Inside the Atoms: Ranking on a Network of Networks

KDD2014
This year's special theme: Data Science for Social Good
20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining
August 24-27, 2014 New York City

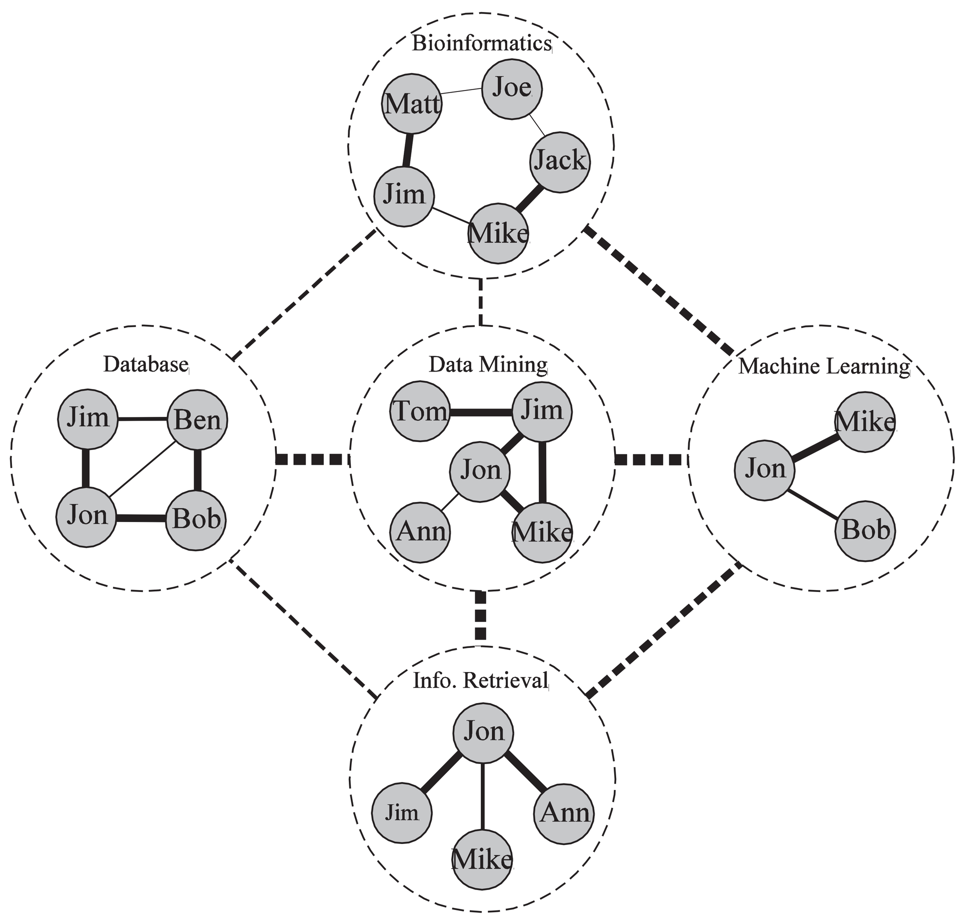
CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY

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

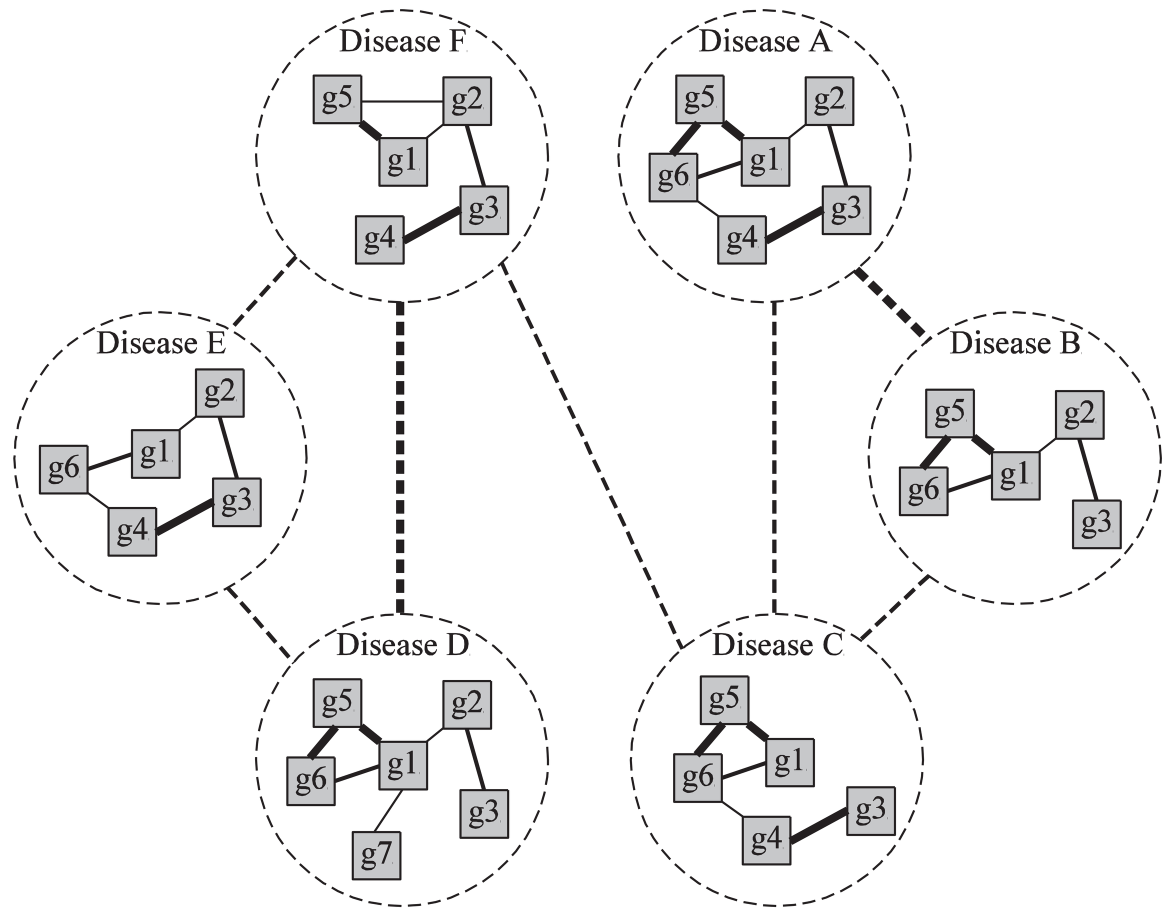
Network of Networks

Motivation: Network is an important and popular data model since real-world data are naturally networks, e.g., web network, social network, biological network, etc. However, networks are not independent. For example, a co-author network of data mining area is highly related to a co-author network of database area. In fact, networks themselves form a network. We do not want to ignore this high level network since it helps us learn more about the data. We call such structure as a **Network of Networks (NoN)**, such as research area network of co-author networks, disease similarity network of protein interaction networks, etc.

Research area network of co-author networks



Disease network of protein interaction networks



Examples of NoN. The main network is represented by dashed nodes and edges. The domain-specific networks are represnetd by solid nodes and edges.

Definition. **Network of Networks (NoN).** Given a $g \times g$ **main** network $\mathbf{G}$, a set of g **domain-specific** networks $\mathcal{A} = \{\mathbf{A}_1, \ldots, \mathbf{A}_g\}$ and a one-to-one mapping function θ, which maps each node in the main network $\mathbf{G}$ to a domain-specific network, a **Network of Networks (NoN)** is defined as the triplet $\mathcal{R} = <\mathbf{G}, \mathcal{A}, \theta>$. Nodes in the main network are referred to as *main nodes*, nodes in the domain-specific networks are called *domain nodes*. Each main node represents a domain-specific network through the mapping function θ. In addition, we represent the nodes in each domain-specific network as $\mathcal{V}_i$ $(i = 1, \ldots, g)$. We define $I_{i,j}$ as the *common nodes* between $\mathbf{A}_i$ and $\mathbf{A}_j$, i.e., $I_{i,j} = \mathcal{V}_i \cap \mathcal{V}_j$.

CrossRank

Given a network, ranking is an important task. People want to quickly identify important nodes (e.g., users, genes, etc.) from a network with thousands or millions of nodes.

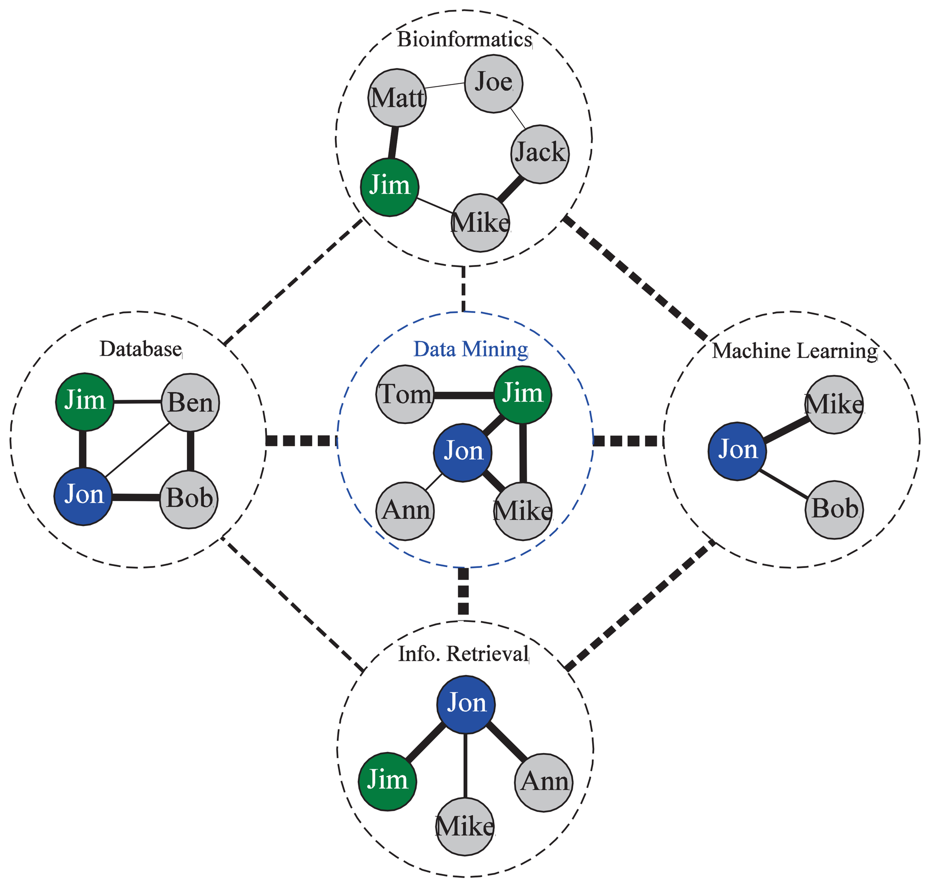
NoN allows us to rank nodes in broader view:

- Who is more important in data mining area? Jon or Jim? If we consider data mining only?
- Who is more important in data mining area? Jon or Jim? If we consider all highly related areas to data mining?

We propose a regularized optimization model to rank domain nodes w.r.t. the main network, i.e., minimizing:

$$J(\mathbf{r}_1, \ldots, \mathbf{r}_g) = \underbrace{c\sum_{i=1}^{g} \mathbf{r}_i'(\mathbf{I}_{n_i} - \tilde{\mathbf{A}}_i)\mathbf{r}_i}_{\text{within-network smoothness}} + \underbrace{(1-c)\sum_{i=1}^{g} \|\mathbf{r}_i - \mathbf{e}_i\|_F^2}_{\text{query preference}}$$

$$\underbrace{+ a\sum_{i,j=1}^{g} \left\| \frac{\mathbf{r}_i(I_{ij})}{\sqrt{d_m(i)}} - \frac{\mathbf{r}_j(I_{ij})}{\sqrt{d_m(j)}} \right\|_F^2 \mathbf{G}(i,j)}_{\text{cross-network consistency}}$$



Jon should be more important than Jim in data mining area since he is a popular researcher in highly related areas to data mining. His overall contribution to data mining is more significant than Jim.

Solution to CrossRank

Matrix form objective function

$$J(\mathbf{r}) = c\mathbf{r}'(\mathbf{I}_n - \tilde{\mathbf{A}})\mathbf{r} + (1-c)\|\mathbf{r} - \mathbf{e}\|_F^2 + 2a\mathbf{r}'\mathbf{X}\mathbf{r}$$

$$\tilde{\mathbf{A}} = \begin{bmatrix} \tilde{\mathbf{A}}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \tilde{\mathbf{A}}_g \end{bmatrix} \quad \mathbf{r} = \begin{bmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_g \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_g \end{bmatrix}$$

$\mathbf{X}$: A normalized Laplacian matrix of cross network links between common nodes.

RWR-like update rule

$$\mathbf{r} = \left(\frac{c}{1+2a}\tilde{\mathbf{A}} + \frac{2a}{1+2a}\tilde{\mathbf{Y}} \right)\mathbf{r} + \frac{1-c}{1+2a}\mathbf{e}$$

Property:

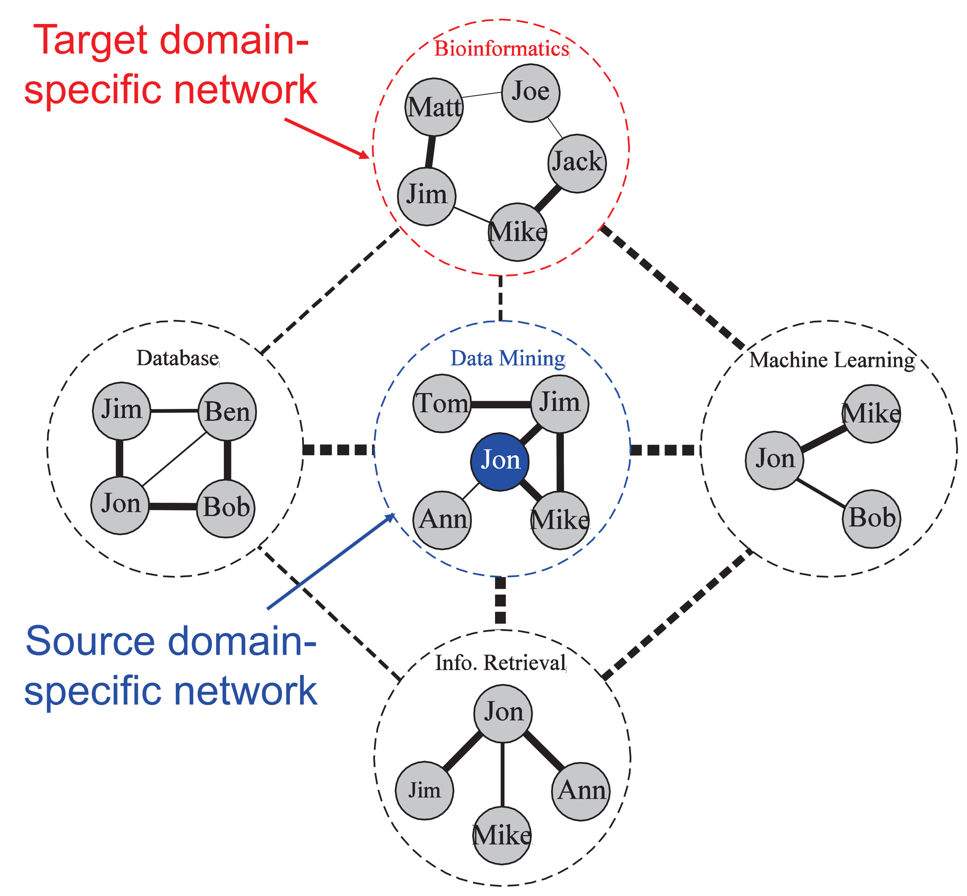
- Eigenvalues of the transition matrix are between −1 and 1
- It converges to the global minimum of the objective function

CrossQuery

Different people have different interests in nodes of a network. They may want to find top-k "similar" nodes in a network w.r.t. a query node. This is a ranking problem with query node.

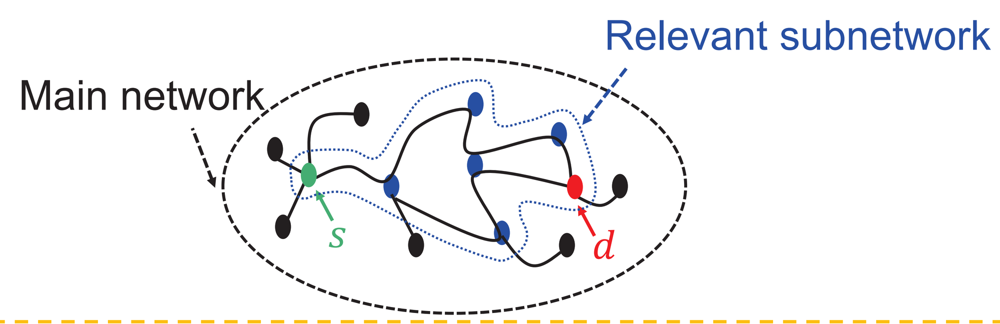
NoN allows us to query a node in a source domain-specific network and retrieve top-k "similar" nodes from a target domain-specific network:

- Which bioinformatics researchers will collaborate with the data mining researcher Jon?



CrossQuery-Basic: RWR-like update rule allows us to apply existing scalable algorithms for RWR.

CrossQuery-Fast: 1. Extract relevant subnetwork w.r.t. main nodes representing source and target domain-specific networks from the main network; 2. Prune NoN; 3. Apply CrossQuery-Basic on the pruned NoN.



Experiments

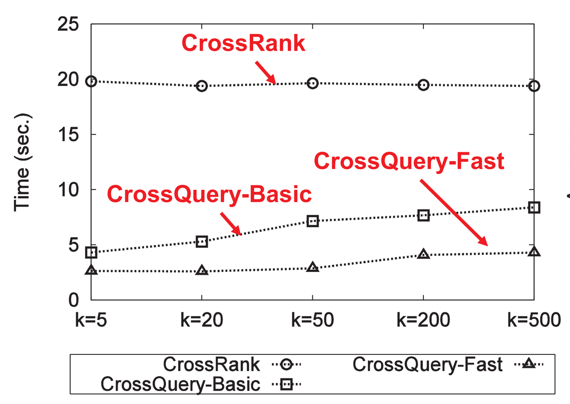
Co-author NoN

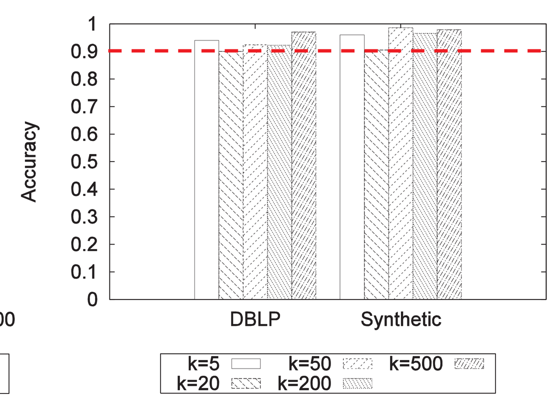
Top ranked authors in the database area when varying a

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

CrossQuery

- *Efficiency*
- *Accuracy*

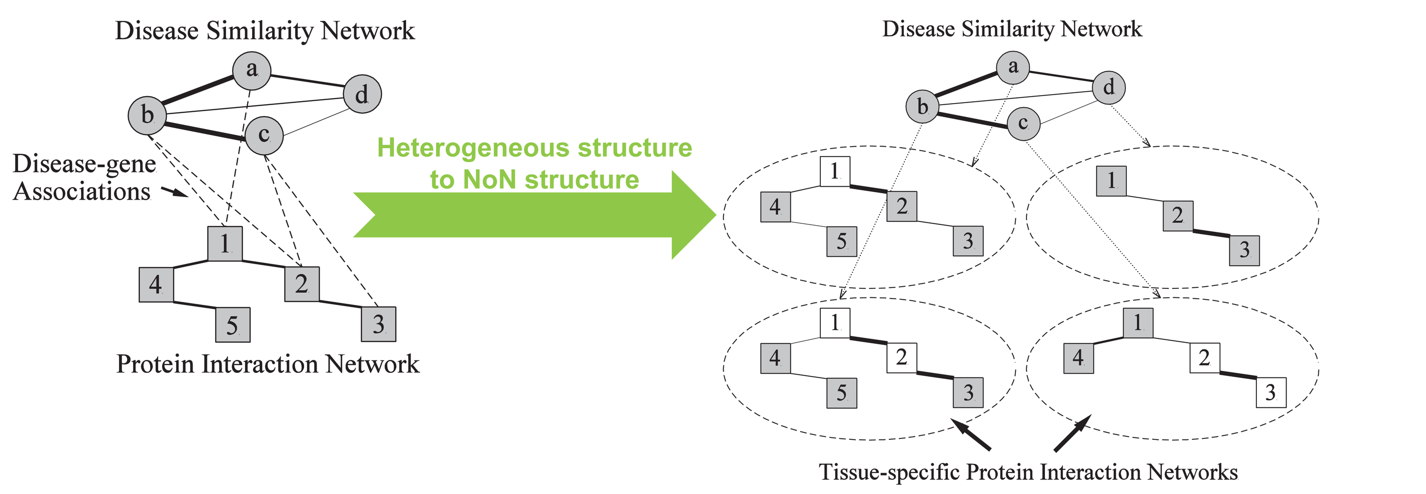




Cross-area co-authorship prediction results

#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**

Protein Interaction NoN



ROC curve comparison