Clustering quality

7th Winter School on Data Analytics (DA 2022)

Pierre Miasnikof

17 November, 2022

Acknowledgements

Joint work

This is joint work with A.Y. Shestopaloff, A. Bonner, Y. Lawryshyn and P.M. Pardalos.

Joint work

This is joint work with A.Y. Shestopaloff, A. Bonner, Y. Lawryshyn and P.M. Pardalos.

Thanks!

We would like to thank:

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

Content based on

Today's lecture is based on

- "A Statistical Performance Analysis of Graph Clustering Algorithms", LNCS (2018), (PM, Shestopaloff, Bonner, Lawryshyn)
- "A density-based statistical analysis of graph clustering algorithm performance", Jnl of Complex Networks (2020), (PM, Shestopaloff, Bonner, Lawryshyn, Pardalos)

Introduction

Objectives for this lecture

1. Introduce the graph clustering quality question

Objectives for this lecture

1. Introduce the graph clustering quality question
2. Briefly discuss the widely used modularity and conductance

Objectives for this lecture

1. Introduce the graph clustering quality question
2. Briefly discuss the widely used modularity and conductance
3. Introduce our own (statistical) test

Objectives for this lecture

1. Introduce the graph clustering quality question
2. Briefly discuss the widely used modularity and conductance
3. Introduce our own (statistical) test
4. Show comparative performances

Objectives for this lecture

1. Introduce the graph clustering quality question
2. Briefly discuss the widely used modularity and conductance
3. Introduce our own (statistical) test
4. Show comparative performances
5. Discuss our technique's mathematical justification

Objectives for this lecture

1. Introduce the graph clustering quality question
2. Briefly discuss the widely used modularity and conductance
3. Introduce our own (statistical) test
4. Show comparative performances
5. Discuss our technique's mathematical justification

Limitations for this talk

For today, let's assume

- Undirected
- Unweighted
- No self-loops

A note on vocabulary: I will use these terms interchangeably

- "graph" and "network"
- "vertex" and "node"
- "edge" and "connection"
- "clustering" and "community detection"

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 clustering algorithm doing a good job?
- Do the clusters identified by this algorithm offer a meaningful summary of the graph?

The focal question

Context:

Graph clustering, also known as community detection

Still open questions: (hopefully partially answered here)

- Is the clustering algorithm doing a good job?
- Do the clusters identified by this algorithm offer a meaningful summary of the graph?

Key facts:

- No formal definition of "graph clusters"

The focal question

Context:

Graph clustering, also known as community detection

Still open questions: (hopefully partially answered here)

- Is the clustering algorithm doing a good job?
- Do the clusters identified by this algorithm offer a meaningful summary of the graph?

Key facts:

- No formal definition of "graph clusters"
- BUT, consensus is they form densely connected subsets of nodes, with sparse connections to the remaining graph (more in than out connetions)

The focal question

Context:

Graph clustering, also known as community detection

Still open questions: (hopefully partially answered here)

- Is the clustering algorithm doing a good job?
- Do the clusters identified by this algorithm offer a meaningful summary of the graph?

Key facts:

- No formal definition of "graph clusters"
- BUT, consensus is they form densely connected subsets of nodes, with sparse connections to the remaining graph (more in than out connetions)
- A few clustering quality functions in the literature

Why do we care?

In the recent literature!

" (...) there is no universally accepted metric for evaluating the performance of community detection algorithms. "
[Prokhorenkova, 2019]

Why do we care?

In the recent literature!

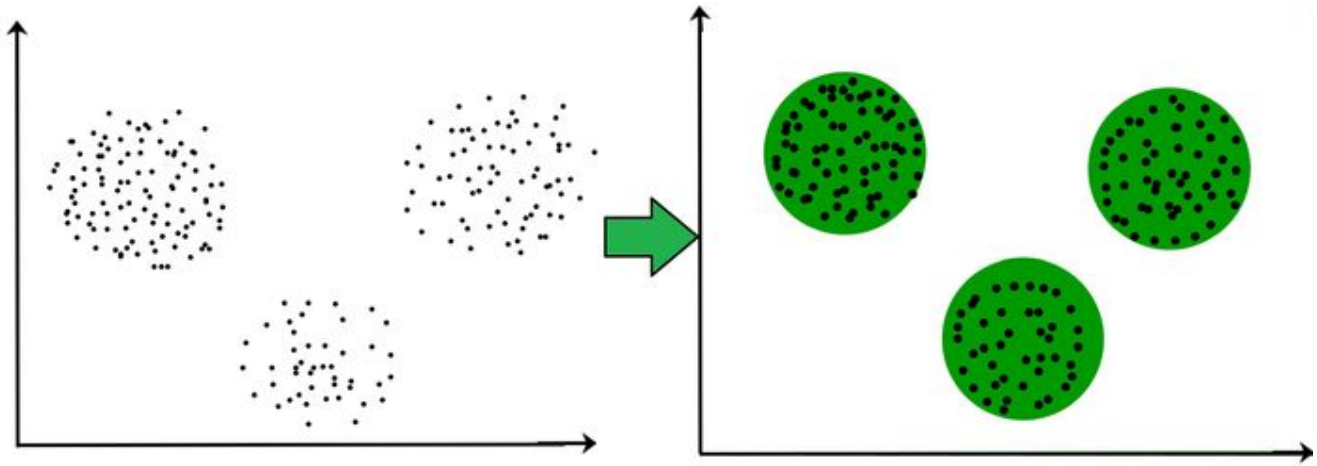
" (...) there is no universally accepted metric for evaluating the performance of community detection algorithms. "
[Prokhorenkova, 2019]

For this reason, we posit that our only quality measure is a thorough examination of the graph's and resulting clusters and connectivity patterns.

Clustering from the beginning

Typical case with numerical data (e.g., $\mathbb{R}^n$)

- Clusters are formed by data points that are deemed similar
- Similarity is typically measured by (Euclidean) distance
- "Clustering" is the process of identifying members of each pocket of similarity (subset)



The graph clustering case is different

Similarity between nodes?

- Things are not as clear-cut

The graph clustering case is different

Similarity between nodes?

- Things are not as clear-cut
- Typically we don't have the "distance" between vertices

The graph clustering case is different

Similarity between nodes?

- Things are not as clear-cut
- Typically we don't have the "distance" between vertices
- In fact, there are many "distances" between vertices (geodesic, commute,...)

The graph clustering case is different

Similarity between nodes?

- Things are not as clear-cut
- Typically we don't have the "distance" between vertices
- In fact, there are many "distances" between vertices (geodesic, commute,...)
- SO, K-means/silhouette type clustering techniques/quality measures do not work... We need something else!

What are vertex clusters?

A still open question

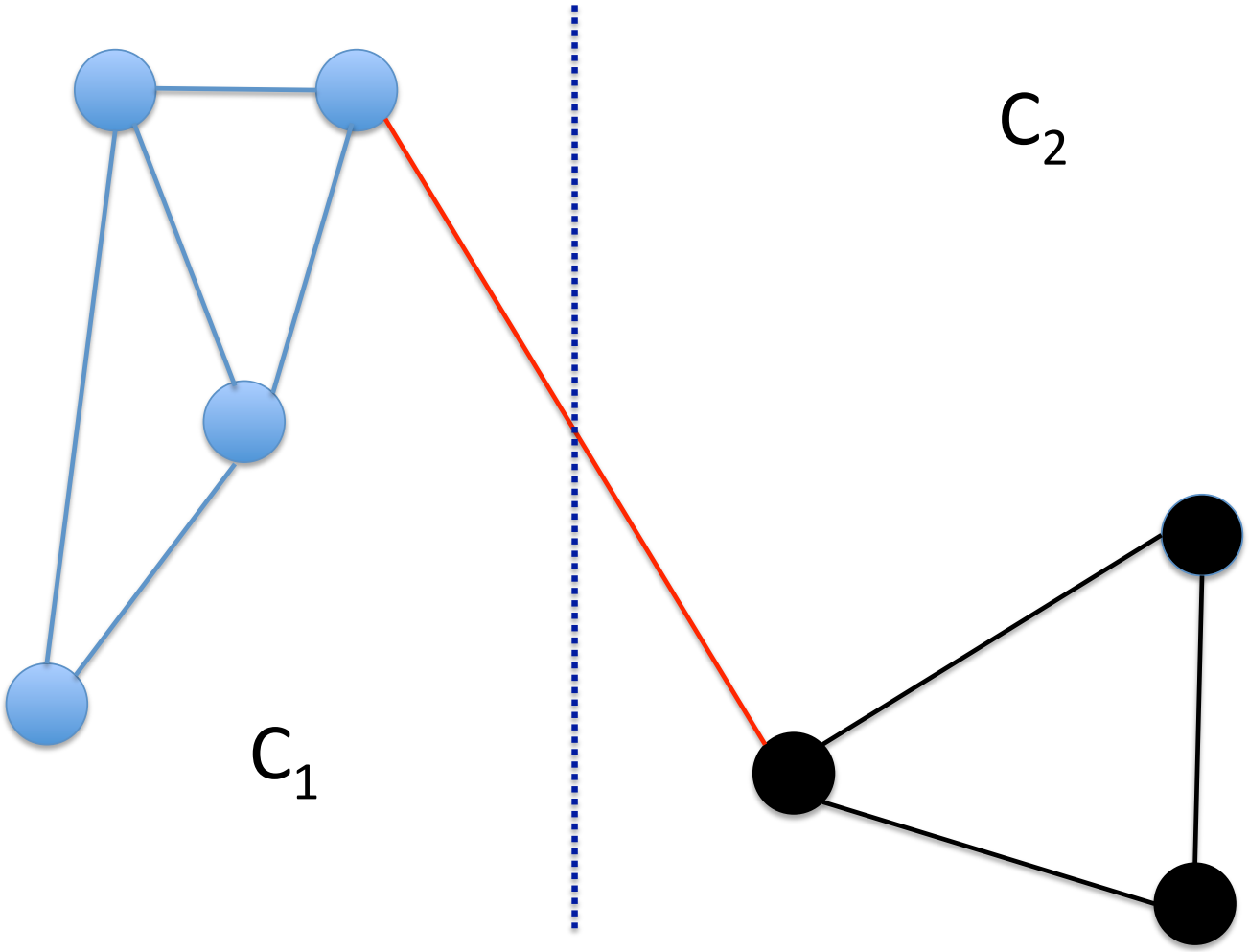
- A formal definitions of "vertex clusters" does not exist (yet)
- However, the consensus is that clusters are
  - Densely connected sets of vertices
  - That display sparse connections to the remaining vertices

What are vertex clusters?

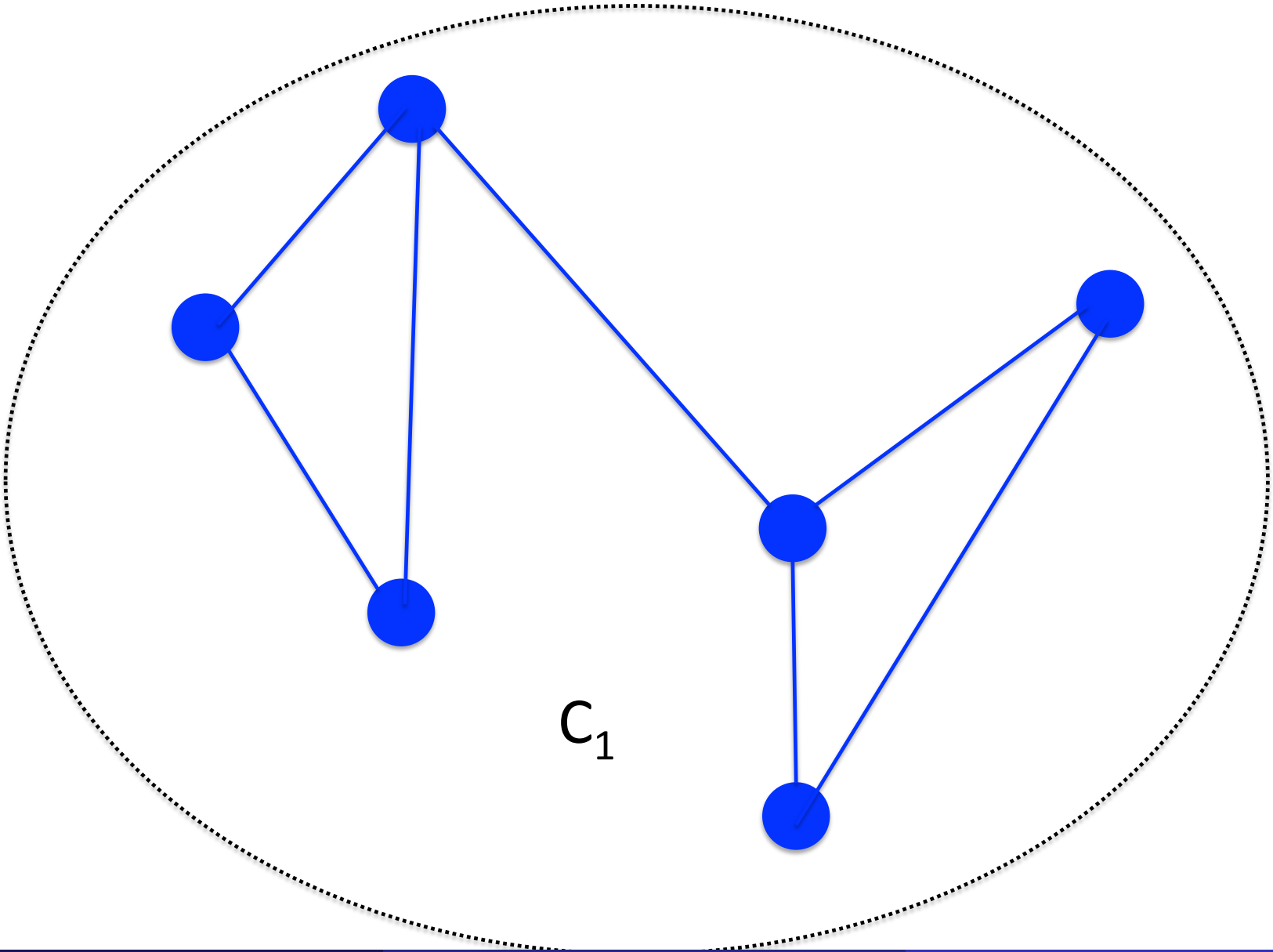
A still open question

- 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

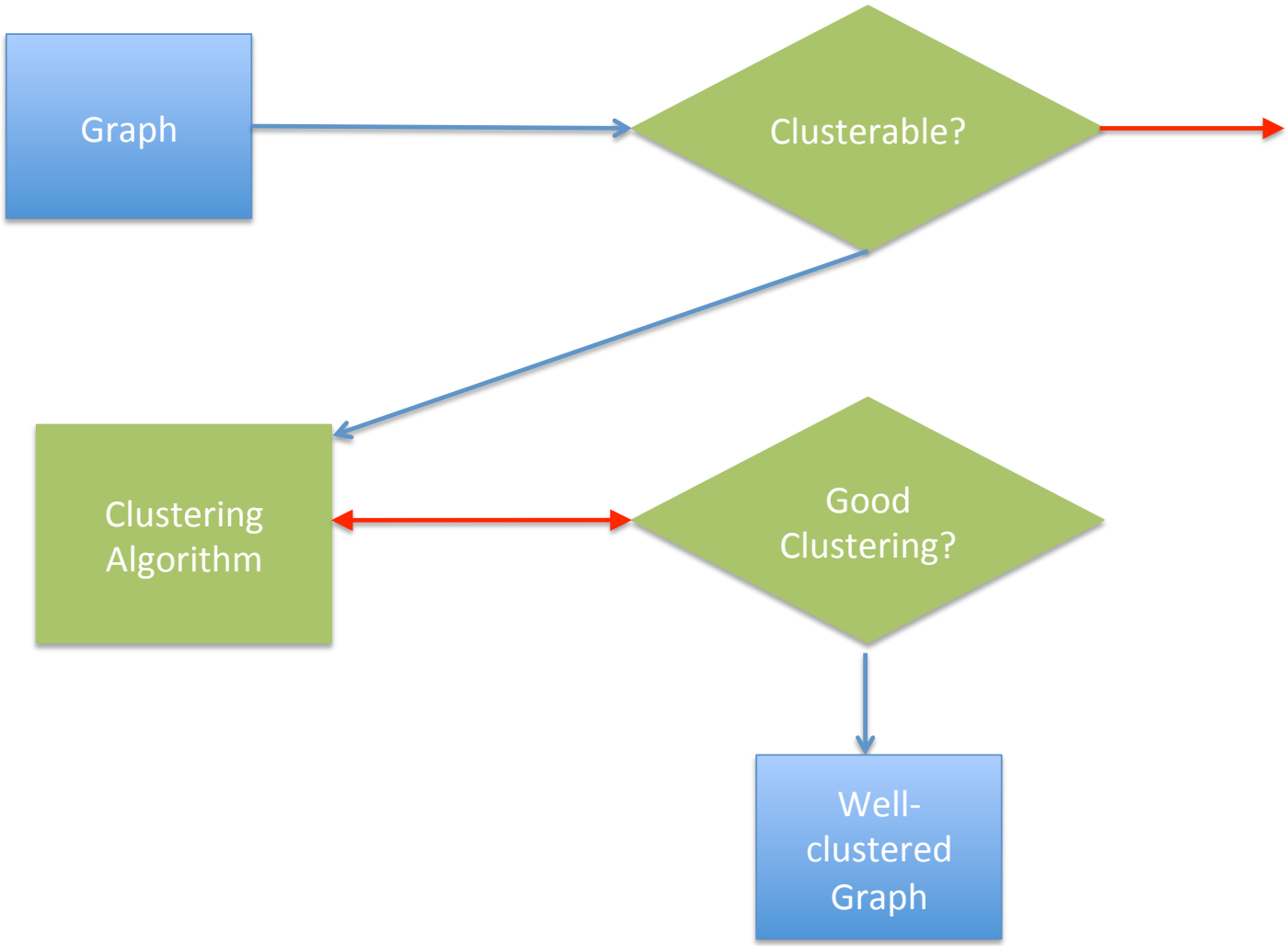
We want to answer: does the algorithm group in this way?



OR like this?



Ideal workflow for graph clustering



Clustering quality functions

What are they?

- A function that takes in a graph G and a set of node clusters $C_G^a = \{c_1, c_2, ...\}$ (for clearer notation, I'll just use C) and returns a real number

What are they?

- A function that takes in a graph G and a set of node clusters $C_G^a = \{c_1, c_2, ...\}$ (for clearer notation, I'll just use C) and returns a real number
- The two most widely used are modularity and conductance

What are they?

- A function that takes in a graph G and a set of node clusters $C_G^a = \{c_1, c_2, ...\}$ (for clearer notation, I'll just use C) and returns a real number
- The two most widely used are modularity and conductance
- Roughly speaking: a function that tells us how good the clustering is

What are they?

- A function that takes in a graph G and a set of node clusters $C_G^a = \{c_1, c_2, ...\}$ (for clearer notation, I'll just use C) and returns a real number
- The two most widely used are modularity and conductance
- Roughly speaking: a function that tells us how good the clustering is
- It tells us how homogeneous the subsets are

Modularity

$$Q = \sum_{i=1}^{k} \left(\underbrace{e_{ii} - a_i^2}_{q_i} \right) (\in [-0.5, 1])$$

where,

$$e_{ii} = \frac{1}{2m} \sum_{v,w} A_{v,w}\, \delta(c_v, i)\delta(c_w, i)$$

$$a_i = \frac{1}{2m} \sum_{v} A_{v,.}\vec{1}\, \delta(c_v, i)\,.$$

Modularity (cont'd)

$$Q = \sum_{i=1}^{k} \left[\underbrace{\frac{1}{2m} \sum_{v,w} A_{v,w}\, \delta(c_v, i)\delta(c_w, i)}_{e_{ii}} - \underbrace{\frac{1}{4m^2} \left(\sum_{v} A_{v,.} \vec{1}\, \delta(c_v, i) \right)^2}_{a_i^2} \right]$$

Modularity (cont'd)

$$Q = \sum_{i=1}^{k} \left[\underbrace{\frac{1}{2m} \sum_{v,w} A_{v,w}\, \delta(c_v, i)\delta(c_w, i)}_{e_{ii}} - \underbrace{\frac{1}{4m^2} \left(\sum_{v} A_{v,.} \vec{1}\, \delta(c_v, i) \right)^2}_{a_i^2} \right]$$

We want HIGH modularity (> 0.3 is considered good)

Conductance

At the individual cluster level, conductance is defined as

$$\phi(c_i) = \frac{\partial(c_i)}{\min\left(d(c_i), d(V \setminus c_i)\right)}\,.$$

While at the graph level, it is defined as

$$\Phi(G) = \min_{c_i} \phi(c_i)\,.$$

Conductance

At the individual cluster level, conductance is defined as

$$\phi(c_i) = \frac{\partial(c_i)}{\min\left(d(c_i), d(V \setminus c_i)\right)}\,.$$

While at the graph level, it is defined as

$$\Phi(G) = \min_{c_i} \phi(c_i)\,.$$

We want low conductance

CAVEAT EMPTOR!

- BOTH these functions are HIGHLY FLAWED

CAVEAT EMPTOR!

- BOTH these functions are HIGHLY FLAWED
- In no small part, because they violate the axioms for a good clustering function (Van Laarhoven and Marchiori, Kehagias and Pitsoulis, ...)

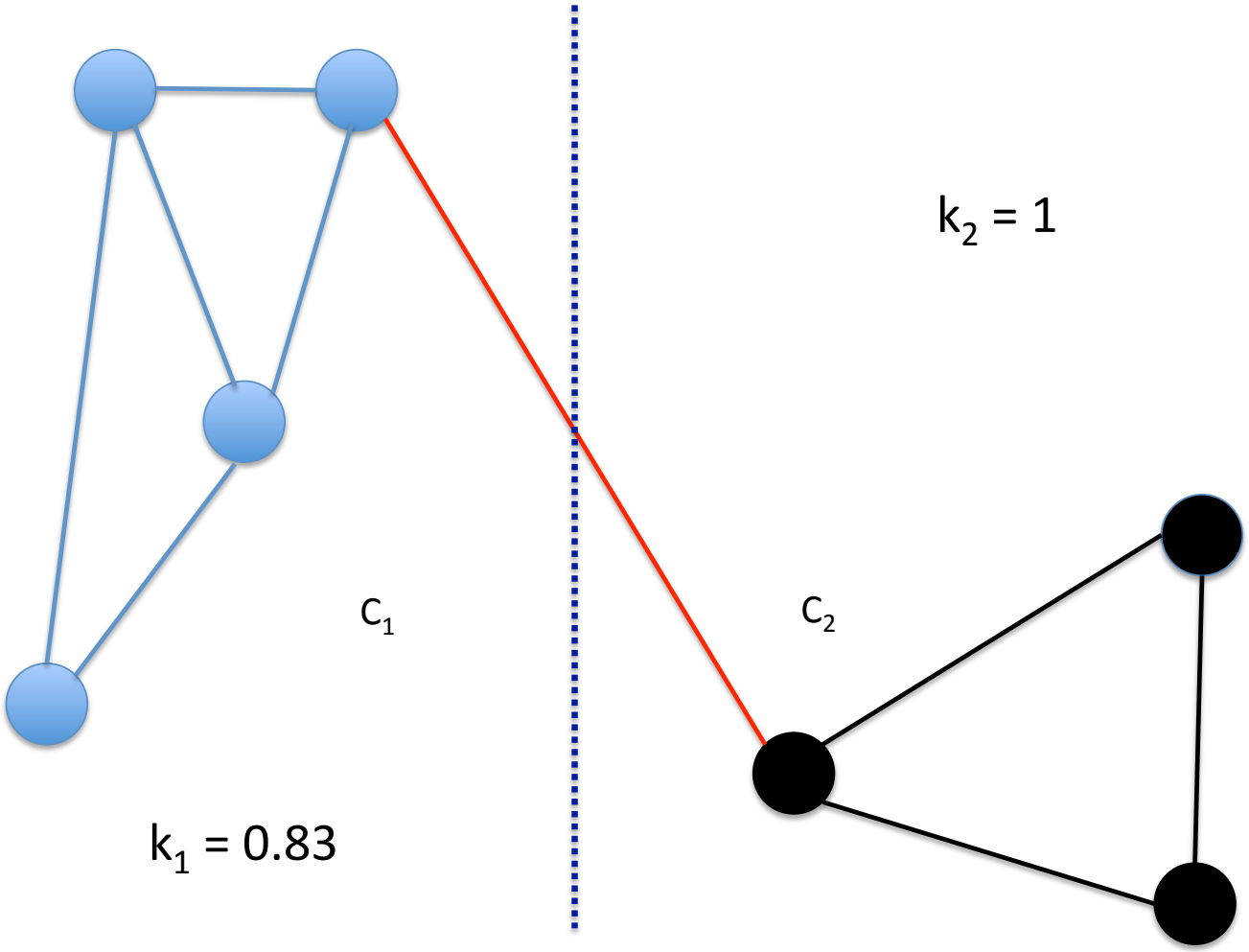
CAVEAT EMPTOR!

- BOTH these functions are HIGHLY FLAWED
- In no small part, because they violate the axioms for a good clustering function (Van Laarhoven and Marchiori, Kehagias and Pitsoulis, ...)
- We will not cover this topic in great detail, today...

CAVEAT EMPTOR!

- BOTH these functions are HIGHLY FLAWED
- In no small part, because they violate the axioms for a good clustering function (Van Laarhoven and Marchiori, Kehagias and Pitsoulis, ...)
- We will not cover this topic in great detail, today...
- BUT we will review results that illustrate some of these flaws

Our tests, intuition



k_{12}= 0.08
K = 0.43

Densities

Graph's overall density

$$K = \frac{|E|}{0.5 \times N(N-1)}$$

Densities

Graph's overall density

$$K = \frac{|E|}{0.5 \times N(N-1)}$$

INTRA-cluster density

$$\kappa_i = \frac{|e_{ii}|}{0.5 \times n_i(n_i - 1)}$$

Densities

Graph's overall density

$$K = \frac{|E|}{0.5 \times N(N-1)}$$

INTRA-cluster density

$$\kappa_i = \frac{|e_{ii}|}{0.5 \times n_i(n_i - 1)}$$

INTER-cluster density

$$\kappa_{ij} = \frac{|e_{ij}|}{n_i \times n_j}$$

To get an aggregate (graph-level) view

We take means:

To get an aggregate (graph-level) view

We take means:

MEAN INTRA-cluster density

$$\bar{K}_{\text{intra}} = \frac{1}{|C|} \sum_{i} \kappa_i$$

To get an aggregate (graph-level) view

We take means:

MEAN INTRA-cluster density

$$\bar{K}_{\text{intra}} = \frac{1}{|C|} \sum_i \kappa_i$$

MEAN INTER-cluster density

$$\bar{K}_{\text{inter}} = \frac{1}{0.5 \times |C| \times (|C| - 1)} \sum_i \sum_{j=i+1} \kappa_{ij}$$

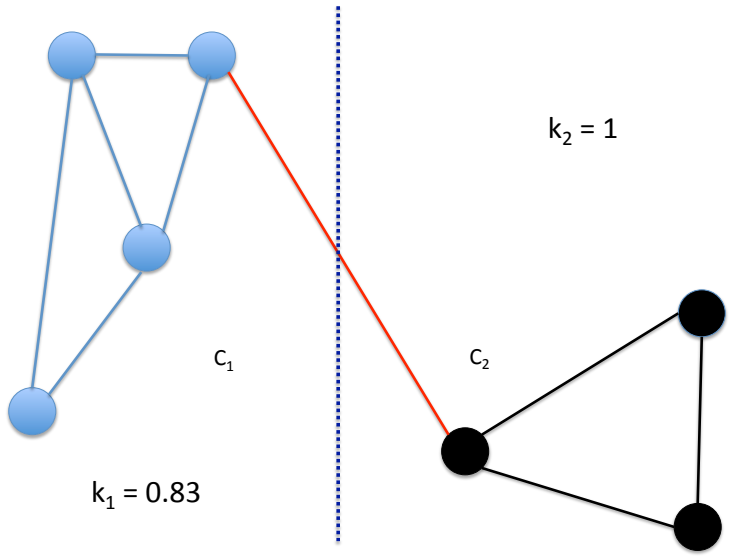
Measuring clustering quality

We extend the "general consensus" and claim that a good clustering should result in the following inequalities holding:

$$\bar{K}_{\text{inter}} < K < \bar{K}_{\text{intra}}$$

(general consensus: more connections within than between)

Earlier example revisited



$$\bar{K}_{\text{intra}} = \frac{1}{2}(1 + 0.83) = 0.92$$

$$\bar{K}_{\text{inter}} = \left(\frac{1}{0.5 \times 2 \times 1}\right)\frac{1}{4 \times 3} = 0.08$$

$$K = \frac{9}{0.5 \times 7 \times 6} = 0.43$$

Comparisons using synthetic data

Varying Intra-Cluster Connectivity, with Noise from Inter-Cluster Connectivity, to study effect on quality functions

Pct Inter = 100, Pct Intra varies					
Pct Intra	0	25	50	75	100
N	10,048	9,725	10,374	9,490	9,700
$\|C\|$	200	200	200	200	200
$\|E\|$	50,142,500	47,052,100	53,631,600	44,950,500	47,040,200
K	0.9934	0.9951	0.9968	0.9983	1.0000
$\bar{K}_{\text{intra}}$	0.0000	0.2640	0.4995	0.7523	0.9900
$\bar{K}_{\text{inter}}$	1.0000	1.0000	1.0000	1.0000	1.0000
Φ	1.0000	0.9974	0.9952	0.9922	0.9899
Q	-0.0067	-0.0050	-0.0033	-0.0018	-0.0001

Probabilistic interpretation: another strength

Let the existence (or not) of an edge between two vertices be given by a Bernoulli trial

Graph's overall density as the empirical estimate of edge probability

$$\begin{aligned} K &= \frac{\text{actual number of edges}}{\text{possible number of edges}} \left(= \frac{\text{number of successes}}{\text{number of trials}} \right) \\ &= \hat{P}(e_{ij}) \end{aligned}$$

Extending to intra/inter

MEAN INTRA-cluster density as the empirical estimate of intra-cluster edge probability

$$\begin{aligned} \kappa_1 &= \hat{P}(e_{ij}|c_i = c_j = 1) \text{ (one example)} \\ \Rightarrow \bar{K}_{\text{intra}} &= \hat{P}(e_{ij}|c_i = c_j) \text{ (under very minor assumptions)} \end{aligned}$$

Extending to intra/inter

MEAN INTRA-cluster density as the empirical estimate of intra-cluster edge probability

$$\begin{aligned} \kappa_1 &= \hat{P}(e_{ij}|c_i = c_j = 1) \text{ (one example)} \\ \Rightarrow \bar{K}_{\text{intra}} &= \hat{P}(e_{ij}|c_i = c_j) \text{ (under very minor assumptions)} \end{aligned}$$

MEAN INTER-cluster density as the empirical estimate of inter-cluster edge probability

$$\begin{aligned} \kappa_{12} &= \hat{P}(e_{ij}|c_i = 1, c_j = 2) \text{ (another example)} \\ \Rightarrow \bar{K}_{\text{inter}} &= \hat{P}(e_{ij}|c_i \neq c_j) \text{ (again, under minor assumptions)} \end{aligned}$$

Hypothesis tests

Test overview

- One of the many problems with modularity (& conductance) is the difficulty to interpret them formally

Test overview

- One of the many problems with modularity (& conductance) is the difficulty to interpret them formally
- Our technique is based on *MEAN* densities and these quantities have a probabilistic interpretation

Test overview

- One of the many problems with modularity (& conductance) is the difficulty to interpret them formally
- Our technique is based on *MEAN* densities and these quantities have a probabilistic interpretation
- They can also be understood as one possible instance (sample) from a population of all possible clusterings

Test overview

- One of the many problems with modularity (& conductance) is the difficulty to interpret them formally
- Our technique is based on *MEAN* densities and these quantities have a probabilistic interpretation
- They can also be understood as one possible instance (sample) from a population of all possible clusterings
- These features allow for formal significance tests

Test overview

- One of the many problems with modularity (& conductance) is the difficulty to interpret them formally
- Our technique is based on *MEAN* densities and these quantities have a probabilistic interpretation
- They can also be understood as one possible instance (sample) from a population of all possible clusterings
- These features allow for formal significance tests
- We offer TWO necessary conditions that, together, are sufficient to identify good clustering

Test overview

- One of the many problems with modularity (& conductance) is the difficulty to interpret them formally
- Our technique is based on *MEAN* densities and these quantities have a probabilistic interpretation
- They can also be understood as one possible instance (sample) from a population of all possible clusterings
- These features allow for formal significance tests
- We offer TWO necessary conditions that, together, are sufficient to identify good clustering
    - The inequalities holding

Test overview

- One of the many problems with modularity (& conductance) is the difficulty to interpret them formally
- Our technique is based on *MEAN* densities and these quantities have a probabilistic interpretation
- They can also be understood as one possible instance (sample) from a population of all possible clusterings
- These features allow for formal significance tests
- We offer TWO necessary conditions that, together, are sufficient to identify good clustering
  - The inequalities holding
  - The γ statistic being statistically significant

TWO necessary conditions to form ONE sufficient condition

Necessary #1

The inequality $\bar{K}_{\text{inter}} < K < \bar{K}_{\text{intra}}$ MUST HOLD

TWO necessary conditions to form ONE sufficient condition

Necessary #1

The inequality $\bar{K}_{\text{inter}} < K < \bar{K}_{\text{intra}}$ MUST HOLD

Necessary #2

The statistic $\gamma = \bar{K}_{\text{intra}} - \bar{K}_{\text{inter}}$ must be statistically significantly greater than zero

The γ statistic in detail

$$\gamma = \bar{K}_{\text{intra}} - \bar{K}_{\text{inter}} \quad (1)$$

$$= \frac{1}{|C|}\sum_i \kappa_i - \frac{1}{0.5 \times |C| \times (|C|-1)} \sum_i \sum_{j=i+1} \kappa_{ij} \quad (2)$$

The γ statistic in detail

$$\gamma = \bar{K}_{\text{intra}} - \bar{K}_{\text{inter}} \quad (1)$$

$$= \frac{1}{|C|}\sum_i \kappa_i - \frac{1}{0.5 \times |C| \times (|C|-1)} \sum_i \sum_{j=i+1} \kappa_{ij} \quad (2)$$

In Equation 2, we see γ is the difference of two sums

The γ statistic in detail

$$\gamma = \bar{K}_{\text{intra}} - \bar{K}_{\text{inter}} \quad (1)$$

$$= \frac{1}{|C|}\sum_i \kappa_i - \frac{1}{0.5 \times |C| \times (|C|-1)} \sum_i \sum_{j=i+1} \kappa_{ij} \quad (2)$$

In Equation 2, we see γ is the difference of two sums

If $|C| > 30$, we can invoke the CLT and declare γ to be "approximately Gaussian"

The γ statistic in detail

$$\gamma = \bar{K}_{\text{intra}} - \bar{K}_{\text{inter}} \quad (1)$$

$$= \frac{1}{|C|}\sum_i \kappa_i - \frac{1}{0.5 \times |C| \times (|C|-1)} \sum_i \sum_{j=i+1} \kappa_{ij} \quad (2)$$

In Equation 2, we see γ is the difference of two sums

If $|C| > 30$, we can invoke the CLT and declare γ to be "approximately Gaussian"

WE CAN USE THE (Student) t-test TO TEST ITS SIGNIFICANCE!

Distribution of γ under the null

Gaussian approximation

- As discussed earlier, in order to use the t-test, we need γ to be (approximately) Gaussian
- This property is garanteed by the CLT

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`

Another important statistical property: the LLN (law of large numbers)

In a few words:

The larger the sample, the closer the mean will be to its expected value

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{x_i}{n} = E(x) = \mu$$

Why do we care about these things?

Because

- We want to justify our test

Why do we care about these things?

Because

- We want to justify our test
- We want to understand the test statistic and its distribiution

Why do we care about these things?

Because

- We want to justify our test
- We want to understand the test statistic and its distribiution
- Sensitivity to sample size is **KEY**

Null and alternative hypotheses

Null

H_0: γ is statistically indistinguishable from 0 (i.e. $\bar{K}_{\text{intra}} \approx \bar{K}_{\text{inter}}$)
$\Rightarrow$ The graph is poorly clustered!

Null and alternative hypotheses

Null

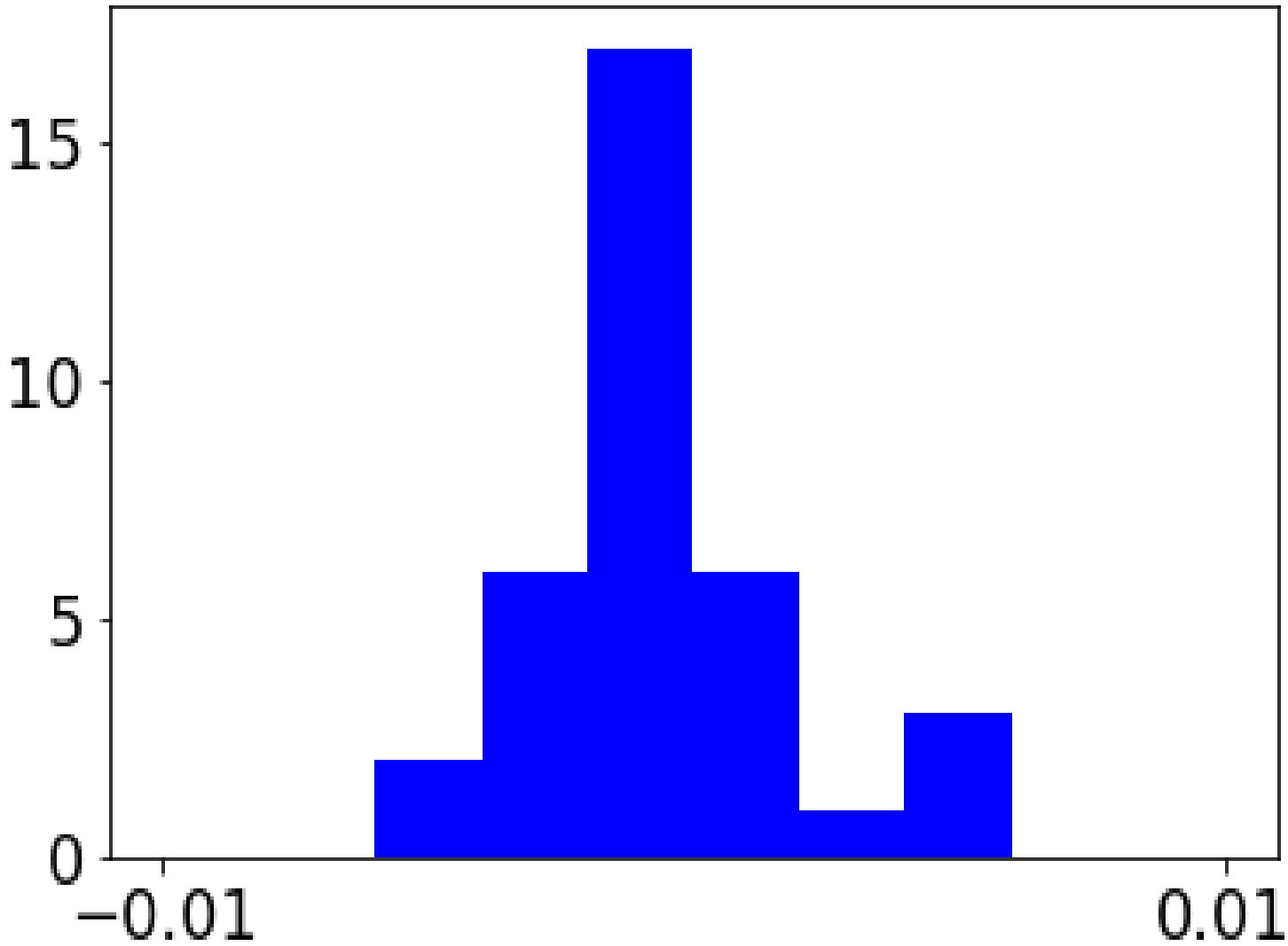
H_0: γ is statistically indistinguishable from 0 (i.e. $\bar{K}_{\text{intra}} \approx \bar{K}_{\text{inter}}$)
$\Rightarrow$ The graph is poorly clustered!

Alternative

H_a: γ is statistically different from 0 (i.e. $\bar{K}_{\text{intra}} > \bar{K}_{\text{inter}}$)
$\Rightarrow$ The graph is well clustered!

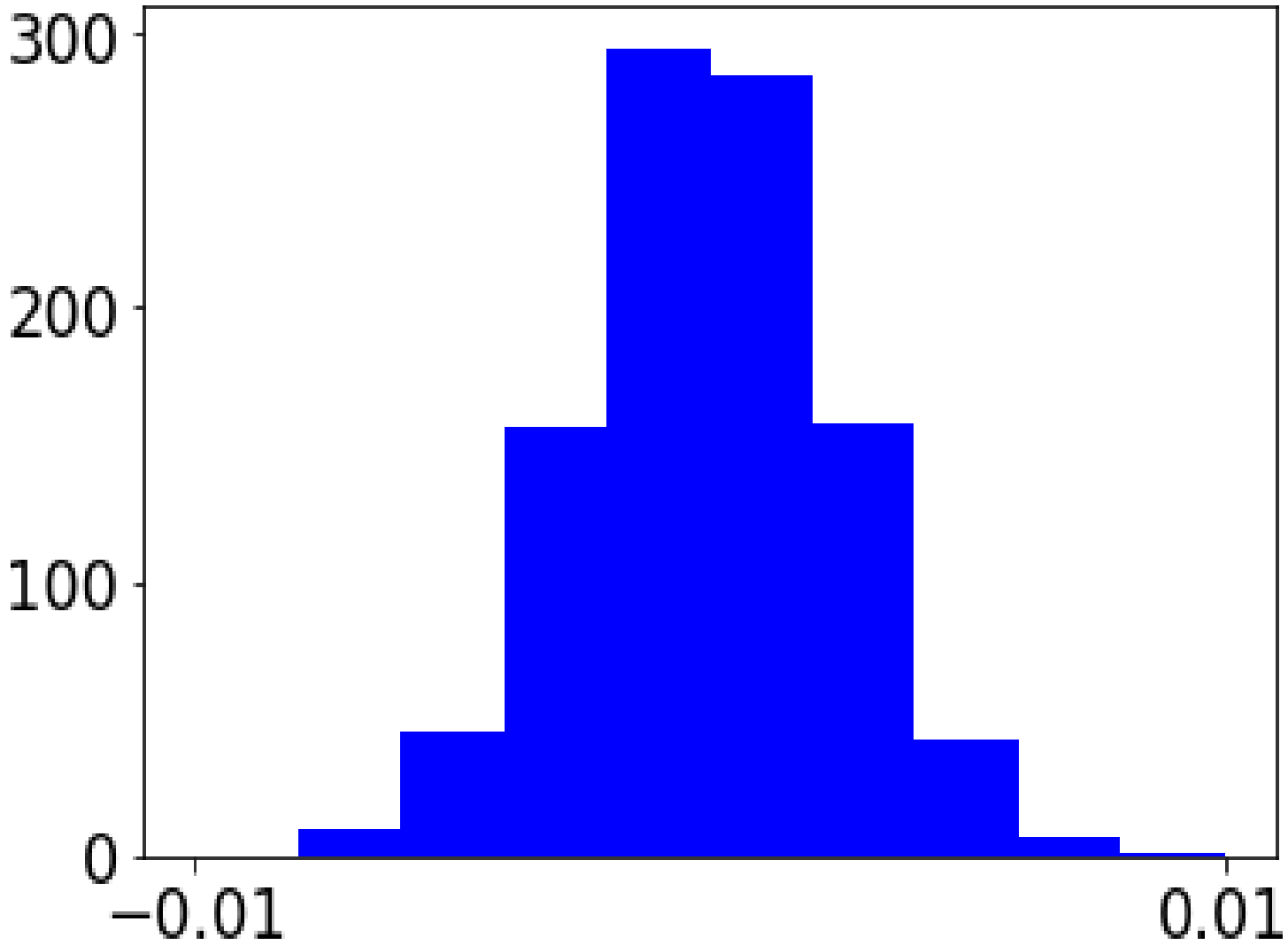
Null distribution, small sample

ER graph random assignment 35 runs ($\gamma \approx 0$)



Null distribution, small sample

ER graph random assignment 1K runs ($\gamma \approx 0$)



SUMMARY: Our proposed test procedure

Clustering Quality Assessment Routine:

- Use clustering algorithm labels to compute the Kappas
- Numerically verify that the inequalities $\bar{K}_{\text{inter}} < K < \bar{K}_{\text{intra}}$ hold (first necessary condition)
- If they don't and if the number of clusters is greater than one, conclude the algorithm has poorly clustered the graph
- If they don't and if the number of clusters is one, use global density to assess clustering quality
- If they do hold and if the number of clusters is sufficiently large, perform statistical test to verify significance (second necessary condition) of the difference

$$\gamma = \bar{K}_{\text{intra}} - \bar{K}_{\text{inter}}$$

Summary cont'd

BOTTOM LINE: DO NOT USE MODULARITY OR CONDUCTANCE!

Summary cont'd

BOTTOM LINE: DO NOT USE MODULARITY OR CONDUCTANCE!

THEY DON'T WORK!

THANK YOU!

Contact info, please feel free!

p.miasnikof@mail.utoronto.ca


Prokhorenkova, L. (2019).
Using synthetic networks for parameter tuning in community detection.
arXiv e-prints, page arXiv:1906.04555.