

THE 18TH ACM SIGKDD CONFERENCE ON Beijing, China **KNOWLEDGE DISCOVERY AND DATA MINING** 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

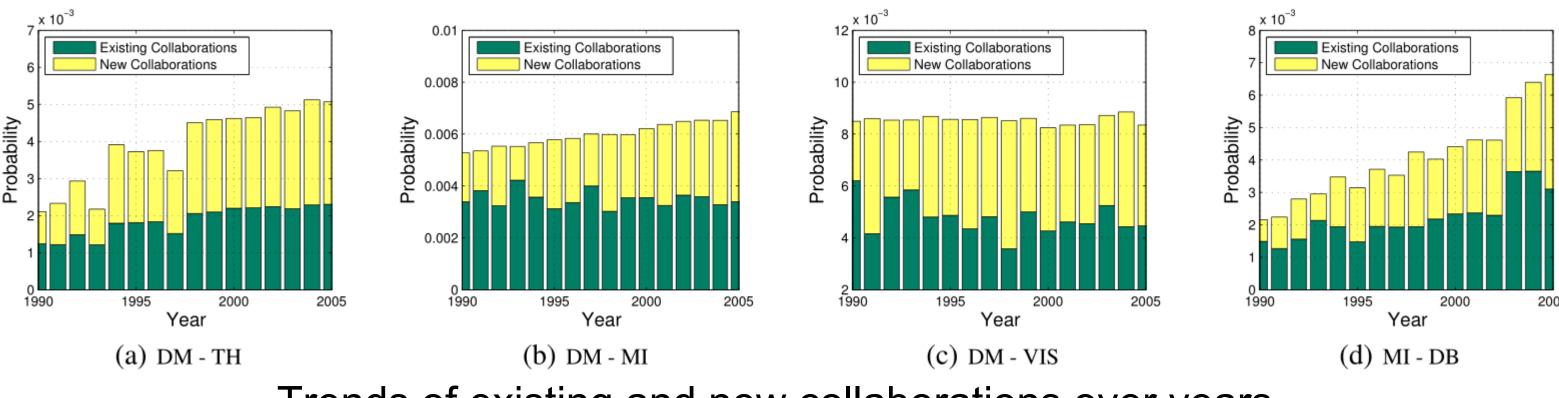
jimeng@us.ibm.com

**Computer Science Tsinghua University** Beijing, China

Hang Su

suhang@sse.buaa.edu.cn





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

### **Cross-Domain Collaboration** Recommendation

Yang Shi



CitedBy 42

Published Year:2004

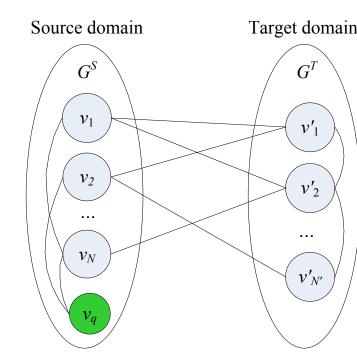
MECHANISMS OF APOPTOSIS THROUGH

Trends of existing and new collaborations over years.

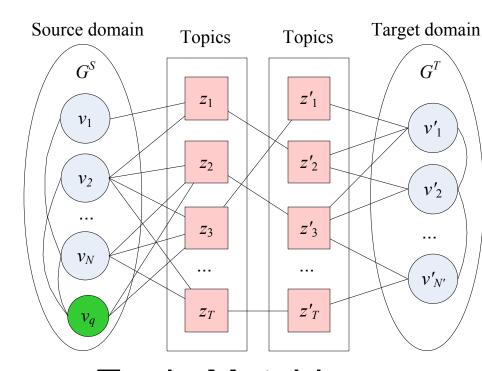
#### Cross-domain collaborations is very different from traditional collaborations:

- 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

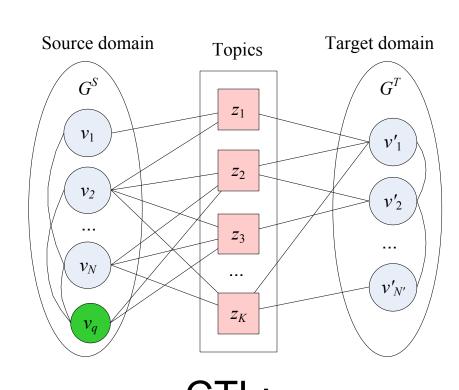
#### **Cross-domain Topic Learning (CTL)**



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

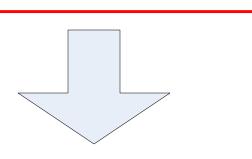
			All Experts					Target		
Neuro	science		iatry & Psycholo	av	Cell	Biology		Domain	Medical Informatics	
Fang Liu		Paul Emery	Claire Bombardier	Keith Soothill	Ch		vid J. nen		Jimeng Sun IBM T.J. Watson Resea Center 19 Skyline Dr H NY 10532, US	
			David L. Paterson	Kevin Howells	Pascale Cossart	Fred C	hang	Interests	Active Learning	
Jue He	Nancy Y. Lp	James F. Fries	Faterson	TIOWOID	Shoichiro Tsukita	Yang Shi			Social Network Analysis	
									Large-scale online social ne	
			Diseases		Genetics & Genealog		DOW.			more interes
	Donald L Price	Shizuo Akira	John G. Bartlett	Weng-Keen Wong	Loralie D. Ma	Mark J.	Katsushi Tokunega		Search	
Kevin Chen								C Vinners Ob	B. B. & B. attace	
	Weng Cho Chew	Hong Lu	Warren Strober	David A. Hafler	Eric S. Lander	Kenneth K. Kidd	Matthew P. Scott	rigong si	i's Publications 1 Order by : relevant	- 5 of 231 publicat
commend c	ollaborators for Jimeng Sun from "Cell Biology" Yigong Shi Professor, School of Life Science, Tsinghua University					â 	JConf: Na Published	tefan J. Riedl, Yigon ture Reviews Molecu	lar Cell Biology CitedBy 2	
( L	and the second se	38, #Papers: 231, #Citations: 5788 ogy, Molecular Biology, Genetics & Genealogy					>>	in apopto	phosphatidylsering tic cells through ph se SCRM-1	
	David J. Chen Professor, Rehabilit H-index: 27, #Papers Cell Biology, Molecu	: 182, #Citations: 2	424					Authors: Xiaochen Wang, Jin Wang, Keiko Gengyo-Ando, Lichuan Gu, Chun-Ling Sun, Chonglin Yang, Yong Shi, Tetsuo Kobayashi, Yigong Shi, Shohei Mitani, Xiao-Song Xie, Din Xue.		
	Pascale Cossar								ture Cell Biology Year:2007	CitedBy 3
	Huazhong Agriculture University							Multiple Apoptotic Caspase Cascades		
250	H-index: 41, #Papers: 208, #Citations: 5794 Molecular Biology, Microbiology, Cell Biology							Drosophil	in Nonapoptotic Ro la Spermatid Individ un R. Huh, Stephanie	lualization
6	Professor, Department of Cell Biology, Kyoto University							Hong Yu, Nieng Yan, Yigong Shi, Ming Guo, Bruce A. Hay. JConf: Plos Biology		
A	H-index: 49, #Papers: 174, #Citations: 7467 Cell Biology, Biochemistry, Molecular Biology							Published Year:2004 CitedBy 4 Transforming Growth Factor -Mediated		
0	Fred Chang Professor of Microbiology & Immunology M.D., Ph.D., University of California at San Francisco H-index: 22, #Papers: 115, #Citations: 1519 Cell Biology, Molecular Biology, Pharmacology				<b>1</b>	Transcriptional Repression of c-myc Is Dependent on Direct Binding of Smad3 to a Novel Repressive Smad Binding Element Authors: Joshua P. Frederick, Nicole T. Liberati, David S. Waddell, Yigong Shi, Xiao-Fan Wang.				

	SYMBOL	DESCRIPTION
	T	number of topics
	d	a collaborated document
S	$A_d$	a set of authors of document $d$
Š	$x_{di}$	the <i>i</i> th attribute (word) in document $d$
ō	$z_{di}$	the topic assigned to attribute $x_{di}$
Ţ.	$s_{di}$	if $x_{di}$ is a word from a single domain or a cross domain
a	$ heta_v$	multinomial distribution over topics specific to author $v$
otati	$\vartheta_{vv'}$	multinomial distribution over topics specific to author
Ž		pair $(v,v')$
	$\phi_z$	multinomial distribution over words specific to topic $z$
	lpha,eta	Dirichlet priors to multinomial distributions $\theta$ , $\theta'$ and $\phi$
	$\lambda$	parameter for sampling the binary variable s
	$\gamma, \gamma_t$	Beta parameters to generate $\lambda$

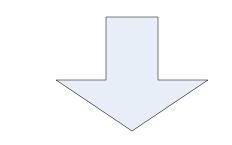
#### **Probabilistic generative process in CTL**

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

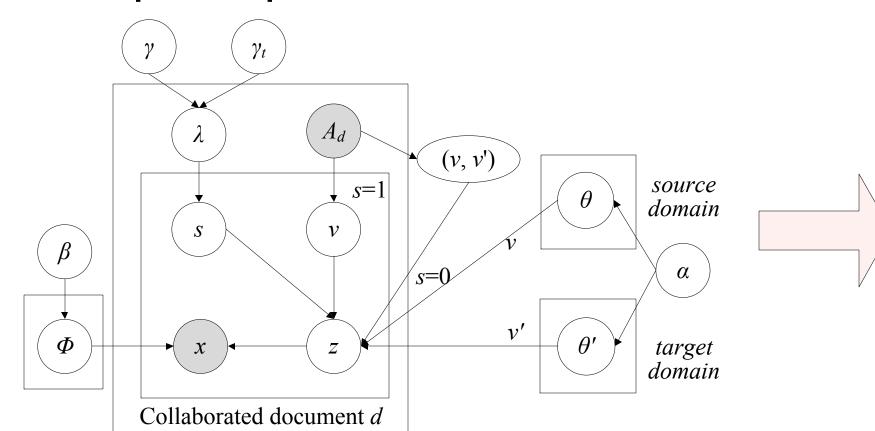
Draw a word  $x_{di} \sim multi(\phi_{z_{di}})$  from  $z_{di}$ -specific word distribution; end

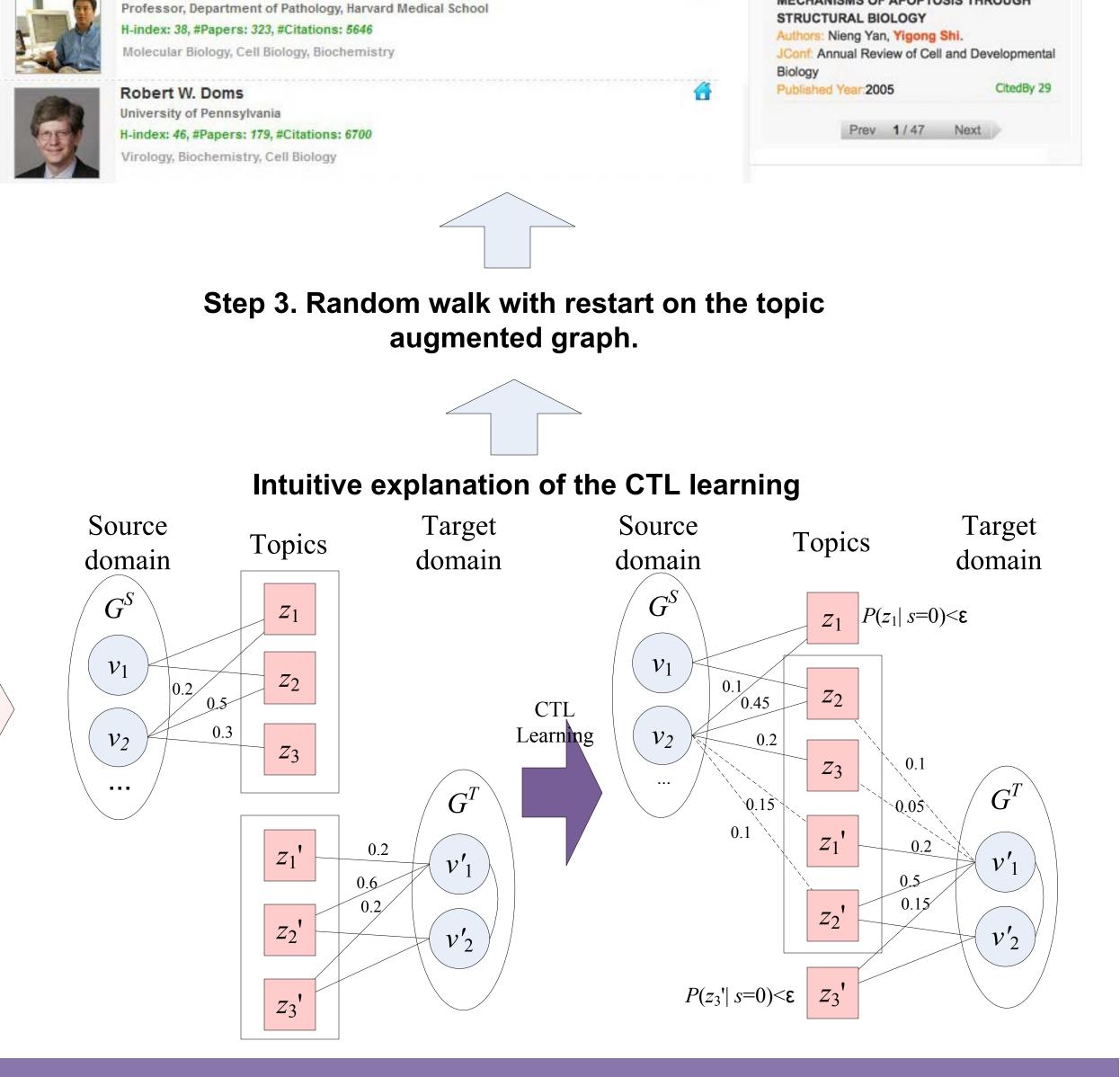


Step 1. Learning LDA or ACT model on the source and the target domain respectively.



Step 2. CTL Learning Graphical representation of CTL model.





#### **Empirical Analysis**

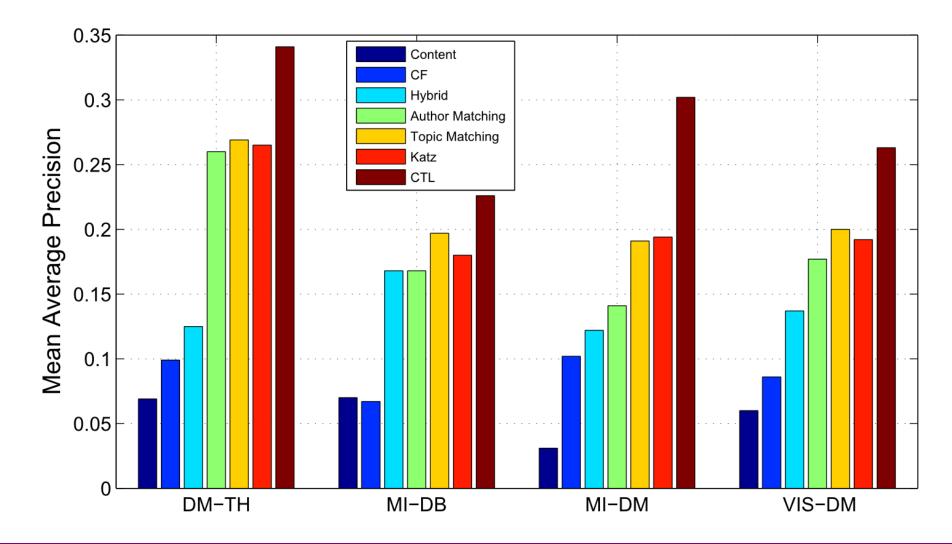
#### **Datasets (from Arnetminer): 5 domains**

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

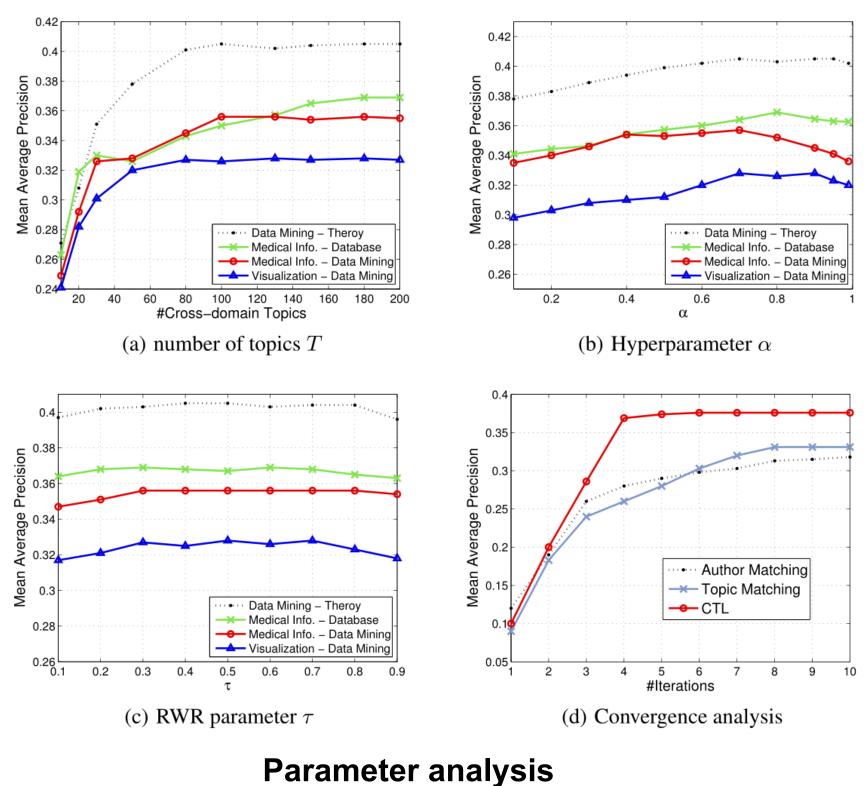
#### Performance on new collaboration prediction of all algorithms



Cross	ALG	P@10	P@20	MAP	R@100	ARHR	ARHR
domain				WIAI	K@100	-10	-20
	Content	10.3	10.2	10.9	31.4	4.9	2.1
Data	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
Mining (S)	Author	27.2	22.3	25.7	32.4	10.1	6.4
to Theory (T)	Topic	28.0	26.0	32.4	33.5	13.4	7.1
Theory (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
	Content	10.1	10.9	12.5	45.9	3.6	2.1
Medical	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
Info. (S) to	Author	26.2	29.6	32.2	54.8	10.5	5.4
Database (T)	Topic	29.4	26.3	34.7	59.3	11.5	5.2
Database (1)	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
	Content	5.8	5.7	9.5	19.8	1.9	0.9
Medical	CF	13.7	17.8	18.9	34.3	2.7	1.3
Info. (S)	Hybrid	18.0	19.0	19.8	36.7	3.4	1.3
to	Author	20.1	23.8	29.3	64.4	5.3	2.1
Data	Topic	26.0	25.0	33.9	48.1	10.7	5.6
Mining (T)	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
	Content	9.6	11.8	13.2	18.9	3.1	1.8
Visual. (S)	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
to Data	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
Mining (T)	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(%)** 

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.

Jie Tang et al. @KEG, Tsinghua; Jimeng Sun @IBM TJ Watson