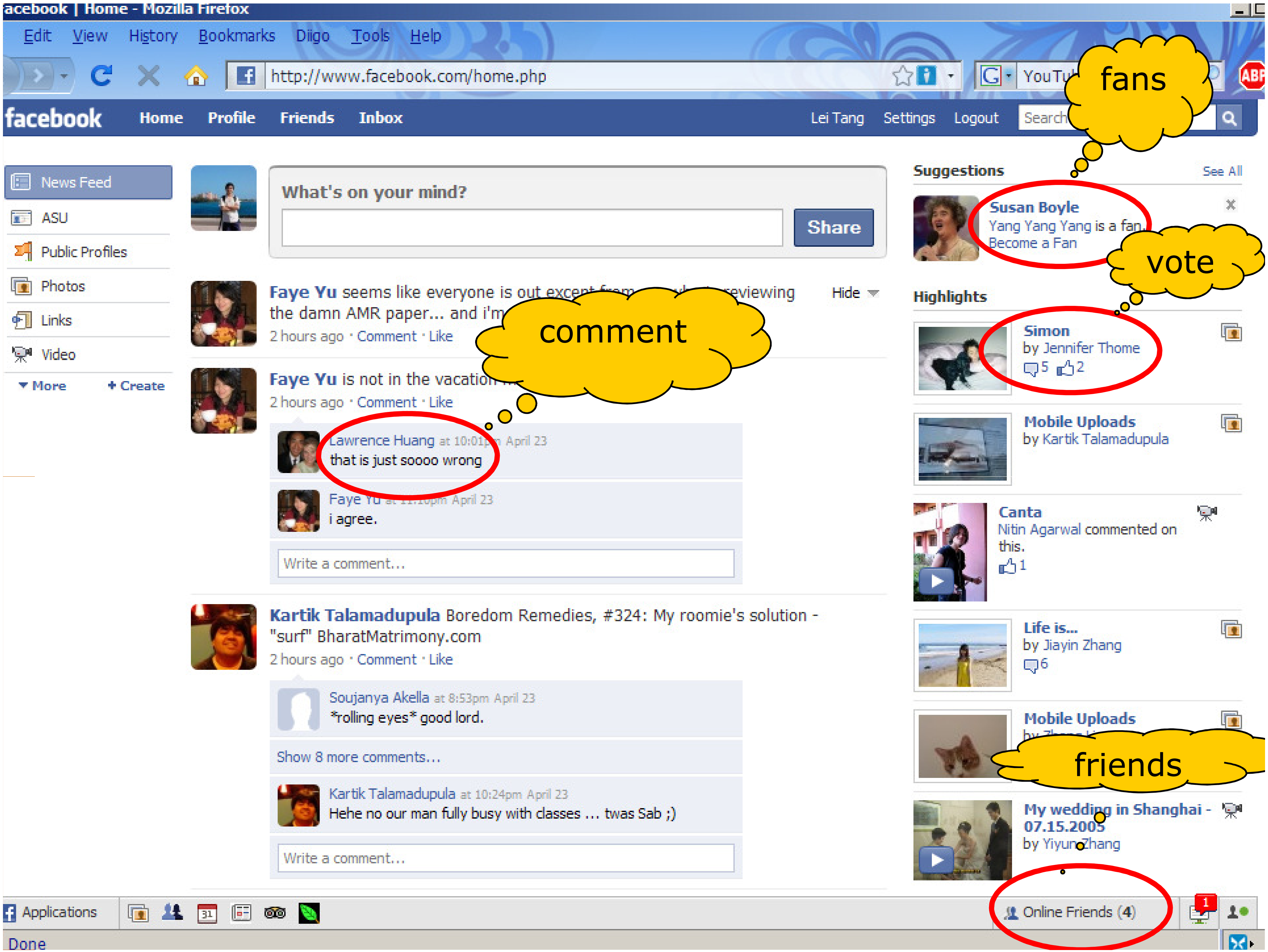
Uncovering Groups via Heterogeneous Interaction Analysis

Lei Tang, Xufei Wang and Huan Liu

Arizona State University

ICDM, Dec. 9th, 2009

YouTube - azcamel007's YouTube - Mozilla Firefox
File Edit View History Bookmarks Diigo Tools Help
http://www.youtube.com/
YouTube
Broadcast Yourself ™
Worldwide | English
(0) azcamel007 | Account | QuickList (0) | Help | Sign Out
Home Subscriptions Videos Shows Channels
Search
Upload
Video File
Quick Capture
Add / Remove Modules
Friend Activity
edit
Your friends haven't done anything yet.
Once your friends start uploading, favoriting, and rating videos, we'll show it here on this page.
Subscriptions (view all)
Your subscriptions haven't added videos lately.
Recommended for You (view all)
Backstreet Boys - I want it that...
1 year ago
139,212 views
MinekeW
Back Dormitory Boys
2 years ago
111,987 views
SakuraCiel
two chinese boys:dadada
2 years ago
2,105,378 views
younotwolf
1 year ago
162,404 views
monkydance
Videos Being Watched Now (view all)
Inbox
0 Personal Messages
0 Shared with you
0 Comments
0 Video Responses
send message
What's New
Watch your favorite videos in HD!
100 Invites to RealTime for Blog Readers
Today is the beta launch of YouTube RealTime, a new way of discovering what your friends are doing on YouTube. Wanna try it? Be one of the first to comment on this post.
Read more in our Blog
comment
Video Response
Message
http://www.youtube.com/my_videos_upload



acebook | Home - Mozilla Firefox
Edit View History Bookmarks Diigo Tools Help
http://www.facebook.com/home.php
facebook
Home Profile Friends Inbox
Lei Tang Settings Logout
fans
News Feed
ASU
Public Profiles
Photos
Links
Video
More
+ Create
What's on your mind?
Share
Suggestions
See All
Susan Boyle
Yang Yang Yang is a fan.
Become a Fan
vote
Highlights
comment
Hide
2 hours ago · Comment · Like
Faye Yu is not in the vacation
2 hours ago · Comment · Like
that is just soooo wrong
i agree.
Write a comment...
Simon
by Jennifer Thome
5 2
Mobile Uploads
by Kartik Talamadupula
Canta
Nitin Agarwal commented on this.
1
Kartik Talamadupula Boredom Remedies, #324: My roomie's solution - "surf" BharatMatrimony.com
2 hours ago · Comment · Like
Soujanya Akella at 8:53pm April 23
rolling eyes good lord.
Show 8 more comments...
Kartik Talamadupula at 10:24pm April 23
Hehe no our man fully busy with classes ... twas Sab ;)
Write a comment...
Life is...
by Jiayin Zhang
6
Mobile Uploads
friends
My wedding in Shanghai - 07.15.2005
by Yiyun Zhang
Applications
Online Friends (4)
Done

Multi-Dimensional Network

- Different behaviors lead to heterogeneous Interactions

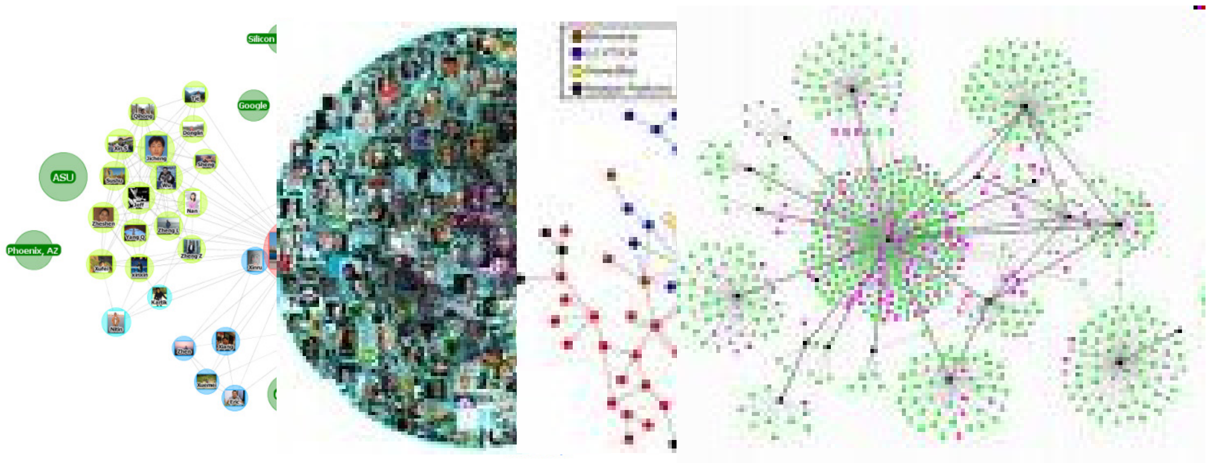
Contacts/friends

Tagging on Social Content

Fans/Subscriptions

Response to Social Content

…………………



Network of Multiple Dimensions

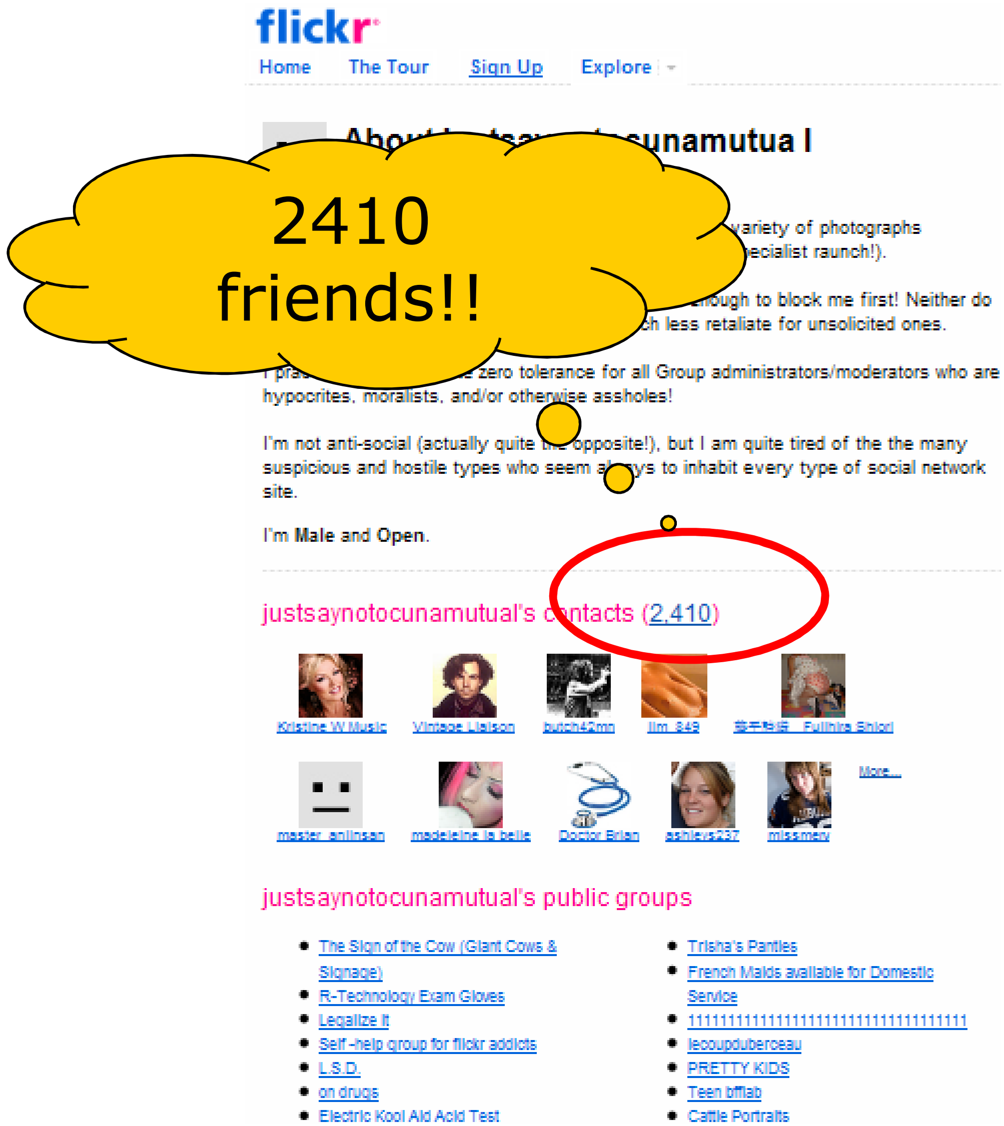
Heterogeneous Interaction Analysis

- A latent community structure is shared in a multi-dimensional network
  - E.g. a group sharing similar interest

- Goal: *Find out the shared community structure by integrating the interactions at different dimensions*

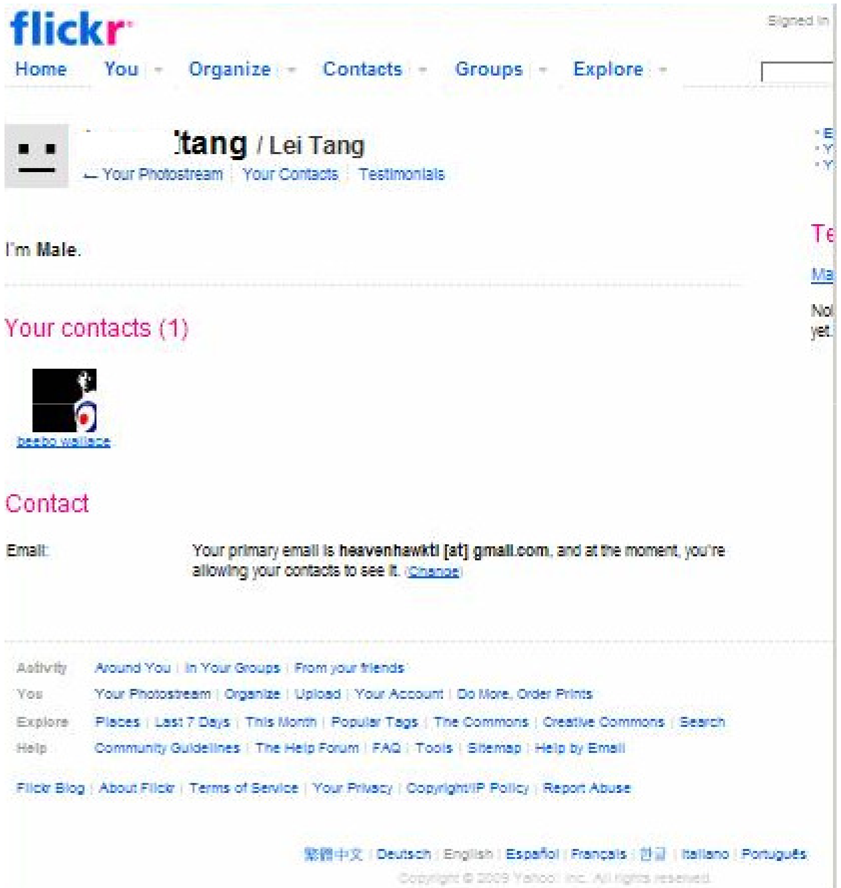
Why not just friends network?

- Too many friends?



Why not just friends network?

- Too many friends?
- Too few friends?
- Friends network tells limited info for some users
- Interaction at other dimensions might help



Recap of Modularity

- **Modularity**: A measure that compares the within group interaction with uniform random graph *with the same node degree distribution*

- In a network of m edges, for two nodes with degree d_i and d_j, respectively, the expected number of edges between them:

$$d_i d_j / 2m$$

- The connection strength in a group: $\sum_{i \in C, j \in C} A_{ij} - d_i d_j / 2m$

- To partition the network into multiple groups, we maximize

$$\frac{1}{2m} \sum_{C} \sum_{i \in C, j \in C} A_{ij} - d_i d_j / 2m$$

Modularity Matrix

- Modularity can be formulated in a matrix form

$$Q = \frac{1}{2m} Tr(S^T B S)$$

- B is the modularity matrix

$$B_{ij} = A_{ij} - d_i d_j / 2m$$

- With spectral relaxation, S corresponds to the top eigenvectors of the modularity matrix B

Modularity in M-D Networks

- Average Modularity Maximization (AMM)
  - Average the network interaction of different dimensions

$$\bar{A} = \frac{1}{D}(A_1 + A_2 + \cdots + A_D)$$

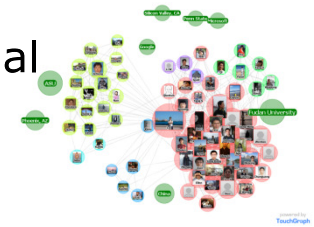
- Total Modularity Maximization (TMM)
  - the total sum of modularity at different dimensions

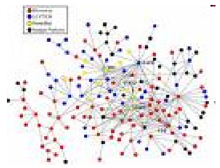
$$\max(Q_1 + Q_2 + \cdots + Q_D)$$

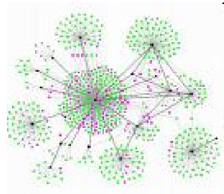
- Potential Cons
  - Not sure whether the interaction or modularity of different dimensions are comparable
  - Can be sensitive to networks of noisy dimensions

Principal Modularity Maximization

Multi-Dimensional Networks







Extract Structural Features via Modularity Maximization

Denoise the interaction at each dimension

- These structural features are not necessarily similar, but are highly correlated.

- Transform these features into a shared space such that their correlation is maximized.

- Solution: Generalized Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis

$$R(i,j) = (S_i w_i)^T (S_j w_j) = w_i^T (S_i^T S_j) w_j = w_i^T C_{ij} w_j$$

$$\max \quad \sum_{i=1}^{d} \sum_{j=1}^{d} w_i^T C_{ij} w_j$$

$$s.t. \quad \sum_{i=1}^{d} w_i^T C_{ii} w_i = 1$$

$\Rightarrow$

$$\begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1d} \\ C_{21} & C_{22} & \cdots & C_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ C_{d1} & C_{d2} & \cdots & C_{dd} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = \lambda \begin{bmatrix} C_{11} & 0 & \cdots & 0 \\ 0 & C_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & C_{dd} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

$$C_{11} = S_1^T S_1 = I$$
$$C_{22} = S_2^T S_2 = I$$
....

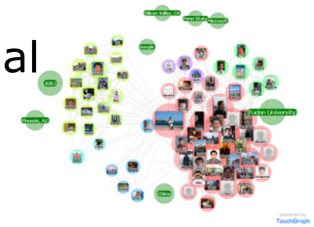
$\Downarrow$

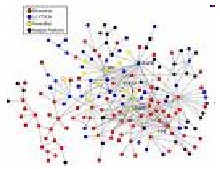
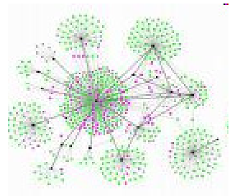
Principal Component Analysis (PCA)

eigvenvector of the covariance matrix

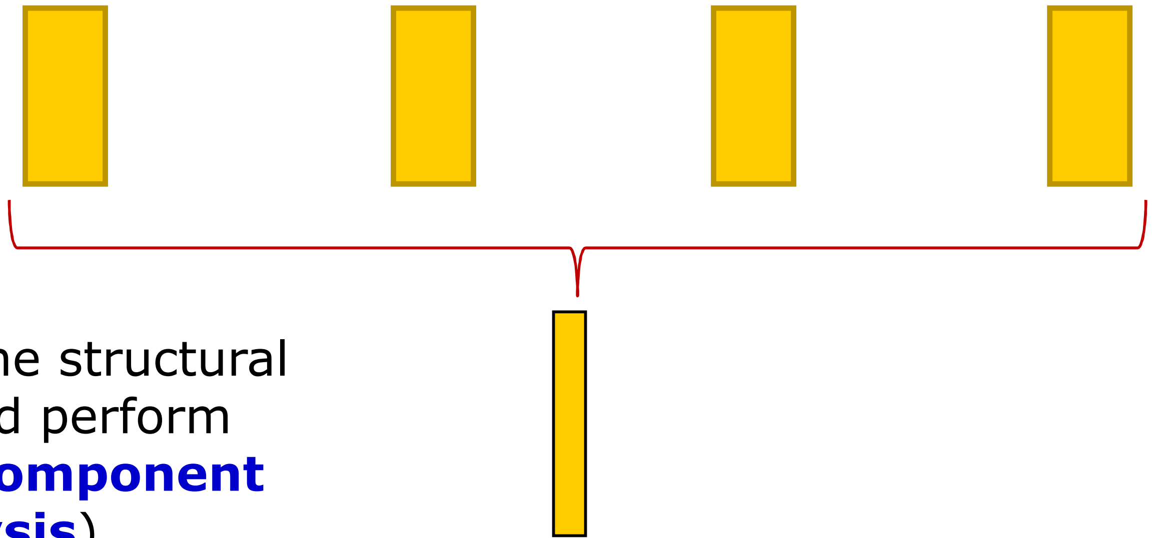
Principal Modularity Maximization

Multi-Dimensional Networks



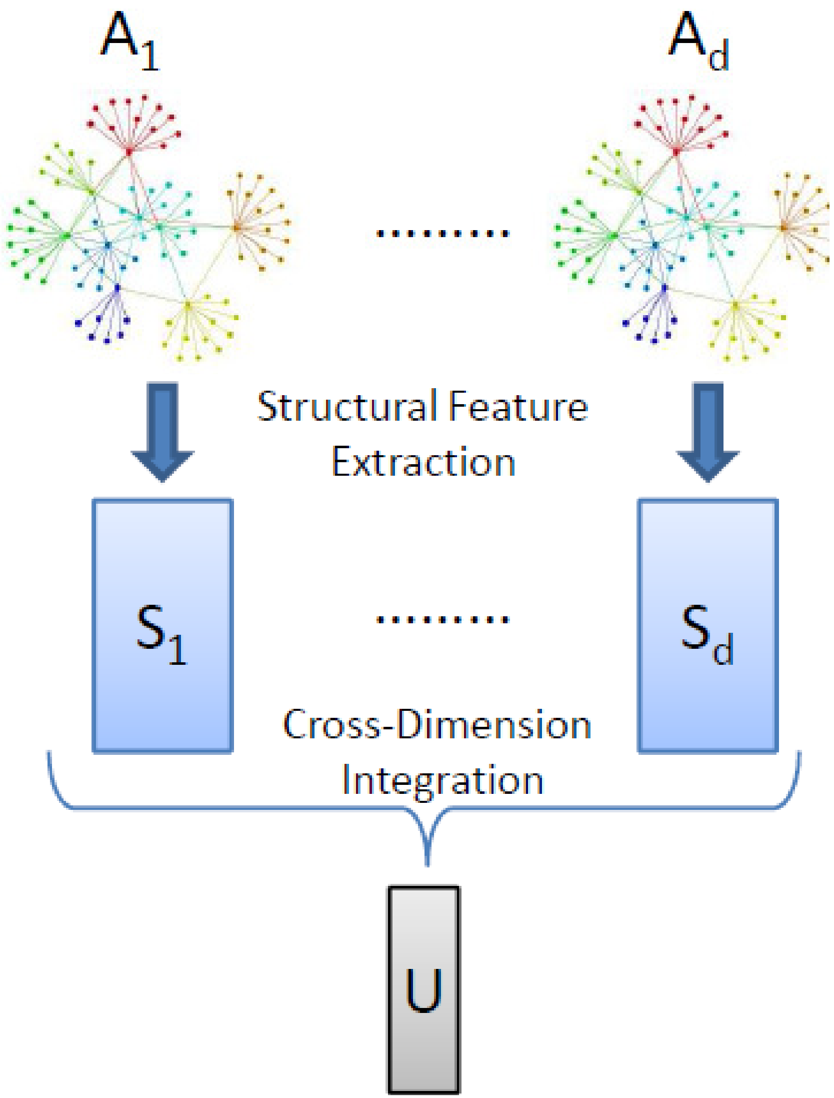



Extract Structural Features via Modularity Maximization



Combine all the structural features and perform **Principal Component Analysis**)

Overview of PMM



PMM Algorithm

- **Given:** a multi-dimensional network
- **Output:** shared group structure

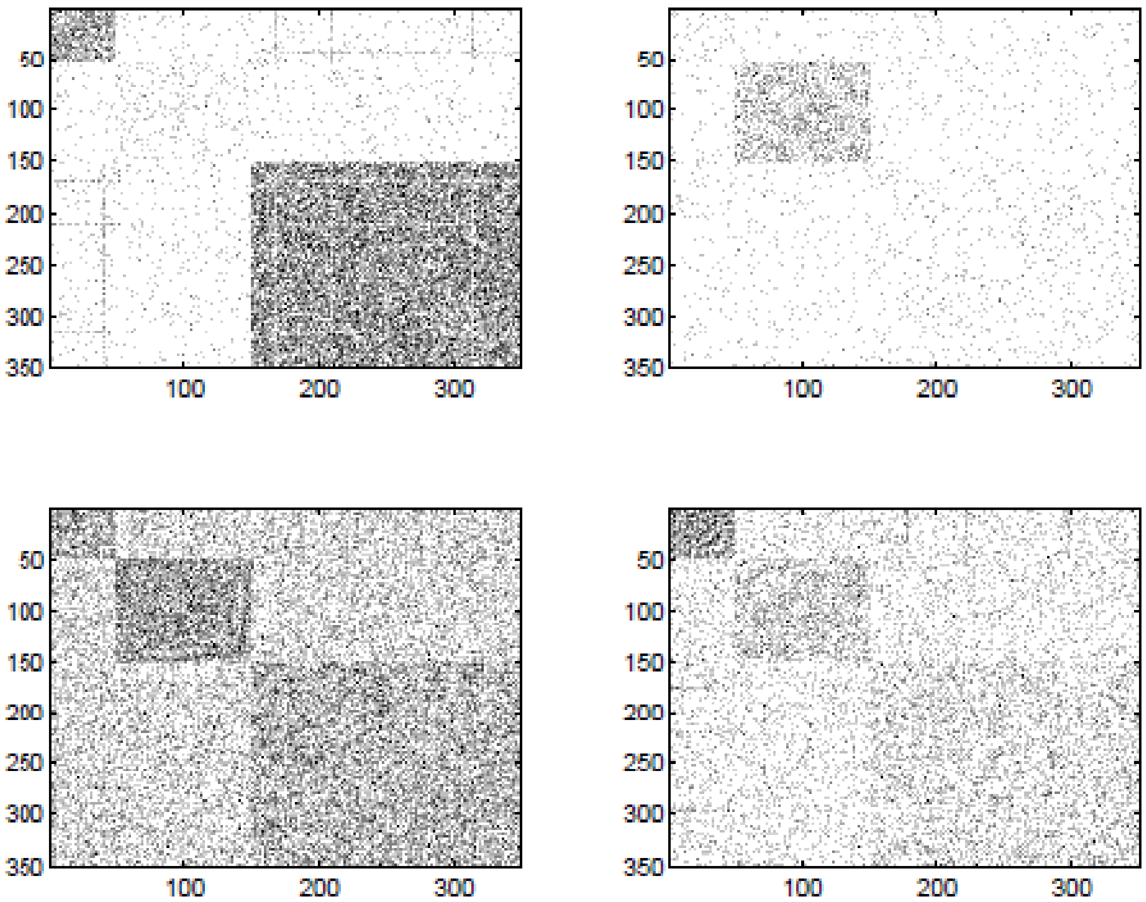
- **Algorithm:**
  - Phase I: Extract structural features from each dimension of the network
  - Phase II: Combine all the extracted features of each dimension and perform PCA
  - Apply K-means to obtain the discrete partition

Experiments

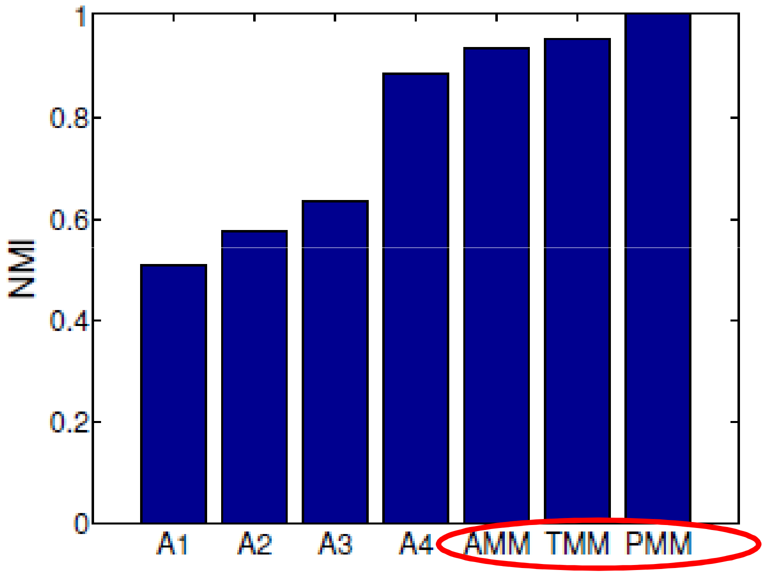
- Compare different community detection strategies
  - AMM, TMM, PMM
  - Modularity maximization on a single dimension
- Verify the sensitivity to noise for different methods
- Data Sets
  - Synthetic Data
    - controlled noise and ground truth information
  - Real-World Data
    - collected from YouTube

Synthetic Data

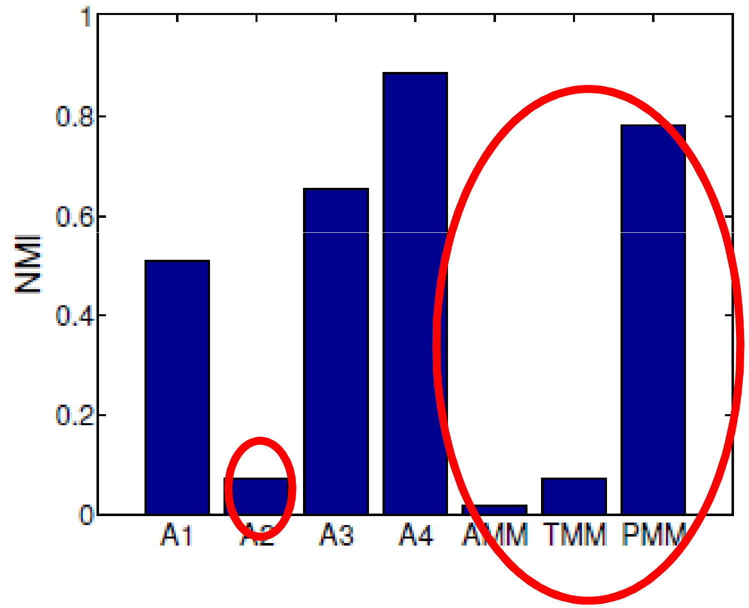
- 4 dimensions, 3 communities
- 50, 100, 150 members respectively



Performance on Synthetic Data



Small Noise



Substantial Noise
in one dimension

Average Performance

Table 1: Average Performance Over 100 Runs

	Strategy	Performance
	A_1	0.7237 ± 0.1924
Single-Dimensional	A_2	0.6798 ± 0.1888
	A_3	0.6672 ± 0.1848
	A_4	0.6906 ± 0.1976
	AMM	0.7946 ± 0.1623
Multi-Dimensional	TMM	0.9157 ± 0.1137
	PMM	$\mathbf{0.9351 \pm 0.1059}$

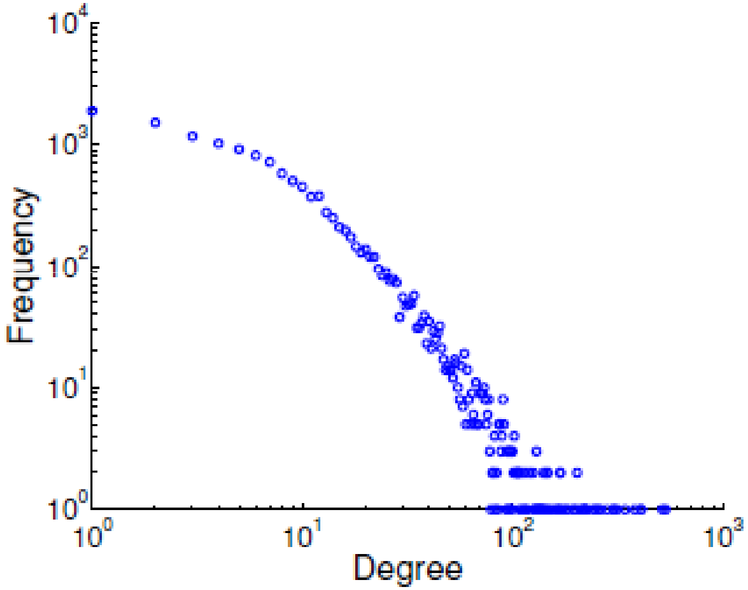
Single < AMM < TMM < PMM

PMM: Low Variance

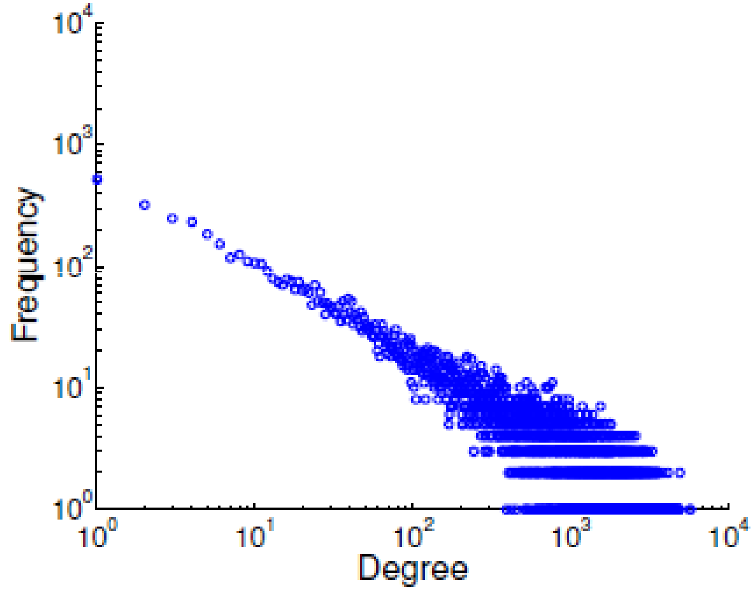
YouTube Data

- Collect contact network, subscription network, and users' favorite videos
- Crawl 30,522 user profiles reaching in total 848,003 users and 1,299, 642 favorite videos
- 15,088 active users
- Construct a 5-dimensional network
  - Contact
  - Share Contacts
  - Share subscription
  - Followed by the same set of people
  - Share favorite videos

Degree Distribution



(a) Contact Network



(b) Favorite Network

Evaluation on Real-World Data

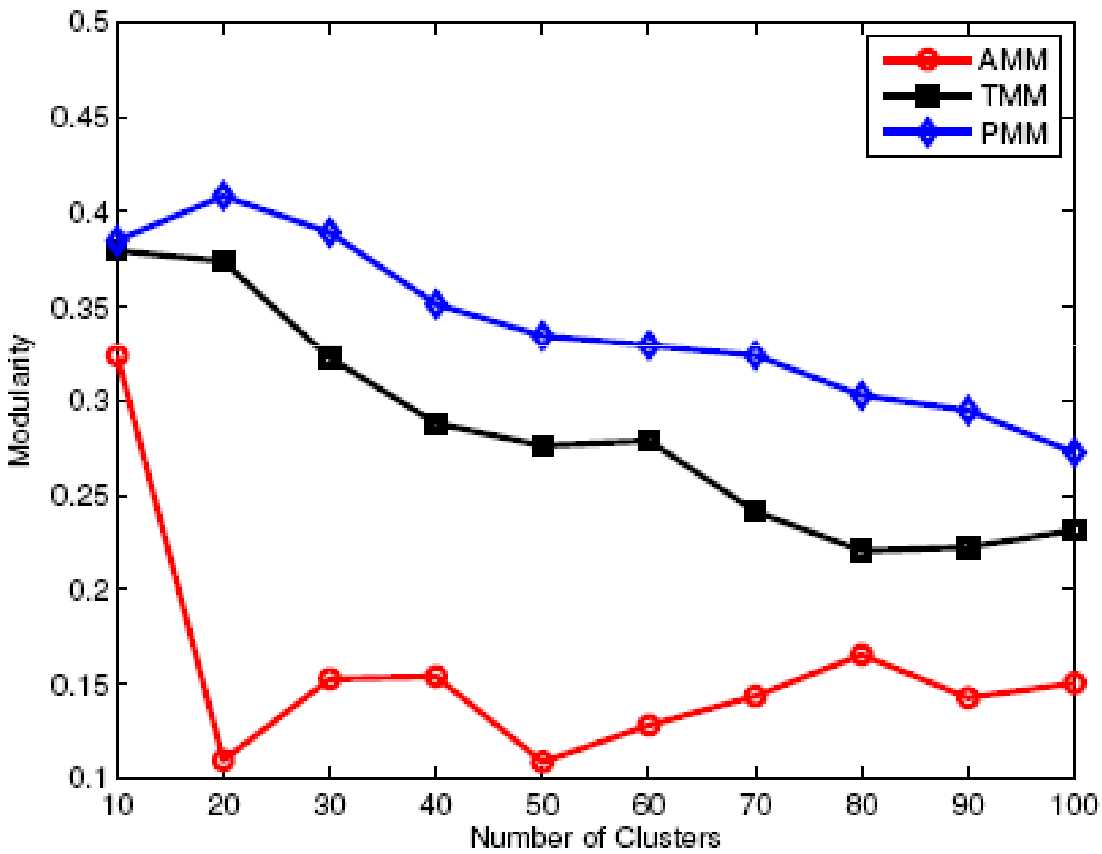
- Challenges
  - No ground truth
  - Need a smart way to do the comparative study
- Evaluation -- Cross Dimension Validation
  - Follow the idea of cross validation
  - For a network of *D* dimensions
    - learn the community structure from (*D-1*) dimensions
    - evaluate it on the remaining dimension in terms of modularity

Performance on YouTube Data

Methods	A_1	A_2	A_3	A_4	A_5
A_1	—	.0007	.0008	.0008	.0002
A_2	.1548	—	.0133	.0361	.0076
A_3	.0712	.0275	—	.0446	.0140
A_4	.0584	.0569	.0186	—	.0108
A_5	.0314	.0135	.0095	.0180	—
AMM	.1096	.0001	.0018	.0053	.0070
TMM	.3740	.1856	.1246	.1800	.0706
PMM	**.4085**	**.2063**	**.1307**	**.1844**	**.0947**

- PMM tends to be the winner

PMM compared with AMM & TMM



AMM < TMM < PMM

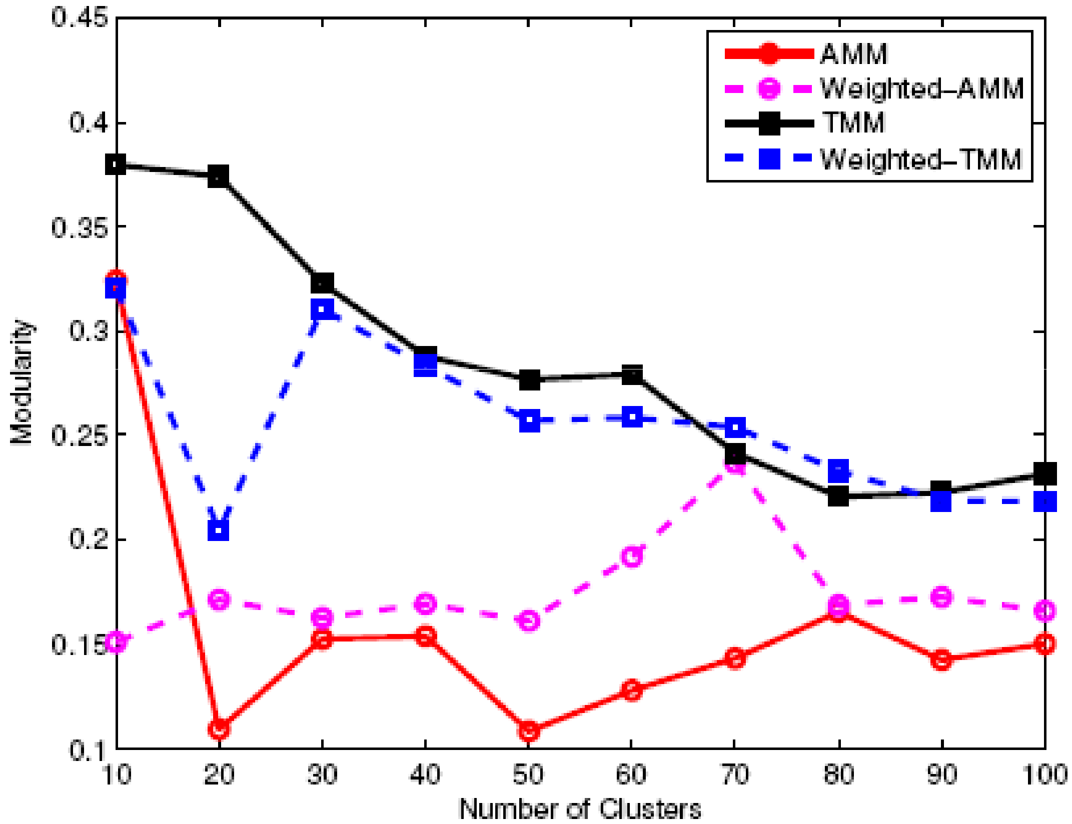
Conclusions & Future Work

- Networks in social media are multi-dimensional and noisy
- Propose an effective Principal Modularity Maximization to extract the shared group structure
  - Extract Structural Features via Modularity Maximization
  - Perform Cross-Dimensional Integration via PCA
- Can be applied similarly to other spectral clustering methods
- Future Work:
  - Determine whether two network dimensions share the same community structure?
  - Need to remove noisy interaction dimensions?
  - One actor assigned to multiple different groups?
  - Scalability?

Acknowledgments

- Thanks to the sponsorship of AFOSR and ONR.

Weighted AMM & TMM



Assigning a proper weight to each dimension is not easy!