[[1]](#footnote-1)

Research case study: Detecting Anomalous Activity on Networks with the Graph Fourier Scan Statistic

Vipulanandan Pragatheeswaran

*Abstract*— To determine whether a single noisy measurement at each vertex of a given graph is constant over the graph or there exists a cluster of vertices with anomalous activation. In this proposal we will evaluate the usability of graph Fourier scan statistic (GFSS) to do the determination. We will demonstrate that the GFSS can efficiently detect a simulated arsenic contamination in ground water.

*Index Terms*— Adaptive signal detection, anomaly detection, graph filters, hypothesis testing

# INTRODUCTION

Signal detection on graphs is a key area of interest as it is relevant to numerous avenues of science such as disease outbreak detection, biomedical imaging, environmental monitoring, and malware detection. Additionally with the recent traction in the field of the spectral graph theory [1], has led to extensive Fourier and wavelet analysis on graphs. Hence its natural to extend the signal or anomaly detection done on regular signal processing domains extended to graph domain with the use of them.

The major caveat with statistical hypothesis testing of graph structures has gained some momentum in the recent years. Popularly the generalized likelihood ratio test (GLRT) also known as the graph scan statistic has been discussed in [3], [4]. Additionally, some works has also been done for specific topologies and specific signal classes [5], [6], [7]. Though this specific graph topology methods have been shown to have near optimal statistical performance, for general graphs and signal classes these detectors are infeasible, due to very high computational complexity or lack of constructive ways.

There has been some recent work which has been on developing fast graph subset scanning methods [10]. The greedy methods sacrifice statistical power. Additionally, there is work on developing Fourier basis and wavelets for graphs. In [11], the authors consider the complete graph and study detection under some combinational classes such as signals supported over cliques, bi-cliques and spanning trees. In this case study we will using a novel approach called the graph Fourier scan statistic (GFSS), to detect arsenic contamination levels in ground water in a graph.

# Methodology

Throughput this case study our focus of interest will be a , fixed, undirected graph with vertices.   
Denoted by the set edges denoted by pairs , and weighted adjacency matrix (where the weight denotes the 'strength' of the connection between vertices .

## Graph Fourier scan statistic

We will be using graph Fourier transform, we will be defining the graph Fourier scan statistic (GFSS).  
Let the diagonal matrix as,

Let combinatorial Laplacian matrix,

Let the eigenvalues and eigenvectors of with respectively, where we order the eigenvalues in increasing order. Hence, if is the matrix where the th column is the eigenvector and then we have

for small are the low frequency components of and for large are the high frequency components

Where, for small are the low frequency components of and for large are the high frequency components. The GFSS can be considered as a low pass filter for the graph G.

where is a tuning parameter, the selection of as . This will ensure the GFSS obtains a P-Value [8].

The above GFSS attenuates the high frequency components of the graph Fourier transformation. As increases with , the attenuation factor, , is 1 for small enough and is nonincreasing in .  
  
Hence we can definite GFSS as, the energy of the attenuated signal with an adjustment for the attenuation.

As the Laplacian matrix is symmetric and has a connected component, the first eigen value (sorted) will be zero. Hence the range of interest for the adjustment will be from the 2nd eigen value.

## Significance testing

The standard derivation and approach used by [8] is to use a likelihood ratio test. They start with the following hypothesis.

Where is the null hypothesis, where we are only receiving some form of noise, and is the case, where we are receiving a signal (arsenic detected) with some noise. Via the derivation they prove the authors simplify the statistics as following in [8].  
  
Under the null hypothesis , with probability at least where

Under the alternative hypothesis, , with probability at least where

Additionally, in [8] the authors show GFSS can asymptotically distinguish from if the SNR is stronger than

# Algorithm

The graph Fourier scan static given a sequence of vectors for a known graph structure, we will be processing it as follows.

1. Compute the kNN graph with signals. The area of contamination was fixed throughout all tests, this ensures that the water body contamination we are interested in is correlated with the values which are obtained from the nearby well.
2. Compute the degree matrix from the adjacency matrix of the graph which was created in step 1.
3. Calculate the eigen vectors and eigen values
4. Calculate the GFSS components and the associated statical value
5. Check the significance test against the threshold for

# Results

In this results section we will be testing the above methodology discussed on a synthetic arsenic dataset generated and tested the methodology out to a single file dataset provided by the authors.

## Synthetic arsenic contamination simulation

For the simulation of arsenic contamination, a total of 219 geolocations were created. And out of that 112 near by geolocations were contaminated. The contamination signal was set as 5, where the gaussian noise variance as 1.  
The test was repeated for 200 iterations to achieve generality.

Chart

Description automatically generated

Figure 1: eigen values & Fourier loadings against ordered eigen pairs

The above image is an evaluation of the eigen values and Fourier loadings obtained from a sample graph obtained via KNN method, with K set as 8.

A total of 2 times the locations of interests were created with half of those scenarios not having any contamination signal while the other half had a contamination signal. With the signal the KNN was created, and the likelihood ratio test was conducted and the following plot on receiver (ROC) generated.

Chart, scatter chart

Description automatically generated

Figure 2: True positive rate vs false positive rate for a signal elevation of value 5

Figure 2 was generated to a range of alpha and gamma values equally spaced between 0.1 and 1.0 with increments at 0.1.  
  
From the above we can see that the methodology works well. Compared to true positive rate (TPR), for certain alpha and gamma selections we can see that the false positive rates drop.

The following TPR vs FPR was generated by reducing, the signal amplitude very much less than the variance of the gaussian distribution, we can sort of see from the following plot that the testing is falling when the signal amplitude is reducing.

Chart, line chart

Description automatically generated

## Test for arsenic dataset

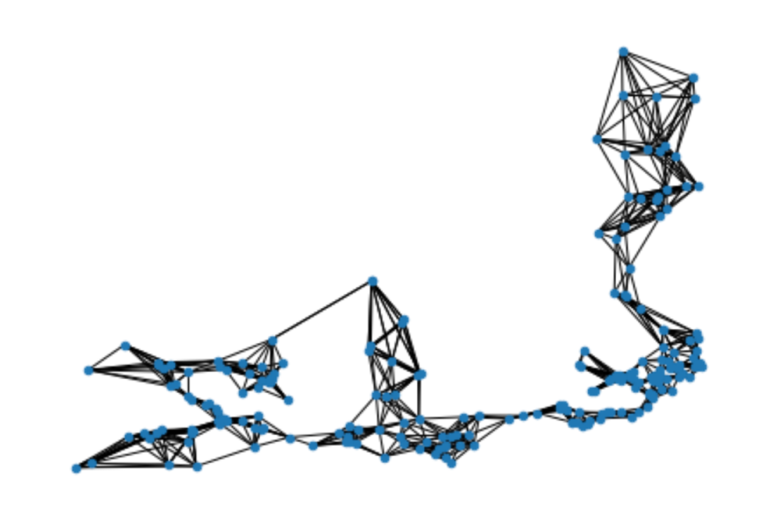


Figure 3: Geolocation plot of the signal obtained from the actual dataset

The dataset which was accessible via the paper/sites had a sample of 20043 geolocations with various states across USA. Since the node count might be too large to run, subsampling was done and only the region of “Idaho” was selected. During further evaluation it was seen that the recorded time series varied between 1979 and 1992 in this dataset. Nevertheless, all recordings of the arsenic level was pooled together disregarding the time aspect.

During the statistical test for the given dataset the arsenic contamination was correctly captured.   
  
Text

Description automatically generated with medium confidence

Figure 4: The statical results and the thresholds used to accept H0 and H1

# Conclusion

In this research case study, we focus on how to tractably detect anomalies in networks under gaussian noise conditions. The proposed Fourier scan statistic was evaluated in two datasets, the actual (subsampled) and synthetic.   
  
From the above results it can be concluded that for a smaller subsample of the dataset, the methodology works with a substantial accuracy, yet it is evident that its not the best. In the analysis section of the authors, it can be seen for certain simulations Aggregated statistics and max statistics outperforming this methodology.  
  
Nevertheless the graph based statistical approach is commendable. Thus, making it wonder whether the approach can be used in other graph-based anomaly detection problems. Specially, the anomaly detection on directed graph seems an interesting for me. Additionally, it could be noted that certain new graph neural networks are implementing eigen vectors or Laplacians as part of their layers (architectures). Hence the usage of a eigen values (used in this approach) as part of these architectures might be another venture of interest too.

# Codes & Data

The codes datasets can be found in the following GitHub link.

https://github.com/pragasv/anomalies-in-graphs

References

1. D. K. Hammond, P. Vandergheynst, and R. Gribonval, “Wavelets on graphs via spectral graph theory,” Appl. Comput. Harmon. Anal., vol.30, no. 2, pp. 129–150, 2011.
2. M. Focazio, A. Welch, S. Watkins, D. Helsel, and M. Horn, “A

retrospective analysis on the occurrence of arsenic in ground-water resources of the united states and limitations in drinking-water-supply characterizations,” Washington, DC, USA, U.S. Geological Survey Water-Resources Investigation Rep. 99-4279, 1999.

1. C. E. Priebe, “Scan statistics on graphs,” Johns Hopkins Univ., Baltimore, MD, USA, Tech. Rep. 650, 2004.
2. H. Wang, M. Tang, Y. Park, and C. E. Priebe, “Locality statistics for anomaly detection in time series of graphs,” IEEE Trans. Signal Process., vol. 62, no. 3, pp. 703–717, 2014
3. E. Arias-Castro, D. Donoho, and X. Huo, “Near-optimal detection of geometric objects by fast multiscale methods,” IEEE Trans. Inf. Theory, vol. 51, no. 7, pp. 2402–2425, 2005
4. E. Arias-Castro, E. Candes, H. Helgason, and O. Zeitouni, “Searching for a trail of evidence in a maze,” Ann. Statist., vol. 36, no. 4, pp.1726–1757, 2008
5. E. Arias-Castro, E. Candes, and A. Durand, “Detection of an anomalous cluster in a network,” Ann. Statist., vol. 39, no. 1, pp. 278–304, 2011.
6. J. Sharpnack, A. Rinaldo and A. Singh, "Detecting Anomalous Activity on Networks With the Graph Fourier Scan Statistic," in IEEE Transactions on Signal Processing, vol. 64, no. 2, pp. 364-379, Jan.15, 2016, doi: 10.1109/TSP.2015.2481866.
7. J. Sharpnack, A. Rinaldo, and A. Singh, “Changepoint detection over graphs with the spectral scan statistic,” in Proc. AISTATS, 2013, pp. 545–553.
8. S. Speakman and D. B. Neill, “Fast graph scan for scalable detection of arbitrary connected clusters,” in Proc. Int. Soc. Disease Surveill. Annu. Conf., 2010.
9. L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi, “On combinatorial testing problems,” Ann. Statist., vol. 38, no. 5, pp. 3063–3092,2010.
10. Ricaud, Benjamin; Borgnat, Pierre; Tremblay, Nicolas; Gonçalves, Paulo; Vandergheynst, Pierre (2019-07-01). ["Fourier could be a data scientist: From graph Fourier transform to signal processing on graphs"](https://doi.org/10.1016%2Fj.crhy.2019.08.003). *Comptes Rendus Physique*. Fourier and the science of today / Fourier et la science d’aujourd’hui. **20** (5): 474–488.

1. [↑](#footnote-ref-1)