Percolation-Based Detection of Anomalous Subgraphs in Complex Networks

Corentin Larroche[1,2(✉)], Johan Mazel[1], and Stephan Clémençon[2]

[1] French National Cybersecurity Agency (ANSSI), Paris, France
{corentin.larroche,johan.mazel}@ssi.gouv.fr
[2] LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France
{corentin.larroche,stephan.clemencon}@telecom-paris.fr

Abstract. The ability to detect an unusual concentration of extreme observations in a connected region of a graph is fundamental in a number of use cases, ranging from traffic accident detection in road networks to intrusion detection in computer networks. This task is usually performed using scan statistics-based methods, which require explicitly finding the most anomalous subgraph and thus are computationally intensive.

We propose a more scalable method in the case where the observations are assigned to the edges of a large-scale network. The rationale behind our work is that if an anomalous cluster exists in the graph, then the subgraph induced by the most individually anomalous edges should contain an unexpectedly large connected component. We therefore reformulate our problem as the detection of anomalous sample paths of a percolation process on the graph, and our contribution can be seen as a generalization of previous work on percolation-based cluster detection. We evaluate our method through extensive simulations.

1 Introduction

Detection of a significant connected subgraph in a larger background network is a ubiquitous task: such significant regions can be indicative of fraudulent behavior in social networks [15] or of the propagation of an intruder in a computer network [22], for instance. Therefore, being able to discern them from ambient noise has valuable applications in a number of settings. This anomaly detection problem is, however, remarkably challenging: the large size and complex structure of real-world graphs make the characterization of normal behavior difficult and the search for non-trivial substructures computationally expensive.

The aim of this paper is to propose a scalable method for anomalous connected subgraph detection in a graph with observations attached to its edges. The null distribution of the observations, or an approximation thereof, is assumed to be known. Building upon this knowledge, the degree of abnormality of each individual edge with respect to the model can be measured, and our goal is to detect a significant concentration of anomalous edges in a connected region of

M. R. Berthold et al. (Eds.): IDA 2020, LNCS 12080, pp. 287–299, 2020.
https://doi.org/10.1007/978-3-030-44584-3_23

the graph. Usual methods for this task are built around scan statistics [14]. Such methods boil down to maximizing a scoring function over the set of connected regions of the graph, then rejecting the null hypothesis (*i.e.* absence of anomalous subgraph) if the maximum exceeds a certain threshold. This implies solving a combinatorial optimization problem over the class of all connected subgraphs, which is expensive due to the exponentially growing size of the latter.

In contrast, our approach does not require explicitly searching for the best candidate subgraph. Instead, we build on the following idea: under the null hypothesis, the most individually anomalous edges are randomly spread out over the graph. Therefore, removing all but the k most anomalous edges from the graph is equivalent to drawing k edges uniformly at random and extracting the subgraph induced by these edges. In other words, this procedure amounts to bond percolation on a graph. On the other hand, when an anomalous subgraph is present, the location of the individual anomalies is no longer random, and thus the largest connected component of the subgraph induced by the k most anomalous edges should contain an unexpectedly large connected component. This link between anomalous subgraph detection and percolation theory has already been introduced in the context of regular lattices [6,19,20], but to the best of our knowledge, it has not yet been studied for arbitrary graphs.

We argue that our method is more scalable than traditional ones while retaining an acceptable detection power, especially when seeking to detect small anomalous regions in large graphs. We assess this detection performance through numerical experiments on several realistic synthetic graphs.

The rest of this paper is structured as follows. In Sect. 2, we introduce the statistical framework for our problem and present some related work. Section 3 describes our detection method, while Sect. 4 is devoted to its empirical evaluation on simulated data. Finally, we discuss our results and some interesting leads for future work in Sect. 5, then briefly conclude in Sect. 6.

2 Problem Formulation and Related Work

We begin with a thorough formulation of our problem as a case of statistical hypothesis testing, then review the main existing approaches to it.

2.1 Problem Formulation – Statistical Hypothesis Testing

Consider an undirected and connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ (resp. $\mathcal{E}$) is the set of vertices (resp. edges) of $\mathcal{G}$. Letting $|\mathcal{A}|$ denote the number of elements of a set $\mathcal{A}$, we write $m = |\mathcal{E}|$, and we use $\mathcal{E}$ and $[m] = \{1, \dots, m\}$ interchangeably to represent the set of edges. We further write $2^{\mathcal{A}}$ for the set of all subsets of $\mathcal{A}$ and $\mathbb{1}\{\cdot\}$ for the indicator function of an event.

Let $\Lambda \subset 2^{\mathcal{E}}$ denote the class of subsets of $\mathcal{E}$ whose induced subgraph in $\mathcal{G}$ is connected. Given a signal $\mathbf{X} = (X_1, \dots, X_m) \in \mathbb{R}^m$ observed on the edges of $\mathcal{G}$ and a known probability distribution F_0, the null hypothesis is defined as

$H_0 : X_i \overset{\text{iid}}{\sim} F_0$. For each $\mathcal{S} \in \Lambda$, we further define the alternative

$$H_{\mathcal{S}} : \begin{cases} \mathbf{X}_{|\mathcal{S}} \sim F_{\mathcal{S}} \\ \forall i \notin \mathcal{S},\ X_i \sim F_0 \end{cases},$$

where $\mathbf{X}_{|\mathcal{S}}$ is the restriction of $\mathbf{X}$ to $\mathcal{S}$ and $F_{\mathcal{S}}$ is a joint probability distribution. $F_{\mathcal{S}}$ is only assumed to be different from $F_0^{\otimes|\mathcal{S}|}$, and it can differ in various ways. In many applications, the observations in $\mathcal{S}$ are simply larger than expected (consider for instance network intrusion detection, where the presence of an intruder results in additional activity in a connected region of the network). The problem considered in this paper can be formulated as

$$H_0 \quad \text{vs.} \quad H_1 = \bigcup_{\mathcal{S} \in \Lambda} H_{\mathcal{S}}\,.$$

That is, we want to know whether there exists a connected subgraph of $\mathcal{G}$ inside of which the observations X_i are drawn from an alternative distribution. Note that we only care about detection, leaving the reconstruction of $\mathcal{S}$ aside.

2.2 Related Work – Scan Statistics and Beyond

A lot of existing work deals with a specific instance of the problem defined above, namely elevated mean detection on a graph. In this setting, the observations are independent standard centered normal random variables under the null, while X_i has mean $\mu_{\mathcal{S}} \mathbb{1}\{i \in \mathcal{S}\}$ under the alternative $H_{\mathcal{S}}$ (for some $\mu_{\mathcal{S}} > 0$). Theoretical conditions for detectability in this case are stated in [1]. A closely related problem arises when the observations are associated with vertices rather than edges, and this setting was studied in [3–5]. However, these papers focus on statistical analysis and do not provide computationally tractable tests.

From a more practical perspective, the most common approach to anomalous subgraph detection is based on scan statistics. Broadly speaking, this method consists in defining a scoring function $f : 2^{\mathcal{E}} \to \mathbb{R}$, computing the test statistic $t = \max_{\mathcal{S} \in \Lambda} f(\mathcal{S})$, then rejecting H_0 if t exceeds a given threshold. This amounts to finding the most anomalous subset $\mathcal{S}^*$ in Λ, and then rejecting the null hypothesis if $\mathcal{S}^*$ is anomalous enough. Defining f requires some hypotheses on the class of alternative distributions $\{F_{\mathcal{S}}\}$. For instance, when $F_{\mathcal{S}}$ has a parametric form, $f(\mathcal{S})$ can be defined as the likelihood ratio between $H_{\mathcal{S}}$ and H_0. In the more general case considered here, however, finding a suitable scoring function is non-trivial. Moreover, computing t implies maximizing f over the combinatorial class Λ, which quickly becomes computationally intensive as the graph grows. Therefore, most related work focuses on making the computation of scan statistics more efficient. Ways to achieve this include the following:

Restriction of the Class Λ. The easiest way to speed up the computation is to simply reduce the size of the search space by considering only a subset of Λ. Such restriction can be based on domain-specific knowledge [17,18,22,25] or more general heuristics [24].

Convex Relaxation. Another classical approach to combinatorial optimization consists in solving a convex relaxation of the problem, and then projecting the solution back onto the original search space. This method was applied to scan statistics [2,26,27], using elements of spectral graph theory [9] to find a relaxed form of the connectivity constraint. Similar ideas were also used in a slightly different context [29–31], where the class Λ consists of subgraphs with low cut size rather than connected ones.

Algorithmic Approaches. Finally, efficient optimization algorithms have been used to find exact or approximate values for the scan statistic, including simulated annealing [11,12], greedy algorithms [28], primal-dual algorithms [28], branch and bound algorithms [32] and dynamic programming algorithms [33].

Despite the popularity of scan statistics, other ideas have also been considered in the literature. We focus on one of these alternative approaches, namely the Largest Open Cluster (LOC) test, which was first studied in the context of object detection in images [19,20]. The idea of this method is to represent an image as a two-dimensional lattice, each node carrying a random variable standing for the value of the associated pixel. Then, after deleting from the lattice every vertex whose pixel value is lower than a suitable threshold, the largest remaining connected component is expected to be small if there is no object in the image. On the other hand, if an object is present, an unexpectedly large connected component should remain in the thresholded lattice. The theory behind the LOC test has since been extended to lattices of arbitrary dimension [6], but to the best of our knowledge, the underlying idea of using percolation theory to detect anomalous connected subgraphs has not yet been applied to complex, arbitrary-shaped networks.

3 Local Anomaly Detection and Percolation Theory

We now describe our method, first introducing some necessary notions of percolation theory, then highlighting their relevance to our anomaly detection problem. Finally, we provide a detailed description of our testing procedure.

3.1 Some Notions of Percolation Theory

An interesting aspect of the LOC test is that the behavior of its test statistic under the null hypothesis can be described using percolation theory. Therefore, we first review some useful results from this field, which motivate our approach. For more details, see for example [10] and references therein. Since our primary interest is in signals associated with edges, we focus on bond percolation, where edges of a connected graph with n vertices are occupied uniformly at random with probability p or unoccupied with probability $1 - p$.

Let $C(p)$ denote the size of the largest connected component of this graph at occupation probability p. The main focus of percolation theory is to find the limit of $C(p)$ as n becomes large. Extremal values of p yield obvious results: for

$p = 0$, $C(p) = 1$ for any n and for $p = 1$, $\lim_{n\to\infty} C(p) = \infty$. For intermediate values of p, however, there are two possible regimes. If p is small enough, only small connected components are present and $C(p)/n$ converges in probability to 0. On the other hand, larger values of p lead to the emergence of a giant connected component, which contains a constant fraction of the vertices. The transition between the two regimes happens for a critical value of p called the percolation threshold p_c. Note that p_c depends on the graph structure and can be vanishingly small. Although this phase transition is only well-defined in the limit of an infinite graph, a somewhat similar behavior can be observed in the finite case [8,16]. In particular, define the percolation process $\{C(p)\}_{0\le p\le 1}$ as follows: assign to each edge e an independent random variable U_e, uniformly distributed on $[0, 1]$. Then, keeping the U_e fixed, let p vary on $[0, 1]$, deleting e from the graph whenever $U_e > p$. A tightly related process is obtained by considering the imbedded Markov chain $\{\mathcal{G}_k\}_{k\ge 0}$, where $\mathcal{G}_k$ is the subgraph induced by the edges associated with the k smallest random variables. Letting C_k denote the size of the largest connected component of $\mathcal{G}_k$, $\{C_k\}_{k\ge 0}$ can be seen as a discretized version of $\{C(p)\}_{0\le p\le 1}$. Even for finite graphs, sample paths of these two processes do not deviate significantly from the mean trajectory, making them suitable candidates for anomaly detection.

3.2 Application to Anomalous Subgraph Detection

We now motivate the idea of mapping a signal $\mathbf{X}$ onto a sample path of the percolation process. For $i \in [m]$, define $P_i = 1 - F_0(X_i)$ as the upper tail p-value associated with X_i. Define also, for $k \in \{0, \dots, m\}$, the subgraph $\mathcal{G}_k$ induced by the edges associated with the k smallest p-values, and let S_k denote the size of its largest connected component. Under the null hypothesis, the random variables $\{P_i\}$ are independent and uniformly distributed on $[0, 1]$. Therefore, S_k has the same distribution as C_k for all $k \in \{0, \dots, m\}$. Under the alternative $H_{\mathcal{S}}$, however, the distribution of the variables $\{P_i\}_{i\in\mathcal{S}}$ is altered, which induces a deviation in the process $\{S_k\}_{0\le k\le m}$ with respect to the normal percolation process. Our test aims to detect this deviation.

Figure 1 illustrates the normal and anomalous behaviors of the percolation process for three graph models: a two-dimensional square lattice, an Erdős-Rényi random graph [13] and a Barabási-Albert preferential attachment graph [7]. For each model, a graph with 1024 vertices and approximately 2000 edges is generated, and the mean and standard deviation of the fraction of vertices in the largest connected component for each value of p is estimated using 10000 Monte Carlo simulations. Then, for each graph, we generate a subtree $\mathcal{S}$ containing a fraction δ of the vertices, assign to each edge e an independent Gaussian random variable $X_e \sim \mathcal{N}(\mu\mathbb{1}\{e \in \mathcal{S}\}, 1)$ and compute the associated sample path of the percolation process. This experiment was repeated 1000 times for each graph, and the mean sample path for different values of δ and μ is displayed. The two regimes of the percolation process can be observed, and the shape and location of the phase transition both clearly depend on the graph model. While the

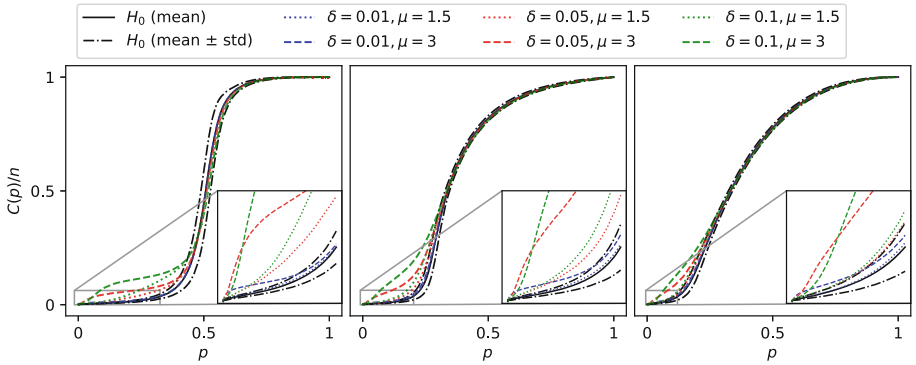

Fig. 1. Evolution of the fraction of vertices in the largest connected component as p varies from 0 to 1, under H_0 and various alternatives, for three kinds of graphs: a two-dimensional square lattice (left), an Erdős-Rényi random graph (center) and a Barabási-Albert preferential attachment graph (right).

separation between the two regimes is quite clear for the lattice and the Erdős-Rényi graph, it is much blurrier for the Barabási-Albert model, which yields more complex structures – most interestingly, heavy-tailed degree distributions. Since such properties are often found in real-world networks, it is important to qualify their impact on the feasibility of percolation-based cluster detection. Figure 1 shows that although the anomalous sample paths become harder to distinguish as the phase transition gets hazier, the normal trajectories are concentrated enough to make even small deviations visible, which motivates our approach.

3.3 Putting It All Together – Description of Our Test

We now proceed with the description of our test. First, define

$$K = \min \left\{ k \leq m,\, \mathbb{E}_0[S_k] \geq \sqrt{|\mathcal{V}|} \right\},$$

where $\mathbb{E}_0$ denotes the expected value under H_0. K can be understood as the index corresponding to the onset of the phase transition. Since we aim to detect the appearance of an unexpectedly large connected component in the early steps of the percolation process, the test statistic we use is

$$\chi = \frac{1}{|\mathcal{V}| \cdot K} \sum_{k=1}^{K} S_k \,.$$

This statistic is equivalent to the area under a piecewise constant interpolation of the sequence of points $\{(k, S_k)\}_{0 \leq k \leq K}$, and is therefore expected to be higher than usual in the presence of an anomalous subgraph.

Estimation of K and calibration of the test are both done through Monte Carlo simulation: using the Newman-Ziff algorithm [23], N random sample paths

of the imbedded Markov chain are computed. Let $\{S_k^{(i)}\}_{0\le k\le m}$ denote the trajectory of the largest connected component's size for the ith realization of the process. We get the following estimates:

$$\hat{K} = \min\left\{k \le m,\ \frac{1}{N}\sum_{i=1}^{N} S_k^{(i)} \ge \sqrt{|\mathcal{V}|}\right\}, \qquad \hat{\chi} = \frac{1}{|\mathcal{V}|\cdot\hat{K}}\sum_{k=1}^{\hat{K}} S_k\,.$$

Finally, the empirical p-value can be expressed as

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{\hat{\chi} \le \hat{\chi}^{(i)}\}, \quad \text{where } \hat{\chi}^{(i)} = \frac{1}{|\mathcal{V}|\cdot\hat{K}}\sum_{k=1}^{\hat{K}} S_k^{(i)} \quad \text{for } i \in \{1,\dots,N\}.$$

4 Experiments

In order to assess the power of our test, we ran it on several synthetic graphs containing random anomalous trees. This section describes the procedure we used to generate the dataset, then presents our results and their interpretation.

4.1 Generation of the Dataset

The dataset is generated using the stochastic Kronecker graph model [21]. Kronecker graphs exhibit similar structural properties as real-world networks, most importantly power law-distributed degrees and small diameter. Hence, this model allows us to evaluate our test in a somewhat realistic setting.

Two parameter matrices are used: $\Theta_1 = [0.9\ 0.5;\ 0.5\ 0.3]$ (core-periphery network) and $\Theta_2 = [0.9\ 0.2;\ 0.2\ 0.9]$ (hierarchical network). For a given matrix and for $i \in \{12, 13, 14, 15\}$, we generate an undirected graph through i iterations of the Kronecker product, and only the largest connected component of this graph is kept in order to obtain a connected network with approximately 2^i vertices. Using this procedure, 10 graphs are generated for each pair of parameters (Θ, i). Thus, we evaluate our test on graphs with sizes ranging from a few thousands to a few tens of thousands of vertices, which covers a wide scope of potential use cases. For each synthetic graph, anomalies are then generated as follows: given $\delta \in (0, 1)$, a random subtree $\mathcal{S}$ containing a fraction δ of the vertices is drawn. Then, a random observation $X_e \sim \mathcal{N}(\mu\mathbb{1}\{e \in \mathcal{S}\}, 1)$ is independently drawn for each edge e of the graph (where μ is a fixed signal strength). For a given graph and a pair of parameters (δ, μ), 1000 anomalous signals $\mathbf{X} = (X_1, \dots, X_m)$ are generated. 1000 signals are also drawn from the null distribution (that is, $\mathbf{X} \sim \mathcal{N}(0, I_m)$, where I_m is the $m \times m$ identity matrix) for comparison. Finally, for each graph, the null distribution of the test statistic is estimated using 10000 random realizations of the percolation process. Using the obtained histogram, the empirical p-values associated with the normal and anomalous samples are derived, and we construct the Receiver Operating Characteristic (ROC) curve for each pair (δ, μ). This procedure exposes the influence of various parameters on the performance of our test, namely the graph size, the generator matrix, the size δ of the anomalous region and the signal strength μ.

4.2 Detectability Conditions – Empirical Study

Our results are displayed in Table 1 and Figs. 2 and 3. Our main interest is in finding out which parameters have the strongest influence on the power of the test, and we provide some key observations and interpretations below.

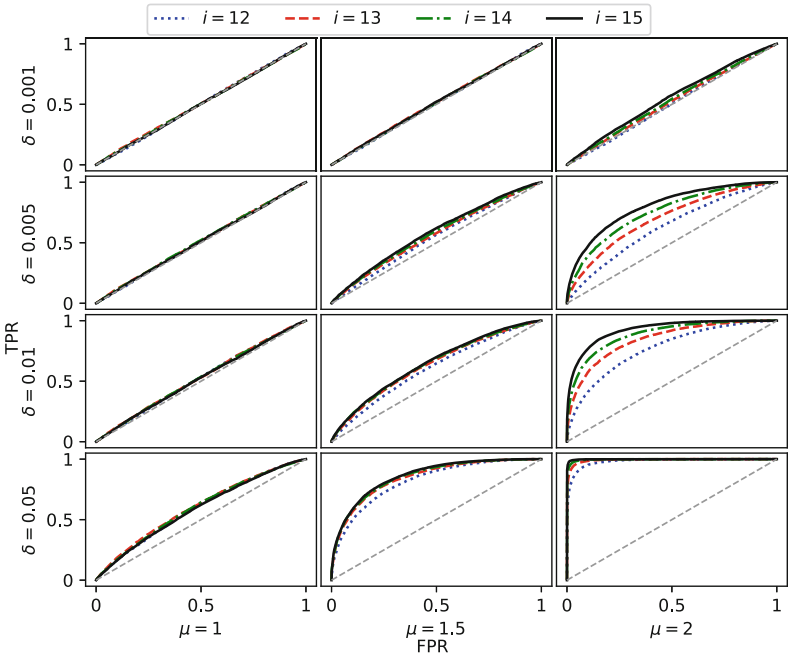

Fig. 2. Aggregated ROC curves of our test for 10 Kronecker graphs with initial matrix $\Theta_1 = [0.9\ 0.5; 0.5\ 0.3]$, for several values of the number of Kronecker product iterations i, the proportion δ of vertices in the anomalous tree and the signal strength μ.

Influence of the Graph Size. The first thing we notice in Figs. 2 and 3 is that for a given pair of parameters (δ, μ), the performance of the test consistently improves as the size of the graph increases. One possible explanation for this comes from percolation theory: before the phase transition, the size of the largest connected component is sublinear in the size of the graph. This implies that, for a fixed ratio of vertices in the anomalous component, the difference between the size of the latter and the expected size of the largest component grows with the graph size. Therefore, the anomalous component becomes more visible as the graph grows. Note, however, that some structural properties of our synthetic graphs (*e.g.* density) might not remain identical for different values of i. It is thus difficult to pinpoint the actual influence of the sole number of vertices.

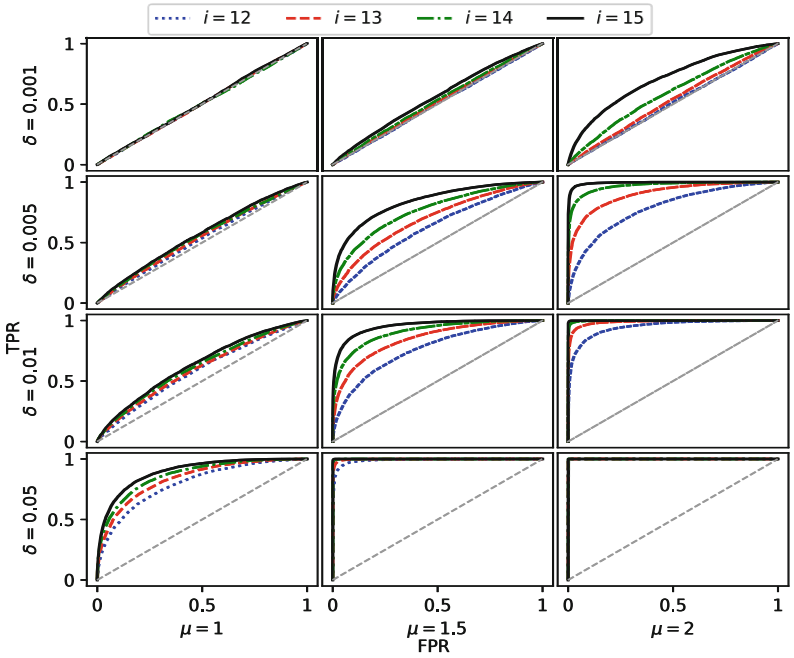

Fig. 3. Aggregated ROC curves of our test for 10 Kronecker graphs with initial matrix $\Theta_2 = [0.9\ 0.2; 0.2\ 0.9]$, for several values of the number of Kronecker product iterations i, the proportion δ of vertices in the anomalous tree and the signal strength μ.

Trade-Off Between δ and μ. As could be intuitively expected, our test performs better for higher values of δ and μ. More interestingly, these two parameters are intertwined: what makes an anomalous subgraph detectable is not only the number of vertices it contains (which is controlled by δ), but also the presence of a sufficient fraction of its edges among the most individually anomalous edges of the graph (which is controlled by μ). In terms of experimental results, this translates to poor performance when at least one of these parameters is too low. However, there seems to be a range of values of δ and μ in which a decrease in one can be made up for by an increase in the other. In particular, this implies that even small subgraphs can be detected by our test as long as the signal is strong enough. This is useful in "needle-in-a-haystack" scenarios such as network intrusion detection, where the anomalies one looks for are often localized.

Influence of the Graph Structure. As evidenced by Fig. 1, structural properties of the graph heavily influence the normal behavior of the percolation process, in turn affecting the viability of percolation-based cluster detection. This explains the observable difference in detection power between the two kinds of graphs we consider. Further analysis shows that the generator Θ_1 yields more heavy-tailed degree distributions, which is a plausible cause for the performance gap.

5 Discussion and Future Work

We now discuss the main properties of our test, identifying some limitations and providing leads for future work.

Table 1. Aggregated AUC score of our test for 10 Kronecker graphs, using various combinations of initial matrix Θ, number of iterations of the Kronecker product i, proportion δ of vertices in the anomalous tree and signal strength μ.

		Θ_1				Θ_2			
		$\delta = 0.001$	0.005	0.01	0.05	0.001	0.005	0.01	0.05
$i = 12$	$\mu = 1$	0.502	0.510	0.525	0.591	0.502	0.527	0.582	0.796
	1.5	0.505	0.542	0.603	0.819	0.502	0.626	0.763	0.990
	2	0.503	0.628	0.769	0.981	0.505	0.785	0.949	1.000
$i = 13$	1	0.507	0.513	0.528	0.602	0.505	0.540	0.595	0.838
	1.5	0.513	0.560	0.631	0.847	0.512	0.694	0.848	0.998
	2	0.518	0.699	0.845	0.993	0.531	0.902	0.988	1.000
$i = 14$	1	0.503	0.515	0.525	0.596	0.503	0.550	0.614	0.867
	1.5	0.508	0.570	0.639	0.855	0.524	0.764	0.908	1.000
	2	0.528	0.752	0.887	0.997	0.590	0.969	0.998	1.000
$i = 15$	1	0.500	0.509	0.522	0.586	0.508	0.565	0.634	0.897
	1.5	0.511	0.584	0.645	0.861	0.555	0.840	0.955	1.000
	2	0.551	0.801	0.925	0.999	0.706	0.994	1.000	1.000

Theoretical Guarantees. From a theoretical perspective, our setting is more complex than that of [6]: we consider arbitrary networks instead of regular lattices, and our test statistic depends on the whole sample path of the percolation process rather than the marginal behavior at a given occupation probability. Therefore, the search for theoretical guarantees for our test was left out of the scope of this work, although it would certainly be of great interest.

Computational Cost. The main advantage of our method is its computational efficiency. Indeed, computing the empirical p-value for a given graph and an observed signal only requires $N + 1$ runs of the Newman-Ziff algorithm, which has a very low cost. In contrast, a scan statistic-based test would require $N + 1$ runs of a combinatorial optimization algorithm (one for the observed data and N additional runs to estimate the distribution of the test statistic under the null). Even with a very efficient optimization method, this is significantly more intensive. In terms of complexity, our test requires sorting the observations X_i, running the Newman-Ziff algorithm $N + 1$ times, computing the mean sample path and the index K, and summing the first K values for each of the $N + 1$ trajectories, resulting in $\mathcal{O}(m(\log m + N))$ operations. Note that the algorithm

can be further optimized using the fact that the test statistic depends only on the first K steps of the percolation process. Although the exact value of K depends on the graph, we empirically observe that it is generally smaller than the number of vertices $|\mathcal{V}|$. Therefore, early stopping of the Newman-Ziff algorithm and partial sorting can reduce the complexity to $\mathcal{O}(m + |\mathcal{V}|(N + \log|\mathcal{V}|))$.

Detection Power. The expected downside of our method's low computational cost is a loss in detection power. Our simulations show, however, that the proposed test can detect reasonably small anomalous subgraphs in large enough ambient graphs, which is our main goal here. Moreover, it does not rely on prior knowledge of the alternative distribution and can be used with only a rough estimate of F_0, which improves its usability in realistic settings.

Although the influence of some factors on the performance of the test was left out of the scope of this work, a wider analysis would be an interesting topic for future work. These factors include the density of the graph and the shape of the anomalous subgraph. More specifically, we only evaluated our test in the case of random anomalous trees, which provides general results but no insight into the influence of the diameter and the density of the anomalous subgraph.

6 Conclusion

By extending previous work on percolation-based cluster detection to a more general setting, we propose a computationally efficient test to detect an anomalous connected subgraph in an edge-weighted network. The underlying intuition is that it is often possible to find out whether such a subgraph is present without explicitly finding it: instead of enumerating all possible candidates, a much faster method can be obtained by looking for properties of the whole graph which are affected by the apparition of an anomalous cluster. Our work suggests that percolation theory can provide such properties.

Since it scales easily to large graphs and does not rely on extensive knowledge of the null and alternative distributions of the observed signal, we argue that our method is applicable to real-world problems. Moreover, we show through extensive simulations that its detection power remains acceptable, and that it can in particular detect small anomalous regions in large graphs. Therefore, we think the link between cluster detection and percolation theory deserves further exploration, both from a theoretical and applied point of view.

References

1. Addario-Berry, L., Broutin, N., Devroye, L., Lugosi, G., et al.: On combinatorial testing problems. Ann. Stat. **38**(5), 3063–3092 (2010)
2. Aksoylar, C., Orecchia, L., Saligrama, V.: Connected subgraph detection with mirror descent on SDPs. In: ICML (2017)
3. Arias-Castro, E., Candes, E.J., Durand, A., et al.: Detection of an anomalous cluster in a network. Ann. Stat. **39**(1), 278–304 (2011)

4. Arias-Castro, E., Candès, E.J., Helgason, H., Zeitouni, O., et al.: Searching for a trail of evidence in a maze. Ann. Stat. **36**(4), 1726–1757 (2008)
5. Arias-Castro, E., Donoho, D.L., Huo, X., et al.: Near-optimal detection of geometric objects by fast multiscale methods. IEEE Trans. Inf. Theory **51**(7), 2402–2425 (2005)
6. Arias-Castro, E., Grimmett, G.R., et al.: Cluster detection in networks using percolation. Bernoulli **19**(2), 676–719 (2013)
7. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
8. Callaway, D.S., Newman, M.E., Strogatz, S.H., Watts, D.J.: Network robustness and fragility: percolation on random graphs. Phys. Rev. Lett. **85**(25), 5468 (2000)
9. Chung, F.: Spectral Graph Theory. American Mathematical Society, Providence (1997)
10. Chung, F., Horn, P., Lu, L.: Percolation in general graphs. Internet Math. **6**(3), 331–347 (2009)
11. Duczmal, L., Assuncao, R.: A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. Comput. Stat. Data Anal. **45**(2), 269–286 (2004)
12. Duczmal, L., Kulldorff, M., Huang, L.: Evaluation of spatial scan statistics for irregularly shaped clusters. J. Comput. Graph. Stat. **15**(2), 428–442 (2006)
13. Erdős, P., Rényi, A.: On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci **5**, 17–60 (1960)
14. Glaz, J., Naus, J., Wallenstein, S.: Scan Statistics. Springer, Berlin (2001)
15. Hooi, B., Song, H.A., Beutel, A., Shah, N., Shin, K., Faloutsos, C.: Fraudar: bounding graph fraud in the face of camouflage. In: KDD (2016)
16. Karrer, B., Newman, M.E., Zdeborová, L.: Percolation on sparse networks. Phys. Rev. Lett. **113**(20), 208702 (2014)
17. Kulldorff, M.: A spatial scan statistic. Commun. Stat. - Theory Methods **26**(6), 1481–1496 (1997)
18. Kulldorff, M., Huang, L., Pickle, L., Duczmal, L.: An elliptic spatial scan statistic. Stat. Med. **25**(22), 3929–3943 (2006)
19. Langovoy, M., Habeck, M., Schölkopf, B.: Spatial statistics, image analysis and percolation theory. arXiv preprint arXiv:1310.8574 (2013)
20. Langovoy, M., Wittich, O.: Robust nonparametric detection of objects in noisy images. J. Nonparametr. Stat. **25**(2), 409–426 (2013)
21. Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., Ghahramani, Z.: Kronecker graphs: an approach to modeling networks. J. Mach. Learn. Res. **11**, 985–1042 (2010)
22. Neil, J., Hash, C., Brugh, A., Fisk, M., Storlie, C.B.: Scan statistics for the online detection of locally anomalous subgraphs. Technometrics **55**(4), 403–414 (2013)
23. Newman, M.E., Ziff, R.M.: Fast Monte Carlo algorithm for site or bond percolation. Phys. Rev. E **64**(1), 016706 (2001)
24. Patil, G., Taillie, C., et al.: Geographic and network surveillance via scan statistics for critical area detection. Stat. Sci. **18**(4), 457–465 (2003)
25. Priebe, C.E., Conroy, J.M., Marchette, D.J., Park, Y.: Scan statistics on enron graphs. Comput. Math. Organ. Theory **11**(3), 229–247 (2005)
26. Qian, J., Saligrama, V.: Efficient minimax signal detection on graphs. In: NeurIPS (2014)
27. Qian, J., Saligrama, V., Chen, Y.: Connected sub-graph detection. In: AISTATS (2014)
28. Rozenshtein, P., Anagnostopoulos, A., Gionis, A., Tatti, N.: Event detection in activity networks. In: KDD (2014)

29. Sharpnack, J., Rinaldo, A., Singh, A.: Detecting anomalous activity on networks with the graph Fourier scan statistic. IEEE Trans. Signal Process. **64**(2), 364–379 (2015)
30. Sharpnack, J., Singh, A., Rinaldo, A.: Changepoint detection over graphs with the spectral scan statistic. In: AISTATS (2013)
31. Sharpnack, J.L., Krishnamurthy, A., Singh, A.: Near-optimal anomaly detection in graphs using Lovasz extended scan statistic. In: NeurIPS (2013)
32. Speakman, S., McFowland III, E., Neill, D.B.: Scalable detection of anomalous patterns with connectivity constraints. J. Comput. Graph. Stat. **24**(4), 1014–1033 (2015)
33. Wu, N., Chen, F., Li, J., Zhou, B., Ramakrishnan, N.: Efficient nonparametric subgraph detection using tree shaped priors. In: AAAI (2016)