Cross-domain Link Prediction and Recommendation

Jie Tang

Department of Computer Science and Technology
Tsinghua University
Networked World

- 1.26 billion users
- 700 billion minutes/month

- 280 million users
- 80% of users are 80-90’s

- 555 million users
- 0.5 billion tweets/day

- 560 million users
- Influencing our daily life

- 79 million users per month
- 9.65 billion items/year

- 500 million users
- 35 billion on 11/11

- 800 million users
- ~50% revenue from network life
Challenge: Big Social Data

• We generate $2.5 \times 10^{18}$ byte big data per day.

• Big social data:
  – 90% of the data was generated in the past 2 yrs
  – Mining in single data center $\rightarrow$ mining deep knowledge from multiple data sources
Understanding the mechanisms of interaction dynamics
Core Research in Social Network

Application
- Information Diffusion
- Search
- Prediction
- Advertise

Social Network Analysis
- Macro
  - BA model
  - ER model
- Meso
  - Community
  - Group Behavior
  - Structural Hole
- Micro
  - Social Action
  - Social Influence
  - Social Tie

Theory
- Social Theories
- Algorithmic Foundations

BIG Social Data
Part A:
Let us start with a simple case
“inferring social ties in single network”

(KDD 2010, PKDD 2011 Best Runnerup)
Real social networks are complex...

• Nobody exists merely in one social network.
  – Public network vs. private network
  – Business network vs. family network

• However, existing networks (e.g., Facebook and Twitter) are trying to lump everyone into one big network
  – FB/QQ tries to solve this problem via lists/groups
  – however...

• Google circles
Even complex than we imaged!

- Only 16% of mobile phone users in Europe have created custom contact groups
  - *users do not* take the time to create it
  - *users do not* know how to circle their friends

- The Problem is that online social network are **black** **white**...
Example 1. From BW to Color (KDD’10)
Example 2. From BW to Color
(PKDD’11, Best Paper Runnerup)

Enterprise email network

How to infer

User interactions may form *implicit groups*
What is behind?

Publication network

Twitter’s following network

Mobile communication network

Adam  2005  Paper 1
Bob  2003  Paper 2
Chris  2008  Paper 3
Danny  2006  Paper 4

From Home 08:40
From Office 11:35
Both in office 08:00 – 18:00
From Office 15:20
From Office 17:55
From Outside 21:30

From Office 11:35
From Office 15:20
From Office 17:55
From Outside 21:30
What is behind?

Questions:
- What are the fundamental forces behind?
- A generalized framework for inferring social ties?
- How to connect the different networks?
inferring social ties in single network

Learning Framework
Problem Analysis

Input: Temporal collaboration network

Dynamic collaborative network

Output: Relationship analysis

(0.9, [/, 1998])
(0.4, [/, 1999])
(0.5, [, 2000])
(0.49, [, 1999])
(0.7, [2000, 2001])
(0.2, [2001, 2003])
(0.65, [2002, 2004])

Output: potential types of relationships and their probabilities:
(type, prob, [s_time, e_time])

The problem is cast as, for each node, identifying which neighbor has the highest probability to be his/her advisor, i.e., $P(y_i=j | x_i, x_{-i}, y)$, where $x_j$ and $x_i$ are neighbors.
Time-constrained Probabilistic Factor Graph (TPFG)

- Hidden variable $y_x$: $a_x$’s advisor
- $st_{x,y_x}$: starting time
- $ed_{x,y_x}$: ending time
- $g(y_x, st_{x}, ed_{x})$ is pairwise local feature
- $f_x(y_x, Z_x) = \max g(y_x, st_{x}, ed_{x})$ under time constraint
- $Y_x$: set of potential advisors of $a_x$
Maximum likelihood estimation

- A general likelihood objective function can be defined as

\[
P(y_1, \cdots, y_N) = \frac{1}{Z} \prod_{i=1}^{N} f_i(y_i | \{y_x | x \in Y_i^{-1}\})
\]

with

\[
f_i(y_i | \{y_x | x \in Y_i^{-1}\}) = g(y_i, st_{ij}, ed_{ij}) \prod_{x \in Y_i^{-1}} \phi(y_x, ed_{ij}, st_{xi})
\]

where \(\phi(.)\) can be instantiated in different ways, e.g.,

\[
\phi(y_x, ed_{ij}, st_{xi}) = \begin{cases} 
1, & y_x \neq i \lor ed_{ij} < st_{xi} \\
0, & y_x = i \land ed_{ij} \geq st_{xi}
\end{cases}
\]
Inference algorithm of TPFG

- \( r_{ij} = \max P(y_1, \ldots, y_{na} | y_i = j) = \exp (\text{sent}_{ij} + \text{recv}_{ij}) \)
Results of Model 1

- DBLP data: 654, 628 authors, 1,076,946 publications, years provided.
- Ground truth: MathGenealogy Project; AI Genealogy Project; Faculty Homepage

<table>
<thead>
<tr>
<th>Datasets</th>
<th>RULE</th>
<th>SVM</th>
<th>IndMAX</th>
<th>Model 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEST1</td>
<td>69.9%</td>
<td>73.4%</td>
<td>75.2%</td>
<td>80.2%</td>
</tr>
<tr>
<td>TEST2</td>
<td>69.8%</td>
<td>74.6%</td>
<td>74.6%</td>
<td>81.5%</td>
</tr>
<tr>
<td>TEST3</td>
<td>80.6%</td>
<td>86.7%</td>
<td>83.1%</td>
<td>88.8%</td>
</tr>
</tbody>
</table>

heuristics
Supervised learning
Empirical optimized parameter parameter
Results

Part B:
Extend the problem to cross-domain
“cross-domain collaboration recommendation”

(KDD 2012, WSDM)
Cross-domain Collaboration

• Interdisciplinary collaborations have generated huge impact, for example,
  – 51 (>1/3) of the KDD 2012 papers are result of cross-domain collaborations between graph theory, visualization, economics, medical inf., DB, NLP, IR
  – Research field evolution

Cross-domain Collaboration (cont.)

- Increasing trend of cross-domain collaborations

Data Mining (DM), Medical Informatics (MI), Theory (TH), Visualization (VIS)
Challenges

1. Sparse Connection: <1%
   - Large graph
   - Heterogeneous network
   - Social network

2. Complementary expertise
   - Automata theory
   - Graph theory
   - Complexity theory

3. Topic skewness: 9%
Related Work - Collaboration recommendation

- Collaborative topic modeling for recommending papers
  - C. Wang and D.M. Blei. [2011]

- On social networks and collaborative recommendation
  - I. Konstas, V. Stathopoulos, and J. M. Jose. [2009]

- CollabSeer: a search engine for collaboration discovery

- Referral web: Combining social networks and collaborative filtering

- Fab: content-based, collaborative recommendation
  - M. Balabanovi and Y. Shoham. [1997]
Related Work - Expert finding and matching

- Topic level expertise search over heterogeneous networks

- Formal models for expert finding in enterprise corpora
  - K. Balog, L. Azzopardi, and M. de Rijke. [2006]

- Expertise modeling for matching papers with reviewers
  - D. Mimno and A. McCallum. [2007]

- On optimization of expertise matching with various constraints
  - W. Tang, J. Tang, T. Lei, C. Tan, B. Gao, and T. Li. [2012]
cross-domain collaboration recommendation

Approach Framework
—Cross-domain Topic Learning
Author Matching

Data Mining

Medical Informatics

$G^S$

$v_1$

$v_2$

$\ldots$

$v_N$

$v_q$

$G^T$

$v_1'$

$v_2'$

$\ldots$

$v_N'$

Cross-domain coauthorships

Author

Coauthorships

Query user

$$r^{(t+1)} = (1 - \tau)S \cdot r^{(t)} + \tau q$$
Recall Random Walk

• Let us begin with PageRank[1]

\[ r = (1 - \alpha)M \cdot r + \alpha U \]

\[ M_{ij} = \frac{1}{\text{outdeg}(v_i)} \]

\[ U_i = \frac{1}{N} \]

\[ \alpha = 0.15 \]

\[ (0.2 + 0.2 \times 0.5 + 0.2 \times 1/3 + 0.2) \times 0.85 + 0.15 \times 0.2 \]

---

Random Walk with Restart\cite{Sun2005}

\[ \mathbf{r}_q = (1 - \alpha) \mathbf{M} \cdot \mathbf{r}_q + \alpha \mathbf{U} \]

\[ M_{ij} = \frac{1}{\text{outdeg}(v_i)} \]

\[ U_i = \begin{cases} 1, & i = q \\ 0, & i \neq q \end{cases} \]

Author Matching

Data Mining

$G^S$

$v_1$

$v_2$

$\ldots$

$v_N$

$v_q$

Medical Informatics

$G^T$

$v'_2$

$\ldots$

$v'_{N'}$

Author

Coauthorships

Cross-domain coauthorship

Query user

\[ r^{(t+1)} = (1 - \tau)S \cdot r^{(t)} + \tau q \]
Topic Matching

Data Mining

Medical Informatics

Topics Extraction

Complementary Expertise!

Topic skewness!

Topics correlations

32
Recall Topic Model

- Usage of a theme:
  - Summarize topics/subtopics
  - Navigate documents
  - Retrieve documents
  - Segment documents
  - All other tasks involving unigram language models

Mixture components | Mixture weights
---|---
Bayesian approach: use priors
Mixture weights $\sim$ Dirichlet($\alpha$)
Mixture components $\sim$ Dirichlet($\beta$)
A generative model for generating the co-occurrence of documents \( d \in D = \{d_1, \ldots, d_D\} \) and terms \( w \in W = \{w_1, \ldots, w_W\} \), which associates latent variable \( z \in Z = \{z_1, \ldots, z_Z\} \).

The generative processing is:

\[
P(d) \rightarrow P(z|d) \rightarrow P(w|z)
\]
Topic Model

\[ P(d) \]

\[ P(z|d) \]

\[ P(w|z) \]

\[ w_1 \]

\[ w_2 \]

\[ w_W \]

\[ z_1 \]

\[ z_2 \]

\[ z_Z \]

\[ d_1 \]

\[ d_2 \]

\[ d_D \]

\[ P(C) = \Phi \Theta \]

\[ \Phi \]

\[ \Theta \]

\[ C \]

normalized co-occurrence matrix

mixture components

mixture weights
Maximum-likelihood

- **Definition**
  - We have a density function $P(x|\Theta)$ that is governed by the set of parameters $\Theta$, e.g., $P$ might be a set of Gaussians and $\Theta$ could be the means and covariances.
  - We also have a data set $X=\{x_1,\ldots,x_N\}$, supposedly drawn from this distribution $P$, and assume these data vectors are i.i.d. with $P$.
  - Then the log-likelihood function is:
    \[
    L(\Theta | X) = \log p(X | \Theta) = \log \prod_i p(x_i | \Theta) = \sum_i \log p(x_i | \Theta)
    \]
  - The log-likelihood is thought of as a function of the parameters $\Theta$ where the data $X$ is fixed. Our goal is to find the $\Theta$ that maximizes $L$. That is
    \[
    \Theta^* = \arg \max_{\Theta} L(\Theta | X)
    \]
Topic Model

• Following the likelihood principle, we determine $P(d)$, $P(z \mid d)$, and $P(w \mid d)$ by maximization of the log-likelihood function:

$$
L(\Theta \mid d, w, z) = \log \prod_d \prod_w P(d, w)^n = \sum_{d \in D} \sum_{w \in W} n(d, w) \log P(d, w)
$$

$$
= \sum_{d \in D} \sum_{w \in W} n(d, w) \log \left( \sum_{z \in Z} P(w \mid z)P(d \mid z)P(z) \right)
$$

co-occurrence times of $d$ and $w$. Which is obtained according to the multi-distribution
Jensen’s Inequality

- Recall that $f$ is a **convex function** if $f''(x) \geq 0$, and $f$ is strictly convex function if $f''(x) > 0$
- Let $f$ be a convex function, and let $X$ be a random variable, then:

$$E[f(X)] \geq f(EX)$$

- Moreover, if $f$ is strictly convex, then $E[f(X)] = f(EX)$ holds true if and only if $X = EX$ with probability 1 (i.e., if $X$ is a constant)
Basic EM Algorithm

• However, Optimizing the likelihood function is analytically intractable but when the likelihood function can be simplified by assuming the existence of and values for additional but missing (or hidden) parameters:

\[ L(\Theta \mid X) = \sum_i \log p(x_i \mid \Theta) = \sum_i \log \sum_z p(x_i, z \mid \Theta) \]

• Maximizing L(\Theta) explicitly might be difficult, and the strategy is to instead repeatedly construct a lower-bound on L(E-step), and then optimize that lower bound (M-step).
  – For each i, let Q_i be some distribution over z (\( \sum_z Q_i(z)=1, Q_i(z)\geq0 \)), then

\[
\sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \Theta) = \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}; \Theta)}{Q_i(z^{(i)})} \geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \Theta)}{Q_i(z^{(i)})}
\]

  – The above derivation used Jensen’s inequality. Specifically, \( f(x) = \log x \) is a concave function, since \( f''(x)=-1/x^2<0 \)
Parameter Estimation-Using EM

• According to Basic EM:

\[ Q_i(z^{(i)}) = p(z^{(i)} | x^{(i)}; \Theta) \]

• Then we define

\[ Q_i(z^{(i)}) = p(z | d, w) \]

• Thus according to Jensen’s inequality

\[
L(\Theta) = \sum_{d \in D} \sum_{w \in W} n(d, w) \log \sum_{z \in Z} p(z | d, w) \frac{p(w | z) p(d | z) p(z)}{p(z | d, w)} \\
\geq \sum_{d \in D} \sum_{w \in W} n(d, w) \sum_{z \in Z} p(z | d, w) \log \frac{p(w | z) p(d | z) p(z)}{p(z | d, w)}
\]
(1) Solve \( P(w|z) \)

- We introduce Lagrange multiplier \( \lambda \) with the constraint that \( \sum_w P(w|z) = 1 \), and solve the following equation:

\[
\frac{\partial}{\partial P(w|z)} \left\{ \sum_{d \in D} \sum_{w \in W} n(d, w) \sum_{z \in Z} p(z \mid d, w) \log \frac{p(w \mid z)p(d \mid z)p(z)}{p(z \mid d, w)} + \lambda \left[ \sum_z P(w \mid z) - 1 \right] \right\} = 0
\]

\[
\sum_{d \in D} n(d, w) P(z \mid d, w) \quad \frac{P(w \mid z)}{P(w \mid z)} + \lambda = 0,
\]

\[
\therefore \quad P(w \mid z) = -\frac{\sum_{d \in D} n(d, w) P(z \mid d, w)}{\lambda},
\]

\[
\sum_w P(w \mid z) = 1,
\]

\[
\therefore \quad \lambda = -\sum_{w \in W} \sum_{d \in D} n(d, w) P(z \mid d, w),
\]

\[
\therefore \quad P(w \mid z) = \frac{\sum_{d \in D} n(d, w) P(z \mid d, w)}{\sum_{w \in W} \sum_{d \in D} n(d, w) P(z \mid d, w)}
\]
The final update Equations

- **E-step:**

\[
P(z \mid d, w) = \frac{P(w \mid z)P(d \mid z)P(z)}{\sum_{z \in Z} P(w \mid z)P(d \mid z)P(z)}
\]

- **M-step:**

\[
P(w \mid z) = \frac{\sum_{d \in D} n(d, w)P(z \mid d, w)}{\sum_{d \in D} \sum_{w \in W} n(d, w)P(z \mid d, w)}
\]

\[
P(d \mid z) = \frac{\sum_{w \in W} n(d, w)P(z \mid d, w)}{\sum_{d \in D} \sum_{w \in W} n(d, w)P(z \mid d, w)}
\]

\[
P(z) = \frac{\sum_{d \in D} \sum_{w \in W} n(d, w)P(z \mid d, w)}{\sum_{w \in W} \sum_{d \in D} n(d, w)}
\]
PLSI(SIGIR’99)

LDA (JMLR’03)

Document specific distribution over topics

Topic distribution over words

Cross-domain Topic Learning

Identify “cross-domain” Topics

Data Mining

Topics

Medical Informatics

$G^S$

$v_1$

$v_2$

$\ldots$

$v_N$

$v_q$

$G^T$

$z_1$

$z_2$

$z_3$

$\ldots$

$z_K$

$v'_1$

$v'_2$

$\ldots$

$v'_{N'}$
Collaboration Topics Extraction

Step 1: Initialize an ACT model in $G^S$ by learning from documents written by authors only from $G^S$.
Similarly, initialize an ACT model for target domain $G^T$.

Step 2:

Input: a source domain $G^S$ and a target domain $G^T$
Output: estimated parameters $\theta, \theta', \phi, \hat{\alpha}, \hat{\beta}$, and $\lambda$

foreach collaborated document $d$ do
  foreach word $x_{di} \in d$ do
    Toss a coin $s_{di}$ according to $\text{bernoulli}(s_{di}) \sim \text{beta}(\gamma_t, \gamma)$, where $\text{beta}(\cdot)$ is a Beta distribution, and $\gamma_t$ and $\gamma$ are two parameters;
    if $s_{di} = 0$ then
      Randomly select a pair $(v, v')$ from $d$'s authors, where $v$ is an author from $G^S$ and $v'$ from $G^T$;
      Draw a topic $z_{di} \sim \text{multi}(\theta_{v'v})$ from the topic mixture $\theta_{v'v}$ specific to $(v, v')$;
    end
    if $s_{di} = 1$ then
      Randomly select a user $v$;
      Draw a topic $z_{di} \sim \text{multi}(\theta_v)$ from the topic model of user $v$;
    end
  end
  Draw a word $x_{di} \sim \text{multi}(\phi_{z_{di}})$ from $z_{di}$-specific word distribution;
end
Intuitive explanation of Step 2 in CTL
cross-domain collaboration recommendation

Experiments
Data Set and Baselines

- Arnetminer (available at [http://arnetminer.org/collaboration](http://arnetminer.org/collaboration))

<table>
<thead>
<tr>
<th>Domain</th>
<th>Authors</th>
<th>Relationships</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Mining</td>
<td>6,282</td>
<td>22,862</td>
<td>KDD, SDM, ICDM, WSDM, PKDD</td>
</tr>
<tr>
<td>Medical Informatics</td>
<td>9,150</td>
<td>31,851</td>
<td>JAMIA, JBI, AIM, TMI, TITB</td>
</tr>
<tr>
<td>Theory</td>
<td>5,449</td>
<td>27,712</td>
<td>STOC, FOCS, SODA</td>
</tr>
<tr>
<td>Visualization</td>
<td>5,268</td>
<td>19,261</td>
<td>CVPR, ICCV, VAST, TVCG, IV</td>
</tr>
<tr>
<td>Database</td>
<td>7,590</td>
<td>37,592</td>
<td>SIGMOD, VLDB, ICDE</td>
</tr>
</tbody>
</table>

- Baselines
  - Content Similarity(Content)
  - Collaborative Filtering(CF)
  - Hybrid
  - Katz
  - Author Matching(Author), Topic Matching(Topic)
## Performance Analysis

**Training:** collaboration before 2001  
**Validation:** 2001-2005

<table>
<thead>
<tr>
<th>Cross Domain</th>
<th>ALG</th>
<th>P@10</th>
<th>P@20</th>
<th>MAP</th>
<th>R@100</th>
<th>ARHR -10</th>
<th>ARHR -20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Mining(S) to Theory(T)</td>
<td>Content</td>
<td>10.3</td>
<td>10.2</td>
<td>10.9</td>
<td>31.4</td>
<td>4.9</td>
<td>2.1</td>
</tr>
<tr>
<td>CF</td>
<td>15.6</td>
<td>13.3</td>
<td>23.1</td>
<td>26.2</td>
<td>4.9</td>
<td>2.8</td>
<td></td>
</tr>
<tr>
<td>Hybrid</td>
<td>17.4</td>
<td>19.1</td>
<td>20.0</td>
<td>29.5</td>
<td>5.0</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>Author</td>
<td>27.2</td>
<td>22.3</td>
<td>25.7</td>
<td>32.4</td>
<td>10.1</td>
<td>6.4</td>
<td></td>
</tr>
<tr>
<td>Topic</td>
<td>28.0</td>
<td>26.0</td>
<td>32.4</td>
<td>33.5</td>
<td>13.4</td>
<td>7.1</td>
<td></td>
</tr>
<tr>
<td>Katz</td>
<td>30.4</td>
<td>29.8</td>
<td>21.6</td>
<td>27.4</td>
<td>11.2</td>
<td>5.9</td>
<td></td>
</tr>
<tr>
<td>CTL</td>
<td><strong>37.7</strong></td>
<td><strong>36.4</strong></td>
<td><strong>40.6</strong></td>
<td><strong>35.6</strong></td>
<td><strong>14.3</strong></td>
<td><strong>7.5</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Content Similarity (Content):** based on similarity between authors’ publications  
**Collaborative Filtering (CF):** based on existing collaborations  
**Hybrid:** a linear combination of the scores obtained by the Content and the CF methods.  
**Katz:** the best link predictor in link-prediction problem for social networks  
**Author Matching (Author):** based on the random walk with restart on the collaboration graph  
**Topic Matching (Topic):** combining the extracted topics into the random walking algorithm
CTL can still maintain about 0.3 in terms of MAP which is significantly higher than baselines.
Parameter Analysis

(a) varying the number of topics $T$

(b) varying $\alpha$ parameter

(c) varying the restart parameter $\tau$ in the random walk

(d) Convergence analysis
Prototype System

http://arnetminer.org/collaborator

Cross-Domain Collaboration Recommendation

Treemap: representing subtopic in the target domain

Recommend Collaborators & Their relevant publications
Part C:
Further incorporate user feedback
“interactive collaboration recommendation”

(ACM TKDD, TIST, WSDM 2013-14)
Example

Finding co-inventors in IBM (>300,000 employers)

Recommend Candidates

Existing co-inventors

Interactive feedback

Refined Recommendations

Challenges

- What are the fundamental factors that influence the formation of co-invention relationships?

- How to design an interactive mechanism so that the user can provide feedback to the system to refine the recommendations?

- How to learn the interactive recommendation framework in an online mode?
interactive collaboration recommendation

Learning framework
RankFG Model

The problem is cast as, for each relationship, identifying which type has the highest probability.
Modeling with exponential family

Partially Labeled Model

\[ P(y_i \mid Y_{-i}) \propto \exp\left\{ \sum_{c_i} \sum_{k} \mu_k h_k (Y_{c_i}) \right\} \]

\[ P(x_i \mid y_i) \propto \exp\left\{ \sum_{j=1}^{d} \alpha_j g_j (x_{ij}, y_i) \right\} \]

Likelihood objective function

\[ P(Y \mid X, G) = \frac{P(X, G \mid Y)P(Y)}{P(X, G)} \]

\[ \propto P(X \mid Y) \cdot P(Y \mid G) = P(Y \mid G) \prod_i P(x_i \mid y_i) \]
Ranking Factor Graphs

• Pairwise factor function:

\[ f(v_q, v_i, y_i) = \frac{1}{Z_a} \exp\left\{ \sum_k \alpha_k \psi_k(x_q, x_i, y_i) \right\} \]

• Correlation factor function:

\[ g(y_i, y_j) = \frac{1}{Z_b} \exp\left\{ \sum_l \beta_l \phi_l(y_i, y_j) \right\} \]

• Log-likelihood objective function:

\[ \log P(Y|X, \theta) = \sum_{y_i \in Y} \sum_k \alpha_k \psi_k(x_q, x_i, y_i) \]
\[ + \sum_{v_i \sim v_j} \sum_l \beta_l \phi_l(y_i, y_j) - \log Z \]

• Model learning

\[ \theta^* = \arg \max_{\theta} \log P(Y|X, \theta) \]
Learning Algorithm

Input: Query inventors \( Q = \{ v_q \} \) with corresponding topics \( \{ q \} \), \( G = (V, E, X) \), and the learning rate \( \eta \);

Output: learned parameters \( \theta \);

\( \theta \leftarrow 0 \);

repeat

    foreach \( v_q \in Q \text{ and } q \) do

        // Initialization;
        \( L \leftarrow \text{initialization list} \);
        Factor graph \( FG \leftarrow BuildFactorGraph(L) \);
        // Learn the parameter \( \theta \) for factor graph model;
        repeat

            foreach \( v_i \in \text{order} \) do

                Update the messages of \( v_i \) by Eqs. 8 and 9;

            end

        until (all messages \( \mu \) do not change);

        foreach \( \theta_i \in \theta \) do

            Calculate gradient \( \nabla_i \) according to Eq. 7;
            Update \( \theta_{\text{new}} = \theta_{\text{old}} + \eta \cdot \nabla_i \);

        end

    until converge;

Algorithm 1: Learning algorithm for RankFG.
Still Challenge

How to incrementally incorporate users’ feedback?
Learning Algorithm

Input: Query inventors $Q = \{v_q\}$ with corresponding topics $\{q\}$, $G = (V, E, X)$, and the learning rate $\eta$;
Output: learned parameters $\theta$;

$\theta \leftarrow 0$;

repeat
    foreach $v_q \in Q$ and $q$ do
        // Initialization;
        $L \leftarrow$ initialization list;
        Factor graph $FG \leftarrow BuildFactorGraph(L)$;
        // Learn the parameter $\theta$ for factor graph model;
        repeat
            foreach $v_i \in$ order do
                Update the messages of $v_i$ by Eqs. 8 and 9;
            end
            until (all messages $\mu$ do not change);
        
    foreach $\theta_i \in \theta$ do
        Calculate gradient $\nabla_i$ according to Eq. 7;
        Update $\theta^{new} = \theta^{old} + \eta \cdot \nabla_i$;
    end
    until converge;

Algorithm 1: Learning algorithm for RankFG.
Interactive Learning

1) add new factor nodes to the factor graph built in the model learning process.

2) \(l\)-step message passing:
   - Start from the new variable node \(y_{N+1}\).
   - Send messages to all of its neighborhood factors.
   - Propagate the messages up to \(l\)-step.
   - Perform a backward messages passing.

3) Calculate an approximate value of the marginal probabilities of the newly factors.

\[
E^{new}[.] = \frac{N}{N+1}E^{old}[.] + \frac{1}{N+1} \sum_k \theta_k \phi_k(x_{N+1}, y_{N+1})
\]
From passive interactive to active

- Entropy
  \[ \mu(v) = \sum_{y \in Y} B_v(y) \log \frac{1}{B_v(y)} \]

- Threshold
  \[ t(v) = \min\{[\eta(\mu_{\max} - \mu(v))d(v)], d(v)\} \]

- Influence model
  \[ f_\tau(v) = \begin{cases} 1 & \text{if } \sum_{u \in NB(v)} f_{\tau-1}(u) \geq t(v) \\ 0 & \text{if } \sum_{u \in NB(v)} f_{\tau-1}(u) < t(v) \end{cases} \]


Active learning via Non-progressive diffusion model

• Maximizing the diffusion

\[
\max \left\{ \max_{V_S \subseteq V_U} |V_T|, \quad |V_S| \leq k \right\}
\]

with the constraints:

\[f_0(v) = 1 \iff v \in V_S\]  \hspace{1cm} (2)
\[
\exists \tau_M \text{ s.t. } \forall v \in V_T \quad \forall \tau > \tau_M \quad f_\tau(v) = 1 \]  \hspace{1cm} (3)
\[
\forall v \in V_U \setminus V_T, \forall u \in V_T, \mu(v) \leq \mu(u) \]  \hspace{1cm} (4)
\[
f_\tau(v) = 1 \iff \sum_{u \in \text{NB}(v)} f_{\tau-1}(u) \geq t(v) \]  \hspace{1cm} (5)

NP-hard!
MinSS

- Greedily expand $V_p$
MinSS (cont.)

14 if \( V_P = \emptyset \) then
15 sort nodes in \( V_T \) in ascending order of \( d(v) \) as
16 \( v_1, v_2, \ldots, v_m \)
17 \( \text{foreach } v \in V_T \text{ do} \)
18 \( w(v) \leftarrow 0 \)
19 \( \text{foreach } i \leftarrow 1 \text{ to } m \text{ do} \)
20 \( \text{if } \exists u \in NB(v_i) \cap V_T \text{ st. } w(u) = d(u) - t(u) \text{ then} \)
21 \( V_S \leftarrow V_S \cup \{v_i\} \)
22 \( \text{else} \)
23 \( \text{foreach } u \in NB(v_i) \cap V_T \text{ do} \)
24 \( w(u) \leftarrow w(u) + 1 \)
**Theorem 3. Lower Bound.** Let $D(V) = \sum_{v \in V} d(v)$, $T(V) = \sum_{v \in V} t(v)$, and suppose $t(v) \leq \beta d(v)$ for all $v \in V$. If $2T(V_U) - D(V_U) > 0$ and $V_T = V_U$, we have an lower bound for optimal solution $|V_{S,\text{opt}}|$ to problem 2.

$$|V_{S,\text{opt}}| \geq \frac{2T(V_U) - D(V_U)}{\beta \Delta} \quad (9)$$

**Theorem 4. Upper Bound.** Suppose $t(v) \leq \beta d(v)$ for all $v \in V$, we can derive an upper bound for MinSS algorithm.

$$|V_S| \leq \frac{\beta \Delta}{1 - \beta + \beta \Delta} |V_U|$$
Corollary 2. Approximation Ratio. Let $V_{S,g}$ denote the solution given by MinSS algorithm, $V_{S,\text{opt}}$ represent the optimal solution and $\Delta$ be the maximum degree in the graph. Suppose $t(v) \leq \beta d(v)$ for all $v \in V$, if $V_T = V_U$ and $2T(V_U) > D(V_U)$, we have

\[
|V_{s,\text{opt}}| \geq \frac{2T(V_U) - D(V_U)}{\beta \Delta}
\]

\[
|V_S| \leq \frac{\beta \Delta}{1 - \beta + \beta \Delta} |V_U|
\]

\[
\frac{|V_{S,g}|}{|V_{S,\text{opt}}|} \leq \frac{(\beta \Delta)^2}{(1 - \beta + \beta \Delta) \cdot \mathbb{E}[2t(v) - d(v)]}
\]

where $\mathbb{E}[\cdot]$ represents the expectation over all samples in the network.
interactive collaboration recommendation

Experiments
Data Set

• PatentMiner (pminer.org)

<table>
<thead>
<tr>
<th>DataSet</th>
<th>Inventors</th>
<th>Patents</th>
<th>Average increase #patent</th>
<th>Average increase #co-invention</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>55,967</td>
<td>46,782</td>
<td>8.26%</td>
<td>11.9%</td>
</tr>
<tr>
<td>Intel</td>
<td>18,264</td>
<td>54,095</td>
<td>18.8%</td>
<td>35.5%</td>
</tr>
<tr>
<td>Sony</td>
<td>8,505</td>
<td>31,569</td>
<td>11.7%</td>
<td>13/0%</td>
</tr>
<tr>
<td>Exxon</td>
<td>19,174</td>
<td>53,671</td>
<td>10.6%</td>
<td>14.7%</td>
</tr>
</tbody>
</table>

• Baselines:
  – Content Similarity (Content)
  – Collaborative Filtering (CF)
  – Hybrid
  – SVM-Rank

### Performance Analysis - IBM

**Training:** collaboration before 2000  
**Validation:** 2001-2010

<table>
<thead>
<tr>
<th>Data</th>
<th>ALG</th>
<th>P@5</th>
<th>P@10</th>
<th>P@15</th>
<th>P@20</th>
<th>MAP</th>
<th>R@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>Content</td>
<td>23.0</td>
<td>23.3</td>
<td>18.8</td>
<td>15.6</td>
<td>24.0</td>
<td>33.7</td>
</tr>
<tr>
<td></td>
<td>CF</td>
<td>13.8</td>
<td>12.8</td>
<td>11.3</td>
<td>11.5</td>
<td>21.7</td>
<td>36.4</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>13.9</td>
<td>12.8</td>
<td>11.5</td>
<td>11.5</td>
<td>21.8</td>
<td>36.7</td>
</tr>
<tr>
<td></td>
<td>SVMRank</td>
<td>13.3</td>
<td>11.9</td>
<td>9.6</td>
<td>9.8</td>
<td>22.2</td>
<td>43.5</td>
</tr>
<tr>
<td></td>
<td>RankFG</td>
<td>31.1</td>
<td>27.5</td>
<td>25.6</td>
<td>22.4</td>
<td>40.5</td>
<td>46.8</td>
</tr>
<tr>
<td></td>
<td>RankFG+</td>
<td>31.2</td>
<td>27.5</td>
<td>26.6</td>
<td>22.9</td>
<td>42.1</td>
<td>51.0</td>
</tr>
</tbody>
</table>

RankFG+: it uses the proposed RankFG model with 1% interactive feedback.
Interactive Learning Analysis

Interactive learning achieves a close performance to the complete learning with only $1/100$ of the running time used for complete training.
Parameter Analysis

Factor contribution analysis

- **RankFG-C**: stands for ignoring referral chaining factor functions.
- **RankFG-CH**: stands for ignoring both referral chaining and homophily.
- **RankFG-CHR**: stands for further ignoring recency.

Convergence analysis
Results of Active Learning
Summaries

• Inferring social ties in single network
  – Time-dependent factor graph model
• Cross-domain collaboration recommendation
  – Cross-domain topic learning
• Interactive collaboration recommendation
  – Ranking factor graph model
  – Active learning via non-progressive diffusion
Inferring social ties

Reciprocity

Triadic Closure

Future Work
References

- Yi Cai, Ho-fung Leung, Qing Li, Hao Han, Jie Tang, Juanzi Li. Typicality-based Collaborative Filtering Recommendation. IEEE Transaction on Knowledge and Data Engineering (TKDE).
- Zhilin Yang, Jie Tang, and Bin Xu. Active Learning for Networked Data Based on Non-progressive Diffusion Model. WSDM’14.
- Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain Collaboration Recommendation. KDD’12, pages 1285-1293. (Full Presentation & Best Poster Award)
- Jie Tang, Tiancheng Lou, and Jon Kleinberg. Inferring Social Ties across Heterogeneous Networks. WSDM’12, pages 743-752.
- Chi Wang, Jiawei Han, Yuntao Jia, Jie Tang, Duo Zhang, Yintao Yu, and Jingyi Guo. Mining Advisor-Advisee Relationships from Research Publication Networks. KDD’10, pages 203-212.
- Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. KDD’08, pages 990-998.
Thank you!

Collaborators: John Hopcroft, Jon Kleinberg (Cornell)
Jiawei Han and Chi Wang (UIUC)
Tiancheng Lou (Google)
Jimeng Sun (IBM)
Jing Zhang, Zhanpeng Fang, Zi Yang, Sen Wu (THU)

Jie Tang, KEG, Tsinghua U,
Download all data & Codes,
http://keg.cs.tsinghua.edu.cn/jietang
http://arnetminer.org/download