SPECTRAL SUBGRAPH DETECTION WITH CORRUPT OBSERVATIONS

Benjamin A. Miller and Nicholas Arcolano

Lincoln Laboratory
Massachusetts Institute of Technology
Lexington, MA, 02420
bamiller@ll.mit.edu, arcolano@post.harvard.edu

ABSTRACT

Recent work on signal detection in graph-based data focuses on classical detection when the signal and noise are both in the form of discrete entities and their relationships. In practice, the relationships of interest may not be directly observable, or may be observed through a noisy mechanism. The effects of imperfect observations add another layer of difficulty to the detection problem, beyond the effects of typical random fluctuations in the background graph. This paper analyzes the impact on detection performance of several error and corruption mechanisms for graph data. In relatively simple scenarios, the change in signal and noise power is analyzed, and this is demonstrated empirically in more complicated models. It is shown that, with enough side information, it is possible to fully recover performance equivalent to working with uncorrupted data using a Bayesian approach, and a simpler cost-optimization approach is shown to provide a substantial benefit as well.

Index Terms— Graph theory, signal detection theory, spectral analysis, subgraph detection, data error and corruption

1. INTRODUCTION

In numerous applications, entities and the relationships between them are of interest, and the detection of anomalous behavior within some subset of the entities is frequently an important problem. This capability could be used, for example, to find strange traffic patterns in a computer network or unlikely connections in a social network.

Across these varied applications, the data of interest are usually encoded in a graph. A graph $G = (V, E)$ is a mathematical object used to represent relational data. It consists of a pair of sets: a set of vertices, V, denoting the entities, and a set of edges E that represent the relationships of interest. While this structure is ubiquitous and intuitive, applying traditional detection theory to graphs can be difficult, as they are combinatorial objects and optimal signal detection may require solving an NP-hard problem. Thus, spectral methods have a computational advantage over combinatorial solutions, as demonstrated in [1, 2].

In practice, perfect knowledge of the network structure is often not possible. Observations may be obtained through imperfect mechanisms, with sensor noise corrupting measurements. In other situations, the network may not be observable at all, and we may have to rely on proxies for the true relationships of interest. Understanding how these phenomena alter subgraph detection performance is important, and in this paper we demonstrate the effects of several models for error on simulated networks.

Previous work on subgraph detection (e.g., [2, 3]) has focused primarily on detection in the presence of random fluctuations in the background graph and the signal subgraph, and have not addressed the problem of additional data corruption. Errors in sampling real-world social networks have been studied in the social science literature [4], as they have in the study of massive social network analysis [5]. Recently, work on the impact of noise and error on the statistics of graphs has been a topic of interest in the community [6, 7], including work on trading off the quantity of observations with the fidelity of the data [8]. In this paper, we add to this line of research, analyzing the effect that observation error has on spectral methods for the detection of anomalous subgraphs.

The remainder of this paper is organized as follows. In Section 2, we review the subgraph detection problem model and propose various models for observation errors. Section 3 provides analytical insight into the impact that some of these models have on detection performance, and empirical results in a simulated setting. In Section 4, we discuss methods for fusion of multiple observations, and analyze the empirical improvement in detection performance. In Section 5, we conclude and outline future research directions.

2. PROBLEM MODEL

2.1. The Subgraph Detection Problem

The subgraph detection problem, as studied in [2, 9, 3], is a classical detection problem with a graph as the observation. The objective is to resolve the binary hypothesis test

$$\begin{cases} H_0: & \text{The observed graph is "noise" } G_B \\ H_1: & \text{The observed graph is "signal+noise" } G_B \cup G_S. \end{cases}$$

Here the observation is a graph $G = (V, E)$, and the union of two graphs $G_B = (V_B, E_B)$ and $G_S = (V_S, E_S)$ is defined as $G_B \cup G_S = (V_B \cup V_S, E_B \cup E_S)$. Under the null hypothesis, H_0, the observed graph is generated entirely by some distribution that dictates typical behavior. Under the alternative hypothesis, H_1, a subgraph is added to the normal background activity. In this paper, we focus on the case where the subgraph vertices are all part of the background graph, i.e., $V_S \subset V_B$.

The analysis framework originally proposed in [2] is based on spectral analysis of graph residuals, which requires the use of ma-

This work is sponsored by the Intelligence Advanced Research Projects Activity (IARPA) under Air Force Contract FA8721-05-C-0002. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.
Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA or the U.S. Government.

trix representations of a graph. The adjacency matrix A of a graph is a $|V| \times |V|$ matrix in which entry in the ith row and jth column is nonzero only if there is an edge from vertex i to vertex j (this requires an arbitrary labeling of vertices with integers). Residuals analysis is performed by considering the distribution of the background graph and computing the eigendecomposition of the difference between the observation and its expected value, i.e., computing

$$U\Lambda U^T = B := A - \mathbb{E}[A],$$

where B is known as the modularity matrix in the community detection literature [1].

2.2. Error Models

This paper focuses on analyzing the performance of the spectral detection framework when errors are present in the data. We will denote by A the adjacency matrix of the latent (true) graph, and $\widehat{A} = \{\hat{a}_{ij}\}$ will denote the adjacency matrix of the observed (corrupted) graph. We consider the effect of the following error models on subgraph detection.

Uniform deletion. Each edge in the latent graph has a fixed probability p of being observed, but no false edges are observed. Thus, $\Pr(\hat{a}_{ij} = 1 | a_{ij} = 1) = p$ and $\Pr(\hat{a}_{ij} = 1 | a_{ij} = 0) = 0$. This model may approximate mechanisms such as random sensor dropouts.

Uniform corruption. For all pairs of vertices, the observed relationship is incorrect with uniform probability p, yielding $\Pr(\hat{a}_{ij} = 1 | a_{ij} = 1) = (1 - p)$ and $\Pr(\hat{a}_{ij} = 1 | a_{ij} = 0) = p$. This may simulate incorrect observations due to sensor noise.

Degree-based corruption. Similar to uniform corruption, but the expected number of errors associated with a given vertex is proportional to its degree. In this case, the probability of an incorrect edge appearing, or an edge being missed, is given by $\alpha k_i k_j / \sum_m k_m$, where α is a scaling constant and k_i is the observed degree of vertex i.

Random subgraph. Only the connections within a randomly selected subset of vertices are observed. This simulates a case in which only a subset of the population of entities is observable. This is similar to egocentric sampling, mentioned in [4].

Similarity-based errors. Each vertex has associated metadata z_i, and an edge connecting to vertex i is mistakenly connected to vertex j based on an decreasing function of some distance metric between their associated metadata, $s(z_j, z_j)$. This model simulates corruption due to clerical errors or ambiguous data.

3. DETECTION PERFORMANCE WITH OBSERVATION ERRORS

Some of the simpler error models enable theoretical justification for performance loss. Here we consider the theoretical impact of several of the models in relatively simple scenarios (specifically Chung–Lu random backgrounds and signals consisting of small clusters), and empirically demonstrate performance in simulation.

3.1. Theoretical Analysis

Using the spectral norm as a metric for signal and noise power, we can derive the change in power due to the uniform deletion and random subgraph error models. Considering the simple background of an Erdős–Rényi random graph (i.e., a graph where each possible edge occurs with equal probability), we can leverage classic results from random matrix theory, recently applied to planted clique detection [10]. Since the residuals matrix of an Erdős–Rényi graph has entries with zero mean and equal variance, as $|V| \to \infty$, the distribution of its eigenvalues will tend to a semicircle centered at 0 with radius $2\sqrt{|V|p(1-p)}$. Thus, if the expected number of edges is changed by a factor of α, the spectral norm will be changed by a factor of $\sqrt{(\alpha p(1-\alpha p))/(p(1-p))}$. When the graph is sparse, as in many applications of interest, the $(1-p)$ terms can be ignored, and the change in spectral norm will be approximately $\sqrt{\alpha}$. Likewise, if the number of vertices is changed by a factor of α, as in the random subgraph model, and the edge probabilities remain the same, the spectral norm will change by a factor of $\sqrt{\alpha}$. If the signal subgraph is a small cluster with a relatively high edge probability, subtraction of the expected value of the background will not significantly alter the signal power, and a factor of α decrease in the number of edges or vertices will reduce the signal power—the number of vertices times the edge probability—by a factor of α. The signal power will decrease much more quickly than the background power, making detection more difficult.

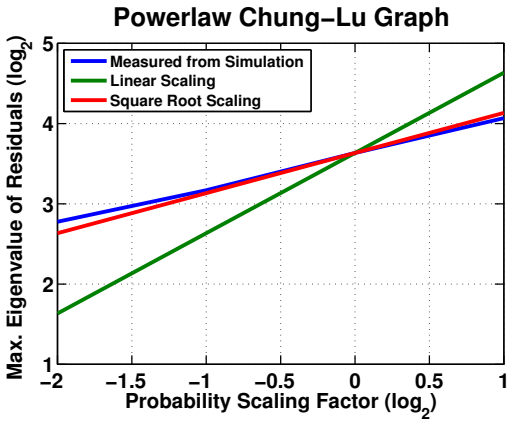

Fig. 1. Spectral norms of sparse Chung–Lu random graphs. As the edge probabilities are changed, the spectral norm of the residuals matrix scales roughly with the square root of the probability scaling factor.

While other models are more complicated to analyze, we empirically observe similar behavior. Fig. 1 demonstrates this with a Chung–Lu random graph [11], i.e., a graph with a rank-1 expected value. The given expected degree sequence of the graph follows a powerlaw distribution, as seen in many real graphs, and is altered by scaling factors of 0.25, 0.5 and 2. As shown in the figure, the actual spectral norm of the residuals matrix scales closely to the square root of the scaling factor, as the graph is quite sparse. (It has the same expected degree sequence as the background graphs in Section 3.2, with an average degree of about 10.) Therefore, in this case, the signal power will also decrease faster than the noise power.

With the uniform and degree-based corruption mechanisms applied to a Chung–Lu background, an edge may occur because it exists in the latent graph and is not removed, or does not exist in the latent graph and is added. This results in probabilities $\hat{p}_{ij} = p_{ij}(1 - p_{ij}^{\text{error}}) + p_{ij}^{\text{error}}(1 - p_{ij})$, yielding probability matrices

$$\mathbb{E}\left[\widehat{A}\right] = ww^T\left(1 - \frac{2\alpha\|w\|_1^2}{|V|^2}\right) + \frac{\alpha\|w\|_1^2}{|V|^2}\mathbf{1}_{|V|\times|V|} \qquad (1)$$

for uniform corruption and

$$\mathbb{E}\left[\widehat{A}\right] = (1+\alpha)ww^T - 2\alpha w^2 {w^2}^T,$$

for degree-based corruption. Here w is a vector of weights that determine the edge probabilities, where the probability of vertex i and vertex j sharing an edge is $w_i w_j$, w^2 is a vector whose ith component is w_i^2, and α is a constant. In powerlaw graphs, w and w^2 will typically be correlated, and the $2\alpha w^2 {w^2}^T$ term will slightly reduce the noise power. Thus, the noise power will be increased by a factor of up to $(1+\alpha)$, and has a relatively minimal impact on signal power, since the signal vertices have relatively low degree. In the case of uniform corruption, we see a "whitening" effect on the noise: while there are more background edges, a more substantial fraction is uniform across the graph (the $\mathbf{1}_{|V|\times|V|}$ term in (1)). Therefore, the additional error will not be correlated with the background and will likely have a minimal effect on detection performance.

With similarity-based errors, we will consider the case in which all vertices are equally likely to be mistaken for all other vertices. The expected value of the corrupted graph will be

$$\left(\alpha I + \frac{(1-\alpha)}{N}\mathbf{1}_{|V|\times|V|}\right) A \left(\alpha I + \frac{(1-\alpha)}{N}\mathbf{1}_{|V|\times|V|}\right).$$

Here $(1-\alpha)/N$ is the probability that either the row or column of the edge entry will be swapped with that of each other vertex. If the background graph comes from a Chung–Lu model, then the expected value of the resulting graph is given by

$$\mathbb{E}\left[\widehat{A}\right] = \left(\alpha I + \frac{(1-\alpha)}{N}\mathbf{1}_{|V|\times|V|}\right) ww^T \left(\alpha I + \frac{(1-\alpha)}{N}\mathbf{1}_{|V|\times|V|}\right).$$

Thus, the expected value remains a rank-1 matrix, with the weight vector w replaced with

$$\hat{w} = \alpha w + \frac{(1-\alpha)\|w\|_1}{N}\mathbf{1}_{|V|}.$$

One way to interpret this change is that it makes the Chung–Lu graph more Erdős–Rényi-like, making the expected degree sequence closer to uniform. While this does not change the overall number of edges in the graph, it does cause a reduction in the spectral norm as the variance in the background becomes more evenly spread across the graph. A cluster embedded into the background will also have a rank-1 expected value, and if the edge probability within the subgraph is p_S, then when the observation is corrupted, the spectral norm of the expected corrupted subgraph is

$$\frac{N_S p_S}{N}\left(N\alpha^2 + (1-\alpha^2)N_S\right),$$

which, assuming the number of subgraph vertices is much smaller than the number of total vertices, is not much larger than $\alpha^2 N_S p_S$, thus roughly reducing the signal power by a factor of α^2.

3.2. Simulation Results

We ran several 10,000-trial Monte Carlo simulations to illustrate the impact of the noted error mechanisms on detection performance. In each experiment, the background consisted of an R-MAT stochastic Kronecker graph [12] with 1024 vertices and an average degree of approximately 10. The modularity matrix introduced in [1] was used for a residuals model, fitting the observed graph to a Chung–Lu model and computing the top eigenvectors of

$$B = A - \frac{kk^T}{\|k\|_1},$$

where k is the observed degree vector. The mild community structure of the R-MAT background provides some mismatch from the Chung–Lu model, yielding a more challenging problem.

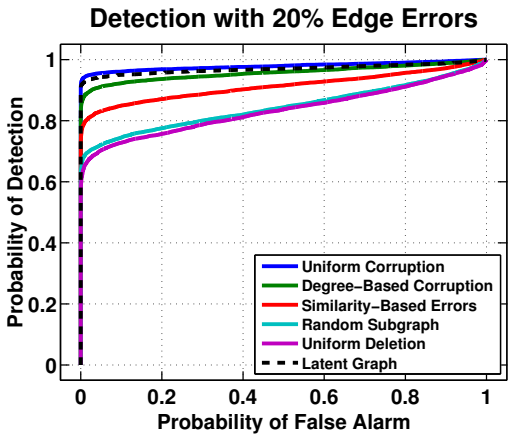

Fig. 2. A comparison of subgraph detection performance with each error mechanism in place.

In each of the simulations, a background graph was created, and a 12-vertex cluster where each pair of vertices has 85% probability of connection was embedded for the cases under the alternative hypothesis. Each error mechanism was applied so that the average number of edge errors (i.e., the number of incorrect edges plus the number of missing edges) was 20% of the total number of edges in the latent graph. For similarity-based errors, similarities were computed by generating a 3-dimensional feature vector for each vertex, drawn independently at random from a uniform distribution over $[0,1]^3$. As a detection statistic, the method introduced in [2] is used, where the residuals are projected into their principal 2-dimensional subspace, and a chi-squared test for independence is performed on a contingency table representing the number of points falling in each quadrant. The statistic is maximized over rotation in the plane, so that more radially symmetric projections have lower statistics.

Results of the experiment are shown in Fig. 2. As predicted by our theoretical analysis, uniform corruption actually improves detection performance, since it has a decorrelating effect on the noise with little impact on the signal. Degree-based corruption has a mildly negative effect on detection performance, since the data corruption is somewhat correlated with the background noise. As expected, uniform deletion and the random subgraph mechanism have extremely similar performance, since both scale signal power by a factor of 0.8 and noise power by approximately $\sqrt{0.8}$. A less significant reduction in performance is demonstrated by the similarity-based errors. As discussed, this is due to the "whitening" of the background noise that occurs, making it more like an Erdős–Rényi graph.

4. MULTI-SOURCE FUSION FOR PERFORMANCE RECOVERY

One technical area of interest is the fusion of data from multiple diverse sources to improve detection performance. In this section,

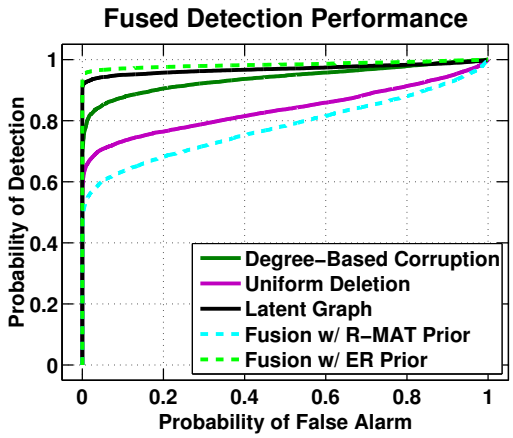

Fig. 3. Detection performance when recovering the original graph via Bayesian inference.

we consider two approaches: a Bayesian inference approach and an optimization approach.

4.1. Bayesian Approach

For the Bayesian approach, we assume knowledge of the background model and the error model. For a given sequence of observations $\widehat{A}_m$, we can compute the expected value of the latent graph. This is especially tractable when the errors are independent, i.e., an error regarding the observation of one edge does not impact the observation of any other. Uniform deletion and the random corruption mechanisms fit most easily in this category, so we use these for a fusion experiment. The expected value of A is computed, given the observations and the model parameters, and detection algorithm is run on the associated weighted graph.

We consider a graph with 50% edge errors via degree-based corruption and one with 20% edge errors by uniform deletion. The background and foreground models are the same as in Section 3.2. Results of this experiment are shown in Fig. 3. Interestingly, when significant information about the background is used—in particular, the true edge probabilities—detection performance actually decreases. This may be because the signal subgraph does not follow the background model, so using additional information helps to reconstruct the background but hinders signal power. Using an Erdős–Rényi (ER) prior, however, fully recovers the detection performance achieved when operating on the latent graph. In fact, performance is slightly better, probably due to the additional information provided: the expected total number of edges.

4.2. Optimization Approach

While the Bayesian approach is satisfying in its statistical rigor, such in-depth knowledge of the error mechanism may not be available. In some applications, judgement calls must be made based on the level of "trust" an analyst has in a particular data source. In this section, we consider fusion via a simple weighting procedure, where if an edge occurs between vertices i and j in any of the n observations, then in the fused graph the edge is given the weight

$$g\left(\theta_0 + \sum_{m=1}^{n} \theta_m \hat{a}_{ij}^{(m)}\right),$$

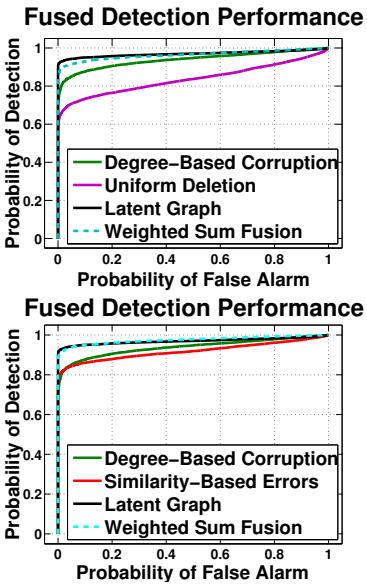

Fig. 4. Detection performance when fusing observations via a weighted sum. Fusion of an observation with uniform corruption and one with uniform deletion nearly recovers performance achieved with the latent graph (top), while fusion of mechanisms with a milder effect on performance come even closer to full recovery (bottom).

where θ_m, $0 \leq m \leq n$, are real-valued coefficients and $g(x) = 1/(1+e^{-x})$ is the logistic function. This technique is applied to the same scenario as in Section 4.1, as well as the same degree-based corruption mechanism fused with data with similarity-based errors.

Results of this experiment are shown in Figure 4. Using the same data as in the Bayesian experiment, while the same performance is not achieved, there is a substantial improvement in detection performance. Also, this technique can be used for models that may be too complicated for full Bayesian inference of the latent graph. As shown in the figure, when the observation with degree-based corruption is fused with the graph with similarity-based errors, performance nearly equivalent to working with the latent graph is achieved.

5. SUMMARY

This paper studies the effect of various data corruption models on spectral methods for anomalous subgraph detection. Five models for data errors are proposed, and each one's influence on signal and noise power is studied for simple background and foreground models. Simulations demonstrate how these error models alter detection performance, with random corruption mechanisms having relatively little impact, random vertex and edge deletions having a much more substantial effect, and similarity-based errors having a somewhat milder impact. When fusing multiple observations from different error mechanisms together, a Bayesian inference technique is shown to fully recover detection performance, while a simple weighting procedure also provides a substantial benefit without necessarily relying on additional knowledge of the model. Future work will focus on a deeper analysis of the signal and noise power metrics, and addressing additional methods, such as snowball sampling and random walk sampling.

6. REFERENCES

[1] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Phys. Rev. E*, vol. 74, no. 3, 2006.

[2] Benjamin A. Miller, Nadya T. Bliss, and Patrick J. Wolfe, "Toward signal processing theory for graphs and non-Euclidean data," in *ICASSP*, 2010, pp. 5414–5417.

[3] Benjamin A. Miller, Nicholas Arcolano, and Nadya T. Bliss, "Efficient anomaly detection in dynamic, attributed graphs," in *Proc. IEEE Int. Conf. Intelligence and Security Informatics*, 2013, pp. 179–184.

[4] Mark S. Handcock and Krista J. Gile, "Modeling social networks from sampled data," *Ann. Appl. Stat.*, vol. 4, no. 1, pp. 5–25, 2010.

[5] Jure Leskovec and Christos Faloutsos, "Sampling from large graphs," in *Proc. KDD*, 2006, pp. 631–636.

[6] Sanjukta Bhowmick, Sriram Srinivasan, and Vladimir Ufimtsev, "Evaluating noise in complex networks," *SIAM Conf. Computational Sci. and Eng.*, 2013.

[7] Eric D. Kolaczyk, "Quantification of uncertainty in network summary statistics," *SIAM Conf. Computational Sci. and Eng.*, 2013.

[8] Carey E. Priebe, Joshua T. Vogelstein, and Davi Bock, "Optimizing the quantity/quality trade-off in connectome inference," Preprint: arXiv:1108.6271v2, 2011.

[9] Benjamin A. Miller, Michelle S. Beard, and Nadya T. Bliss, "Matched filtering for subgraph detection in dynamic networks," in *Proc. IEEE Statistical Signal Process. Workshop*, 2011, pp. 509–512.

[10] Raj Rao Nadakuditi, "On hard limits of eigen-analysis based planted clique detection," in *Proc. IEEE Statistical Signal Process. Workshop*, 2012, pp. 129–132.

[11] Fan Chung, Linyuan Lu, and Van Vu, "The spectra of random graphs with given expected degrees," *Proc. of National Academy of Sciences of the USA*, vol. 100, no. 11, pp. 6313–6318, 2003.

[12] Deepayan Chakrabarti, Yiping Zhan, and Christos Faloutsos, "R-MAT: A recursive model for graph mining," in *Proc. SIAM Int. Conf. Data Mining*, 2004, vol. 6, pp. 442–446.