# Probabilistic Community and Role Model for Social Networks

Yu Han[1] and Jie Tang[1,2,3]

[1]*Department of Computer Science and Technology, Tsinghua University*

[2]*Tsinghua National Laboratory for Information Science and Technology (TNList)*

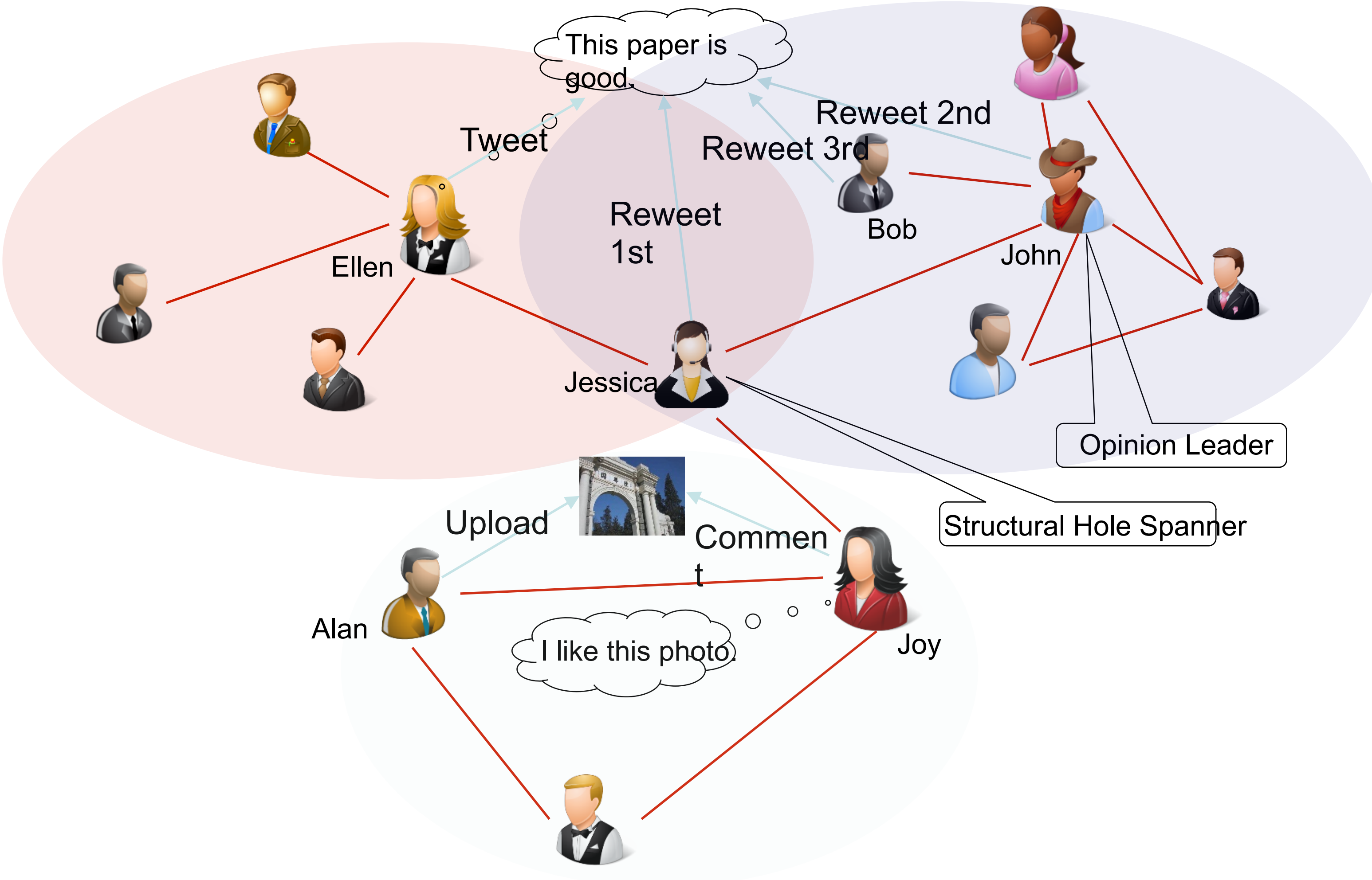
[3]*Jiangsu Collaborative Innovation Center for Language Ability, Jiangsu Normal University, China*

*yuhanthu@126.com, jietang@tsinghua.edu.cn*

# Social Networks

- There are visible and invisible elements in social networks
  - visible elements: *users*, *links*, *actions*
  - invisible elements: *communities*, *roles*
- Visible and invisible elements interact and affect each other
  - users may have closer relationships within a community than across communities
  - users' actions depend both on the attributes of themselves and on the influence of their communities
  - …

# Social Networks



# Problems:

- How should we model a complex social network so that the model can capture the intrinsic relations between all these elements, such as conformity influence, individual attributes, and actions?
- How do we use a social network model to handle issues such as community detection and behavior prediction without changing model itself?

# Limitations of existing work:

- Utilizing only portions of the available social network information.
- Focusing only on a few aspects of social networks, missing the global view.
- Basing on discriminative methods, ignoring the nature of social networks.
- Using deterministic method. Can not handle uncertain cases.

# Our goal:

To propose a unified probabilistic framework to model a social network, which can exactly reflect the intrinsic relationships between all visible and invisible elements of a social network, and can be used to handle practical issues in a social network.

# Intuitions and Assumptions

## Intuitions

☺ Links.
- ✓ Locally inhomogeneous.
- ✓ Each node may belong to several communities.

☺ Attributions.
- ✓ Each node has many attributes, such as in-degree, out-degree, etc.
- ✓ Based on these attributes, we can classify the nodes into clusters.
- ✓ Each cluster can be regarded as a role that nodes play.

☺ Actions.
- ✓ Whether a node takes a specific action partly depends on the community it belongs to.
- ✓ Whether a node takes an action may also depend on the role it plays.

## Assumptions

Assumption 1: Each node has a distribution over the communities.

Assumption 2: Each community has a distribution over the links.

Assumption 3: The attributes of each role satisfy a specific distribution—such as a Gaussian distribution.

Assumption 4: Each node has a distribution over roles according to its attributes.

Assumption 5: Community and role have a distribution over actions.

# CRM

### For each node $v$ in the graph:

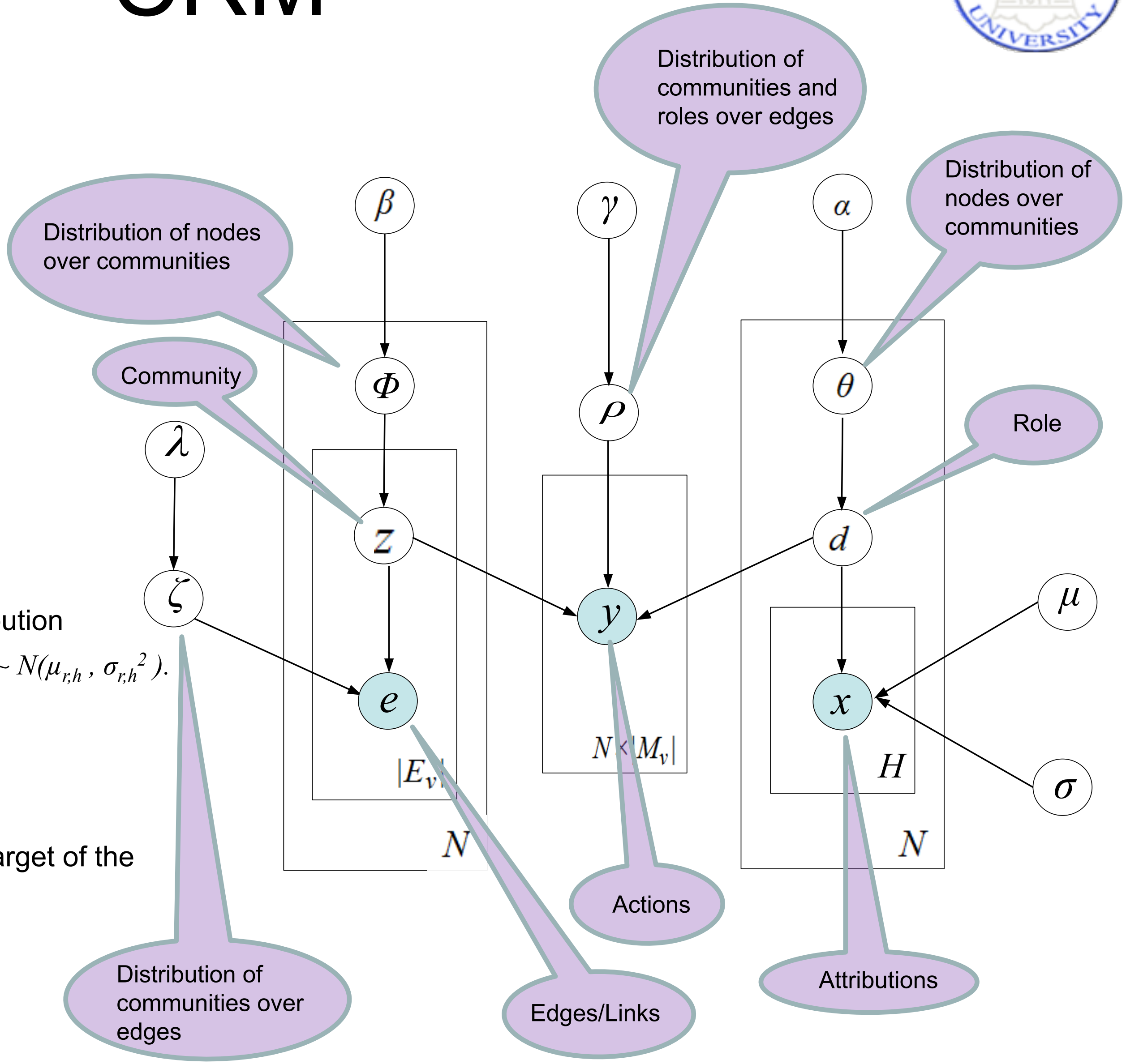
1. Draw $\zeta$ from $Dirichlet(\lambda)$;
2. Draw a $\varphi_v$ from $Dirichlet(\beta)$ prior;
3. For each edge $e_{v,i}$ :
   - Draw a community $z_{v,i} = c$ from multinomial distribution $\varphi_v$ ;
   - Draw an edge $e_{v,i}$ from a multinomial distribution $\zeta^{(c)}$ specific to community $c$.

### For each node $v$ in the graph:

1. Draw a $\theta_v$ from $Dirichlet(\alpha)$ prior;
2. Draw a role $d_v = r$ from multinomial distribution $\theta_v$ ;
3. For each attribute of $v$, draw a value $x_h^{(r)} \sim N(\mu_{r,h}, \sigma_{r,h}^2)$.

### For each action $y_m$ :

1. Draw $\rho$ from $Dirichlet(\gamma)$ prior;
2. Draw a community $c_v$ for $v$ from $\varphi_v$ ;
3. Draw a community $c_u$ for $u$, which is the target of the action, from $\varphi_u$ ;
4. Draw a role $r$ from $\theta_v$ ;
5. Draw $y_m \sim Multinomial(\rho^{\tau,r})$.



# Experiments

We first use a real dataset to learn the parameters of CRM. Then we use the parameters to generate a synthetic social network. Then we evaluate CRM by the following three tasks:

- **Structure recovery.**

  We compare the difference of structures between the generated synthetic network and the real network by means of six metrics: degree distribution, cluster coefficient, etc.

- **Behavior prediction.**

  CRM can predict users' actions by parameter $\rho$.

- **Community detection.**

  CRM can mine communities by parameter $\zeta$.

# Datasets

- **Coauthor**

  1,765 nodes, 13,415 links.
- **Facebook**

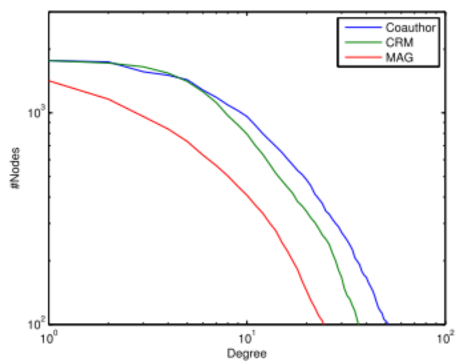
  4,039 nodes, 88,234 links.
- **Weibo**

  1,776,950 nodes, 308,489,739 links.

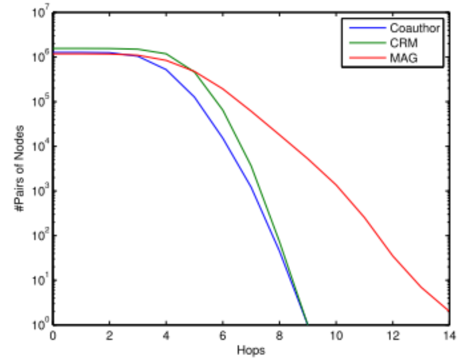
# Structural Recovery

- Baseline: MAG (UAI'11)
- Datasets:
  - Coauthor
  - Facebook
- Metrics
  - Degree is the degree of nodes versus the number of corresponding nodes.
  - Pairs of Nodes is the cumulative number of pairs of nodes that can be reached in ≤ h hops.
  - Eigenvalues are eigenvalues of the adjacency matrix representing the given network versus their rank.
  - Eigenvector is the components of the leading eigenvector versus the rank.
  - Clustering Coefficient is the average local clustering coefficient of nodes versus their degree.
  - Triangle Participation Ratio is the number of triangles that a node is adjacent to versus the number of nodes.

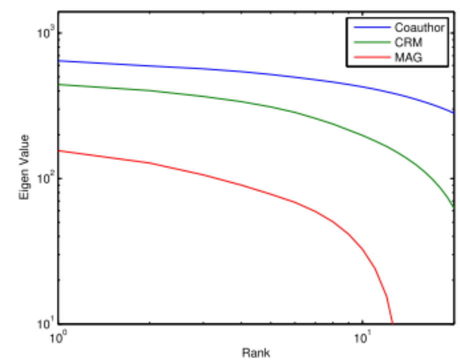
# Structural Recovery



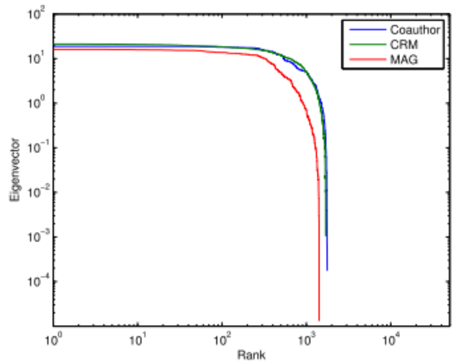

(a) Degree



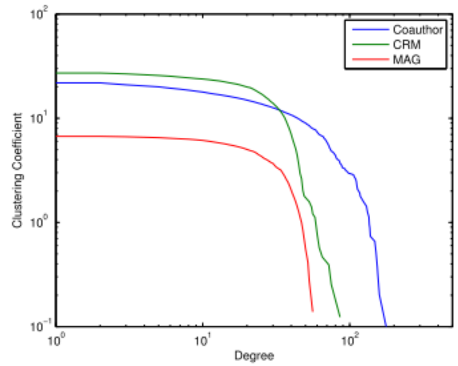

(b) Pairs of Nodes



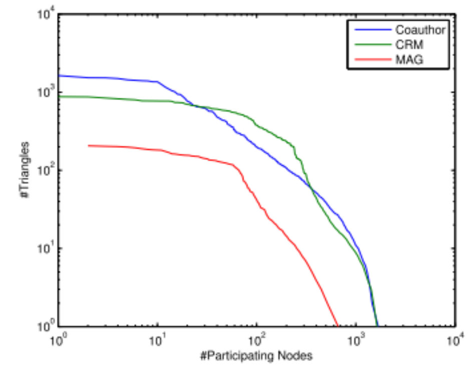

(c) Eigenvalues




(d) Eigenvector



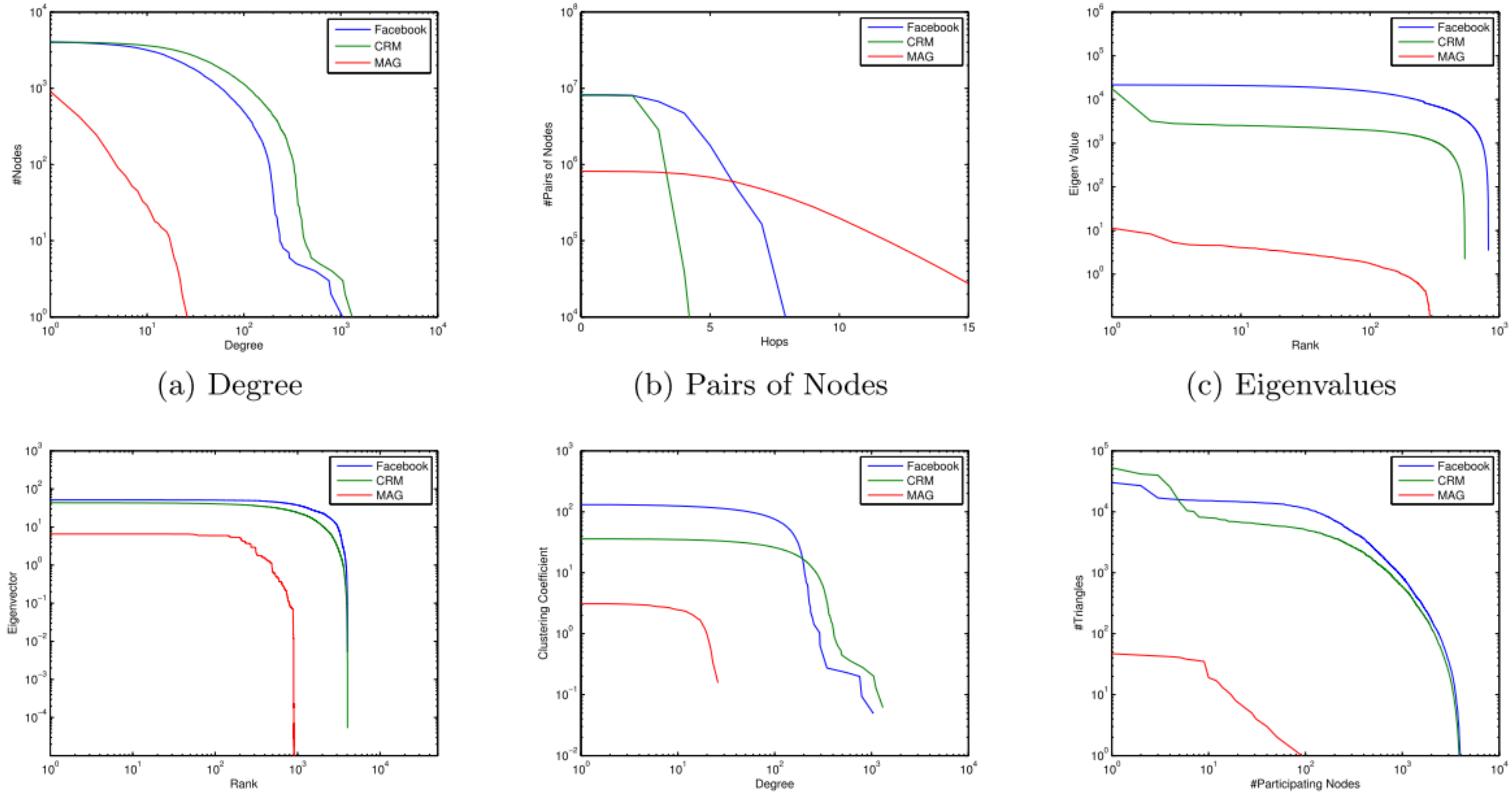

(e) Clustering Coefficient




(f) Triangle Participation Ratio

Metric values of the Coauthor network and the two networks generated by CRM and MAG. CRM outperforms MAG for every metric.

# Structural Recovery




(a) Degree (b) Pairs of Nodes (c) Eigenvalues

(d) Eigenvector (e) Clustering Coefficient (f) Triangle Participation Ratio

Metric values of the Facebook network and the two networks generated by CRM and MAG. CRM outperforms MAG for every metric.

# Behavior Prediction

- Baseline: SVM, SMO, LR, NB, RBF, C4.5
- Datasets:
  - Coauthor
  - Weibo
- Metrics: Precision, Recall, F1, AUC

| Date set | Method | Precision | Recall | F1-measure | AUC |
|---|---|---|---|---|---|
| Coauthor | SVM | **0.8838(0.1725)** | 0.5562(0.3183) | 0.6827(0.2054) | 0.7360(0.1111) |
| | SMO | 0.8647(0.1218) | 0.8142(0.1260) | 0.8387(0.1138) | 0.9218(0.0366) |
| | LR | 0.8668(0.1242) | 0.8292(0.1022) | 0.8476(0.1016) | 0.9642(0.0196) |
| | NB | 0.8183(0.1830) | 0.8115(0.1444) | 0.8149(0.1549) | 0.9417(0.0335) |
| | RBF | 0.8552(0.1058) | 0.8353(0.1165) | 0.8451(0.1081) | 0.9477(0.0271) |
| | C4.5 | 0.8328(0.0518) | 0.8015(0.1286) | 0.8169(0.1478) | 0.9065(0.1165) |
| | CRM | 0.8562(0.1490) | **0.8630(0.0598)** | **0.8596(0.1013)** | **0.9800(0.0199)** |
| Weibo | SVM | 0.5067(0.1405) | 0.5027(0.1185) | 0.5047(0.1150) | 0.6068(0.1113) |
| | SMO | 0.5074(0.1464) | 0.5209(0.1099) | 0.5141(0.1271) | 0.6145(0.0363) |
| | LR | 0.5199(0.1306) | 0.5469(0.1073) | 0.5331(0.1157) | 0.6330(0.0377) |
| | NB | 0.5112(0.1245) | 0.5692(0.1083) | 0.5386(0.1172) | 0.6397(0.0394) |
| | RBF | 0.5225(0.1361) | 0.4679(0.1117) | 0.4937(0.1217) | 0.5945(0.0085) |
| | C4.5 | 0.5237(0.1367) | 0.5322(0.1114) | 0.5279(0.1211) | 0.6271(0.1083) |
| | CRM | **0.7017(0.1300)** | **0.7305(0.1079)** | **0.7158(0.1149)** | **0.8174(0.0233)** |

# Community Detection

- Datasets:
  - Coauthor
- Result:

| Comm. | Name | Affiliation |
|---|---|---|
| 1 | Jiawei Han | UIUC |
| | Jian Pei | SFU |
| | Philip S. Yu | UIC |
| | Hong Cheng | CUHK |
| | Wei Wang | UNC |
| 2 | Thomas S. Huang | UIUC |
| | Yun Raymond Fu | UB |
| | Shuicheng Yan | NUS |
| | Mark A. Hasegawa-Johnson | UIUC |
| | Xiaoou Tang | CUHK |
| 3 | Philip A. Bernstein | Microsoft |
| | Nathan Andrew Goodman | UA |
| | David Dewitt | UW-Madison |
| | Erhard Rahm | U. of Leipzig |
| | Michael Stonebraker | MIT |

# Future Work

- Mining more factors
- Integrating nonparametric methords

# Thank you!
# Q&A