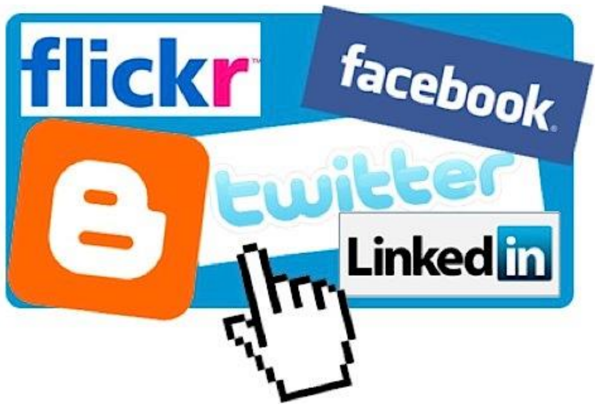
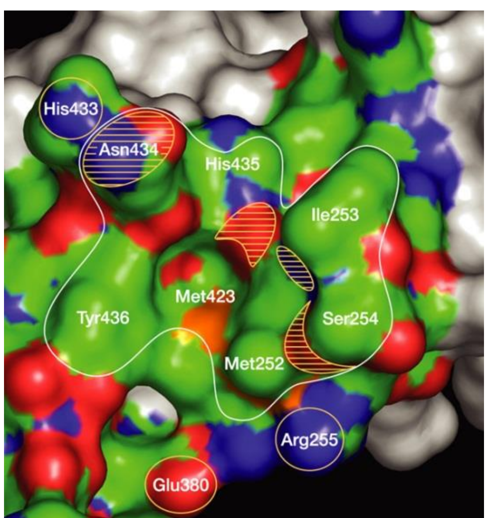
Community Detection in Graphs through Correlation

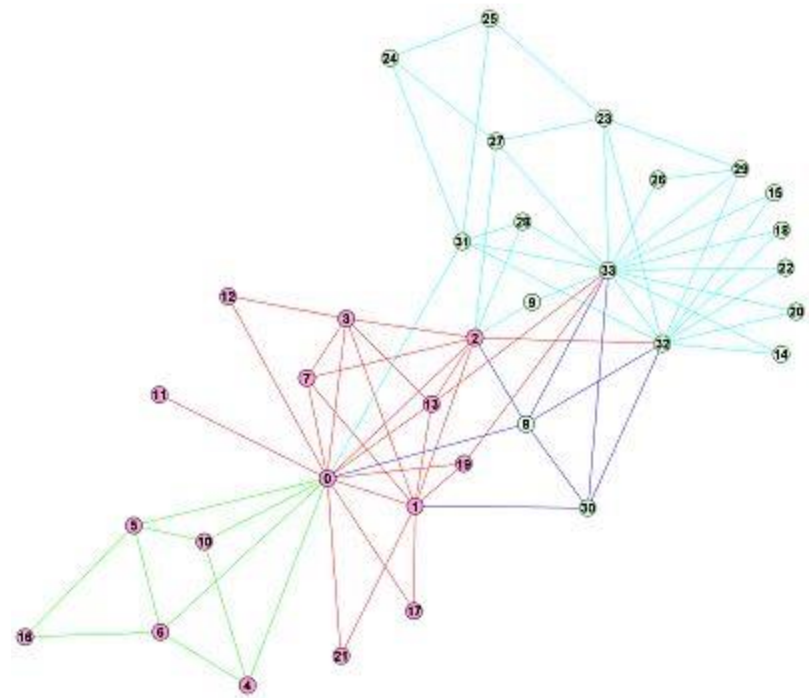
Lian Duan, New Jersey Institute of Technology
W. Nick Street, University of Iowa
Yanchi Liu, New Jersey Institute of Technology
Haibing Lu, Santa Clara University

Community Detection

- What and why?







Related Work

- Methods
  - Cut-based (M. Girvan and M. E. J. Newman, 2002)
  - Spectral-based (U. Luxburg, 2007)
  - Density-based (S. Mancoridis, et al., 1998)
  - Modularity-based (A. Clauset, et al.,2004)
  - Statistical-inference-based (M. E. J. Newman, 2013)

- Goal
  - More connections inside each community
  - Fewer connections across different communities

Opportunities for improvement

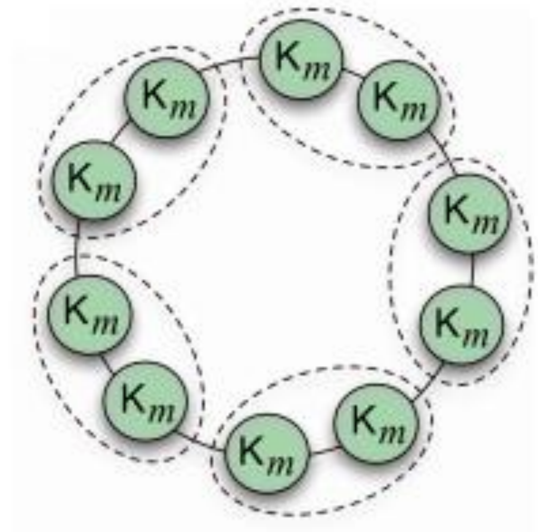
- Feature selection
  - Spectral-based

- Objective function
  - Cut-based
  - Modularity-based (Our focus in this paper)
  - Statistical-inference-based

- Search procedure
  - Greedy search
  - EM
  - Simulated annealing

Major problem of modularity

- Resolution problem
  (Lancichinetti & Fortunato 2011)
  - K_m is an m-clique
  - The detected communities are marked by circles with dash lines.



- Multi-resolution
  (Reichardt & Bornholdt 2006)
  - Further divide each detected community
  - Bias (Xiang et al. 2012)
    - Merge small communities
    - Split large communities

Connection with itemset search

- Graph communities: number of internal edges is greater than expected under assumption of random partition

- Correlated itemsets: occur more than expected under the assumption of item independence

- Connection: modularity = leverage

Correlated Itemsets (Duan, et al. 2014)

Given: itemset $S = \{I_1, I_2, \dots, I_m\}$ with m items in a dataset with n transactions

- True probability: $tp_s = P(S)$
- Expected probability: $ep_s = \prod_{i=1}^{m} P(I_i)$
- Correlation measure: $M_s = f(tp_s, ep_s)$
  - Chi-square: $\frac{(tp_s - ep_s)^2}{ep_s}$
  - Probability ratio / Lift: $\frac{tp_s}{ep_s}$
  - Leverage: $tp_s - ep_s$
  - Likelihood ratio: $\frac{tp_s^{tp_s} * (1 - tp_s)^{1 - tp_s}}{ep_s^{tp_s} * (1 - ep_s)^{1 - tp_s}}$

Correlated itemset example

t1: Beef, Chicken, Milk
t2: Beef, Cheese
t3: Cheese, Boots
t4: Beef, Chicken, Cheese
t5: Beef, Chicken, Clothes, Cheese, Milk

- For the itemset {Beef, Chicken}
  - $tp = \frac{3}{5}, ep = \frac{4}{5} * \frac{3}{5}, Leverage = tp - ep = \frac{3}{25}$

Modularity Function

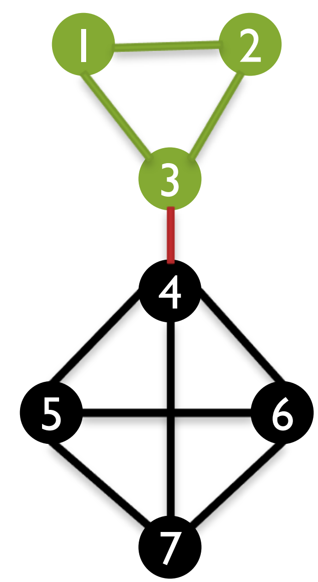
$$Q = \frac{1}{2m}\sum\nolimits_{i\in G, j\in G}\left(w_{ij} - \frac{k_i * k_j}{2m}\right) * \delta(v_i, v_j)$$

- n: the number of nodes
- m: the number of links
- w_{ij}: the edge between node i and j
- k_i: the degree of node i
- $\delta(v_i, v_j)$: the Kronecker delta function
  - $\delta(v_i, v_j)$=1 when v_i and v_j are in the same community
  - $\delta(v_i, v_j)$=0 otherwise

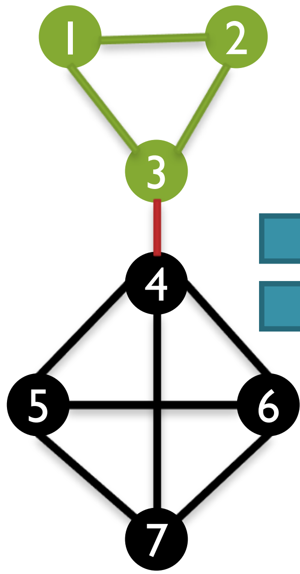
Transforming modularity function

For partition $\{G_1, G_2, \dots, G_l\}$ on graph G

- k_i: degree of node i
- $k_i^{internal}$: number of nodes in the same group of node i that connect to node i
- For the green partition:
  - $\sum_{i \in G_p} k_i = (2+2+3) = 7$
  - $\sum_{i \in G_p} k_i^{internal} = (2+2+2) = 6$
  - Total number of edges: $m = 10$



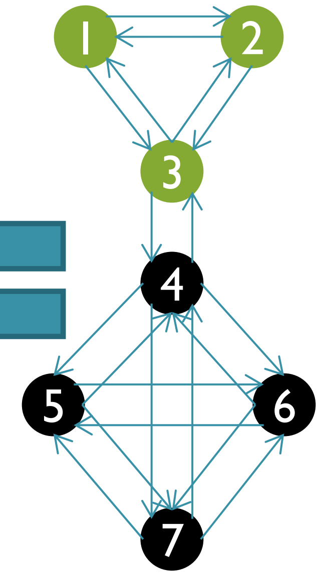
Transforming modularity function



=

	1	2	3	4	5	6	7
1	0	1	1	0	0	0	0
2	1	0	1	0	0	0	0
3	1	1	0	1	0	0	0
4	0	0	1	0	1	1	1
5	0	0	0	1	0	1	1
6	0	0	0	1	1	0	1
7	0	0	0	1	1	1	0

=



For the green partition:

(1) $\sum_{i \in G_p} k_i = (2+2+3) = 7$

(2) $\sum_{i \in G_p} k_i^{internal} = (2+2+2) = 6$

(3) Total number of edges: $m = 10$

The probability of the edge

(1) Both ends in the green partition: 6/20=$\sum_{i \in G_p} k_i^{internal} / 2\mathrm{m}$

(2) Started from the green partition: 7/20=$\sum_{i \in G_p} k_i / 2m$

(3) Ended in the green partition: 7/20=$\sum_{i \in G_p} k_i / 2m$

Transforming modularity function

If we randomly select an edge from the doubly-directed graph,

- The true probability of the edge in G_p:
  $$tp = \frac{\sum_{i \in G_p} k_i^{internal}}{2m}$$
- Probability the edge started from G_p: $\frac{\sum_{i \in G_p} k_i}{2m}$
- Probability the edge ended in G_p: $\frac{\sum_{j \in G_p} k_j}{2m}$

- The expected probability of the edge in G_p under the assumption of independence:
  $$ep = \frac{\sum_{i \in G_p} k_i}{2m} * \frac{\sum_{j \in G_p} k_j}{2m}$$

Transforming modularity function

For partition $\{G_1, G_2, \dots, G_l\}$ on graph G

- $Q = \frac{1}{2m} \sum_{i \in G, j \in G} \left(w_{ij} - \frac{k_i * k_j}{2m} \right) * \delta(v_i, v_j)$
- We define Q_p as the partial modularity for the group p where

$$Q_p = \frac{1}{2m} \sum_{i \in G_p, j \in G} \left(w_{ij} - \frac{k_i * k_j}{2m} \right) * \delta(v_i, v_j)$$

- $Q = \sum_{p=1}^{l} Q_p$

Transforming modularity function

- Partial modularity
  - $Q_p = \frac{\sum_{i \in G_p} k_i^{internal}}{2m} - \frac{\sum_{i \in G_p} k_i}{2m} * \frac{\sum_{j \in G_p} k_j}{2m}$

- If we randomly select an edge from the doubly-directed graph,
  - The true probability of the edge in G_p:

$$tp_p = \frac{\sum_{i \in G_p} k_i^{internal}}{2m}$$

  - Expected probability of the edge in G_p under the assumption of independence:

$$ep_p = \frac{\sum_{i \in G_p} k_i}{2m} * \frac{\sum_{j \in G_p} k_j}{2m}$$

Transforming modularity function

- Connecting correlation with modularity
  - For a given partition G_p, partial modularity $Q_p = tp_p - ep_p$
  - For a given itemset S, leverage = $tp_s - ep_s$

- For any correlation function $f(tp, ep)$, the partial modularity function can be changed accordingly.
  - Chi-square: $\frac{(tp_s - ep_s)^2}{ep_s}$
  - Probability ratio / Lift: $\frac{tp_s}{ep_s}$
  - Leverage: $tp_s - ep_s$
  - Likelihood ratio: $\frac{tp_s^{tp_s} * (1 - tp_s)^{1 - tp_s}}{ep_s^{tp_s} * (1 - ep_s)^{1 - tp_s}}$

Outline

- Basic concepts
  - Correlated itemset search
  - Modularity function
- Connecting modularity with leverage in correlated itemset search
- Upper bound analysis
- Evaluation

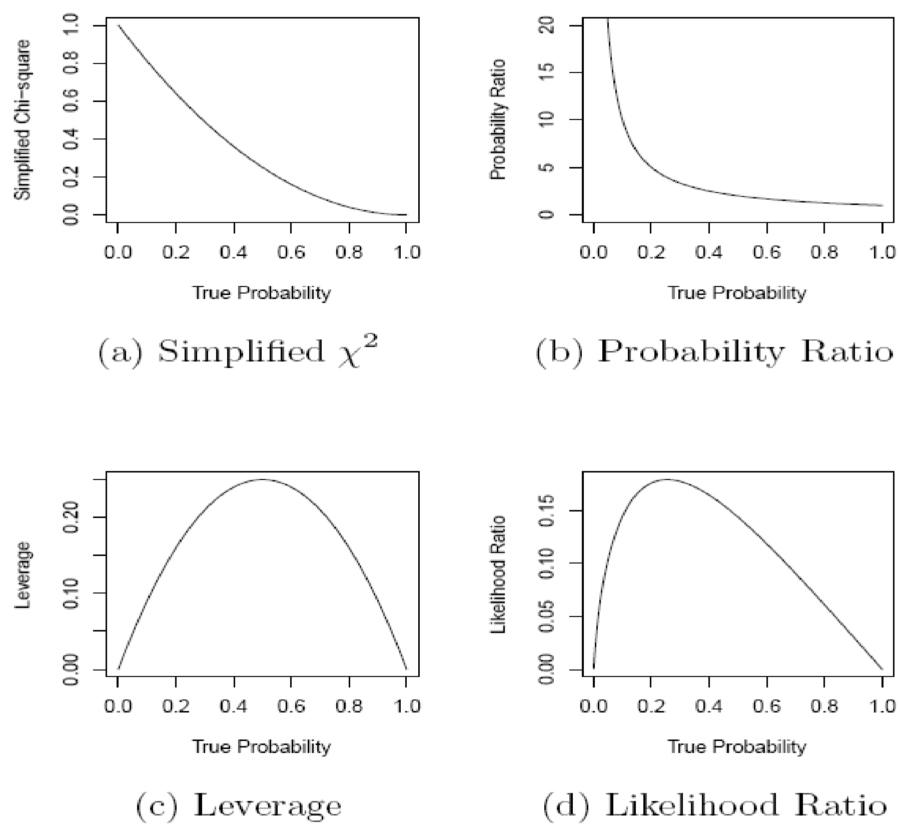
Upper bound analysis

- Correlation Property
  - The correlation function monotonically increases with the decrease of ep when tp remains the same.

- Understanding the bias to the community size
  - Given a group G_p with fixed $tp = \frac{\sum_{i \in G_p} k_i^{internal}}{2m}$
  - Partial modularity has the highest possible value when $ep = \frac{\sum_{i \in G_p} k_i}{2m} * \frac{\sum_{j \in G_p} k_j}{2m}$ reaches its lowest value tp^2

Upper bound analysis

- The highest possible partial modularity: $f(tp, ep = tp^2)$



Outline

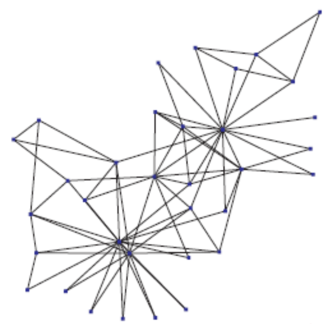
- Basic concepts
  - Correlated itemset search
  - Modularity function
- Connecting modularity with leverage in correlated itemset search
- Upper bound analysis
- Evaluation

Experiments

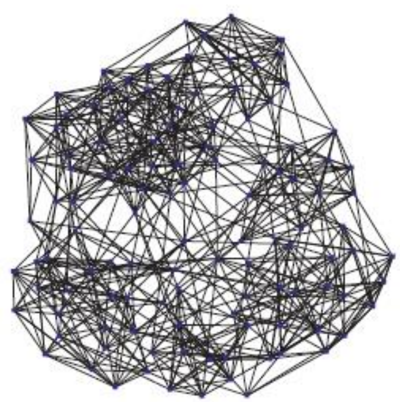
- Modify the objective function
- Greedy search (hierarchical clustering)
- Baseline:
  - Modularity-based methods (Leverage)
- Datasets:
  - Real life
  - Simulated with LFR model (Lancichinetti et al. 2008)
- Evaluation measures:
  - Rand Index (Rand1971), Jaccard, F-measure, Normalized mutual information (Danon 2005)

Real life datasets

- Karate club: two equal size communities

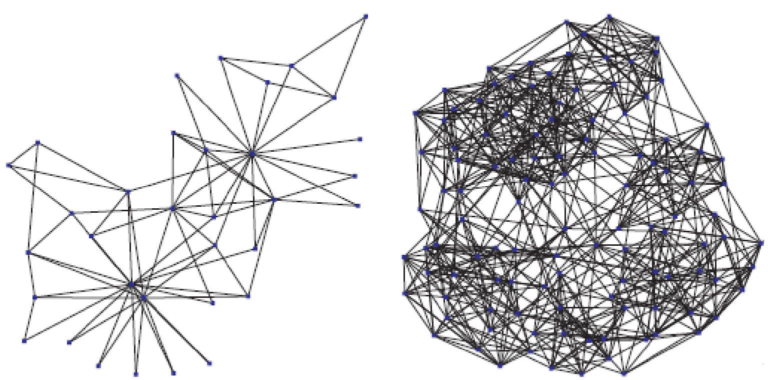


- College football: 12 equal size communities

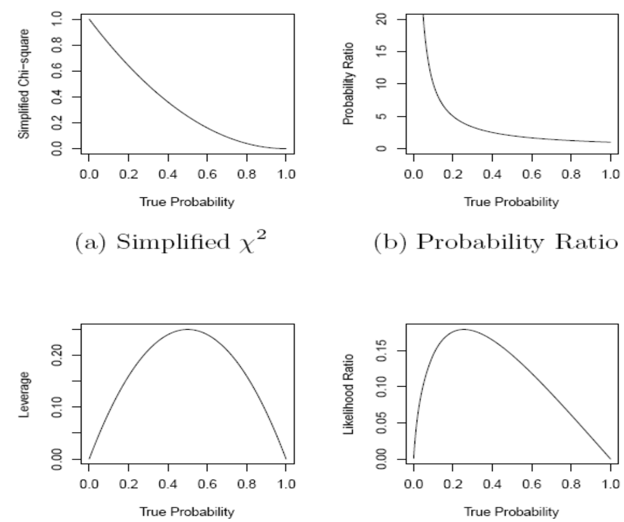


Real life datasets

Data Set	Measure	NMI	Jaccard	RI	F-measure	DNC	ANC
Karate	χ^2	0.4852	0.2842	0.6453	0.4426	7	2
	PR	0.3868	0.0945	0.5561	0.1728	14	2
	Leverage	**0.6925**	**0.6833**	**0.8414**	**0.8118**	3	2
	LR	0.5385	0.3958	0.6952	0.5671	5	2
Football	χ^2	**0.9141**	0.7571	0.9793	0.8618	14	12
	PR	0.6864	0.0829	0.9240	0.1531	55	12
	Leverage	0.6977	0.3622	0.8807	0.5317	6	12
	LR	0.9086	**0.7897**	**0.9812**	**0.8825**	12	12



(a) Karate (b) Football

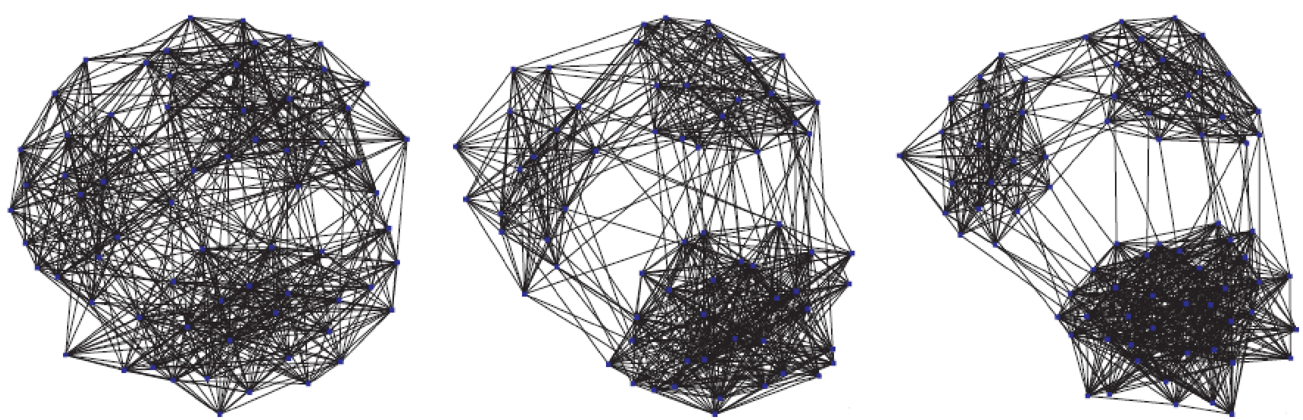


(a) Simplified χ^2 (b) Probability Ratio

(c) Leverage (d) Likelihood Ratio

Simulated datasets

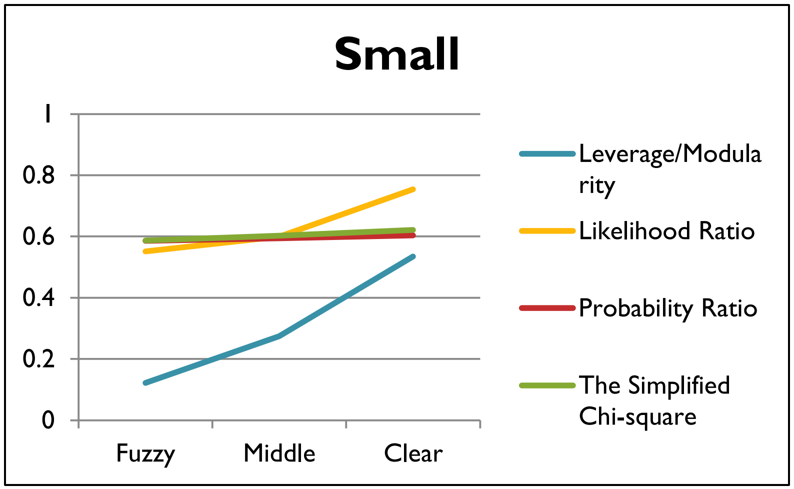
- Parameters:
  - Minimal community size: 50, 500, or 5000
  - Community structure ratio β: 5, 10, or 20

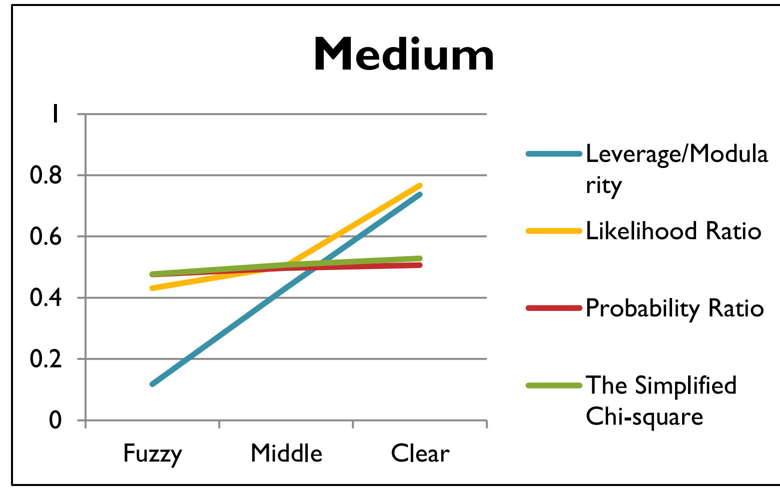


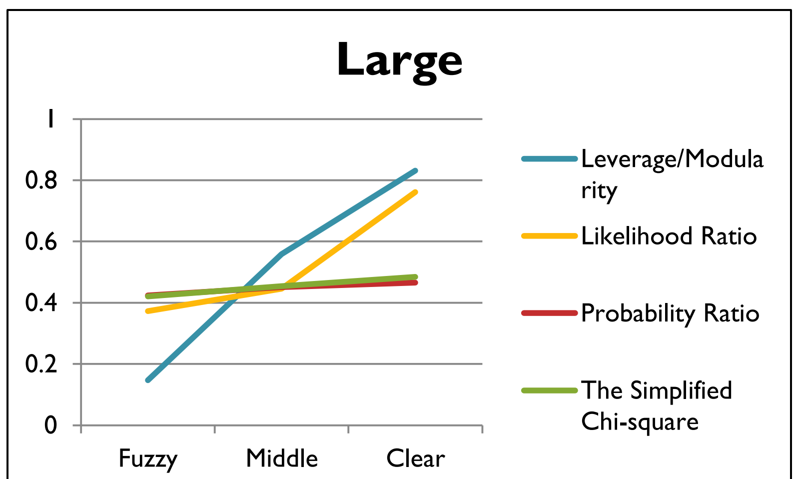
(a) β is 5 (b) β is 10 (c) β is 20

Experiments

- NMI when fixing Min-community-size

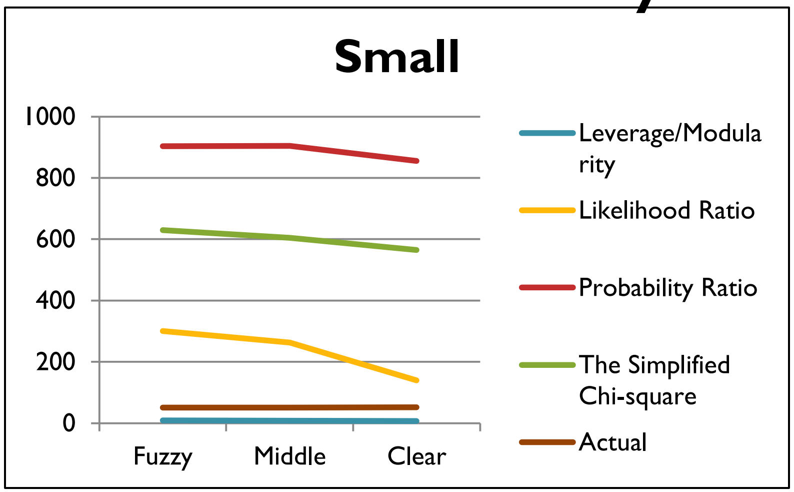


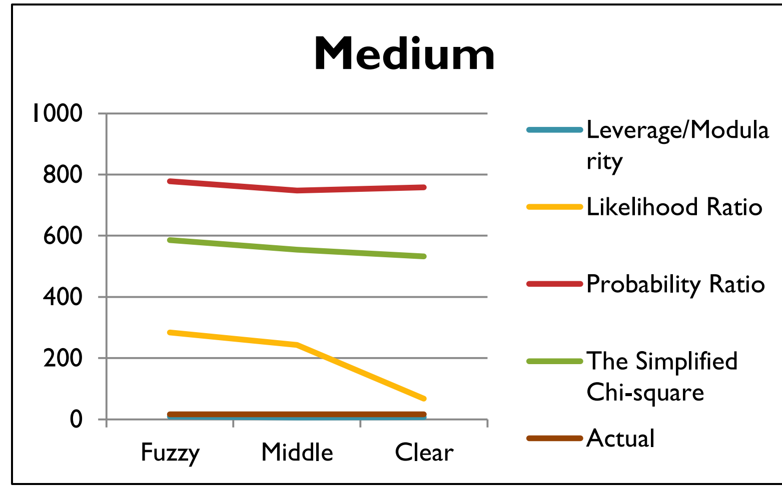


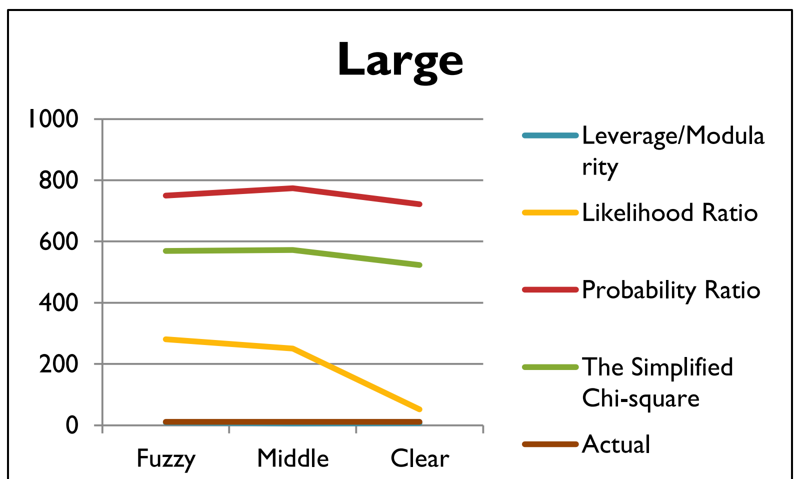


Experiments

- Number of detected groups when fixing Min-community-size







Summary

- Connection between community detection and correlation search
- Conclusion
  - Modularity is good only when there are large and clear communities
  - Likelihood ratio is robust to any type of communities
  - Probability ratio partitions the whole graph into small communities with 2 or 3 objects

References

1. A. Clauset, M. E. J. Newman, and C. Moore. (2004) Finding community structure in very large networks. Physical Review E, 70(6):066111+.
2. L. Danon, A. D. Guilera, J. Duch, and A. Arenas. (2005) Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, 2005(9):P09008–09008.
3. L. Duan, W. N. Street, Y. Liu, S. Xu and B. Wu. (2014) Selecting the right correlation measure for binary data. ACM Transactions on Knowledge Discovery from Data.
4. M. Girvan and M. E. J. Newman. (2002) Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12):7821–7826.
5. A. Lancichinetti, S. Fortunato, and F. Radicchi. (2008) Benchmark graphs for testing community detection algorithms. Phys. Rev. E, 78:046110.
6. A. Lancichinetti and S. Fortunato. (2011) Limits of modularity maximization in community detection. Phys. Rev. E, 84:066122.
7. U. Luxburg. (2007) A tutorial on spectral clustering. Statistics and Computing, 17:395–416.
8. S. Mancoridis, B. S. Mitchell, and C. Rorres. (1998) Using automatic clustering to produce high-level system organizations of source code. In Proc. 6th Intl. Workshop on Program Comprehension, pages 45–53.
9. M. E. J. Newman. (2013) Community detection and graph partitioning. CoRR, abs/1305.4974.
10. W. M. Rand. (1971) Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association, 66(336):846–850.
11. J. Reichardt and S. Bornholdt. (2006) Statistical mechanics of community detection. Phys. Rev. E, 74:016110.
12. J. Xiang, X. G. Hu, X.Y. Zhang, et al. (2012) Multi-resolution modularity methods and their limitations in community detection. The European Physical Journal B, Vol. 85, No. 10., pp. 1-10.