Mining Advisor-advisee Relationship from Research Publication Network

Chi Wang¹, Jiawei Han¹, Yuntao Jia¹, Jie Tang², Duo Zhang¹, Yintao Yu¹, Jingyi Guo²

¹ University of Illinois at Urbana-Champaign
{chiwang1, hanj, yjia3, dzhang22, yintao}@illinois.edu
² Tsinghua University {jietang, guojy07@mails}.tsinghua.edu.cn
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 – \((r, [st, ed])\)

![Diagram showing the input temporal collaboration network and the output relationship analysis with visualized chronological hierarchies.](image-url)
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

- **ai**: author i
- **pj**: paper j
- **py**: paper year
- **pn**: paper#
- **lj**: local feature
- **st_{i,yi}**: start time
- **ed_{i,yi}**: end time
- **ri,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
    - $S_{ij} =$ the start time of coauthorship
    - $E_{ij} =$ the time point when correlation drops
      - YEAR1: $K_{ij}^t > K_{ij}^{t+1}$
      - YEAR2: $\max(K_{ij}^t - K_{ij}^{t+1})$
  - 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: interdependence

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

\[ r_{ij} = \max P(y_1, \ldots, y_{na} | y_i = j) = \exp (sent_{ij} + recv_{ij}) \]
A Running Example

The diagram illustrates a network with nodes labeled $a_0$, $a_1$, $a_2$, $a_3$, $a_4$, and $a_5$. The edges between the nodes are labeled with numerical values, indicating connections and weights in the network.
A Running Example (cont’d)

\[ \log r = \text{sent} + \text{recv} \]

- \[ \log r_{10} = -3.4 + 3.4 = 0 \]
- \[ \log r_{20} = -8.5 + 1.8 = -6.7 \]
- \[ \log r_{21} = -1.8 + 1.8 = 0 \]
- \[ \log r_{30} = -3.8 + 1.6 = -2.2 \]
- \[ \log r_{31} = -1.6 + 1.6 = 0 \]
- \[ \log r_{32} = -3.6 - 0.65 = -4.25 \]
- \[ \log r_{40} = -1.5 + 0.94 = -0.6 \]
- \[ \log r_{43} = -0.94 - 0.16 = -1.1 \]
- \[ \log r_{50} = -4.8 + 0.25 = -4.6 \]
- \[ \log r_{53} = -0.25 + 0.25 = 0 \]

Gather answers:

- \[ y_1 = 0 \]
- \[ y_2 = 1, \text{st}_2 = 1999, \text{ed}_2 = 2000 \]
- \[ y_3 = 1, \text{st}_3 = 2000, \text{ed}_3 = 2001 \]
- \[ y_4 = 3, \text{st}_4 = 2001, \text{ed}_4 = 2003 \]
- \[ y_5 = 3, \text{st}_5 = 2002, \text{ed}_5 = 2004 \]
Experiment Results

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

<table>
<thead>
<tr>
<th>Datasets</th>
<th>RULE</th>
<th>SVM</th>
<th>TPFG</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEST1</td>
<td>69.9%</td>
<td>73.4%</td>
<td>80.2%</td>
</tr>
<tr>
<td>TEST2</td>
<td>69.8%</td>
<td>74.6%</td>
<td>81.5%</td>
</tr>
<tr>
<td>TEST3</td>
<td>80.6%</td>
<td>86.7%</td>
<td>88.8%</td>
</tr>
</tbody>
</table>

Heuristics | Supervised learning | Empirical parameter | Optimized parameter
# Case Study & Scalability

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

![Graphs](image-url)
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
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,

\( R4: py_{ij}^1 + 2 > py_{ij}^1 \),

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

18
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