Statistical testing of clusterability

Pierre Miasnikof, University of Toronto

17 November, 2022

Acknowledgements

Joint work

This is joint work with A.Y. Shestopaloff, L. Prokhorenkova and A.M. Raigorodskii.

Joint work

This is joint work with A.Y. Shestopaloff, L. Prokhorenkova and A.M. Raigorodskii.

Thanks!

We would like to thank:

- The seminar organizers, for the invitation!
- You, the audience, for your time and interest!

Objectives for today

1. Introduce the clusterability question

Objectives for today

1. Introduce the clusterability question
2. Introduce two tests "on the market"

Objectives for today

1. Introduce the clusterability question
2. Introduce two tests "on the market"
3. Show comparative performances

Objectives for today

1. Introduce the clusterability question
2. Introduce two tests "on the market"
3. Show comparative performances
4. Introduce our own test, the "δ test"

Objectives for today

1. Introduce the clusterability question
2. Introduce two tests "on the market"
3. Show comparative performances
4. Introduce our own test, the "δ test"
5. **Describe our test's statistical power** (latest work)

Objectives for today

1. Introduce the clusterability question
2. Introduce two tests "on the market"
3. Show comparative performances
4. Introduce our own test, the "δ test"
5. **Describe our test's statistical power** (latest work)

Limitations for this talk

For today,

- Undirected
- Unweighted
- No self-loops

Goal of this work

The focal question

Context:

Graph clustering, also known as community detection

The focal question

Context:

Graph clustering, also known as community detection

Still open questions: (hopefully partially answered here)

Is the graph "cluster-able"? Does it even have a clustered structure? (this question was raised recently in the literature)

The focal question

Context:

Graph clustering, also known as community detection

Still open questions: (hopefully partially answered here)

Is the graph "cluster-able"? Does it even have a clustered structure? (this question was raised recently in the literature)

Key facts:

- Not all graphs are suited for clustering (aka. community detection)

The focal question

Context:

Graph clustering, also known as community detection

Still open questions: (hopefully partially answered here)

Is the graph "cluster-able"? Does it even have a clustered structure? (this question was raised recently in the literature)

Key facts:

- Not all graphs are suited for clustering (aka. community detection)
- Not all graphs (or data sets) are formed by homogeneous subsets within a heterogeneous whole

The focal question

Context:

Graph clustering, also known as community detection

Still open questions: (hopefully partially answered here)

Is the graph "cluster-able"? Does it even have a clustered structure? (this question was raised recently in the literature)

Key facts:

- Not all graphs are suited for clustering (aka. community detection)
- Not all graphs (or data sets) are formed by homogeneous subsets within a heterogeneous whole
- Bottom line: Not all graphs (or data sets) can be meaningfully summarized through clusters

What are vertex clusters?

A still open question

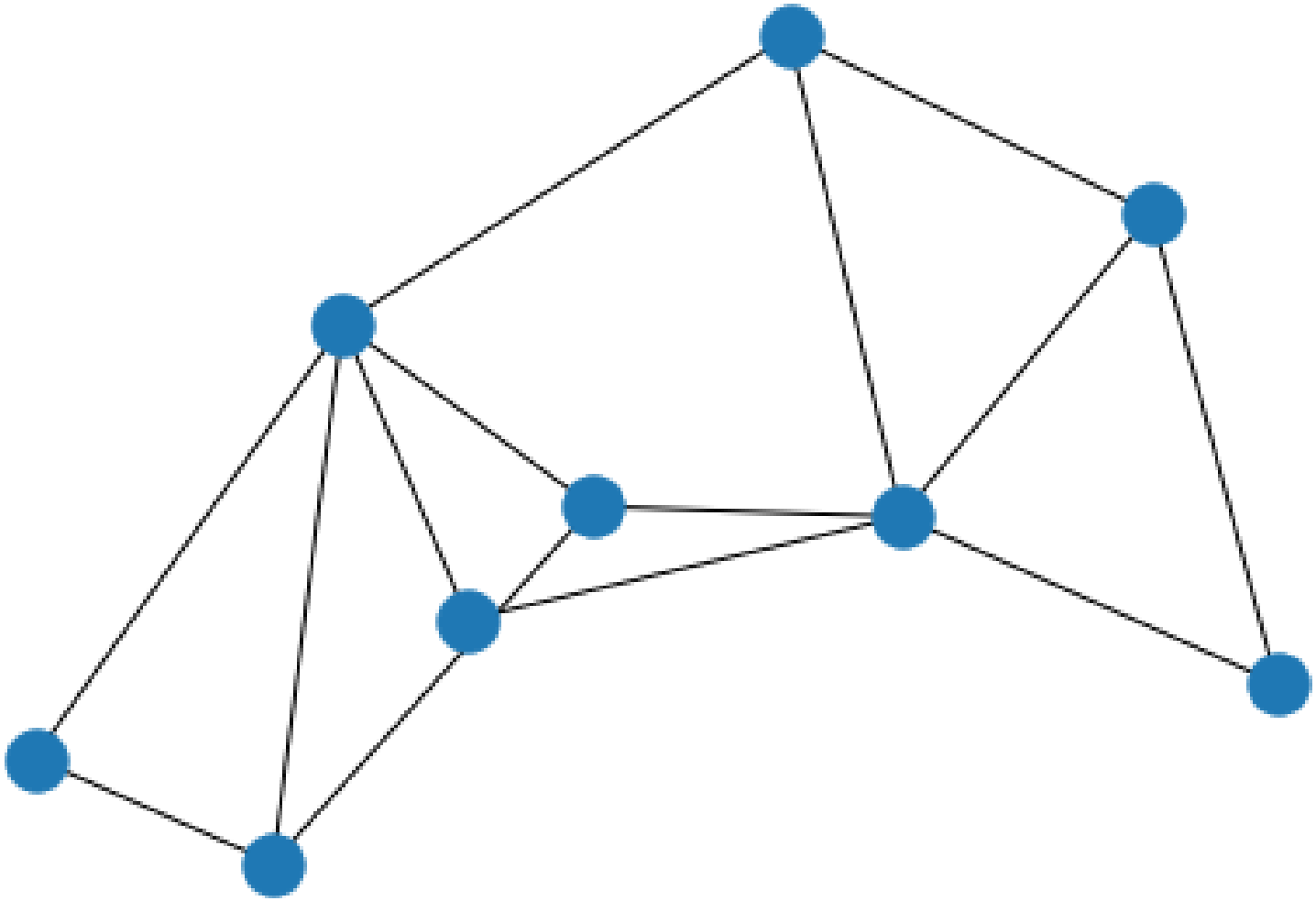
- A formal definitions of "vertex clusters" does not exist (yet)
- However, the consensus is that clusters are
  - Densely connected sets of vertices
  - That display sparse connections to the remaining vertices

What are vertex clusters?

A still open question

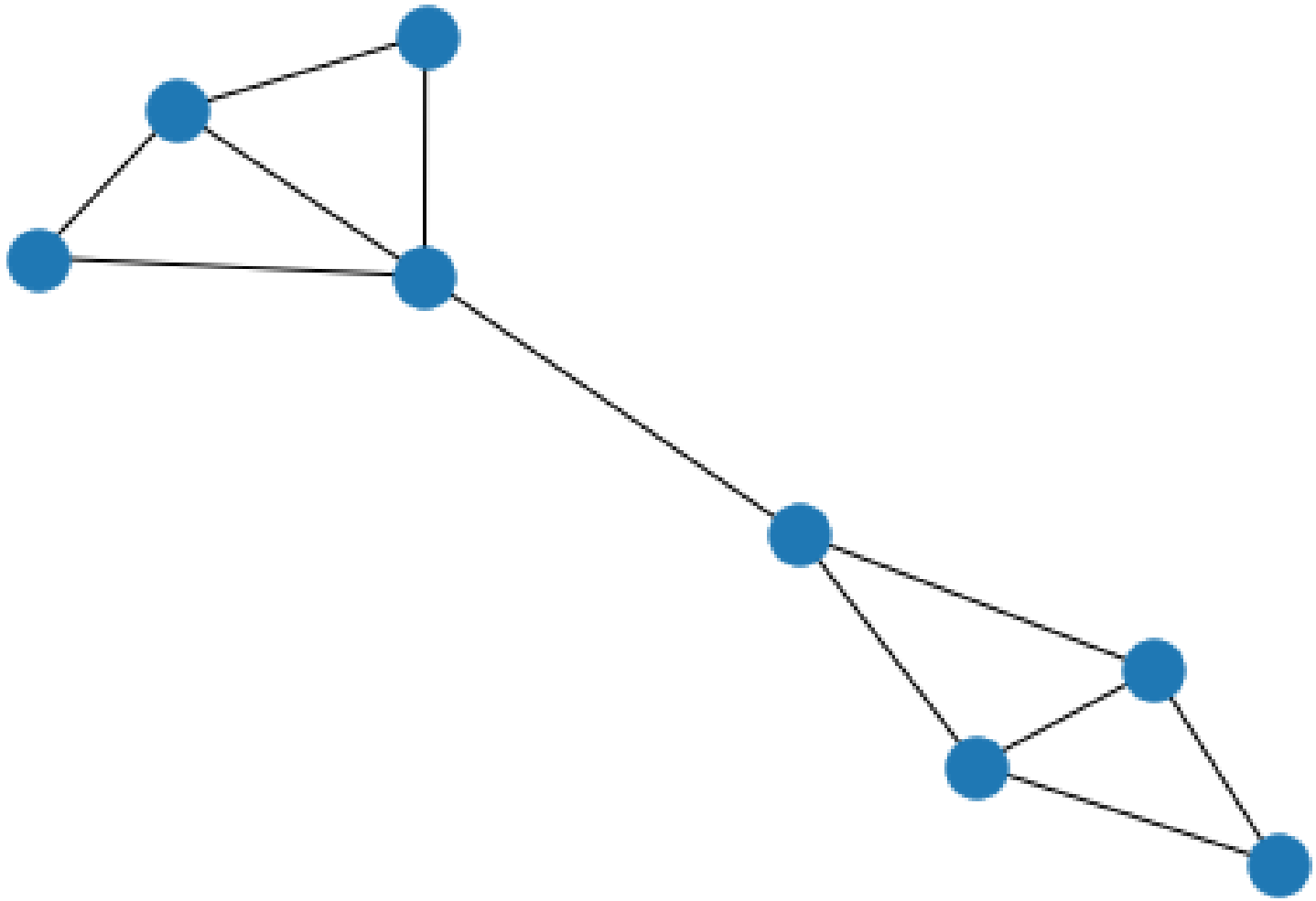
- A formal definitions of "vertex clusters" does not exist (yet)
- However, the consensus is that clusters are
  - Densely connected sets of vertices
  - That display sparse connections to the remaining vertices
- In summary, clusters form dense induced subgraphs

We want to answer: does the graph look like this?



ERG graph with $n=9$ & $p = 1/3$

OR does it look like this?



Motivations

Why do we care?

In the literature, back in 2006!

"(...) running a clustering algorithm over a set of randomly generated data points will always produce clusters which, however, have little meaning." [Reichardt and Bornholdt, 2006]

Why do we care?

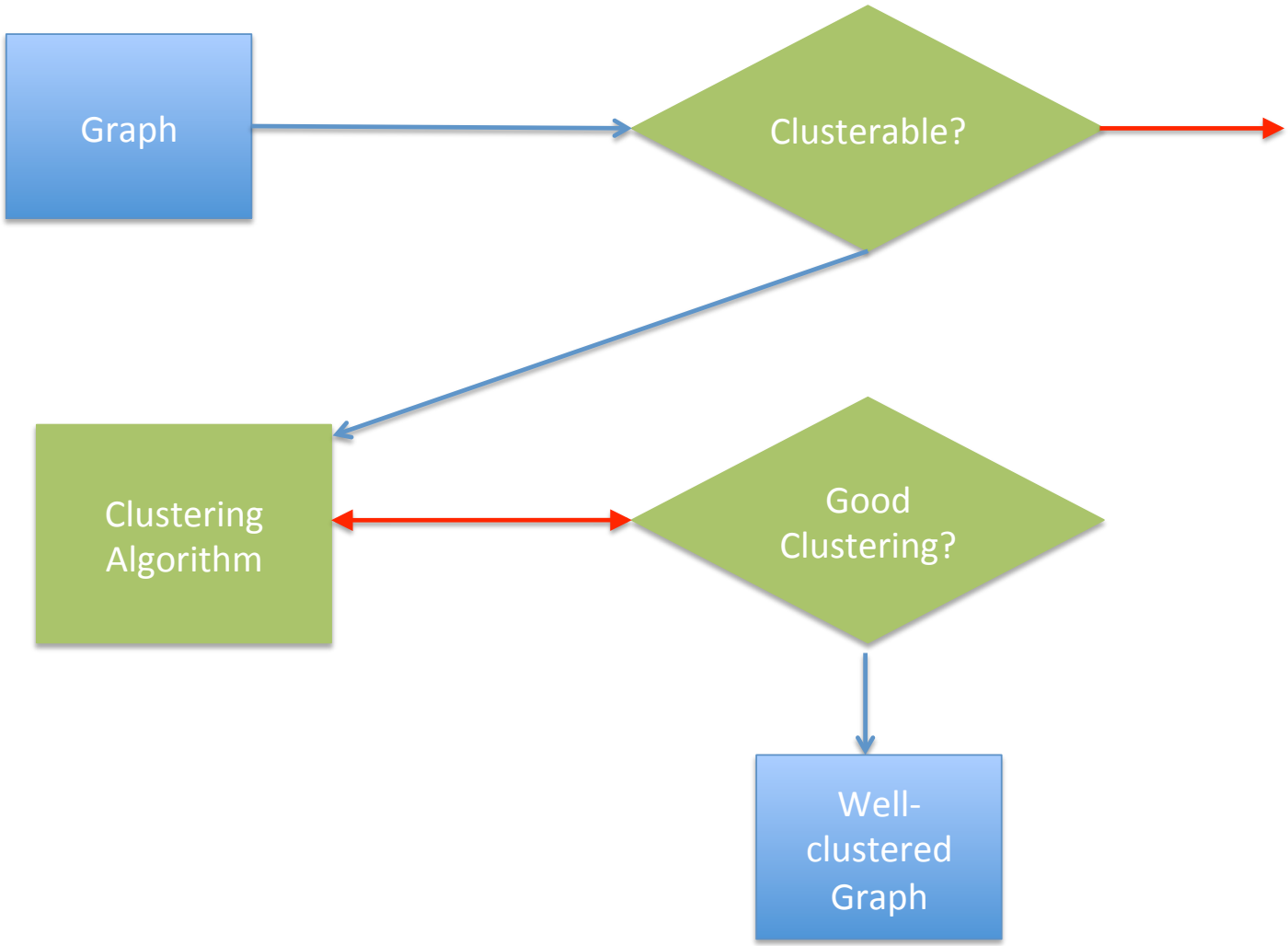
In the literature, back in 2006!

"(...) running a clustering algorithm over a set of randomly generated data points will always produce clusters which, however, have little meaning." [Reichardt and Bornholdt, 2006]

More recently:

"(...) given access to a graph G = (V,E), can we quickly determine whether the graph can be partitioned into a few clusters with good inner conductance (...)? " [Chiplunkar et al., 2018]

Ideal workflow for graph clustering



Previous work (recent)

- Chiplunkar et al. pose the question "(...) given access to a graph G = (V,E), can we quickly determine whether the graph can be partitioned into a few clusters with good inner conductance (...)?"
  [Chiplunkar et al., 2018]
- Gao & Lafferty use sampling and statistical testing to determine if a network has community structure
  [Gao and Lafferty, 2017b, Gao and Lafferty, 2017a]

$\rightarrow$ BUT, these authors' tests rely on strong assumptions about the graph AND their results are lacking [Miasnikof et al., 2019]

More on previous work I

Our answer

The "δ test" (a statistical significance test, more to follow on this topic) [Miasnikof et al., 2019]

Methods

The "δ test"

- Main idea: If the graph is clusterable, then we should observe heterogeneous local densities (on average)

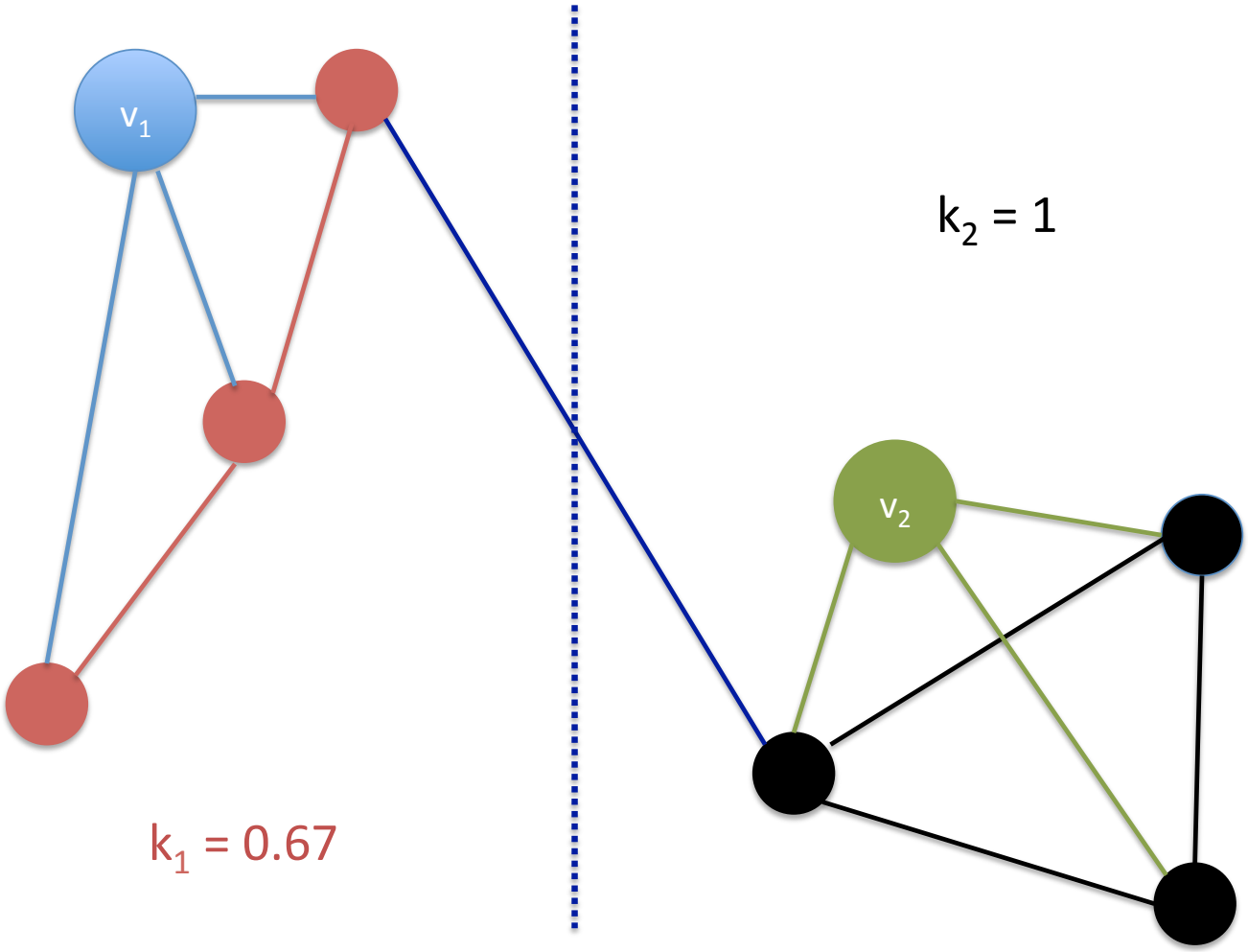
The "δ test"

- Main idea: If the graph is clusterable, then we should observe heterogeneous local densities (on average)
- We should have denser (than overall) connections in neighborhood induced subgraphs (sample & test)

The "δ test"

- Main idea: If the graph is clusterable, then we should observe heterogeneous local densities (on average)
- We should have denser (than overall) connections in neighborhood induced subgraphs (sample & test)
- We use the δ test to assess the statistical significance of the heterogeneity of local densities ("A Statistical Test of Heterogeneous Subgraph Densities To Assess Clusterability") [Miasnikof et al., 2019]

Heterogeneous local densities



K = 0.43

Statistical hypotheses

Our test is based on the hypothesis that a clusterable graph will display, on average, a local neighborhood induced subgraph density that is greater than the graph's overall density. (one-sided t-test)

Statistical hypotheses

Our test is based on the hypothesis that a clusterable graph will display, on average, a local neighborhood induced subgraph density that is greater than the graph's overall density. (one-sided t-test)

- Null hypothesis: the graph is not clusterable
  - Mean local densities is statistically indistinguishable from the global graph density (two-sample t-test, with unequal variance)
- Alternative hypothesis: the graph **MAY BE** clusterable
  - Mean local densities is statistically distinguishable ($>$) from the global graph density (one-sided)

Statistical hypotheses

Our test is based on the hypothesis that a clusterable graph will display, on average, a local neighborhood induced subgraph density that is greater than the graph's overall density. (one-sided t-test)

- Null hypothesis: the graph is not clusterable
  - Mean local densities is statistically indistinguishable from the global graph density (two-sample t-test, with unequal variance)
- Alternative hypothesis: the graph **MAY BE** clusterable
  - Mean local densities is statistically distinguishable ($>$) from the global graph density (one-sided)

N.B.

- We test a **NECESSARY** (not sufficient) condition
- If we reject the null, we cannot conclude the graph is clusterable
- On the other hand, if we do not reject the null, we can conclude the graph is not clusterable

Statistical hypotheses

- Because we are taking means (of local densities),
  - We do not need to specify a statistical distribution
  - Under CLT (applicable if $n > 30$), the mean follows a Gaussian distribution
- Under the null this Gaussian is centered at K (global density)
- Under the alternative, it is centered at a significantly higher point (one-sided test)

Test procedure

1. Sample s, with $30 < s \ll |V|$ vertices

Test procedure

1. Sample s, with $30 < s \ll |V|$ vertices
2. Compute local densities for each neighborhood, κ_i

Test procedure

1. Sample s, with $30 < s \ll |V|$ vertices
2. Compute local densities for each neighborhood, κ_i
3. Take mean of local densities, $\bar{\kappa} = \frac{1}{s} \sum_i \kappa_i$

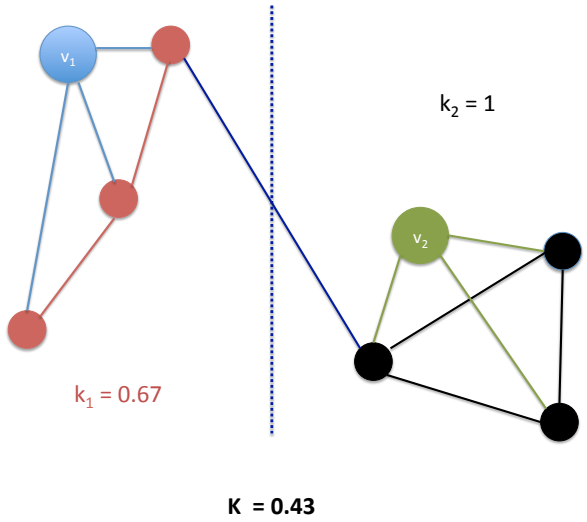
Test procedure

1. Sample s, with $30 < s \ll |V|$ vertices
2. Compute local densities for each neighborhood, κ_i
3. Take mean of local densities, $\bar{\kappa} = \frac{1}{s} \sum_i \kappa_i$
4. Compute $\delta = (\bar{\kappa}/K) - 1$

Test procedure

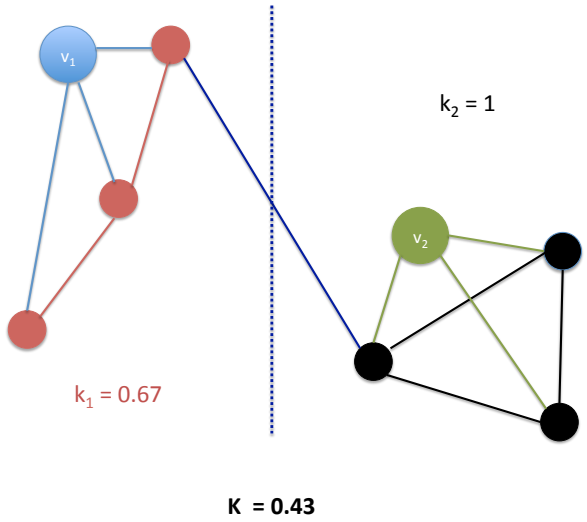
1. Sample s, with $30 < s \ll |V|$ vertices
2. Compute local densities for each neighborhood, κ_i
3. Take mean of local densities, $\bar{\kappa} = \frac{1}{s} \sum_i \kappa_i$
4. Compute $\delta = (\bar{\kappa}/K) - 1$
5. Test significance of δ: if $s > 30$, then
   - Under the null, δ follows a Gaussian distribution centered at 0
   - Under the alternative, δ follows a Gaussian distribution centered at a point > 0

Sampling procedure and mean (a two sample, $s = 2$ example)



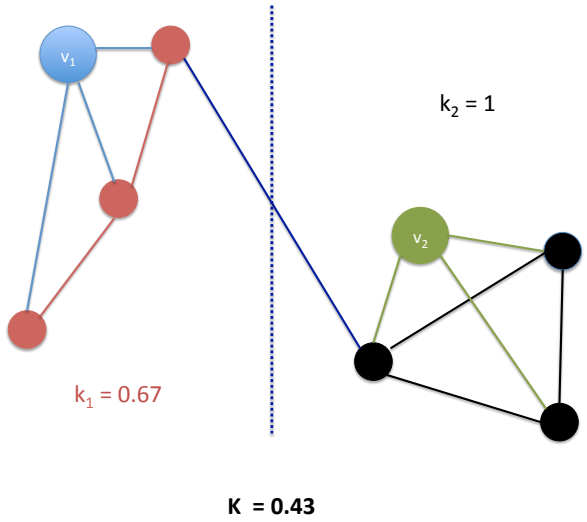
- Sample v_1, get density of it's neighborhood induced subgraph (κ_1)

Sampling procedure and mean (a two sample, $s = 2$ example)



- Sample v_1, get density of it's neighborhood induced subgraph (κ_1)
- Sample v_2, repeat (κ_2)

Sampling procedure and mean (a two sample, $s = 2$ example)



- Sample v_1, get density of it's neighborhood induced subgraph (κ_1)
- Sample v_2, repeat (κ_2)
- compute mean: $\bar{\kappa} = \frac{1}{2}\kappa_1 + \kappa_2 = \frac{1}{2}(0.67 + 1)$

Invoking the central limit theorem

CLT (a very standard definition)

"The central limit theorem states that the sample mean $\bar{X}$ follows approximately the normal distribution with mean and standard deviation $\frac{\sigma}{\sqrt{n}}$, where μ and σ are the mean and standard deviation of the population from where the sample was selected. The sample size n has to be large (usually $n \geq 30$) if the population from where the sample is taken is nonnormal."

Source: `http://www.stat.ucla.edu/~nchristo/introeconometrics/introecon_central_limit_theorem.pdf`

δ test in comparison

Table: A Comparison of Recently Published Clustering Tests

	$\kappa - \phi$ **Test**	**G-L Tests**	δ **Test**
Req num clust	Y	N	N
Assumes gen mod	Y	N	N
Stat conf int	N	Y	Y
Imposes dist'n on null	NA	Y	N
Weighted and unweighted	Y	N	Y

Empirical comparisons, initial test scenarios (2019)

Graph	Graph Characteristics
ER	$N = 1,000, p = 0.333$
SBM	$N = 1,000$, $\lvert c_i \rvert = 91, 21, 333, 555$
CC	num cliques $= 10$, size of cliques $= 100$
BA3	$N = 1,000$, out-degree $= 3$
BA5	$N = 1,000$, out-degree $= 5$
WS	$N = 1,000$, $k = 14$, $p = 0.2$
CM	$N = 1,000$, exp $= 3$
MD	One ER and two CC, $N = 1,000 = 850 + 2 \times (3 \times 25)$
EUC	$N = 1,005$ in 42 known clusters

Empirical comparisons, test results (2019)

Table: p-value comparisons to both G-L tests

Graph	EZ Test	T^2 Test	δ Test
ER	1.00	0.82	0.64
WS	1.00	0.00	0.00
BA3	1.00	0.00	0.00
BA5	1.00	0.00	0.00
CC	0.99	0.00	0.00
SBM	0.97	0.00	0.00
EUC	1.00	0.00	0.00
CM	1.00	0.00	0.25
MD	1.00	0.00	0.08

δ test in summary

- We developped a test to detect heterogeneity in local densities
- We posit this heterogeneity is a necessary condition for clusterability
- Our test was shown to be more accurate and to rely on fewer hypotheses than recently introduced competing tests

Two graph test

The question

Given two graphs G_1 & G_2, is one more clusterable than the other?

Two graph test

The question

Given two graphs G_1 & G_2, is one more clusterable than the other?

The answer

- For each graph, sample and compute δ_1 & δ_2
- Use the TWO-SAMPLE t-test with unequal variances to compare
- Null is $\delta_1 = \delta_2$
- Alternative is $\delta_1 > \delta_2$

Statistical power

Objectives

- In our initial tests, we arbitrarily sampled 25% of nodes, computed local densities and tested their mean
- Our results were meaningful
- The question is can we sample fewer nodes/neighborhoods and still obtain meaningful conclusions? (looking for scalability, "big data" applications, etc...)

This (very recent & under review) work is presented in "Statistical power, accuracy, reproducibility and robustness of a graph clusterability test", [Miasnikof et al., 2022]

Definition of "statistical power"

Textbook definition:

The power of a test is the probability of **CORRECTLY** rejecting a null hypothesis. (typically tied to sample size, or loosely "robustness to sample size")

- With 25%, our results were meaningful (100% power)
- Can we sample fewer nodes/neighborhoods and still obtain meaningful conclusions (maintain 100% power)?
- To answer this question, we applied our test to samples of 0.5, 1 and 10% of nodes ($|V| \geq 10K$) and conducted 500 tests

Synthetic graphs (2022 tests)

- CC (connected caveman): $|V| = 10,000$ nodes divided in 200 cliques with 50 vertices per clique (one randomly selected edge is reassigned to connect to another clique)
- SBM (stochastic block model):
  - Mean intra-cluster edge probability $P_{\text{intra}} = 0.75$ (range $[0.68, 0.99]$)
  - Mean inter-cluster edge probability $P_{\text{inter}} = 0.30$ (range $[0.27, 0.33]$)
  - Mean vertices per clusters of $\bar{n}_i = 100$ (range $[80, 120]$)
- ER (Erdős-Rényi-Gilbert): Edge probability $p = 0.333$, $n\ (|V|) = 10,000$
- CM (configuration model) : $|V| = 10,000$, exponent$= 3$

	$\mathcal{K}$ (density)	$\lvert E\rvert$	$\lvert V\rvert$
CC	4.90E-03	245,000	10,000
SBM	0.30	21,043,009	11,752
ER	0.33	16,647,645	10,000
CM	2.46E-04	12,315	10,000

Real-world graphs (2022 tests, data from SNAP repository)

- DIMACS10 ("(...) snapshot of the structure of the Internet at the level of autonomous systems, reconstructed from BGP tables posted by the University of Oregon Route Views Project")
- Lancichinetti, Fortunato, Radicchi (LFR) graph
- Astro Physics collaboration network (arXiv ASTRO-PH)
- Enron email network
- DBLP collaboration network (co-authorship)

	$\mathcal{K}$ (density)	$\|E\|$	$\|V\|$
DIMACS10	1.84E-04	48,436	22,963
LFR	2.44E-04	1,220,023	100,000
Astro	1.12E-03	198,110	18,772
Enron	2.73E-04	183,831	36,692
DBLP	2.09E-05	1,049,866	317,080

Power analysis summary

- Even with very small sample sizes ($s = 0.5\% \times |V|$), our test retains perfect power
- Under some scenarios, our Gaussian null hypothesis is not accurate (esp. CM graphs)
- Nevertheless, our experiments demonstrate that our tests conclusions remain valid, even under severe departures from this null hypothesis

Robustness & departures from the test hypotheses

Recall

Hypotheses

- Under the null, δ follows a Gaussian distribution centered at 0
- Under the alternative, δ follows a Gaussian distribution centered at a point > 0

Recall

Hypotheses

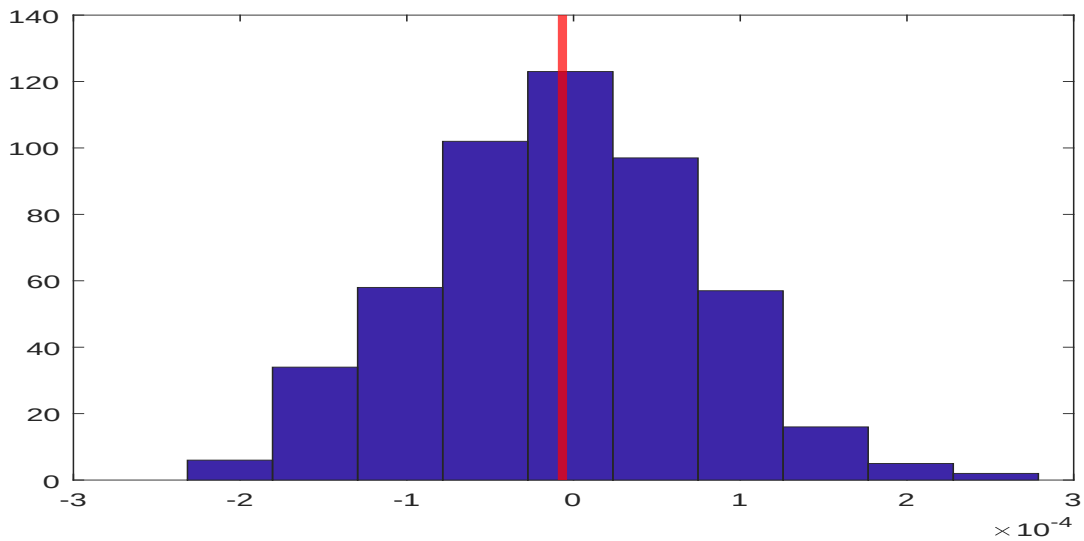
- Under the null, δ follows a Gaussian distribution centered at 0
- Under the alternative, δ follows a Gaussian distribution centered at a point > 0

Does it hold in reality?

- If it doesn't, what happens to the conclusions?
- After all, the CLT describes asymptotic properties...

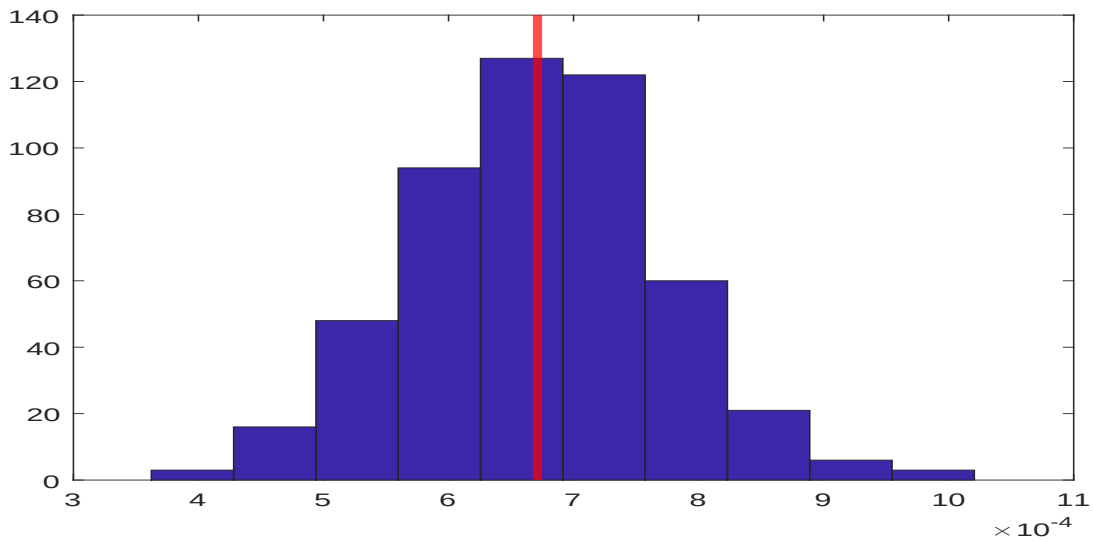
Empirical results with ERG graph (δ)

ERG graph with $p = 0.333$, 0.5% of nodes sampled, 500 experiments



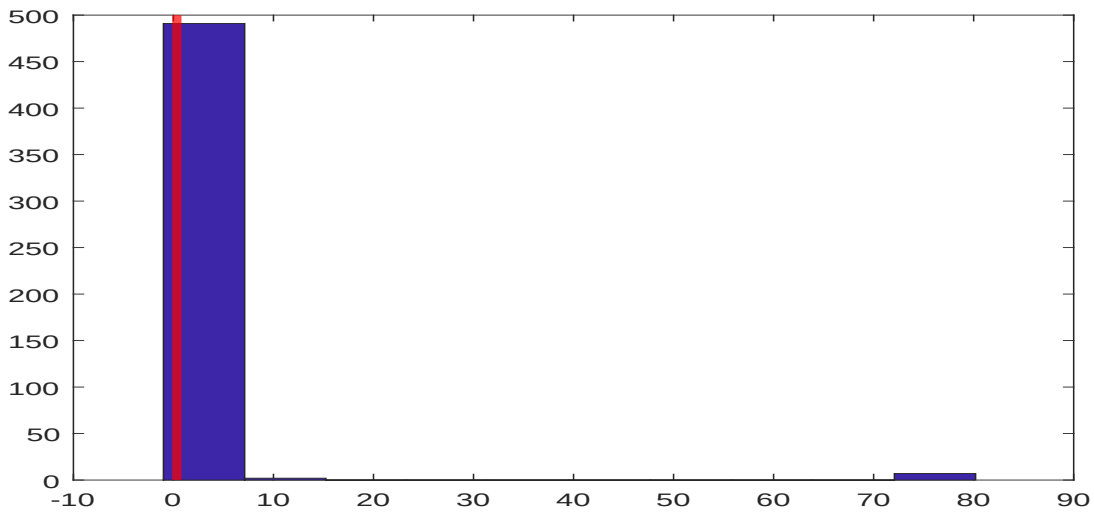
Empirical results with SBM graph (δ)

SBM, 0.5% of nodes sampled, 500 experiments



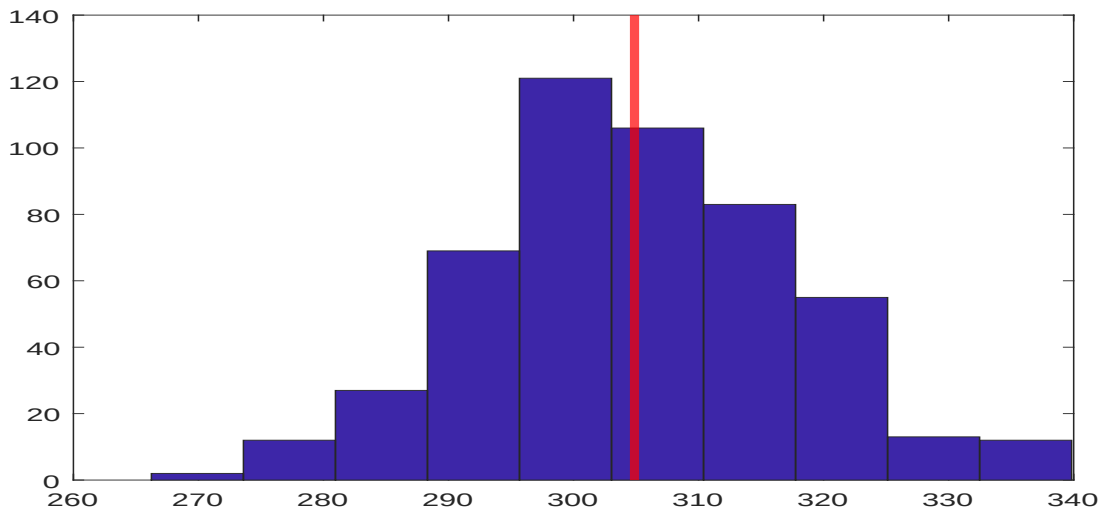
Empirical results with CM graph (δ)

Configuration model, 0.5% of nodes sampled, 500 experiments



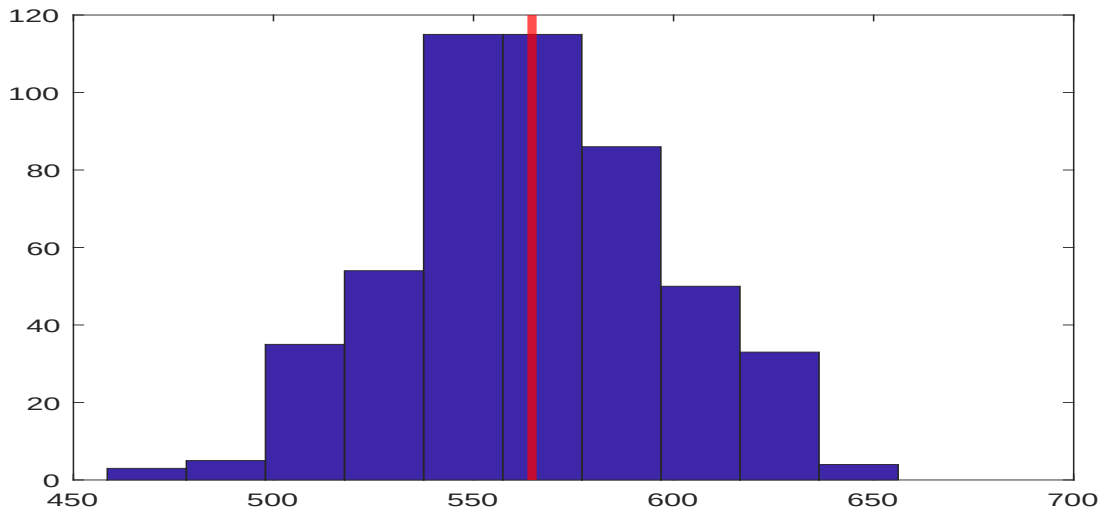
Empirical results with LFR graph (δ)

LFR, 0.5% of nodes sampled, 500 experiments



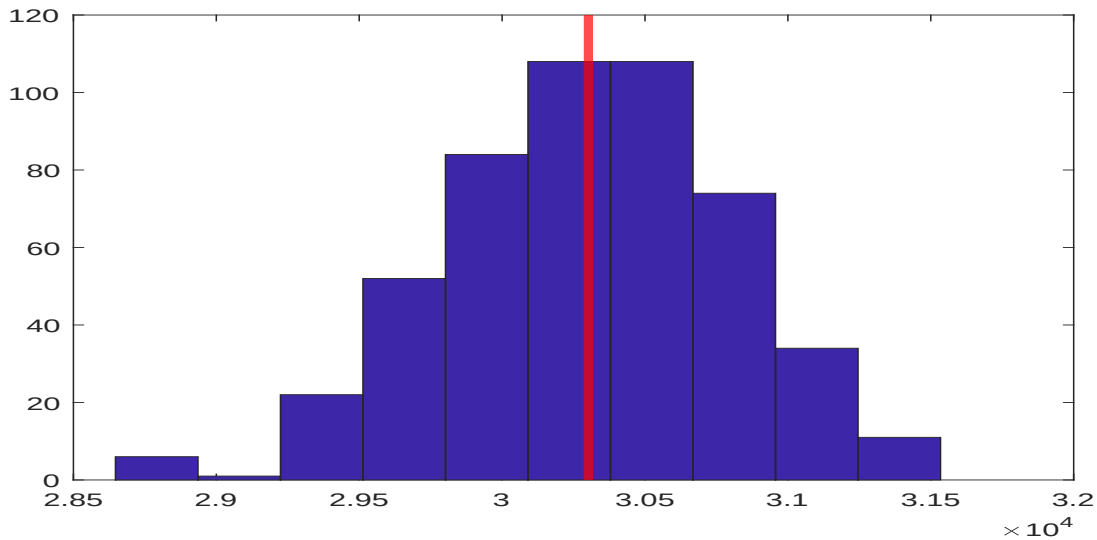
Empirical results with real-world graph (Astrophysics co-authorship) (δ)

ASTRO PH, 0.5% of nodes sampled, 500 experiments



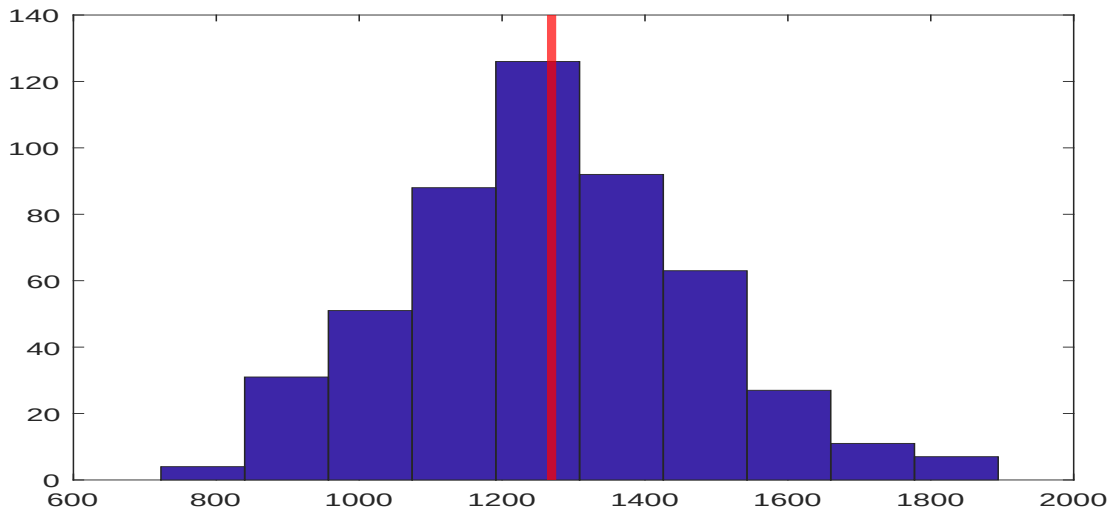
Empirical results with real-world graph (DBLP co-authorship) (δ)

DBLP co-authorship, 0.5% of nodes sampled, 500 experiments



Empirical results with real-world graph (DIMACS10 internet autonomous systems) (δ)

DIMACS10, 0.5% of nodes sampled, 500 experiments



Summary

Assumptions of the δ test

- Clusterable graphs display heterogenous local (induced subgraph) density

Assumptions of the δ test

- Clusterable graphs display heterogenous local (induced subgraph) density
- The existence of such heterogeneity is a NECESSARY (not sufficient) condition for clusterability

Assumptions of the δ test

- Clusterable graphs display heterogenous local (induced subgraph) density
- The existence of such heterogeneity is a NECESSARY (not sufficient) condition for clusterability
- Unlike other tests in the literature, our test is centered on very weak, transparent & verifiable (and reasonable) assumptions

Empirical test results

- Our test correctly identifies non-clusterable graphs

Empirical test results

- Our test correctly identifies non-clusterable graphs
- It remains accurate even with very small sample sizes

Empirical test results

- Our test correctly identifies non-clusterable graphs
- It remains accurate even with very small sample sizes
- Even under severe departures from the underlying hypotheses (Gaussian null & alternative), the test remains accurate

THANK YOU!


Chiplunkar, A., Kapralov, M., Khanna, S., Mousavifar, A., and Peres, Y. (2018).
Testing Graph Clusterability: Algorithms and Lower Bounds.
ArXiv e-prints.


Gao, C. and Lafferty, J. (2017a).
Testing for Global Network Structure Using Small Subgraph Statistics.
ArXiv e-prints.


Gao, C. and Lafferty, J. (2017b).
Testing Network Structure Using Relations Between Small Subgraph Probabilities.
ArXiv e-prints.


Miasnikof, P., Prokhorenkova, L., Shestopaloff, A., and Raigorodskii, A. (2019).
A statistical test of heterogeneous subgraph densities to assess clusterability.
Springer LNCS.


Miasnikof, P., Shestopaloff, A., and Raigorodskii, A. (2022).

Statistical power, accuracy, reproducibility and robustness of a graph clusterability test.
under review.

Reichardt, J. and Bornholdt, S. (2006).
When are networks truly modular?
Physica D: Nonlinear Phenomena, 224(1):20 – 26.
Dynamics on Complex Networks and Applications.