Calibrated Nonparametric Scan Statistics for Anomalous Pattern Detection in Graphs

Chunpai Wang,[1] **Daniel B. Neill,** [2] **Feng Chen** [3]
[1] University at Albany - SUNY
[2] New York University
[3] The University of Texas at Dallas
cwang25@albany.edu, daniel.neill@nyu.edu, feng.chen@utdallas.edu

Abstract

We propose a new approach, the calibrated nonparametric scan statistic (CNSS), for more accurate detection of anomalous patterns in large-scale, real-world graphs. Scan statistics identify connected subgraphs that are interesting or unexpected through maximization of a likelihood ratio statistic; in particular, nonparametric scan statistics (NPSSs) identify subgraphs with a higher than expected proportion of individually significant nodes. However, we show that recently proposed NPSS methods are *miscalibrated*, failing to account for the maximization of the statistic over the multiplicity of subgraphs. This results in both reduced detection power for subtle signals, and low precision of the detected subgraph even for stronger signals. Thus we develop a new statistical approach to recalibrate NPSSs, correctly adjusting for multiple hypothesis testing and taking the underlying graph structure into account. While the recalibration, based on randomization testing, is computationally expensive, we propose both an efficient (approximate) algorithm and new, closed-form lower bounds (on the expected maximum proportion of significant nodes for subgraphs of a given size, under the null hypothesis of no anomalous patterns). These advances, along with the integration of recent core-tree decomposition methods, enable CNSS to scale to large real-world graphs, with substantial improvement in the accuracy of detected subgraphs. Extensive experiments on both semi-synthetic and real-world datasets are demonstrated to validate the effectiveness of our proposed methods, in comparison with state-of-the-art counterparts.

1 Introduction

Detecting "hotspots" or anomalous patterns in graphs is an important but challenging problem, with numerous critical applications in areas such as epidemiology, law enforcement, finance, and security. Among the powerful and widely used methods, the paradigm of scan statistics is one of the few that has a sound and general statistical basis (for related surveys see Glaz, Pozdnyakov, and Wallenstein (2009); Akoglu, Tong, and Koutra (2015); and Cadena, Chen, and Vullikanti (2018)). Graph-based scan statistics (Speakman, McFowland III, and Neill 2015; Speakman, Zhang, and Neill 2013; Chen and Neill 2014; Cadena, Chen, and Vullikanti 2019) identify connected subgraphs that are interesting or unexpected through maximization of a likelihood ratio statistic. The connectivity constraint is important because it ensures that subgraphs reflect changes due to localized anomalous processes (e.g., disease outbreaks, water pollution events). In particular, nonparametric scan statistic (NPSS) methods (Neill and Lingwall 2007; McFowland III, Speakman, and Neill 2013; Chen and Neill 2014) are designed without assuming any known background process on the graph. These approaches use historical data (assuming no anomalous patterns are present) to compute an empirical p-value for each graph node, and then compare the observed and expected number of significantly low p-values contained in connected subgraphs. Those with the largest scores are returned as the most anomalous subgraphs. However, as we show, NPSSs fail to account for the multiple hypothesis testing effects of searching over the huge space of connected subgraphs, reducing detection performance. In this work, we conduct a systematic study of this challenging problem and make the following **key contributions:**

- We show that recently proposed NPSS methods are *miscalibrated*, failing to account for the maximization of the statistic over the multiplicity of subgraphs. This results in both reduced detection power for subtle signals, and low precision of the detected subgraph.
- We develop a new statistical approach to recalibrate NPSS, correctly adjusting for multiple hypothesis testing and taking the underlying graph structure into account, substantially improving detection performance.
- We propose an efficient (approximate) algorithm and new, closed-form lower bounds on the expected maximum proportion of significant nodes for subgraphs of a given size, under the null hypothesis of no anomalous patterns. These advances, along with integration of recent core-tree decomposition methods, enable the CNSS approach to scale to large real-world graphs, with substantial improvement in the accuracy of detected subgraphs.
- Extensive experiments on semi-synthetic and real-world datasets show that our methods can detect anomalous subgraphs more accurately than state-of-the-art counterparts, while maintaining comparable time efficiency.

2 Related Work

As anomaly detection in graphs has a large literature, we refer to Akoglu, Tong, and Koutra (2015) and Cadena, Chen,

and Vullikanti (2018) for comprehensive surveys on this topic. For brevity, we will discuss the methods based on scan statistics for detecting anomalous connected subgraphs, including those based on parametric scan statistics and NPSSs.

Parametric scan statistics are defined as the likelihood ratio statistics of the hypothesis test, where under the null hypothesis $\mathcal{H}_0$, the attribute data of nodes within a candidate connected subgraph $\mathcal{S}$ are generated by a parameterized background process. Under the alternative hypothesis $\mathcal{H}_1(\mathcal{S})$, the attribute data are generated by a different parameterized distribution (a localized anomalous process). Depending on the assumptions on these two distributions, a variety of scan statistics are formulated, such as Positive Elevated Mean (Qian, Saligrama, and Chen 2014) and Expectation-based Poisson and Gaussian (Neill 2009), in addition to the Kulldorff Scan Statistic (Kulldorff 1997). While these methods have been shown to achieve high detection power across a variety of spatio-temporal graph datasets, they make strong parametric model assumptions, and performance degrades when these models are incorrect.

In comparison, *NPSSs* use historical data with no anomalous patterns to calibrate an empirical p-value for each node and are defined as likelihood ratio statistics of the nonparametric hypothesis test. Under the null hypothesis of no anomalous patterns ($\mathcal{H}_0$), the empirical p-values of nodes within a candidate connected subgraph ($\mathcal{S}$) follow a uniform distribution between 0 and 1. Under the alternative hypothesis ($\mathcal{H}_1$), the empirical p-values follow a different distribution. Depending on the specific form of the distribution under $\mathcal{H}_1$, different NPSSs are formulated, such as the Berk-Jones (Berk and Jones 1979), Higher Criticism (Donoho and Jin 2004), Kolmogorov–Smirnov (Massey Jr 1951), and Anderson-Darling scan statistics (Eicker 1979).

Optimizing scan statistics is challenging in the presence of connectivity constraints. A number of heuristic algorithms have been proposed for parametric scan statistics, such as additive GraphScan based on shortest paths in graphs (Speakman, Zhang, and Neill 2013), Steiner tree heuristics (Rozenshtein et al. 2014), and simulated annealing (Duczmal and Assuncao 2004). Qian, Saligrama, and Chen (2014) used linear matrix inequalities as a way to characterize the connectivity constraint and designed efficient iterative algorithms with convergence guarantees to optimize scan statistics (e.g., Positive Elevated Mean) that are convex after relaxation. Sharpnack, Rinaldo, and Singh (2015) proposed a computationally tractable algorithm with consistency guarantees. Several heuristics have been proposed for NPSSs, such as greedy growth (Chen and Neill 2014) and Steiner tree heuristics based on approximation of the underlying graph with trees (Wu et al. 2016). A depth-first-search based algorithm, named GraphScan, was proposed to exactly identify the most anomalous connected subgraphs for scan statistics (e.g., Kulldorff, Berk-Jones) that satisfy the "linear time subset scanning" (LTSS) property (Neill 2012), but has an exponential time complexity in the worst case (Speakman, McFowland III, and Neill 2015). An approximate algorithm built based on the color-coding technique (Alon, Yuster, and Zwick 1995) was designed for a large class of scan statistics with rigorous guarantees (Cadena, Chen, and Vullikanti 2019). Although it has the performance bound of $1-\epsilon$, its run time scales exponentially with the size of the most anomalous connected subgraphs.

Recent work by Reyna et al. (2021) and Chitra et al. (2021) demonstrates the miscalibration of the Gaussian scan statistic and presents a Gaussian mixture modeling approach to reduce this bias. As discussed in Appendix A.4 , the nonparametric scan statistics that we consider here differ fundamentally from the Gaussian scan, both in their assumptions about the true signal (distribution of p-values under $\mathcal{H}_1$) and in their maximization over a range of significance levels α.

3 Limitations of Nonparametric Scan

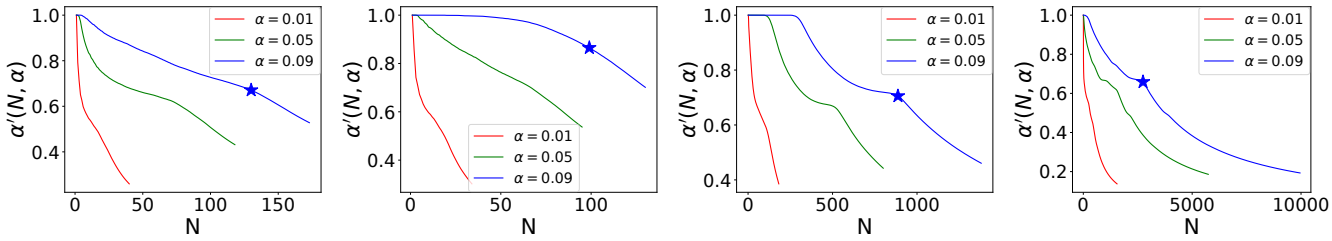
Given a graph $\mathbb{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V}$ being a set of n vertices and $\mathcal{E}$ being a set of m edges. Each node $v_i \in \mathcal{V}$ is associated with a feature vector $\mathbf{x}_i \in \mathbb{R}^N$ and its historical observations $\{\mathbf{x}_i^{(1)}, \cdots, \mathbf{x}_i^{(T)}\}$. We use the historical observations to convert the feature vector of each node ($\mathbf{x}_i$) to a single empirical p-value (p_i), using the two-stage empirical calibration procedure introduced in Chen and Neill (2014). Additional details on the computation of empirical p-values are provided in Appendix A.2 . Critically, under the null hypothesis $\mathcal{H}_0$, the current observations are assumed to be exchangeable with the null distribution of interest, and thus the p-values (computed by ranking the current observation against the historical observations and then normalizing the ranks) are asymptotically uniform on [0,1] under the null.

For instance, the graph could be a geospatial network, in which each node represents a county, two nodes are connected via an edge if they are spatially adjacent, and each node has a single feature, $x_i \in \mathbb{R}$, that is the number of confirmed Covid-19 disease cases for the current week. The goal is to detect the most anomalous cluster or connected subgraph (representing a potential Covid-19 outbreak). In this case, the empirical p-value p_i is simply the proportion of the historical observations with case counts that are greater than or equal to the current observation.

We denote by $\mathbb{G}_{\mathcal{S}} = (\mathcal{S}, \mathcal{E}_{\mathcal{S}})$ the subgraph induced by the subset $\mathcal{S} \subseteq \mathcal{V}$ and $\mathbb{M} = \{\mathcal{S} \mid \mathcal{S} \subseteq \mathcal{V}, \mathbb{G}_{\mathcal{S}} \text{ is connected in } \mathbb{G}\}$ the set of all possible connected subsets. The problem of NPSS-based anomalous pattern detection is defined as the connected subgraph optimization problem:

$$\begin{aligned}\max_{\mathcal{S}\in\mathbb{M}} F(\mathcal{S}) &= \max_{\mathcal{S}\in\mathbb{M}} \max_{\alpha\le\alpha_{\max}} \Phi\left(\alpha, N_\alpha(\mathcal{S}), N(\mathcal{S})\right) \\ &= \max_{\alpha\le\alpha_{\max}} \max_{\mathcal{S}\in\mathbb{M}} \Phi\left(\alpha, N_\alpha(\mathcal{S}), N(\mathcal{S})\right)\end{aligned} \tag{1}$$

where $F(\mathcal{S}) := \max_{\alpha\le\alpha_{\max}} \Phi\left(\alpha, N_\alpha(\mathcal{S}), N(\mathcal{S})\right)$ refers to the general form of NPSS defined by McFowland III, Speakman, and Neill (2013), $\mathcal{S} \subseteq \mathcal{V}$ is a connected set of nodes, $N_\alpha(\mathcal{S}) = \sum_{v\in\mathcal{S}} \mathbf{1}\{p(v) \le \alpha\}$ refers to the number of p-values in subset $\mathcal{S}$ that are significant at level α, and $N(\mathcal{S}) = \sum_{v\in\mathcal{S}} 1$ refers to the total number of p-values in subset $\mathcal{S}$. The function $\Phi\left(\alpha, N_\alpha(\mathcal{S}), N(\mathcal{S})\right)$ compares the observed number of significant p-values $N_\alpha(\mathcal{S})$ at level α to the expected number of significant p-values $\mathbb{E}\left[N_\alpha(\mathcal{S})\right] = \alpha N(\mathcal{S})$ under the null hypothesis $\mathcal{H}_0$. Critically, NPSSs optimize the significance level α between 0 and some constant $\alpha_{\max} < 1$. Maximization over a range of α values allows



(a) $|\mathcal{V}| = 1000, p = 0.01$. (b) $|\mathcal{V}| = 1000, p = 0.02$. (c) WikiVote. (d) CondMat.

Figure 1: Simulation results (a) and (b) on Erdos-Renyi graphs and (c) and (d) on real world graphs under $\mathcal{H}_0$. Each curve ends at the value of N such that all significant nodes in the graph are included. The starred point is the combination of N and α for which $N \times \mathtt{KL}(\alpha', \alpha)$ is maximized. Descriptive statistics of the `WikiVote` and `CondMat` datasets can be viewed in Appendix C.1.

accurate detection of signals with either a small number of highly significant p-values or a larger number of moderately significant p-values. In practice, rather than considering all $\alpha \leq \alpha_{\max}$, we consider a discrete set of α values, $\mathcal{L} = \{0.001, \cdots, 0.009, 0.01, \cdots, 0.09\}$, and solve the constrained optimization $\max_{S \in \mathbb{M}} \Phi\left(\alpha, N_\alpha(\mathcal{S}), N(\mathcal{S})\right)$ for each α in $\mathcal{L}$ to find the most anomalous subgraphs.

Here we focus on the Berk-Jones (BJ) nonparametric scan statistic, without loss of generalization to other NPSSs (see Appendix A.3). The BJ statistic is defined as

$$\Phi_{BJ}\left(\alpha, N_\alpha(\mathcal{S}), N(\mathcal{S})\right) = N(\mathcal{S}) \times \mathtt{KL}\left(\frac{N_\alpha(\mathcal{S})}{N(\mathcal{S})}, \alpha\right), \quad (2)$$

where `KL` is the Kullback-Liebler divergence between the observed and expected proportions of significant p-values: $\mathtt{KL}(a, b) = a \log(a/b) + (1-a) \log((1-a)/(1-b))$. The BJ statistic is the log-likelihood ratio statistic for testing whether the empirical p-values follow the `Uniform`[0, 1] distribution or a piecewise constant distribution where $\Pr(p < \alpha) > \alpha$. Please see Appendix A.1 for details.

Despite their effectiveness for anomalous pattern detection in graphs, NPSSs were originally designed without taking into consideration the multiple hypothesis testing effect resulting from the multiplicity of subgraphs. In particular, it follows from the assumption of uniform p-values under $\mathcal{H}_0$ made by NPSSs that the expected proportion of individually significant nodes within a connected subset $\mathcal{S}$ under $\mathcal{H}_0$ is $\mathbb{E}[N_\alpha(\mathcal{S})/N(\mathcal{S})] = \alpha$. However, this is true for a randomly selected connected subset, but not for connected subsets $\mathcal{S}$ that are identified by maximizing the NPSS score. Even when the null hypothesis $\mathcal{H}_0$ holds, and p-values are uniform on [0,1], we find that the expected proportion of individually significant nodes within the highest-scoring connected subsets, denoted as α', is typically much larger than α, which we refer to as *miscalibration*. More precisely, we define α' as the expected maximum proportion of significant nodes for all connected subgraphs of a given size N: $\alpha'(N, \alpha) = \mathbb{E}[\max_{\mathcal{S} \in \mathbb{M}, |\mathcal{S}| = N} N_\alpha(\mathcal{S})/N]$.

To illustrate the relationship between α' and α, we conduct simulations on both Erdos-Renyi (ER) random graphs and two real world graphs: WikiVote ($|\mathcal{V}| = 7066, |\mathcal{E}| = 100736$, density $= 0.004$) and CondMat ($|\mathcal{V}| = 21363, |\mathcal{E}| = 91286$, density $= 0.0004$). First, we generate 100 random graphs with $|\mathcal{V}| = 1,000$ and edge probabilities $p \in \{0.01, 0.02\}$, respectively. We simulate p-values under $\mathcal{H}_0$ for each ER random graph, and calculate the average α' for each $N \in \{1, 2, ..., |\mathcal{V}|\}$ and $\alpha \in \{.01, .05, .09\}$ among all the 100 random graphs, as shown in Figure 1(a)-(b). We simulate p-values under $\mathcal{H}_0$ on the graphs WikiVote and CondMat for 100 times and calculate the average α' for each N and α, as shown in Figure 1(c)-(d). The results indicate that the expected maximum proportion $\alpha'(N, \alpha)$ for given values of N and α is much higher than the expected proportion $\mathbb{E}[N_\alpha(\mathcal{S})/N] = \alpha$. The implication is that, even when the null hypothesis $\mathcal{H}_0$ holds and there are no true subgraphs of interest, there exist subgraphs $\mathcal{S}$ with $N_\alpha(\mathcal{S}) \gg \alpha N(\mathcal{S})$, and thus very high NPSS scores. The amount of difference between $\alpha'(N, \alpha)$ and α is a function of N, α, and the graph structure. We observe that $\alpha'(N, \alpha)$ decreases with N but remains much higher than α for the entire range of N.

The results of this discrepancy between α' and α are threefold. First, the maximum NPSS score under the null hypothesis $\mathcal{H}_0$ will be large. To see this, we compute the expected score $N \times \mathtt{KL}(\alpha', \alpha)$ for each combination of α and N for each of Figure 1(a)-(d), and show the highest-scoring combination on each graph as a star icon. The corresponding scores range from 131.7 for Figure 1(a) to 2677 for Figure 1(d). These large scores under the null result in *reduced detection power*, since the NPSS score of the true anomalous subgraph must exceed a larger threshold (i.e., the 95th percentile of the maximum NPSS scores under $\mathcal{H}_0$) to be considered significant. Second, NPSS will be *biased toward detecting clusters at larger α thresholds*, even if the true signal is for a much smaller α. We observe that, for all four graphs in Figure 1, the null score is maximized at the largest of the three α values considered. Third, NPSS will identify overly large clusters which include many nodes that have significant p-values just by chance, resulting in *reduced precision* of the detected cluster. We observe that, for all four graphs in Figure 1, the null score is maximized for a large value of N, where almost all of the significant nodes in the graph are included in the detected cluster. This observation is also supported by low precision (and low F-scores) for all uncalibrated scan methods in our evaluation results below. An additional, concrete example of miscalibration for the (uncalibrated) BJ statistic is provided in Appendix B.1 .

4 Calibrated Nonparametric Scan Statistics

Thus we have shown that uncalibrated NPSS methods discover large, high-scoring anomalous connected subsets even under the null hypothesis $\mathcal{H}_0$, resulting in reduced detection power and precision. This observation motivates us to develop a new approach to recalibrating the non-parametric scan statistic that accounts for multiple testing (and the resulting, large differences between α' and α), for improved detection performance. Hence, we propose *Calibrated Nonparametric Scan Statistics (CNSS)*, where $F(\mathcal{S}) = \max_{\alpha \leq \alpha_{\max}} \Phi\left(\alpha, N_\alpha(\mathcal{S}), N(\mathcal{S})\right)$ as above, but the expected proportion of significant p-values ($\mathbb{E}[N_\alpha(\mathcal{S})/N(\mathcal{S})] = \alpha$) used in $\Phi(\cdot)$ is replaced with the expected *maximum* proportion of significant p-values $\alpha'(N, \alpha)$ over all subgraphs of size N under the null hypothesis $\mathcal{H}_0$. For example, our proposed Calibrated Berk Jones (CBJ) statistic is defined as

$$\Phi_{\text{CBJ}}\left(\alpha, N_\alpha(\mathcal{S}), N(\mathcal{S})\right) = N(\mathcal{S}) \times \text{KL}\left(\frac{N_\alpha(\mathcal{S})}{N(\mathcal{S})}, \alpha'(N(\mathcal{S}), \alpha)\right) \tag{3}$$

where

$$\alpha'(N, \alpha) = \frac{\mathbb{E}\left[\max_{\mathcal{S} \in \mathbb{M}, |\mathcal{S}|=N} N_\alpha(\mathcal{S})\right]}{N}. \tag{4}$$

Critically, this approach guarantees that, under the null hypothesis $\mathcal{H}_0$ that current and historical observations are exchangeable, for any N and α, the expected ratio $N_\alpha(\mathcal{S})/N$ for the highest-scoring subgraph $\mathcal{S}$ of size N is equal to α', thus adjusting for the multiplicity of subgraphs and correctly calibrating across N and α. See Appendix B.1 for more explanation on the correctness of the calibration approach.

As shown in Figure 1, the expected maximum proportion of significant nodes α' depends on the subgraph size N, the significance level α, and the graph structure. To estimate $\alpha'(N, \alpha)$ for a given graph, we use a *randomization test* to estimate $\mathbb{E}[\max_{\mathcal{S} \in \mathbb{M}, |\mathcal{S}|=N} N_\alpha(\mathcal{S})]$ for each N and α. In particular, we create a large number ($K = 200$) of replica datasets under the null hypothesis $\mathcal{H}_0$, where each node of the input graph $\mathbb{G}$ has its p-value redrawn uniformly at random from $[0, 1]$. We then apply the efficient approximate algorithm proposed in Section 4.1 to solve the constrained set optimization problem $\max_{\mathcal{S} \in \mathbb{M}, |\mathcal{S}|=N} N_\alpha(\mathcal{S})$ for each combination $(N, \alpha) \in \{1, \cdots, |\mathcal{V}|\} \times \mathcal{L}$. Based on the K replica datasets, for each combination of N and α, we collect K samples of the maximum number of significant nodes $\max_{\mathcal{S} \in \mathbb{M}, |\mathcal{S}|=N} N_\alpha(\mathcal{S})$ and use the samples to estimate $\alpha'(N, \alpha)$ under $\mathcal{H}_0$. The same algorithm is applied to the original dataset to identify the most significant subgraph $\max_{\mathcal{S} \in \mathbb{M}, |\mathcal{S}|=N} N_\alpha(\mathcal{S})$ for each $(N, \alpha) \in \{1, \cdots, |\mathcal{V}|\} \times \mathcal{L}$, and then the subgraph with the highest score $F(\mathcal{S}) = \max_{\alpha \leq \alpha_{\max}} \Phi_{\text{CBJ}}(\alpha, N_\alpha(\mathcal{S}), N(\mathcal{S}))$ is returned. More details are provided in Algorithm 1 in Appendix B.3 .

4.1 An Efficient Approximate Algorithm

The fundamental computational challenge of CNSS is to find the maximum number of significant nodes, $\max_{\mathcal{S} \in \mathbb{M}, |\mathcal{S}|=N} N_\alpha(\mathcal{S})$, for connected subgraphs of every size $N \in \{1, \cdots, |\mathcal{V}|\}$. One approach for doing so would be, for each N and each $\alpha \in \mathcal{L}$, to separately run a prize-collecting Steiner tree (PCST) algorithm to identify the maximum N_α. The PCST is NP-hard but can be approximated in $O(|\mathcal{V}|^2 \log |\mathcal{V}|)$ time; however, computing the PCST for each N would then result in an insufficiently scalable $O(|\mathcal{V}|^3 \log |\mathcal{V}|)$ algorithm. As an alternative, we propose a novel algorithm which approximates the maximum N_α for each $N \in \{1, \ldots, |\mathcal{V}|\}$, for a given value of α, in a single, efficient run. This process must then be repeated for each value of α under consideration.

The pseudocode of estimating the maximum N_α for each N under a given significance threshold α is described in Algorithm 2 in Appendix B.4 . The approach is based on repeated merging of nodes with the highest proportion of significant p-values. Given a graph with node-level p-values, we first merge all adjacent significant nodes, and maintain a list $\mathcal{Z}$ of merged nodes sorted by significance ratio $N_\alpha(\mathcal{S})/N(\mathcal{S})$. We repeatedly choose the merged node with highest significance ratio and perform one of the following three merge steps: (1) add an adjacent node which contains some or all significant p-values; (2) add an adjacent non-significant node that is also adjacent to at least one other significant node; or (3) add the highest-degree non-significant neighbor. At each merge step, our method will try all three options and utilize the one leading to a merged node with the highest $N_\alpha(\mathcal{S})/N(\mathcal{S})$ ratio; this is repeated until the list $\mathcal{Z}$ only contains a single merged node. The advantage of this merging process is that we can keep track of the maximum $N_\alpha(\mathcal{S})$ for each $N(\mathcal{S})$ and iteratively update these values throughout the entire merging process. In the end, we have a list of estimated max $N_\alpha(\mathcal{S})$ for $N(\mathcal{S}) \in \{1, \cdots, |\mathcal{V}|\}$.

The overall time complexity of this algorithm is $\mathcal{O}(k|\mathcal{V}| + |\mathcal{V}| \log |\mathcal{V}|)$ where k is the largest degree of a node in the graph. See Appendix B.6 for a more detailed analysis.

4.2 Lower Bounds for the Expected Maximum Proportion of Significant Nodes, $\alpha'(N, \alpha)$

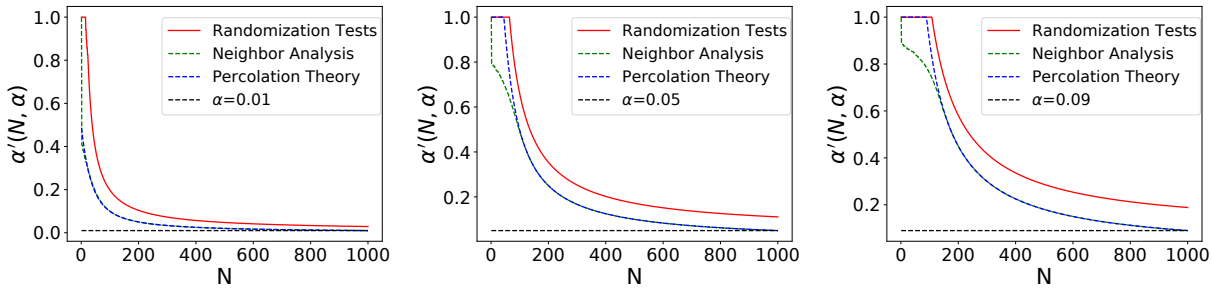
One limitation of our proposed calibration method is that it requires randomization tests to calibrate $\alpha'(N, \alpha)$ which are time-consuming for large graphs. Here we explore two strategies for obtaining closed-form lower bounds of $\alpha'(N, \alpha)$, thus avoiding the time-consuming randomization.

Lower Bound from Network Neighborhood Analysis The first approach is based on neighborhood analysis, and we denote the obtained lower bound of α' as α'_1. We lower bound the maximum number of significant nodes N_α for any given subgraph size N under $\mathcal{H}_0$, by identifying a subgraph of size N with expected number of significant nodes $\mathbb{E}[N_\alpha]$. Given any subgraph $\mathcal{S}$, let the exterior ("ext") degree of $\mathcal{S}$ be the number of edges between vertices $v_i \in \mathcal{S}$ and $v'_i \notin \mathcal{S}$.

Theorem 1. *For each $c \in \{1, ..., |\mathcal{V}|\}$, let k_c be the largest ext-degree of a connected subgraph of size c. Then for any $N \in \{1, ..., |\mathcal{V}|\}$ such that $c \leq N \leq c + k_c$, a lower bound for $\mathbb{E}[\max_{\mathcal{S} \in \mathbb{M}, |\mathcal{S}|=N} N_\alpha(\mathcal{S})]$ is: $c\alpha + \min(k_c\alpha, N - c)$.*

Proof. See Appendix B.2 . □

Given that high ext-degree subgraphs are more likely to connect more significant nodes, we first select the highest



(a) ER, $p = 0.05, \alpha = 0.01$ (b) ER, $p = 0.05, \alpha = 0.05$ (c) ER, $p = 0.05, \alpha = 0.09$

Figure 2: Lower Bounds of α' Compared with Empirical Distribution by Randomization Tests.

ext-degree node for $c = 1$ and count the its neighbors as k_c, and continue the process by increasing c and adding the highest ext-degree neighbor (i.e., with the highest number of neighbors not in $\mathcal{S}$) into $\mathcal{S}$. This approximation potentially underestimates k_c but cannot overestimate it, thus remaining a lower bound. For each N, we obtain multiple lower bounds α'_1 due to all the values of c under consideration, and we choose the largest (tightest) lower bound for each N.

Lower Bound from Percolation Theory The second lower bound, α'_2, is based on percolation theory on Erdos-Renyi (ER) graphs (Erdős and Rényi 1960; Achlioptas, D'Souza, and Spencer 2009). Given a large ER graph with $|\mathcal{V}|$ nodes and edge probability p, the average node degree is $\langle k \rangle = (|\mathcal{V}| - 1)p$. Percolation theory states that if a sufficiently large fraction of the graph nodes, $\rho > \frac{1}{\langle k \rangle}$, are "marked", then with high probability, there exists a connected subgraph $\mathcal{S}$ consisting of only marked nodes, with $|\mathcal{S}|$ equal to a constant fraction P_∞ of $|\mathcal{V}|$. More precisely, as shown by Erdős and Rényi (1960) and Bollobás and Erdos (1976), P_∞ is the solution to the equation, $P_\infty = \rho(1 - \exp(-\langle k \rangle P_\infty)$. We apply this result by "marking" both significant and (as needed) non-significant graph nodes to reach the percolation threshold, allowing us to prove:

Theorem 2. *For an Erdos-Renyi $(|\mathcal{V}|, p)$ graph with average degree $\langle k \rangle = (|\mathcal{V}| - 1)p$, with high probability,*

$$\alpha' \geq \min\left(1, \frac{\alpha|\mathcal{V}|}{N}\left(1 - \exp\left(-\langle k \rangle \frac{N}{|\mathcal{V}|}\right)\right)\right).$$

Proof. See Appendix B.2 . □

We show averaged α' lower bounds on 100 Erdos-Renyi graphs with size 1000 and $p = 0.05$ in Figure 2 for $\alpha \in [0.01, 0.05, 0.09]$. Compared to the true α' obtained from randomization testing, we observe empirically that the lower bounds α'_2 from percolation theory are tighter than the lower bounds α'_1 from neighbor analysis. However, we do not have theoretical results on the tightness of these bounds. We also note that the percolation bound is only guaranteed to be a lower bound on α' when the graph is Erdos-Renyi, while the neighbor analysis guarantees a lower bound for all graphs.

4.3 Core-Tree Decomposition

Core-whiskers (or core-periphery) structure commonly exists in many real-world networks, such as social networks, transportation networks, and the World Wide Web (Rombach et al. 2014; Leskovec et al. 2009). That is, real-world networks can be viewed as a set of low tree-width periphery surrounding a core consisting of a small fraction of nodes. The core tends to be an expander graph and has similar properties to random graphs (Leskovec et al. 2008). We first apply core-tree decomposition (Maehara et al. 2014) to decompose the graph into a small, dense core and a low-treewidth periphery. One benefit is that the small core keeps the general skeleton and connectivity of the entire graph, enabling adjacent, significant nodes from the whiskers to be incorporated into the detected subgraph. Thus we apply tree-node compression which merges the significant nodes in each single tree into an adjacent core node for follow-up optimization in a smaller core. If multiple core nodes are adjacent to a significant tree node, then we compress the significant tree node into the most significant (lowest p-value) core node. See Appendix B.5 for details of the compression procedure.

5 Experiments

In this section, we investigate four main research questions:

Q1. Subgraph Detection: Does our proposed `CNSS` have a better performance than state-of-the-art baselines on the task of anomalous subgraph detection?

Q2. Calibration: How does calibration affect detection performance, as a function of signal strength and graph structure?

Q3. Lower Bounds: How does the use of lower bounds of α', instead of α' obtained via randomization tests, affect detection performance?

Q4. Core Tree Decomposition: How does integrating core-tree decomposition into `CNSS` affect the detection performance and run time?

5.1 Experiment Setup

Datasets: We use five semi-synthetic datasets from the Stanford Network Analysis Project (SNAP [1]), including 1) `WikiVote`; 2) `CondMat`; 3) `Twitter`; 4) `Slashdot`; and 5) `DBLP`. We leverage the graph structure of these five networks, and simulate the true subgraph $\mathcal{S}$ using a random walk with size $\approx 0.01|\mathcal{V}|$. We generate the p-value of each graph node assuming Gaussian signals, $x_i \sim N(\mu_i, 1)$

[1] https://snap.stanford.edu/data/

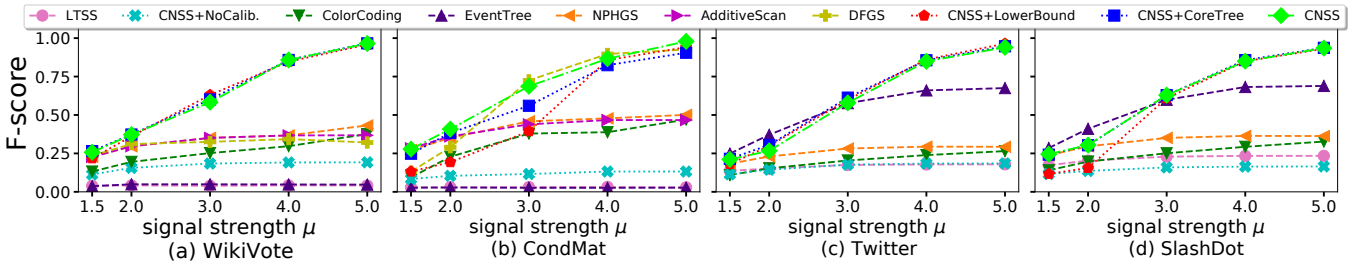


Figure 3: F-score of each method under different signal strengths and network structures (best viewed in color). Experiments of AdditiveScan and DFGS on Twitter and SlashDot datasets took over 2 weeks of clock time to run on 250 CPUs, therefore we do not report them. See Appendix C.4 , Table 6 , for the corresponding numerical results with significance tests.

and $p_i = 1 - \texttt{CDF}(x_i)$, where $\mu_i = \mu$ for $v_i \in \mathcal{S}$, and $\mu_i = 0$ (and thus $p_i \sim \texttt{Uniform}[0,1]$) for $v_i \in \mathcal{V} \setminus \mathcal{S}$. Here $\mu \in [1.5, 2, 3, 4, 5]$ is the signal strength and $\texttt{CDF}(\cdot)$ assumes the standard normal distribution. We report the average performance over 50 runs of simulations of true subgraphs and p-values on each network structure. See Appendix C.1 for more details of datasets, and Appendix C.5 for simulation results with piecewise constant p-values rather than Gaussian signals (i.e., assuming the BJ model is correctly specified).

Baseline Methods: We compare CNSS with 6 baselines, including 1) Linear Time Subset Scanning (LTSS) (Neill 2012); 2) EventTree (Rozenshtein et al. 2014); 3) Non-parametric Heterogeneous Graph Scan (NPHGS) (Chen and Neill 2014); 4) AdditiveScan (Speakman, Zhang, and Neill 2013) 5) Depth First Graph Scan (DFGS) (Speakman, McFowland III, and Neill 2015); and 6) ColorCoding (Cadena, Chen, and Vullikanti 2019). We summarize the limitations and time complexity of each competing method in Appendix C.5 .

Ablation Study: We also validate the effectiveness of our proposed components by comparing CNSS with methods: 1) CNSS+NoCalib, which removes calibration from CNSS, performing the same search but using the original α instead of α' in the score function; 2) CNSS+LowerBound, which replaces the randomization test with the tightest lower bound, $\max(\alpha'_1, \alpha'_2)$, of α'; and 3) CNSS+CoreTree, which integrates core-tree decomposition into CNSS.

Evaluation Metrics: We evaluate the detection performance of the CNSS and competing methods on running time, detection power, precision, recall, and F-score (see definitions in Appendix C.3). We report the results of F-score in the paper, and the remaining metrics are shown in Appendix C.4.

5.2 Results

Subgraph Detection: As shown in Figure 3, CNSS outperforms all baselines in terms of F-score when the event signal is strong, and it consistently has good and stable performance for different strengths of event signal and network structures. Competing methods have low precision, and thus low F-score, even when the signal is strong. In addition, we average the F-score over all signal strengths and network structures under consideration for each method, and we observe the following performance order: CNSS > CNSS+CoreTree > CNSS+LowerBound > DFGS > AdditiveScan > NPHGS > EventTree > ColorCoding > CNSS+NoCalib > LTSS. The average F-score of CNSS is 0.603, while the best-performing baseline method DFGS has average F-score 0.451. See Appendix C.4 for additional performance results. Appendix C.5 shows very similar results for piecewise constant signals.

Calibration: Calibration significantly improves detection performance across different signal strengths and various network structures. Specifically, the calibrated BJ scan statistic helps to pinpoint the true cluster as the strength of signal increases. On the contrary, all baselines, as well as the uncalibrated version of CNSS, fail to achieve accurate detection (as measured by F-score) for all network structures under consideration. These results demonstrate that calibration, rather than the search procedure for detecting anomalous subgraphs, is driving the difference in performance between methods. Our proposed search procedure simply enables calibration by making it computationally feasible to find $\max_{S:|S|=N} N_\alpha(S)$ for each combination of N and α.

Lower Bounds: Based on the empirical results on real-world networks, we find that our derived lower bounds provide substantial performance improvement on real-world networks, as shown as CNSS+LowerBound in Figure 3. Overall performance of the lower bound is lower than that of the randomization testing-based CNSS approach, particularly for low signal strengths, but CNSS+LowerBound substantially outperforms the baselines with respect to precision and F-score, particularly for stronger signals. Most importantly, computing lower bounds of α' is much faster than computing α' using randomization tests, resulting in a 400x to 2200x speedup for the various network structures under consideration. See Table 4 in Appendix C.4 .

Core Tree Decomposition: Core-tree decomposition substantially reduces run time for all datasets and does not significantly change detection performance. We see that CNSS+CoreTree is 2x faster on WikiVote dataset and 20x faster on CondMat dataset than CNSS. With core-tree decomposition, CNSS is more scalable than baseline methods including ColorCoding, NPHGS, AdditiveScan, and DFGS. While it is still more computationally expensive than LTSS and EventTree, our proposed method has much better detection performance. See Appendix C.4 for details.

5.3 Case Studies

We now compare the anomalous subgraphs detected by our CNSS method to those identified by two of the compet-

	# of weeks detected	avg. # of counties detected per week	avg. population of detected counties	avg. confirmed cases per week	avg. deaths per week (2 weeks lag)	avg. death rate (2 weeks lag) $\times 10^{-5}$
CNSS 1st	16	294.19	49369759.69	86596.81	4166.44	**8.44**
CNSS 2nd	15	60.67	10151920.33	14001.60	520.6	5.13
CNSS 3rd	13	7.69	4480384.39	10877.31	207	4.62
LTSS 1st	17	632.24	111861408.00	138212.47	5986	5.35
LTSS 2nd	14	5.14	802079.71	678.43	8.71	1.09
LTSS 3rd	4	9.25	2505224.25	1935.50	34.25	1.37
EventTree 1st	16	566.13	96492336.44	134612.50	5739.69	5.95
EventTree 2nd	7	2.14	762258.57	579.43	32.14	4.22
EventTree 3rd	1	2	299612.00	262	13	4.34

Table 1: COVID-19 Case Study: Top-3 Detected Subgraphs for Each Method

ing methods (LTSS and EventTree) on two real-world datasets, COVID-19 infection rates and Twitter data related to the Black Lives Matter movement. We note that the ColorCoding, NPHGS, AdditiveScan, and DFGS approaches were not able to scale to these large real-world datasets. We show the COVID-19 case study in the paper and BlackLivesMatter case study in Appendix D.1 .

COVID-19 Confirmed Cases Subgraph Discovery We study our proposed method on COVID-19 data[2] to discover significant infected regions over time. This dataset contains the daily confirmed cases for 3,234 counties in the USA across over 25 weeks from January 22-July 8, 2020. We build a spatial-temporal graph with 80,850 nodes and 850,725 edges based on the weekly confirmed cases and county adjacency (see Appendix D.2 for more details), where each node represents a county in one week. In addition to the edges that represent adjacency between counties (which are identical for each week t), we add an undirected temporal edge from each node i in week t to node i in week $t+1$ as well as undirected edges from each node i in week t to all neighboring nodes j in week $t+1$. The p-value of each node is generated based on the rank of the weekly confirmed cases to county population ratio divided by the total number of nodes in the graph. Therefore, a higher ratio of the number of weekly confirmed cases to the county population indicates a higher rank and thus a smaller p-value.

We apply our proposed method on this spatial-temporal graph and discover three subgraphs that are significant (as identified using randomization tests on 100 runs under the null hypothesis). The statistics of these three discovered subgraphs are shown in Table 11 in Appendix D.2 .

As shown in Table 1, our CNSS method detects a significant connected subgraph of counties that have a 42% higher death rate two weeks later, as compared with the top-1 subgraphs detected by LTSS and EventTree. The use of the two-week-lagged death rate as an evaluation metric better identifies the anomaly in true COVID-19 cases than the confirmed cases rate, which was highly affected in many areas by insufficient testing resources. (Note that death rate data is not provided to the detection algorithms.) The visualization of the highest-scoring subgraphs detected by different methods is shown in D.2 . We see that the baseline methods cannot discover a cohesive subgraph due to the poorly calibrated objective function, instead showing a dispersed pattern across much of the country. In contrast, our method is capable of detecting more impacted geographic regions, for better targeting of needed health resources.

[2]https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

6 Limitations and Conclusions

While CNSS achieves state of the art performance for anomalous pattern detection on graphs, it has two main limitations. First, the randomization test-based calibration approach is time-consuming, particularly for large-scale graphs. Although our proposed closed-form lower bounds of $\alpha'(N, \alpha)$ avoid the need for randomization tests and hence reduce the time cost of CNSS significantly, detection power is reduced when the anomalous signal strength is low, as shown in Figure 3. Second, our proposed efficient algorithm is heuristic rather than exact, and thus is not guaranteed to discover the maximum number of significant nodes N_α for each subgraph size N. However, as discussed in Section 2, subgraph detection is very challenging in the presence of connectivity constraints and no methods exist that have rigorous guarantees and at the same time are scalable to large graphs. For the calibrated scan, the computational problem is even more difficult: we must identify the subgraph with the largest number of significant p-values N_α for each subgraph size N and significance level α, which prevents us from using previous methods that search for a single highest-scoring subgraph. Finally, since the problem of pattern detection in graphs is general, detection approaches could be used for negative as well as beneficial social impacts, such as monitoring of social media by an oppressive government.

In summary, we demonstrated that existing nonparametric scan statistic methods are miscalibrated for anomalous pattern detection in graphs, and developed a new statistical approach to recalibrate NPSSs to account for the multiple hypothesis testing effect of the graph structure. We proposed a more efficient algorithm and new, closed-form lower bounds, and integrated recent core-tree decomposition methods, to enable our proposed CNSS approach to scale to large, real-world graphs. We observed outstanding performance of our method compared with six state-of-the-art baselines on five real-world datasets under various signal strengths and network structures. Finally, we applied CNSS to two real-world applications, and found more meaningful subgraphs compared with competing methods.

Acknowledgements

The work of Feng Chen is supported by the National Science Foundation (NSF) under Grant Number #1815696 and #1750911.

References

Achlioptas, D.; D'Souza, R. M.; and Spencer, J. 2009. Explosive percolation in random networks. *Science*, 323(5920): 1453–1455.

Akoglu, L.; Tong, H.; and Koutra, D. 2015. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29(3): 626–688.

Alon, N.; Yuster, R.; and Zwick, U. 1995. Color-coding. *Journal of the ACM*, 42(4): 844–856.

Berk, R. H.; and Jones, D. H. 1979. Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(1): 47–59.

Bollobás, B.; and Erdos, P. 1976. Cliques in random graphs. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 80, 419–427. Cambridge University Press.

Cadena, J.; Chen, F.; and Vullikanti, A. 2018. Graph anomaly detection based on Steiner connectivity and density. *Proceedings of the IEEE*, 106(5): 829–845.

Cadena, J.; Chen, F.; and Vullikanti, A. 2019. Near-optimal and practical algorithms for graph scan statistics with connectivity constraints. *ACM Transactions on Knowledge Discovery from Data*, 13(2): 1–33.

Chen, F.; and Neill, D. B. 2014. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1166–1175. ACM.

Chitra, U.; Ding, K.; Lee, J. C. H.; and Raphael, B. J. 2021. Quantifying and Reducing Bias in Maximum Likelihood Estimation of Structured Anomalies. In *Proc. 38th Intl. Conf. on Machine Learning, PMLR 139*, 1908–1919.

Donoho, D.; and Jin, J. 2004. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3): 962–994.

Duczmal, L.; and Assuncao, R. 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis*, 45(2): 269–286.

Eicker, F. 1979. The asymptotic distribution of the suprema of the standardized empirical processes. *The Annals of Statistics*, 116–138.

Erdős, P.; and Rényi, A. 1960. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(1): 17–60.

Glaz, J.; Pozdnyakov, V.; and Wallenstein, S. 2009. *Scan statistics: Methods and applications*. Springer Science & Business Media.

Kulldorff, M. 1997. A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6): 1481–1496.

Leskovec, J.; Lang, K. J.; Dasgupta, A.; and Mahoney, M. W. 2008. Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th international conference on World Wide Web*, 695–704. ACM.

Leskovec, J.; Lang, K. J.; Dasgupta, A.; and Mahoney, M. W. 2009. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1): 29–123.

Maehara, T.; Akiba, T.; Iwata, Y.; and Kawarabayashi, K.-i. 2014. Computing Personalized PageRank Quickly by Exploiting Graph Structures. *Proceedings of the VLDB Endowment*, 7(12): 1023–1034.

Massey Jr, F. J. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253): 68–78.

McFowland III, E.; Speakman, S.; and Neill, D. B. 2013. Fast generalized subset scan for anomalous pattern detection. *Journal of Machine Learning Research*, 14: 1533–1561.

Neill, D. B. 2009. Expectation-based scan statistics for monitoring spatial time series data. *International Journal of Forecasting*, 25(3): 498–517.

Neill, D. B. 2012. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2): 337–360.

Neill, D. B.; and Lingwall, J. 2007. A nonparametric scan statistic for multivariate disease surveillance. *Advances in Disease Surveillance*, 4: 106.

Qian, J.; Saligrama, V.; and Chen, Y. 2014. Connected subgraph detection. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, 796–804. PMLR.

Reyna, M. A.; Chitra, U.; Elyanow, R.; and Raphael, B. J. 2021. NetMix: A Network-Structured Mixture Model for Reduced-Bias Estimation of Altered Subnetworks. *Journal of Computational Biology*, 28(5): 469–484.

Rombach, M. P.; Porter, M. A.; Fowler, J. H.; and Mucha, P. J. 2014. Core-periphery structure in networks. *SIAM Journal on Applied Mathematics*, 74(1): 167–190.

Rozenshtein, P.; Anagnostopoulos, A.; Gionis, A.; and Tatti, N. 2014. Event detection in activity networks. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1176–1185. ACM.

Sharpnack, J.; Rinaldo, A.; and Singh, A. 2015. Detecting anomalous activity on networks with the graph Fourier scan statistic. *IEEE Transactions on Signal Processing*, 64(2): 364–379.

Speakman, S.; McFowland III, E.; and Neill, D. B. 2015. Scalable detection of anomalous patterns with connectivity constraints. *Journal of Computational and Graphical Statistics*, 24(4): 1014–1033.

Speakman, S.; Zhang, Y.; and Neill, D. B. 2013. Dynamic pattern detection with temporal consistency and connectivity constraints. In *Proceedings of the 13th IEEE International Conference on Data Mining*, 697–706. IEEE.

Wu, N.; Chen, F.; Li, J.; Zhou, B.; and Ramakrishnan, N. 2016. Efficient nonparametric subgraph detection using tree shaped priors. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, volume 30. The AAAI Press.