Mining Organizational Structure in Social Network

Presenter: Zhang Bo

Organizational Structure

- More than simply related or not.
- Reveals the direction of supervision and influence.
- Examples:
  - Advisor-advisee relationship
  - Terrorist organization hierarchy

Background

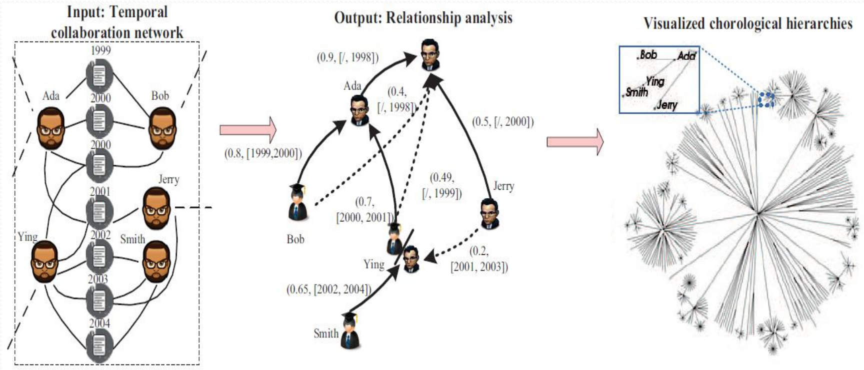
- Community Discovery
  - Goal: discover related groups that have denser intra-group communication
  - Often reveals interesting properties. Common hobbies, social functions, etc.
  - Fail to show power of members and their scope of influence.
- Organizational Structure Discovery
  - Good for finding members influential power within the structure.
  - Useful in many applications.

Advisor-Advisee Relationship

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.

- Given: publication data with co-author list
- Target: Among those co-authors, find advisor-advisee pairs.
- Used to find experts, or to see students of an expert.

Example



Preliminaries

- a_i: author i
- a_{yi}: advisor of a_i
- [st_{ij}, ed_{ij}]: time interval that i's advisor is j, i.e., [2003, 2007]
- [st_i, ed_i]: (briefly) time interval that i is advised

- py_i: pub_year_vector of i, i.e., [2003, 2004, 2005]
- pn_i: pub_num_vector of i, i.e., [2, 3, 4]

- py_{ij}: pub_year_vector of co-author i and j; link property
- pn_{ij}: pub_num_vector of co-author i and j; link property

- py^1_i: first component of py_i

Assumptions

1) $ed_j < st_i < ed_i$
   - *j* can only advise *i* after j graduated.

1) $py^1_j < py^1_{ij}$
   - Advisor *j* should always have a longer publication history than advisee *i*.

More Assumptions

- $Kulc_{ij}$: Kulczynski ratio. Correlation of two authors' publications
- IR_{ij}: Imbalance ratio between $(j|i)$ and $(i|j)$

- *j* is not *i*'s advisor if
  - $IR_{ij} < 0$ during the collaboration period. Advisor should have more publications than advisee
  - $Kulc_{ij}$ does not increase during the collaboration period
  - The collaboration period lasts for only one year
  - $py^1_j + 2 > py^1_{ij}$

Approach Step 1

- Step 1: preprocessing
  - Remove unlikely pairs;
  - Generate candidate graph, which is a DAG

Approach Step 2

- TPFG: Time-constrained Probabilistic Factor Graph model
- Let y_i be advisor of a_i; we need to decide tuple (y_i, st_i, ed_i)
- Suppose a local feature function g(y_i, st_i, ed_i). Joint probability is defined as

$$P(\{y_i, st_i, ed_i\}_{a_i \in V^a}) = \frac{1}{Z} \prod_{a_i \in V^a} g(y_i, st_i, ed_i)$$

- With assumption 1 as the constraint

Approach Step 2

- To find most possible relations, maximize the joint probability
- Exhaustive search: $O((CT^2)^n)$, C candidates/author, with period variable in range T.
- Optimize local feature function to find best advising time $[st_i, ed_i]$ for i. Only $\{y_i\}$ is left for optimization

Performance

data set	RULE	SVM	IndMAX		TPFG	
TEST1	69.9%	73.4%	75.2%	78.9%	80.2%	84.4%
TEST2	69.8%	74.6%	74.6%	79.0%	81.5%	84.3%
TEST3	80.6%	86.7%	83.1%	90.9%	88.8%	91.3%

TRAIN1=Colleague(491)+PHD(100)
TEST1=Teacher(257)+MathGP(1909)+Colleague(2166)
TRAIN2=TRAIN3=Teacher(257)+Colleague(2166)
TEST2=PHD(100)+MathGP(1909)+Colleague(4351)
TEST3=AIGP(666)+Colleague(459)

Issues:

- Need the insight of relationship characteristics. Difficult to be generalized for other kind of relationships
- How to appropriately interpret the result probabilities: 95%, 5%, 51%
- Real world scenario:
  - A is B's advisor in Computer Science;
  - B is A's advisor in music;
  - Similar amount of publications;
  - All possible relations between st_A, st_B, ed_A, ed_B, etc.

Relative Importance in Networks

Scott White and Padhraic Smyth.
Algorithms for estimating relative importance in networks. KDD '03.

- Given a relationship network, rank nodes' importance
- Focus: How much "importance" node *t* inherited from node *r*

K-Short Node-Disjoint paths

- Why not shortest/closeness/betweenness: longer paths may play important role
- Why node-disjoint: otherwise nodes and edges may appear multiple times in different paths.
- *P(r, t)* : set of paths from *r* to *t*.
- P_i : the i^{th} path in *P*
- λ :scaling factor

$$I(t|r) = \sum_{i=1}^{|\mathcal{P}(r,t)|} \lambda^{-|p_i|}$$

Markov Centrality

- *n*: number of steps taken
- $f^n{}_{rt}$: probability the chain first return to *t* in exactly *n* steps
- m_{rt}: mean first passage time from *r* to *t*
- R: given root set

$$m_{rt} = \sum_{n=1}^{\infty} n f_{rt}^{(n)}$$

$$I(t|R) = \frac{1}{\frac{1}{|R|}\sum_{r \in R} m_{rt}}$$

PageRank with Priors

- $P_R = \{p_1,\ldots,p_V\}$: prior probabilities(importances) attached to roots, i.e., $p_1 = \ldots = p_V = 1/|R|$
- $0 \leqslant \beta \leqslant 1$: probability that we jump back to R
- Iterative stationary probability equation:

$$\pi(v)^{(i+1)} = (1-\beta)\left(\sum_{u=1}^{d_{in}(v)} p(v|u)\pi^{(i)}(u)\right) + \beta p_v$$

- After converge:

$$I(v|R) = \pi(v)$$

HITS with Priors

- Similar assumption

$$H^{(i)} = \sum_{v=1}^{|V|} \sum_{u=1}^{d_{in}(v)} h^{(i)}(u)$$
$$A^{(i)} = \sum_{v=1}^{|V|} \sum_{u=1}^{d_{out}(v)} a^{(i)}(u)$$

$$a^{(i+1)}(v) = (1-\beta)\left(\sum_{u=1}^{d_{in}(v)} \frac{h^{(t)}(u)}{H^{(i)}}\right) + \beta p_v$$
$$h^{(i+1)}(v) = (1-\beta)\left(\sum_{u=1}^{d_{out}(v)} \frac{a^{(t)}(u)}{A^{(i)}}\right) + \beta p_v$$

K-Step Markov

- Random walk starting from R
- Back probability β
- Fixed-length K
- Compute: Relative probability that the system spend time at any node, after K steps
- A: Markov transition matrix

$$I(t|R) = \left[\mathbf{A}\mathbf{p}_R + \mathbf{A}^2\mathbf{p}_R \ldots \mathbf{A}^K\mathbf{p}_R\right]_t$$

911 European Al Qaeda terrorist network

- Known fact:
  - Djamal Beghal has been a leader
  - Key roles: Khemais, Maaroufi, Daoudi, and Moussaoui
  - 911 leader: Mohammed Atta

Rank	PRankP		HITSP		WKPaths		MarkovC		KSMarkov	
1:	Khemais	0.221	Khemais	0.173	Beghal	0.045	Atta	0.063	Khemais	0.115
2:	Beghal	0.218	Beghal	0.166	Khemais	0.045	Al-Shehhi	0.041	Beghal	0.108
3:	Moussaoui	0.044	Atta	0.038	Moussaoui	0.045	al-Shibh	0.037	Moussaoui	0.065
4:	Maaroufi	0.039	Moussaoui	0.029	Maaroufi	0.044	Moussaoui	0.036	Maaroufi	0.059
5:	Qatada	0.036	Maaroufi	0.026	Bensakhria	0.037	Jarrah	0.030	Qatada	0.052
6:	Daoudi	0.035	Qatada	0.025	Daoudi	0.037	Hanjour	0.028	Daoudi	0.049
7:	Courtaillier	0.032	Bensakhria	0.023	Qatada	0.036	Al-Omari	0.026	Bensakhria	0.045
8:	Bensakhria	0.031	Daoudi	0.023	Walid	0.031	Khemais	0.025	Courtaillier	0.045
9:	Walid	0.030	Courtaillier	0.022	Courtaillier	0.031	Qatada	0.025	Walid	0.040
10:	Khammoun	0.025	Khammoun	0.021	Khammoun	0.029	Bahaji	0.024	Khammoun	0.034

Coauthership Network

- R = {Brin, Page, Kleinberg}

Rank	PRankP		HITSP		WKPaths		KSMarkov	
1:	Brin	0.2014	Brin	0.1119	Kleinberg	0.0023	Brin	0.1045
2:	Page	0.1352	Kleinberg	0.1107	Brin	0.0019	Motwani	0.0627
3:	Kleinberg	0.1137	Page	0.1087	Motwani	0.0017	Ullman	0.0536
4:	Motwani	0.0474	Motwani	0.0184	Raghavan	0.0016	Silverstein	0.0467
5:	Ullman	0.0429	Raghavan	0.0147	Page	0.0014	Page	0.0394
6:	Silverstein	0.0392	Ullman	0.0136	Silverstein	0.0014	Kleinberg	0.0194
7:	Raghavan	0.0111	Silverstein	0.0119	Ullman	0.0014	Raghavan	0.0138
8:	Lynch	0.0086	Williamson	0.0113	Williamson	0.0012	Zhang	0.0109
9:	Kedem	0.0086	Papadimitriou	0.0110	Vempala	0.0012	Guibas	0.0106
10:	Williamson	0.0085	Lynch	0.0108	Indyk	0.0010	Robertson	0.0101

Evolving Networks

Jiangtao Qiu, Zhangxi Lin, Changjie Tang, and Shaojie Qiao.
Discovering Organizational Structure in Dynamic Social Network
ICDM '09

- Algorithm
  - Random walk to find the community tree
  - Modified PageRank algorithm for m-score computation
- Novalty: min-distance-error evolving tree
  - Good for observing power changes
- Insufficient and prelimary results. No comparison to state-of-art.

Thank You!