


Mining Advisor-advisee Relationship from Research Publication Network

Chi Wang[1], Jiawei Han[1], Yuntao Jia[1], Jie Tang[2], Duo Zhang[1], Yintao Yu[1], Jingyi Guo[2]

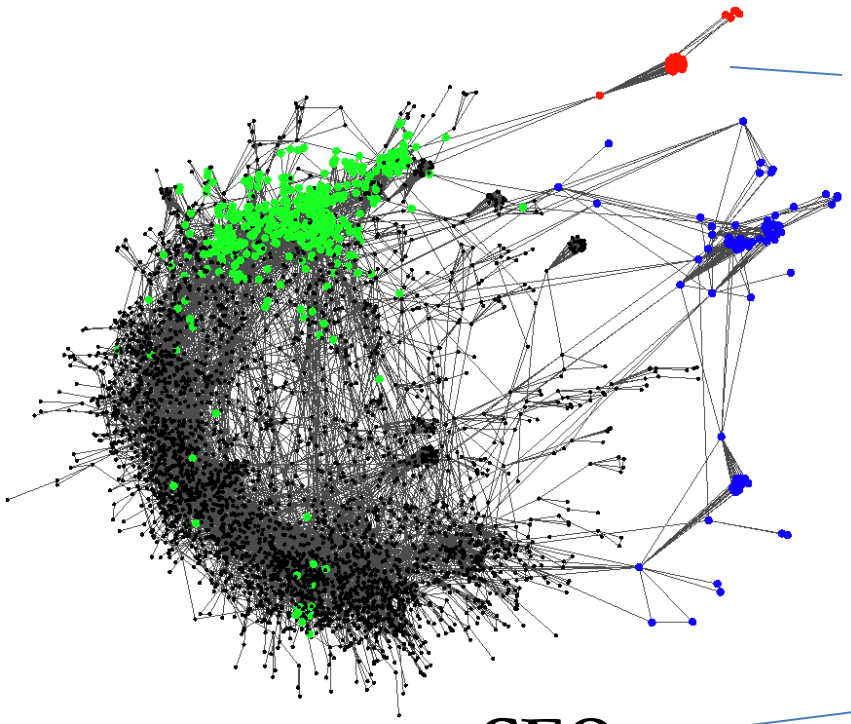
[1] University of Illinois at Urbana-Champaign

{chiwang1, hanj, yjia3, dzhang22, yintao}@illinois.edu

[2] Tsinghua University {jietang, guojy07@mails}.tsinghua.edu.cn

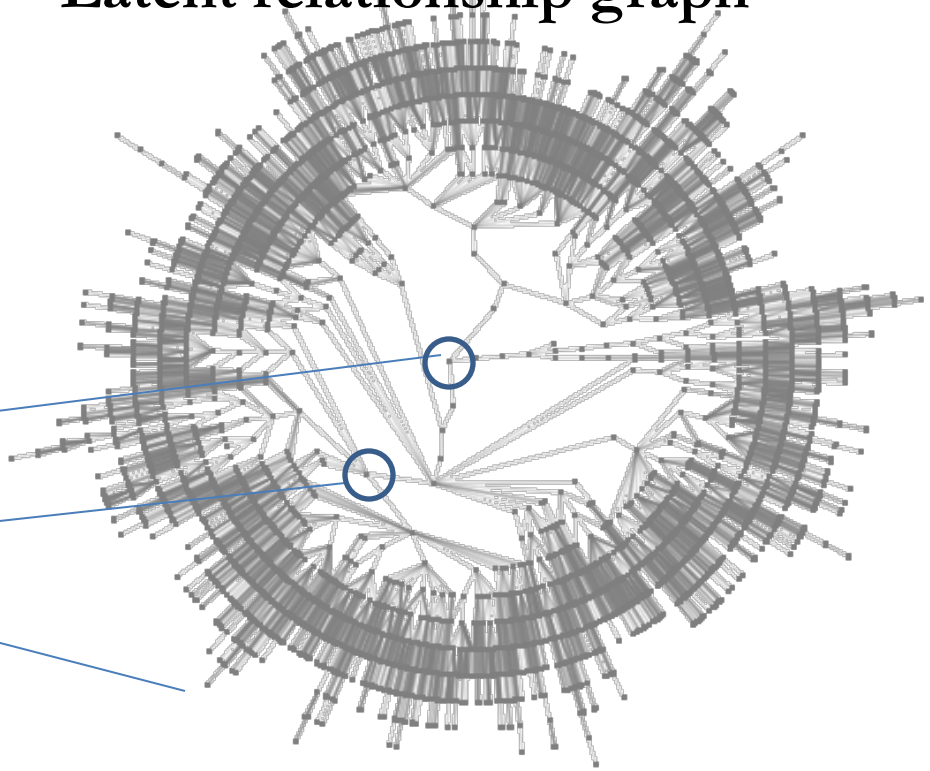
KDD 2010

Role Discovery



Information network without role/relationship info, e.g. a company's email network

Latent relationship graph

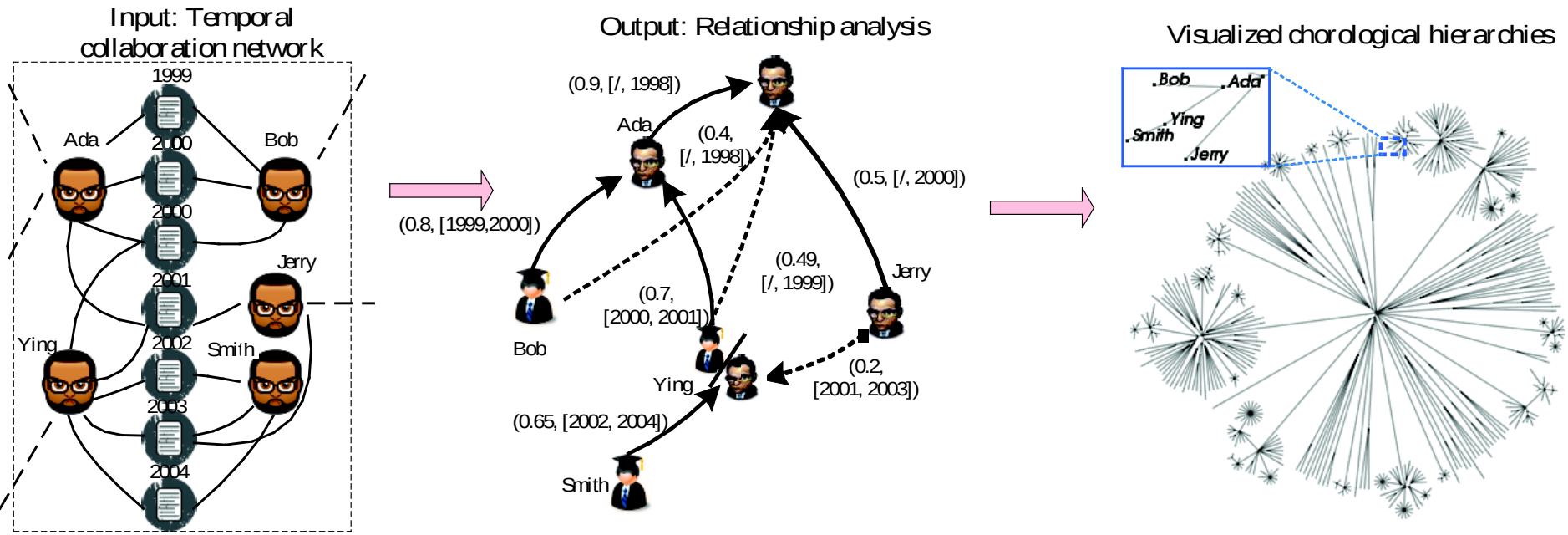


How to infer

- CEO
- Manager
- Employee

This Work: Advisor-advisee

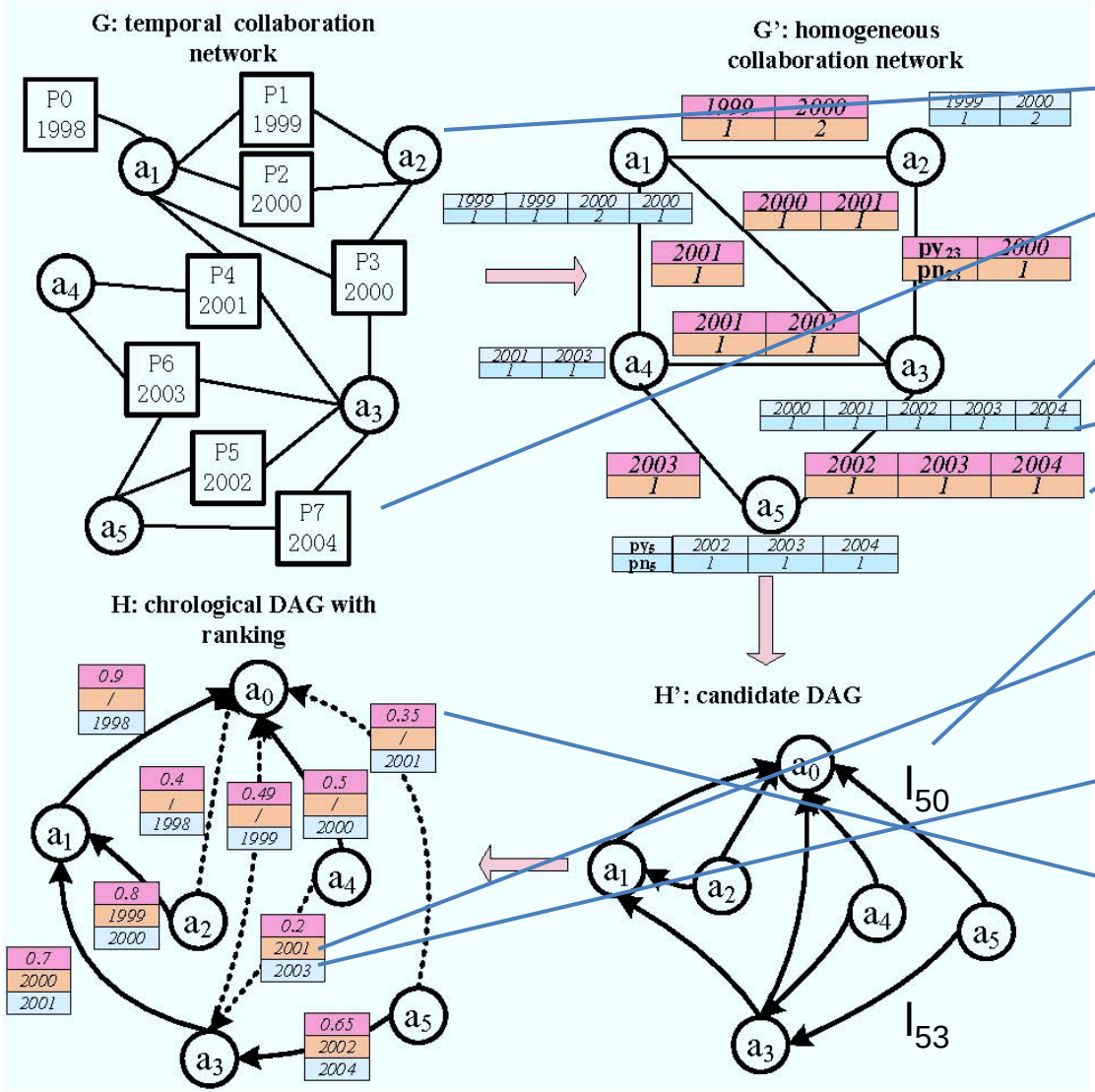
- **Input: research publication network.**
- **Output: potential advising relationship and their ranking – (r, [st, ed])**



Problem Analysis

- **Objective: predict relationship type from plain links**
  - **Challenge?**
    - Time-dependent
    - Interdependency on network
    - Scalability
  - **Opportunity?**
    - Rules, though soft
    - Crosscheck using network
    - Sparsity
- **Methodology: propagate simple intuitive rules and constraints over the whole network**

Overall Framework



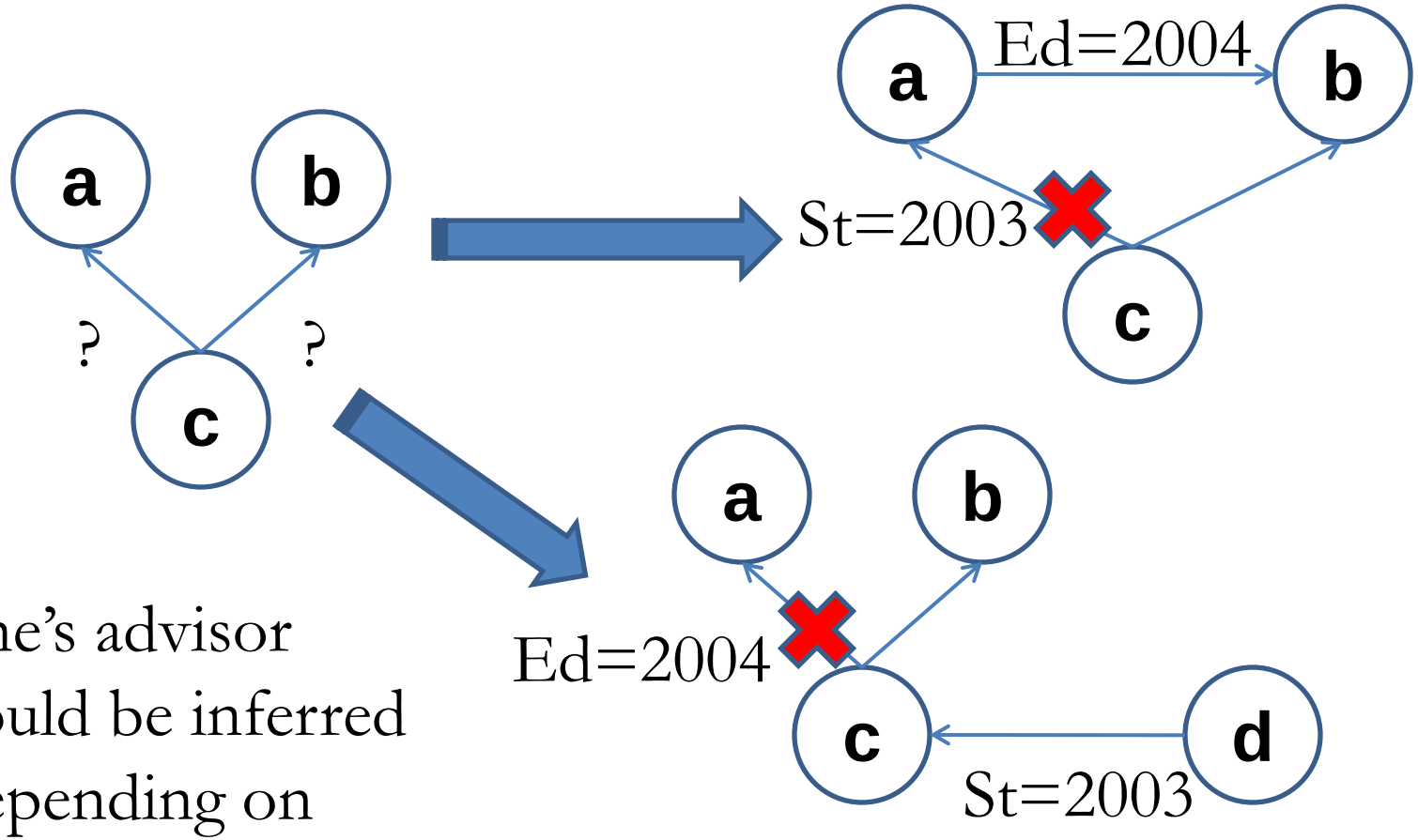
- a_i: author i
- p_j: paper j
- py: paper year
- pn: paper#
- l_{jj}: local feature
- $st_{i,yi}$: start time
- $ed_{i,yi}$: end time
- $r_{i,yi}$: ranking score

Local Features - Preprocess

- **For every pair of coauthors a_i and a_j**
  - Create a potential link from a_i to a_j if a_j has a longer publication history than a_i
  - Compute Kulczynski and Imbalance Ratio measure for the coauthored publications at different time t
  - Estimate the advising time
    - St_{ij} = the start time of coauthorship
    - Ed_{ij} = the time point when correlation drops
      YEAR1: $Kul_{ij}^{t} > Kul_{ij}^{t+1}$, YEAR2: $\max(Kul_{ij}^{t} - Kul_{ij}^{t+1})$
  - Remove the link if certain rules apply, o.w. sum average Kul and IR as a rough likelihood

Why is network structure helpful?

- More than pairwise features: interdependency



one's advisor could be inferred depending on others' advisor!

Basic Constraints

- **If a_y advises a_x since the year st_x**
  - a_y can only advise a_x after it graduated

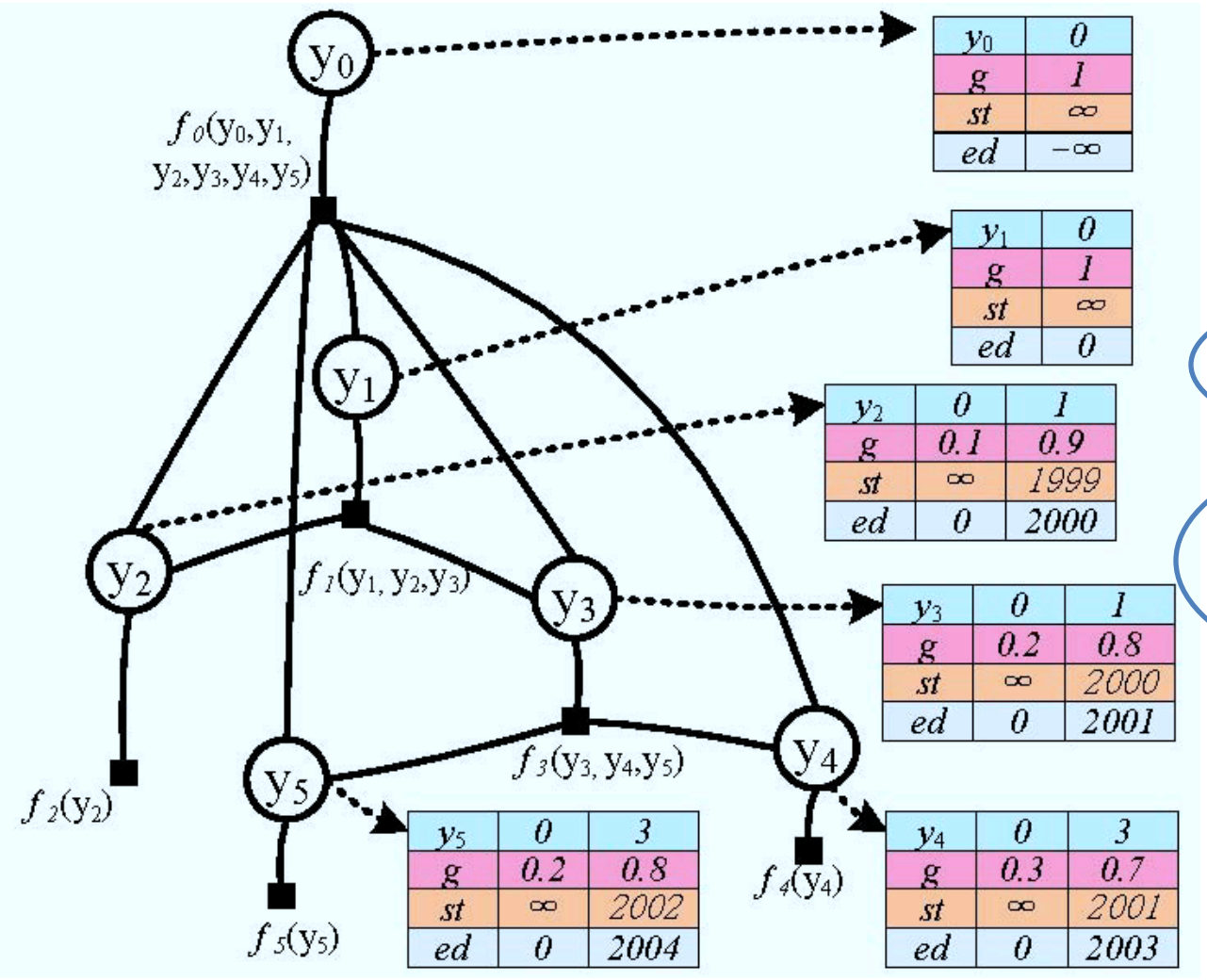
  ➔ $ed_y < st_x < ed_x$

  - a_y must have a longer history of publication than a_x before st_x.

  ➔ The candidate graph H' is a DAG.

The model can incorporate other intuitions as factor functions.

Time-constrained Probabilistic Factor Graph (TPFG)



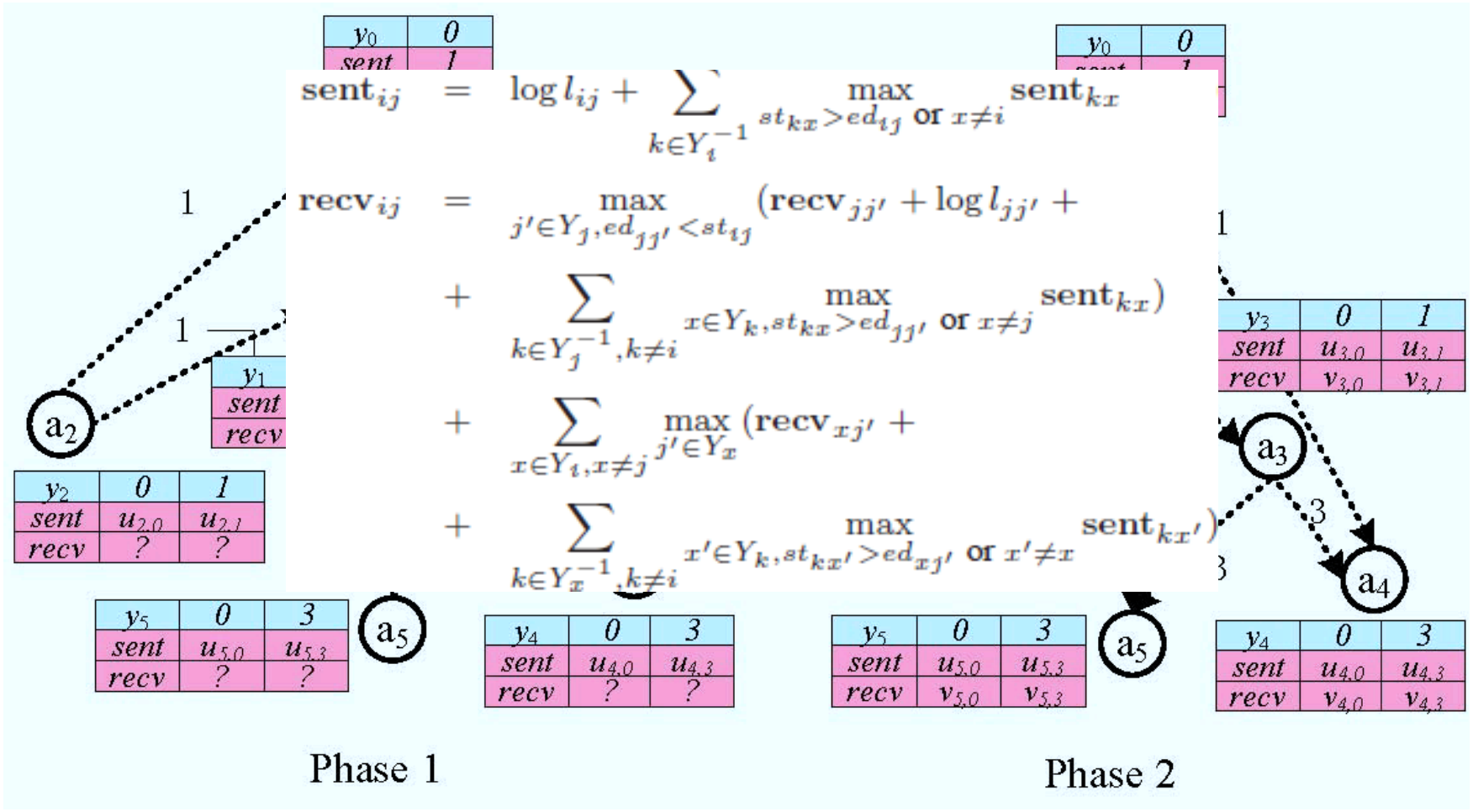
- Hidden variable y_x - a_x's advisor
- $st_{x,yx}$ - start time $ed_{x,yx}$ - end time
- $g(y_x, st_x, ed_x)$ - pairwise local feature $l_{x,yx}$
- $f_x(y_x,Z_x)= g(y_x , st_x, ed_x)$ *if time constraint is s.f., 0 otherwise*
- Objective function $P(\{y_x\})=\prod_x f_x(y_x,Z_x)$
- Z_x - set of potential advisees of a_x

Inference Algorithm of TPFG

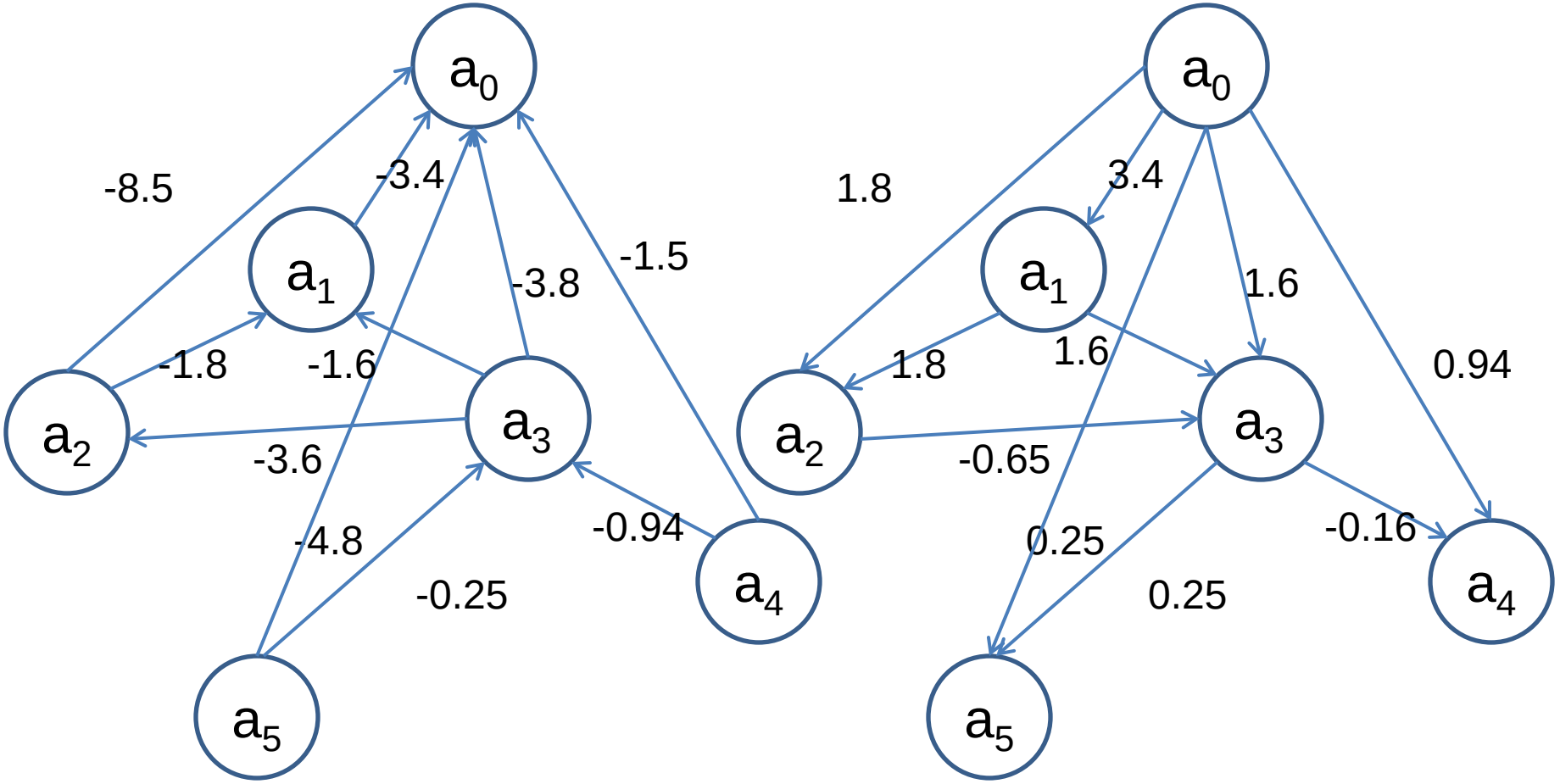
- $r_{ij} = \max P(y_1, \ldots, y_n | y_i = j) = \exp(\text{sent}_{ij} + \text{recv}_{ij})$

$$\text{sent}_{ij} = \log l_{ij} + \sum_{k \in Y_i^{-1}} \max_{st_{kx} > ed_{ij} \text{ or } x \neq i} \text{sent}_{kx}$$

$$\text{recv}_{ij} = \max_{j' \in Y_j, ed_{jj'} < st_{ij}} (\text{recv}_{jj'} + \log l_{jj'} + \sum_{k \in Y_j^{-1}, k \neq i} \max_{x \in Y_k, st_{kx} > ed_{jj'} \text{ or } x \neq j} \text{sent}_{kx}) + \sum_{x \in Y_i, x \neq j} \max_{j' \in Y_x} (\text{recv}_{xj'} + \sum_{k \in Y_x^{-1}, k \neq i} \max_{x' \in Y_k, st_{kx'} > ed_{xj'} \text{ or } x' \neq x} \text{sent}_{kx'})$$



A Running Example



A Running Example (cont'd)

log r = sent + recv

- **log r_{10} = -3.4 + 3.4 = 0**
- log r_{20} = -8.5 + 1.8 = -6.7

  log **r_{21} = -1.8 + 1.8 = 0**
- log r_{30} = -3.8 + 1.6 = -2.2

  log **r_{31} = -1.6 + 1.6 = 0**

  log r_{32} = -3.6 - 0.65 = -4.25
- **logr_{40} = -1.5 + 0.94 = -0.6**

  logr_{43} = -0.94– 0.16 = -1.1
- log r_{50} = -4.8 + 0.25 = -4.6

  log **r_{53} = -0.25 + 0.25 = 0**

Gather answers

- y_1 = 0
- y_2 = 1, st_2 = 1999, ed_2 = 2000
- y_3 = 1, st_3 = 2000, ed_3 = 2001

$sent_{32} > sent_{30}$, but $r_{30} > r_{32}$

- y_4 = 3, st_4 = 2001, ed_4 = 2003

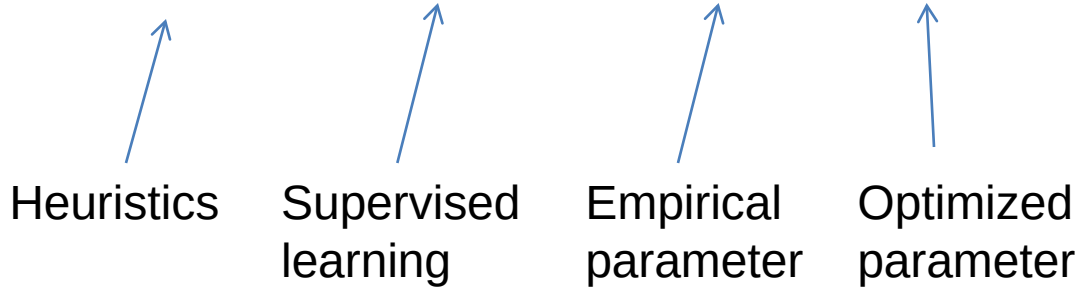
$sent_{43} > sent_{40}$, but $r_{40} > r_{43}$

- y_5 = 3, st_5 = 2002, ed_5 = 2004

Experiment Results

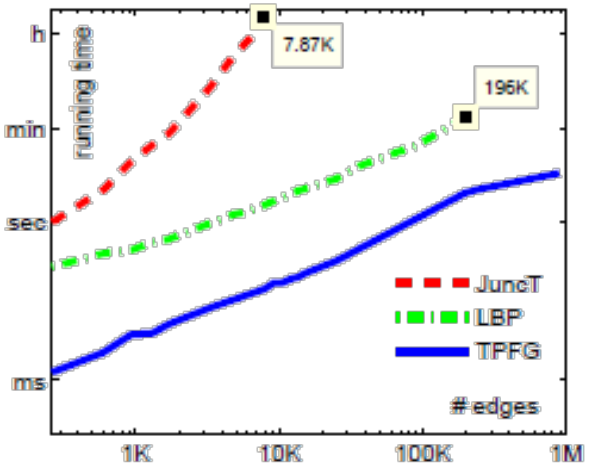
- **DBLP data: 654, 628 authors, 1076,946 publications, publishing time provided.**
- **Labeled data: MathGenealogy Project; AI Gealogy Project; Faculty Homepage**

Datasets	RULE	SVM	TPFG	
TEST1	69.9%	73.4%	80.2%	**84.4%**
TEST2	69.8%	74.6%	81.5%	**84.3%**
TEST3	80.6%	86.7%	88.8%	**91.3%**

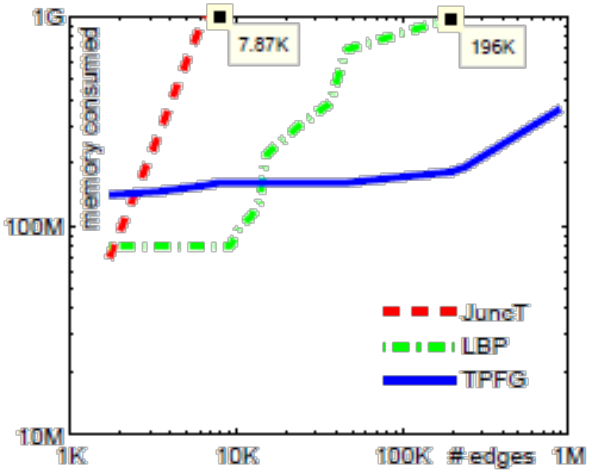


Case Study & Scalability

Advisee	Top Ranked Advisor	Time	Note
David M. Blei	1. Michael I. Jordan	01-03	PhD advisor, 2004 grad
	2. John D. Lafferty	05-06	Postdoc, 2006
Hong Cheng	1. Qiang Yang	02-03	MS advisor, 2003
	2. Jiawei Han	04-08	PhD advisor, 2008
Sergey Brin	1. Rajeev Motawani	97-98	"Unofficial advisor"



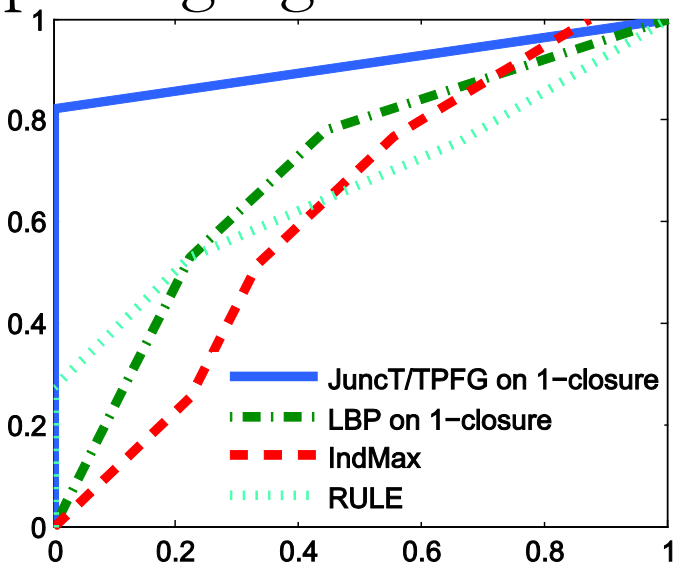
(a) Time



(b) Space

Exact VS Approximate Inference

- **Exact inference of TPFG**
  - JuncT: Junction Tree + Sum-Product
- **Approximate inference**
  - LBP: Loopy Belief Propagation
  - TPFG: the proposed message passing algorithm
  - IndMax: local features only



Effect of Rules - ROC Curve

- **Filtering rules in TPFG**

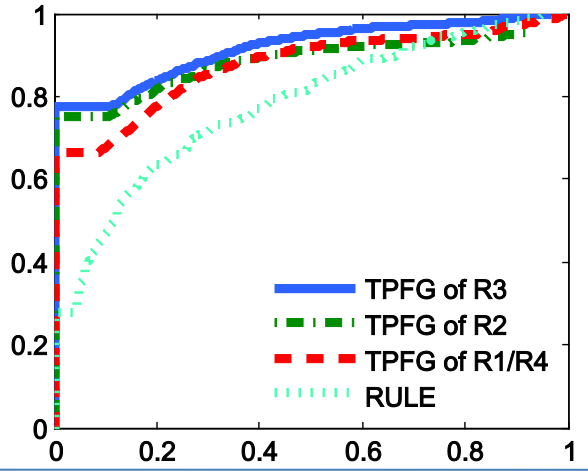
  $R1$: $IR_{ij}^t < 0$ in the sequence $\{IR_{ij}^t\}_t$ during the collaboration period of a_i and a_j,

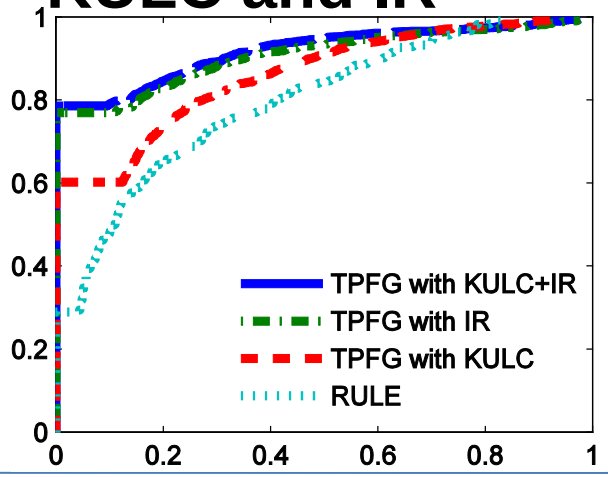
  $R2$: there is no increase in the sequence $\{kulc_{ij}^t\}_t$ during the collaboration period,

  $R3$: the collaboration period of a_i and a_j lasts only for one year,

  $R4$: $py_j^1 + 2 > py_{ij}^1$,

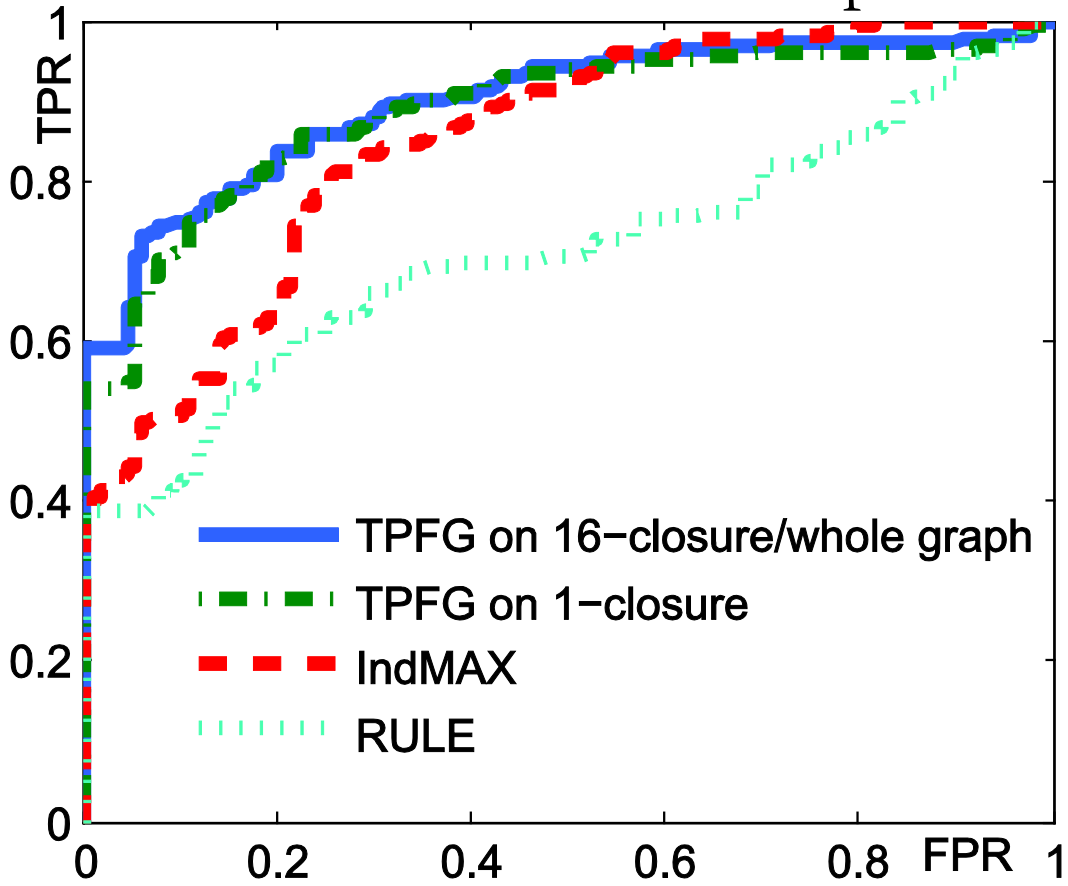
- **Local feature measure: KULC and IR**





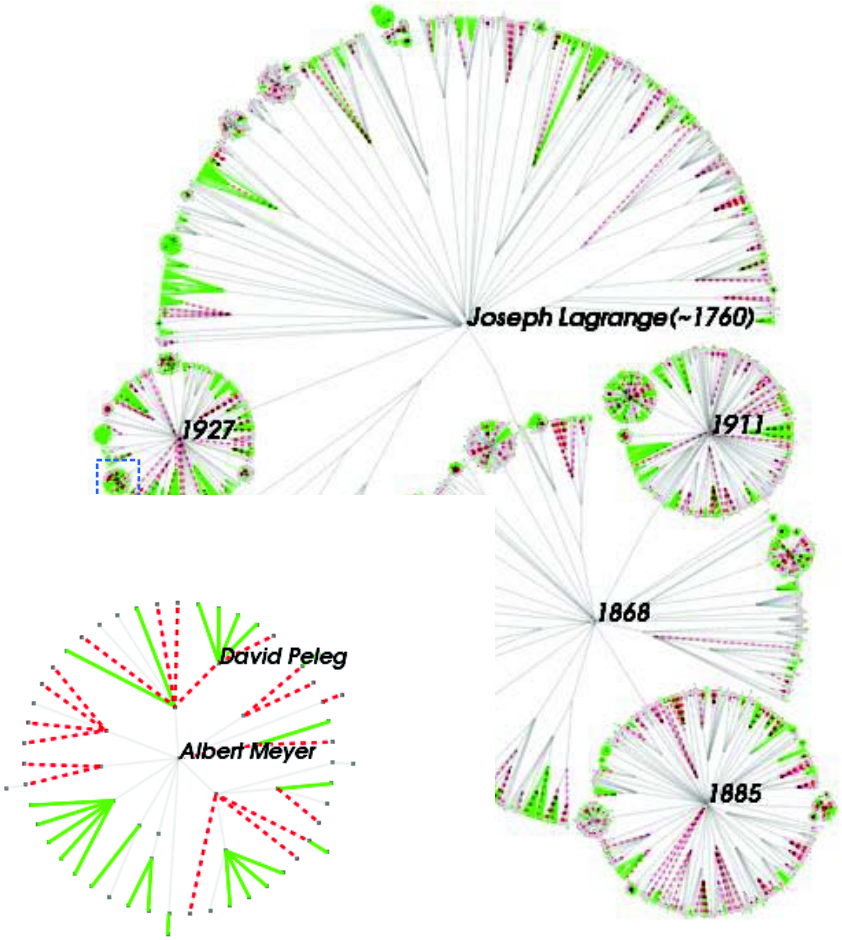
Effect of Network Depth

- **Different closures of given set of nodes**
  - DFS with bounded maximal depth d: d-closure

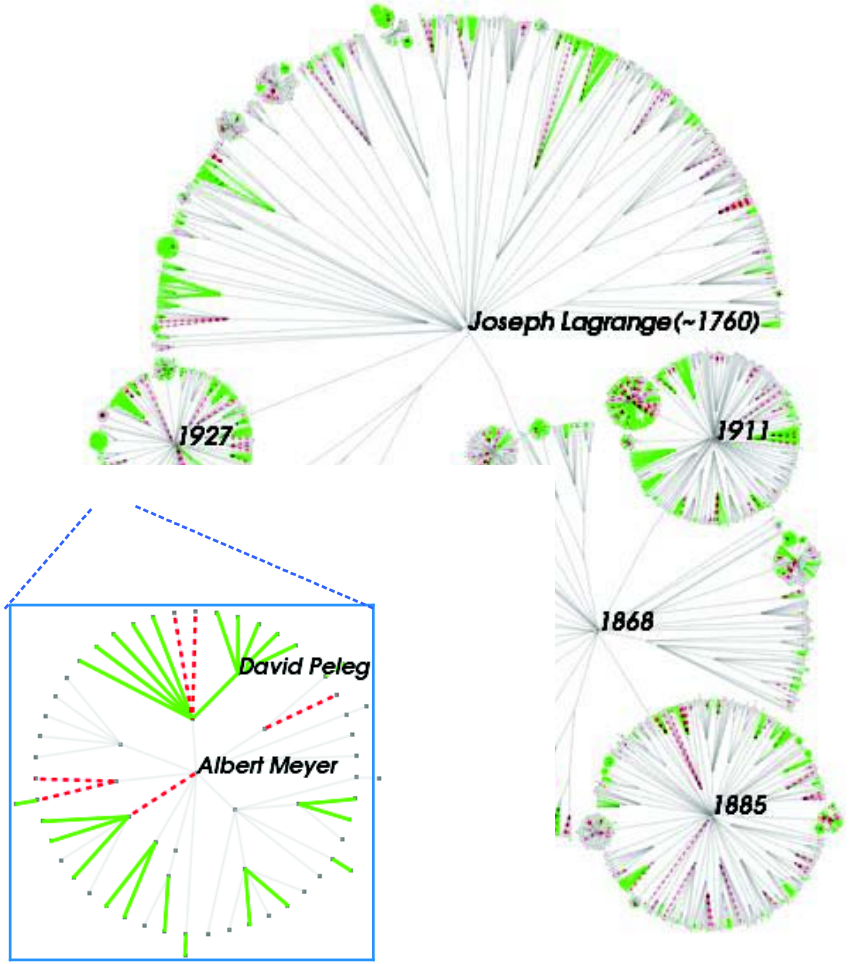


Application: Visualization

RULE

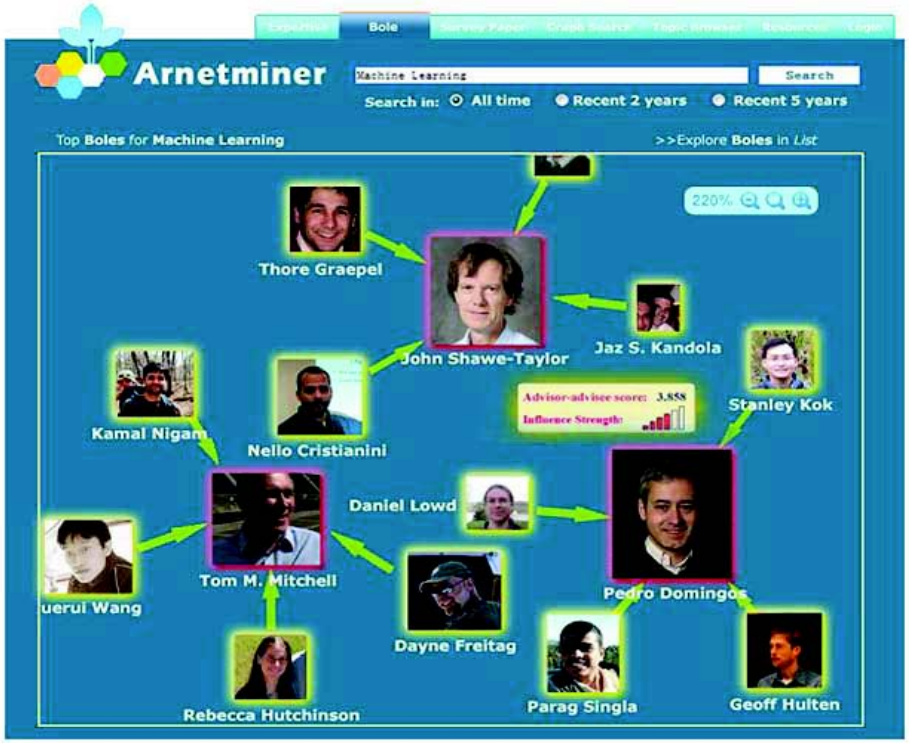


TPFG

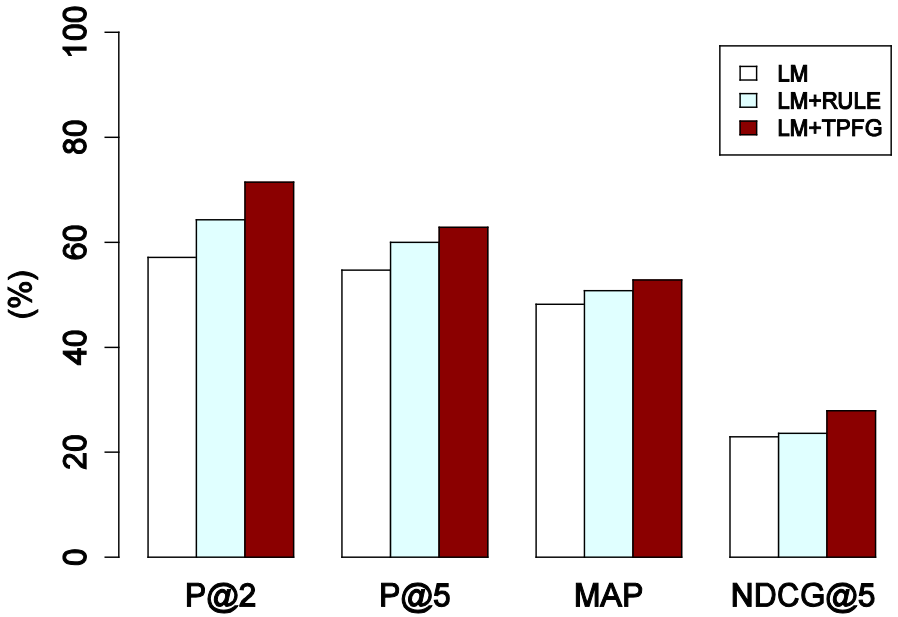


Application: Expert Finding

An example on a real system: Arnetminer



Performance improvement



Related work

- **"Relation Mining" [Kadri 03, Rinaldi 06, Coppola 08]**
  - Mainly text mining and language processing on text data and structured data.
- **"Relational Learning" [Getoor 07, Tang 09]**
  - The classification when objects and entities are presented in multiple relations
- **Relationship with semantic meaning**
  - [Diehl 07]: a supervised approach
  - Our approach: for network with neither text nor labeled data

Thank you

Washington, D.C. July, 2010.