Technical Appendix

In this section, we provide theoretical and empirical justification for reducing the edge density of the graph, and specifically, for our proposed approach of using the $\alpha_{\max}$ parameter of NPHGS as a threshold for edge removal. For example, if $\alpha_{\max} = 0.15$, we would include a directed edge from a prescriber node p to a dispenser node d if at least 15% of p's prescriptions were filled by d, while the reverse edge from d to p would be included if at least 15% of the prescriptions filled by d were prescribed by p. Removing directed edges that correspond to a low proportion of a provider's prescription volume has several intuitive advantages: first, it avoids including edges that correspond to rare or chance occurrences, thus including only those edges that represent a real and sustained connection between prescriber and dispenser. Second, such low-volume edges are unlikely to correspond to illicit collusion between prescriber and dispenser. Third, and perhaps most importantly, we argue that removing low-volume edges increases the power and accuracy of the NPHGS method for detecting true patterns of high-risk and potentially illicit prescribing, avoiding false positives that could result from overfitting in cases when no signal is present.

Recent work by Wang, Neill, and Chen (2022) demonstrates that the nonparametric scan statistic is often poorly calibrated, picking out large, high-scoring subsets of graph nodes even under the null hypothesis H_0 that all nodes' empirical p-values are uniformly distributed on [0,1]. Under H_0, for a sufficiently large value of the significance threshold α, and a sufficiently connected graph structure, the scan is likely to find a large, high-scoring connected subgraph that includes almost all of the significant nodes in the graph. This reduces the detection power of the scan (since this false positive subgraph might outscore a true, smaller pattern of interest, and thus be detected instead of that pattern) as well as its precision (since the detected subgraph might include the true pattern but also many other nodes that are significant at level α just by chance). Wang, Neill, and Chen (2022) propose a new approach using a randomization test to re-calibrate the nonparametric scan statistic, but their approach is computationally expensive and thus infeasible for massive graphs like the ones we consider here.

As an alternative, we propose the much simpler approach of reducing the graph's edge density sufficiently so that, with high probability, a large, high-scoring false positive subgraph will no longer be detected. Our approach, and the specific choice of $\alpha_{\max}$ as a threshold, builds on recent results in percolation theory (Krivelevich 2016). We note that NPHGS, unlike the more complex scan approach considered by Wang, Neill, and Chen (2022), performs a greedy subgraph expansion search, starting from each of a chosen set of seed nodes $\{s_{(i)}\}$ (here, the five nodes of each type with individually most significant p-values), for a range of α values between 0 and $\alpha_{\max}$. With only a single p-value per node, which can either be significant or not significant at level α, this greedy search corresponds to finding the connected components containing each seed node $s_{(i)}$ in the induced graph G_α, where G_α consists of all and only those nodes with p-values less than α. The largest connected component found through this procedure will be the highest-scoring subgraph S^*_α for the given value of α, with corresponding BJ score $F_\alpha(S^*_\alpha) = |S^*_\alpha| \times KL(1, \alpha) = |S^*_\alpha| \log\left(\frac{1}{\alpha}\right)$. Then the overall maximum subgraph score returned by NPHGS is $F^* = \max_{0<\alpha\leq\alpha_{\max}} F_\alpha(S^*_\alpha)$.

We can now use percolation theory to estimate the size $|S^*_\alpha|$ as a function of the number of graph nodes n, the significance threshold α, and the connectivity of the graph.

In particular, by requiring at least proportion $\alpha_{\max}$ of a provider's prescriptions as a threshold to include a directed edge of the graph, we guarantee that every graph node will have degree no larger than $\left\lfloor \frac{1}{\alpha_{\max}} \right\rfloor$. We assume that this quantity is strictly less than $\frac{1}{\alpha_{\max}}$, since it is unlikely in practice that a provider will have exactly $\alpha_{\max}$ proportion of their prescriptions correspond to each of $\frac{1}{\alpha_{\max}}$ neighbors. We can then apply Theorem 1 of Krivelevich (2016), which states:

Theorem 1. *Let $\epsilon > 0$. Let $G = (V, E)$ be a graph of maximum degree at most d on n vertices. Form a random subset $R \subseteq V$ by including each vertex $v \in V$ in R independently and with probability p. If $p = \frac{1-\epsilon}{d}$, then* w.h.p.*, all connected components of the induced subgraph $G[R]$ are of size less than $\frac{4}{\epsilon^2} \log n$.*

In our case, $d = \left\lfloor \frac{1}{\alpha_{\max}} \right\rfloor$, and under the null hypothesis H_0, the induced subgraph $G[R] = G_\alpha$ is formed by including each vertex independently with probability $p = \alpha$, for $\alpha \leq \alpha_{\max}$. Then Theorem 1 implies that, *w.h.p.*, $|S^*_\alpha| < \frac{4}{\epsilon^2} \log n$, where $\epsilon = 1 - \alpha \left\lfloor \frac{1}{\alpha_{\max}} \right\rfloor > 0$, and thus the corresponding score $F_\alpha(S^*_\alpha)$ is $O(\log n)$ under H_0.

If, on the other hand, we did not restrict the edge density of the graph, or used a less restrictive threshold, then it is likely that the score $F_\alpha(S^*_\alpha)$ under H_0 would be much larger. Krivelevich (2016), Theorem 3, implies that if the node degrees d are greater than $\frac{1}{\alpha}$, under mild pseudorandomness conditions on the spectral ratio of the graph, then *w.h.p.*, there exists a connected component of G_α of size $\Theta(n)$. Since it is very likely that at least one of the seed nodes is part of this giant component, we expect that the score $F_\alpha(S^*_\alpha)$ will be large, $\Theta(n)$, in this case.

To support this reasoning, we perform experiments with a semi-synthetic dataset based on our first quarter of PDMP data, but with all p-values generated under the null hypothesis (i.e., uniform on [0,1]). We consider fixed α values of 0.01, 0.02, ..., 0.25 for graphs with thresholds of 1%, 1.5%, 2%, 5%, 10%, and 15% of prescriptions for an edge to be included, along with the original graph without edge restriction (i.e., there is an bidirectional edge between a prescriber and a dispenser if they have at least one prescription in common). We then compute the maximum subgraph score $F_\alpha(S^*_\alpha)$ for each combination of graph and α, and average these scores over ten runs, with different sets of randomly generated p-values for each run.

As shown in Figure 3, the score $F_\alpha(S^*_\alpha)$ remains small and relatively constant for α values below the edge threshold, but increases rapidly for α values that are sufficiently large compared to the threshold. We note that our edge threshold is conservative in practice: as shown in Table 6, the average node out-degree for the prescriber-dispenser graph, using

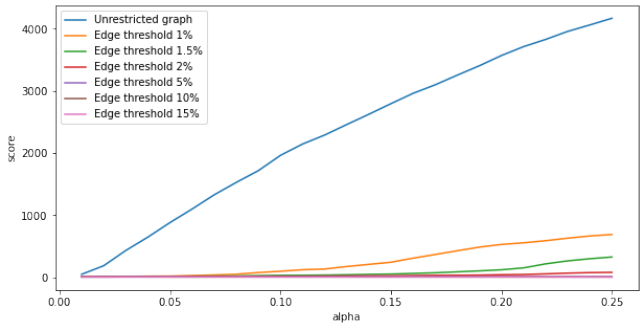

Figure 3: Score $F_\alpha(S^*_\alpha)$ as a function of α, for various edge thresholds and the original graph without edge restriction.

Edge threshold	Average out-degree	Maximum out-degree	Theoretical upper bound
1%	3.0	86	100
1.5%	2.6	52	66
2%	2.3	39	50
5%	1.6	19	20
10%	1.3	10	10
15%	1.1	6	6

Table 6: Comparison of average and maximum node out-degrees with the theoretical upper bound on maximum out-degree, for various edge thresholds. In total, the unrestricted graph has 71,048 nodes and 522,756 directed edges.

$\alpha_{\max}$ as the edge threshold, is much smaller than the theoretical upper bound on maximum node out-degree, $\left\lfloor \frac{1}{\alpha_{\max}} \right\rfloor$. As a result, for higher edge thresholds and the resulting sparse graphs, the score remains small even for α values exceeding the edge threshold by a substantial margin.

Since the NPHGS score is optimized over $\alpha \leq \alpha_{\max}$, these results demonstrate both theoretically and empirically that the score will remain small under the null hypothesis for our choice of $\alpha_{\max}$ as the edge threshold. Our evaluation experiments and case studies in the main paper confirm that this methodological contribution enables anomalous clusters of high-risk and potentially illicit prescribing behavior to be found in practice, thus providing state law enforcement and regulatory agencies with a new tool in their fight against the serious public health threats of opioid misuse and overdose.