THE 18TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING
KDD 2012 BEIJING
Beijing, China
August 12-16, 2012

Cross-domain Collaboration Recommendation

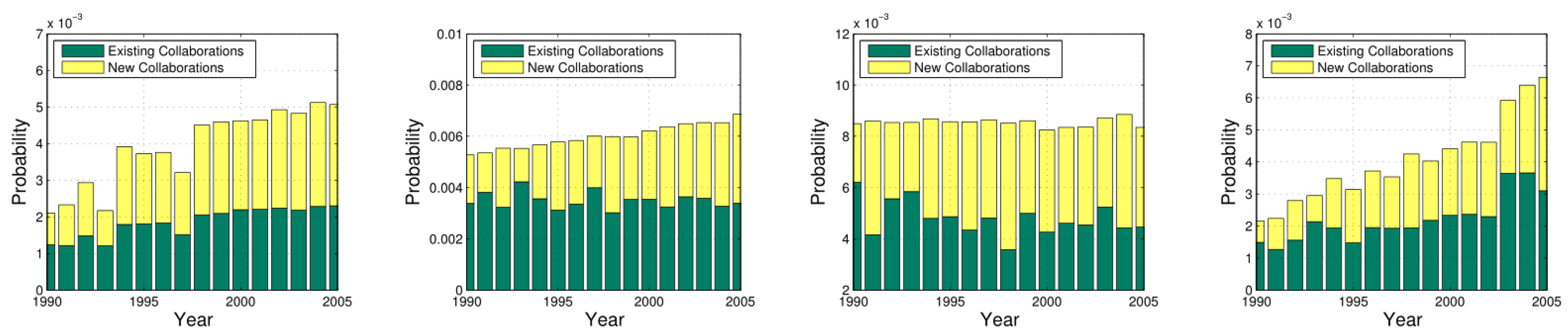
Jie Tang
Computer Science
Tsinghua University
Beijing, China
jietang@tsinghua.edu.cn

Sen Wu
Computer Science
Tsinghua University
Beijing, China
ronaldosen@gmail.com

Jimeng Sun
IBM TJ Watson
Research Center
Hawthorne, NY, USA
jimeng@us.ibm.com

Hang Su
Computer Science
Tsinghua University
Beijing, China
suhang@sse.buaa.edu.cn

Interdisciplinary collaborations have generated huge impact to society.

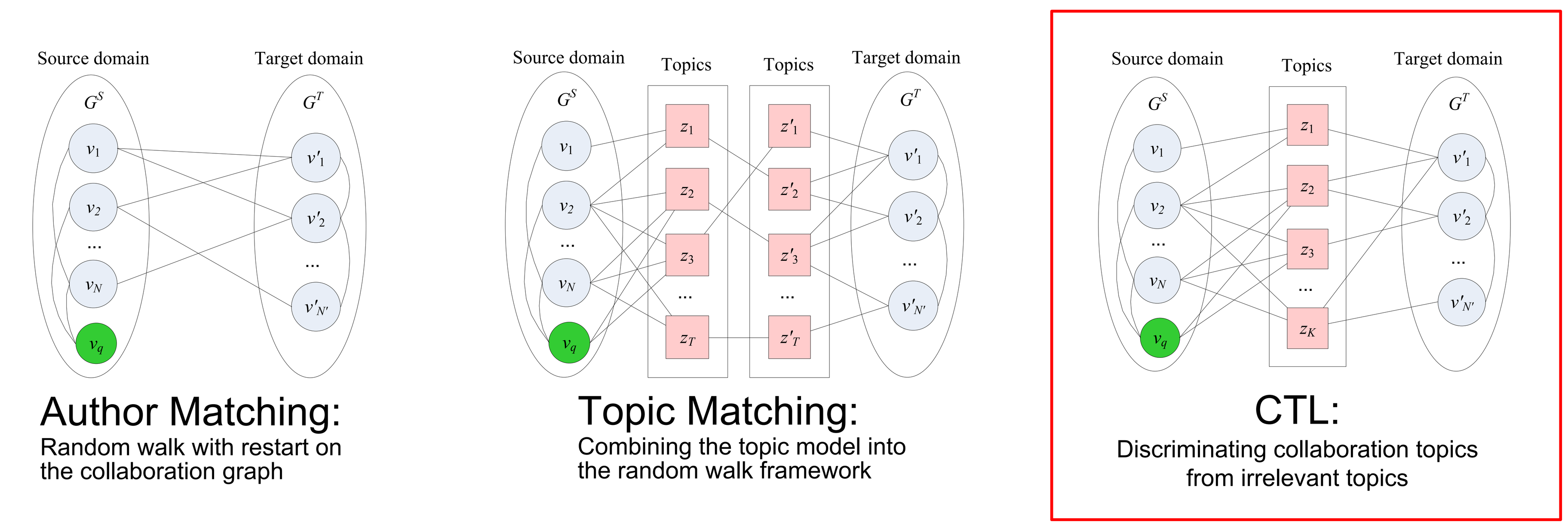


(a) DM - TH (b) DM - MI (c) DM - VIS (d) MI - DB

Trends of existing and new collaborations over years.

Cross-domain collaborations is very different from traditional collaborations:
1) sparse connection: cross-domain collaborations are rare
2) complementary expertise: cross-domain collaborators have different expertise
3) topic skewness: cross-domain collaboration topics are focused on a subset of topics

Cross-domain Topic Learning (CTL)



Author Matching: Random walk with restart on the collaboration graph

Topic Matching: Combining the topic model into the random walk framework

CTL: Discriminating collaboration topics from irrelevant topics

Notations

SYMBOL	DESCRIPTION
T	number of topics
d	a collaborated document
A_d	a set of authors of document d
x_{di}	the ith attribute (word) in document d
z_{di}	the topic assigned to attribute x_{di}
s_{di}	if x_{di} is a word from a single domain or a cross domain
θ_v	multinomial distribution over topics specific to author v
$\vartheta_{vv'}$	multinomial distribution over topics specific to author pair (v,v')
ϕ_z	multinomial distribution over words specific to topic z
α, β	Dirichlet priors to multinomial distributions θ, θ' and ϕ
λ	parameter for sampling the binary variable s
γ, γ_t	Beta parameters to generate λ

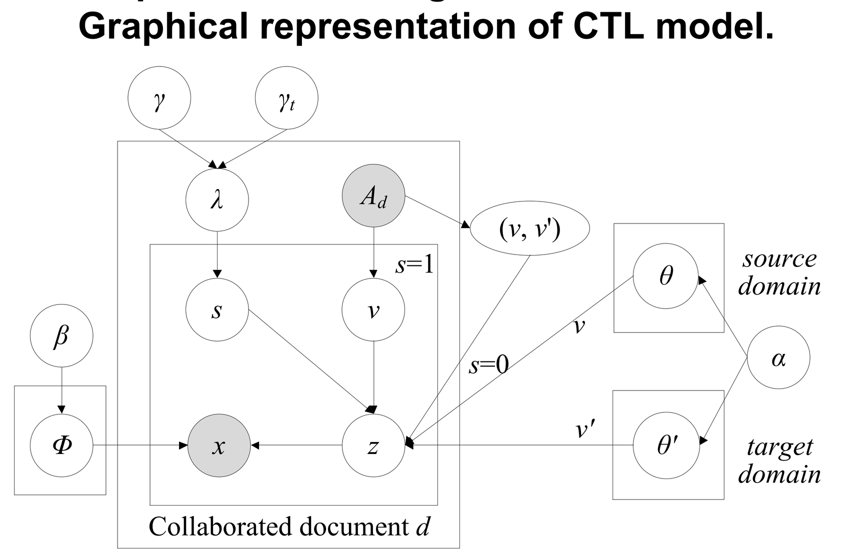
Probabilistic generative process in CTL

Input: a source domain G^S and a target domain G^T
Output: estimated parameters θ,θ',ϕ, ϑ, and λ

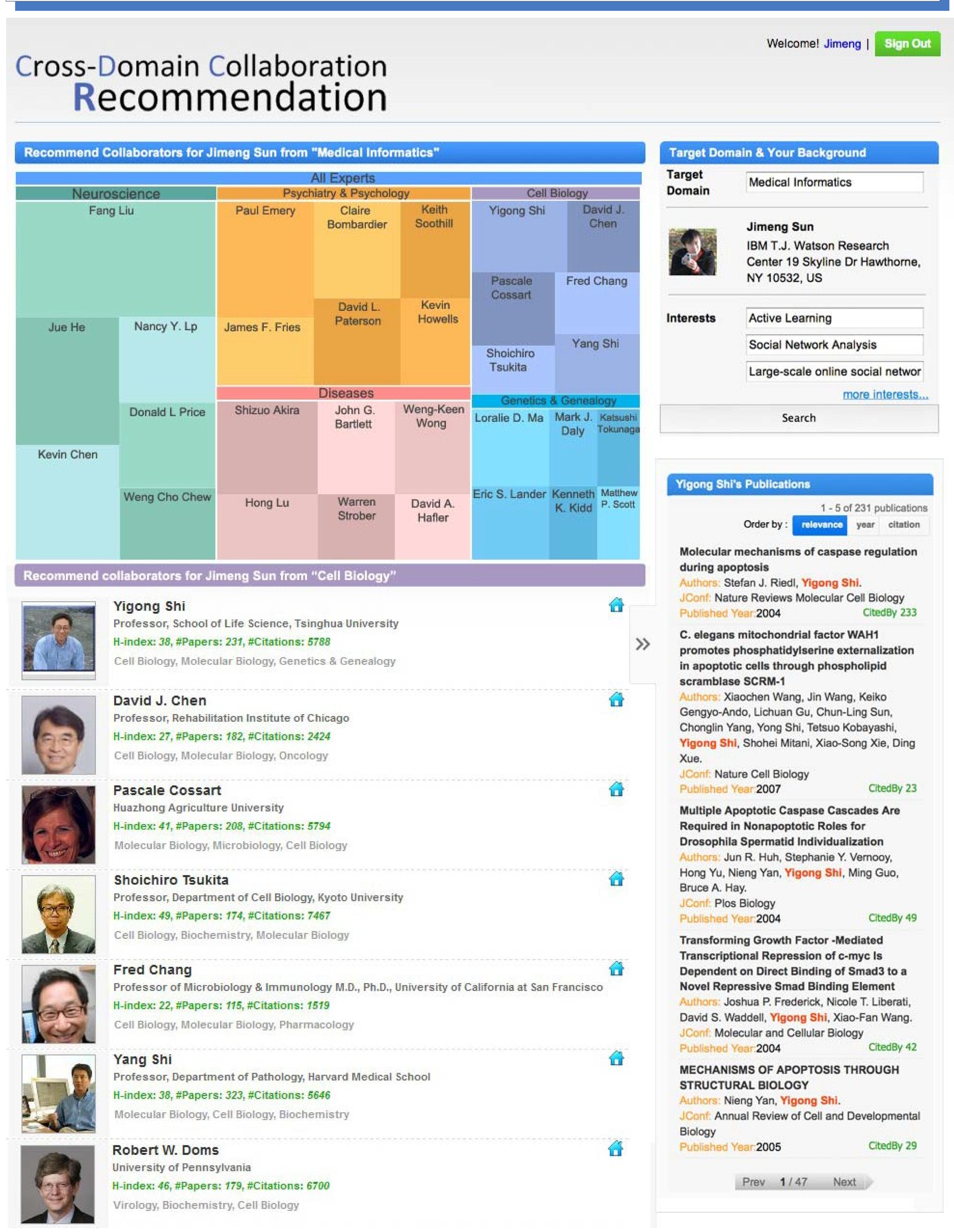
Initialize an ACT model in G^S by learning from documents written by authors only from G^S;
Similarly, initialize an ACT model for target domain G^T;
foreach *collaborated document d* **do**
  foreach *word* $x_{di} \in d$ **do**
    Toss a coin s_{di} according to $bernoulli(s_{di}) \sim beta(\gamma_t, \gamma)$, where $beta(.)$ is a Beta distribution, and γ_t and γ 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 $z_{di} \sim multi(\vartheta_{vv'})$ from the topic mixture $\vartheta_{vv'}$ specific to (v, v');
    end
    if $s_{di} = 1$ **then**
      Randomly select a user v;
      Draw a topic $z_{di} \sim multi(\theta_v)$ from the topic model of user v;
    end
  end
  Draw a word $x_{di} \sim multi(\phi_{z_{di}})$ from z_{di}-specific word distribution;
end

Step 1. Learning LDA or ACT model on the source and the target domain respectively.

Step 2. CTL Learning
Graphical representation of CTL model.

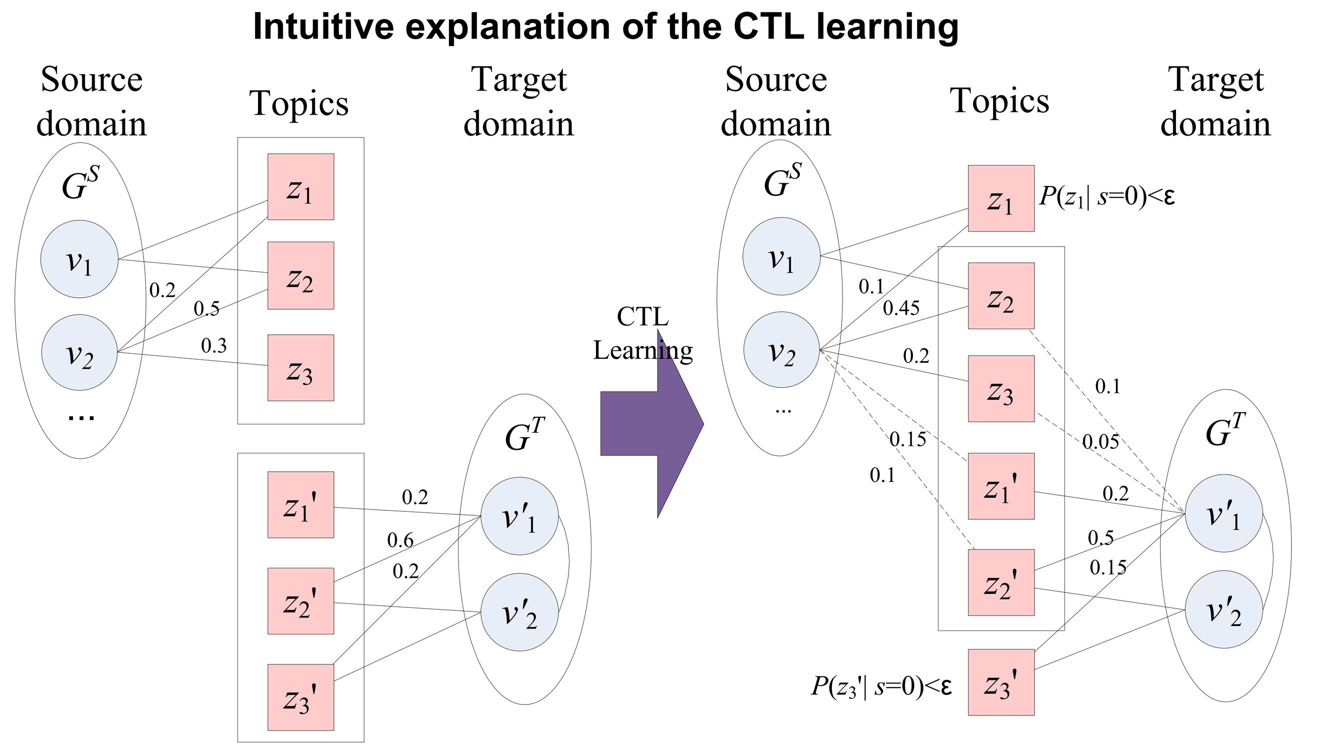


Prototype system: http://arnetminer.org/collaborator



Step 3. Random walk with restart on the topic augmented graph.

Intuitive explanation of the CTL learning



Empirical Analysis

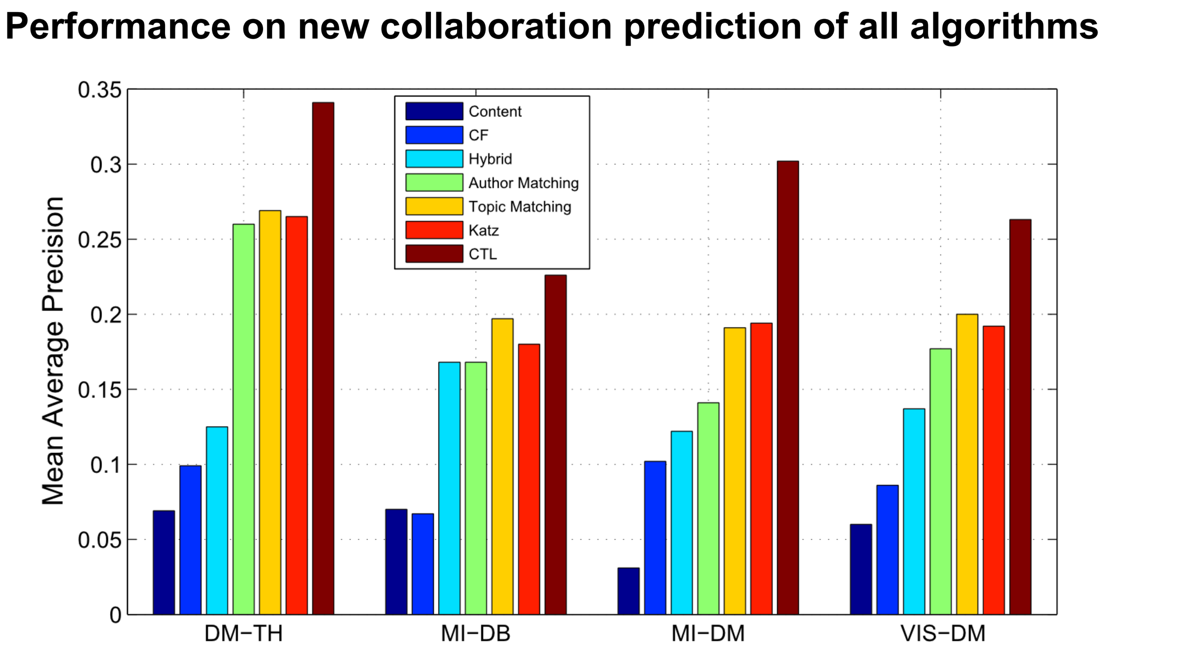
Datasets (from Arnetminer): 5 domains

Data Mining(DM)—6,282 authors and 22,862 relationships.
Medical Informatics(MI)—9,150 authors and 31,851 relationships.
Theory(TH)—5,449 authors and 27,712 relationships.
Visualization(VIS)—5,268 authors and 19,261 relationships.
Database(DB)—7,590 authors and 37,592 relationships.

Baselines:

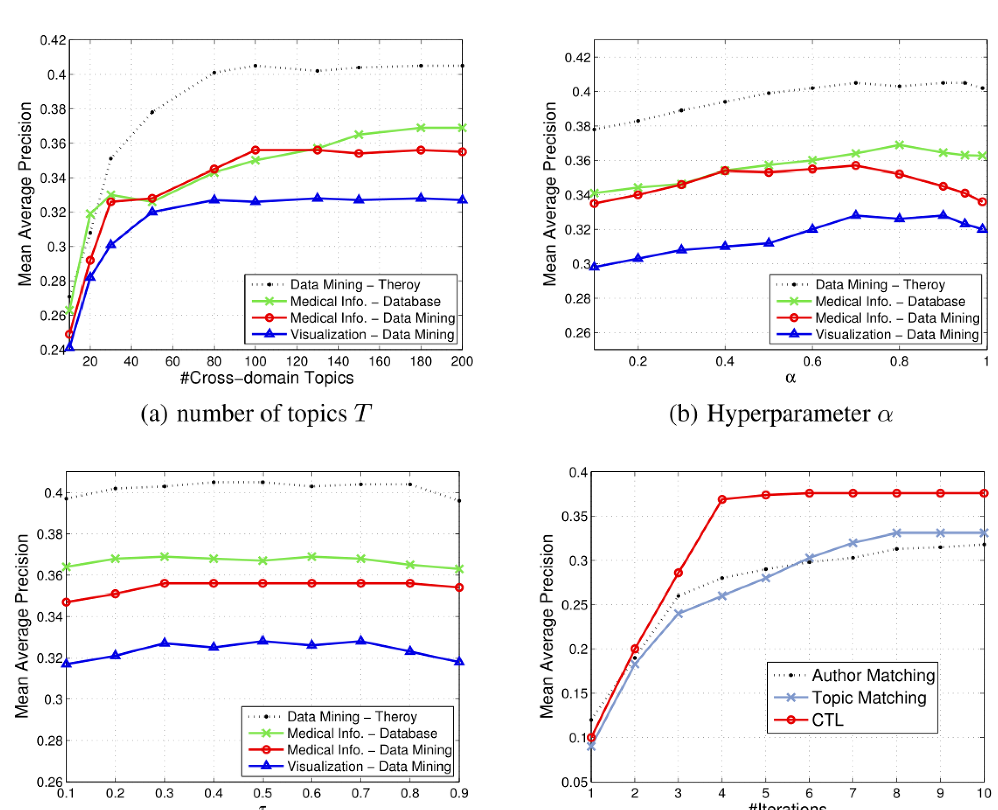
Content Similarity(Content)—based on similarity between authors's publications
Collaborative Filtering(CF)—based on existing collaborations
Hybrid— a linear combination of the scores obtained by the Content and the CF methods.
Katz—the best link predictor in link-prediction problem for social networks
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

Performance on new collaboration prediction of all algorithms



Cross domain	ALG	P@10	P@20	MAP	R@100	ARHR-10	ARHR-20
Data Mining (S) to Theory (T)	Content	10.3	10.2	10.9	31.4	4.9	2.1
	CF	15.6	13.3	23.1	26.2	4.9	2.8
	Hybrid	17.4	19.1	20.0	29.5	5.0	2.4
	Author	27.2	22.3	25.7	32.4	10.1	6.4
	Topic	28.0	26.0	32.4	33.5	13.4	7.1
	Katz	30.4	29.8	31.6	27.4	11.2	5.9
	CTL	**37.7**	**36.4**	**40.6**	**35.6**	**14.3**	**7.5**
Medical Info. (S) to Database (T)	Content	10.1	10.9	12.5	45.9	3.6	2.1
	CF	18.3	20.2	21.4	47.6	5.3	3.9
	Hybrid	25.0	26.5	28.4	59.1	6.4	4.2
	Author	26.2	29.6	32.2	54.8	10.5	**5.4**
	Topic	29.4	26.3	34.7	59.3	**11.5**	5.2
	Katz	27.5	28.3	30.7	57.2	10.5	5.0
	CTL	**32.5**	**30.0**	**36.9**	**59.8**	11.4	**5.4**
Medical Info. (S) to Data Mining (T)	Content	5.8	5.7	9.5	19.8	1.9	0.9
	CF	13.7	17.8	18.9	34.3	2.7	1.3
	Hybrid	18.0	19.0	19.8	36.7	3.4	1.3
	Author	20.1	23.8	29.3	**64.4**	5.3	2.1
	Topic	26.0	**25.0**	33.9	48.1	10.7	5.6
	Katz	21.2	23.8	32.4	48.1	10.2	4.8
	CTL	**30.0**	24.0	**35.6**	49.6	**12.2**	**6.0**
Visual. (S) to Data Mining (T)	Content	9.6	11.8	13.2	18.9	3.1	1.8
	CF	14.0	20.8	26.4	29.4	6.9	4.3
	Hybrid	16.0	20.0	27.6	30.1	6.3	4.4
	Author	22.0	25.2	27.7	31.1	11.9	6.7
	Topic	26.3	25.0	32.3	31.4	13.2	8.8
	Katz	23.0	25.1	29.3	30.2	10.4	5.4
	CTL	**28.3**	**26.0**	**32.8**	**36.3**	**14.0**	**9.1**

Recommendation performance(%)



(a) number of topics T (b) Hyperparameter α (c) RWR parameter τ (d) Convergence analysis

Parameter analysis

Paper ID: 535

References

J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In WWW'10, pages 641-650, 2010.
D. Liben-Nowell and J. M. Kleinberg. The link-prediction problem for social networks. JASIST, 58(7):1019-1031, 2007.
J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In ICDM'05, pages 418-425, 2005.
J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In KDD'08, pages 990-998, 2008.