Inside the Atoms: Ranking on a Network of Networks

Jingchao Ni[1], Hanghang Tong[2], Wei Fan[3], Xiang Zhang[1]

[1]Department of Electrical Engineering and Computer Science, Case Western Reserve University

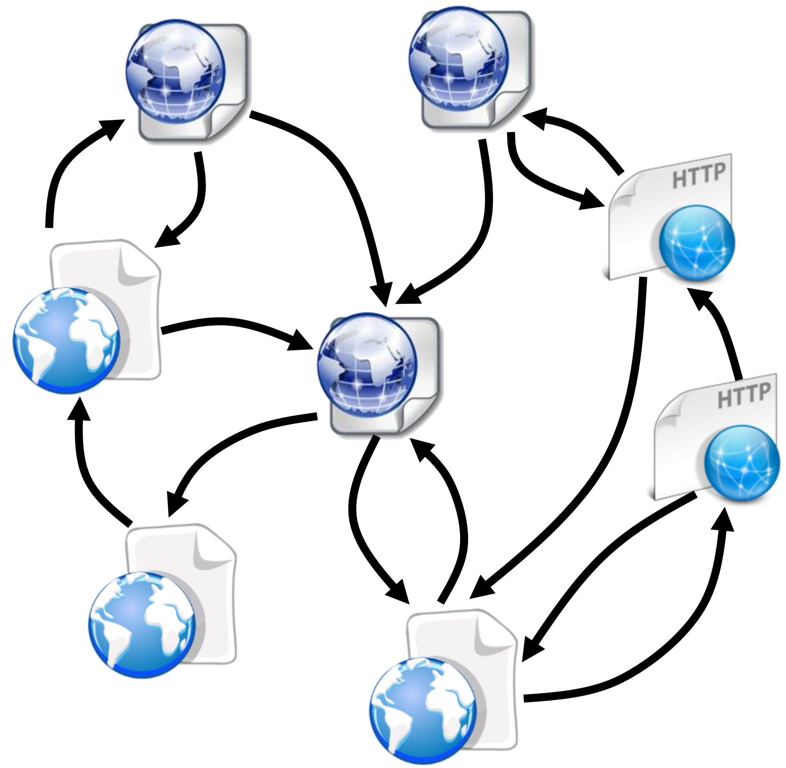
[2]School of Computing, Informatics, Decision Systems Engineering, Arizona State University

[3]Huawei Noahs Ark Lab

The 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining

Background: Ranking in a Network

- **Network: Data are naturally networks**
  - Webs are linked by hyperlink
  - Users are linked by friendship
  - Proteins are linked by interactions
- **Ranking without query**
  - Rank all nodes based on certain measures, e.g., Pagerank, HITS
  - Who are most popular users?
- **Ranking with query**
  - Find top-k most "similar" nodes for a query node based on certain measure, e.g., Personalized Pagerank, Simrank
  - Who are potential friends of Jon?



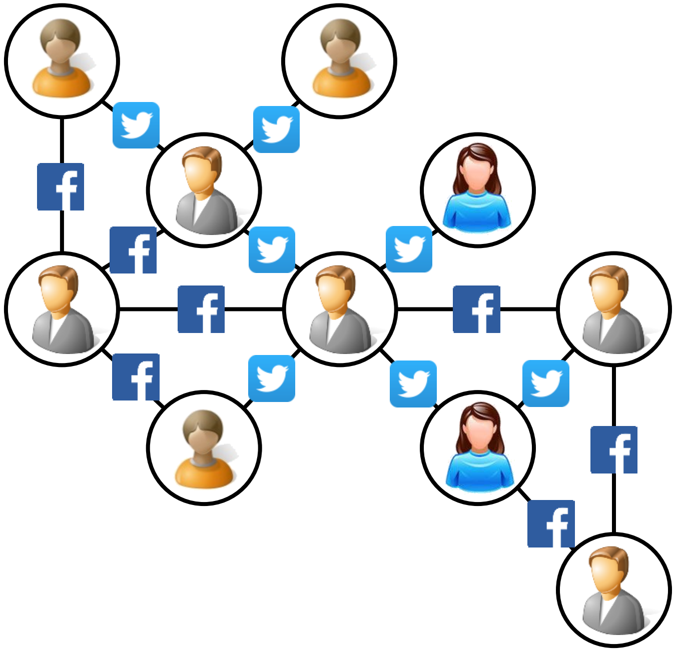
Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

Background: Ranking in a Network

- **Network: Data are naturally networks**
  - Webs are linked by hyperlink
  - Users are linked by friendship
  - Proteins are linked by interactions

- **Ranking without query**
  - Rank all nodes based on certain measures, e.g., Pagerank, HITS
  - Who are most popular users?

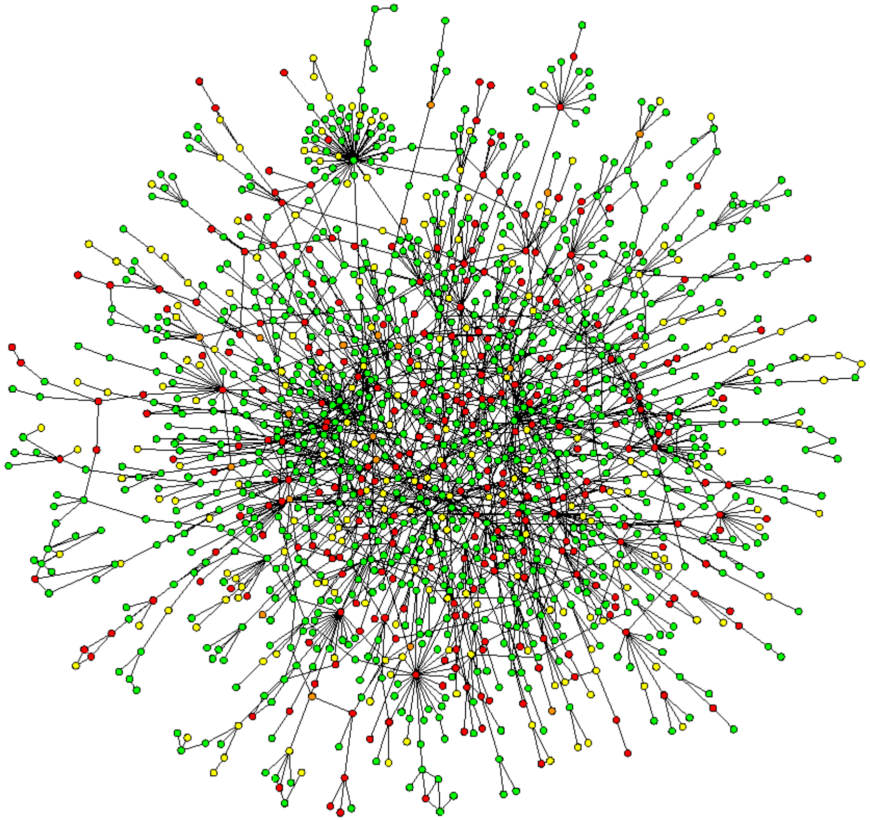
- **Ranking with query**
  - Find top-k most "similar" nodes for a query node based on certain measure, e.g., Personalized Pagerank, Simrank
  - Who are potential friends of Jon?



Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

Background: Ranking in a Network

- **Network: Data are naturally networks**
  - Webs are linked by hyperlink
  - Users are linked by friendship
  - Proteins are linked by interactions

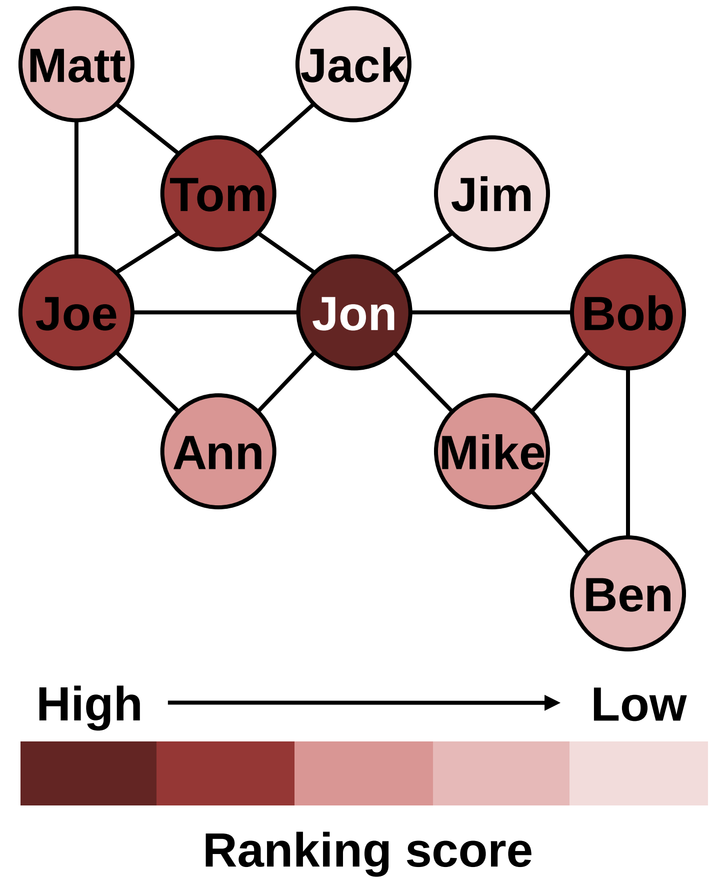


- **Ranking without query**
  - Rank all nodes based on certain measures, e.g., Pagerank, HITS
  - Who are most popular users?
- **Ranking with query**
  - Find top-k most "similar" nodes for a query node based on certain measure, e.g., Personalized Pagerank, Simrank
  - Who are potential friends of Jon?

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang. **Inside the Atoms: Ranking on a Network of Networks.** In KDD, 2014.

Background: Ranking in a Network

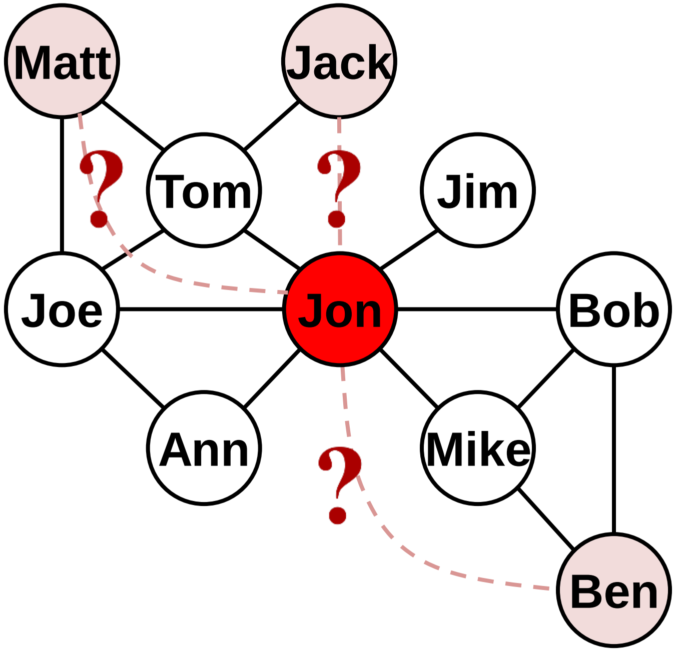
- ❑ **Network: Data are naturally networks**
  - Webs are linked by hyperlink
  - Users are linked by friendship
  - Proteins are linked by interactions
- ❑ **Ranking without query**
  - Rank all nodes based on certain measures, e.g., Pagerank, HITS
  - Who are most popular users?
- ❑ **Ranking with query**
  - Find top-k most “similar” nodes for a query node based on certain measure, e.g., Personalized Pagerank, Simrank
  - Who are potential friends of Jon?



Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

Background: Ranking in a Network

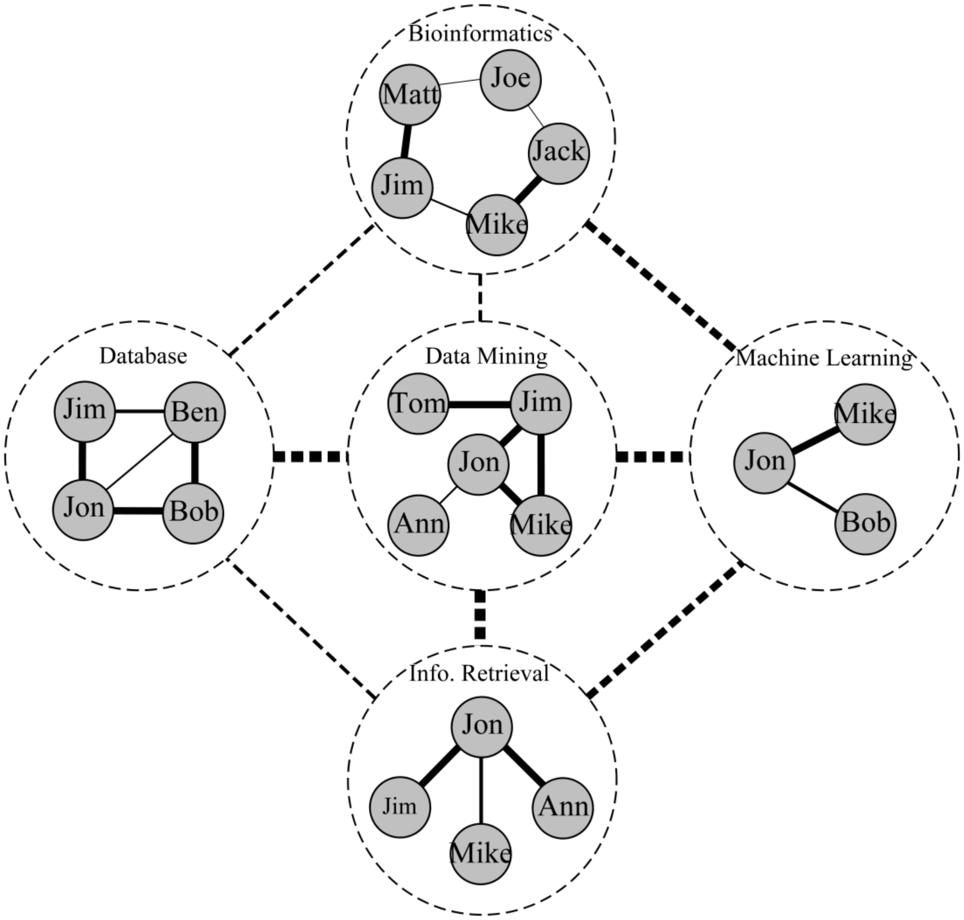
- ❑ **Network: Data are naturally networks**
  - ➢ Webs are linked by hyperlink
  - ➢ Users are linked by friendship
  - ➢ Proteins are linked by interactions
- ❑ **Ranking without query node**
  - ➢ Rank all nodes based on certain measures, e.g., Pagerank, HITS
  - ➢ Who are most popular users?
- ❑ **Ranking with query node**
  - ➢ Find top-k most "similar" nodes for a query node based on certain measure, e.g., Personalized Pagerank, Simrank
  - ➢ Who are potential friends of Jon?



Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang. **Inside the Atoms: Ranking on a Network of Networks.** In KDD, 2014.

Motivation: Network of Networks (NoN)

Research Area Network of Co-author Networks

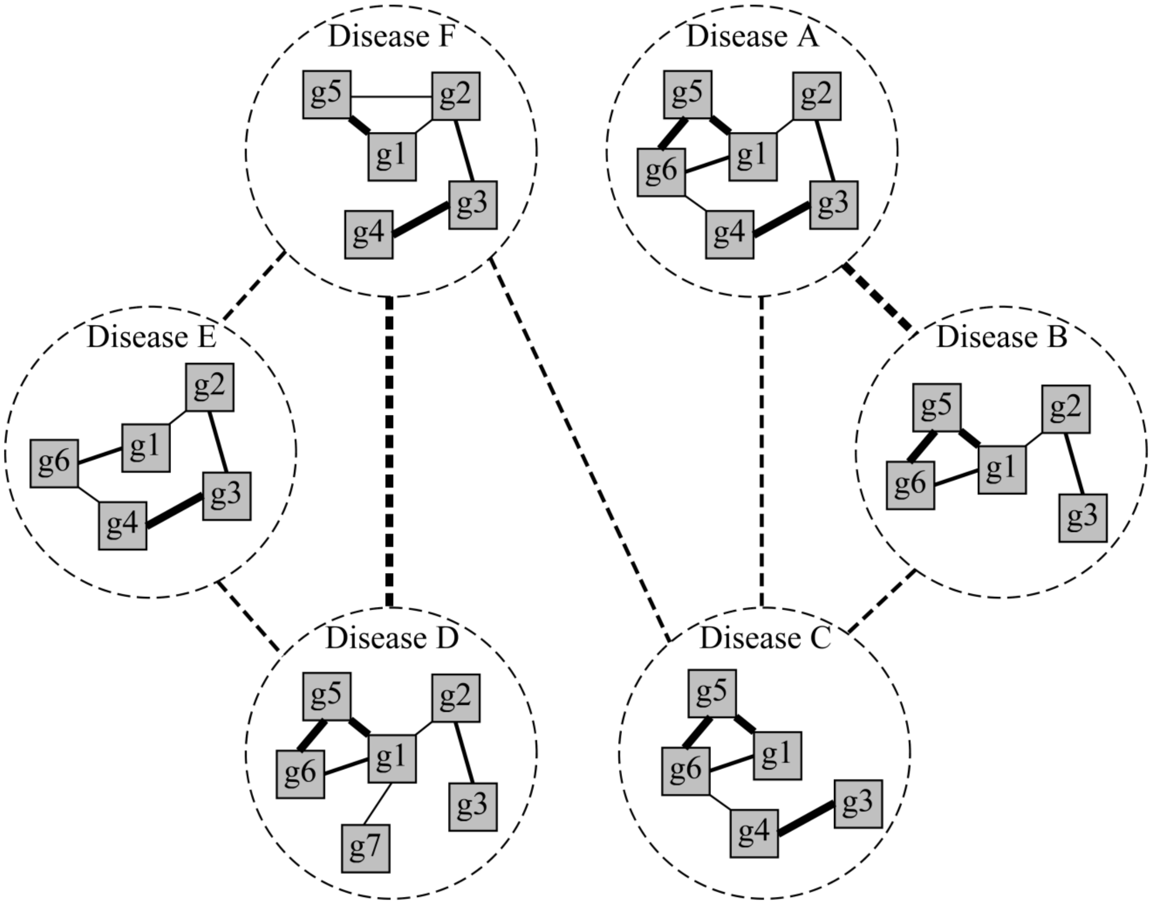


CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

Motivation: Network of Networks (NoN)

Disease Network of Protein Interaction Networks

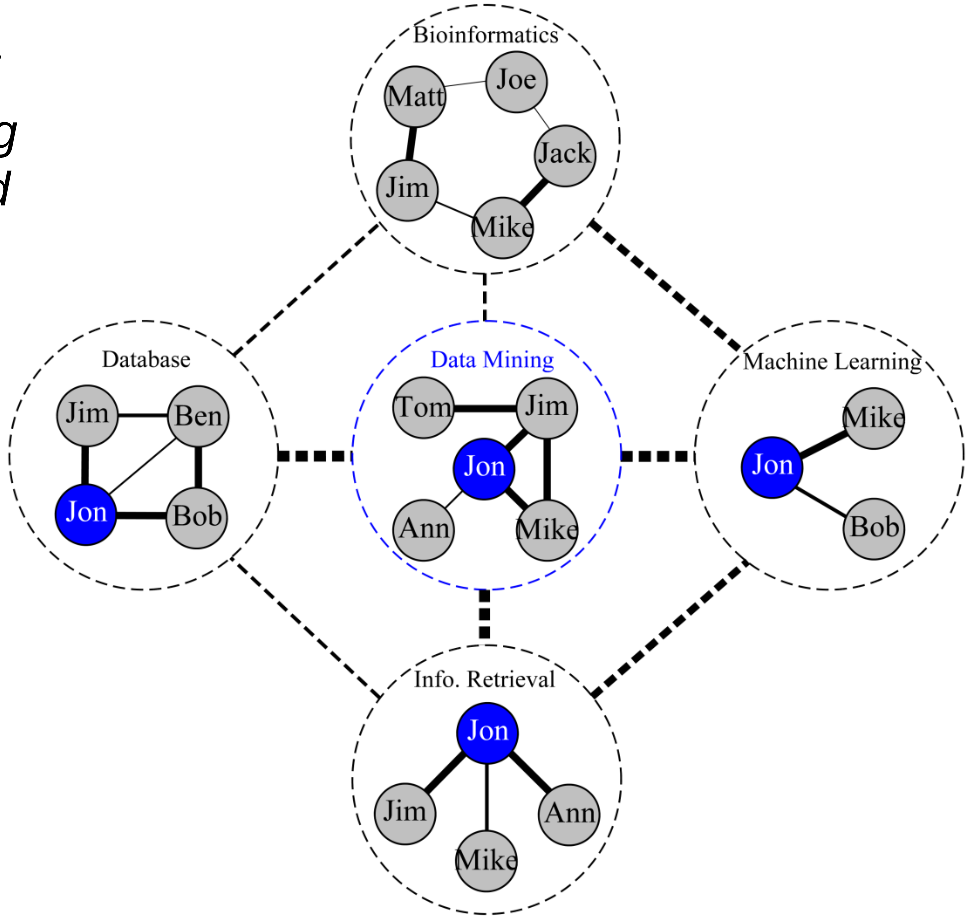


Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

Motivation: Ranking in NoN

Research Area Network of Co-author Networks

How to identify the importance of Jon in Data Mining by considering his overall contributions in related areas?

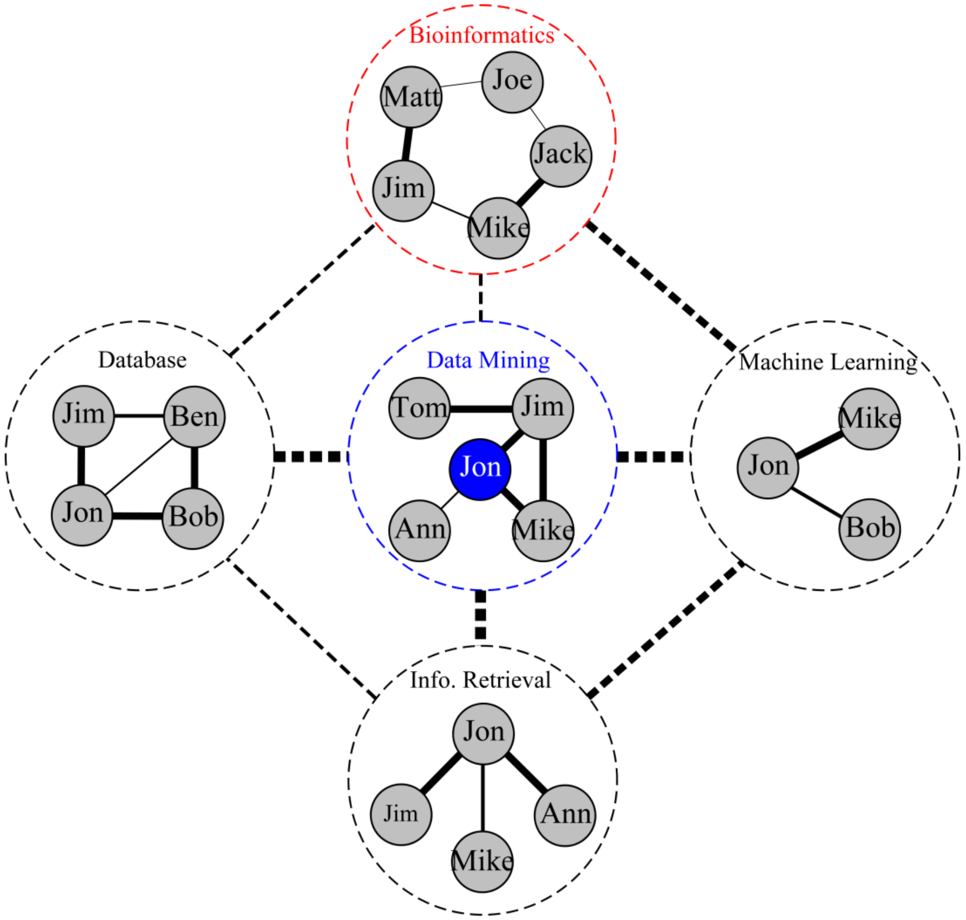


Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

Motivation: Query in NoN

Research Area Network of Co-author Networks

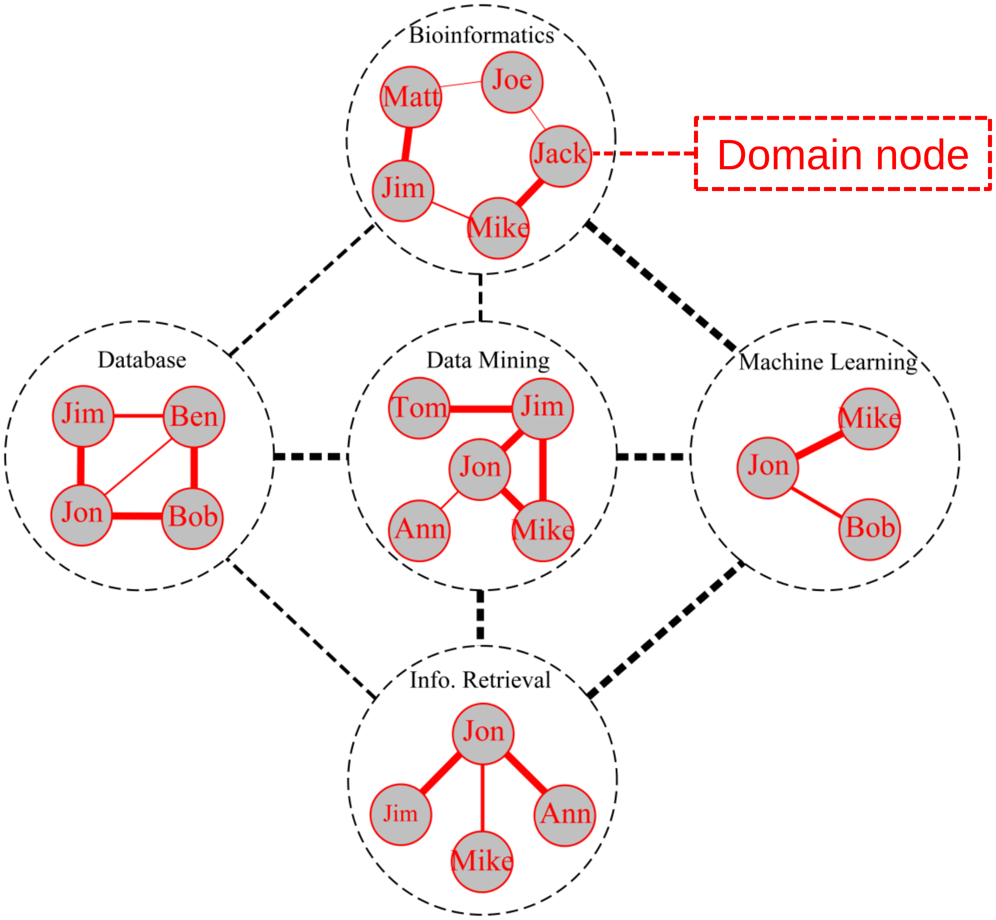
Which Bioinformatics researcher are most likely to collaborate with Data Mining researcher Jon?



Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

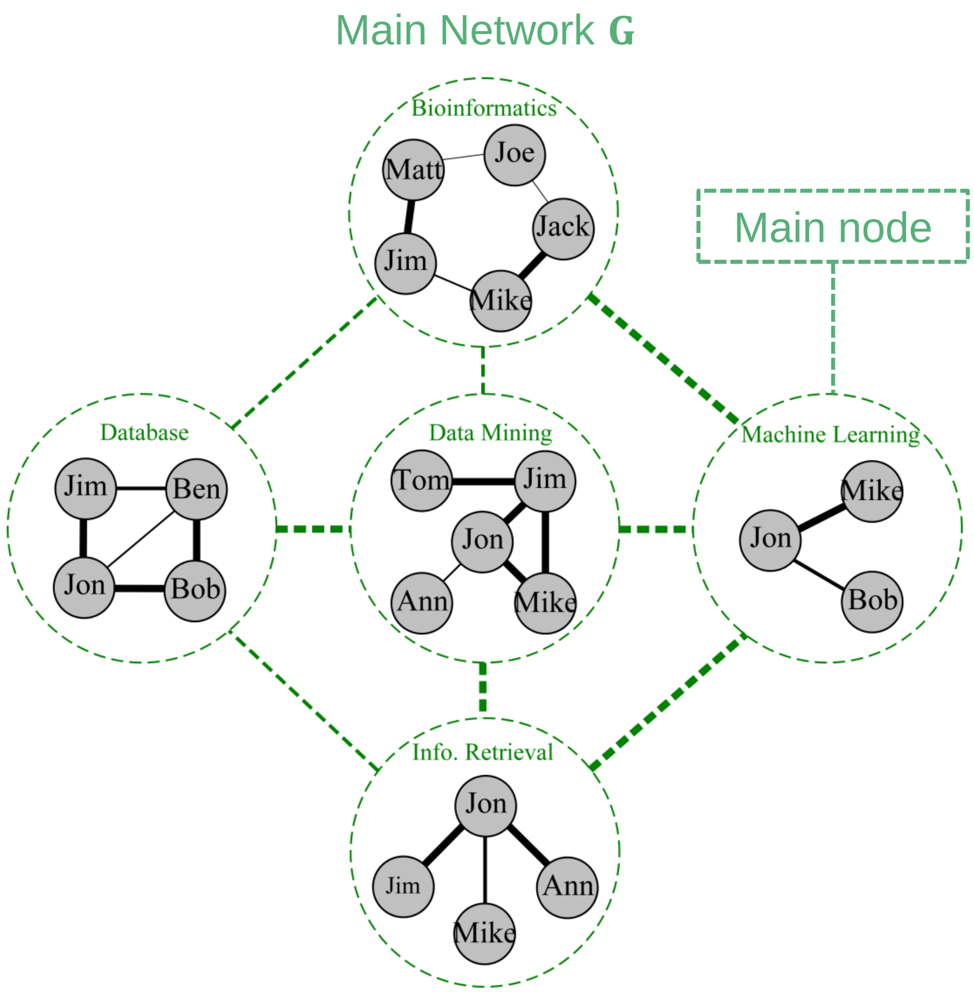
Problem Definitions

Domain-specific Network $\boldsymbol{A}_i$



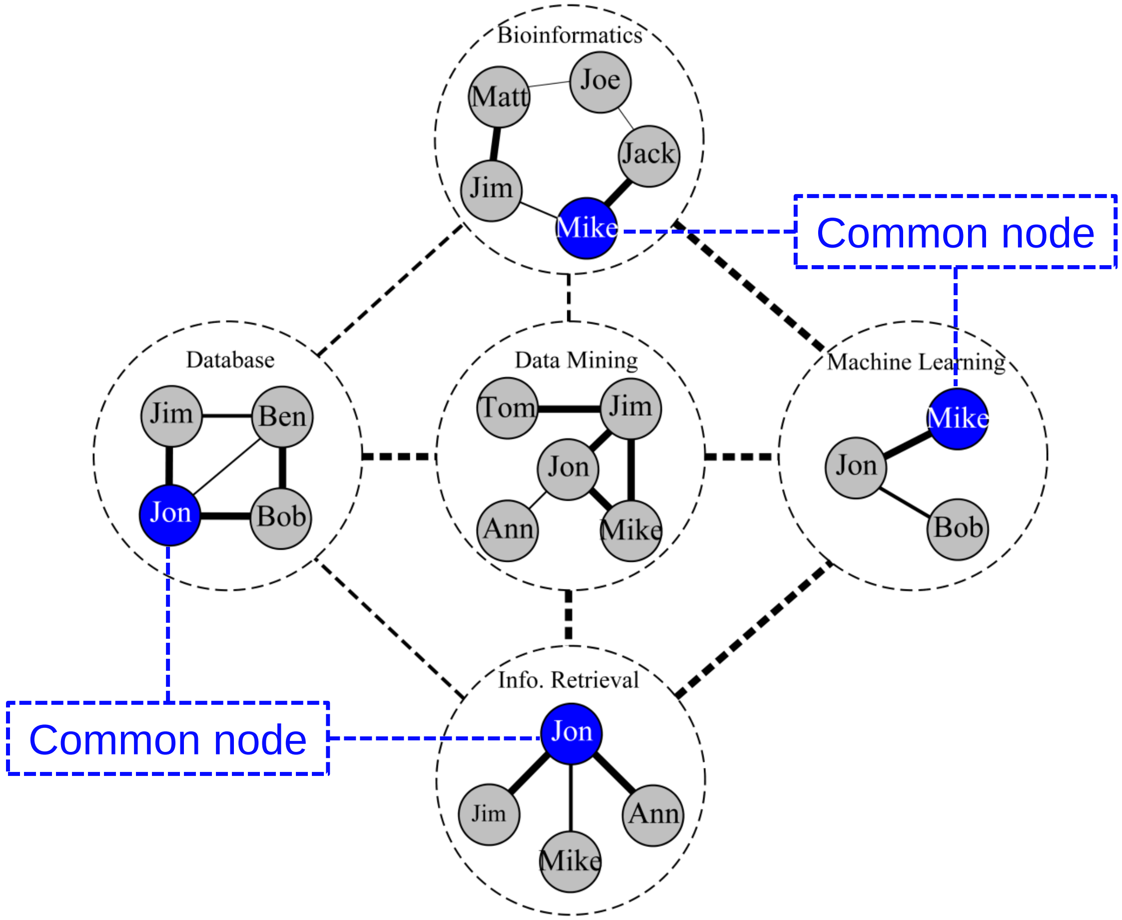
Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang. **Inside the Atoms: Ranking on a Network of Networks.** In KDD, 2014.

Problem Definitions



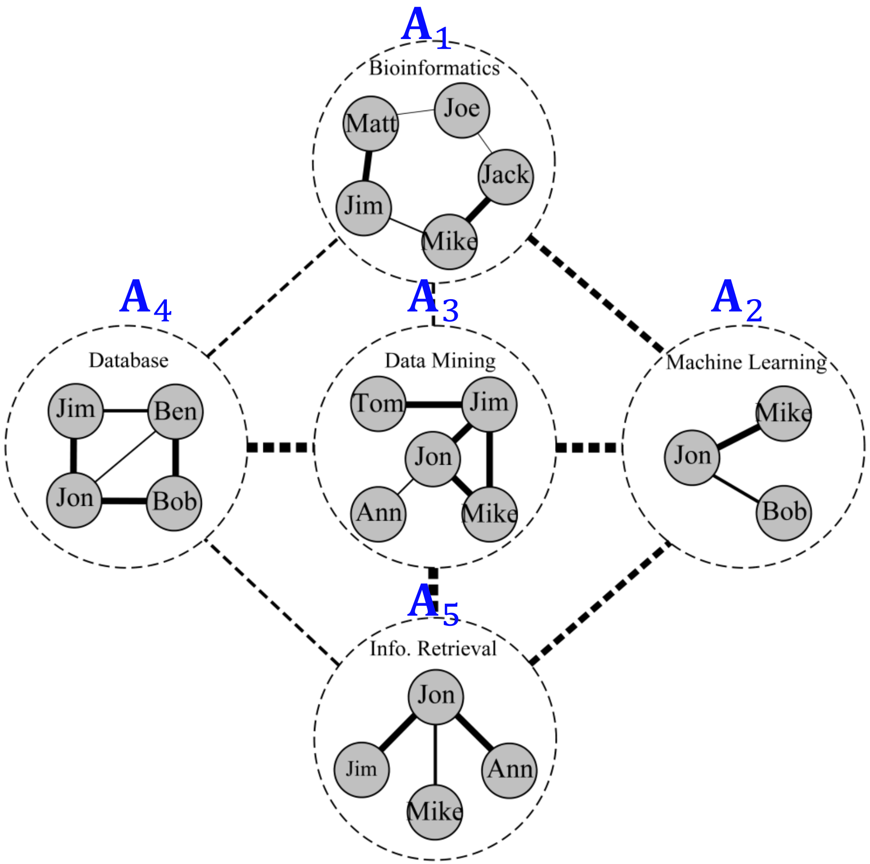
Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

Problem Definitions



Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

CrossRank



Problem 1: CrossRank

Given: (1) an NoN, and (2) the query vectors $\mathbf{e}_\mathrm{i}$ $(\mathrm{i} = 1, \ldots, \mathrm{g})$;

Find: ranking vectors $\mathbf{r}_\mathrm{i}$ for the nodes in the domain-specific networks $\mathbf{A}_\mathrm{i}$ $(\mathrm{i} = 1, \ldots, \mathrm{g})$.

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

CrossRank

Regularized Optimization Problem

$$J(\mathbf{r}_1, \dots, \mathbf{r}_g) = \underbrace{c \sum_{i=1}^{g} \mathbf{r}_i'(\mathbf{I}_{n_i} - \widetilde{\mathbf{A}}_i)\mathbf{r}_i}_{\text{within-network smoothness}} + \underbrace{(1-c)\sum_{i=1}^{g} \|\mathbf{r}_i - \mathbf{e}_i\|^2}_{\text{query preference}} + \underbrace{a \sum_{i,j=1}^{g} \left\| \frac{\mathbf{r}_i(\mathcal{I}_{ij})}{\sqrt{d_m(i)}} - \frac{\mathbf{r}_j(\mathcal{I}_{ij})}{\sqrt{d_m(j)}} \right\|^2 \mathbf{G}(i,j)}_{\text{cross-network consistency}}$$

- $\mathbf{r}_i$ is the ranking vector of the domain-specific network $\mathbf{A}_i$
- $d_m(i)$ is the degree of main node i in the main network $\mathbf{G}$
- $\widetilde{\mathbf{A}}_\mathbf{i}$ is the symmetric normalized adjacency matrix $\mathbf{A}_i$
- $\mathcal{I}_{ij}$ is the set of common nodes between $\mathbf{A}_i$ and $\mathbf{A}_j$

CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

CrossRank

Matrix Form of the Objective Function

$$J(\mathbf{r}) = c\mathbf{r}'\left(\mathbf{I_n} - \widetilde{\mathbf{A}}\right)\mathbf{r} + (1-c)\|\mathbf{r} - \mathbf{e}\|^2 + 2a\mathbf{r}'\mathbf{X}\mathbf{r}$$

$$\widetilde{\mathbf{A}} = \begin{bmatrix} \widetilde{\mathbf{A}}_\mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \widetilde{\mathbf{A}}_\mathbf{g} \end{bmatrix} \qquad \mathbf{r} = \begin{bmatrix} \mathbf{r_1} \\ \vdots \\ \mathbf{r_g} \end{bmatrix} \qquad \mathbf{e} = \begin{bmatrix} \mathbf{e_1} \\ \vdots \\ \mathbf{e_g} \end{bmatrix}$$

$\mathbf{X}$ encodes the cross-network consistency

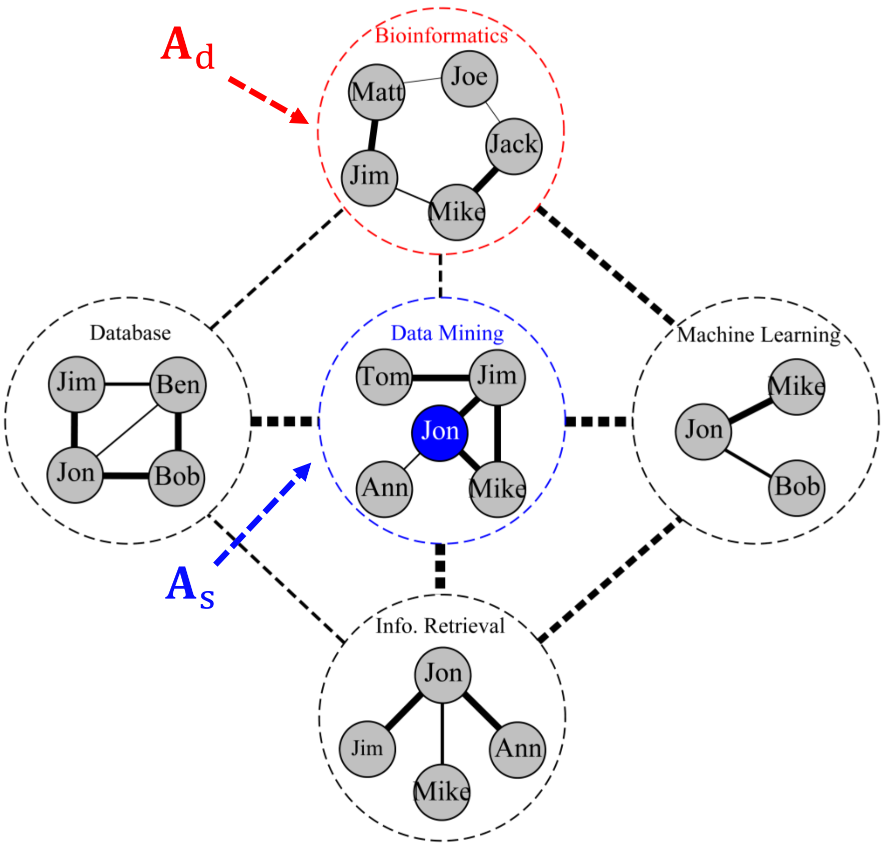
RWR-like Update rule

$$\mathbf{r} = \left(\underbrace{\frac{c}{1+2a}\widetilde{\mathbf{A}}}_{\text{within-network walk}} + \underbrace{\frac{2a}{1+2a}\widetilde{\mathbf{Y}}}_{\text{cross-network walk}}\right)\mathbf{r} + \frac{1-c}{1+2a}\mathbf{e}$$

CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang. **Inside the Atoms: Ranking on a Network of Networks.** In KDD, 2014.

CrossQuery



Problem 2: CrossQuery

Given: (1) an NoN, (2) a query node from a *source* domain-specific network $\mathbf{A}_\mathrm{s}$, (3) a *target* domain-specific network $\mathbf{A}_\mathrm{d}$, and (4) an integer k;

Find: the top-k most relevant nodes from the target domain-specific network $\mathbf{A}_\mathrm{d}$ w.r.t. the query node

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

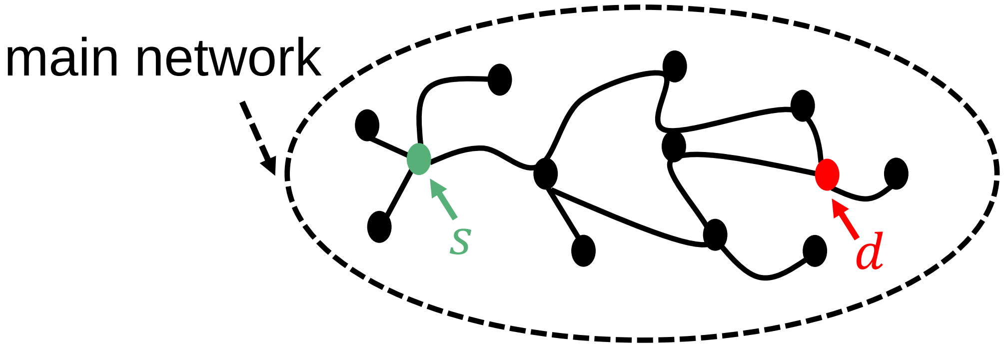
CrossQuery

CrossQuery-Basic

- Idea: our RWR-like update rule allows us to apply existing fast random walk with restart algorithm[1] where there is no accuracy loss. The candidate nodes can be restricted to those in the target domain-specific network.

CrossQuery-Fast

- Idea: given source and target domain-specific networks $\mathbf{A_s}$ and $\mathbf{A_d}$ of main nodes s and d respectively, prune less relevant main nodes[2]. Then apply CrossQuery-Basic on the pruned NoN.



1. Y. Fujiwara, et al., Efficient ad-hoc search for personalized pagerank. In SIGMOD, pages 445-456, 2013
2. Y. Koren et al., Measuring and extracting proximity in networks. In KDD, pages 245-255, 2006

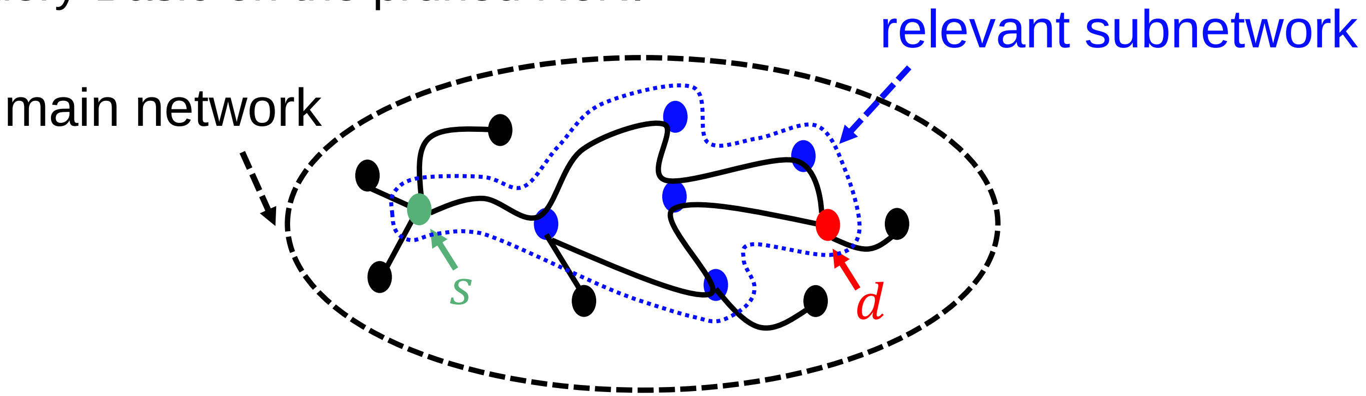
CrossQuery

CrossQuery-Basic

- Idea: our RWR-like update rule allows us to apply existing fast random walk with restart algorithm[1] where there is no accuracy loss. The candidate nodes can be restricted to those in the target domain-specific network.

CrossQuery-Fast

- Idea: given source and target domain-specific networks $\mathbf{A_s}$ and $\mathbf{A_d}$ of main nodes s and d respectively, prune less relevant main nodes[2]. Then apply CrossQuery-Basic on the pruned NoN.



1. Y. Fujiwara, et al., Efficient ad-hoc search for personalized pagerank. In SIGMOD, pages 445-456, 2013
2. Y. Koren et al., Measuring and extracting proximity in networks. In KDD, pages 245-255, 2006

CrossRank Effectiveness

Co-Author NoN

Areas in the main network

Area	Conference included
DM	KDD, ICDM, SDM, PKDD, PAKDD
ML	ICML, NIPS, AAAI, IJCAI, UAI, ECML
DB	VLDB, SIGMOD, ICDE, ICDT, EDBT, PODS
IR	SIGIR, WWW, ACL, ECIR, CIKM
BIO	ISMB, RECOMB, ECCB, BIBE, BIBM, WABI

CrossRank Effectiveness

Top ranked authors in the database area when varying a

Rank	$a = 0$	$a = 0.05$	$a = 0.1$	$a = 0.3$	$a = 0.5$
1	Divesh Srivastava	**Jiawei Han**	**Jiawei Han**	**Jiawei Han**	**Jiawei Han**
2	**Jiawei Han**	Divesh Srivastava	Divesh Srivastava	**Philip S. Yu**	**Philip S. Yu**
3	**Philip S. Yu**	**Philip S. Yu**	**Philip S. Yu**	Divesh Srivastava	**Christos Faloutsos**
4	Hector Garcia-Molina	Hector Garcia-Molina	Hector Garcia-Molina	**Christos Faloutsos**	Michael Stonebraker
5	Raghu Ramakrishnan	Raghu Ramakrishnan	**Christos Faloutsos**	Michael Stonebraker	Divesh Srivastava
6	Gerhard Weikum	Gerhard Weikum	Michael Stonebraker	Hector Garcia-Molina	Hector Garcia-Molina
7	Beng Chin Ooi	**Christos Faloutsos**	Raghu Ramakrishnan	Michael J. Carey	Michael J. Carey
8	H. V. Jagadish	Michael Stonebraker	Gerhard Weikum	Raghu Ramakrishnan	Gerhard Weikum
9	Michael J. Carey	Michael J. Carey	Michael J. Carey	Gerhard Weikum	Raghu Ramakrishnan
10	Michael Stonebraker	Beng Chin Ooi	Beng Chin Ooi	Elke A. Rundensteiner	Elke A. Rundensteiner

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

CrossRank Effectiveness

Co-Author NoN

Are

Area	Conference i
DM	KDD, ICD
ML	ICML, NIPS
DB	VLDB, SIGMO
IR	SIGIR, WWW, ACL,
BIO	ISMB, RECOMB, ECCB,

$$J(\mathbf{r}_1,\dots,\mathbf{r}_g) = c\sum_{i=1}^{g} \mathbf{r}_i'(\mathbf{I}_{n_i}-\widetilde{\mathbf{A}}_i)\mathbf{r}_i + (1-c)\sum_{i=1}^{g} \|\mathbf{r}_i-\mathbf{e}_i\|^2 + a\sum_{i,j=1}^{g} \left\| \frac{\mathbf{r}_i(I_{ij})}{\sqrt{d_m(i)}} - \frac{\mathbf{r}_j(I_{ij})}{\sqrt{d_m(j)}}\right\|^2 \mathbf{G}(i,j)$$

CrossRank Effectiveness

Top ranked authors in the database area when varying a

Rank	$a = 0$	$a = 0.05$	$a = 0.1$	$a = 0.3$	$a = 0.5$
1	Divesh Srivastava	**Jiawei Han**	**Jiawei Han**	**Jiawei Han**	**Jiawei Han**
2	**Jiawei Han**	Divesh Srivastava	Divesh Srivastava	**Philip S. Yu**	**Philip S. Yu**
3	**Philip S. Yu**	**Philip S. Yu**	**Philip S. Yu**	Divesh Srivastava	**Christos Faloutsos**
4	Hector Garcia-Molina	Hector Garcia-Molina	Hector Garcia-Molina	**Christos Faloutsos**	Michael Stonebraker
5	Raghu Ramakrishnan	Raghu Ramakrishnan	**Christos Faloutsos**	Michael Stonebraker	Divesh Srivastava
6	Gerhard Weikum	Gerhard Weikum	Michael Stonebraker	Hector Garcia-Molina	Hector Garcia-Molina
7	Beng Chin Ooi	**Christos Faloutsos**	Raghu Ramakrishnan	Michael J. Carey	Michael J. Carey
8	H. V. Jagadish	Michael Stonebraker	Gerhard Weikum	Raghu Ramakrishnan	Gerhard Weikum
9	Michael J. Carey	Michael J. Carey	Michael J. Carey	Gerhard Weikum	Raghu Ramakrishnan
10	Michael Stonebraker	Beng Chin Ooi	Beng Chin Ooi	Elke A. Rundensteiner	Elke A. Rundensteiner

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

CrossQuery Effectiveness

CrossQuery Effectiveness

Cross-area Co-authorship prediction results

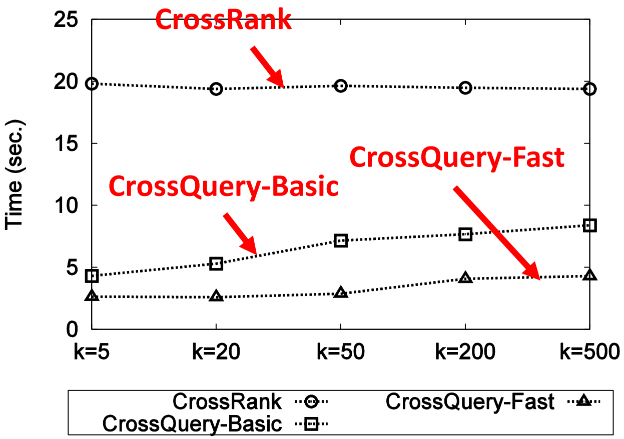
- *Which DB authors are most likely to collaborate with a DM author?*

#Papers	Hops	#Pairs	Methods	AUC	Accuracy
≥ 3	[3, 4]	45	PC	0.7196	0.4444
			Katz	0.7439	0.5556
			PropFlow	0.7558	0.6222
			PathSim	0.5636	0.2444
			PageRank	0.7417	0.5333
			CrossQuery	**0.7685**	**0.6444**
≥ 3	[3, 6]	70	PC	0.6009	0.3000
			Katz	0.6243	0.3714
			PropFlow	0.6268	0.4429
			PathSim	0.5278	0.2143
			PageRank	0.6378	0.3714
			CrossQuery	**0.6632**	**0.4571**
≥ 5	[3, 4]	23	PC	0.6521	0.2609
			Katz	0.6717	**0.3478**
			PropFlow	0.6850	**0.3478**
			PathSim	0.4279	0.1304
			PageRank	0.6743	**0.3478**
			CrossQuery	**0.7099**	**0.3478**
≥ 5	[3, 6]	38	PC	0.5692	0.2105
			Katz	0.5786	0.2368
			PropFlow	0.5950	**0.2895**
			PathSim	0.4362	0.1053
			PageRank	0.5880	0.2368
			CrossQuery	**0.6308**	**0.2895**

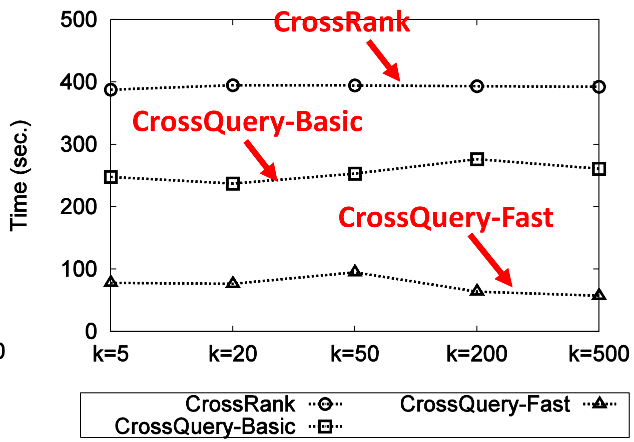
Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

CrossQuery Efficiency

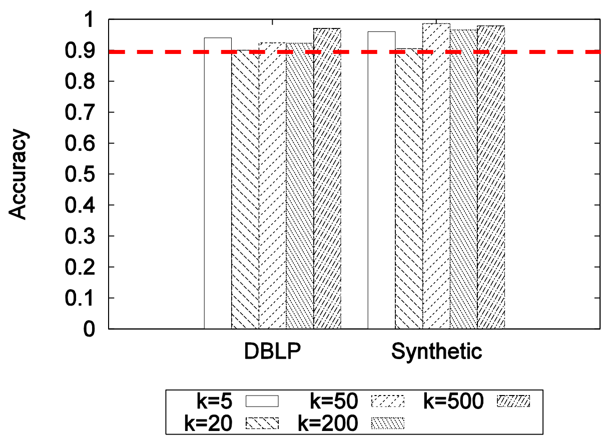
CrossQuery Efficiency



Query time on DBLP NoN



Query time on synthetic NoN



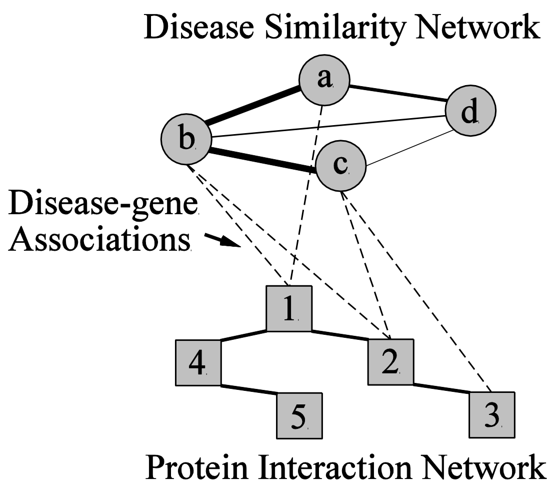
Accuracy of CrossQuery-Fast

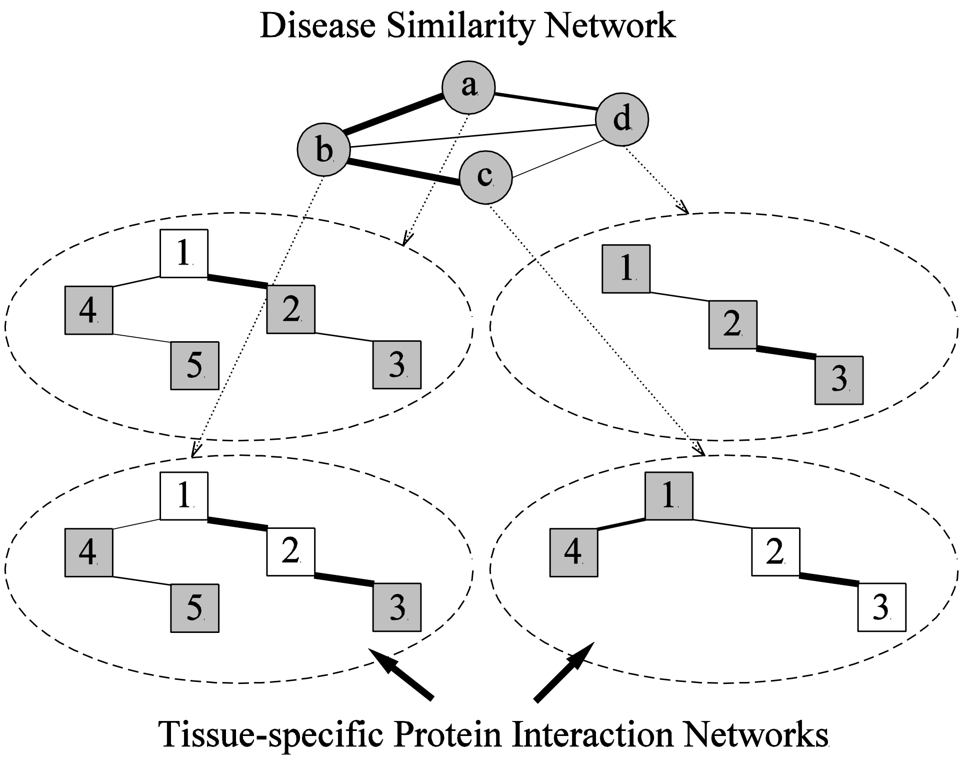
Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

Candidate Gene Prediction

Protein Interaction NoN

Heterogeneous network structure	Network of Networks structure

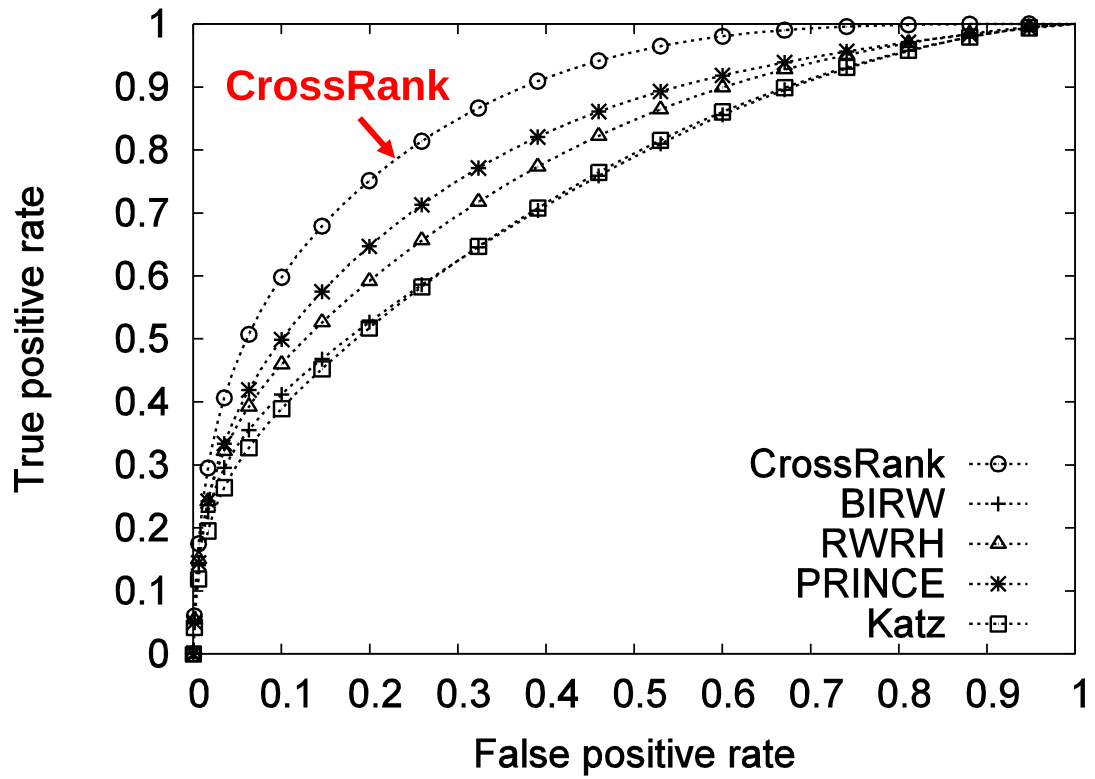




Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

Candidate Gene Prediction

Protein Interaction NoN



ROC curve comparison

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

Candidate Gene Prediction

Protein Interaction NoN

Ranking results comparison

Method	p-value
CrossRank vs. BIRW	1.82×10^{-11}
CrossRank vs. RWRH	2.04×10^{-11}
CrossRank vs. PRINCE	1.08×10^{-10}
CrossRank vs. Katz	2.32×10^{-12}

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Inside the Atoms: Ranking on a Network of Networks.
In KDD, 2014.

Conclusion

- **New Data Model:** Network of Networks (NoN)
- **New Ranking Algorithm:** CrossRank
- **Efficient Query Algorithm:** CrossQuery

Thank you!

Questions?