Parameter-free Identification of Cohesive Subgroups in Large Attributed Graphs

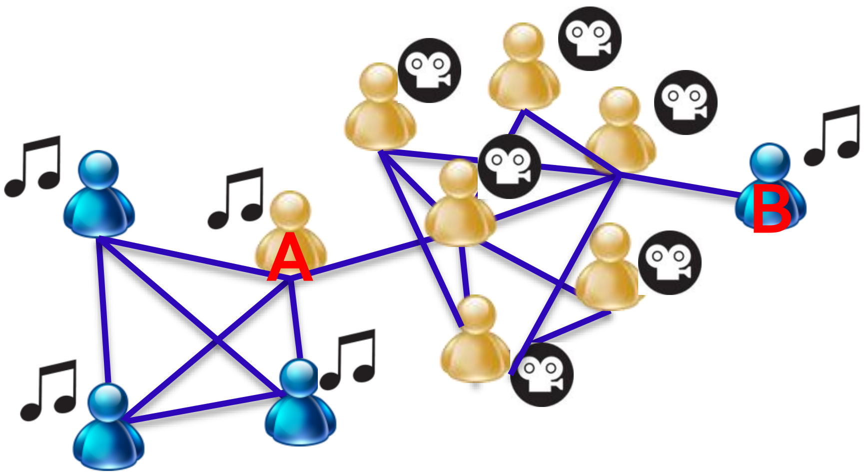
Leman Akoglu — Carnegie Mellon University
Hanghang Tong — IBM T. J. Watson
Brendan Meeder — Carnegie Mellon University
Christos Faloutsos — Carnegie Mellon University

PICS: problem

Given a graph with node attributes (features)

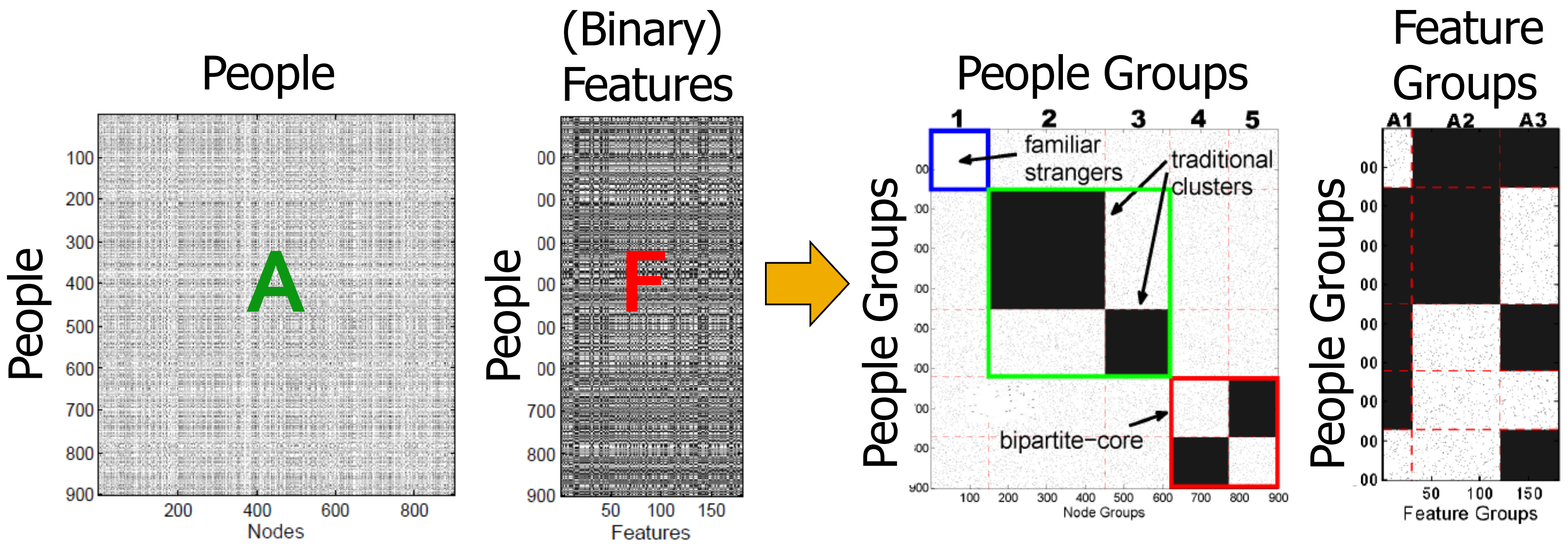
social networks + user interests
phone call networks + customer demographics
gene interaction networks + gene expression info

Find cohesive clusters, bridges, anomalies



cohesive cluster: similar connectivity & attribute coherence

PICS: problem sketch



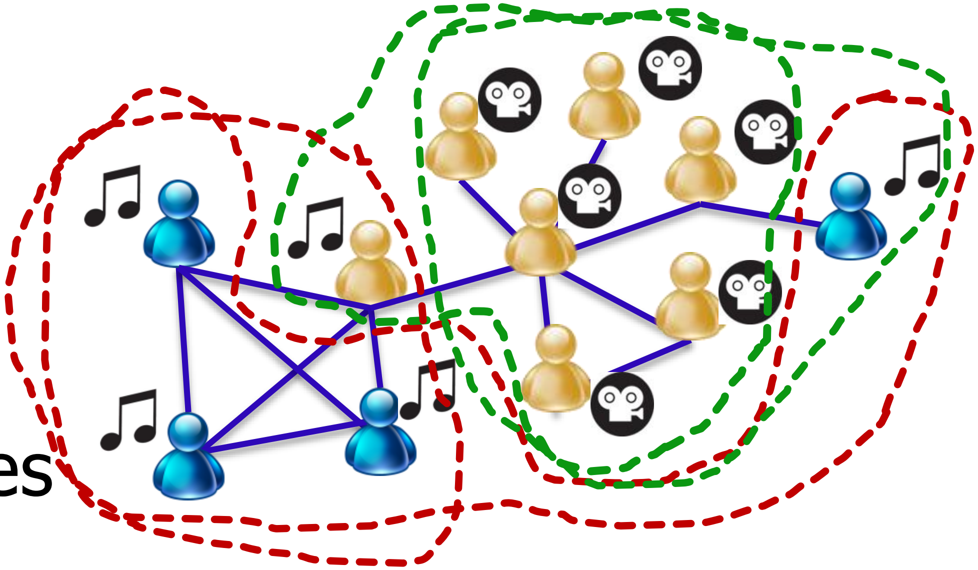
Given adjacency matrix **A** and feature matrix **F**

Find homogeneous blocks (clusters) in **A** and **F**

- * parameter-free
- * scalable

Simple extensions: why not?

- Flat clustering
- Graph clustering
  - Additional feature nodes
    - heterogeneous graph
  - Weighted edges by both connectivity and feature similarity
    - quadratic pairwise computations!
    - choice of similarity function



Related Work

	Graph structure	Node attributes	Parameter-free	Linear scalability
Flat clustering (e.g. k-means) [Kriegel+] [Leeuwen+]		✓		✓
METIS [Karypis and Kumar], [Flake+] [Girvan and Newman] [Andersen+] spectral [Ng+], co-clustering [Dhillon+]	✓			✓
SA-cluster [Zhou+], Spect. rel. clus. [Long+]	✓	✓		
CoPaM [Moser+], Gamer [Gunneman+]	✓	✓		?,✓
Autopart and cross-assoc.s [Chakrabarti+], GraphScope [Sun+], PaCK [He+]	✓		✓	✓

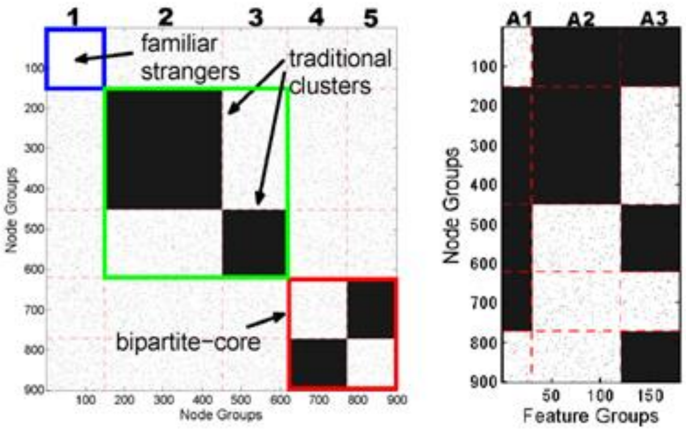
DETAILS

PICS: approach

1. How many node- & attribute-clusters?
2. How to assign nodes and attributes to clusters?

Main idea: employ Minimum Description Length

$$\underbrace{L\,(M)}_{\text{encoding length of clustering}} + \underbrace{L\,(D|M)}_{\text{encoding length of blocks}}$$



Good Clustering ⟺ implies ⟺ Good Compression

Minimum Description Length

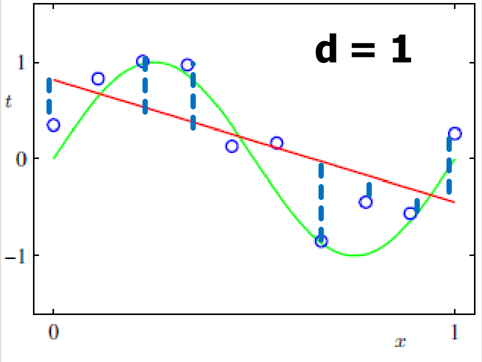
Given database D and set of models for D, MDL selects model M that minimizes

$$\underbrace{L\,(M)}_{\text{length in bits: description of } \textbf{model } M} + \underbrace{L\,(D|M)}_{\text{length in bits: } \textbf{data}\text{, encoded by } M}$$

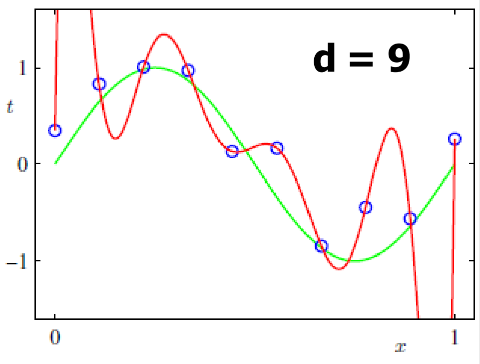
$\Downarrow$ a_1x+a_0 $\Downarrow$ deltas

vs.

$a_9x^9+\ldots+a_1x+a_0$ {}



vs.



Bishop: PR&ML

PICS: formulation

DETAILS

- L (M) : Model description cost

1. $\boxed{\log^\star n + \log^\star f}$ n: #nodes f: #attributes

2. $\boxed{\log^\star k + \log^\star l}$ k: #node-clus. l: #attribute-clus.

$$\log^\star(k) = \log(k) + \log\log(k) + \cdots$$

3. $\boxed{nH(P) + fH(Q)}$

$$p_i = \frac{r_i}{n}$$ ← size of node cluster i

$$q_j = \frac{c_j}{f}$$ ← size of attr. cluster j

$$optimal\ \#bits = -\log\frac{r_i}{n} = -\log p_i$$

$$node\,clus.\,cost = \sum_i r_i\,.-\log\frac{r_i}{n} = n.-\sum_i \frac{r_i}{n}\log\frac{r_i}{n} = nH(P)$$

PICS: formulation

DETAILS

- L(D|M): Data description cost given Model

1. For each block in A and F , #1s: $\boxed{\log^\star n_1(B_{ij})}$

2. Encoding cost of a block

$$\boxed{\begin{aligned} E(B_{ij}) &= -n_1(B_{ij})\log_2(P_{ij}(1)) - n_0(B_{ij})\log_2(P_{ij}(0)) \\ &= n(B_{ij})H(P_{ij}(1)). \end{aligned}}$$

where

$$P_{ij}(1) = n_1(B_{ij})/n(B_{ij})$$

$r_i c_j$ or $r_i r_j$

PICS: total cost objective

DETAILS

- L (M) : Model description cost

1. $\log^\star n + \log^\star f$ n: #nodes, f: #attributes
2. $\log^\star k + \log^\star l$ k: #node-clusters, l: #attribute-clusters
3. $nH(P) + fH(Q)$ $p_i = \frac{r_i}{n}$ ← size of node-cluster i

   $q_j = \frac{c_j}{f}$ ← size of attribute-cluster j

A similar problem (column re-ordering for minimum total run length) is shown to be NP-hard [Johnson+]. (reduction from Hamiltonian Path)

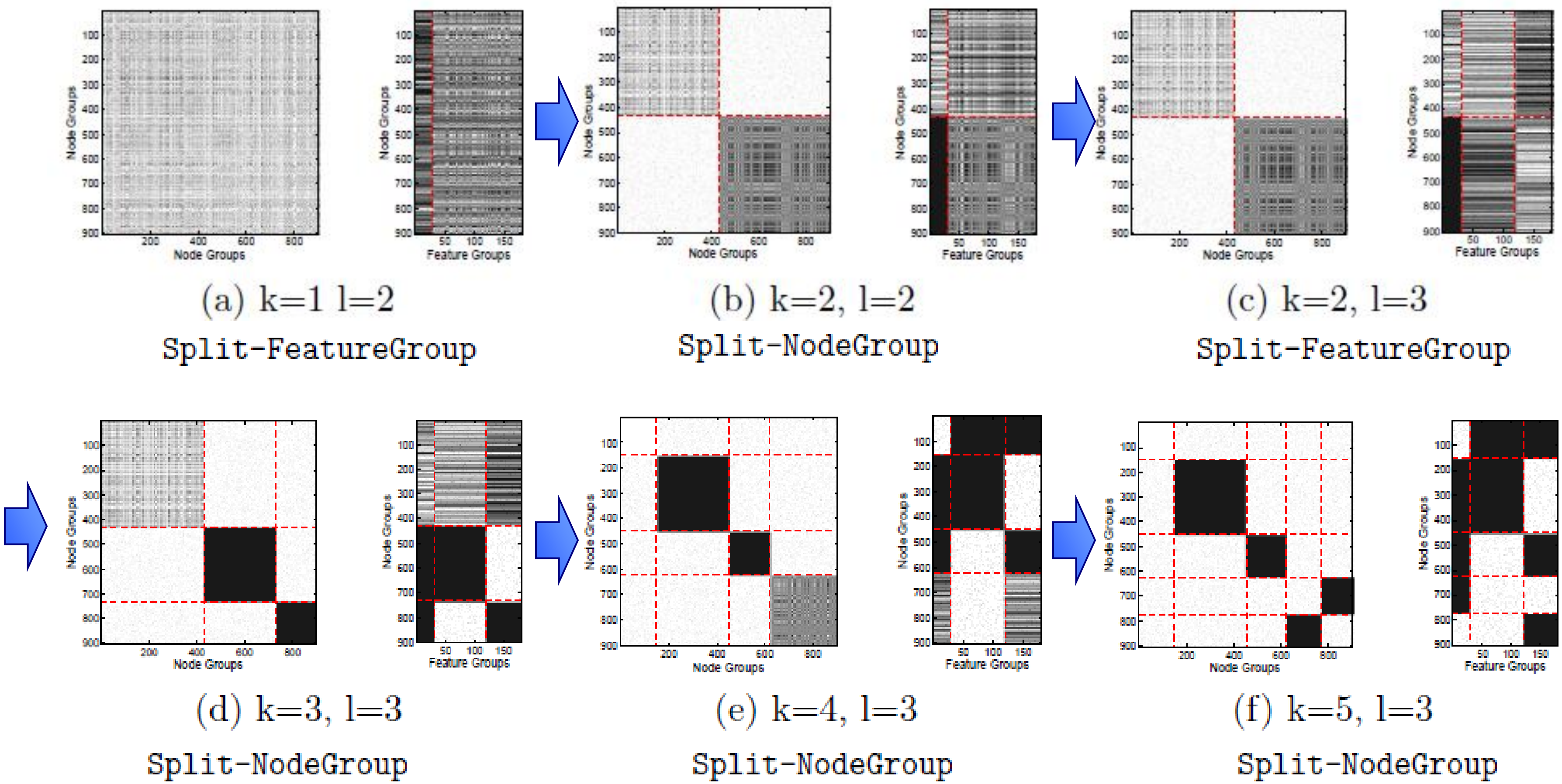
- L(D|M): Data description cost given Model

1. For each block in A and F, # 1s: $\log^\star n_1(B_{ij})$
2. $$E(B_{ij}) = -n_1(B_{ij})\log_2(P_{ij}(1)) - n_0(B_{ij})\log_2(P_{ij}(0))$$
   $$= n(B_{ij})H(P_{ij}(1)).$$

where $P_{ij}(1) = n_1(B_{ij})/n(B_{ij})$

$r_i c_j$ or $r_i r_j$

PICS: algorithm sketch



The algorithm is iterative and monotonic – will converge to local optimum

PICS: objective and algorithm

Total Encoding Cost (Length in bits)

$$L(\mathbf{A},\mathbf{F};R,C) = \log^\star n + \log^\star f + \log^\star k + \log^\star l$$
$$-\sum_{i=1}^{k} r_i \log_2\left(\frac{r_i}{n}\right) - \sum_{j=1}^{l} c_j \log_2\left(\frac{c_j}{f}\right)$$
$$+\sum_{i=1}^{k}\sum_{j=1}^{l}\left(\log^\star n_1(B_{ij}^F) + E(B_{ij}^F)\right)$$
$$+\sum_{i=1}^{k}\sum_{j=1}^{k}\left(\log^\star n_1(B_{ij}^A) + E(B_{ij}^A)\right).$$

Algorithm PICS

Input: $n \times n$ link matrix $\mathbf{A}$, $n \times f$ feature matrix $\mathbf{F}$

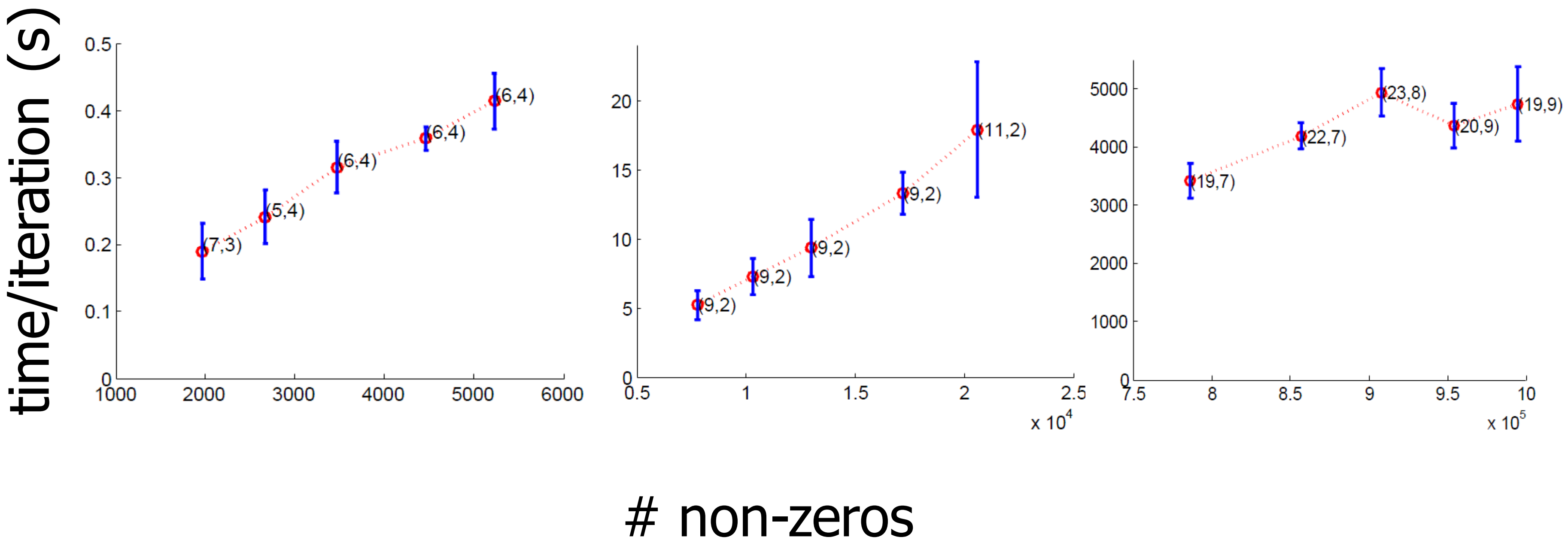
Output: A heuristic solution towards minimizing total encoding $L(\mathbf{A},\mathbf{F};R,C)$: number of row and column groups (k^*, l^*), associated mapping (R^*, C^*)

```
Set k^0 = l^0 = 1 as we start with a single node and feature cluster.
Set R^0 := {1,2,...,n} → {1,1,...,1}
Set C^0 := {1,2,...,f} → {1,1,...,1}
Let T denote the outer iteration index. Set T = 0.
repeat
    C^{T+1}, l^{T+1} := Split-FeatureGroup(F, C^T, l^T)
    (R^{T+1}, C^{T+1}) := Shuffle(A, F, (R^T, C^{T+1}), (k^T, l^{T+1}))
    R^{T+1}, k^{T+1} := Split-NodeGroup(A, F, (R^{T+1}, C^{T+1}), (k^T, l^{T+1}))
    (R^{T+1}, C^{T+1}) := Shuffle(A, F, (R^{T+1}, C^{T+1}), (k^{T+1}, l^{T+1}))
    if L(A, F; R^{T+1}, C^{T+1}) ≥ L(A, F; R^T, C^T) then
        return (k*, l*) = (k^T, l^T), (R*, C*) = (R^T, C^T)
    else
        Set T = T + 1
    end if
until convergence
```

PICS: scalability

Computational complexity:

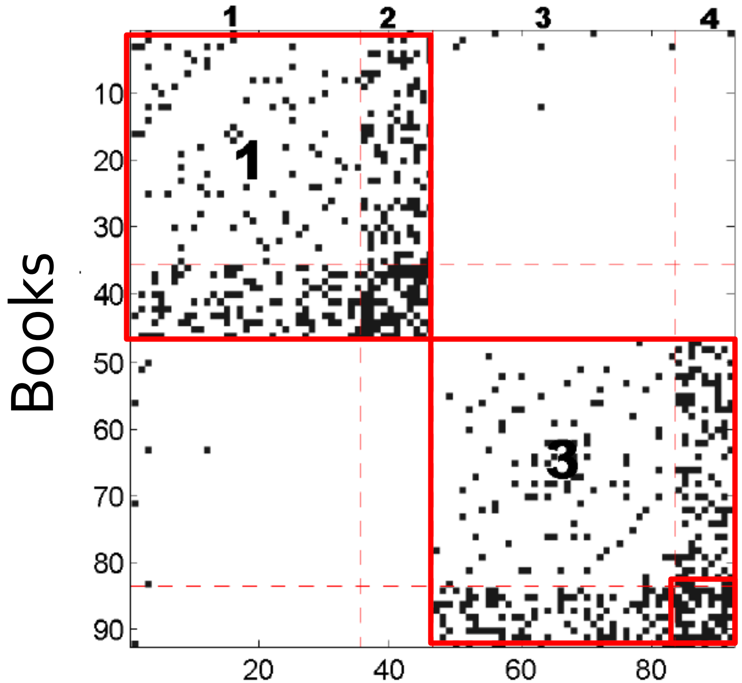
$$O(\max(k^*, l^*) * [2n_1(A)k^* + n_1(F)(k^* + l^*)] * \hat{t})$$



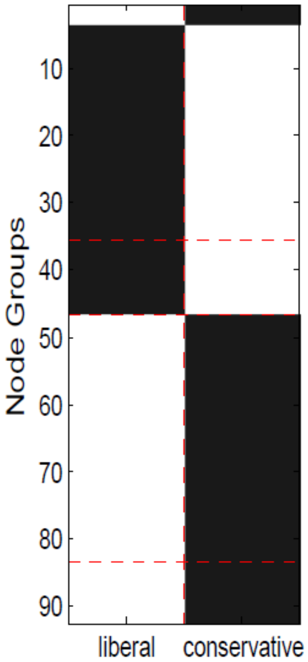
PICS: datasets

Graphs	Description	n	f	nnz
1. **Phone call**	users, titles	94	7	391
2. **Device**	users, titles	94	7	5K
3. PolBooks	books, incl.	92	2	840
4. **PolBlogs**	blogs, incl.	1.5K	2	20K
5. **Twitter**	users, h-tags	9.6K	10K	82K
6. YouTube	users, groups	77K	30K	1M
7. YeastGene	genes, articles	844	17K	64K

PICS at work (Political books)



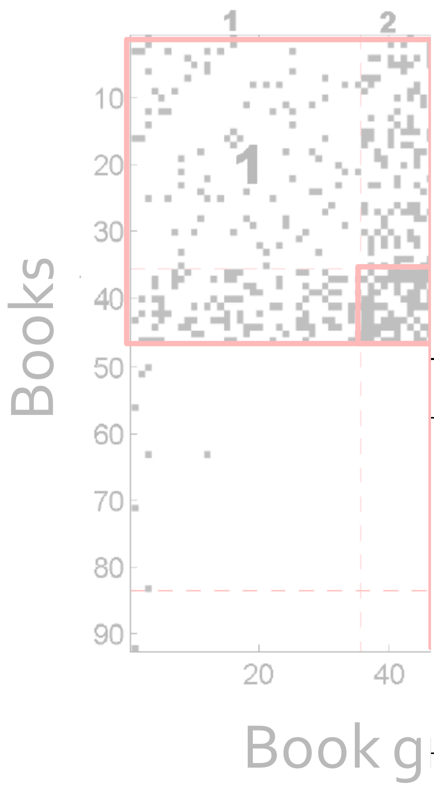
Book groups



liberal vs. conservative

"core and periphery"

PICS at work (Political books)



Examples of "core" liberal and conservative books

Liberal

–Lies and the Lying Liars Who Tell Them: *A Fair and Balanced Look at the Right*
–Big Lies: *The Right-Wing Propaganda Machine and How It Distorts the Truth*
–The Lies of George W. Bush
–Dude, Where's My Country?

Conservative

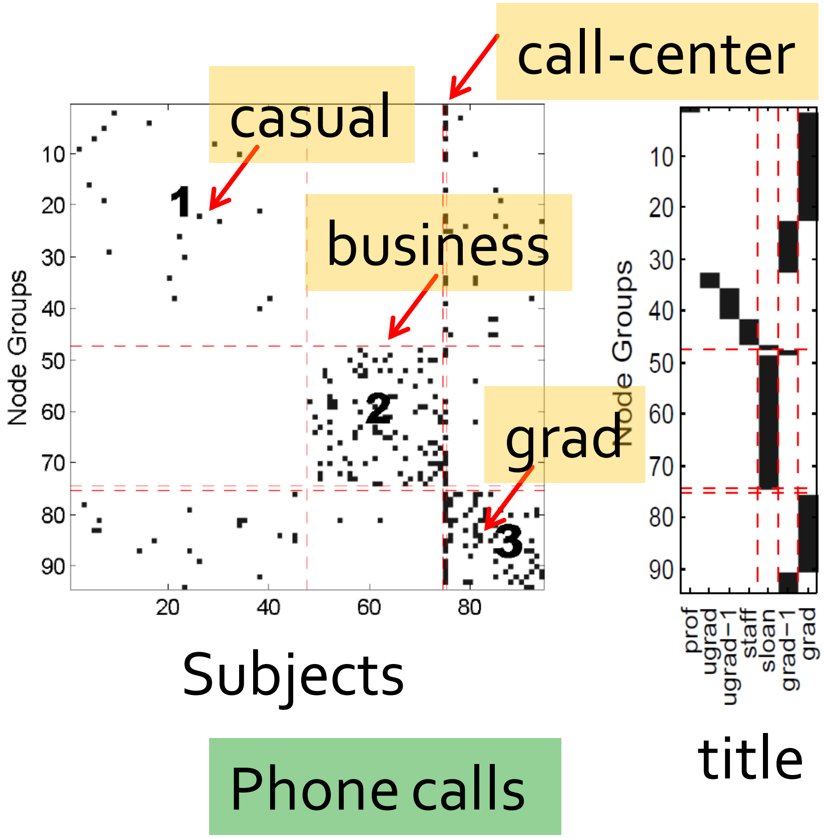
–Persecution: *How Liberals Are Waging War Against Christianity*
–Deliver Us from Evil: *Defeating Terrorism, Despotism, and Liberalism*
–Tales from the Left Coast
–A National Party No More

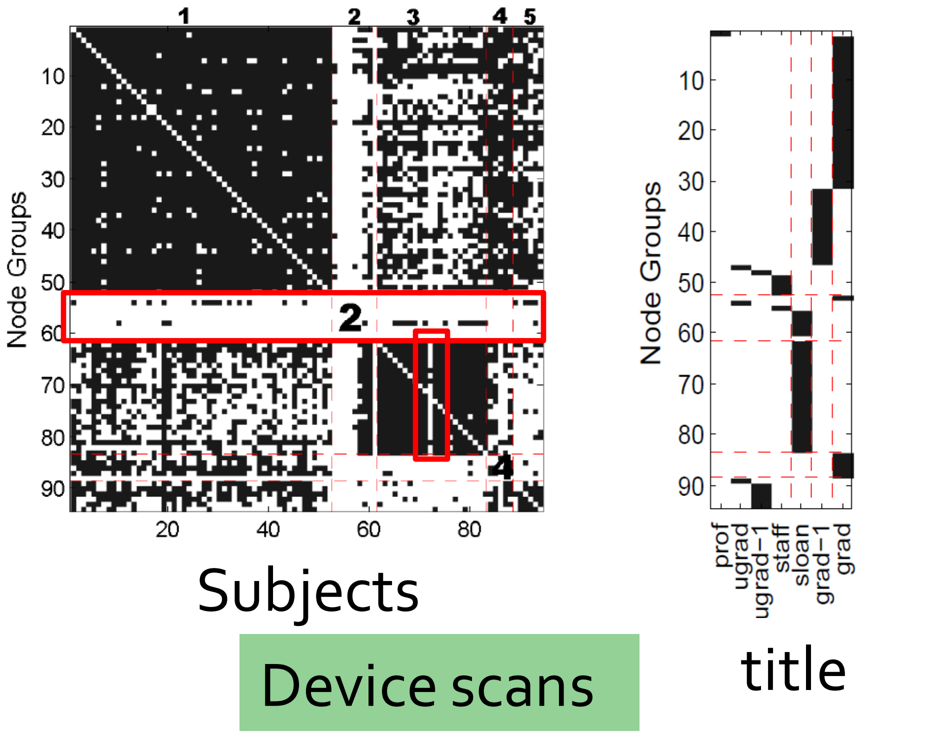
Examples of bridging 'conservative' books

–*Bush at War*
–*The Bushes: Portrait of a Dynasty*
–*Rise of the Vulcans: The History of Bush's War Cabinet*

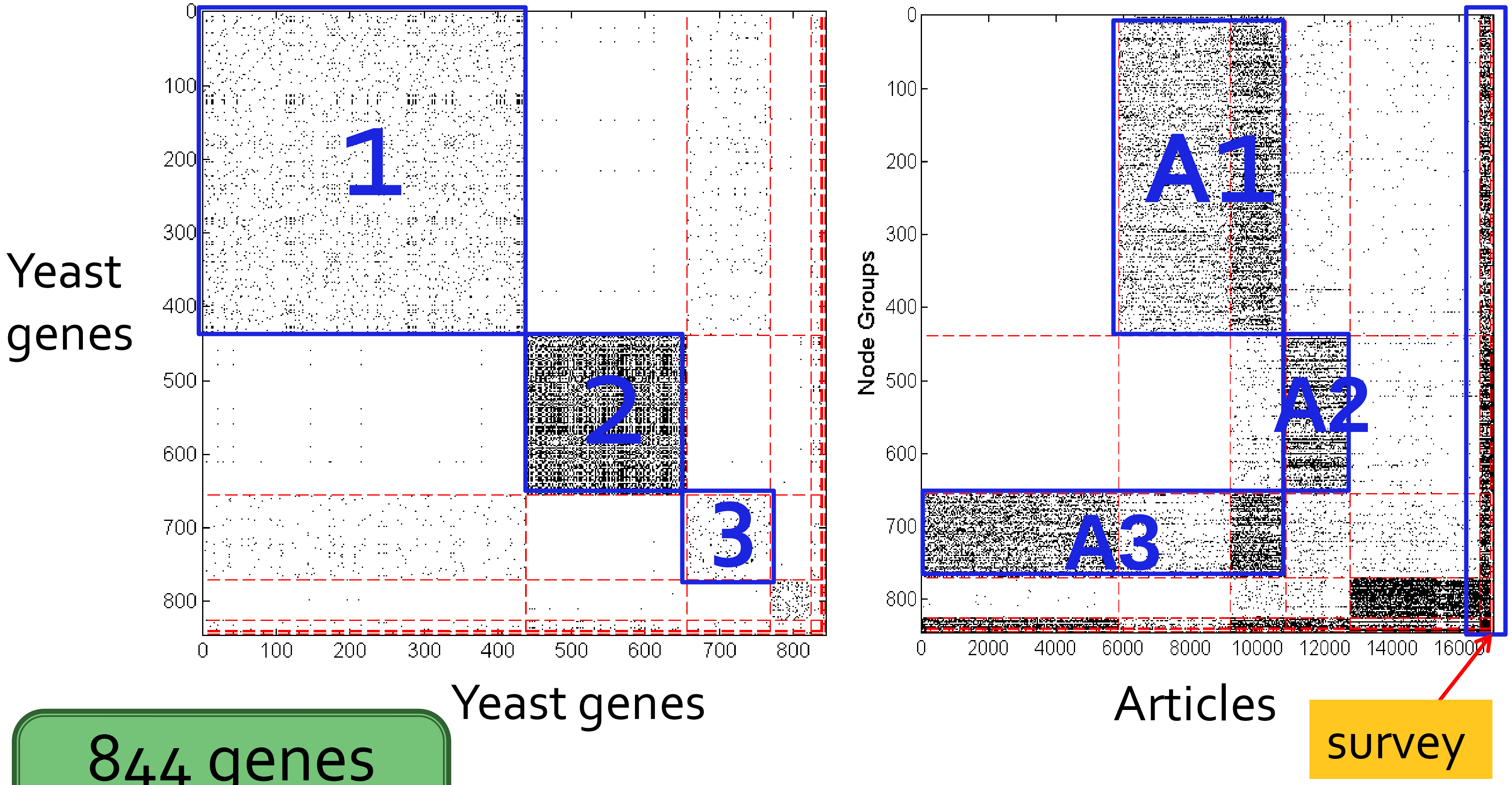
"core and perip

PICS at work (Reality mining)



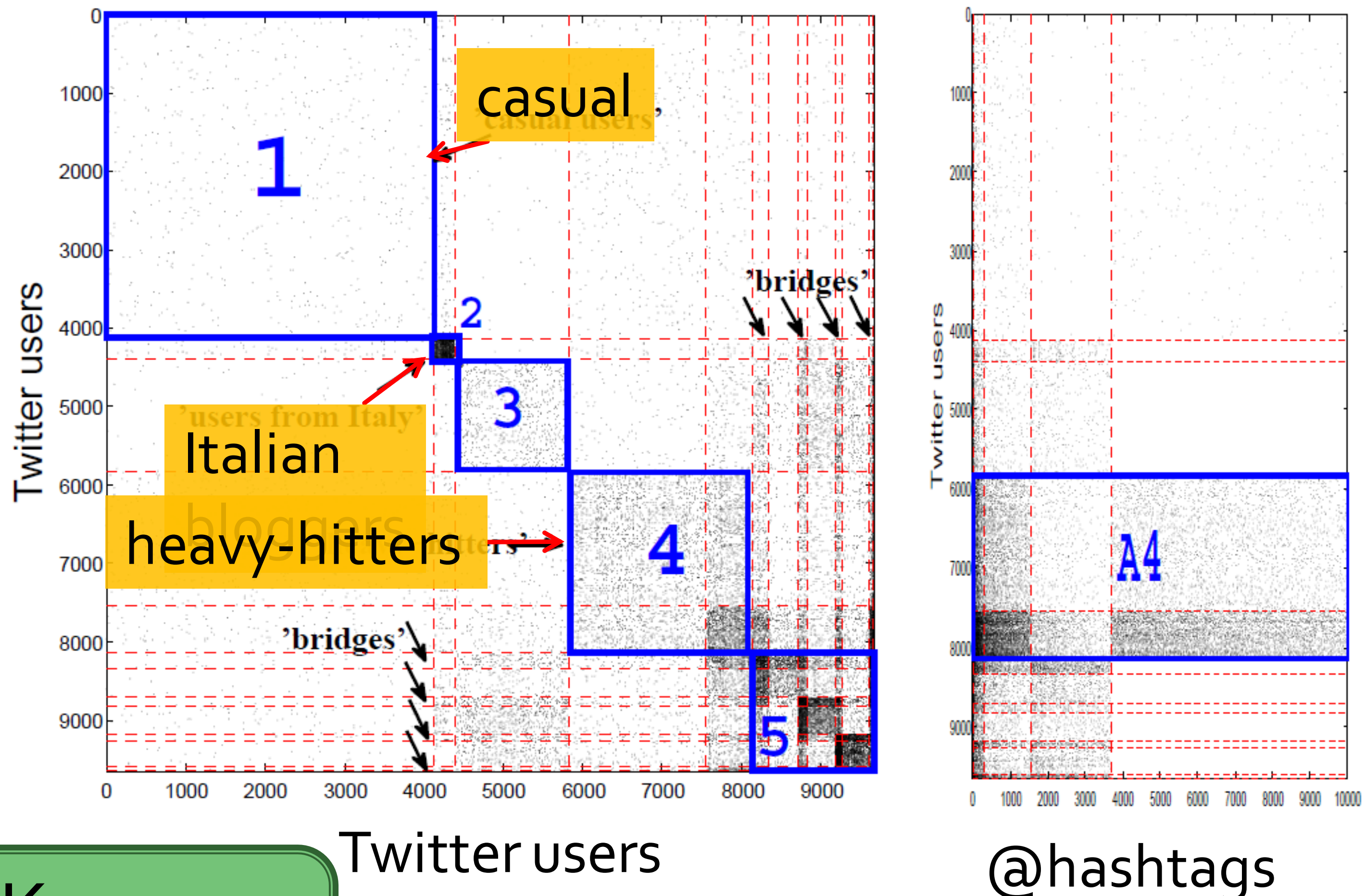


PICS at work (YeastGene)



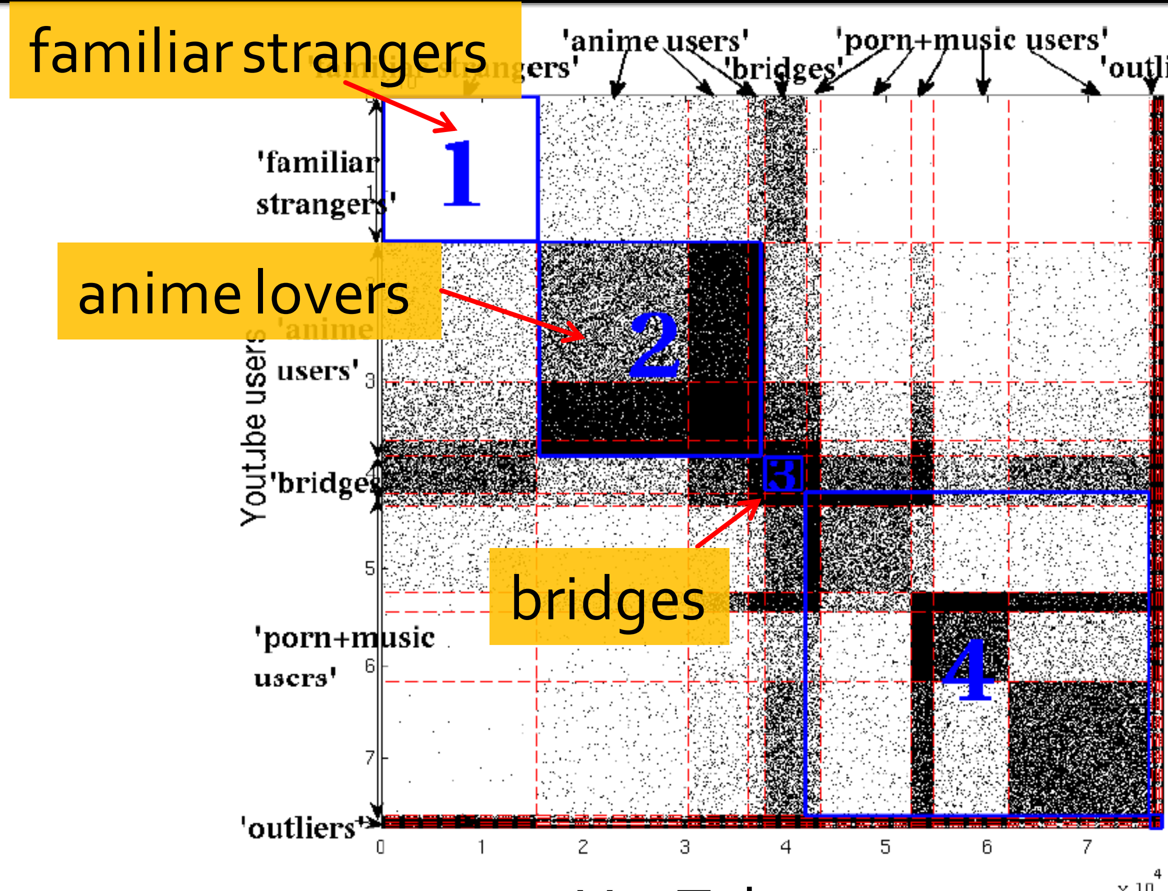
844 genes
17K articles

PICS at work (Twitter)

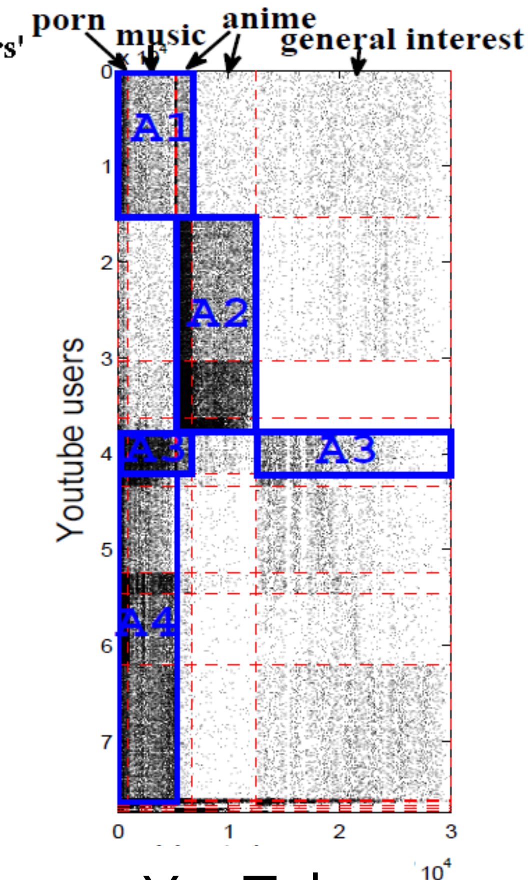


9,6K users
10K hashtags

PICS at work (YouTube)



YouTube users



YouTube groups

77K users
30K groups

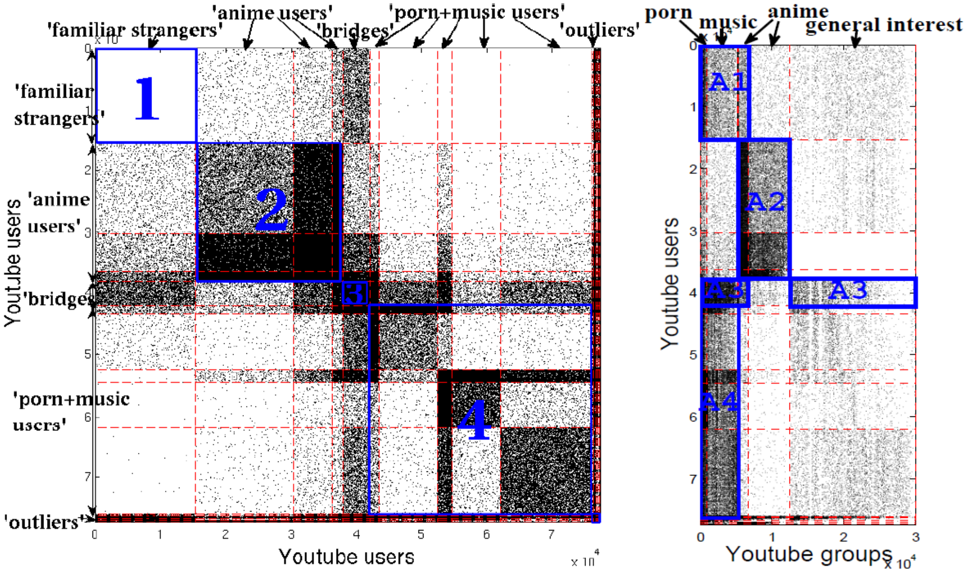
Summary of contributions

- Novel clustering model:
  - PICS finds groups of nodes in an attributed graph with (1) similar connectivity, and (2) attribute homogeneity.
  - It also groups the node attributes into attribute-clusters.
- Parameter-free nature:
  - No user input, e.g. number of clusters, similarity functions/thresholds
- Effectiveness:
  - Insightful clusters, bridges and outliers in diverse real-world datasets including YouTube and Twitter.
- Scalability:
  - Linearly growing run time with graph + attribute size

Thank you!

lakoglu@cs.cmu.edu

http://www.cs.cmu.edu/~lakoglu/



Source code: **www.cs.cmu.edu/~lakoglu/#pics**