} \quad \sum_{y_i\in Y_i} \lambda_i^f = 0, \ \ \forall f \in \mathcal{F}$$

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

Results

Connecting AMiner with …

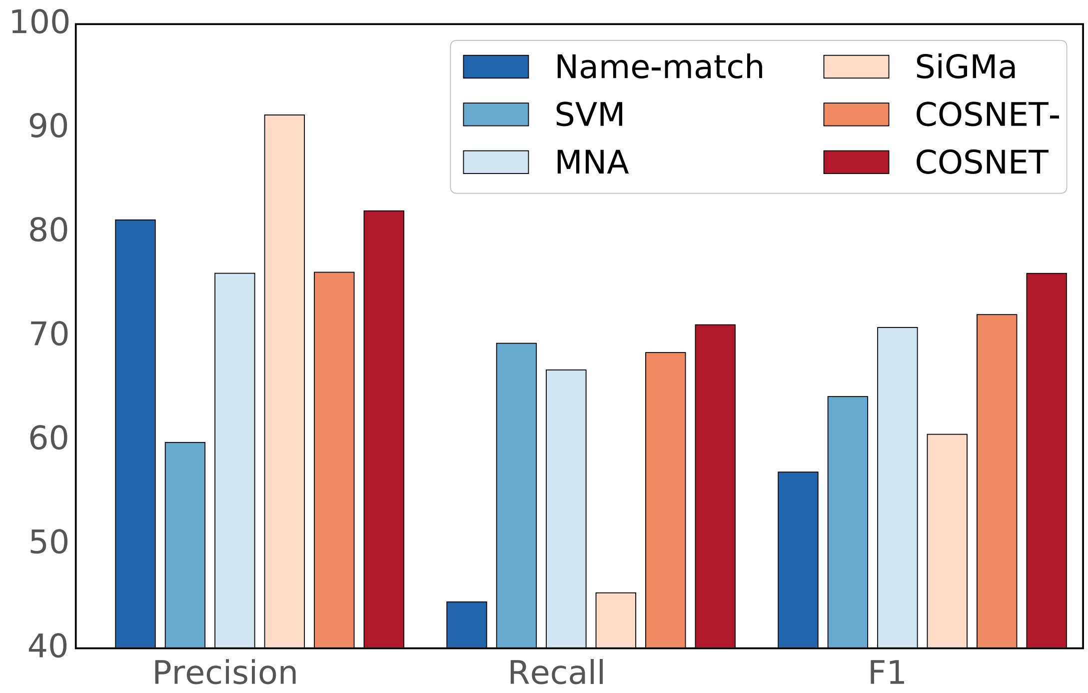
- LinkedIn and VideoLectures



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

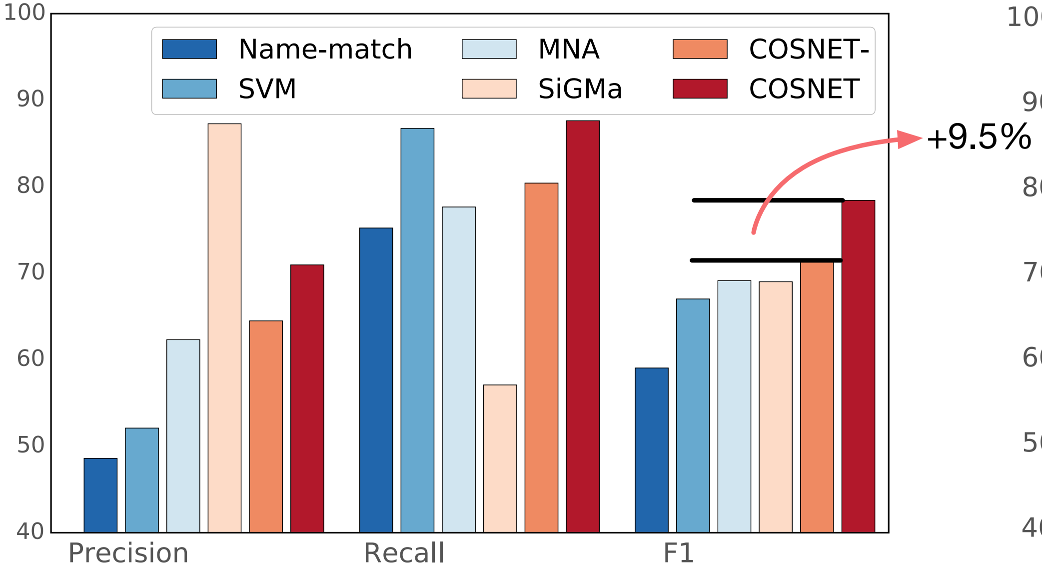


Name-match: match name only;
SVM: use classifier to identify the same user;
MNA: an optimization method;

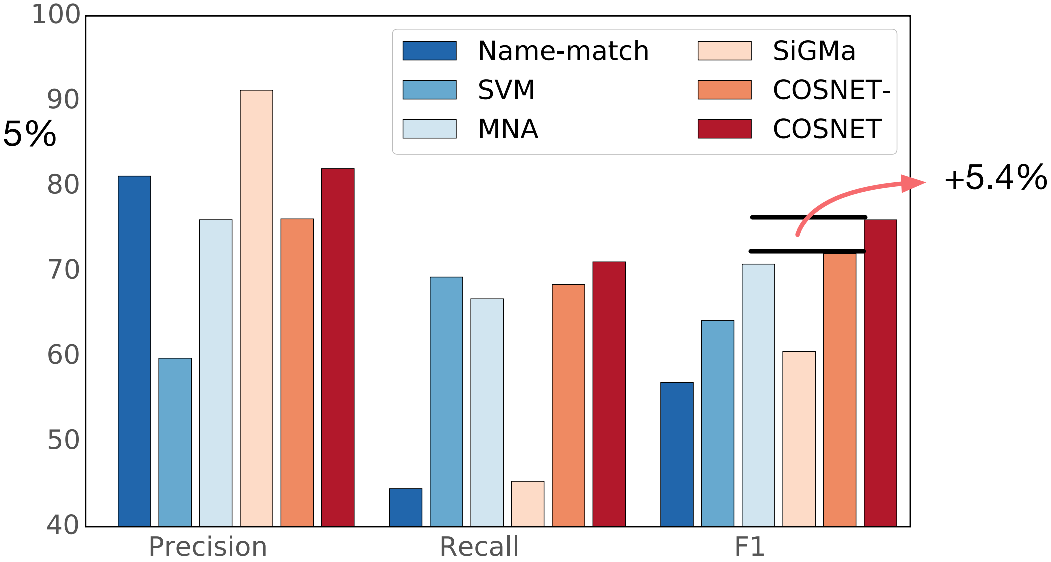
SiGMa: local propagation;
COSNET: our method;
COSNET-: w/o global consistency.

Effects of Global Consistency

COSNET-: w/o global consistency.



Academia Collection

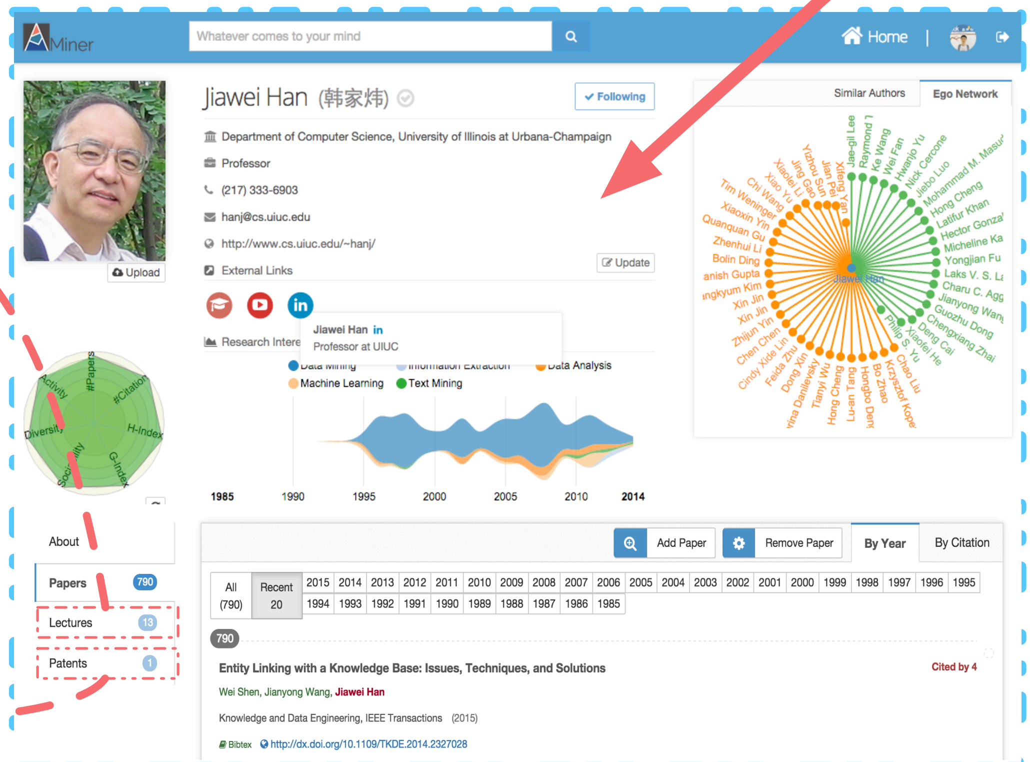


SNS Collection

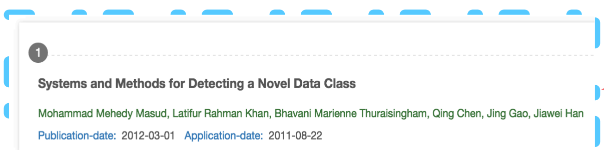
Application in AMiner

- Video contents

- Personal profiles
- Business connections
- Skills and expertise



- Patents data



Thanks!

AMiner

Data & source code

http://aminer.org

http://aminer.org/cosnet