International Journal of Data Science and Analytics (2023) 15:379–390
https://doi.org/10.1007/s41060-023-00389-6

REGULAR PAPER

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

Pierre Miasnikof[1] · **Alexander Y. Shestopaloff**[2,3] · **Andrei Raigorodskii**[4,5]

Received: 28 June 2022 / Accepted: 5 March 2023 / Published online: 16 April 2023

Abstract
Not all graphs are clusterable. Not all graphs have a clustered structure and can be meaningfully summarized through vertex clustering. Clusterable graphs are characterized by pockets of densely connected vertices that are only sparsely connected to the remaining graph. In this article, we re-introduce a very simple and intuitive, yet highly informative, statistical hypothesis test for graph clusterability that is based on vertex and neighborhood samples. The goal of this test is to determine if a graph meets the necessary structural conditions to be summarized meaningfully through vertex clusters. 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. The test is also applied to graph comparisons, to test whether one graph has a stronger clustered structure than another. Significance is assessed using the t-statistic. Since it is based on sampling, we provide a focused examination of our test's sensitivity to sample size. The main contribution of this article is a detailed examination of our test's accuracy, sensitivity to sample size, conclusion reproducibility and robustness. Our empirical results remain consistent with our earlier conclusions and demonstrate the almost perfect accuracy of our test, even with very small samples of the graph. They also reveal that our test remains robust even under severe departures from the null hypothesis.

Keywords Clusterability · Clustering · Graph clustering · Data sciences · Complex networks · Significance testing

1 Introduction

Not all graphs are clusterable. Not all graphs have a clustered structure and can be meaningfully summarized through vertex clustering. Clusterable graphs are characterized by pockets of densely connected vertices that are only sparsely connected to the remaining graph. In this article, we re-introduce a very simple and intuitive, yet highly informative, statistical hypothesis test for graph clusterability that is based on vertex and neighborhood samples [24]. The goal of this test is to determine if a graph meets the pre-requisite (necessary) structural conditions to be summarized meaningfully through vertex clusters. 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. The test is also applied to graph comparisons, to test whether one graph has a more pronounced clustered structure than another. Significance is assessed using the t-statistic. Since it is based on sampling, we provide a focused examination of our test's sensitivity to sample size. The main contribution of this article is a detailed examination of our test's accuracy, sensitivity to sample size, conclusion reproducibility and robustness. Our empirical results remain consistent with our earlier conclusions [24] and demonstrate the almost perfect accuracy of our test, even with very small samples of the graph.

To determine if a graph meets the necessary structural conditions to be summarized meaningfully through clusters, we seek to answer a question posed in the literature by Chiplunkar et al. [6], *"(...) 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*

✉ Alexander Y. Shestopaloff
a.shestopaloff@qmul.ac.uk

Pierre Miasnikof
p.miasnikof@mail.utoronto.ca

1 University of Toronto, Toronto, ON, Canada

2 Queen Mary University of London, London, United Kingdom

3 Memorial University of Newfoundland, St. John's, NL, Canada

4 Moscow Institute of Physics and Technology, Dolgoprudny, Russia

5 Yandex, Moscow, Russia

(...)?", through sampling and statistical testing. Here, it is important to specify that while graph "partitions" often designate fixed-sized subsets of vertices, our work focuses on more general "clusters" which can have varying sizes.

Ensuring vertex clusters can provide a meaningful summary of the graph is the first step in any clustering exercise. It is very important to ensure a graph meets the pre-requisite (necessary) conditions for having a clustered structure, for being meaningfully summarizable by vertex clusters, before undertaking any vertex grouping effort. Clustering algorithms will always group vertices, even when they arguably do not form meaningful clusters. In such cases, the clustering process is not only a waste of time, it inevitably leads to misleading conclusions.

In this article, we revisit our previous clusterability testing procedure [24]. While our test was shown to be very accurate, our previous work did not examine sample size sensitivity. Here, we complete our previous work, by assessing its sensitivity to sample size, through an examination of statistical power.

Our test relies on the fact that clusterable graphs are composed of pockets of densely interconnected vertices with sparse connections to the remaining vertices. By definition, clusterable graphs will display mean neighborhood induced subgraph densities that are higher than the graph's overall density. We sample a small subset of the graph's vertices, compute the density of the induced subgraphs formed by their neighborhoods and compare the mean of these densities to the graph's global density. Our numerical experiments reveal that our test is accurate even with very small vertex samples and that it remains robust even under severe departures from the null hypothesis.

The remainder of this article is organized as follows. After a quick overview of the literature published since our previous work was completed, we present our test statistic for single graphs and graph pairs. We then explore its distribution and show test results obtained with varying sample sizes.

2 Previous work

In light of the fact this article constitutes an update on the topic of graph clusterability testing, we only consider developments since the submission of our earlier article [24]. However, just as in our past work, this follow-up remains motivated and inspired by the statistical tests of Gao and Lafferty [12,13]. In our earlier work, we demonstrated these tests' unresponsiveness to graph structure and described their unsuitability to comparisons between graphs. Our much simpler test was demonstrated to be far superior, in all comparisons. It was also shown to be more transparent and not reliant on restrictive underlying assumptions [24].

We are also motivated by the more recent work of Gao and Ma [14]. In fact, these authors build upon the earlier work of Gao and Lafferty [12,13]. This work also presents a statistical test for graph structure. However, the suggested test relies on more restrictive assumptions than ours. The test is also less scalable than ours, as it relies on the graph in its entirety. Ours only uses neighborhood samples.

While not focused on the specific topic of graph vertex clustering but rather general (Euclidian space) clustering, the work of Adolfsson et al. [2], is also a benchmark for us. These authors establish the need for tests of clusterability that are independent of clustering techniques. They begin by asserting that *"(...) an even more fundamental issue than algorithm selection is when clustering should, or should not, be applied."* They then go on to state that *"Clustering with realistic aims, which is our focus here, is only appropriate when cluster structure is present in the data. Otherwise, the results of any clustering technique become necessarily arbitrary and consequently potentially misleading."*

In spite of these very forceful statements, some authors still go through the effort of clustering through trial and error. For example, Filan et al. [9] cluster neural networks using spectral clustering and then assess the quality of their resulting clusters ex-post, using what they call "absolute clusterability" and "relative clusterability", measures that are based on the normalized-cut (n-cut). This state of affairs in clustering only illustrates the need to broadcast the importance of clusterability tests more widely.

We also note that graph testing and sampling from graphs to infer various structural properties remain current topics in the literature. For example, Bogerd et al. [5] test for the presence of a planted community within a graph. Antunes et al. [4] use sampling to estimate triangle distributions.

Another category of tests which continues to be studied in the literature is the $\kappa - \phi$ family of tests, which was first introduced by Czumaj et al. [7] and which was recently tailored to signed graphs by Adriaens and Apers [3]. Tests in this family are more restrictive than the test described in this article. They seek to determine if a graph can be partitioned into, at most, k sets with a conductance of at least ϕ.

In closing this short literature review, it must be mentioned that graph testing was initially introduced by Goldreich et al. [16] in the late 1990 s. These authors' seminal work introduced the practice of sampling vertices and testing for specific properties.

3 Statistical hypothesis test

While there is no formal definition of a cluster of vertices in the literature, there is a clear agreement on its key characteristics. Most authors describe a cluster (or community) as a subset of vertices that exhibit a high-level of interconnec-

tion between themselves and a low-level of connection to the remaining vertices [10,11,25,26,28–30,33] (we quote these authors, but their description is virtually universal across the literature). Consequently, clusterable graphs are composed of a non-trivial number of strongly inter-connected sets of vertices which form dense induced subgraphs with sparse connections to the remaining graph. On the contrary, unclusterable graphs, graphs without clusters, display a constant connectivity pattern. For example, this consistency is very obvious in the case of Erdős-Rény-Gilbert (ERG) graph [8,15]. It is also obvious in complete graphs. These graphs, whose edge probability is constant for all vertex pairs are arguably canonical cases of unclusterable graphs.

In accordance with this quasi-universal agreement on cluster characteristics, we devise a statistical test to detect the presence of a non-trivial amount of dense induced subgraphs with sparse connections to the remaining vertices. Naturally, this heterogeneity (or lack of) in connectivity is reflected in local (neighborhood) densities that are (are not) significantly greater than the graph's overall density, on average. Therefore, we posit that clusterable graphs, graphs composed of clusters, will contain a non-trivial amount of locally dense induced subgraphs of non-trivial size. Our test rests on the hypothesis that, on average, a clusterable graph will have neighborhoods that form induced subgraphs with a density greater than the graph's overall density figure. Conversely, unclusterable graphs are expected to have a mean local density that is indistinguishable from the graph's global density. This fully transparent and intuitive hypothesis is the only underlying assumption of our test. Unlike other tests in the literature, we do not impose restrictive assumptions on the tested graph's generative model, on the null distribution or on the number of clusters.

We also extend our test to graph pairs. In this case, the question we attempt to answer is whether a graph is more significantly clustered than another. We attempt to answer the question "*Are the nodes of graph G more strongly clustered than those of graph H*?".

3.1 Sampling and sampling statistics

To conduct our test, we only sample a small portion of the graph. We sample a portion $s \in (0, 1)$ (ideally with $s \ll 1$) of all nodes, for a total of $L = \lfloor s \times |V| \rfloor$ node samples. We then focus on the induced subgraphs formed by each sampled node's neighborhood and compute the density of each of the L induced subgraphs. For each of the L subgraphs, we denote the local density as κ_i, where $i \in \{1, 2, \ldots, L\}$. To gain a graph level view, we examine the mean neighborhood density ($\bar{\kappa}$). By convention, we set the neighborhood induced subgraph density to zero, for sampled vertices with less than two neighbors.

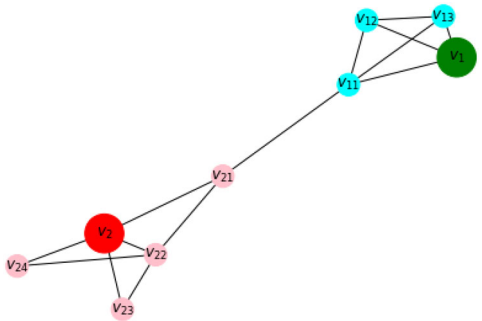


Fig. 1 Sampling example

Figure 1 provides an illustration of our sampling procedure and summary statistic. Assuming we sample verices v_1 (green) and v_2 (red), we take the induced subgraphs formed by each sampled node's neighborhood. In this specific case, we take the graph formed by vertices v_{11}, v_{12}, v_{13} (cyan) and all edges connecting them. For convenience, we label this graph $g_1 = (v_{g_1}, e_{g_1})$. We also obtain the graph formed by vertices $v_{21}, v_{22}, v_{23}, v_{24}$ (pink) and all edges connecting them. We label this graph $g_2 = (v_{g_2}, e_{g_2})$. We then compute the induced subgraph densities for g_1 and g_2, which are denoted by κ_1 and κ_2, respectively. In the equations below, the set of nodes in g_1 (g_2) is denoted as v_{g_1} (v_{g_2}). The set of edges connecting nodes in g_1 (g_2) is denoted as e_{g_1} (e_{g_2}).

$$\kappa_1 = \frac{|e_{g_1}|}{0.5 \times |v_{g_1}| \times (|v_{g_1}| - 1)} = \frac{3}{0.5 \times 3 \times 2} = \frac{3}{3} = 1$$

$$\kappa_2 = \frac{|e_{g_2}|}{0.5 \times |v_{g_2}| \times (|v_{g_2}| - 1)} = \frac{3}{0.5 \times 4 \times 3} = \frac{3}{6} = 0.5$$

With,

$$v_{g_1} = \{v_{11}, v_{12}, v_{13}\} \Rightarrow |v_{g_1}| = 3$$
$$e_{g_1} = \{(v_{11}, v_{12}), (v_{11}, v_{13}), (v_{12}, v_{13})\} \Rightarrow |e_{g_1}| = 3$$
$$v_{g_2} = \{v_{21}, v_{22}, v_{23}, v_{24}\} \Rightarrow |v_{g_2}| = 4$$
$$e_{g_2} = \{(v_{21}, v_{22}), (v_{22}, v_{23}), (v_{22}, v_{24})\} \Rightarrow |e_{g_2}| = 3$$

To obtain a graph-wide picture, we compute the mean neighborhood induced subgraph density, which we denote as $\bar{\kappa}$. In the example in Fig. 1, we have

$$\bar{\kappa} = \frac{1}{2}(\kappa_1 + \kappa_2) = \frac{1}{2}(1 + 0.5) = 0.75.$$

3.2 A probabilistic interpretation of density

Graph and neighborhood induced subgraph densities can be interpreted as the probability that two vertices are connected

by an edge. A graph's global density can be understood as the probability that two arbitrarily selected vertices are connected by an edge. Similarly, at the neighborhood level, local density can also be interpreted as a probability. It can be understood as the conditional probability that two nodes are connected, given they are both in the same induced subgraph formed by their common neighborhood.

The equations below present a mathematical description of this probabilistic interpretation. Equation 1 shows the computation of a graph's global density, which we denote as $\mathcal{K}$. The set of edges is denoted by the usual E and the set of vertices by V.

$$\mathcal{K} = P(e_{ij}) = \frac{|E|}{0.5 \times |V| \times (|V| - 1)} \tag{1}$$

$$\kappa_{\tilde{\nu}} = P(e_{ij} | \nu_i = \nu_j = \tilde{\nu}) = \frac{|e|}{0.5 \times n \times (n - 1)} \tag{2}$$

In Eq. 2, we compute the neighborhood density for an arbitrary neighborhood $\tilde{\nu}$ containing n vertices. The probability two nodes i and j with a neighborhood $\nu_i = \tilde{\nu}$ and $\nu_j = \tilde{\nu}$ respectively are connected in the induced subgraph formed by this neighborhood is given by the ratio of the total number of edges in this subgraph ($|e|$) over the total number of possible connections.

Our statistical test is grounded in this probabilistic interpretation. We posit that a clusterable graph will have a non-trivial amount of densely connected neighborhoods. Under our postulate, vertices of a clusterable graph have a higher probability of sharing an edge with vertices having common neighbors than with those with which they don't. In the average case, we expect that random vertex samples drawn from a given neighborhood will have a greater connection probability than vertices that are arbitrarily sampled across the graph.

3.3 Trivially dense subgraphs and sizes and test limitations

Before proceeding further with our presentation, it is important to note that even an ERG graph, arguably a prototypical unclusterable graph, will often contain a non-trivial number of trivially dense induced subgraphs. For example, the number of triangles (T) in an ERG graph with N vertices and edge probability p can be very large for even moderate sized graphs. Indeed, its expected value, expressed as

$$E(T) = \binom{N}{3} \times p^3,$$

is a non-trivial quantity, in most cases. Similarly, the presence of other dense motifs is also likely.

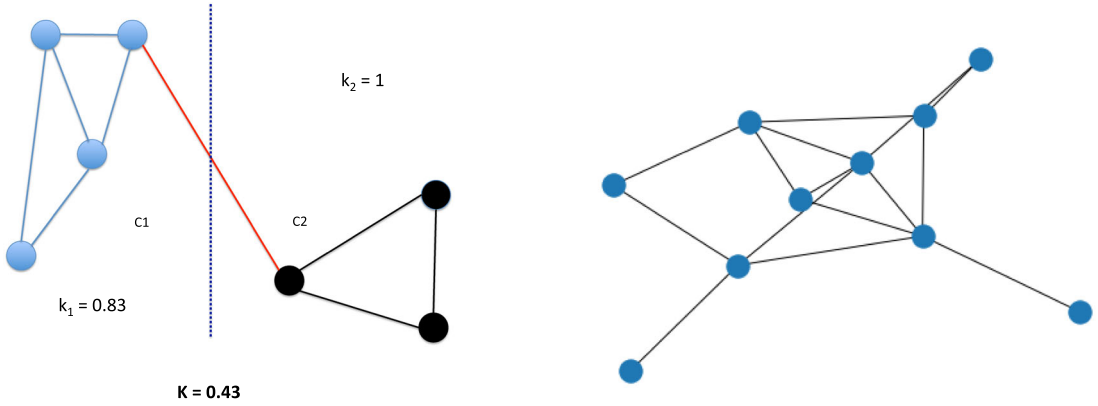
Our test, by design, does not detect such small-scale motifs, regardless of their internal density. It is designed to detect the existence of a statistically significant number of dense subgraphs that are on the scale of neighborhood sizes. It also does not detect statistically insignificant numbers of dense subgraphs. Our test is designed to detect graphs whose structure can be summarized in a meaningful way by dense neighborhoods that are only sparsely connected to the remaining graph. Arguably, a graph containing a small number of dense neighborhoods in an otherwise uniformly connected graph is not meaningfully summarized by clusters. Similarly, a randomly connected graph (e.g., ERG) containing a non-trivial amount of dense motifs (e.g., triangles or other cliques) within broader neighborhoods is also not meaningfully summarized by these features.

As in our previous work, our test rests on the hypothesis that, on average, a clusterable graph will have neighborhoods that form induced subgraphs with a density greater than the graph's global density figure. Figure 2 illustrates this idea. In Fig. 2a, we see a graph composed of two clusters (C_1, C_2). While the graph's global density is $\mathcal{K} = 0.43$, the induced subgraph formed by the nodes in cluster C_1 has a density of $\kappa_1 = 0.83$. The induced subgraph formed by the nodes in cluster C_2 has a density of $\kappa_2 = 1$. In contrast, Fig. 2b displays a graph that is arguably not formed by clusters, in spite of it containing several triangles and other dense motifs. In fact, that graph was generated using the Erdős-Rényi-Gilbert $G(n, p)$ random graph model (ERG) [8,15].

3.4 Single graph test statistic and null hypothesis

Under the null hypothesis, the mean local density ($\bar{\kappa}$) is statistically indistinguishable from the graph's global density ($\mathcal{K}$). This sameness indicates a uniform density structure and the absence of a clustered structure which would be characterized by pockets of strong density. Under this hypothesis, the graph is categorized as *not clusterable*. In the alternative, where $\bar{\kappa}$ is significantly greater than $\mathcal{K}$, we classify the graph as *probably clusterable* because it meets the prerequisite conditions for being clusterable. In such cases, the graph displays heterogeneous local densities which are, on average, greater than the graph's overall density.

Formally, we test whether the mean density over the population of local subgraphs is equal to $\mathcal{K}$. For this comparison of the densities, we use the δ statistic, which transforms $\bar{\kappa}$ into a scaled value that follows a distribution centered at zero. For a sample of $L = \lfloor s \times |V| \rfloor$ neighborhoods, where s represents the proportion of nodes sampled, our test statistic δ measures the difference between between the mean density of the L samples of neighborhood induced subgraphs ($\bar{\kappa}$) and the graph's overall density ($\mathcal{K}$), as a proportion of the graph's



(a) Graph with Clustered Structure

(b) Graph without Clustered Structure

Fig. 2 Graphs Displaying Clustered and Unclustered Structure

overall density.

$$\delta = \frac{\bar{\kappa}}{\mathcal{K}} - 1 \tag{3}$$

Where,

$$\bar{\kappa} = \frac{1}{L}\sum_{i=1}^{L} \kappa_i \tag{4}$$

It is expressed as a centered ratio of mean local densities over global density, in order to make comparisons across different graphs possible. Because it is a scaled and centered mean, we assume δ approximately follows a Gaussian distribution centered at 0, under the null hypothesis. For samples of $L > 30$ neighborhoods, this assumption is justified by the Central Limit Theorem (CLT) [32]. Under the alternative hypothesis, we expect the δ statistics to still follow a Gaussian distribution, but one that is not centered about 0.

Because the true variance of δ is unknown and estimated empirically from the data, we approximate the Gaussian distribution of the δ statistic using a t distribution (centered at 0). Therefore, our significance test uses a one-tailed t-test. The t-statistic is computed as follows:

$$t = \frac{\delta}{s.e(\delta)},$$
$$s.e(\delta) = \sqrt{\frac{Var(\kappa)/(\mathcal{K}^2)}{L}}.$$

3.5 Two graph test statistic and null hypothesis

In the two graph case, we are given two graphs G and H and we want to determine if one displays a more clusterable structure than the other. As in the single-graph case, our test begins with the scaled and centered mean local density, δ. We compute this statistic for each graph. We denote these summary statistics δ_G and δ_H. They are computed exactly as in the single graph case. Here again, they are the scaled and centered local mean densities for graphs G and H, respectively.

Under the null hypothesis, both graphs have the same structure, one is not significantly more clustered than the other. Therefore, the difference $d_{GH} = \delta_G - \delta_H$ is statistically indistinguishable from zero. Under the alternative hypothesis, the difference d_{GH} is statistically significant. Once again, we assume that, under the null hypothesis, the differences d_{GH} follow a Gaussian distribution centered at zero. Correspondingly, we also assume that, under the alternative, the differences d_{GH} also follow a Gaussian distribution, but one that is not centered at zero. Here too, we approximate these Gaussian distributions through a Student's t distribution and assess the statistical significance of the differences using a t-test. In this case, however, we use a two-sample test (unpaired with unequal variance), which makes its computation a bit more convoluted.

In this two-sample case, with neighborhood sample of sizes L_G and L_H, the t-statistic for the difference d_{GH} is computed as follows:

$$t = \frac{d_{GH}}{s.e(d_{GH})},$$
$$s.e(d_{GH}) = \sqrt{s_p^2\left(\frac{1}{L_G} + \frac{1}{L_H}\right)}.$$

Here, s_p^2 is the pooled sample variance of the neighborhood densities κ of each graph scaled by each graph's squared density. For each graph, these sample variances are computed just as in the one-sample case (i.e., $s^2 = Var(\kappa)/\mathcal{K}^2$).

3.6 Test algorithm

As mentioned in the previous section, our δ test statistic can be used to answer two related but distinct questions:

1. Does a graph display heterogeneity in its density?
2. Does a graph G have a more heterogeneous density than a graph H?

In the first question, we ask whether a graph meets the necessary condition to have a clustered structure. In the second, we ask if one graph meets this condition more strongly than another. These questions can be answered by following these steps:

- Sample $L = \lfloor s \times |V| \rfloor$ vertices and extract the L induced subgraphs formed by the neighborhood of each sampled node; (s is the percentage of nodes sampled whose neighborhoods' densities are computed in the next step)
- Compute the local densities $\kappa_i = \frac{|E_i|}{0.5 \times n_i(n_i-1)}$ for each of the L subgraphs (n_i is the number of nodes in the i-th subgraph);
- Compute graph density $\mathcal{K} = \frac{|E|}{0.5 \times |V|(|V|-1)}$, for graph $G = (V, E)$;
- Compute the mean of the local densities: $\bar{\kappa} = \frac{1}{L}\sum_{i=1}^{L} \kappa_i$;
- Normalize the mean and obtain the test statistic: $\delta = \frac{\bar{\kappa}}{\mathcal{K}} - 1$;
- Under the null, the test-statistic δ follows a Gaussian distribution centered at 0, which is approximated by a Student's t distribution;
- Under the alternative hypothesis, we expect the δ statistics to still follow a Gaussian distribution, but one that is not centered about zero;
- In the case of a single-graph test, perform a one-tailed t-test; the null hypothesis is $\mathrm{E}(\delta) = 0$, the alternative is $\mathrm{E}(\delta) > 0$;
- In the case of a two-graph test, perform a two-sample (unpaired with unequal variance) one-tailed t-test on the difference $d_{GH} = \delta_G - \delta_H$. In this case, the null hypothesis is $\mathrm{E}(\delta_G) = \mathrm{E}(\delta_H) \Leftrightarrow E(d_{GH}) = 0$, the alternative is $\mathrm{E}(\delta_G) > \mathrm{E}(\delta_H) \Leftrightarrow E(d_{GH}) > 0$

4 Empirical results

We repeatedly apply our test to several synthetic graphs whose clusterability (or lack of) is known a-priori. The goal of these repetitions is to empirically assess the probability of rejecting the null under various scenarios. We apply our test to samples (sampled without replacement) of 0.5, 1 and 10% of nodes of each graph and, again, repeat the process for 500 iterations. We thus obtain empirical estimates of the probability of rejecting the null hypothesis, for each sample size and under various graph structures.

After determining that samples of 0.5% of nodes provide adequate results, we also apply our test to two-graph trials. Here again, we repeat the process for 500 iterations on each graph pair. These trials yield empirical estimates of the probability of rejecting the null hypothesis, under two different scenarios with known clusterability.

On the basis of our results with samples of 0.5% of nodes, we also estimate our null rejection probabilities on five so-called real-world graphs. Graph details and results are reported in Sect. 4.5.

Table 1 Graph details

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

4.1 Synthetic graphs

To assess the sensitivity of our test to various graph structures and sample sizes, we begin with four synthetic graphs. These graphs, whose clusterability (or lack of) is known a-priori, were simulated using the Python NetworkX library [17]. We generate two clusterable graphs. One is a connected cave man graph (CC) [31], the other has a stochastic block model structure (SBM) [18]. We also generate two unclusterable graphs. The first is a $G(n, p)$ Erdős-Rényi-Gilbert graph (ERG) [8,15] and the second a configuration model graph (CM) [27]. For clarity, generative model details are listed below.

- CC: $|V| = 10,000$ nodes divided in 200 cliques with 50 vertices per clique (one randomly selected edge is reassigned to connect to another clique)
- SBM:
  - 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])
- ERG: Edge probability $p = 0.333$, n ($|V|$) $= 10,000$
- CM: $|V| = 10,000$, exponent $= 3$

Resulting graph characteristics are listed in the Table 1.

Table 2 Synthetic graph test statistics (δ) distributions, by sample size expressed as pct of total nodes (s)

Graph	Mean	Stdev	Min	Max	Pct sampled (s)
Clusterable					
CC	202.56	0.05	202.38	202.70	10
CC	202.56	0.15	202.00	202.89	1
CC	202.56	0.22	201.91	202.90	0.5
SBM	6.73E−04	2.49E−05	5.85E−04	7.50E−04	10
SBM	6.70E−04	7.33E−05	3.99E−04	8.80E−04	1
SBM	6.71E−04	1.01E−04	3.62E−04	1.00E−03	0.5
Unclusterable					
ERG	−1.02E−05	1.86E−05	−6.49E−05	4.23E−05	10
ERG	−8.21E−06	5.58E−05	−1.87E−04	1.56E−04	1
ERG	−6.66E−06	8.30E−05	−2.32E−04	2.79E−04	0.5
CM	0.73	2.46	−1.00	11.71	10
CM	0.54	7.27	−1.00	46.36	1
CM	0.34	9.59	−1.00	80.19	0.5

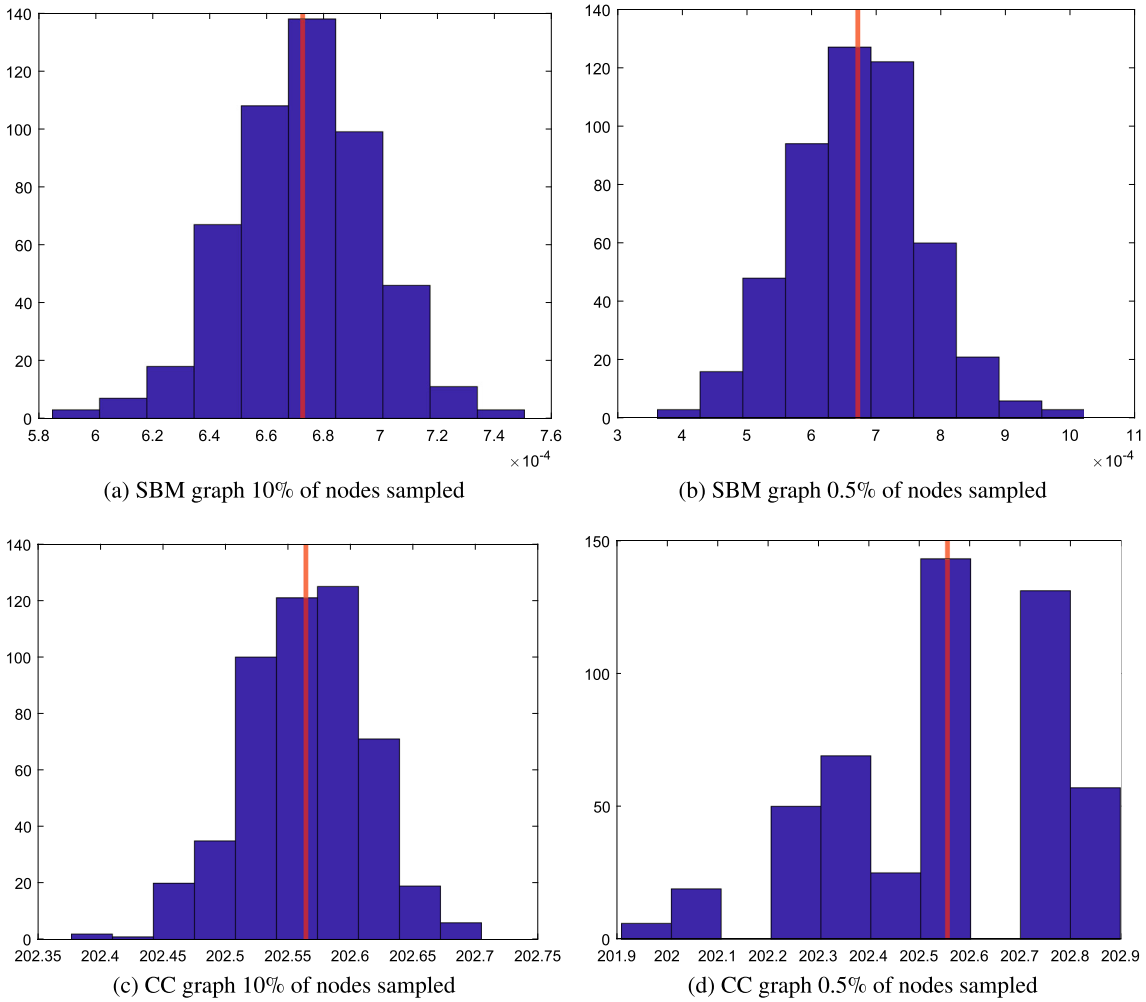
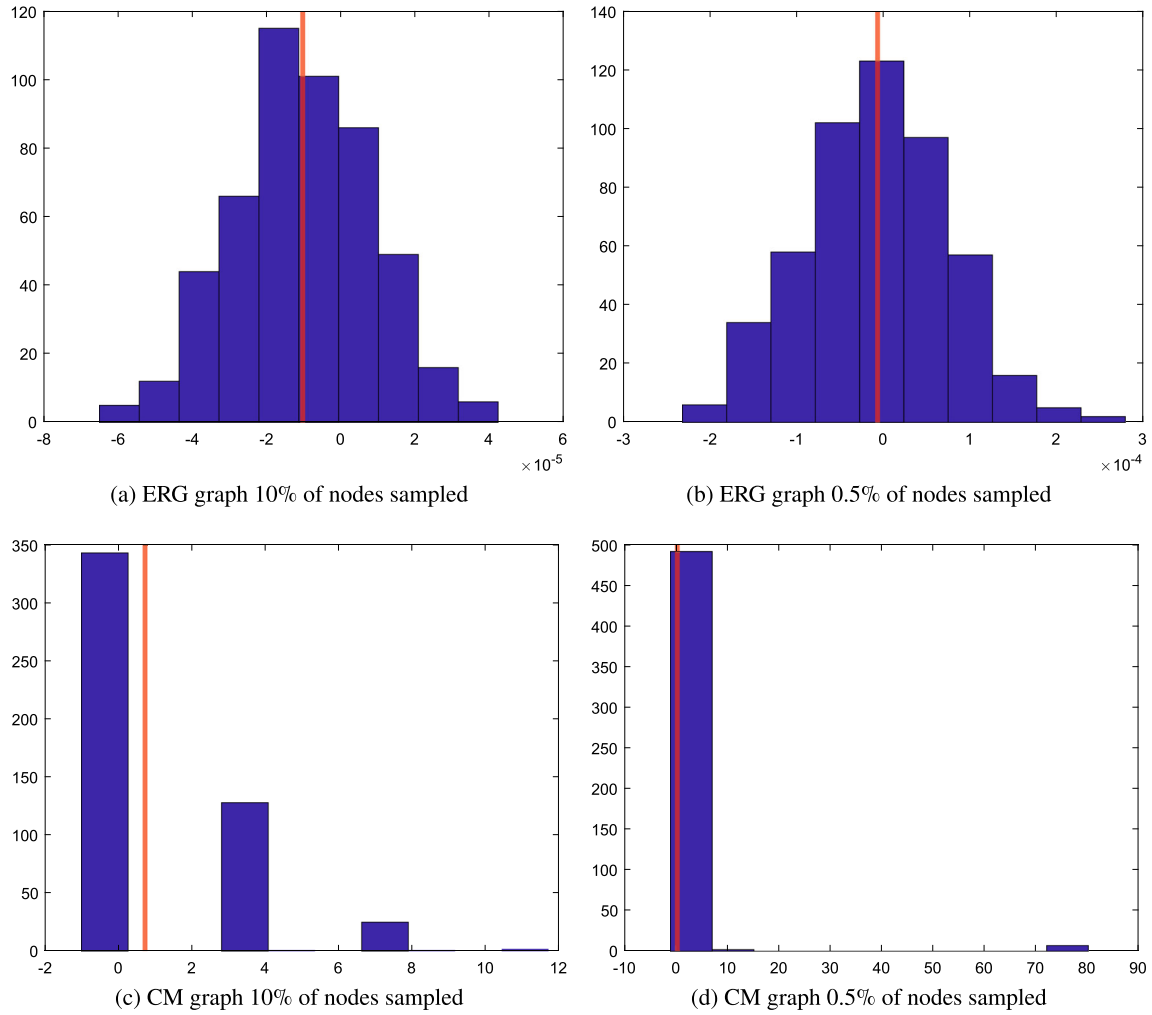


Fig. 3 Synthetic graph test statistic (δ) distributions, clusterable structure



(a) ERG graph 10% of nodes sampled

(b) ERG graph 0.5% of nodes sampled

(c) CM graph 10% of nodes sampled

(d) CM graph 0.5% of nodes sampled

Fig. 4 Synthetic graph test statistic (δ) distributions, clusterable structure

4.2 Test statistic distribution

As mentioned earlier, we empirically examine the effect of sample size on our test statistic δ under the null and alternative hypotheses, using synthetic graphs whose clusterability is known a-priori. We apply our sampling and testing algorithm 500 times (500 iterations), to each graph. At each iteration, we compute the δ statistic and record its value. We then examine the resulting distribution of these 500 trials. The process follows the algorithm described in Sect. 3.6. This process is repeated with samples of 10%, 1% and 0.5% of vertices. Under the null hypothesis, these distributions are expected to be Gaussian and centered at zero. Under the alternative hypothesis, these distributions are also expected to be Gaussian, but centered at a point significantly greater than zero.

Mean, standard deviation, minimum and maximum of the δ statistic are reported for each distribution, in Table 2. Histograms for the sampling distributions of the 0.5% and 10% of nodes experiments are shown in Fig. 4 (known unclusterable graphs) and Fig. 3 (known clusterable graphs). In reviewing these comparisons, it is important to note that these are the distributions of a sample of 500 δ statistics (one per iteration). The histograms, means, standard deviations, minima and maxima are for these 500 δ statistics. They are not summaries of the L ($= \lfloor s \times |V| \rfloor$) randomly selected κ_i data points (local densities) used to compute each of the δ statistics.

In the cases of the SBM (clusterable) and ERG (unclusterable) graphs, we observe a very strong distributional stability of the δ statistics across sampling sizes. Indeed, as observed in Table 2, Figs. 3 and Fig. 4, the distributions of the δ statistic do not appear to be sensitive to sampling sizes. In contrast, in the cases of the CC (clusterable) graph and especially the CM (unclusterable) graph, we note a greater variation in the distributions. However, while the symmetry of the CC graph's distribution diminishes with sampling size, the mean and min-max range remain unaffected. Meanwhile, the case

Table 3 Empirical rejection probabilities based on 500 trials, by sample size expressed as pct of total nodes (s)

Graph	Num rejects	Pct sampled (s)	P(reject)
Clusterable			
CC	500	10%	1
CC	500	1%	1
CC	500	0.5%	1
SBM	500	10%	1
SBM	500	1%	1
SBM	500	0.5%	1
Unclusterable			
ERG	7	10%	0.01
ERG	9	1%	0.02
ERG	20	0.5%	0.04
CM	1	10%	0.00
CM	0	1%	0
CM	0	0.5%	0

Table 4 Empirical rejection probabilities based on 500 trials, 2 graph test

G	H	Ideal	Num rejects	P(reject)
CC	ERG	Reject	500	1
SBM	ERG	Reject	500	1
CM	Cave	No reject	0	0
CM	SBM	No reject	0	0

Table 5 Real-world graph characteristics

	$\mathcal{K}$ (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

Table 6 Real-world graph test-statistic (δ) distribution for 500 trials with 0.5% of nodes sampled, summary

Graph	Mean	Stdev	Min	Max
DIMACS10	1269.30	203.21	722.51	1894.40
LFR	304.88	12.40	266.24	339.84
Astro	564.62	33.22	458.46	656.13
Enron	1824.80	114.48	1479.20	2204.70
DBLP	30,302.00	499.13	28,647.00	31,534.00

of the CM graph is notable. The distribution of its δ statistic varies quite sharply across sampling size experiments.

4.3 Null rejection probability (single graph)

To obtain empirical estimates of rejection probabilities under various scenarios, we repeat our test for 500 iterations. At each iteration, we record the acceptance or rejection conclusion of our test. We also examine the sensitivity of these rejection probabilities to sample size. Results are reported in Table 3.

Results in Table 3 demonstrate that our test's conclusions are unaffected by sampling size. We also note that our null hypothesis rejection probabilities for the case of unclusterable graphs are slightly less than the expected 0.05. Indeed, given that under the null hypothesis the δ should follow a Gaussian distribution centered at zero, we expect erroneous rejections of the null (type I errors) in approximately 5% of cases.

This unexpectedly low type I error rate is especially marked for the CM graph. It is likely attributable to the very wide deviation from the Gaussian null hypothesis of its δ statistic's distribution. In fact, this strong departure from the Gaussian null hypothesis is clearly observable in Fig. 4 (subfigures c and d). In contrast, the null distribution has a Gaussian-like distribution, in the case of the ERG graph. In this latter case, we attribute the slightly lower than expected type I error rate to random noise and sampling error.

Meanwhile, we note a consistently high power, for all sample sizes. At all sampling levels, type II error remains non existent. Indeed, the null hypothesis is always correctly rejected in all clusterable graph experiments.

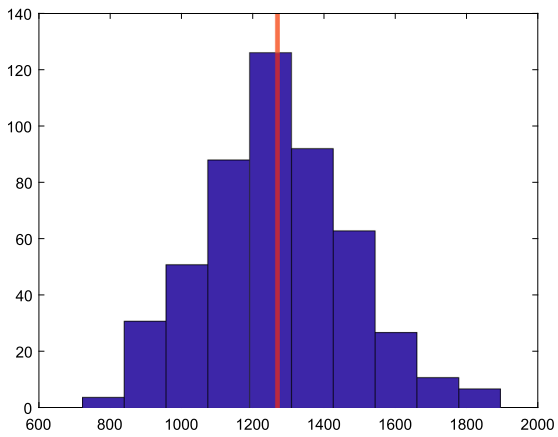
4.4 Null rejection probability (graph pairs)

We also perform the same probability estimation exercise with graph pairs. Here, the goal is to determine the probability that the null hypothesis that graphs G and H are equally clusterable. Given the strong results obtained with a sample of 0.5% of nodes in the single-graph case and in the interest of brevity, we only report results with samples of 0.5% of nodes of each graph. Results are reported in Table 4.

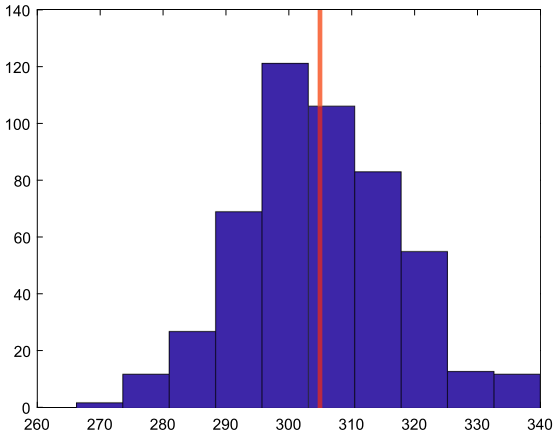
Our test has perfect accuracy. Here again, it does however have a lower than expected (under the Gaussian null hypothesis) type I error. Here again, we attribute this unexpected power to a severe departure from the Gaussian null-hypothesis. As highlighted earlier, the distribution of the CM graph's δ are highly non-Gaussian.

4.5 Illustrative examples using real-world graphs

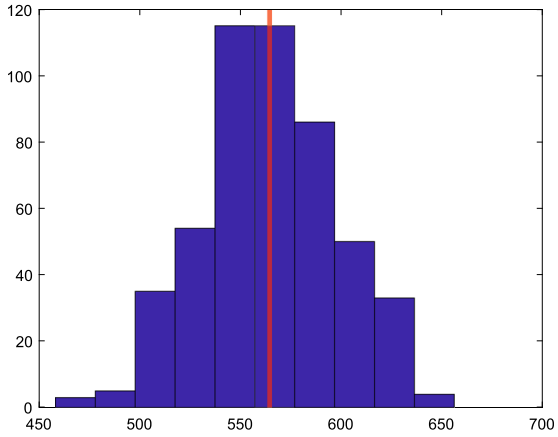
As mentioned earlier, we also estimate our null rejection probabilities on five different so-called real-world graphs,



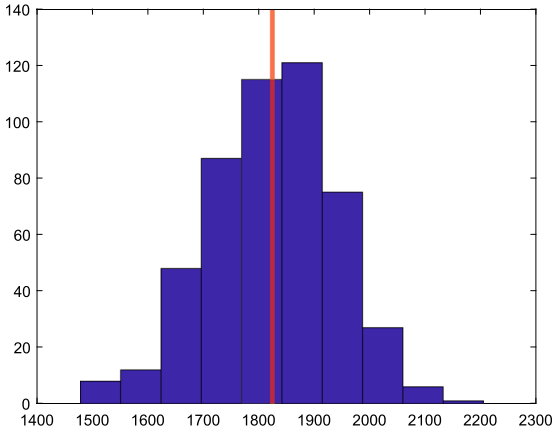
(a) DIMACS10 graph 0.5% of nodes sampled



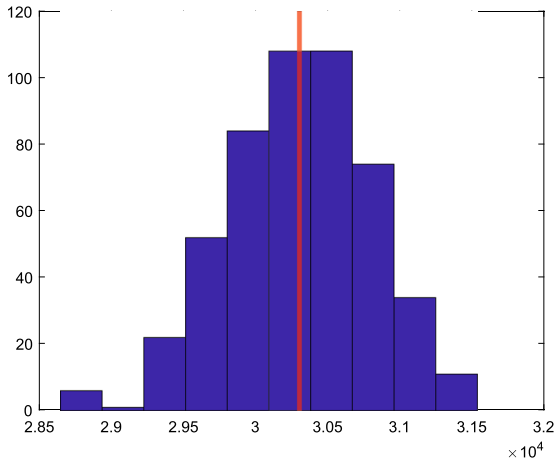
(b) LFR graph 0.5% of nodes sampled



(c) Astro graph 0.5% of nodes sampled



(d) Enron graph 0.5% of nodes sampled



(e) DBLP graph 0.5% of nodes sampled

Fig. 5 Real-world graph test-statistic (δ) distribution for 500 trials with 0.5% of nodes sampled

Table 7 Null rejection probabilities based on 500 trials, single real-world graphs

Graph	Num reject	nodes sampled	P(reject)
DIMACS10	500	0.5%	1
LFR	500	0.5%	1
Astro	500	0.5%	1
Enron	500	0.5%	1
DBLP	500	0.5%	1

using distributions of samples of 0.5% of nodes. Graph details are listed below and also summarized in Table 5:

- dimacs10-as-22july06 network (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") [1,21]
- Lancichinetti, Fortunato, Radicchi (LFR) graph [22]
- Astro Physics collaboration network (Astro, "Arxiv ASTRO-PH (Astro Physics) collaboration network (...)") [19]
- Enron email network (Enron, "Enron email communication network") [20,23]
- DBLP collaboration network (DBLP, "(...) a co-authorship network where two authors are connected if they publish at least one paper together (...)") [33]

Here again, we repeated our sample and test algorithm for 500 iterations. Table 6 and Fig. 5 show the distribution of the test statistic δ. In all five cases, we observe Gaussian-like distributions of δ that are centered about means that are significantly greater than zero, indicating strong support for the rejection of the null hypothesis of these graphs being *not clusterable*. Indeed, our repeated significance test results shown in Table 7 confirm these observations.

Finally, we also conduct two-graph tests for 500 repeated trials. We use the known unclusterable ER and CM graphs as comparison benchmarks. To avoid redundancy, we restrict the pairwise (graph pairs) comparisons to the "dimacs10-as-22july06" (DIMACS10) [1,21] and LFR [22] graphs. In the first set of experiments, we test if the ER and DIMACS10 graphs are equally clusterable. In these experiments, the alternative hypothesis, which we know to be false, is that the ERG graph is more clusterable than the DIMACS10 graph (ER isn't clusterable, DIMACS10 is). Therefore, we realistically expect the null not to be rejected. In the second set of experiments, the null hypothesis, which we know to be to be false and expect to be rejected, is that the LFR and CM graphs are equally clusterable (in fact, LFR is clusterable, CM is not).

Once again, these experiments show that our test has perfect accuracy, as reported in Table 8. Given the very sizable

Table 8 Null rejection probabilities based on 500 trials, two-graph test

G	H	Ideal	Num rejects	P(reject)
ERG	DIMACS10	No reject	0	0
LFR	CM	Reject	500	1

difference between the means and moderately sized standard deviations of the δ test statistics of each graph experiments, this accuracy is completely expected. These summary statistics are shown in Table 2 (synthetic graph experiments) and Table 6 (real-world graph experiments). (Once again, we wish to highlight that these are the means and standard deviations of 500 δ statistics, one per iteration. They are not the sample statistics used to compute each of the the δ statistics.)

5 Conclusion

We have subjected our δ test statistic to several experiments, in order to assess its accuracy under various scenarios. Our experiments also aimed at determining the test statistic's sensitivity to sampling size.

Our results reveal that our test offers valid conclusions, even with very small samplings of the graph's vertices. We do note, however, that under certain scenarios our Gaussian null hypothesis is not accurate. Nevertheless, our experiments demonstrate that our test's conclusions remain valid, even under severe departures from this null hypothesis. In fact, under such departures, our test has been shown to be more accurate than expected.

Acknowledgements PM thanks Prof. Ali Sheikholeslami of the University of Toronto, for his support and guidance during the production of this work.

Author Contributions All authors contributed to this work. They all reviewed its contents and stand by them. PM initiated the study. PM and AS devised the experiments, implemented the computations, collected results and drafted the article. AR provided supervision and guidance.

Funding None.

Declarations

Conflict of interest The authors state they have no competing interests.

References

1. Autonomous systems (DIMACS10) network dataset—KONECT. Available at http://konect.cc/networks/dimacs10-as-22july06 (2018)
2. Adolfsson, A., Ackerman, M., Brownstein, N.C.: To cluster, or not to cluster: an analysis of clusterability methods. Pattern Recogn. **88**, 13–26 (2019)
3. Adriaens, F., Apers, S.: Testing properties of signed graphs. arXiv:2102.07587 (2021)
4. Antunes, N., Guo, T., Pipiras, V.: Sampling methods and estimation of triangle count distributions in large networks. Netw. Sci. **9**(S1), S134–S156 (2021)
5. Bogerd, K., Castro, R.M., van der Hofstad, R., Verzelen, N.: Detecting a planted community in an inhomogeneous random graph. Bernoulli **27**(2), 1159–1188 (2021)
6. Chiplunkar, A., Kapralov, M., Khanna, S., Mousavifar, A., Peres, Y.: Testing graph clusterability: algorithms and lower bounds. In: 2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS), pp. 497–508 (2018)
7. Czumaj, A., Peng, P., Sohler, C.: Testing Cluster Structure of Graphs. In: Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing (2015)
8. Erdös, P., Rényi, A.: On random graphs I. Publ. Math. Debrecen **6**, 290–297 (1959)
9. Filan, D., Casper, S., Hod, S., Wild, C., Critch, A., Russell, S.: Clusterability in neural networks. arXiv:2103.03386, archivePrefix=arXiv, primaryClass=cs.NE, (2021)
10. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)
11. Fortunato, S., Hric, D.: Community detection in networks: a user guide. Phys. Rep. **659**, 1–44 (2016)
12. Gao, C., Lafferty, J.: Testing for global network structure using small subgraph statistics. arXiv, arXiv:1710.00862 (2017)
13. Gao, C., Lafferty, J.: Testing network structure using relations between small subgraph probabilities. arXiv:1704.06742 (2017)
14. Gao, C., Ma, Z.: Minimax rates in network analysis: graphon estimation, community detection and hypothesis testing. page arXiv:1811.06055 (November 2018)
15. Gilbert, E.N.: Random graphs. Ann. Math. Stat. **30**(4), 1141–1144 (1959)
16. Goldreich, O., Goldwasser, S., Ron, D.: Property testing and its connection to learning and approximation. J. ACM **45**(4), 653–750 (1998)
17. Hagberg, A.A., Schult, D.A., Swart, P.J.: Exploring network structure, dynamics, and function using network. In: Varoquaux, G., Vaught, T., Millman, J. (eds) Proceedings of the 7th Python in Science Conference, pp. 11–15, Pasadena, CA USA (2008)
18. Holland, P.W., Laskey, K.B., Leinhardt, S.: Stochastic blockmodels: first steps. Soc. Netw. **5**(2), 109–137 (1983)
19. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution: densification and shrinking diameters. ACM Trans. Knowl. Discov. Data **1**(1), 2-es (2007)
20. Klimt, B., Yang, Y.: Introducing the enron corpus. In: CEAS (2004)
21. Kunegis, J.: KONECT—The Koblenz network collection. In: Proceedings of the 22nd international conference on world wide web, pp. 1343–1350 (2013)
22. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E **78**(4), 046110 (2008)
23. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. arXiv e-prints, page arXiv:0810.1355 (October 2008)
24. Miasnikof, P., Prokhorenkova, L., Shestopaloff, A.Y., Raigorodskii, A.: A statistical test of heterogeneous subgraph densities to assess clusterability. In: Matsatsinis, Nikolaos F., Marinakis, Yannis, Pardalos, Panos (eds.) Learning and Intelligent Optimization, pp. 17–29. Springer International Publishing, Cham (2020)
25. Miasnikof, P., Shestopaloff, A.Y., Bonner, A.J., Lawryshyn, Y.: A Statistical Performance Analysis of Graph Clustering Algorithms, chapter 11. Lecture Notes in Computer Science, vol. 6. Springer Nature, Berlin (2018)
26. Miasnikof, P., Shestopaloff, A.Y., Bonner, A.J., Lawryshyn, Y., Pardalos, P.M.: A density-based statistical analysis of graph clustering algorithm performance. J. Complex Networks **8**(3), 08 (2020)
27. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**(2), 167–256 (2003)
28. Ostroumova Prokhorenkova, L., Prałat, P., Raigorodskii, A.: Modularity of complex networks models. In: Bonato, A., Graham, F.C., Prałat, P. (eds.) Algorithms and Models for the Web Graph, pp. 115–126. Springer International Publishing, Cham (2016)
29. Ostroumova Prokhorenkova, L., Prałat, P., Raigorodskii, A.: Modularity in several random graph models. Electron. Notes Discrete Math., **61**, 947–953 (2017). The European Conference on Combinatorics, Graph Theory and Applications (EUROCOMB'17)
30. Schaeffer, S.E.: Survey: graph clustering. Comput. Sci. Rev. **1**(1), 27–64 (2007)
31. Weisstein, E.W.: Caveman graph. https://mathworld.wolfram.com/CavemanGraph.html
32. Weisstein, E.W.: Central limit theorem. https://mathworld.wolfram.com/CentralLimitTheorem.html
33. Yang, J., Leskovec, J.: Defining and Evaluating Network Communities based on Ground-truth. CoRR, arXiv:1205.6233 (2012)