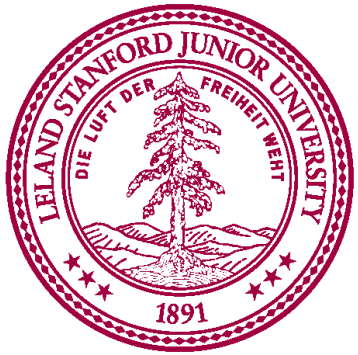
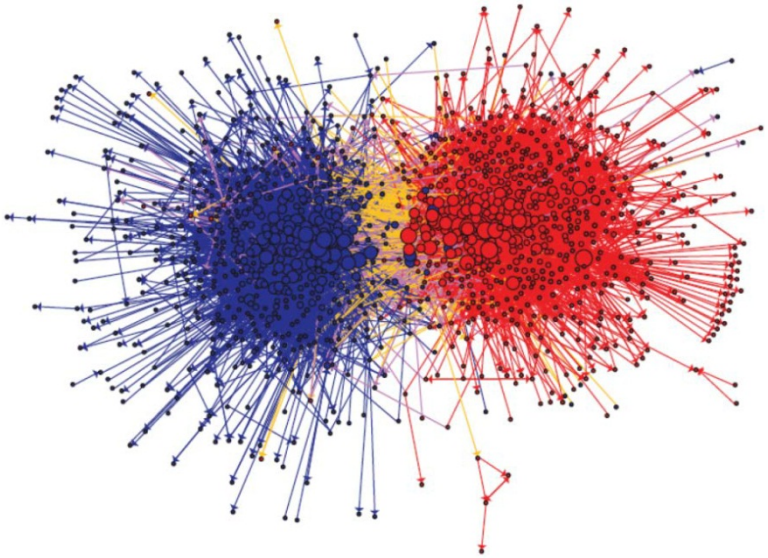
Networks, Communities and the Ground-Truth

Jaewon Yang and Jure Leskovec
Stanford University



Network Clusters

- **Networks are not uniformly/homogeneously linked but we observe formation of clusters**



Blogosphere [Adamic&Glance]

- **Why clusters? What do they correspond to?**

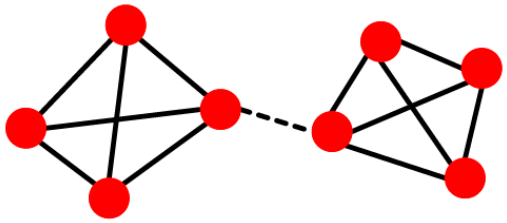
From Clusters to Communities

- **Idea: Clusters form communities**
  - **Cluster**: nodes with a certain connectivity structure
  - **Community**: nodes with a shared latent property

- **Many reasons why communities form:**
  - World Wide Web
  - Citation networks
  - Social networks
  - Metabolic networks

Basis for Community Formation

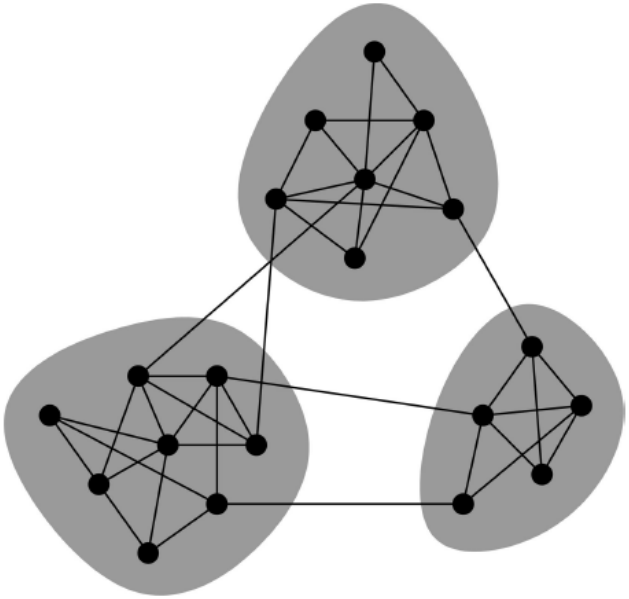
- **How and why do communities form?**
- **Granovetter's *Strength of weak ties* suggest and the models of small-world suggest:**
  - Strong ties are well embedded in the network
  - Weak ties span long-ranges

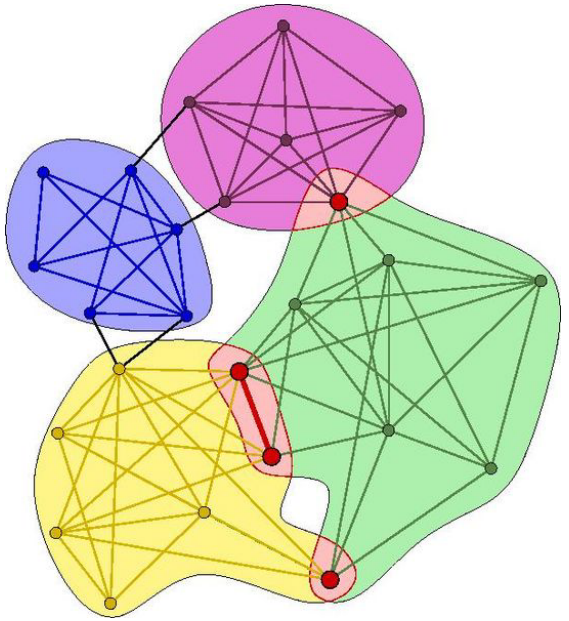


- **Given a network, how to find communities?**
  - Find weak ties and then identify the "boundary" of communities

Overlapping Communities

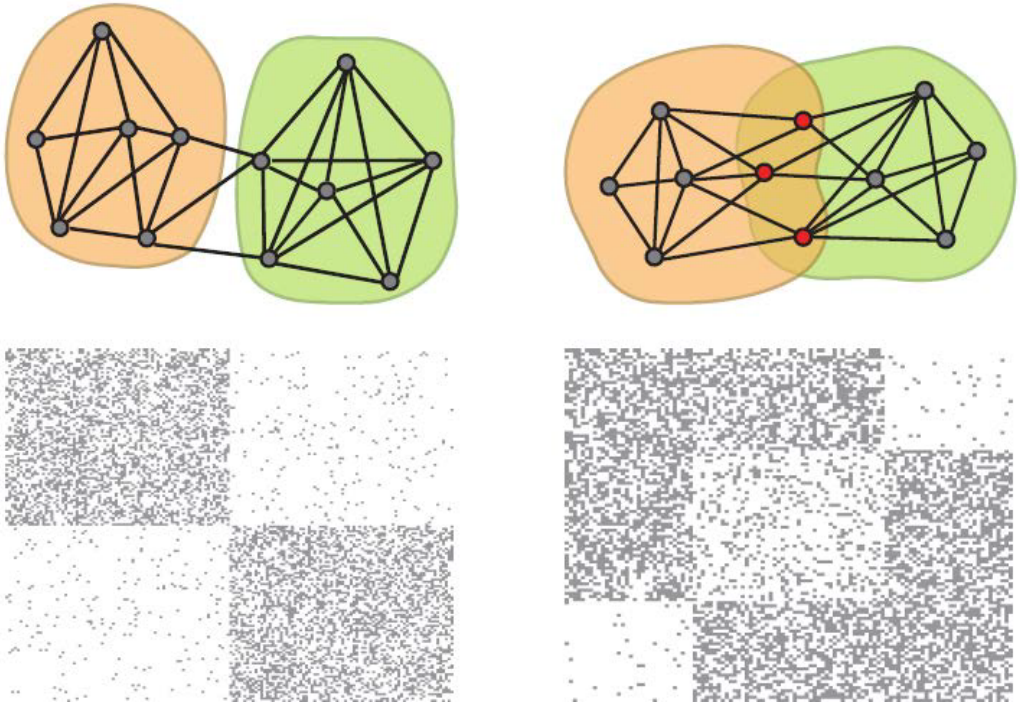
- Communities can overlap
  - The notion of weak-ties is extended for overlapping communities.





Communities in Networks

- **Assumptions about the structure of communities**



Granovetter and all non-overallping methods

Overlapping methods (CPM, MMSB, and so on)

Step Back: Community Detection

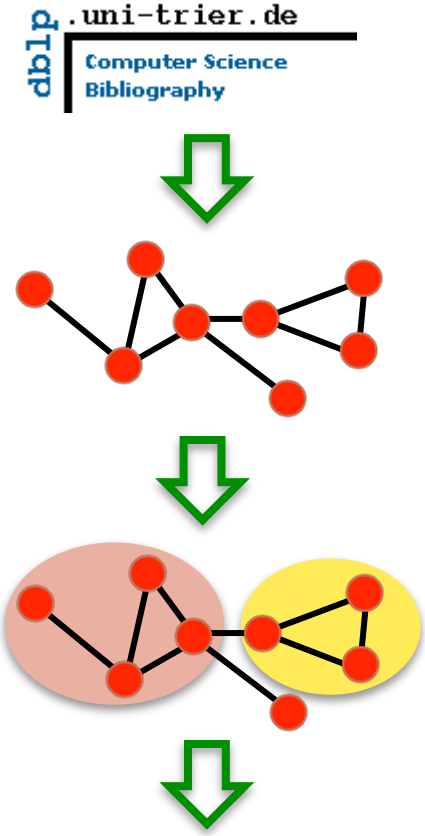
(1) Take a complex system

(2) Represent it as a graph

(3) Identify communities

(really, clusters)

(4) Interpret clusters as "real" communities

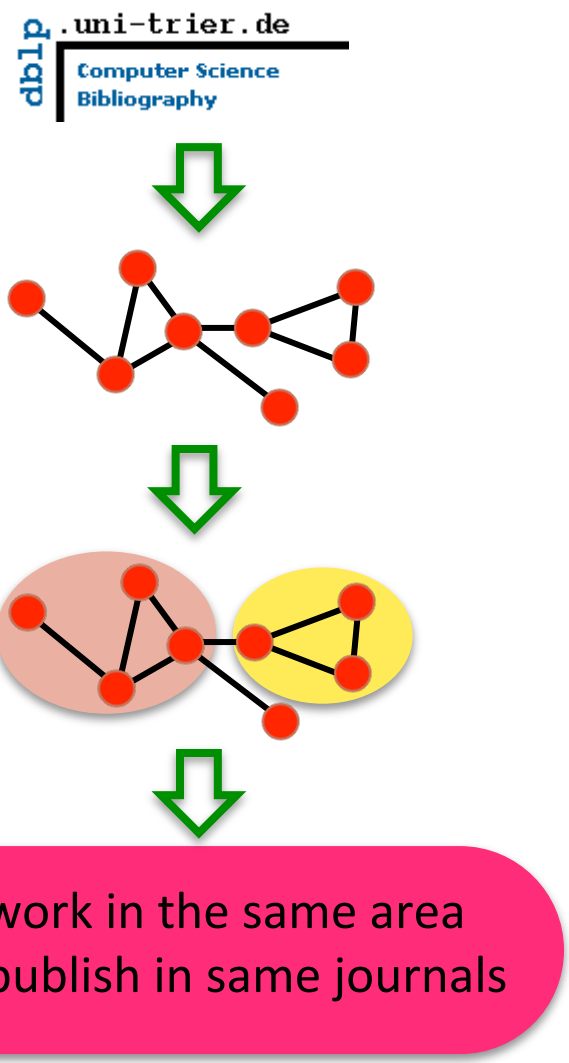


work in the same area

publish in same journals

Ground-Truth

- **Networks with a an explicit notion of Ground-Truth:**
  - **Collaborations:** Conferences & Journals as proxies for areas
  - **Social Networks:** People join to groups, create lists
  - **Information Networks:** Users create topic based groups

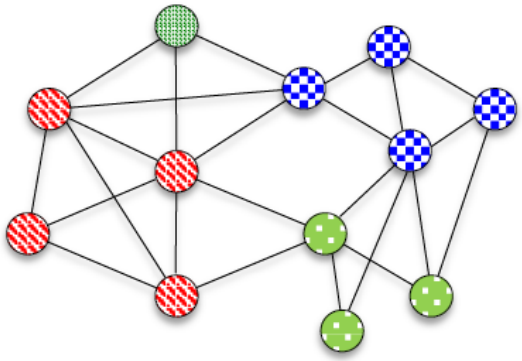


Example of Ground-Truth

- **LiveJournal social network**
  - Users create and join to **groups** created around culture, entertainment, expression, fandom, life/style, life/support, gaming, sports, student life and technology
- **TuDiabetes network**
  - **Groups** form around specific types of diabetes, different age groups, emotional and social support, arts and crafts groups, different geo regions
- A user can be a member of 0 or more groups

Networks with Ground-Truth

Dataset	N	E	C	S	A
LiveJournal	4.0 M	34.9 M	311,782	40.06	3.09
Friendster	117 M	2,586.1 M	1,449,666	26.72	0.33
Orkut	3.0 M	117.2 M	8,455,253	34.86	95.93
DBLP	0.4 M	1.3 M	2547	429.79	2.57
IMDB	1.3 M	39.8 M	205	6688.78	1.00
Amazon	0.3 M	0.9 M	49,732	99.86	14.83



Youtube social network

- N … # of nodes
- E … # of edges
- C … # of ground-truth communities
- S … average community size
- A … memberships per node

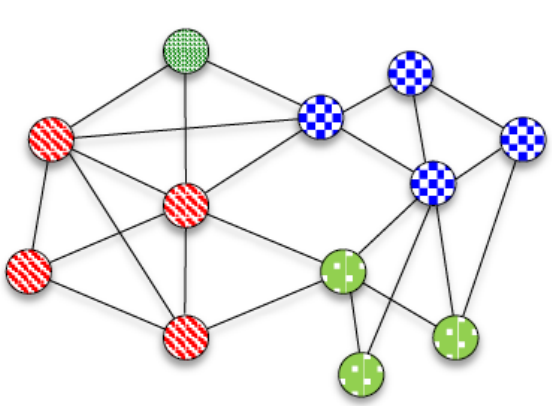
For example:

… fans of Real Madrid

… subscribe to Lady Gaga videos

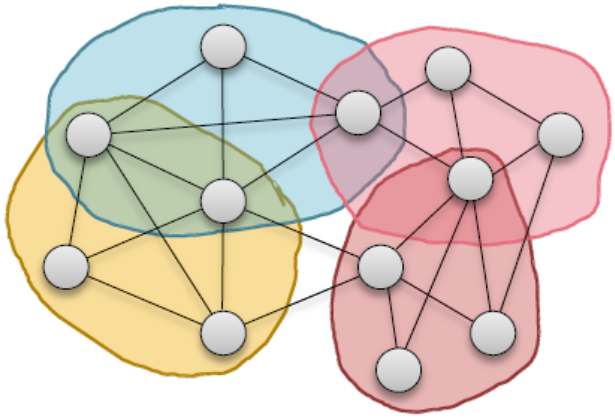
… follow Volvo Ocean Race

Ground-Truth: Consequences



Ground-truth groups

≈

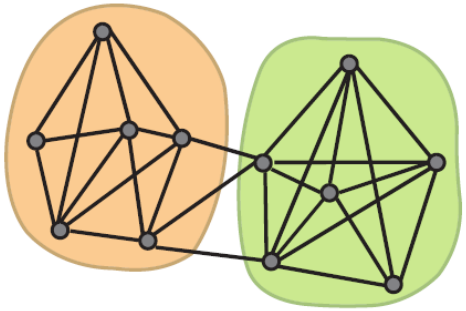


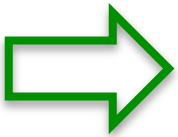
Inferred communities

- **How real groups map on the network?**
  - ⇒ **Insights for Better Algorithms**
- **How to evaluate and interpret?**
  - ⇒ **"Precision" of Algorithms**

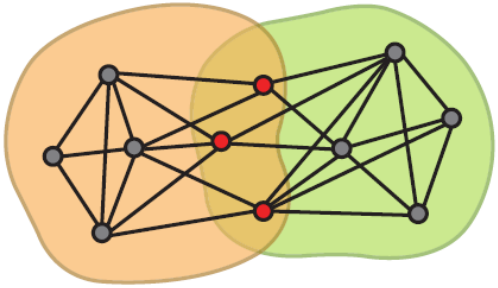
Groups and Networks

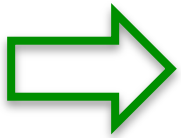
- Nodes *u* and *v* share *k* groups
- **What is edge prob. P(edge | k) as a func. of k?**
- **Today's wisdom:**





P(edge | k) = 0

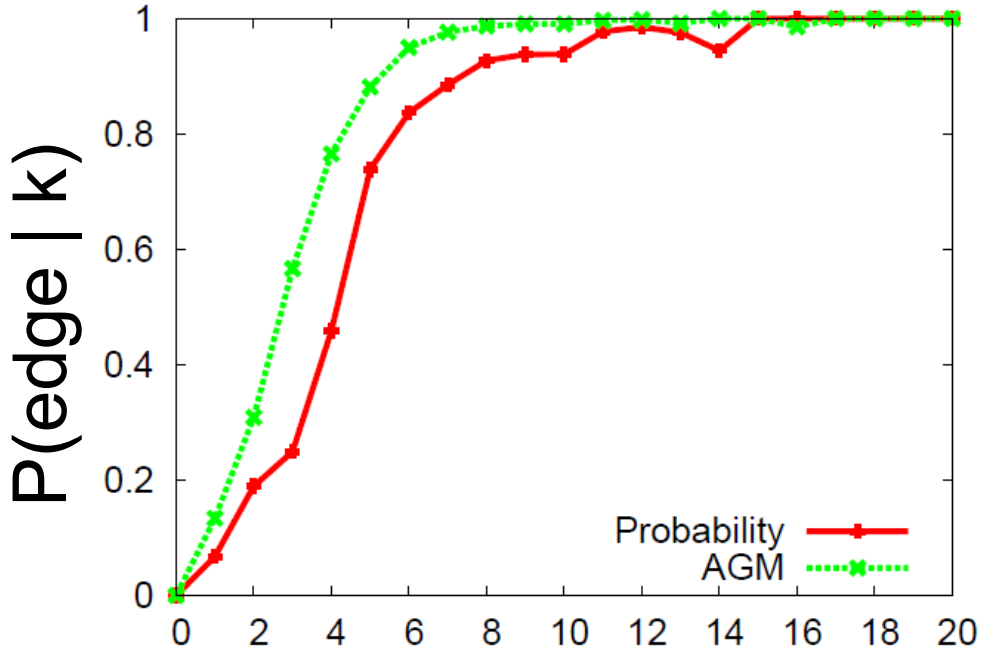




P(edge | k) = decreasing

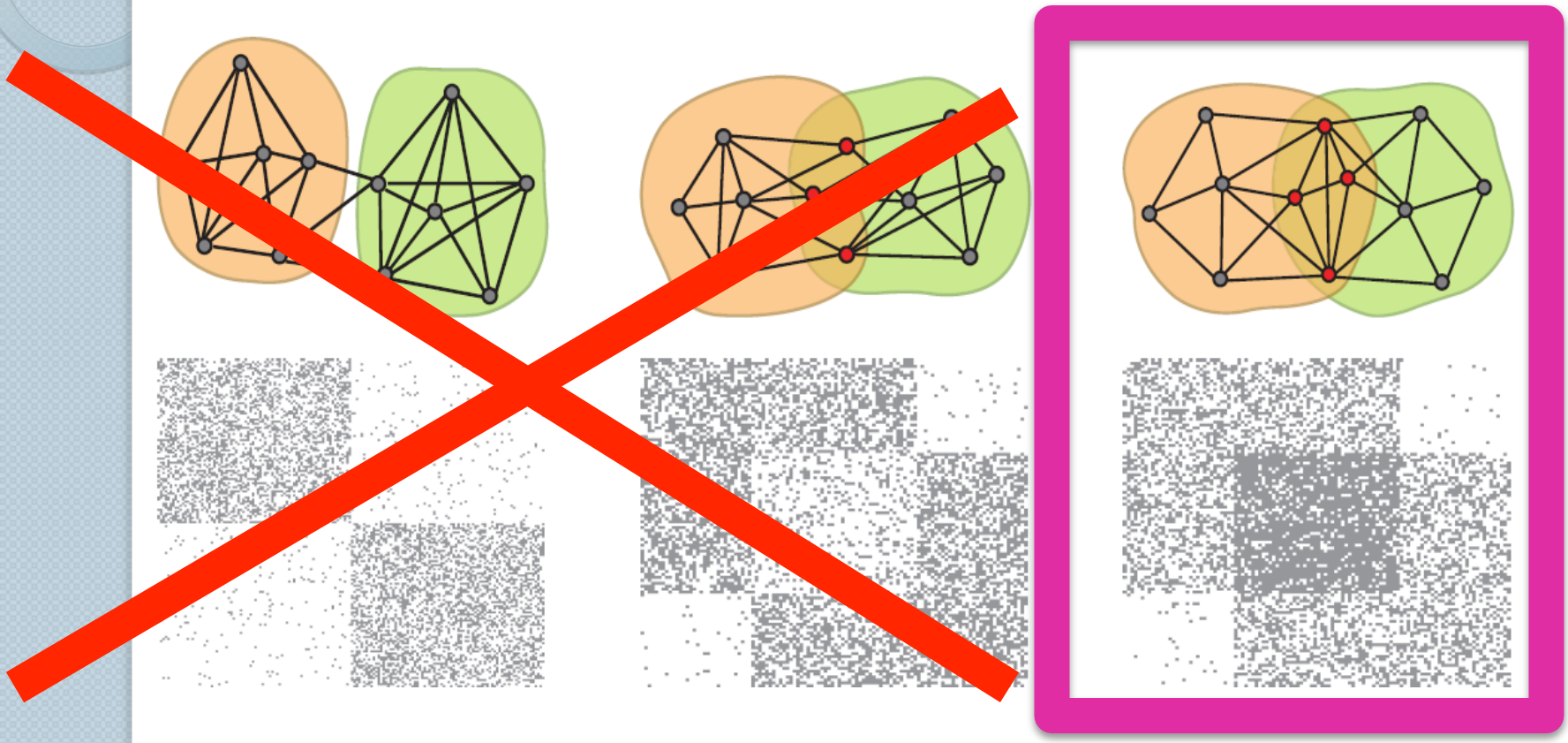
Edge Probability

- Nodes *u* and *v* share *k* groups
- **What is edge prob. P(edge | k) as a func. of k?**



Overlaps are DENSER!

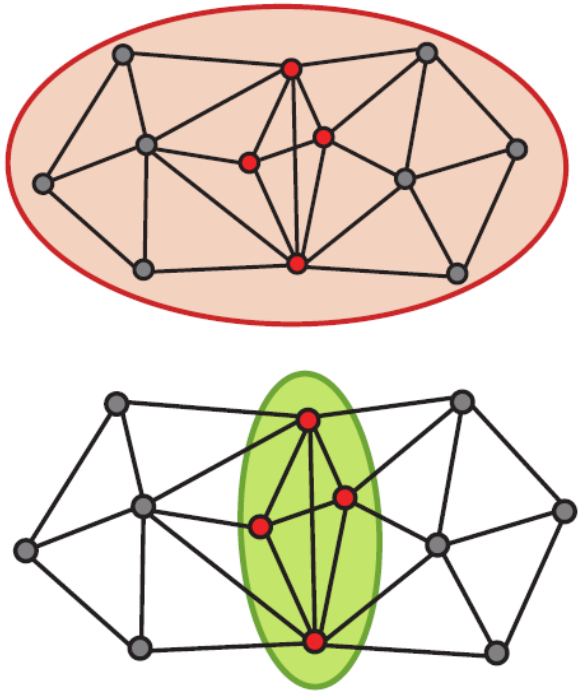
Communities in Networks

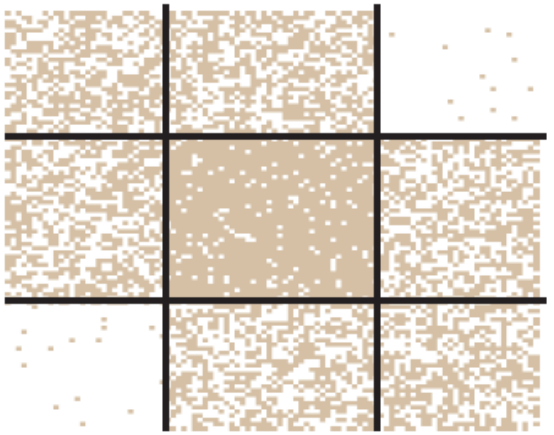


DOES IT MATTER?

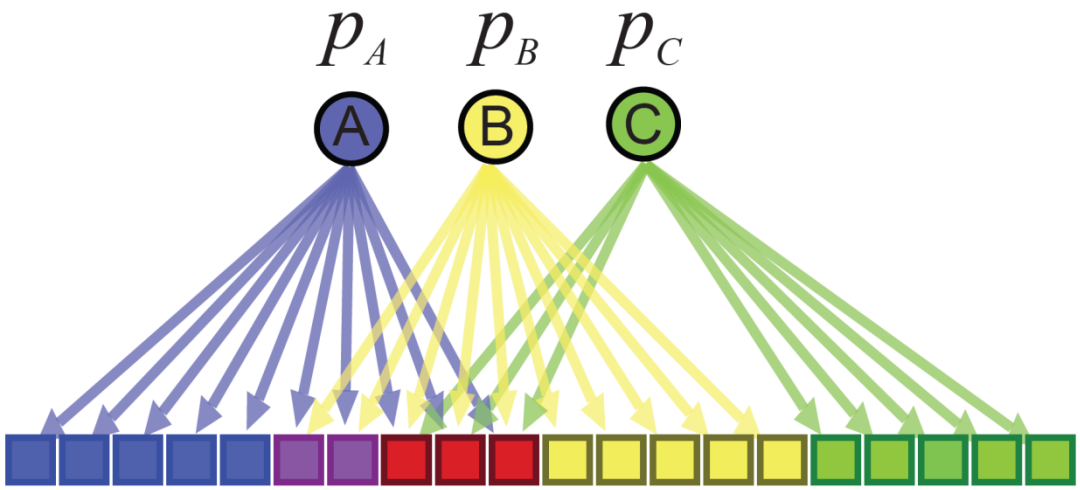
Detecting Dense Overlaps

- Can present community detection methods detect dense overlaps? **No!**





Natural Model



Communities C

Memberships M

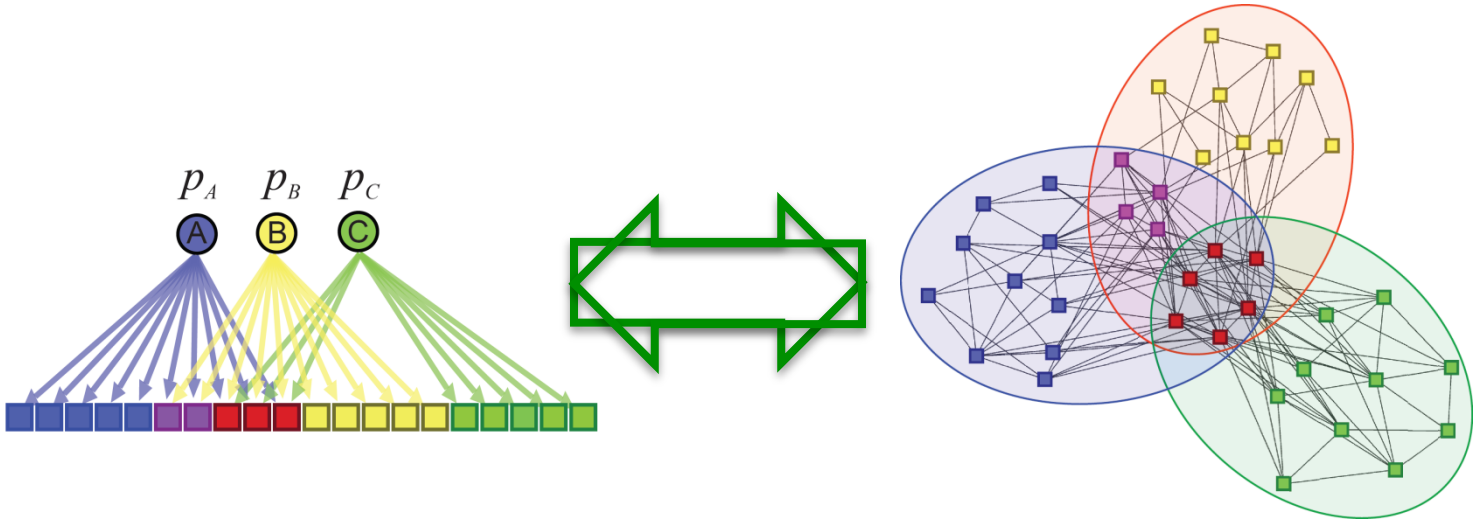
Nodes V

$$P(i,j) = 1 - \prod_{c \in M_i \cap M_j} (1 - p_c)$$

Community-Affiliation Graph Model

Provably generates power-law degree distributions and other patterns real-world networks exhibit. [Lattanzi, Sivakumar, STOC '09]

Model-based Community Detection

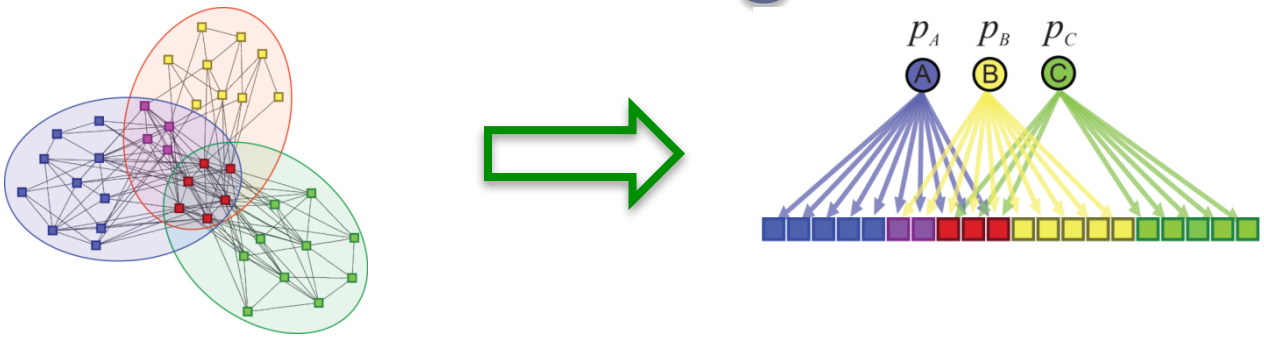


Given a Graph, find the Model

1) Affiliation Graph B(V,C,M)
2) Number of communities
3) Parameter p_i

Yes, we can!

AGM Model Fitting



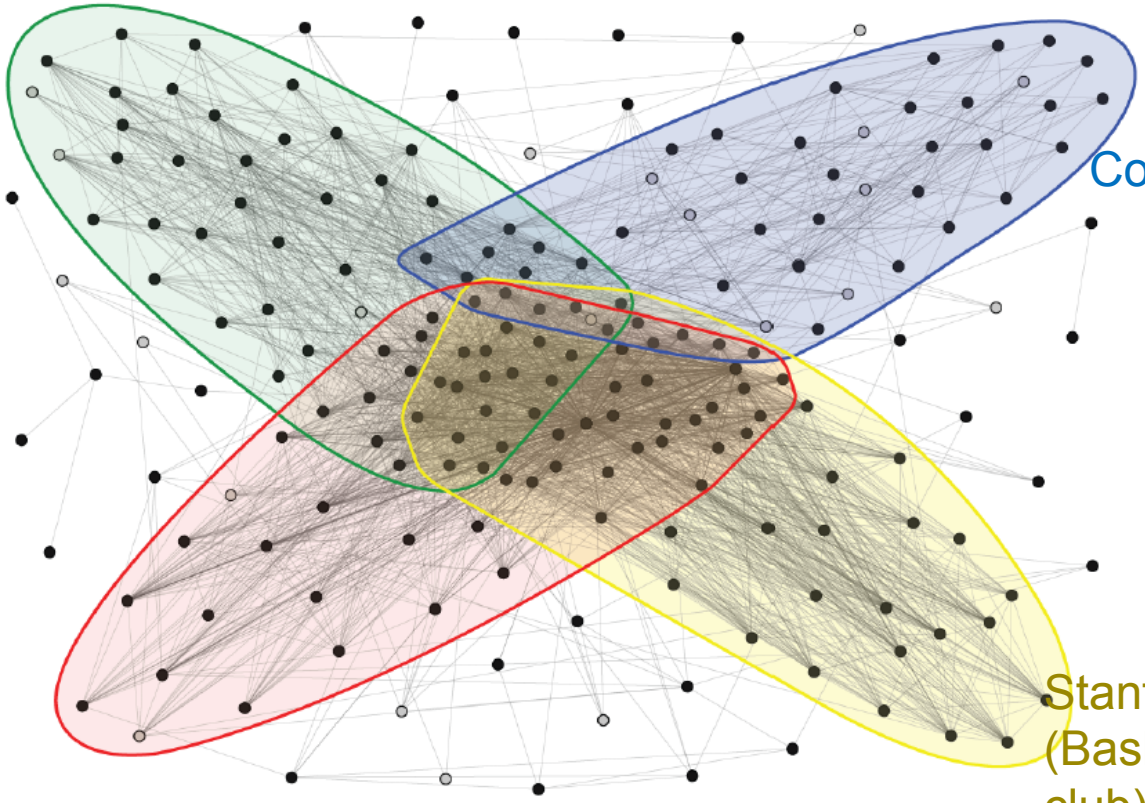
- **Task:**
  - Given network G(V,E), Find B(V,C,M) and $\{p_c\}$
- **Optimizing Likelihood (MLE)**

$$\arg\max_B P(G \mid B) = \prod_{(i,j)\in E} P(i,j) \prod_{(i,j)\notin E} (1 - P(i,j))$$

$$P(i,j) = 1 - \prod_{c \in M_i \cap M_j} (1 - p_c)$$

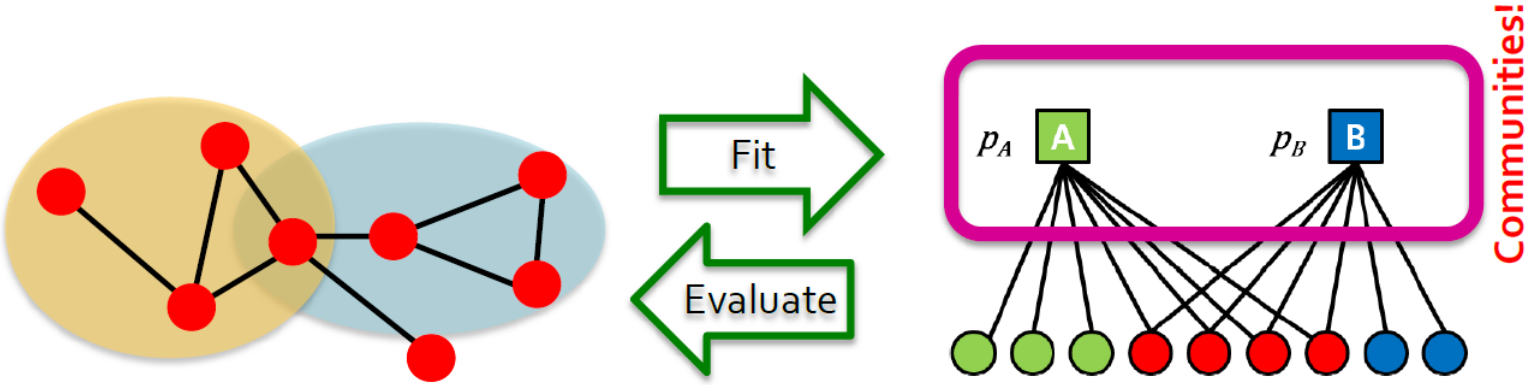
- **How to solve?**
  - Approach: Coordinate ascent
    - (1) Stochastic search over B, while keeping $\{p_c\}$ fixed
    - (2) Optimize $\{p_c\}$, while keeping B fixed (convex!)
  - Works well in practice!

Facebook example



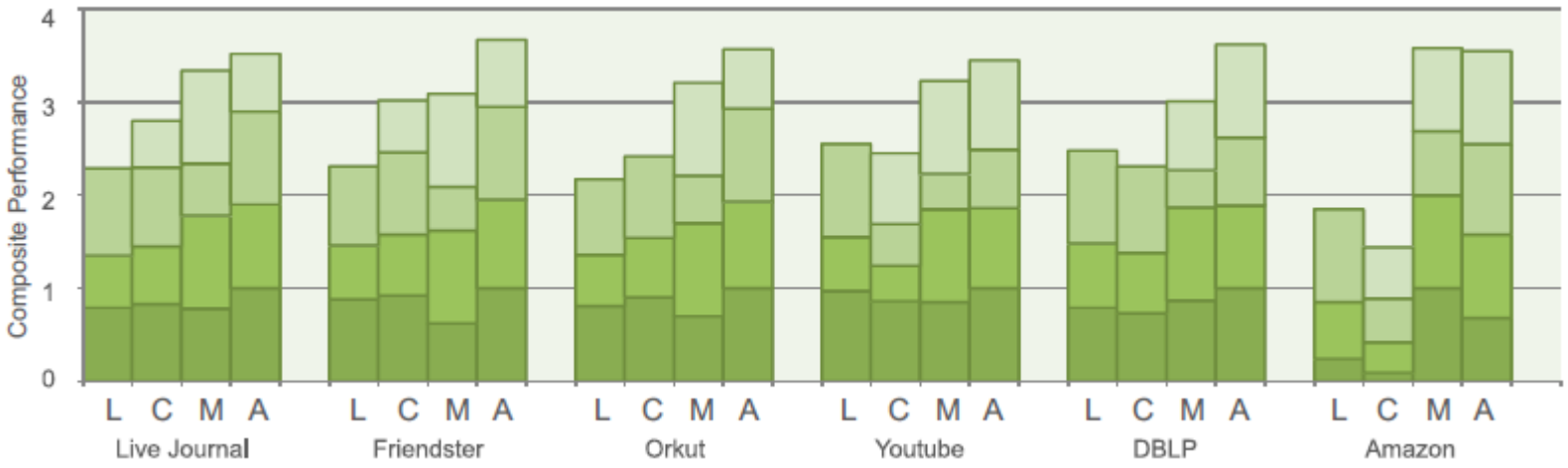
Accuracy: 89%

Experimental Setup



- **Evaluation:**
  - F-score: Precision, Recall
  - Mutual Information [Lancichinetti&Fortunato, PR-E '09]
  - Ω - index [Gregory, J of Stat. Mech. '11]
  - The number of communities
- **Methods for comparison:**
  - Clique Percolation [Palla et al., Nature '05]
  - Link Clustering [Ahn et al., Nature '10]
  - Mixed Membership Stochastic Blockmodels [Airoldi et al., JMLR '08]

Experimental Results: Ground-Truth



- **Overall (only overlaps) AGM improves (F1≈0.6)**
  - 57% (21%) over Link clustering
  - 48% (22%) over CPM
  - 10% (26%) over MMSB

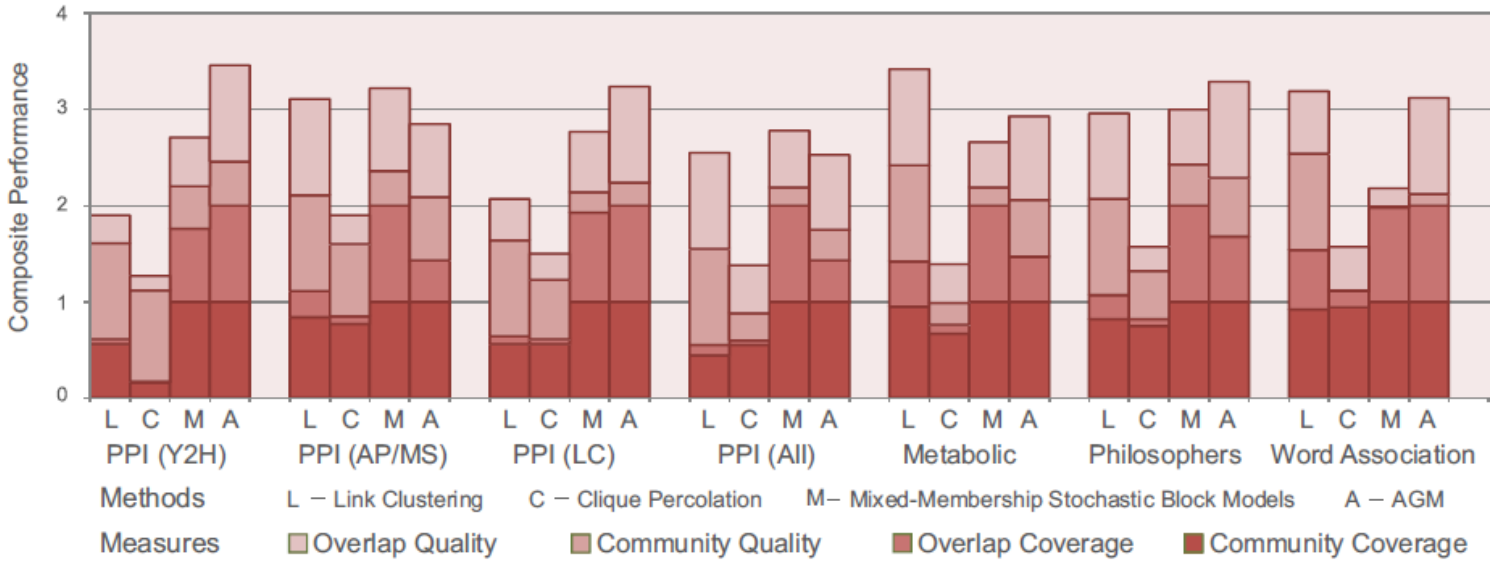
Methods

L – Link Clustering
C – Clique Percolation
M – Mixed-Membership Stochastic Block Model
A – AGM

Measures

Number of Communities
Normalized Mutual Information
F1-score
Ω-index

Experimental Results: Meta data-based



- Evaluation based on node metadata [Ahn et al. '10]
- Similar level of improvement

Conclusion

- **Ground-Truth Communities**
  - ⇒ Overlaps are **denser**
  - Present methods can't detect such overlaps
- **Community-Affiliation Graph Model**
  - ⇒ Model-based Community Detection
  - **Outperforms state-of-the-art**

References

- J. Yang, J. Leskovec. Structure and Overlaps of Communities in Networks. http://arxiv.org/abs/1205.6228
- J. Yang, J. Leskovec. *Defining and Evaluating Network Communities based on Ground-truth.* http://arxiv.org/abs/1205.6233
- J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney. *Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters.* http://arxiv.org/abs/0810.1355

Thank you!

- Code & Data: http://snap.stanford.edu