Calibrated Nonparametric Scan Statistics for Anomalous Pattern Detection in Graphs

Chunpai Wang, Daniel B. Neill, and Feng Chen

UNIVERSITY AT ALBANY
State University of New York

NYU

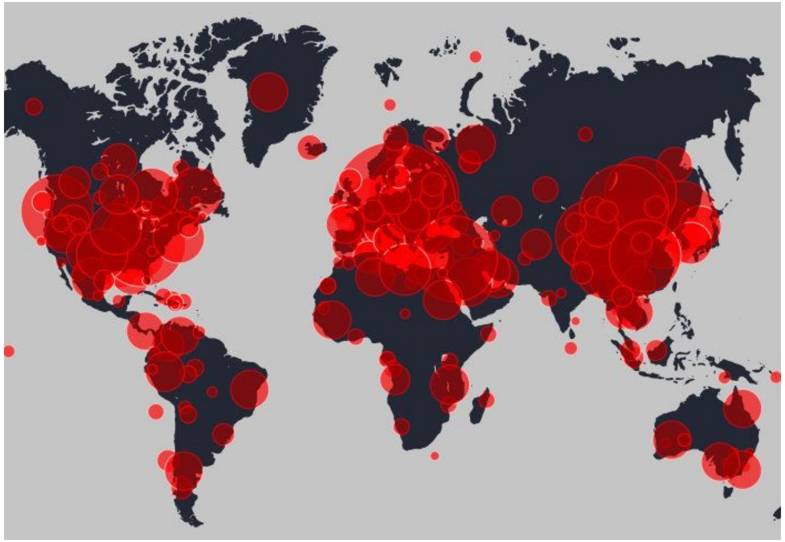
UT DALLAS
The University of Texas at Dallas

Outlines

- Introduction
- Parametric Scan Statistics
- Nonparametric Scan Statistics
- Limitations of Nonparametric Scan
- Motivating Example
- Calibrated Nonparametric Scan Statistics (CNSS)
    - An Efficient Approximate Algorithm
    - Low Bounds for the Expected Maximum Proportion of Significant Nodes
    - Core Tree Decomposition
- Experiments
- Case Studies
- Conclusions

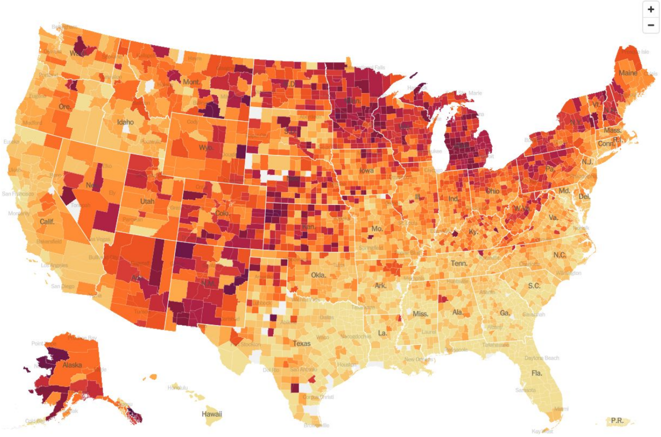
Introduction

- Detecting “hotspots” or anomalous patterns in graphs is an important but challenging problem.
- Disease outbreak detection, network intrusion detection, etc.
- Problem: anomalous connected subgraph detection.



Introduction

- Detecting “hotspots” or anomalous patterns in graphs is an important but challenging problem.
- Disease outbreak detection, network intrusion detection, etc.
- Problem: anomalous connected subgraph detection.
- Given a graph $\mathbb{G} = (\mathcal{V}, \mathcal{E})$,
  - each node $v_i \in \mathcal{V}$ is associated with a feature vector $\mathbf{x}_i \in \mathbb{R}^N$.
  - historical observation $\{\mathbf{x}_i^{(1)}, \cdots, \mathbf{x}_i^{(T)}\}$ for each $v_i \in \mathcal{V}$.
- Find a subgraph $\mathbb{G}_{\mathcal{S}} = (\mathcal{S}, \mathcal{E}_{\mathcal{S}})$ such that
  - $\mathcal{S} \subseteq \mathcal{V}$ and $\mathcal{E}_{\mathcal{S}} \subseteq \mathcal{E}$.
  - $\mathbb{G}_{\mathcal{S}}$ is connected in $\mathbb{G}$.
  - $\mathbb{G}_{\mathcal{S}}$ is anomalous.



Overview of Parametric Scan Statistics

- Parametric scan statistics:
  - likelihood ratio statistics of the hypothesis test.
  - $\mathcal{H}_0$: the $\mathbf{x}_i \in \mathbb{R}^N$ of nodes $\mathcal{S}$ within a candidate subgraph $\mathbb{G}_\mathcal{S}$ are generated by a parameterized *background* process.
  - $\mathcal{H}_1$: the $\mathbf{x}_i \in \mathbb{R}^N$ are generated by a different parameterized distribution (a localized anomalous process).
  - Kulldorff Scan Statistic (Kulldorff 1997).
  - Positive Elevated Mean (Qian, Saligrama, and Chen 2014).
  - Expectation-based Poisson and Gaussian (Neill 2009).
- Achieve high detection power across many spatio-temporal graph datasets.
- Limitations:
  - strong parametric model assumptions.
  - performance degrades when these models are incorrect.

Overview of Nonparametric Scan Statistics

- Nonparametric scan statistics (NPSSs):
  - likelihood ratio statistics of the *nonparametric* hypothesis test.
  - feature vector $\mathbf{x}_i \longrightarrow$ empirical p-value p_i based on $\{\mathbf{x}_i^{(1)}, \cdots, \mathbf{x}_i^{(T)}\}$.

$$p_i = \frac{1 + \sum_{t=1\ldots T} \mathbf{1}\{x_i^{(t)} \geq x_i\}}{1 + T}$$

  - $\mathcal{H}_0$: $p_i \sim \mathrm{Uniform}(0, 1)$ for each node $v_i \in \mathcal{S}$ within a candidate connected subgraph $\mathbb{G}_{\mathcal{S}}$.
  - $\mathcal{H}_1$: the empirical p-values follow a different distribution.
    - different distributions $\longrightarrow$ different NPSSs are formulated.
    - piecewise constant distribution $\longrightarrow$ Berk Jones (Berk and Jones 1979)
    - Higher Criticism (Donoho and Jin 2004)
    - Kolmogorov-Smirnov (Massey Jr 1951)
    - Anderson-Darling (Eicker 1979)

Overview of Nonparametric Scan Statistics

- NPSS-based anomalous pattern (subgraph) detection:
  - $\mathbb{M} = \{\mathcal{S} \mid \mathcal{S} \subseteq \mathcal{V}, \mathbb{G}_{\mathcal{S}} \text{ is connected in } \mathbb{G}\}$.
  - connected subgraph optimization problem:

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

  - $F(\mathcal{S}) := \max_{\alpha \le \alpha_{\max}} \Phi\left(\alpha, N_\alpha(\mathcal{S}), N(\mathcal{S})\right)$ refers to the general form of NPSS.
  - $N_\alpha(\mathcal{S}) = \sum_{v \in \mathcal{S}} \mathbf{1}\{p(v) \le \alpha\}$, and $N(\mathcal{S}) = \sum_{v \in \mathcal{S}} 1$.
  - under the null hypothesis, $\mathbb{E}[N_\alpha(\mathcal{S})] = \alpha N(\mathcal{S})$.
  - $\Phi\left(\alpha, N_\alpha(\mathcal{S}), N(\mathcal{S})\right)$: compares observed $N_\alpha(S)$ with $\mathbb{E}[N_\alpha(S)]$.
  - $0 < \alpha \le \alpha_{\max} < 1$, and $\alpha_{\max}$ is a constant.
  - in practice, $\alpha \in \mathcal{L} = \{0.001, \cdots, 0.009, 0.01, \cdots, 0.09\}$.

Overview of Nonparametric Scan Statistics

- Berk-Jones:

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

  - log-likelihood ratio statistic of the *nonparametric* hypothesis test.
  - $\mathcal{H}_0$: the empirical p-values follow the $\mathrm{Uniform}[0,1]$.
  - $\mathcal{H}_1$: the empirical p-values follow a piecewise constant distribution.
  - $\mathrm{KL}(a, b) = a \log(a/b) + (1-a)\log((1-a)/(1-b))$

- Higher Criticism: $\Phi_{HC}\left(\alpha, N_\alpha(\mathcal{S}), N(\mathcal{S})\right) = \frac{N_\alpha(\mathcal{S}) - \alpha N(\mathcal{S})}{\sqrt{N(\mathcal{S})\alpha(1-\alpha)}}$

- Kolmogorov-Smirnov: $\Phi_{KS}\left(\alpha, N_\alpha(\mathcal{S}), N(\mathcal{S})\right) = \sqrt{N(\mathcal{S})} \cdot \left(\frac{N_\alpha(\mathcal{S})}{N(\mathcal{S})} - \alpha\right)$

- **Assumption**: under $\mathcal{H}_0$, $\mathbb{E}[N_\alpha(\mathcal{S})/N(\mathcal{S})] = \alpha$.

Limitations of Nonparametric Scan

- Assumption of NPSSs: under $\mathcal{H}_0$, $\mathbb{E}[N_\alpha(\mathcal{S})/N(\mathcal{S})] = \alpha$.
- For anomalous pattern (subgraph) detection:
  - the assumption is true for a randomly selected connected subset.
  - but not for connected subsets that are identified by maximizing the NPSS score.

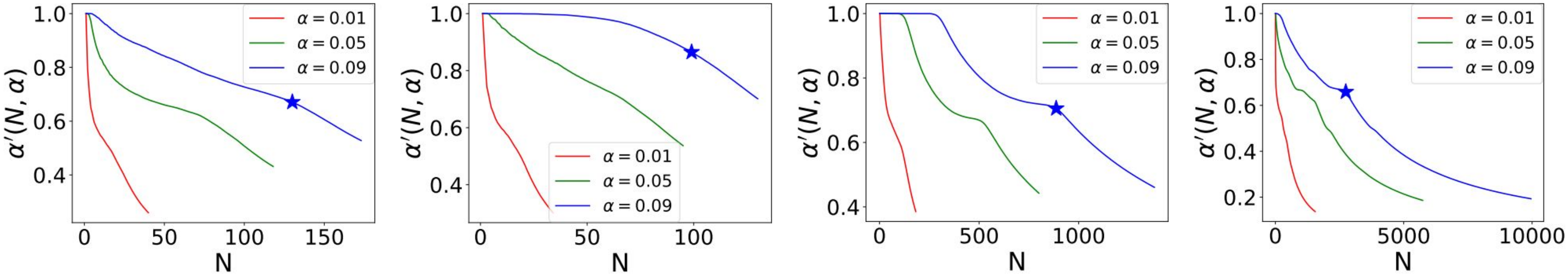
- **Miscalibration:**
  - expected maximum proportion of significant nodes for all connected subgraphs of a given size N:

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

  - we find $\alpha'(N,\alpha) \gg \alpha$.

Limitations of Nonparametric Scan

- Justification of $\alpha'(N, \alpha) \gg \alpha$
  - simulate p-values under $\mathcal{H}_0$ for 100 times on Erdos-Renyi and real graphs.
  - calculate the average α'
    - for each $N \in \{1, 2, \cdots, |\mathcal{V}|\}$ and $\alpha \in \{.01, .05, .09\}$



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

  - the starred point is the combination of N and α for which $N \times \mathrm{KI}(\alpha', \alpha)$ is maximized.
  - $\alpha'(N, \alpha)$ decreases with N but remains much higher than α.

Limitations of Nonparametric Scan

- Issues of $\alpha'(N, \alpha) \gg \alpha$:
  - even under $\mathcal{H}_0$ and there are no true subgraphs of interest, there exists subgraphs $\mathcal{S}$ with $N_\alpha(\mathcal{S}) \gg \alpha N(\mathcal{S})$, and thus very high NPSS scores.
  - These large scores under the null result in *reduced detection power*, since the NPSS scores of the true anomalous subgraph must exceed a larger threshold to be considered significant.
  - NPSS will *biased toward detecting clusters at larger α threshold*, even if the true signal is for a much smaller α.
  - 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.

Motivating Example

- Consider a single instantiation of WikiVote graph $(|\mathcal{V}| = 7066)$ under $\mathcal{H}_1(\mathcal{S})$.
- True subgraph $\mathbb{G}_{\mathcal{S}}$:
  - generated using a random walk with $|\mathcal{S}| = 100$.
  - 75% of the p-values in $\mathcal{S}$ are significant at $\alpha = 0.01$.
  - $BJ = N(\mathcal{S})\mathrm{KL}(\frac{N_\alpha(\mathcal{S})}{N(\mathcal{S})}, \alpha) = 100\ \mathrm{KL}(0.75, 0.01) \approx 289$.
- Another subgraph $\mathbb{G}_{\mathcal{Z}}$ could have an even higher score, corresponding to a high significance threshold α:
  - consider $\alpha = 0.09$
  - uncalibrated BJ picks out a subgraph with $N(\mathcal{Z}) = 900$ and $N_\alpha(\mathcal{Z}) = 670$
  - $BJ = N(\mathcal{Z})\mathrm{KL}(\frac{N_\alpha(\mathcal{Z})}{N(\mathcal{Z})}, \alpha) = 900\ \mathrm{KL}(0.744, 0.09) \approx 1100$.
  - Precision=0.08, Recall=0.75, F-score=0.15
- Uncalibrated BJ:
  - is biased toward detecting clusters at larger α threshold and identifies overly large cluster, resulting in reduced precision and poor detection.

Calibrated Nonparametric Scan Statistics (CNSS)

- CNSS:
  - $F(\mathcal{S}) := \max_{\alpha \le \alpha_{\max}} \Phi\left(\alpha, N_\alpha(\mathcal{S}), N(\mathcal{S})\right)$
  - replace the threshold reference $(\mathbb{E}[N_\alpha(\mathcal{S})/N(\mathcal{S})] = \alpha)$ with

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

- Calibrated Berk-Jones (CBJ):

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

Calibrated Nonparametric Scan Statistics (CNSS)

- Previous example on WikiVote:
  - for true subgraph $\mathbb{G}_{\mathcal{S}}$, $BJ = N(\mathcal{S})\mathrm{KL}(\frac{N_\alpha(\mathcal{S})}{N(\mathcal{S})}, \alpha) = 100\ \mathrm{KL}(0.75, 0.01) \approx 289$
  - for a candidate $\mathbb{G}_{\mathcal{Z}}$, $BJ = N(\mathcal{Z})\mathrm{KL}(\frac{N_\alpha(\mathcal{Z})}{N(\mathcal{Z})}, \alpha) = 900\ \mathrm{KL}(0.744, 0.09) \approx 1100$

Calibrated Nonparametric Scan Statistics (CNSS)

- Previous example on WikiVote:
  - for true subgraph $\mathbb{G}_{\mathcal{S}}$, $BJ = N(\mathcal{S})\mathrm{KL}(\frac{N_\alpha(\mathcal{S})}{N(\mathcal{S})}, \alpha) = 100\ \mathrm{KL}(0.75, 0.01) \approx 289$
  - for a candidate $\mathbb{G}_{\mathcal{Z}}$, $BJ = N(\mathcal{Z})\mathrm{KL}(\frac{N_\alpha(\mathcal{Z})}{N(\mathcal{Z})}, \alpha) = 900\ \mathrm{KL}(0.744, 0.09) \approx 1100$
  - we found $\alpha'(900, 0.09) = 0.699$ then

    $$CBJ = N(\mathcal{Z})\mathrm{KL}(\frac{N_\alpha(\mathcal{Z})}{N(\mathcal{Z})}, \alpha'(N, \alpha)) = 900\ \mathrm{KL}(0.744, 0.699) = 4.47$$
  - allow a subgraph $\mathbb{G}_{\mathcal{W}}$ closer to the true subgraph to be found instead with

    $N(\mathcal{W}) = 202, \frac{N_\alpha(\mathcal{W})}{N(\mathcal{W})} = 0.733 \text{ at } \alpha = 0.01, \alpha'(202, 0.01) = 0.347, \text{ and } CBJ = 62.26$

    Precision=0.72, Recall=0.69, and F-score=0.70.

Calibrated Nonparametric Scan Statistics (CNSS):

An Efficient Approximate Algorithm

- How to compute $\alpha'(N,\alpha) = \frac{\mathbb{E}\left[\max_{\mathcal{S}\in M, |\mathcal{S}|=N} N_\alpha(\mathcal{S})\right]}{N}$ for each $N$and α ?
  - possible solution: run PCST for each N and α.
  - time complexity is $\mathcal{O}(|\mathcal{V}|^3 \log|\mathcal{V}|)$.

Calibrated Nonparametric Scan Statistics (CNSS):

An Efficient Approximate Algorithm

- How to compute $\alpha'(N,\alpha) = \frac{\mathbb{E}[\max_{\mathcal{S}\in M, |\mathcal{S}|=N} N_\alpha(\mathcal{S})]}{N}$ for each N and α ?
  - possible solution: run PCST for each N and α.
  - time complexity is $\mathcal{O}(|\mathcal{V}|^3 \log|\mathcal{V}|)$.
- Our algorithm (randomization test on an efficient approximate algorithm)
  - randomization test to estimate $\mathbb{E}[\max_{\mathcal{S}\in\mathbb{M}, |\mathcal{S}|=N} N_\alpha(\mathcal{S})]$
    - create *K* replica of datasets under $\mathcal{H}_0$, with p-values redrawn uniformly at random from [0, 1].
  - apply an efficient algorithm to solve the constrained set optimization problem $\max_{\mathcal{S}\in\mathbb{M}, |\mathcal{S}|=N} N_\alpha(\mathcal{S})$ for each combination (N,α) .
    - for each value of α, approximates the maximum N_α for each $N \in \{1, \cdots, |\mathcal{V}|\}$ in a single, efficient run.
    - based on repeated merging of nodes with the highest proportion of significant p-values.

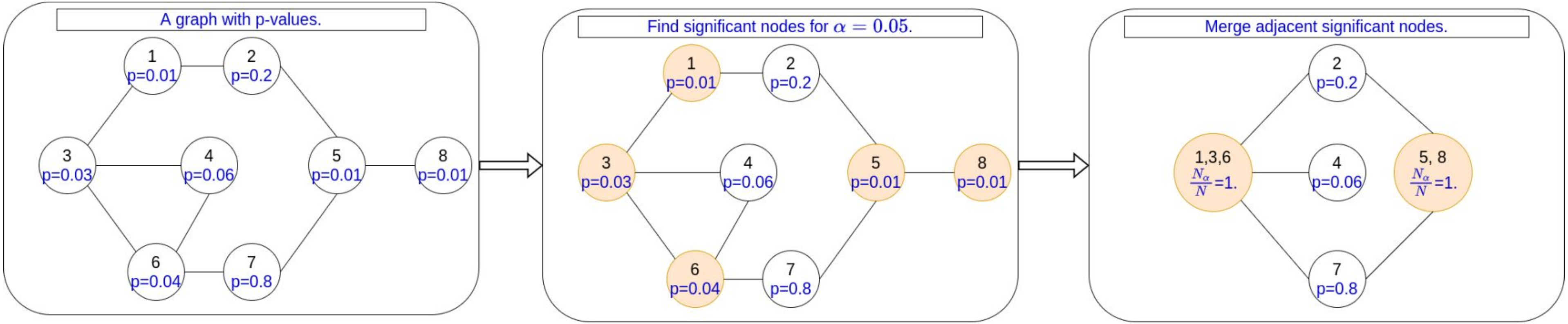
Calibrated Nonparametric Scan Statistics (CNSS):

An Efficient Approximate Algorithm

- Estimate $\max_{\mathcal{S}\in\mathbb{M},|\mathcal{S}|=N} N_\alpha(\mathcal{S})$ for $N \in \{1,\cdots,|\mathcal{V}|\}$ under each α :
  - given a graph with node-level p-values.
  - merge all adjacent significant nodes, and maintain a list $\mathcal{Z}$ of merged nodes sorted by significance ratio $N_\alpha(\mathcal{S})/N(\mathcal{S})$.
  - repeatedly choose the merged node with highest significance ratio and performance as one of the following three merge steps:
    - add an adjacent node which contains some or all significant p-values;
    - add an adjacent non-significant node that is also adjacent to at least one other significant node; or
    - 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.

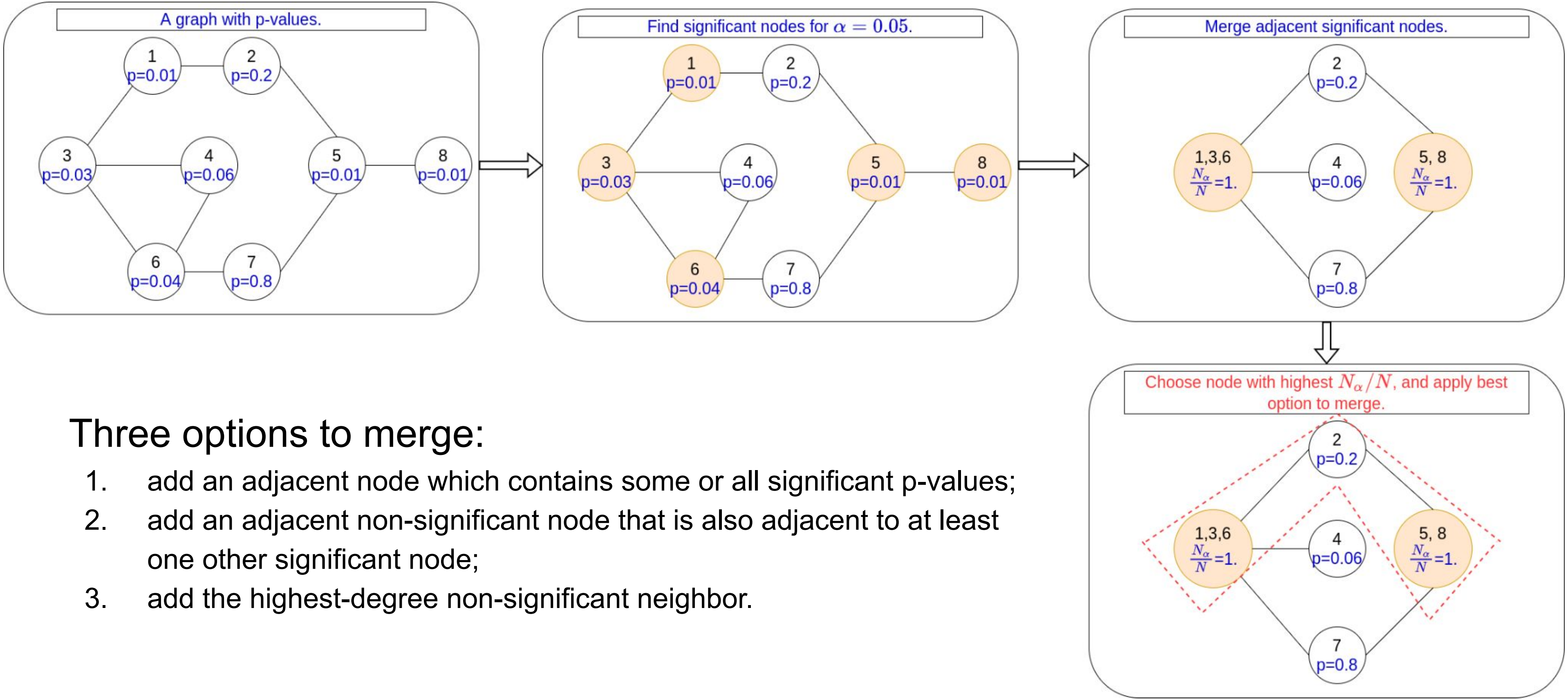
Calibrated Nonparametric Scan Statistics (CNSS):

Estimate $\max_{\mathcal{S}\in\mathbb{M},|\mathcal{S}|=N} N_\alpha(\mathcal{S})$



Calibrated Nonparametric Scan Statistics (CNSS):

Estimate $\max_{\mathcal{S}\in\mathbb{M},|\mathcal{S}|=N} N_\alpha(\mathcal{S})$



A graph with p-values.

Find significant nodes for $\alpha = 0.05$.

Merge adjacent significant nodes.

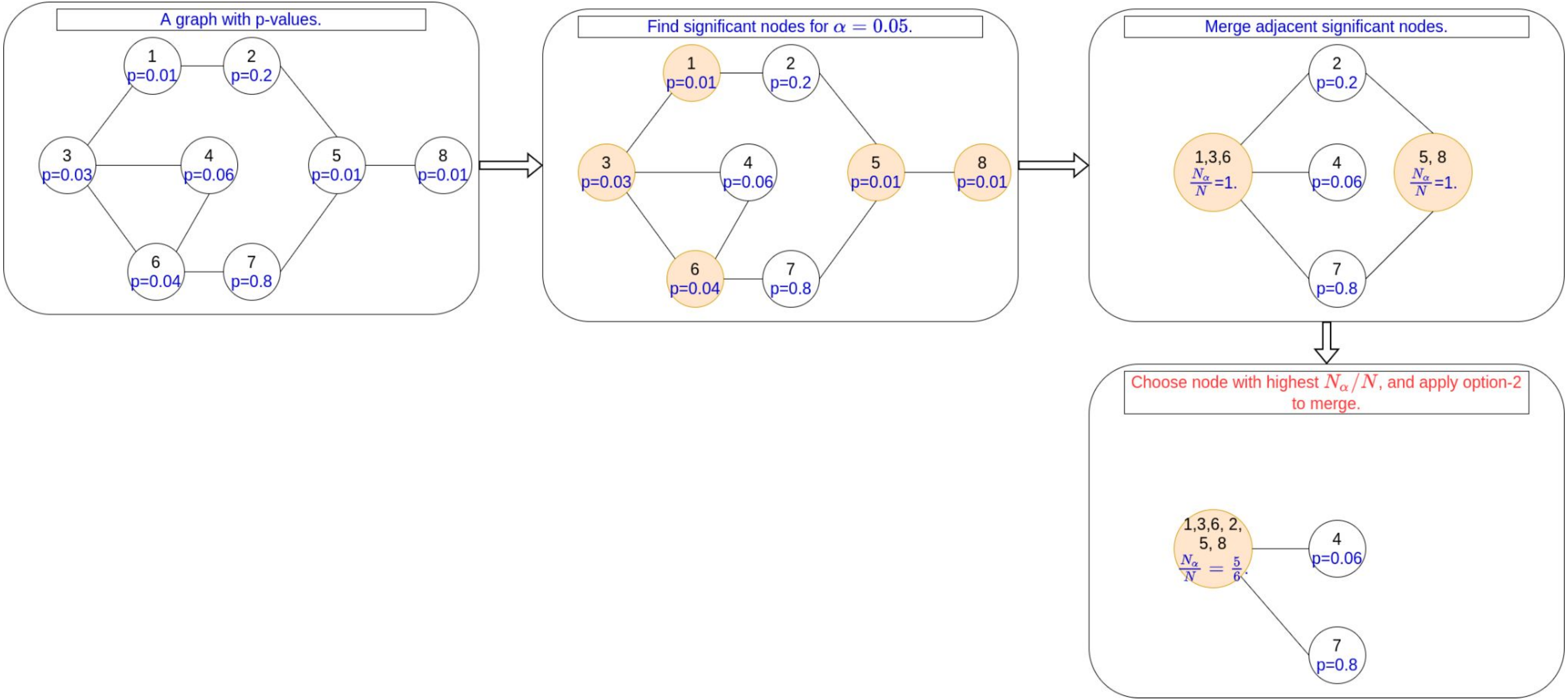
Choose node with highest N_α/N, and apply best option to merge.

Three options to merge:

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;
3. add the highest-degree non-significant neighbor.

Calibrated Nonparametric Scan Statistics (CNSS):

Estimate $\max_{\mathcal{S}\in\mathbb{M},|\mathcal{S}|=N} N_\alpha(\mathcal{S})$



A graph with p-values.

Find significant nodes for $\alpha = 0.05$.

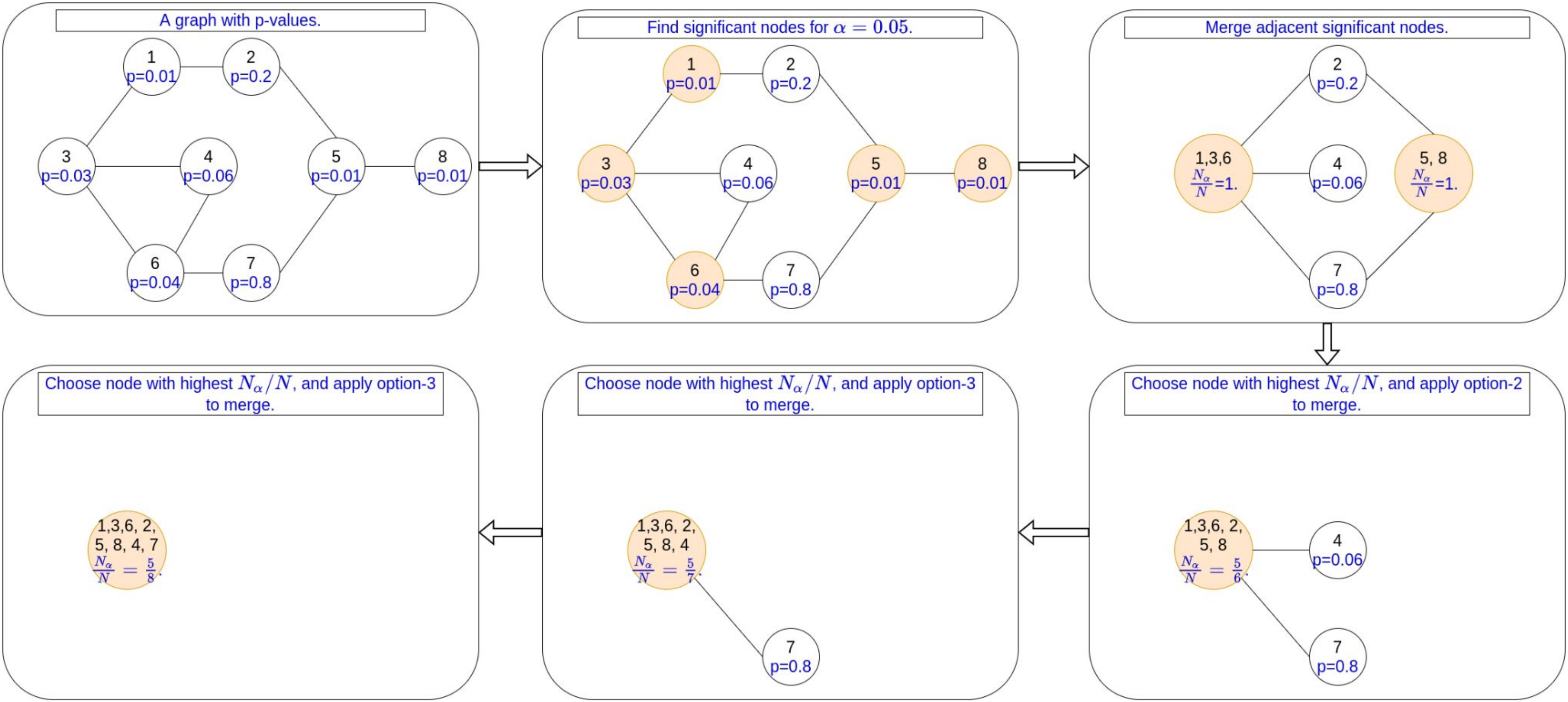
Merge adjacent significant nodes.

Choose node with highest N_α/N, and apply option-2 to merge.

1 p=0.01
2 p=0.2
3 p=0.03
4 p=0.06
5 p=0.01
8 p=0.01
6 p=0.04
7 p=0.8
1,3,6 $\frac{N_\alpha}{N}$=1.
5, 8 $\frac{N_\alpha}{N}$=1.
1,3,6, 2, 5, 8 $\frac{N_\alpha}{N} = \frac{5}{6}$.

Calibrated Nonparametric Scan Statistics (CNSS):

Estimate $\max_{\mathcal{S}\in\mathbb{M},|\mathcal{S}|=N} N_\alpha(\mathcal{S})$



A graph with p-values.

Find significant nodes for $\alpha = 0.05$.

Merge adjacent significant nodes.

Choose node with highest N_α/N, and apply option-2 to merge.

Choose node with highest N_α/N, and apply option-3 to merge.

Choose node with highest N_α/N, and apply option-3 to merge.

1 p=0.01
2 p=0.2
3 p=0.03
4 p=0.06
5 p=0.01
8 p=0.01
6 p=0.04
7 p=0.8
1,3,6 $\frac{N_\alpha}{N}$=1.
5, 8 $\frac{N_\alpha}{N}$=1.
1,3,6, 2, 5, 8 $\frac{N_\alpha}{N} = \frac{5}{6}$.
1,3,6, 2, 5, 8, 4 $\frac{N_\alpha}{N} = \frac{5}{7}$.
1,3,6, 2, 5, 8, 4, 7 $\frac{N_\alpha}{N} = \frac{5}{8}$.

Calibrated Nonparametric Scan Statistics (CNSS):

Estimate $\max_{\mathcal{S}\in\mathbb{M},|\mathcal{S}|=N} N_\alpha(\mathcal{S})$

- During this merge procedure for this graph with $\alpha = 0.05$, we get a list of (N, N_α):
  - when $N = 3$, the max $N_\alpha = 3$;
  - when $N = 6$, the max $N_\alpha = 5$;
  - when $N = 7$, the max $N_\alpha = 5$;
  - when $N = 8$, the max $N_\alpha = 5$;
- If we apply this to the target graph under $\mathcal{H}_0$:
  - apply interpolation to estimate the max N_α for *N*=4 and *N*=5.
  - still need to apply it for *K* replica of datasets with p-values redrawn uniformly at random from [0, 1] to compute $\mathbb{E}[\max_{\mathcal{S}\in\mathbb{M},|\mathcal{S}|=N} N_\alpha(\mathcal{S})]$ under various αs.
  - compute

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

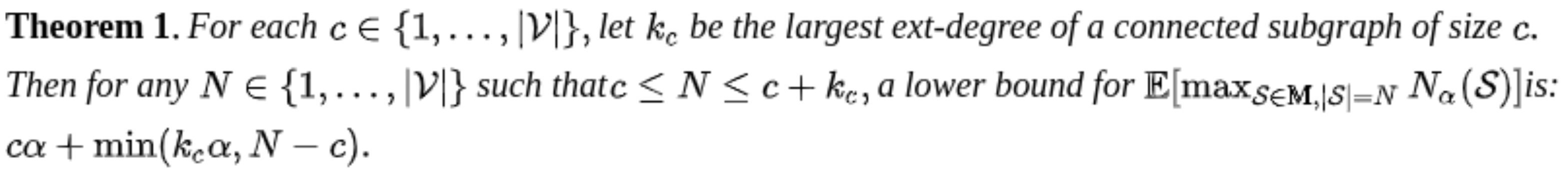
  for all $N \in \{1, \cdots, |\mathcal{V}|\}$ and α under consideration.
- If we apply this to the target graph for detection under $\mathcal{H}_1$:
  - the list of (N, N_α) corresponds to the list of candidate subgraphs (merged super-nodes).
  - still need to apply it under various αs.
  - for each candidate subgraph, we could compute:

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

Calibrated Nonparametric Scan Statistics (CNSS):

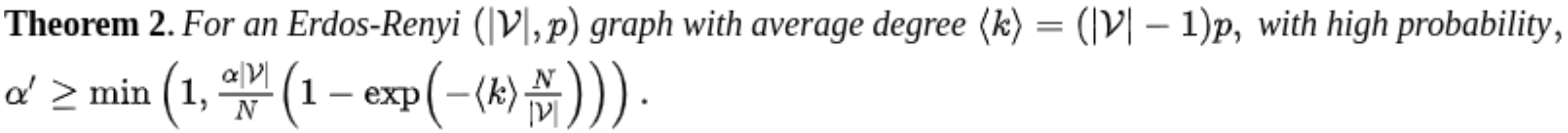
Low Bounds for the Expected Maximum Proportion of Significant Nodes

- Calibration with randomization test is time-consuming for large graphs.
- Two closed-form lower bounds of $\alpha'(N, \alpha)$:
  - lower bound α'_1 from network neighborhood analysis.



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

-
  - low bound α'_2 from percolation theory.

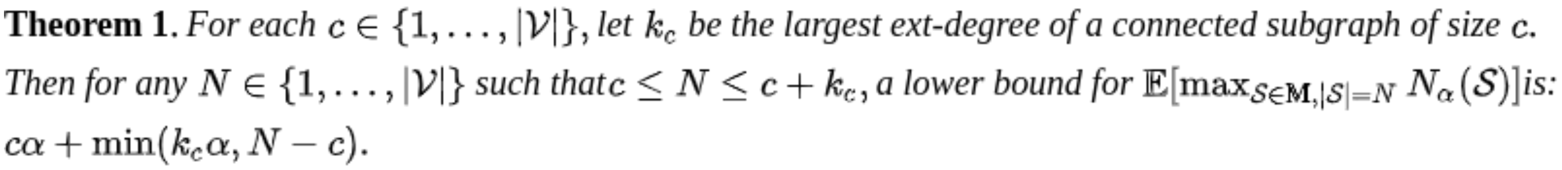


Theorem 2. *For an Erdos-Renyi* $(|\mathcal{V}|, p)$ *graph with average degree* $\langle k \rangle = (|\mathcal{V}| - 1)p$, *with high probability*, $\alpha' \geq \min\left(1, \frac{\alpha|\mathcal{V}|}{N}\left(1 - \exp\left(-\langle k \rangle \frac{N}{|\mathcal{V}|}\right)\right)\right)$.

Calibrated Nonparametric Scan Statistics (CNSS):

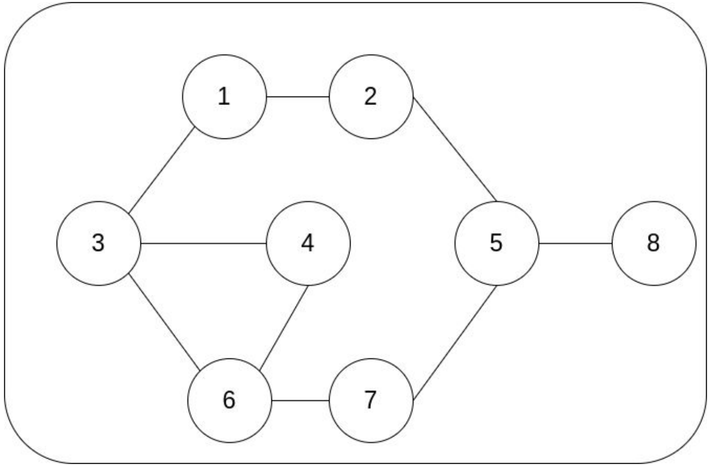
Low Bounds for the Expected Maximum Proportion of Significant Nodes

- lower bound α'_1 from network neighborhood analysis.



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

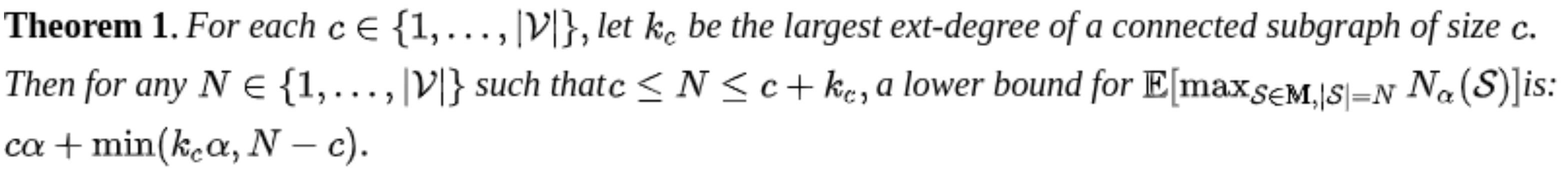
  - ○ Only consider the network structure without the p-values.
  - ○ For any N, what is the $\mathbb{E}[\max_{\mathcal{S} \in \mathbb{M}, |\mathcal{S}|=N} N_\alpha(\mathcal{S})]$ under $\mathcal{H}_0$?



Calibrated Nonparametric Scan Statistics (CNSS):

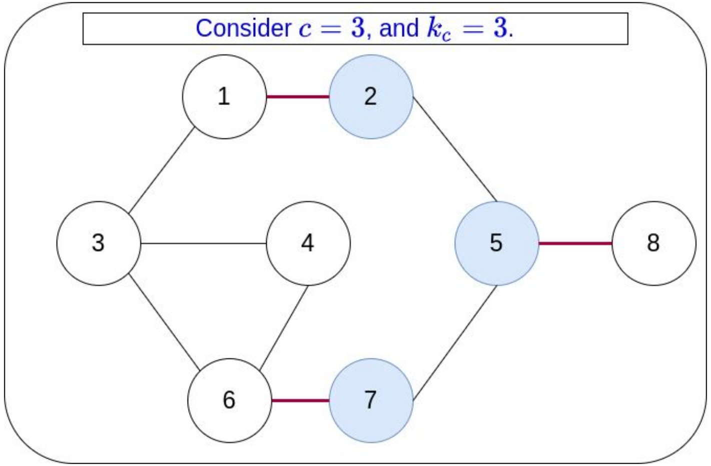
Low Bounds for the Expected Maximum Proportion of Significant Nodes

- lower bound α'_1 from network neighborhood analysis.



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

  - Only consider the network structure without the p-values.
  - For any *N*, what is the $\mathbb{E}[\max_{\mathcal{S} \in \mathbb{M}, |\mathcal{S}|=N} N_\alpha(\mathcal{S})]$ under $\mathcal{H}_0$?
  - For example, let $\alpha = 0.5$, with *S*=[2,5,7]
    - $c\alpha = 1.5$ and $k_c\alpha = 1.5$
    - for N=3, we have $\mathbb{E}[\max_{\mathcal{S} \in \mathbb{M}, |\mathcal{S}|=N} N_\alpha(\mathcal{S})] \geq 1.5$.
    - for N=4, we could add one significant node from the neighbor, thus $\mathbb{E}[\max_{\mathcal{S} \in \mathbb{M}, |\mathcal{S}|=N} N_\alpha(\mathcal{S})] \geq 2.5$.
    - for N=5, $\mathbb{E}[\max_{\mathcal{S} \in \mathbb{M}, |\mathcal{S}|=N} N_\alpha(\mathcal{S})] \geq 3$.
    - for N=6, $\mathbb{E}[\max_{\mathcal{S} \in \mathbb{M}, |\mathcal{S}|=N} N_\alpha(\mathcal{S})] \geq 3$.

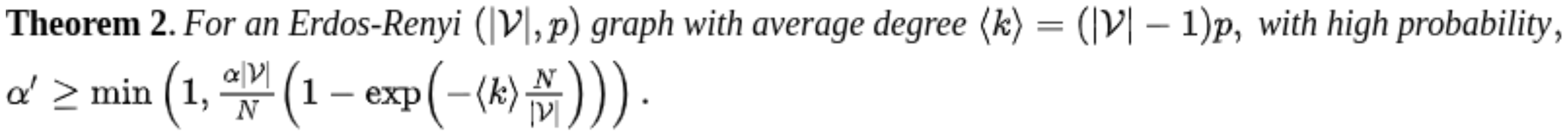


Consider $c = 3$, and $k_c = 3$.

Calibrated Nonparametric Scan Statistics (CNSS):

Low Bounds for the Expected Maximum Proportion of Significant Nodes

- low bound α_2' from percolation theory.



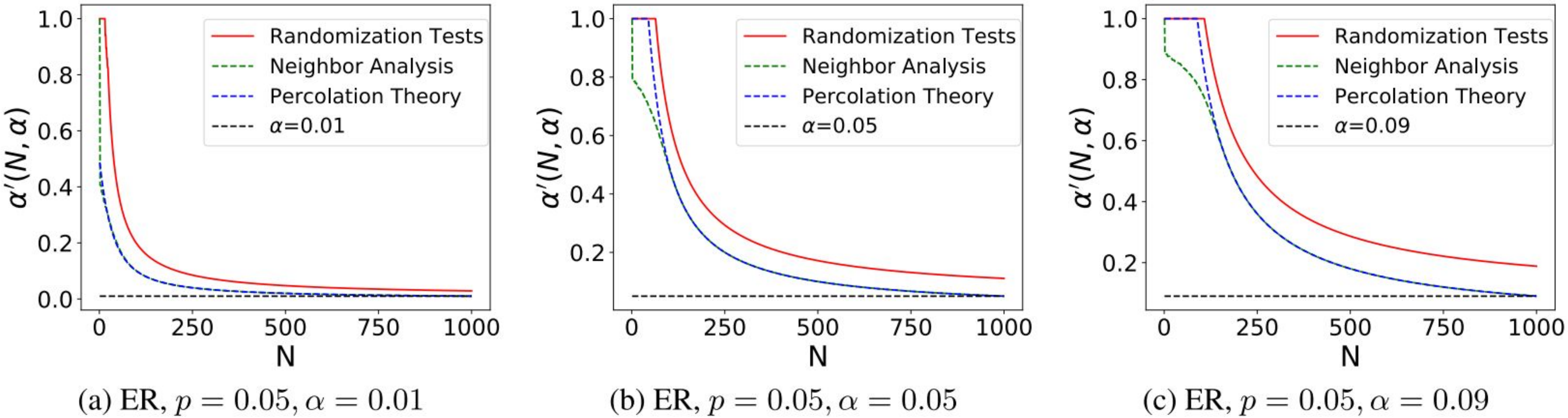
Theorem 2. *For an Erdos-Renyi* $(|\mathcal{V}|, p)$ *graph with average degree* $\langle k \rangle = (|\mathcal{V}| - 1)p$, *with high probability,*

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

  - 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 *S* consisting of only marked nodes, with |*S*| equal to a constant fraction P_∞ of |V|.
  - P_∞ is the solution to the equation $P_\infty = \rho(1 - \exp(-\langle k \rangle P_\infty))$.
  - “Marking” both significant and (as needed) insignificant nodes to reach the percolation threshold.
    - based on the number of marked significant nodes, we could use the formula to find out the number of insignificant nodes are needed to connected all significant nodes.

Calibrated Nonparametric Scan Statistics (CNSS):

Low Bounds for the Expected Maximum Proportion of Significant Nodes

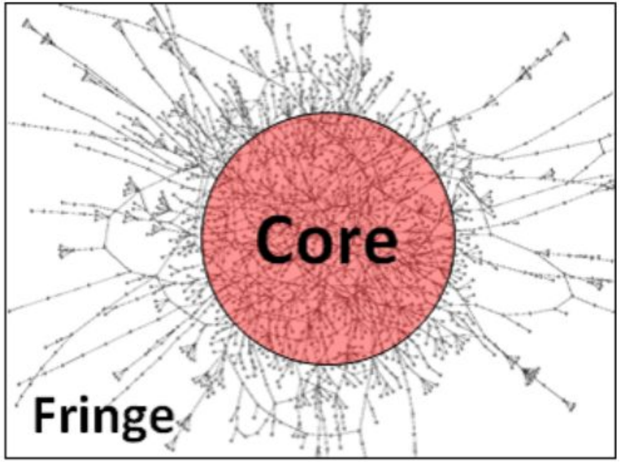


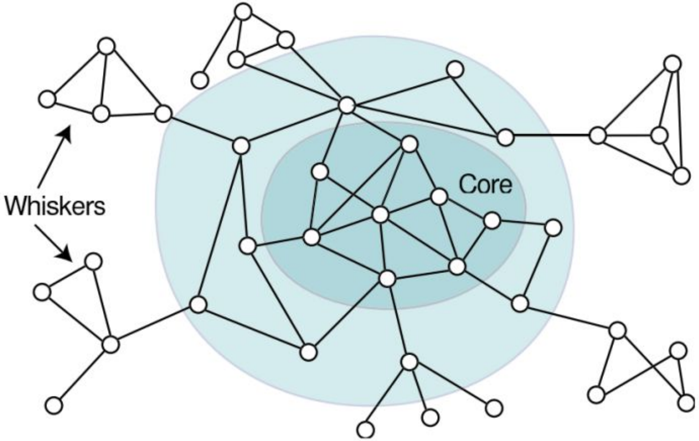
(a) ER, $p = 0.05, \alpha = 0.01$ (b) ER, $p = 0.05, \alpha = 0.05$ (c) ER, $p = 0.05, \alpha = 0.09$

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

Core Tree Decomposition

- Randomization test on large graphs is time-consuming.
  - Solution 1: lower bounds.
  - Solution 2: core tree decomposition.
- Core-whiskers (or core-periphery) structure commonly exists in many real-world networks:
  - the core keeps the general skeleton of the entire graph.



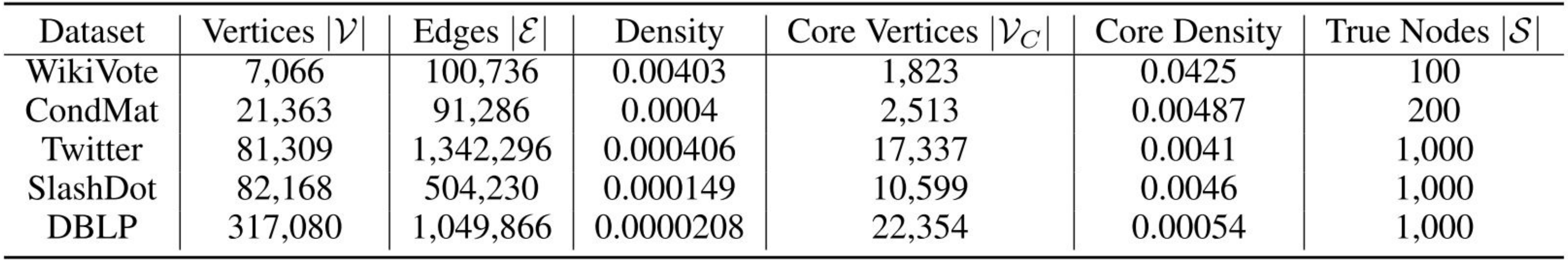


Core Tree Decomposition

- Randomization test on large graphs is time-consuming.
  - Solution 1: lower bounds.
  - Solution 2: core tree decomposition.
- Core-whiskers (or core-periphery) structure commonly exists in many real-world networks:
  - the core keeps the general skeleton of the entire graph.
- Core-tree decomposition:
  - decompose the graph into a small, dense core and a low-treewidth periphery.
  - compress significant tree-nodes into core.
  - apply randomization test or lower bounds on the core.

Experiments

- **Five Semi-synthetic Datasets:**

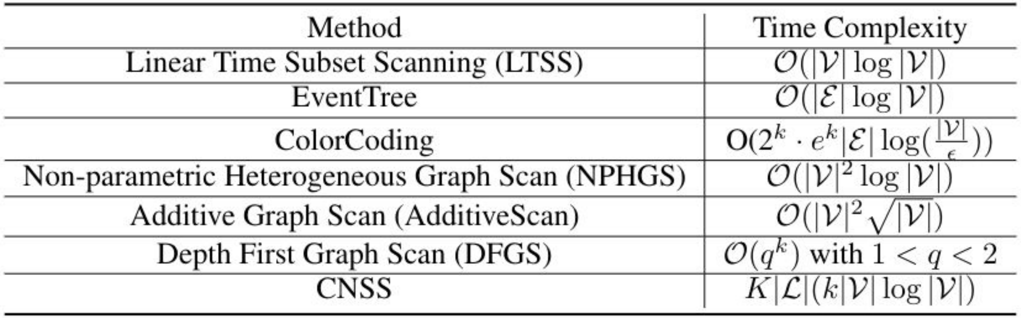


Dataset	Vertices $\lvert\mathcal{V}\rvert$	Edges $\lvert\mathcal{E}\rvert$	Density	Core Vertices $\lvert\mathcal{V}_C\rvert$	Core Density	True Nodes $\lvert\mathcal{S}\rvert$
WikiVote	7,066	100,736	0.00403	1,823	0.0425	100
CondMat	21,363	91,286	0.0004	2,513	0.00487	200
Twitter	81,309	1,342,296	0.000406	17,337	0.0041	1,000
SlashDot	82,168	504,230	0.000149	10,599	0.0046	1,000
DBLP	317,080	1,049,866	0.0000208	22,354	0.00054	1,000

  - leverage the graph structure of real networks.
  - simulate the true subgraph $\mathbb{G}_{\mathcal{S}}$ using a random walk.
    - assume Gaussian signal $x_i \sim \text{Normal}(\mu, 1)\ \forall\ v_i \in \mathcal{S}$
    - generate p-value $p_i = 1 - \text{CDF}(x_i)$
  - $p_i \sim \text{Uniform}\ [0, 1]\ \forall\ v_i \in \mathcal{V} \setminus \mathcal{S}$
  - use $\mu \in [1.5, 2, 3, 4, 5]$ for experiments.

Experiments

- **Baseline Methods:**



Method	Time Complexity
Linear Time Subset Scanning (LTSS)	$\mathcal{O}(\|\mathcal{V}\| \log \|\mathcal{V}\|)$
EventTree	$\mathcal{O}(\|\mathcal{E}\| \log \|\mathcal{V}\|)$
ColorCoding	$\mathrm{O}(2^k \cdot e^k \|\mathcal{E}\| \log(\frac{\|\mathcal{V}\|}{\epsilon}))$
Non-parametric Heterogeneous Graph Scan (NPHGS)	$\mathcal{O}(\|\mathcal{V}\|^2 \log \|\mathcal{V}\|)$
Additive Graph Scan (AdditiveScan)	$\mathcal{O}(\|\mathcal{V}\|^2 \sqrt{\|\mathcal{V}\|})$
Depth First Graph Scan (DFGS)	$\mathcal{O}(q^k)$ with $1 < q < 2$
CNSS	$K\|\mathcal{L}\|(k\|\mathcal{V}\| \log \|\mathcal{V}\|)$

- **Ablation Study:**
  - CNSS+NoCalib: removes the calibration from CNSS, performing the same search but using the original α instead of α' in the score function.
  - CNSS+LowerBound: replace the randomization test with the tightest lower bound $\max(\alpha'_1, \alpha'_2)$.
  - CNSS+CoreTree: integrates the core tree decomposition into CNSS.

Experiments

- **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?*

Experiments

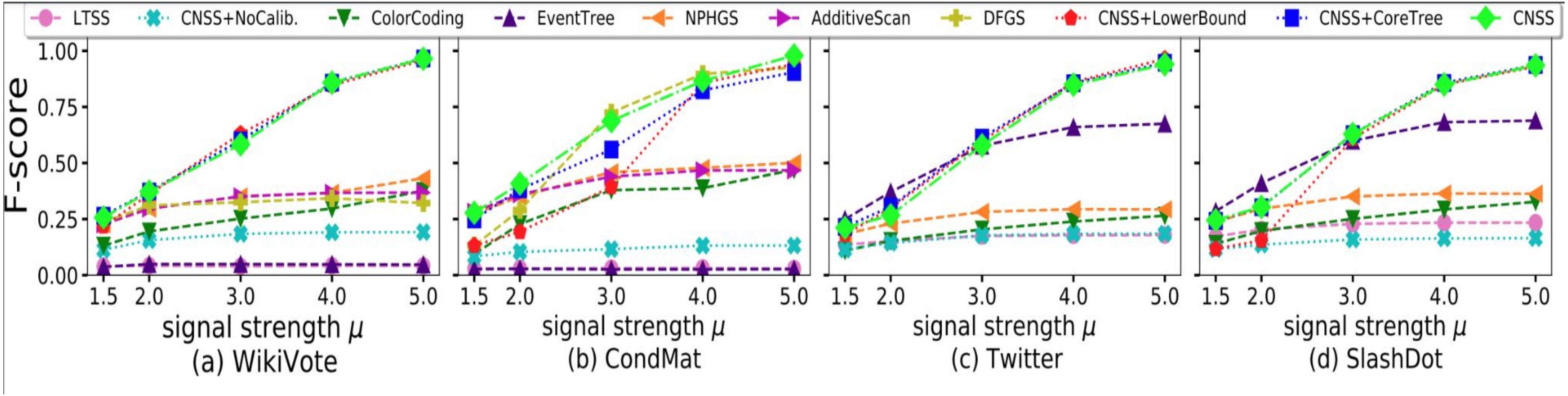
- **Evaluation Metrics**
  - Detection power: measures the ability to distinguish between graphs with or without an affected subgraph.
    - step 1. compute BJ score for each detected subgraph
    - step 2. for each alternative run, we conduct a hypothesis test with significance level $\alpha = 0.05$ by setting p-value as the proportion of null runs that have higher BJ score than the alternative run
    - step 3. compute the proportion of hypothesis tests (for each method, for each real-world graph, for each signal strength μ) that reject the null hypothesis.
  - Detection performance: $\text{Precision} = \frac{|\mathcal{R} \cap \mathcal{S}|}{|\mathcal{S}|}. \quad \text{Recall} = \frac{|\mathcal{R} \cap \mathcal{S}|}{|\mathcal{R}|}.$

$$\text{F-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

  - Run time.

Experiments

★ Detection performance results



(a) WikiVote

(b) CondMat

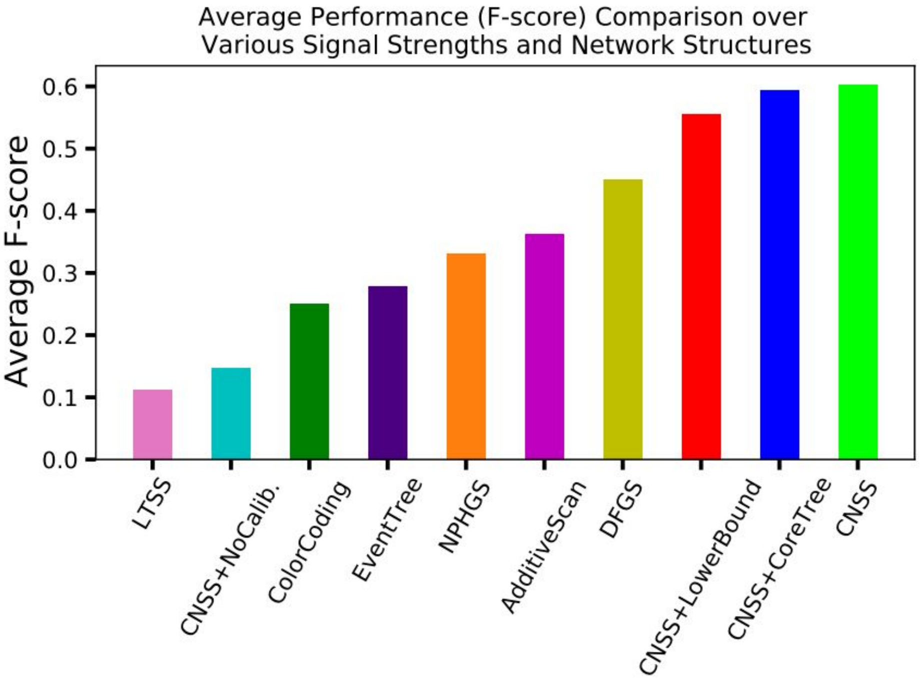
(c) Twitter

(d) SlashDot

- 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.
- CNSS+LowerBound < CNSS, but it’s better than baselines particularly for stronger signals.
- CNSS+CoreTree does not significantly change detection performance.

Experiments

★ Detection performance results



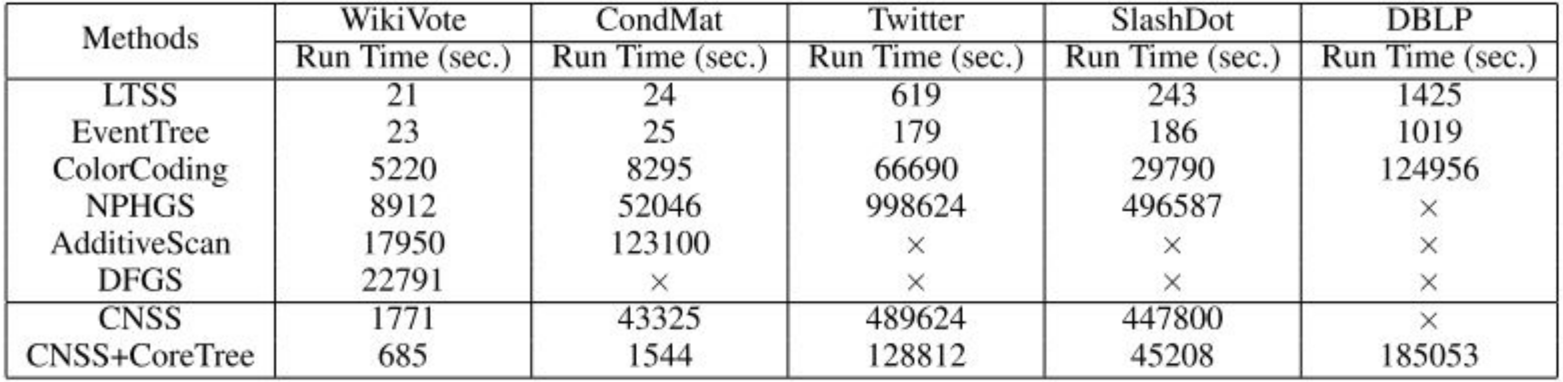
- Our proposed CNSS and its variants have higher average F-score over all networks and signal strength under consideration.

Experiments

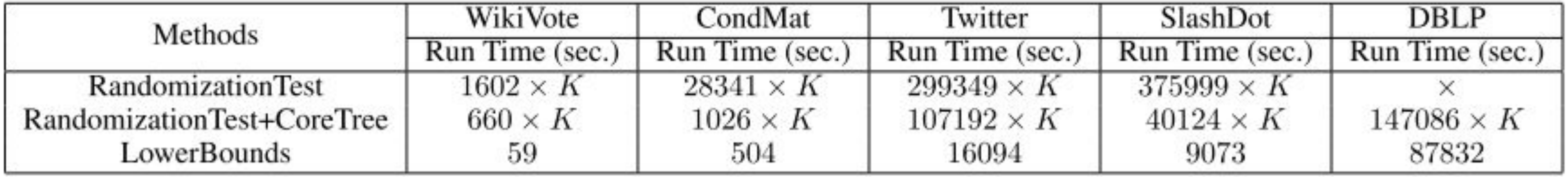
- **Detection power results:**
  - CNSS outperforms baseline methods under different signal strengths on the various network structures.
  - calibrated BJ score helps to precisely pinpoint the true affected subgraph as the strength of signal increases.
  - the use of core-tree decomposition and lower bounds do not have substantial effects on detection performance for these five real-world datasets.
  - the baseline methods do not have consistent performance over different values of μ with different network structures.

Experiments

★ **Run time results:**



Methods	WikiVote	CondMat	Twitter	SlashDot	DBLP
	Run Time (sec.)	Run Time (sec.)	Run Time (sec.)	Run Time (sec.)	Run Time (sec.)
LTSS	21	24	619	243	1425
EventTree	23	25	179	186	1019
ColorCoding	5220	8295	66690	29790	124956
NPHGS	8912	52046	998624	496587	×
AdditiveScan	17950	123100	×	×	×
DFGS	22791	×	×	×	×
CNSS	1771	43325	489624	447800	×
CNSS+CoreTree	685	1544	128812	45208	185053



Methods	WikiVote	CondMat	Twitter	SlashDot	DBLP
	Run Time (sec.)	Run Time (sec.)	Run Time (sec.)	Run Time (sec.)	Run Time (sec.)
RandomizationTest	$1602 \times K$	$28341 \times K$	$299349 \times K$	$375999 \times K$	×
RandomizationTest+CoreTree	$660 \times K$	$1026 \times K$	$107192 \times K$	$40124 \times K$	$147086 \times K$
LowerBounds	59	504	16094	9073	87832

- observe substantial speedups for CNSS+CoreTree.
- lower-bounds save huge preprocessing time.

Case Study: COVID-19 Confirmed Cases Subgraph Discovery

- Dataset: Covid-19 daily confirmed cases for 3,234 counties in the USA across over 25 weeks from January 22 to July 8, 2020.
- Build a spatial-temporal graph with 80,850 nodes and 850,725 edges based on the weekly confirmed cases and county adjacency.
    - each node represents a county in one week.
    - undirected spatial edge represents adjacency between counties.
    - undirected temporal edges:
        - from node i in week t to node i in week t+1.
        - from node i in week t to all neighboring nodes j in week t+1.
- P-value of each node: generated based on the rank of the weekly confirmed cases to county population ratio divided by the total number of nodes in the graph.
    - a higher ratio of the number of weekly confirmed cases to the county population indicates a higher rank and thus a smaller p-value.

Case Study: COVID-19 Confirmed Cases Subgraph Discovery

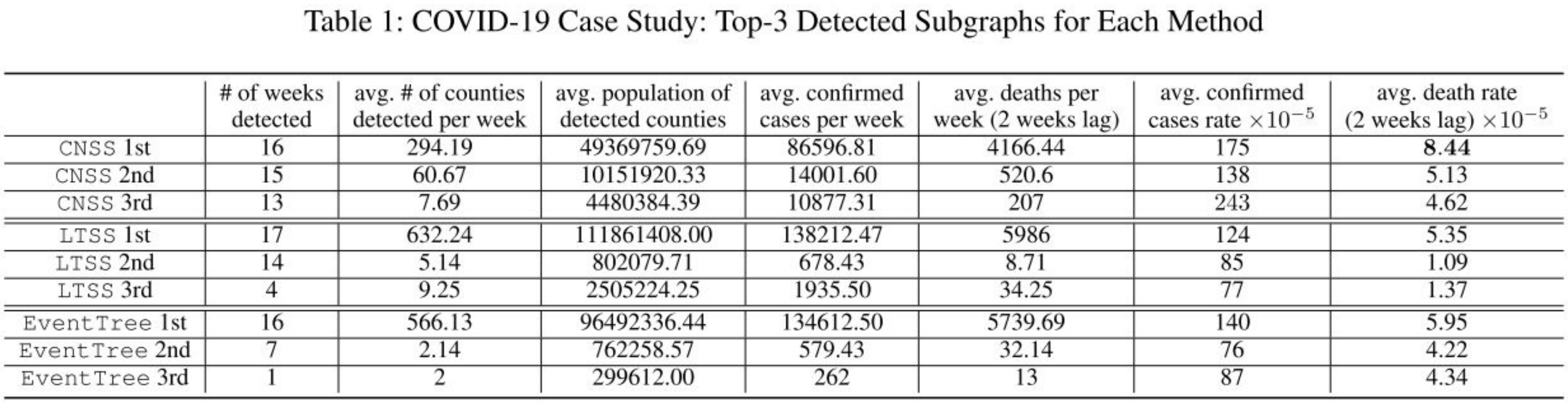


Table 1: COVID-19 Case Study: Top-3 Detected Subgraphs for Each Method

	# 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. confirmed cases rate $\times 10^{-5}$	avg. death rate (2 weeks lag) $\times 10^{-5}$
`CNSS` 1st	16	294.19	49369759.69	86596.81	4166.44	175	**8.44**
`CNSS` 2nd	15	60.67	10151920.33	14001.60	520.6	138	5.13
`CNSS` 3rd	13	7.69	4480384.39	10877.31	207	243	4.62
`LTSS` 1st	17	632.24	111861408.00	138212.47	5986	124	5.35
`LTSS` 2nd	14	5.14	802079.71	678.43	8.71	85	1.09
`LTSS` 3rd	4	9.25	2505224.25	1935.50	34.25	77	1.37
`EventTree` 1st	16	566.13	96492336.44	134612.50	5739.69	140	5.95
`EventTree` 2nd	7	2.14	762258.57	579.43	32.14	76	4.22
`EventTree` 3rd	1	2	299612.00	262	13	87	4.34

- ★ 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.
- ★ death rate data is not provided to the detection algorithms.

Limitations and Conclusions

- We show NPSS methods are mis-calibrated, failing to account for the maximization of the statistic over the multiplicity of subgraphs.
- We develop CNSS 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.
- The randomization test-based calibration approach is time-consuming, particularly for large-scale graphs.
- The closed-form lower bounds avoid the need of randomization test, but detection power is reduced when the anomalous signal strength is low.
- Core-tree decomposition methods enable the CNSS approach to scale to large real-world graphs without significant loss of detection performance.

Thank You!

This paper is based upon work supported by the National Science Foundation under Grant No. .

Presenter: Chunpai Wang
Email: cwang25@albany.edu

UNIVERSITY AT ALBANY
State University of New York

NYU

UT DALLAS
The University of Texas at Dallas

NSF