Ensemble Clustering for Graphs (ECG)

Valérie Poulin
François Théberge*
theberge@ieee.org

August 2019

Outline

Notation

Let $G = (V, E)$, $V = \{1, 2, \ldots, n\}$, undirected.

Edges can have weights $w(e) > 0$ for each $e \in E$, else consider all $w(e) = 1$.

Let $P_i = \{C_i^1, \ldots, C_i^{l_i}\}$ be a partition of V of size l_i.

We use $\mathbf{1}_{C_i^j}(v)$ to denote the indicator function for $v \in C_i^j$.

Wish list

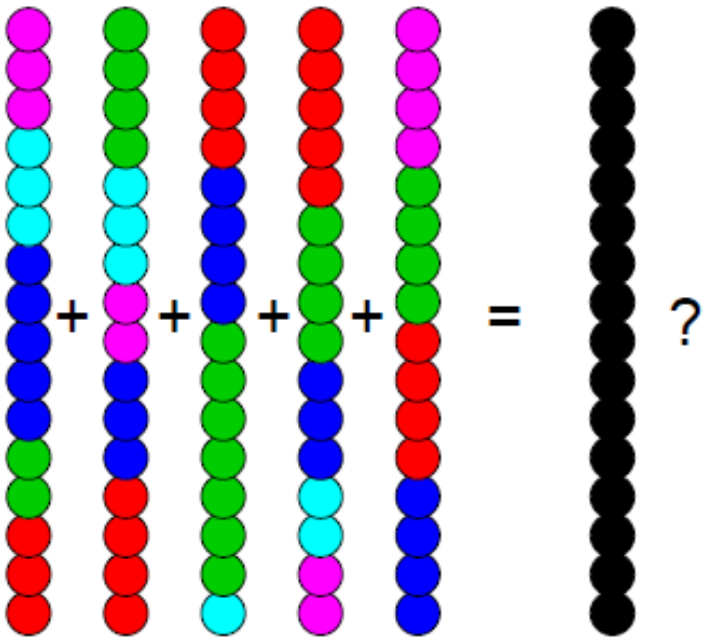
Wish list for practical graph clustering:

- "good", scalable, all-purpose algorithm
  - .. remember this is unsupervised learning!
- measures of association **strength**
- **hierarchy** of clusters
- **no** or minimal parameter **tuning**

} EDA tool for analysts

Ensemble learning

Ensemble learning is a powerful approach, but...
How to merge multiple graph partitions?



ECG Algorithm

The ECG algorithm is a consensus clustering algorithm for graphs. The steps are:

- generation step: k randomized level-1 partitions from Louvain (ML) algorithm: $\mathcal{P} = \{P_1, \ldots, P_k\}$.
- integration step: run ML on a re-weighted version of the initial graph $G = (V, E)$. The ECG weights are obtained through co-association.

ECG Algorithm

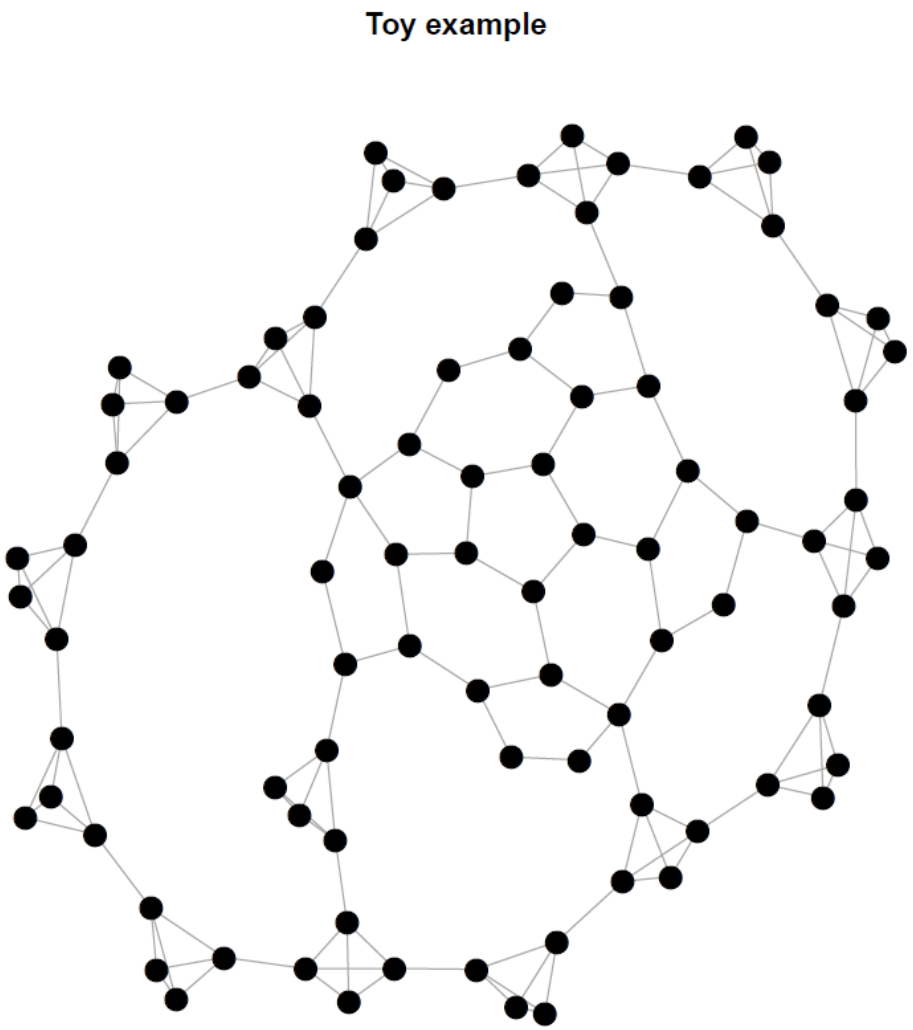
The ECG weight of an edge $e = (u, v) \in E$ is defined as:

$$W_{\mathcal{P}}(u,v) = \begin{cases} w_* + (1 - w_*) \cdot \left(\frac{\sum_{i=1}^{k} \alpha_{P_i}(u,v)}{k} \right), & (u,v) \in \text{2-core} \\ w_*, & \text{otherwise} \end{cases}$$

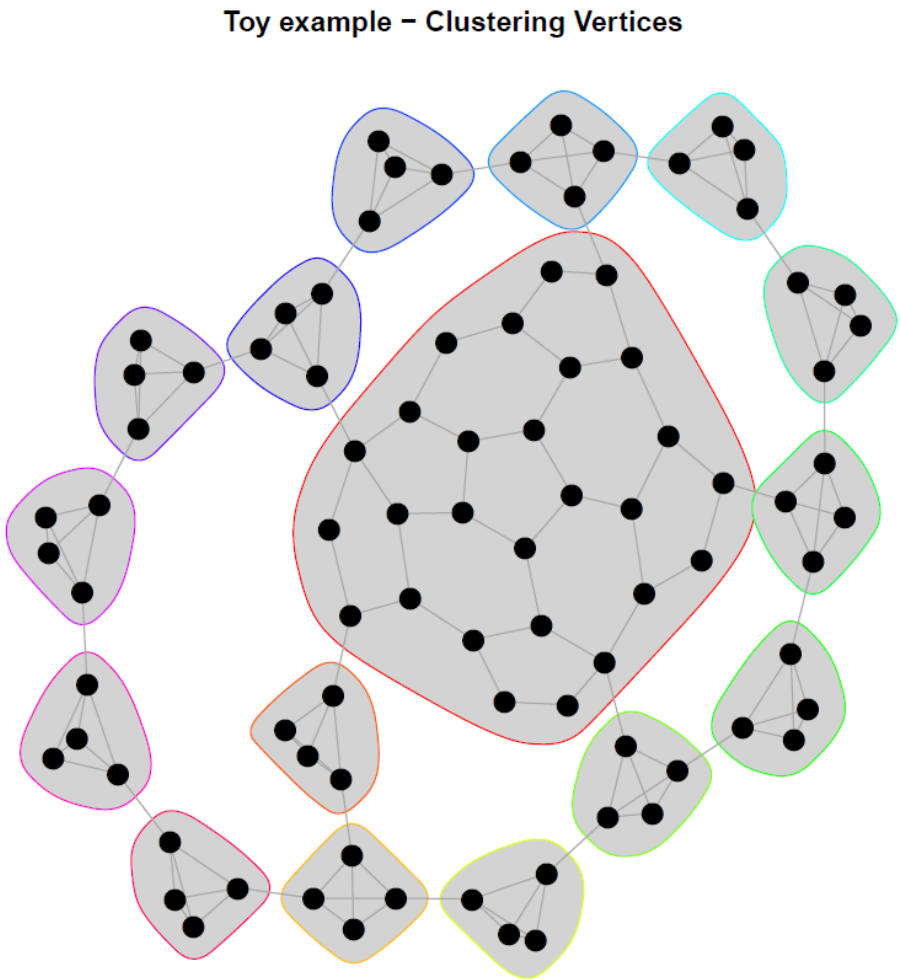
$0 < w_* < 1$ is some minimum weight

$\alpha_{P_i}(u,v) = \sum_{j=1}^{l_i} \mathbf{1}_{C_i^j}(u) \cdot \mathbf{1}_{C_i^j}(v)$ indicates co-occurrence in a cluster of P_i or not.

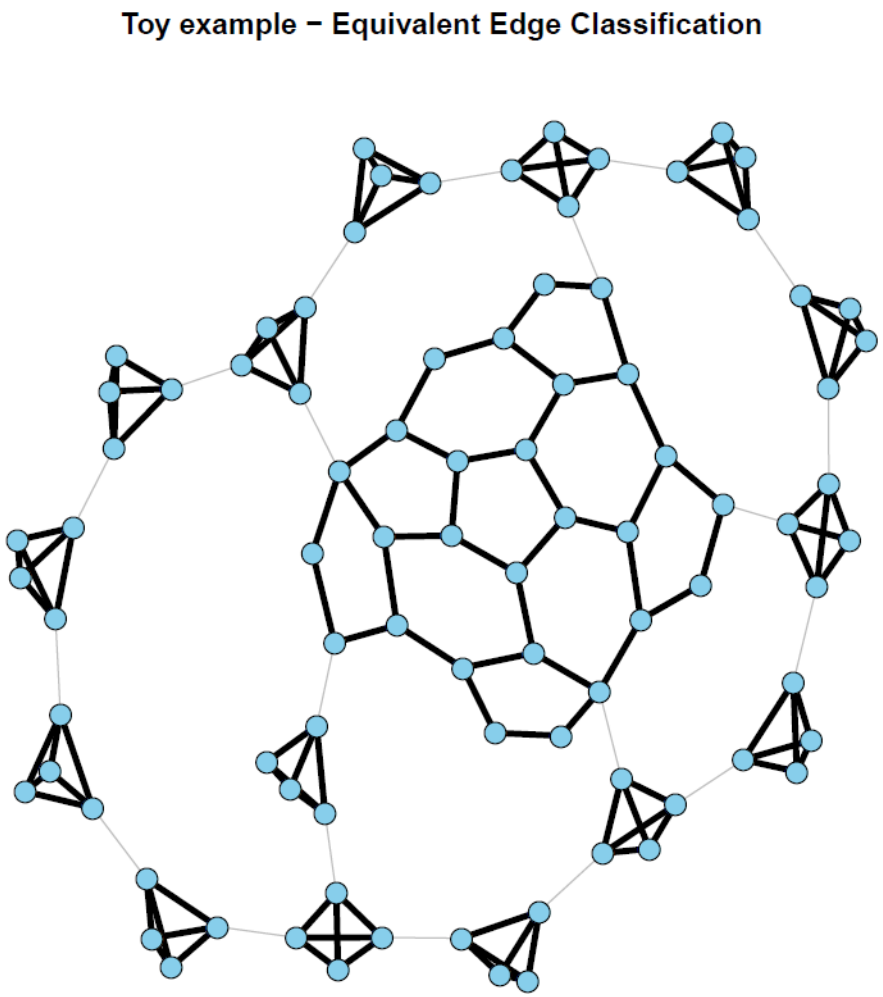
Toy graph example



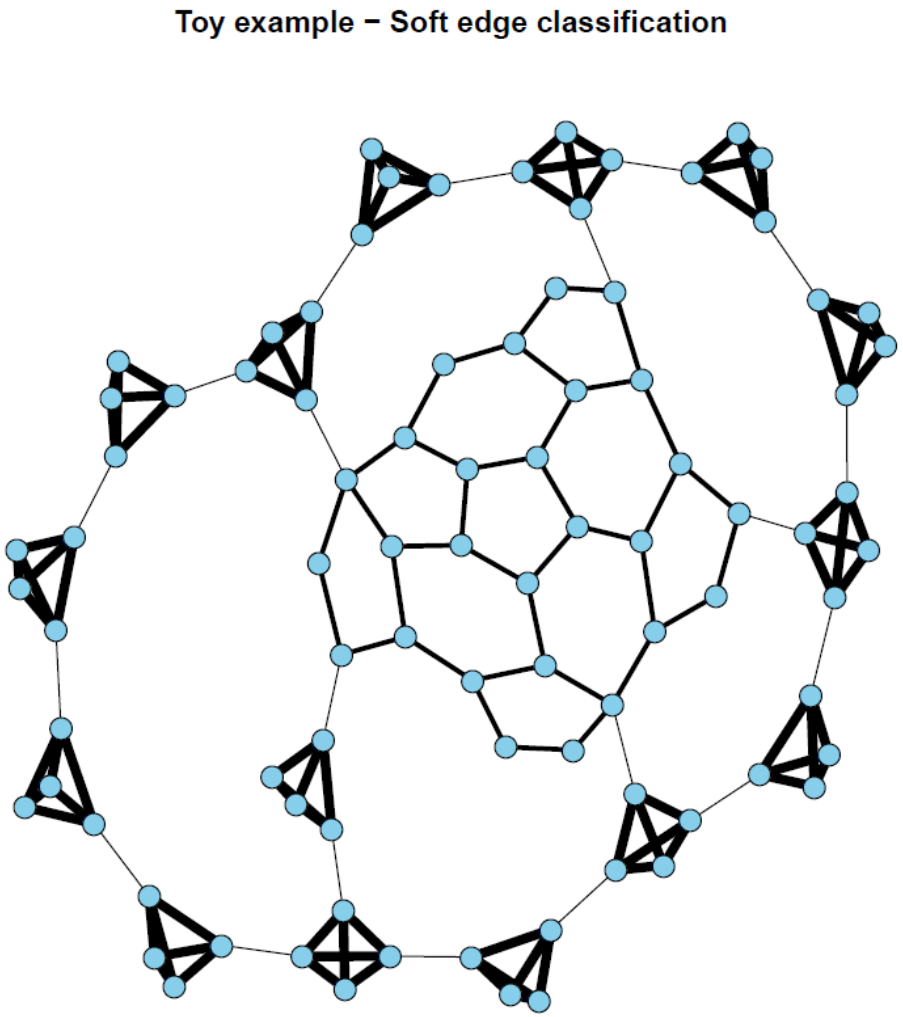
Toy graph example



Toy graph example

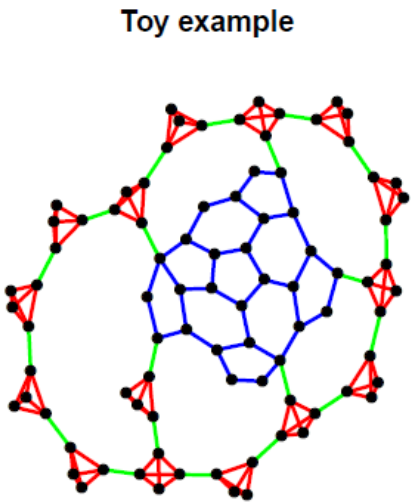


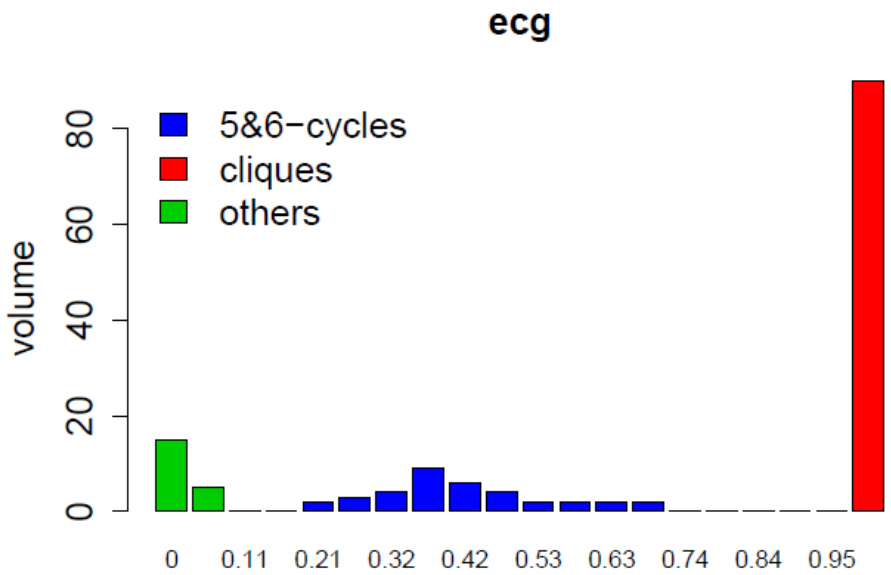
Toy graph example



Toy graph example

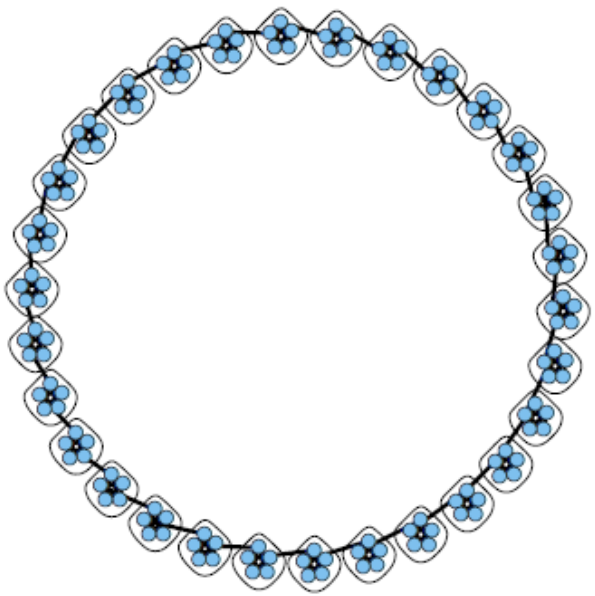
Resulting ECG weights:



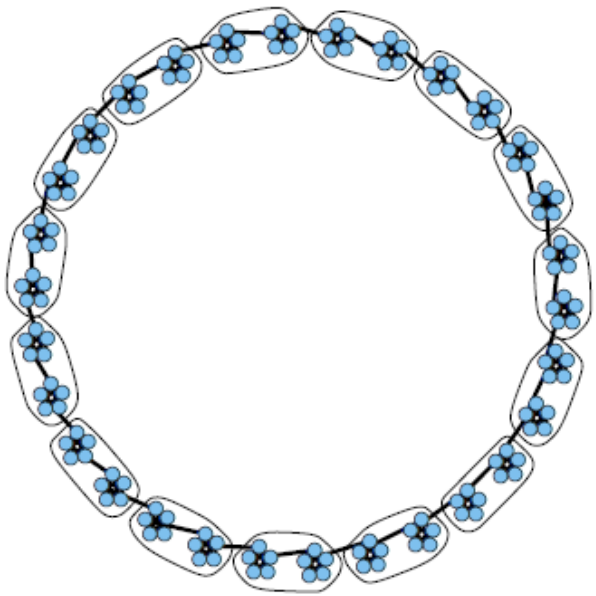


Resolution Issue

Modularity-based algorithm suffer from the resolution limit issue (ring of cliques example):



Q = 0.876



Q = 0.888

Resolution Issue

Consider a ring of l cliques of size m, with $n = l \cdot m$

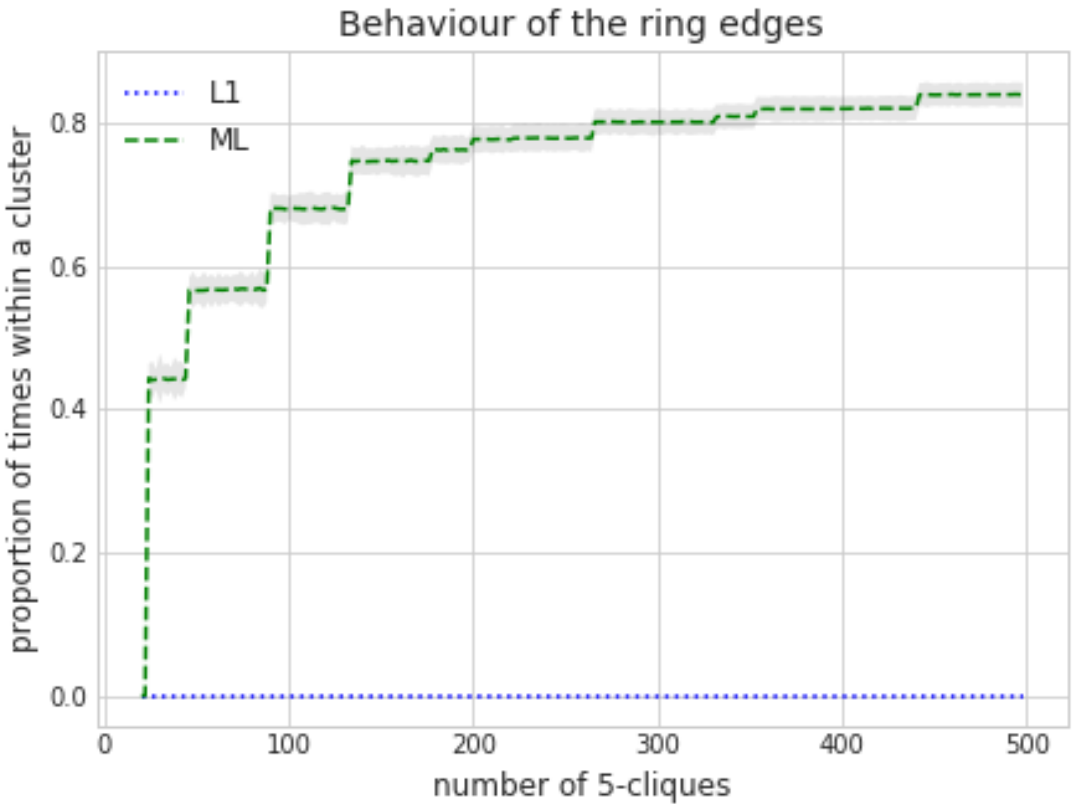
Grouping pairs of adjacent cliques yields a higher modularity than the natural choice of each clique forming its own cluster when $m(m-1) < l - 2$.

ECG with a small value for w_* can alleviate this issue. In particular, choosing $w_* < 1/n$ avoids the issue altogether.

ECG also helps w.r.t. generalized ring of cliques.

Resolution Issue

Using level-1 Louvain as weak leerners is the key:



Resolution Issue

Results over generalized ring-of-cliques:

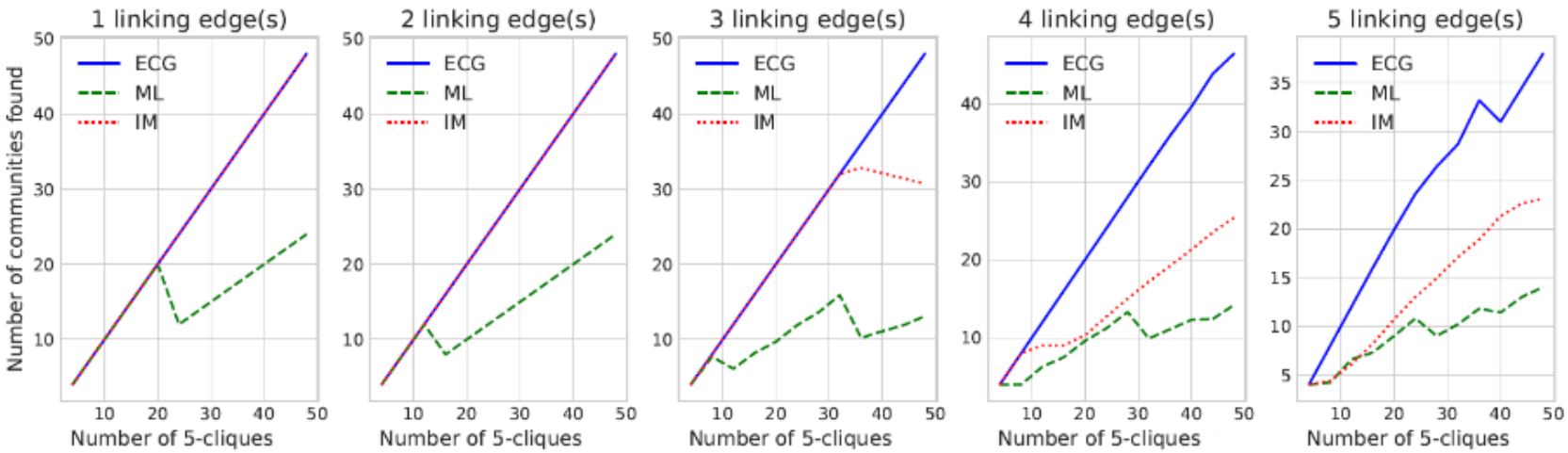


Figure 2 In each plot, we consider l cliques of size $m = 5$ where contiguous cliques are linked by 1 to 5 edges, respectively. We compare the number of communities found by the InfoMap (IM), Louvain (ML) and ECG algorithms. The resolution limit phenomenon is clearly seen with the ML algorithm. The IM algorithm fails to find the right number of communities when we increase the number of edges between the cliques, while ECG remains more stable.

Resolution Issue

Weights remain significant even with high noise:

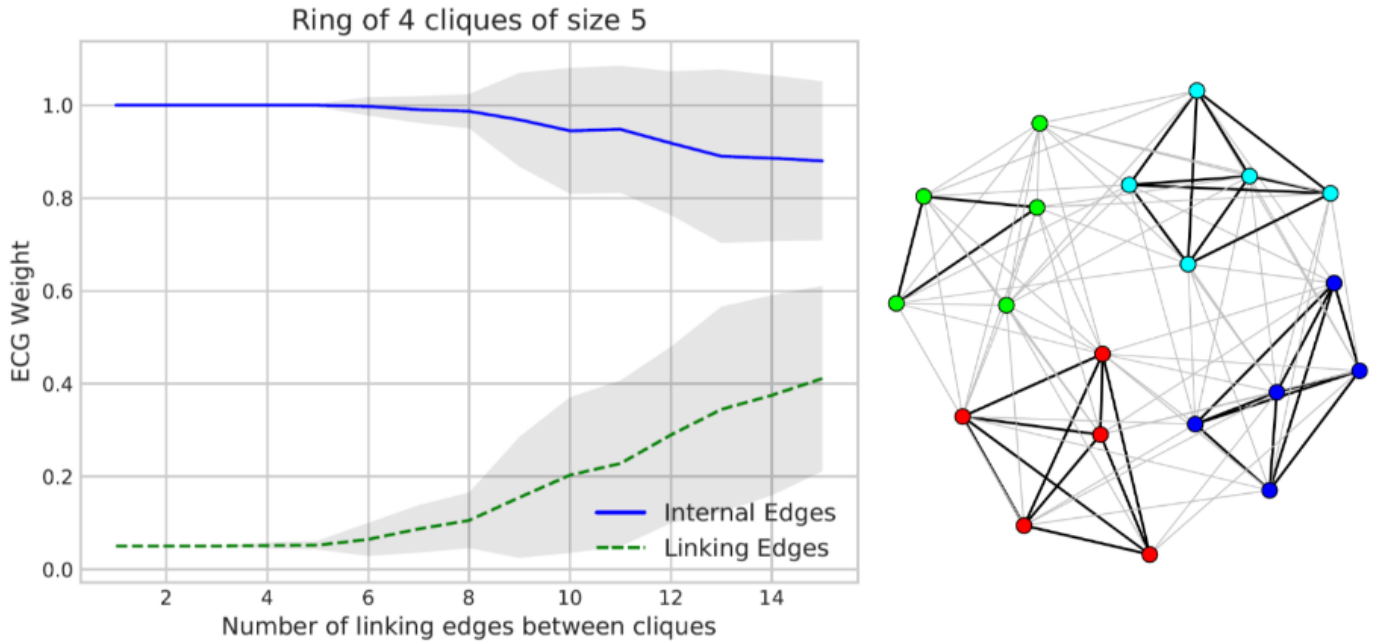
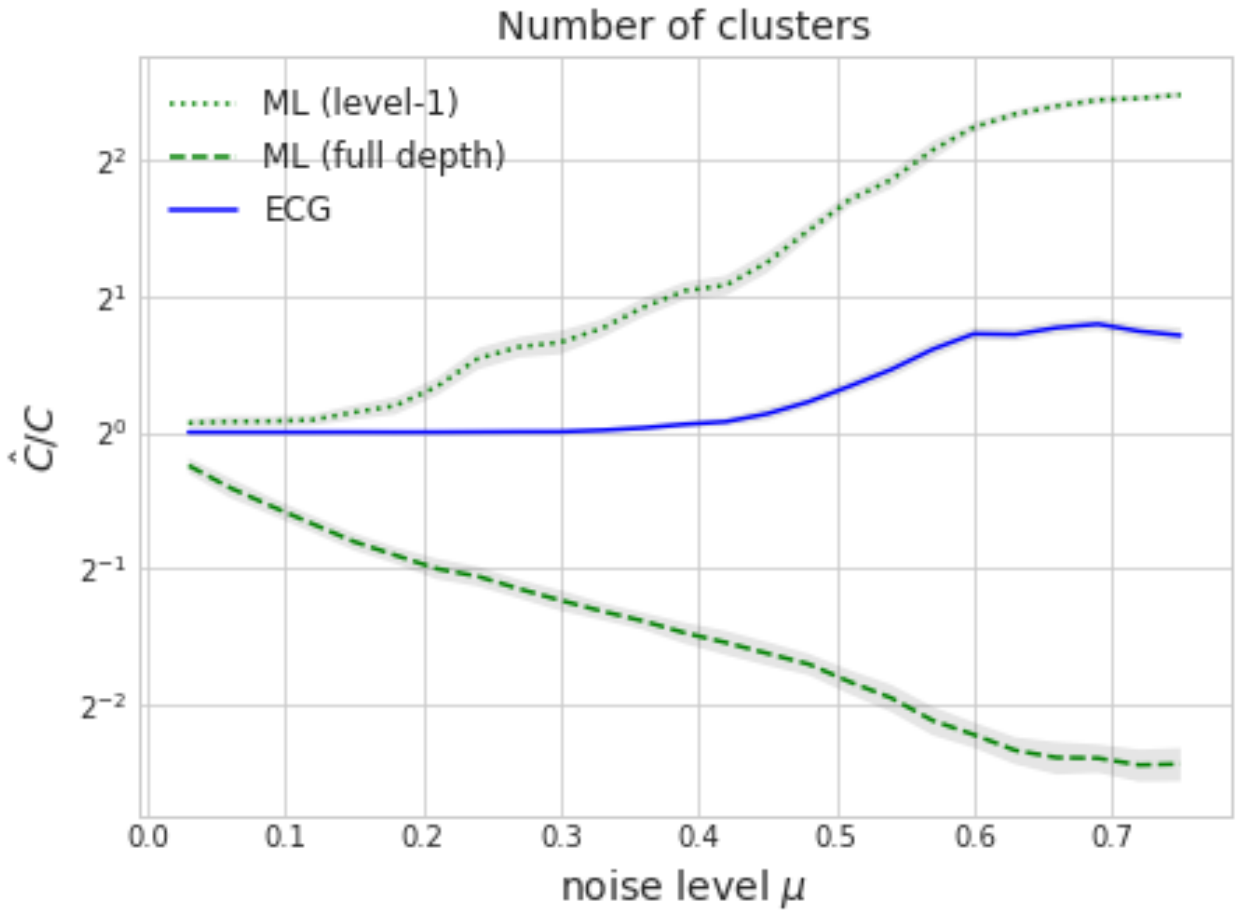


Figure 3 We add 1 to 15 edges between contiguous cliques in a ring of 4 cliques of size 5, and we look at the effect on the ECG edge weights for edges internal to the cliques, or external edges linking the contiguous cliques. In the right plot, we look at the case with 15 edges between cliques; thick edges are the ones where the ECG weight is 0.8 or above.

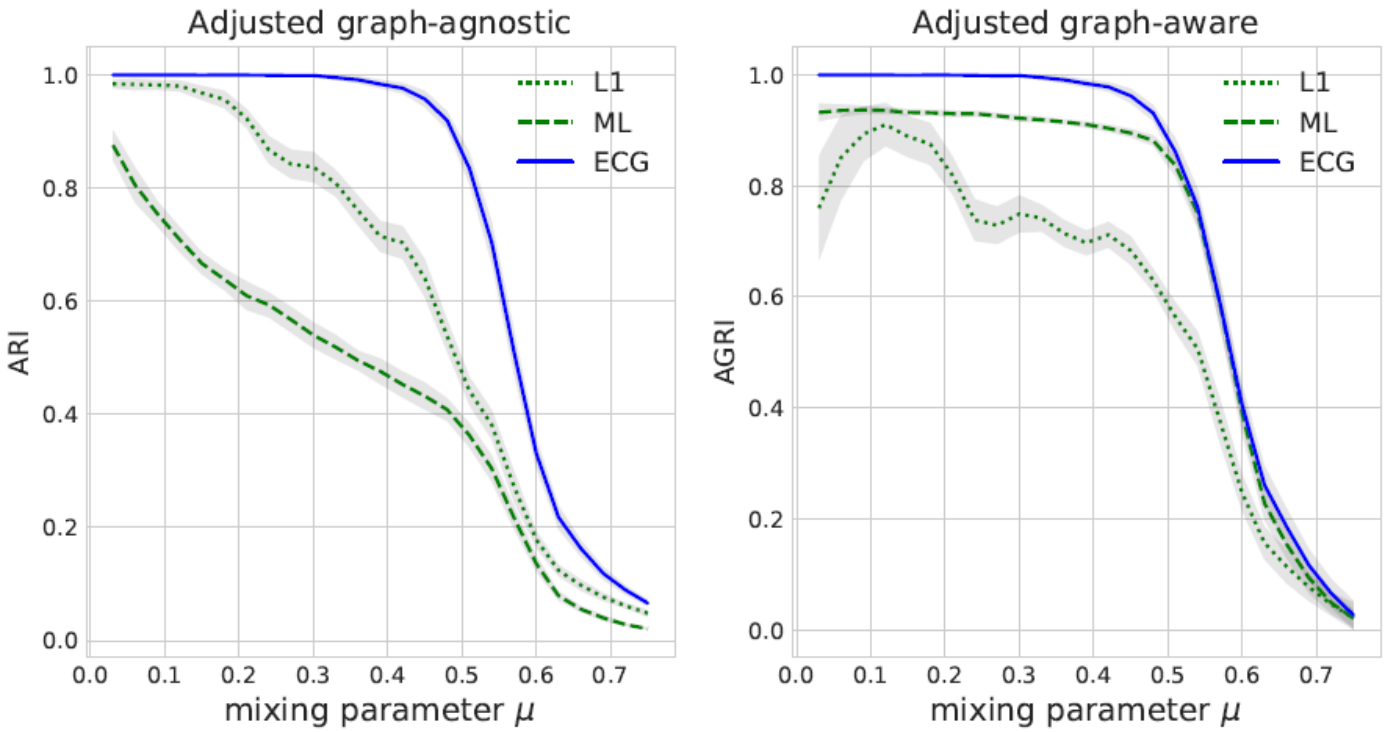
Resolution Issue

We also see the advantage of ECG by looking at the number of clusters found on some LFR graphs:



Resolution Issue

Same behaviour can be seen looking at graph-agnostic and graph-aware measures:



Stability

Different results can be obtained when re-running the same algorithm multiple times

This is a known issue with Louvain and other algorithms

We quantify stability by running each algorithm twice and applying some comparison measure such as ARI; this can be repeated multiple times for each graph

ECG shown a great improvement in stability vs Louvain

Stability

Stability results over LFR graphs:

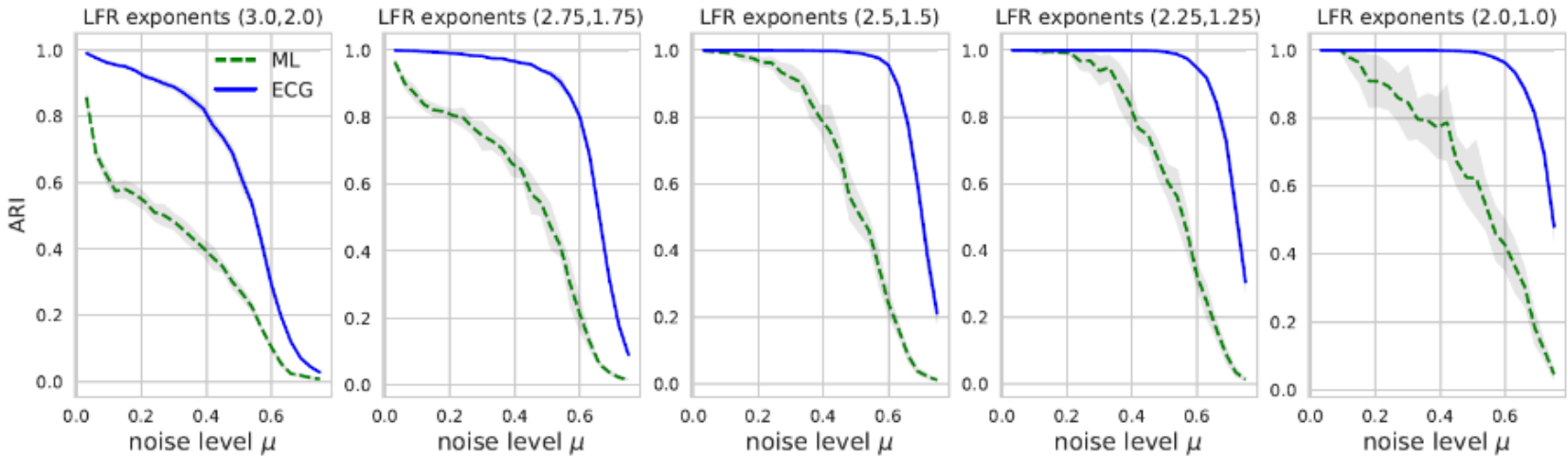
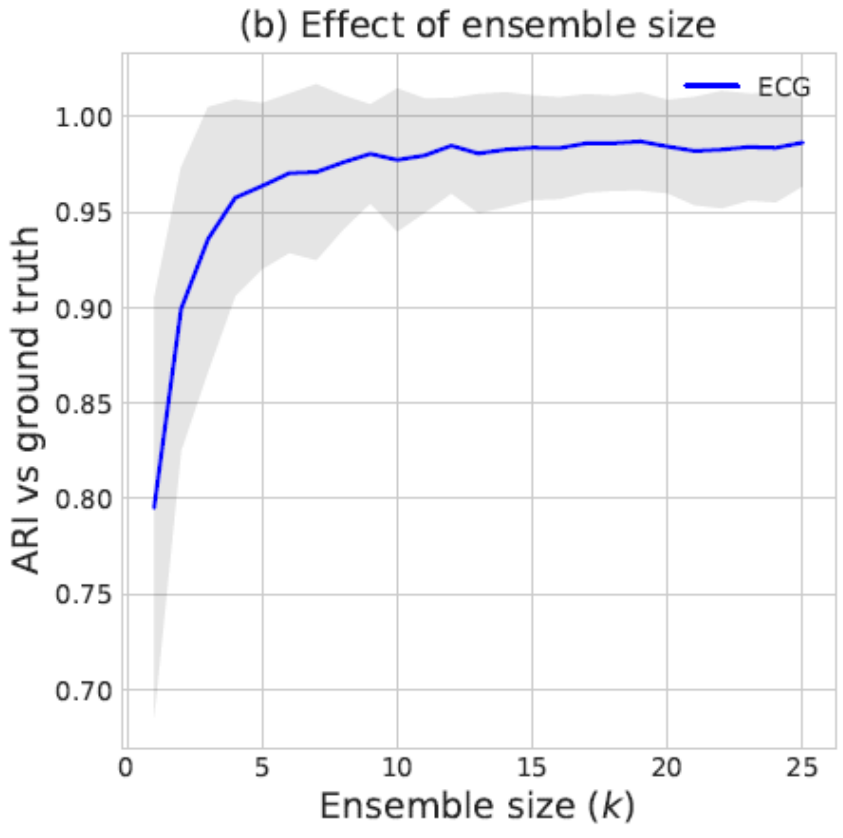
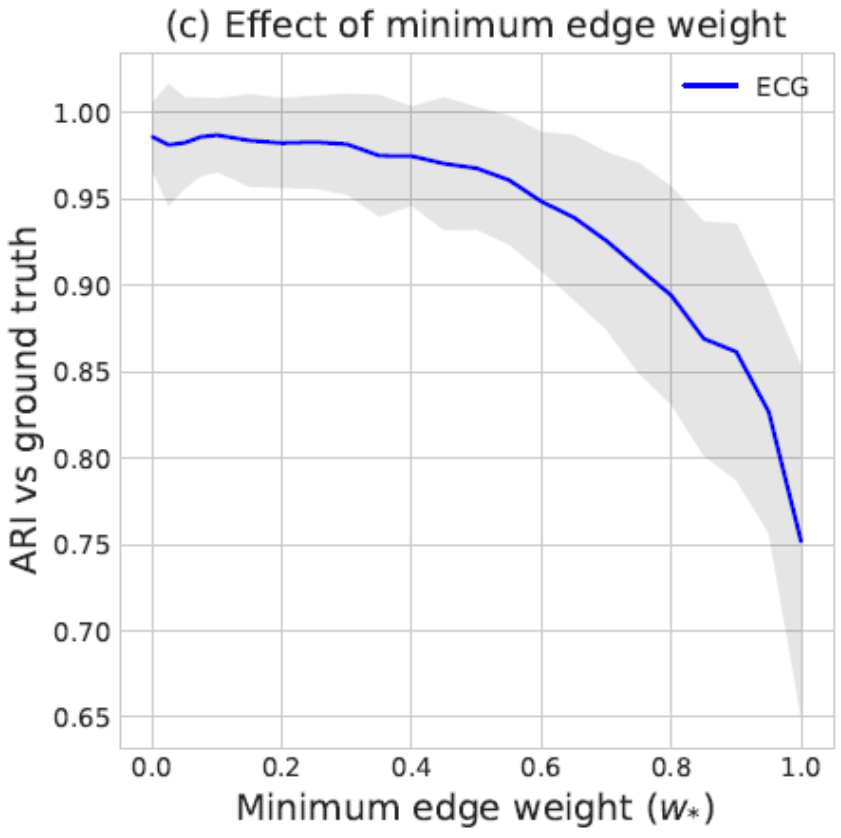


Figure 4 We compare the stability of the communities found by the Louvain (ML) and ECG algorithms over LFR graphs with 5 different choice of power law exponents. Partitions obtained in distinct runs for each algorithm are compared via the ARI measure. We see the much improved stability with ECG. Conclusions are the same with the AGRI measure (not shown).

Stability

Empirical studies indicated that the results are not very sensitive to the choice of parameters.





Comparison Study over LFR Graphs

www.nature.com/scientificreports

OPEN

A Comparative Analysis of Community Detection Algorithms on Artificial Networks

Received: 31 March 2016
Accepted: 07 July 2016
Published: 01 August 2016

Zhao Yang, René Algesheimer & Claudio J. Tessone

Many community detection algorithms have been developed to uncover the mesoscopic properties of complex networks. However how good an algorithm is, in terms of accuracy and computing time, remains still open. Testing algorithms on real-world network has certain restrictions which made their insights potentially biased: the networks are usually small, and the underlying communities are not defined objectively. In this study, we employ the Lancichinetti-Fortunato-Radicchi benchmark graph to test eight state-of-the-art algorithms. We quantify the accuracy using complementary measures and algorithms' computing time. Based on simple network properties and the aforementioned results, we provide guidelines that help to choose the most adequate community detection algorithm for a given network. Moreover, these rules allow uncovering limitations in the use of specific algorithms given macroscopic network properties. Our contribution is threefold: firstly, we provide actual techniques to determine which is the most suited algorithm in most circumstances based on observable properties of the network under consideration. Secondly, we use the mixing parameter as an easily measurable indicator of finding the ranges of reliability of the different algorithms. Finally, we study the dependency with network size focusing on both the algorithm's predicting power and the effective computing time.

Comparison Study over LFR Graphs

Comparing 8 algorithms over thousands of LFR graphs

Parameter	Value
Number of nodes N	$233 \sim 31948$
Maximum degree	$0.1N$
Maximum community size	$0.1N$
Average degree	20
Degree distribution exponent	-2
Community size distribution exponent	-1
Mixing coefficient μ	$[0.03, 0.75]$

Table 1. Parameters of LFR benchmark graphs. To deal with possible discrepancies in the network properties, we have randomly generated 100 network for every set of parameters. Due to the slow computing speed, Spinglass and Edge betweenness algorithms have been tested only on small networks with $N \leq 1000$.

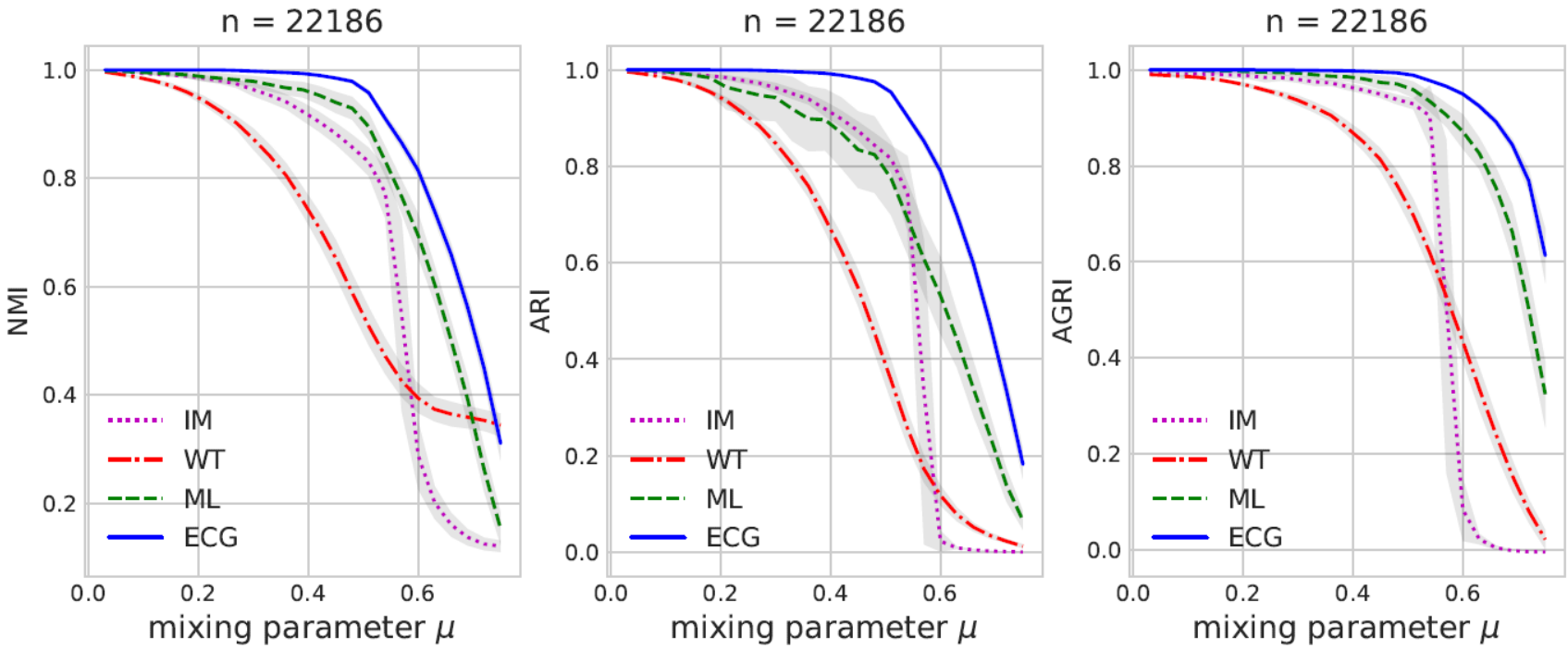
Comparison Study over LFR Graphs

Let n be the number of nodes, e number of edges
The following algorithms were considered:

algorithm	complexity
Louvain (ML)	$O(n \cdot \log n)$
Infomap (IM)	$O(e)$
WalkTrap (WT)	$O(n^2 \cdot \log n)$
Label Propagation (LP)	$O(e)$
SpinGlass (SG)	$O(n^{3.2})$
FastGreedy (FG)	$O(n \log^2 n)$
Leading e-vectors (LE)	$O(n^2)$
Edge betweenness (EB)	$O(e^2 n)$

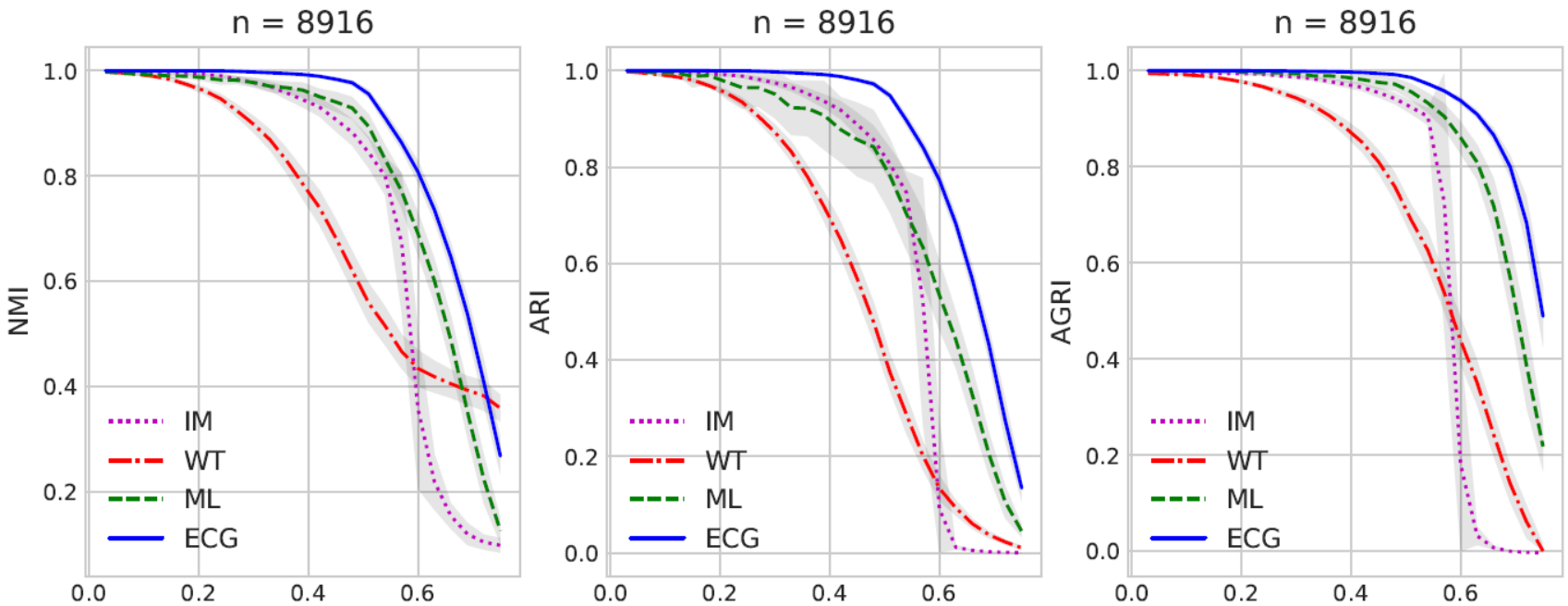
Comparison Study over LFR Graphs

LFR graphs with 22,186 nodes:



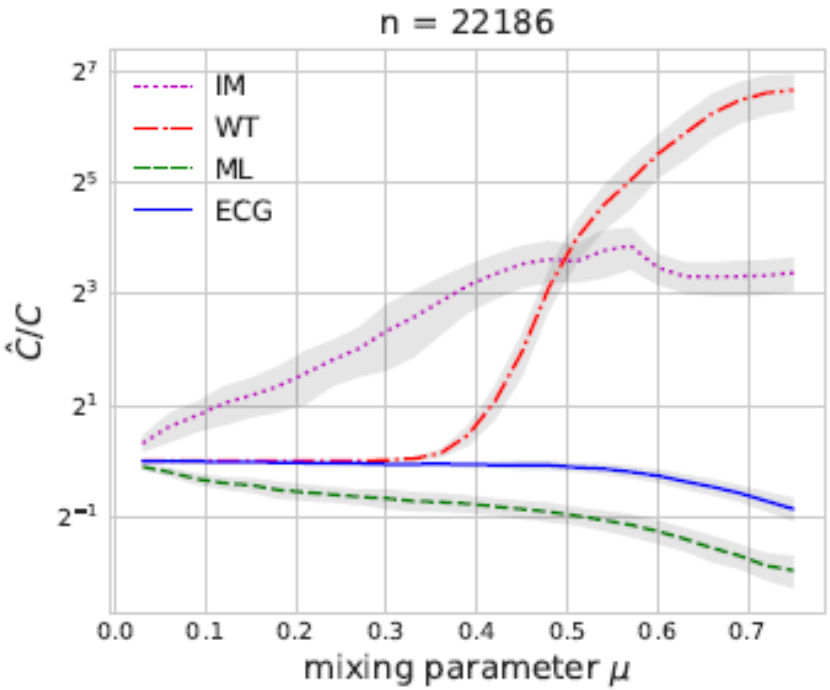
Comparison Study over LFR Graphs

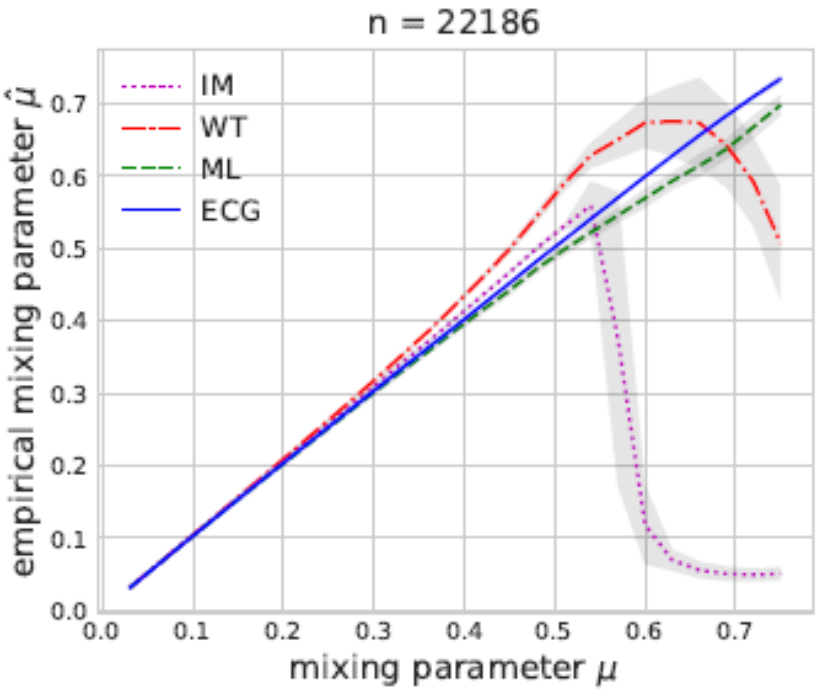
LFR graphs with 8,916 nodes:



Comparison Study over LFR Graphs

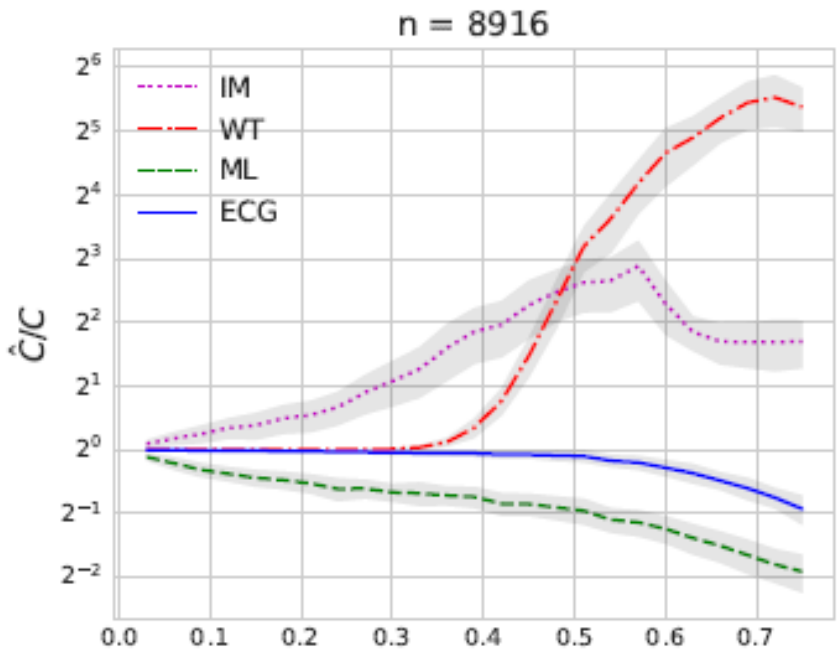
Further properties:

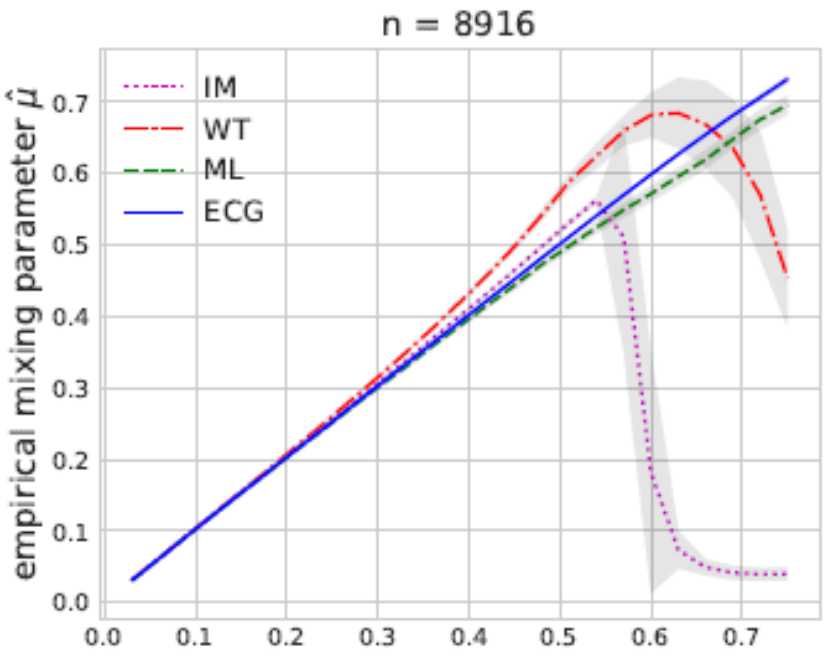




Comparison Study over LFR Graphs

Further properties:





Comparison Study over LFR Graphs

But this study only considered $\gamma_1 = 2, \gamma_2 = 1$.

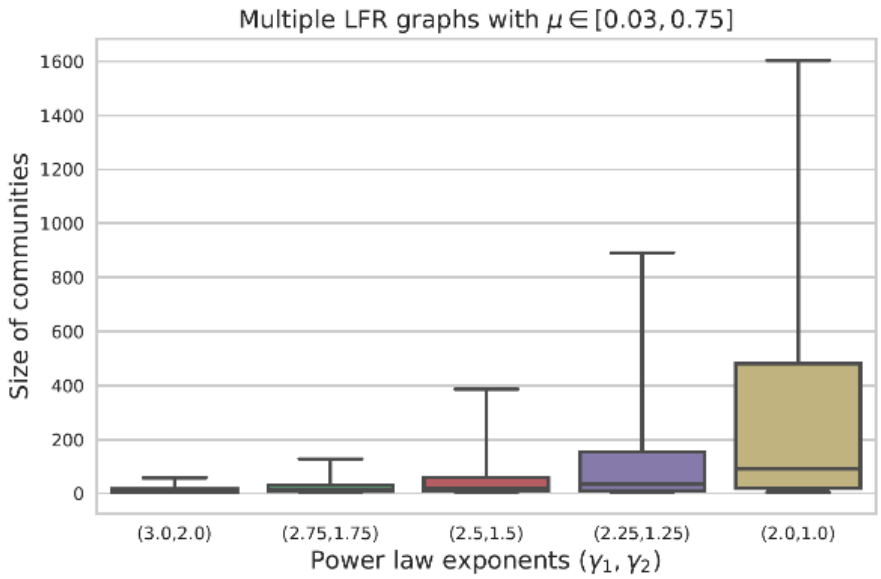
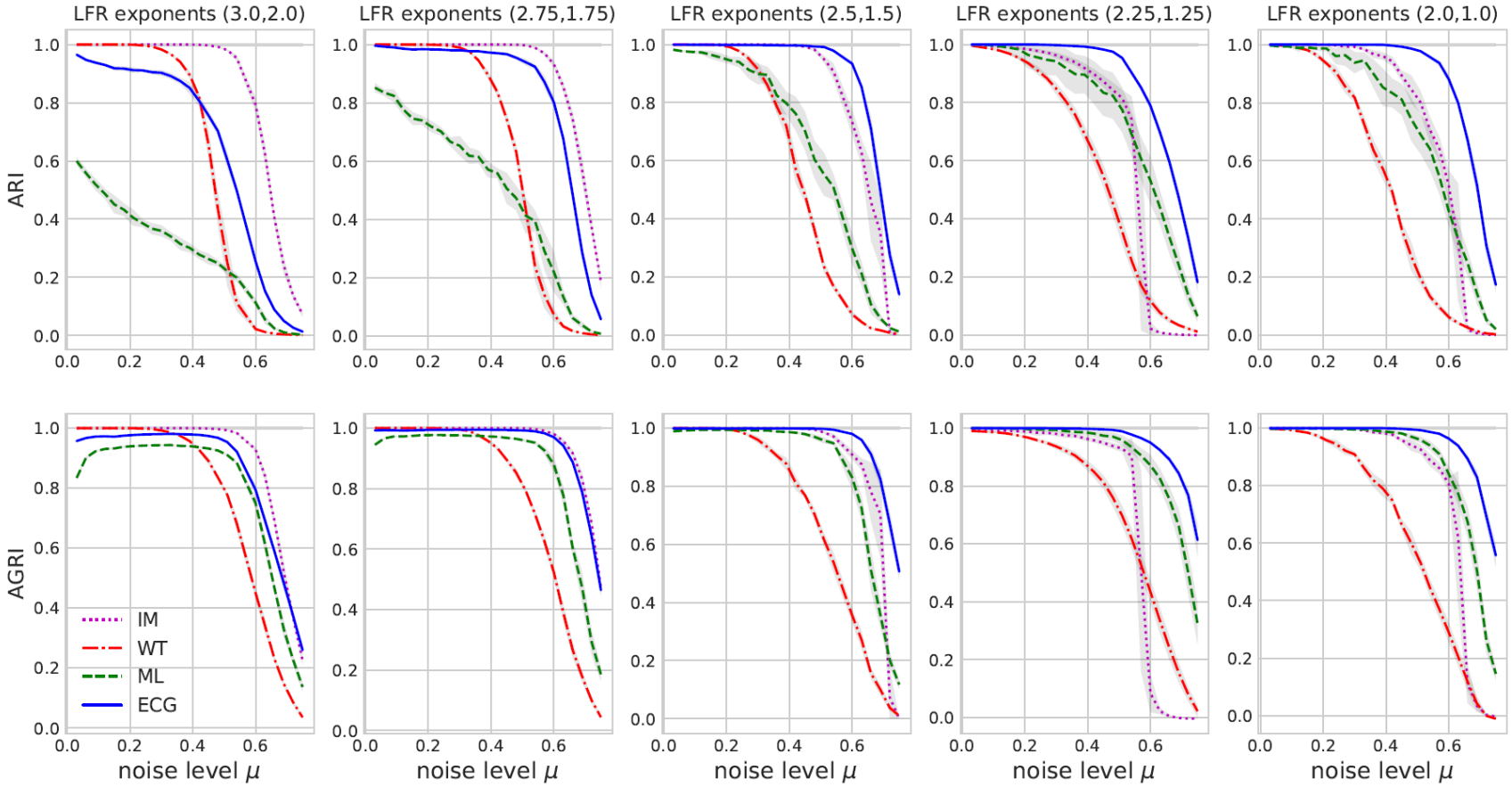


Figure 6 We selected 5 choices for the power law parameters (γ_1, γ_2) which are representative of various types of networks obtained with the LFR benchmark, and we look at the distribution of the sizes of communities. We see that with the largest recommended values $(\gamma_1, \gamma_2) = (3, 2)$, we get small communities of homogeneous size. As the exponents decrease, the sizes of the communities get more heterogeneous. All results were obtained by averaging over 10 graphs with 22,186 nodes for every choice of parameters $(\mu, \gamma_1, \gamma_2)$.

Comparison Study over LFR Graphs

Results over diverse LFR graphs:



Comparison Study over LFR Graphs

General observations:

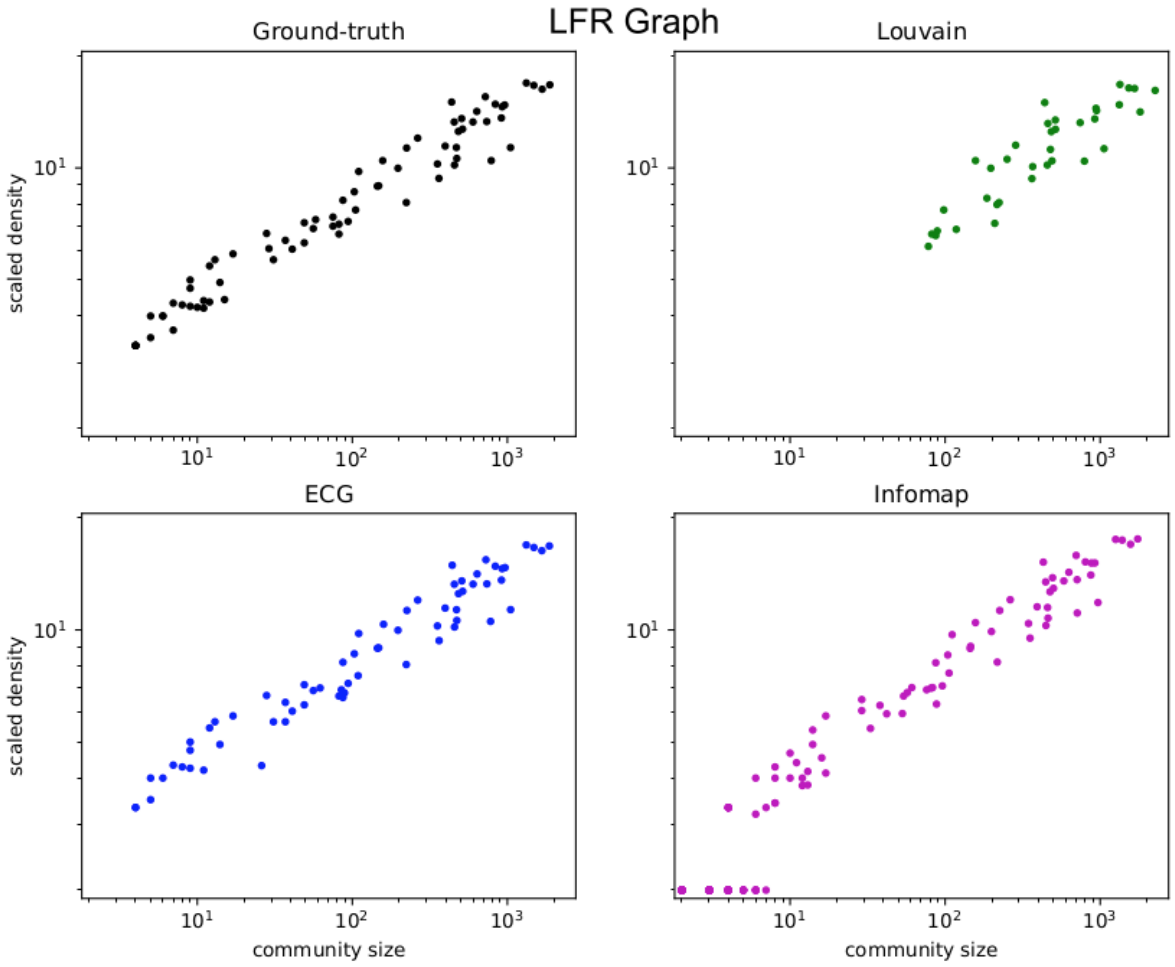
InfoMap gives best results over small communities of homogeneous size

ECG gives best results in other cases

ECG always improves over single Louvain (ML)

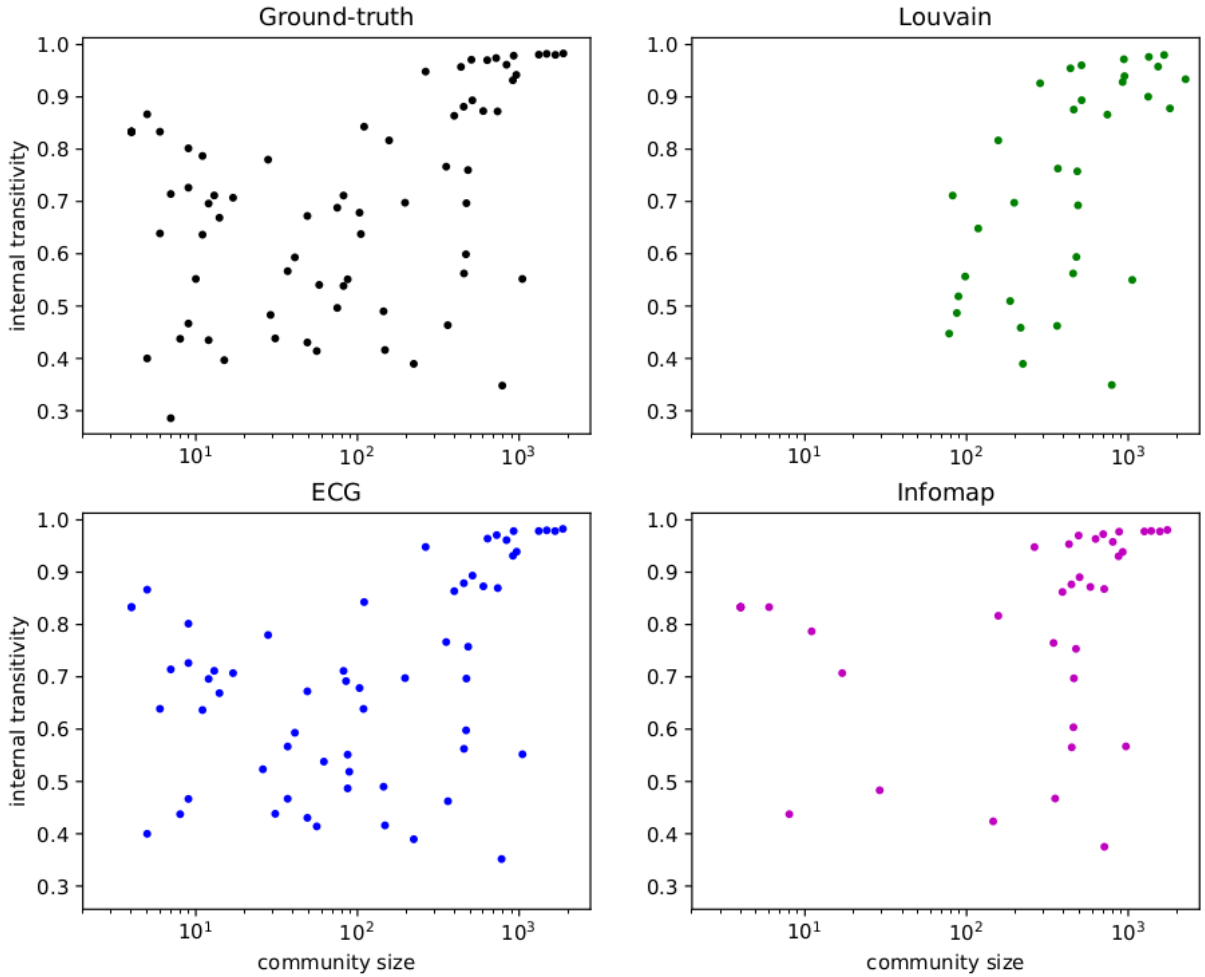
Topological features

Scaled density over LFR graph with $\mu = .39$:



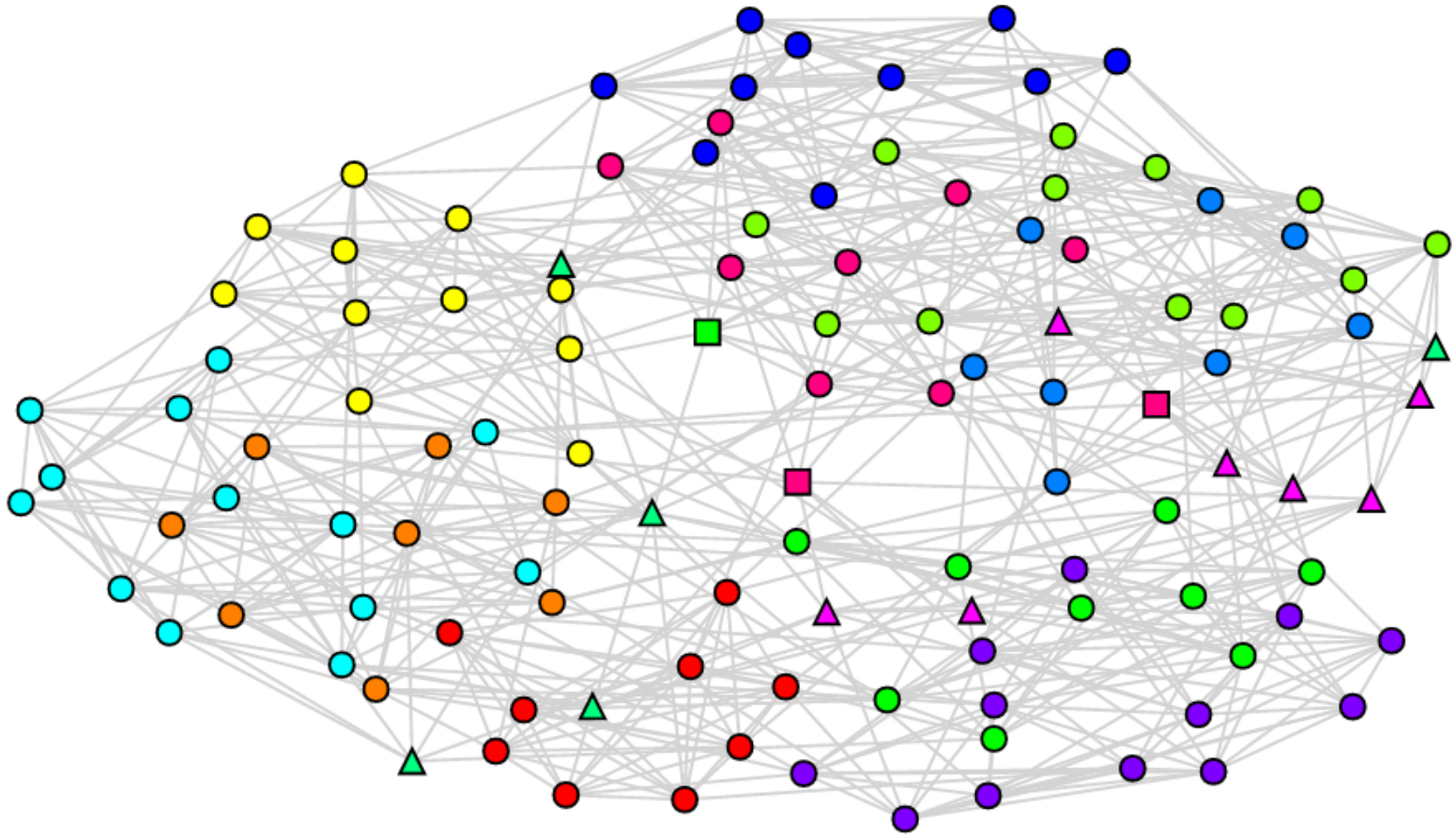
Topological features

Internal transitivity over LFR graph with $\mu = .39$:



The Football Graph

Recall the college football graph:



The Football Graph

Best results with ECG and InfoMap:

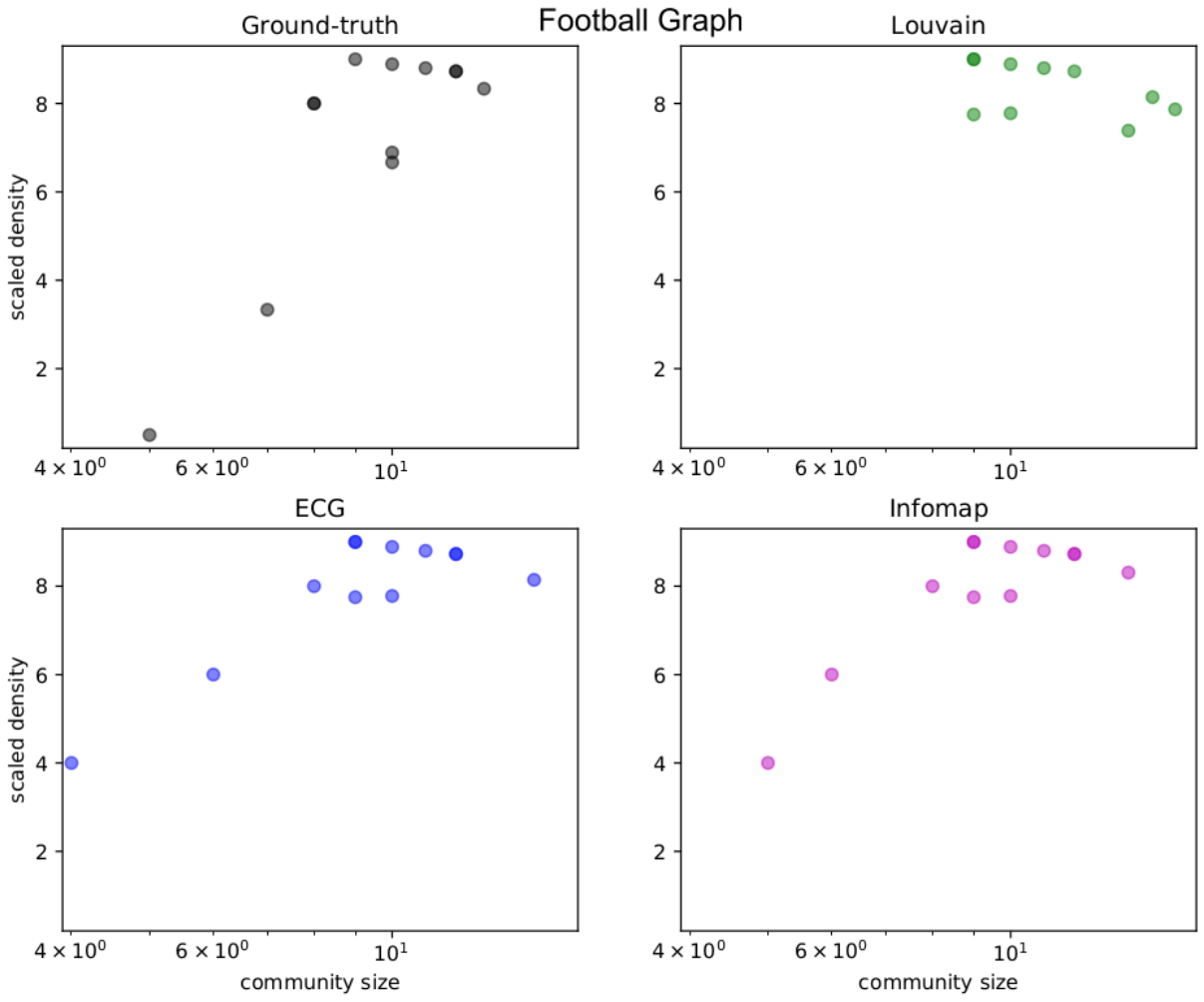
Results for the college football graph

	ECG	ML	WT	IM
ARI	**.889±.016**	.763 ± .052	.815	**.897**
AMI	**.900±.005**	.843 ± .020	.856	**.899**
AGRI	**.869±.005**	.815 ± .019	.837	**.872**

Table 1 We run each clustering algorithm 100 times on the college football dataset, namely: ECG, Louvain (ML), WalkTrap (WT) and InfoMap (IM). We compute three different measures with respect to the ground-truth communities (12 conferences). Where significant, we also report the standard deviation.

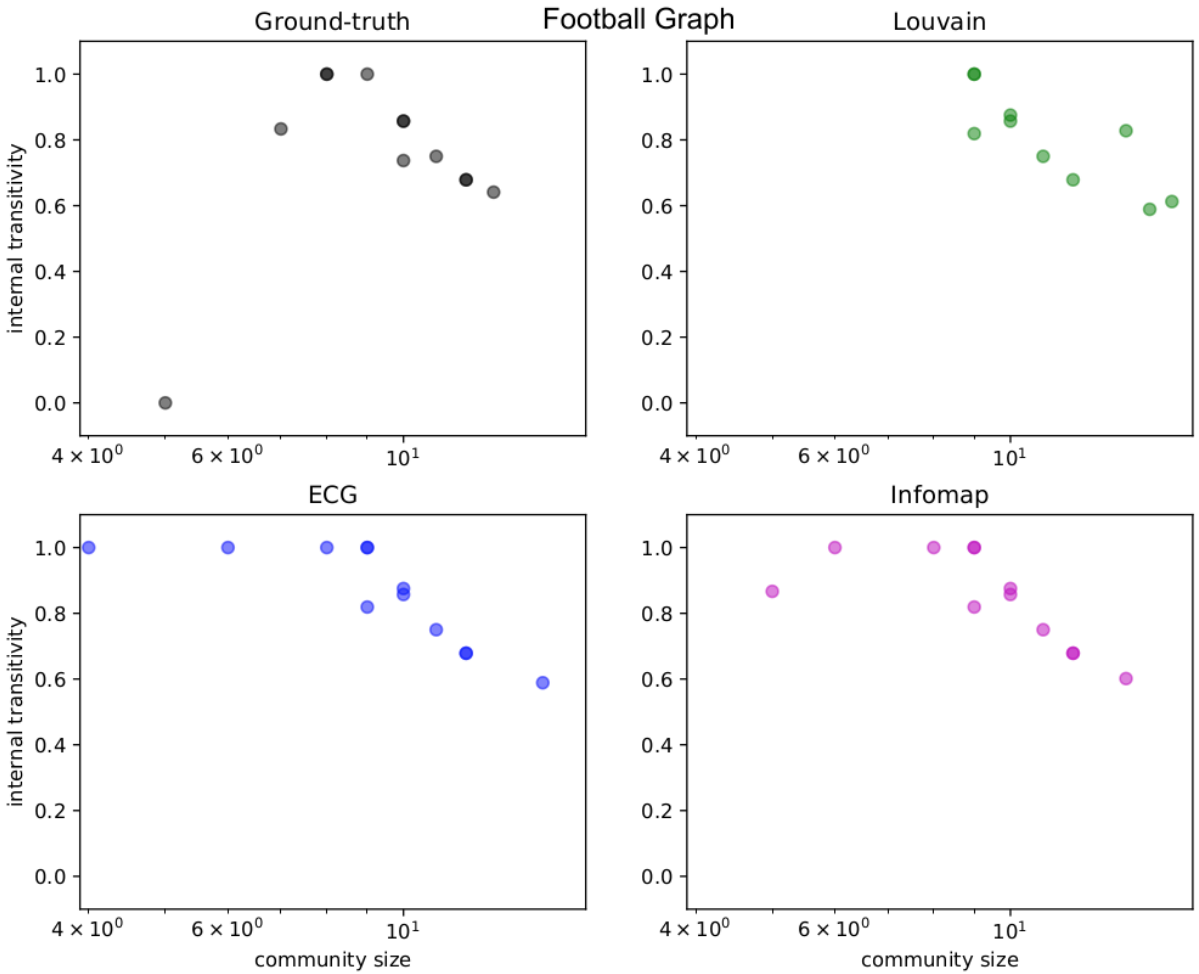
The Football Graph

Topological features: scaled density



The Football Graph

Topological features: internal transitivity



The YouTube Graph

YouTube friendship graph from `snap.standford.edu/data`

1,134,890 nodes (users) and 2,987,624 edges (friendship)

2-core covers only 41.1% of vertices

8,385 communities are declared user groups

Those communities are very weak from topological view

Only 12 qualify as *weak communities*, with ratio of external to internal degree under 0.5

We extend this ratio to 0.75 (similar to μ in LFR graphs)

The YouTube Graph

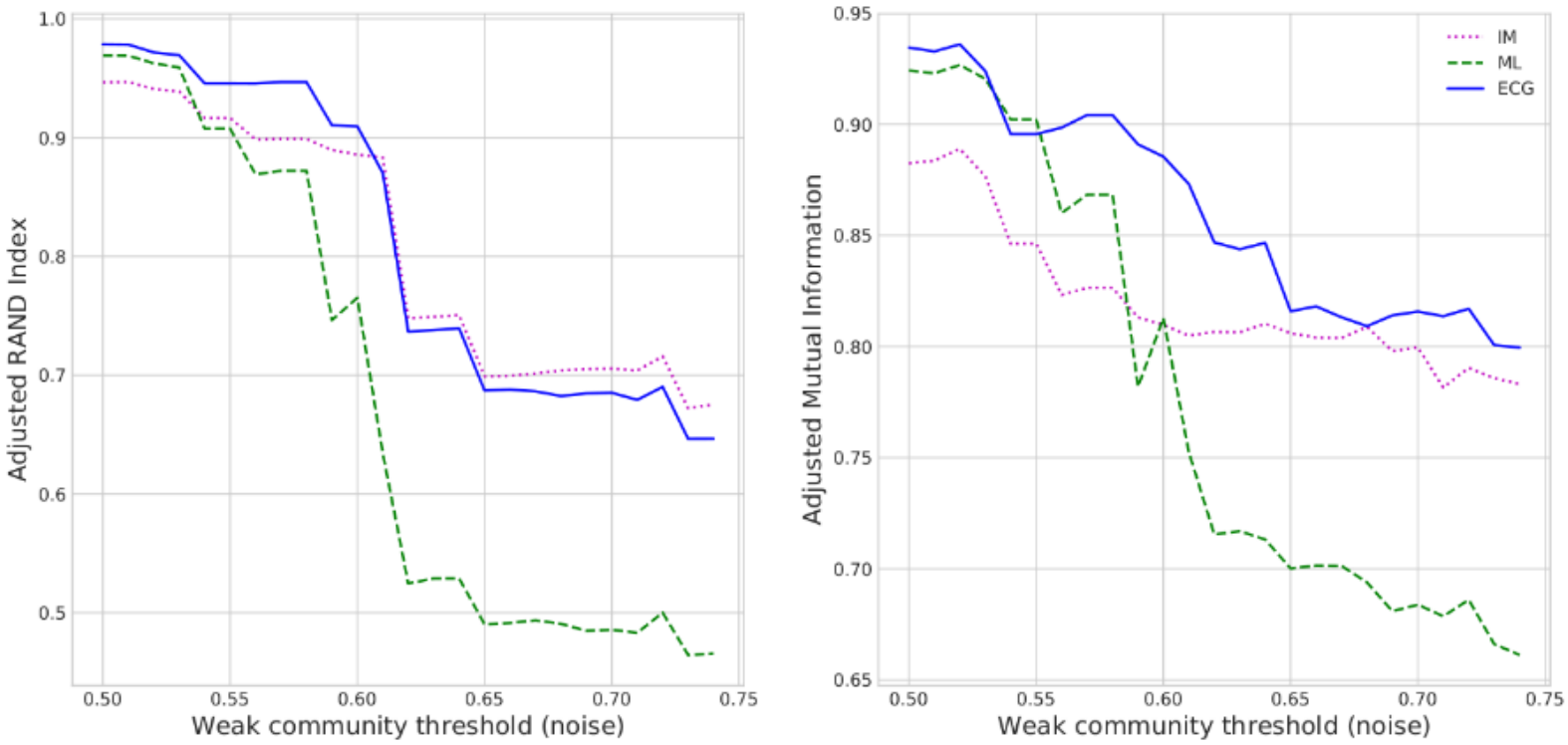


Figure 8 We compare the partitions obtained with ECG, Louvain (ML) and Infomap (IM) over the Youtube graph. We only consider the groud-truth communities where the ratio of external to total degree is below some threshold, which we vary from 0.5 to 0.75 on the x-axis. Results are compared via the ARI and AMI measures.

ECG Weights

We saw that ECG re-weighting helps with the resolution issue and with stability

We discuss a few other applications of the computed ECG weights:

- we define a new *Community Strength Index* (CSI)
- we show how to use the weights to zoom-in around seed vertices

Community Strength Index

We noticed that bi-modal distribution of the ECG weights near the boundaries (0 and 1) is indicative of strong community structure.

We propose a simple community strength indicator (CSI) based on the point-mass Wasserstein distance

For all edges $(u, v) \in E$, with $W_{\mathcal{P}}(u, v)$ from ECG, we define:

$$CSI = 1 - 2 \cdot \frac{1}{|E|} \sum_{(u,v) \in E} \min\left(W_{\mathcal{P}}(u, v), 1 - W_{\mathcal{P}}(u, v)\right)$$

such that $0 \leq CSI \leq 1$.

Community Strength Index

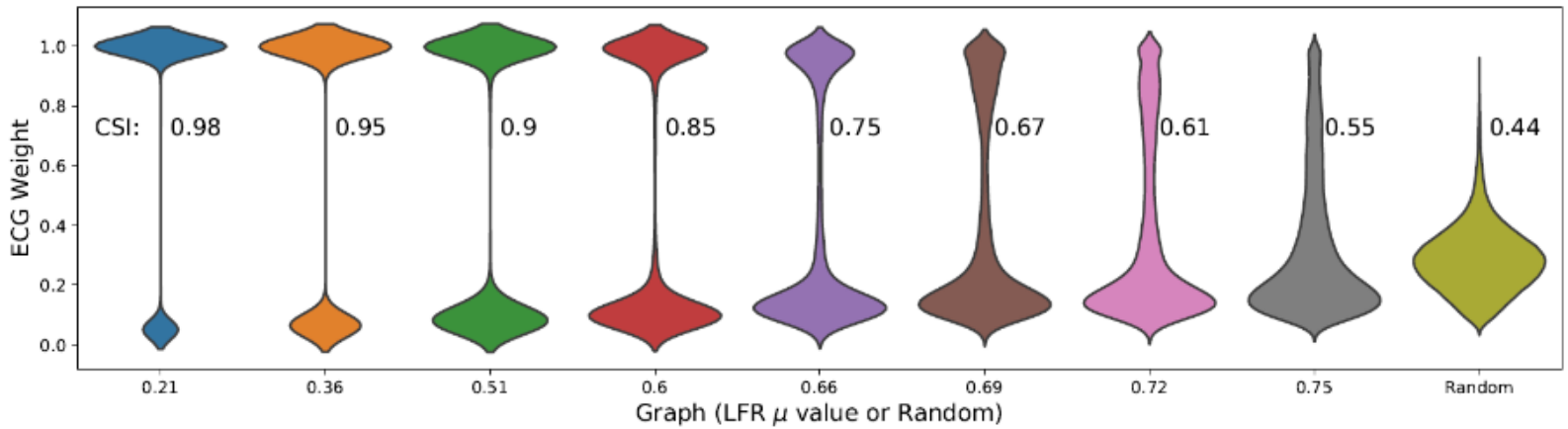
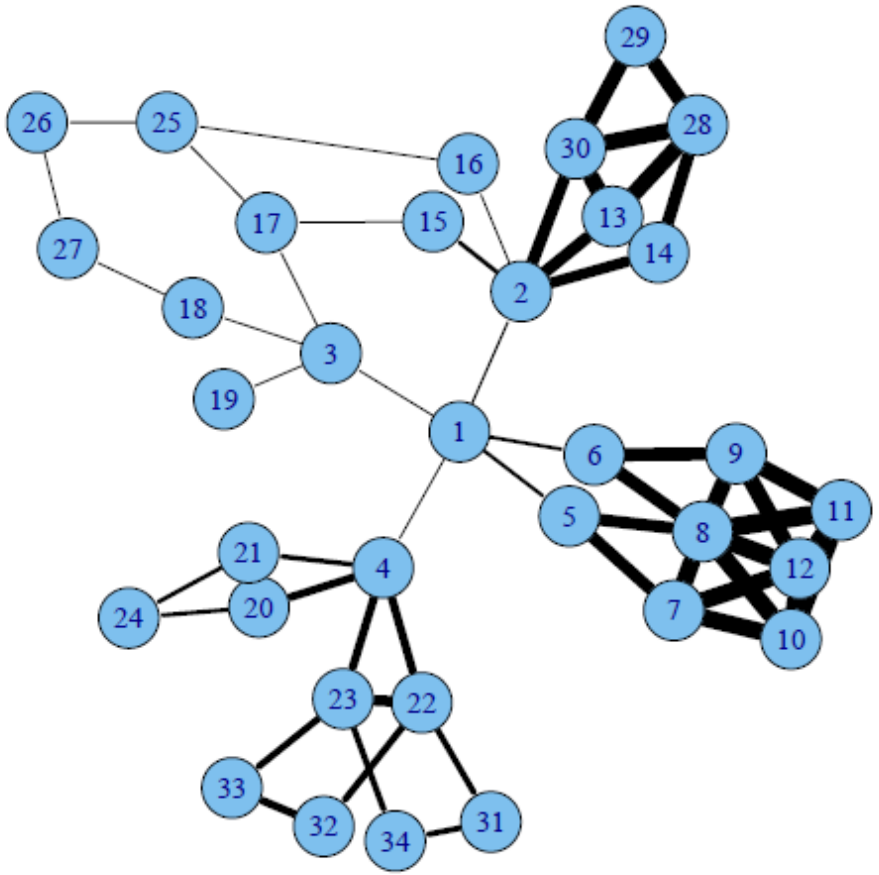


Figure 5 Violin plots of the ECG weight distribution for a family of LFR graphs with $n = 22,186$ nodes, parameters $\gamma_1 = 2$, $\gamma_2 = 1$ and $.21 \leq \mu \leq .75$. We also compare with a random graph of the same size and degree distribution as the graph with $\mu = 0.21$. We see the bi-modal distribution over LFR graphs up to a very high noise level. For large μ, the signal gets weaker. It is even weaker for the random graph. The Community Strength Indicator (CSI) is also reported.

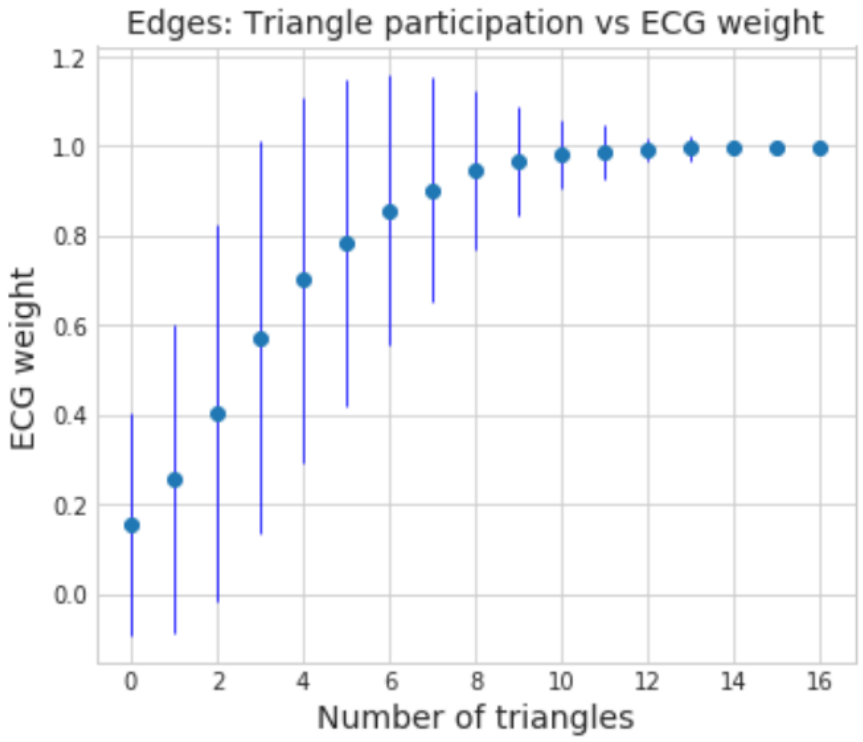
Strength of association

High ECG weights are indicative of strong associations:



Strength of association

Empirically comparing ECG weights and triangle participation:



Strength of association

We can use ECG weights as an alternative to ego-nets to zoom-in around seed nodes

Given a seed node v:

- identify the cluster it belongs to
- delete all edges with ECG weights below some threshold τ
- zoom-in on the connected component containing v
- increase τ to zoom-in even more

We look at the Amazon graph from SNAP (925,872 edges)

Strength of association

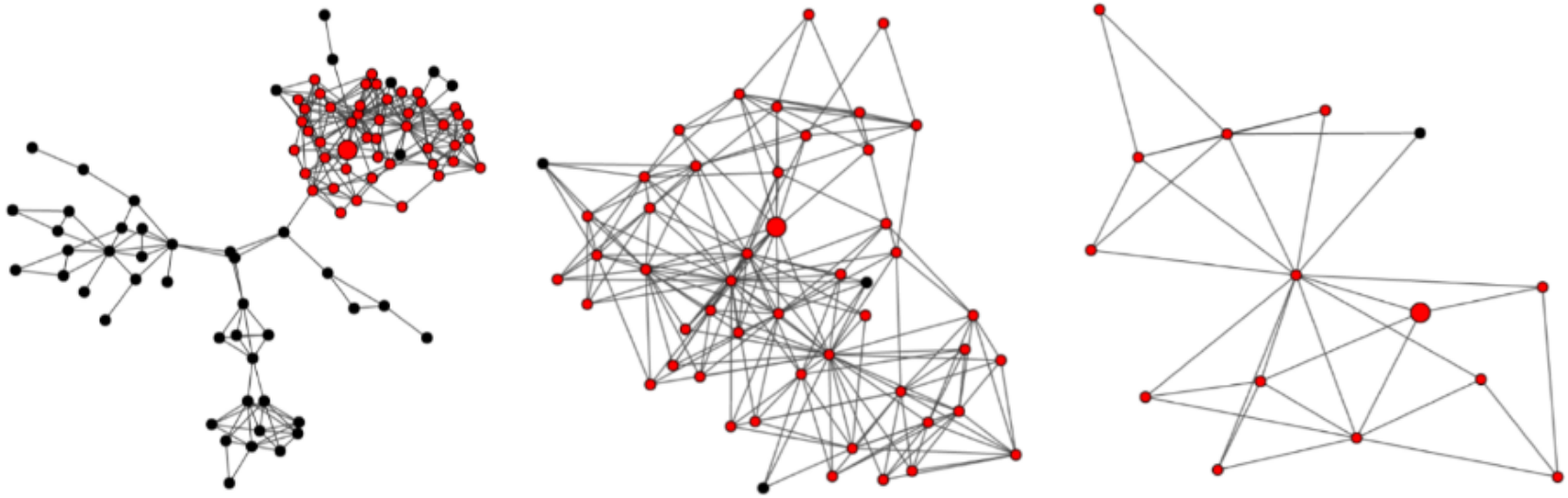


Figure 12 We consider a seed node (shown with larger size) from the Amazon co-purchasing graph. Nodes from the same ground truth communities are displayed in red, and other nodes are displayed in black. From left to right, we display respectively (i) the entire sub-graph obtained from the ECG part that contains the seed, (ii) a connected sub-graph with ECG edge weights above 0.1 containing the seed, and (iii) a connected sub-graph with ECG edge weights above 0.72 containing the seed. While the first plot has many spurious nodes, as we zoom in, most nodes we retain are in the same true community as the seed node.

Anomaly detection

CADA (community-aware anomaly detection) was recently proposed

For each node $v \in V$, let:
$N(v)$: the number of neighbors of v,
$N_c(v)$ the number of neighbors of v that belong to the most represented community (via graph clustering).

The original paper uses Infomap or Louvain:

$$CADA_x(v) = \frac{N(v)}{N_c(v)}$$

where $x \in \{IM, ML\}$.

REF: Helling, Scholtes and Takes, *A community-aware approach for identifying node anomalies in complex networks*

Anomaly detection

They validated their algorithm over LFR graphs with $\gamma_1 = 3$ and $\gamma_2 = 2$ only.

This choice corresponds to small communities of homogeneous size.

We re-visited this approach with ECG, and more values for the power law exponents.

For each graph, we added 200 random anomalous nodes (random edges) with the same degree distribution as in LFR.

Anomaly detection

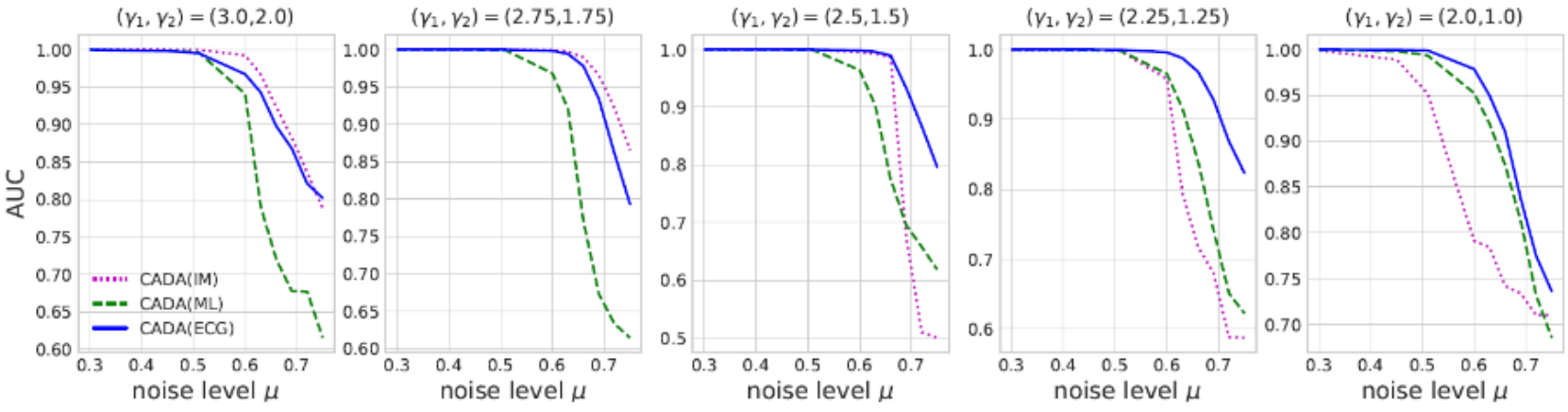


Figure 11 We compare three flavours of the CADA algorithm, using the InfoMap (IM), Louvain (ML) and ECG. For each value of $.3 \leq \mu \leq .75$, we generated 10 LFR graphs of size 22,186, along with 200 random anomalous nodes with the same degree distribution. We considered 5 different choices for the LFR power law exponents. Results are compared via the area under the ROC curve (AUC).

References

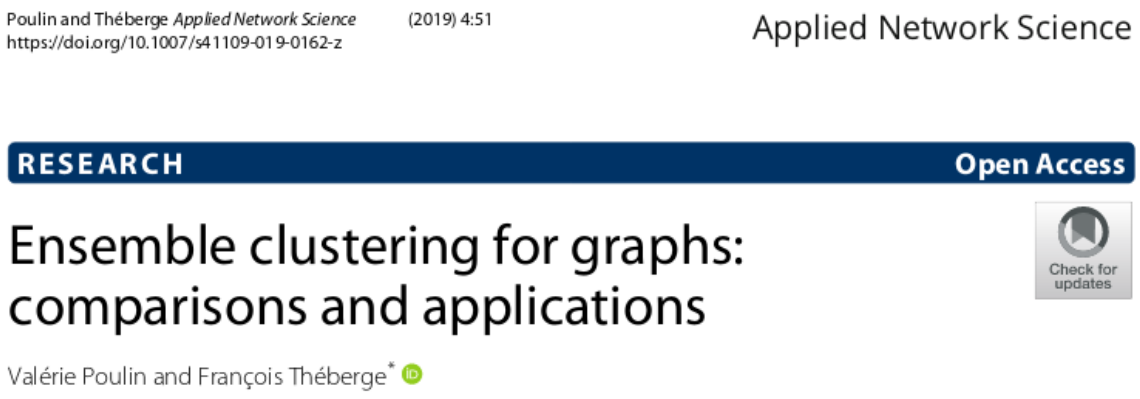
Valérie Poulin and François Théberge,
Ensemble Clustering for Graphs, Complex Networks and Their Applications VII, Studies in Computational Intelligence, vol 812. Springer (2019)
DOI:10.1007/978-3-030-05411-3_19,
pre-print **arXiv:1809.05578**

Code:
https://codeocean.com/capsule/3898939/tree/v1
on **codeocean.com**.

References

Further results and applications in:

Valérie Poulin and François Théberge,
Ensemble Clustering for Graphs: Comparisons and Applications,
pre-print, **arXiv:1903.08012**, 2019; recently published in
Applied Network Science (`https://rdcu.be/bLn9i`):



Poulin and Théberge *Applied Network Science* (2019) 4:51
https://doi.org/10.1007/s41109-019-0162-z

Applied Network Science

RESEARCH Open Access

Ensemble clustering for graphs: comparisons and applications

Check for updates

Valérie Poulin and François Théberge*

Notebook #6