Connected Subgraph Detection with Mirror Descent on SDPs

Cem Aksoylar [1] Lorenzo Orecchia [1] Venkatesh Saligrama [1]

Abstract

We propose a novel, computationally efficient mirror-descent based optimization framework for subgraph detection in graph-structured data. Our aim is to discover anomalous patterns present in a connected subgraph of a given graph. This problem arises in many applications such as detection of network intrusions, community detection, detection of anomalous events in surveillance videos or disease outbreaks. Since optimization over connected subgraphs is a combinatorial and computationally difficult problem, we propose a convex relaxation that offers a principled approach to incorporating connectivity and conductance constraints on candidate subgraphs. We develop a novel efficient algorithm to solve the relaxed problem, establish convergence guarantees and demonstrate its feasibility and performance with experiments on real and very large simulated networks.

1. Introduction

We consider the problem of connected subgraph detection, motivated by statistical anomaly detection on networks where the aim is to determine whether there exists a set of connected nodes that exhibit anomalous signal values. One example of network anomaly detection is disease outbreak detection (Patil et al., 2003), where the nodes are associated with counties that are linked by geographical neighborhood and signal values on nodes depict the number of patients related to a disease. In the existence of a disease outbreak that spreads geographically, higher signal values would be present on certain counties that are neighbors of each other, therefore constituting a subgraph structure. Similar problems in different research areas also exist, such as detection of intrusions in communication or sensor networks, community detection or video surveillance.

The detection or estimation of arbitrary connected subgraphs over graph-structured signals is an example of a structured signal recovery problem and generalizes many useful types of structures such as intervals or paths (Addario-Berry et al., 2010; Arias-Castro et al., 2008). While the existence of structure in terms of connectivity leads to better statistical complexity in detecting or recovering the anomalous sets compared to arbitrary subsets, efficient characterization of sets obeying the connectivity constraint is important for obtaining practical algorithms for detection and estimation. In this paper we aim to characterize the space of arbitrary connected subsets of nodes in a graph via spectral relaxation and propose efficient optimization algorithms that exploit this characterization.

Related work and contributions: Subgraph detection is a difficult problem since connected subgraphs represent a combinatorial structure and systematic approaches to characterizing the space of connected subgraphs of a given graph are relatively recent. Traditional approaches to this problem usually consider parametric methods, which originate from the scan statistics literature (Glaz et al., 2001) and consider scanning for specific shapes such as rectangles, circles or neighborhood balls on graphs (Patil et al., 2003; Kulldorff et al., 2006; Priebe et al., 2005). More recently nonparametric approaches have been considered for subgraphs with arbitrary shapes on general graphs such as the simulated annealing approach of (Duczmal & Assuncao, 2004), however it is a heuristic method without statistical or computational guarantees. There is also a line of work focused on statistical analysis with nonparametric shapes (Addario-Berry et al., 2010; Arias-Castro et al., 2008; 2012) but are computationally intractable.

More recently, (Qian et al., 2014; Qian & Saligrama, 2014) proposed linear matrix inequalities as a way to characterize the connectivity of subsets of nodes exactly. This approach is the similar to ours where we also consider an SDP relaxation with LMI constraints, however we take a different approach to formulating the problem and its relaxation, with the goal to obtain a convex optimization program that is amenable to efficient iterative methods. In contrast, the aforementioned method is only applicable to small problem sizes. Another notable work in this area is the spectral scan statistic approach proposed by (Sharpnack et al., 2016), which presents a computationally tractable al-

*Equal contribution [1]Boston University, Massachusetts, USA. Correspondence to: Venkatesh Saligrama <srv@bu.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

gorithm with consistency guarantees. However this method aims to obtain graph partitions with small conductance and balanced sizes, in contrast to our formulation that guarantees connected subgraphs. In recent work, (Wu et al., 2016) consider nonparametric statistics for signals in addition to nonparametric shapes, obtaining a computationally tractable algorithm by heuristically approximating the underlying graph with trees.

The contribution of this paper is twofold: First, we develop a convex relaxation of the subgraph detection problem that results in an semidefinite optimization formulation, with provable guarantees on the connectivity of the resulting solutions related to the internal conductance of the subgraph. Second, we propose an efficient iterative framework for optimizing the SDP that scales well with large problem sizes, and show computational guarantees. One of the major differences of our formulation to those of (Qian et al., 2014; Qian & Saligrama, 2014) is that prior work enforce a number of constraints that scale with the problem size, whereas ours only considers a constant number of constraints. Also, while aforementioned work utilized generalized convex optimization solvers, our formulation allows us to propose specialized and efficient iterative algorithms.

2. Connected Subgraph Detection

In this section we define the notation and introduce the two statistical models that we consider for the connected subgraph detection problem. Let $G = (V, E)$ denote an undirected unweighted connected graph with n nodes that is provided as input to the problem. For $i \in V$, we write d_i for the degree of vertex i in G and let d be an upper bound on all d_i. For a subset $S \subseteq V$, the notation $\mathrm{Vol}(S)$ indicates the volume measure of S, i.e., $\mathrm{Vol}(S) = \sum_{i \in S} d_i$. We also denote by $G_S = (S, E_S)$ the subgraph induced by the subset S. For an input root vertex $r \in V$, we write $\Lambda_r = \{S \subseteq V : r \in S,\, G_S \text{ is connected in } G\}$. Indicator vectors with notation 1_S are defined as $n \times 1$ vectors with i-th index 1 if $i \in S$ and zero otherwise.

We consider observations $x_v \in \mathbb{R}^p$ associated with each node $v \in V$ in the graph G. We are concerned with optimization problems of the form

$$\max_{S \in \Lambda_r} c(S), \tag{1}$$

for a cost function $c(\cdot)$ which depends on $x_S = \{x_v\}_{v \in S}$. We remark that this is a difficult problem due to the combinatorial nature of the constraint, in fact variants of the prize-collecting Steiner tree problem which is known to be NP-hard (Johnson et al., 2000) can be reduced to the above formulation. Below we provide two examples of the setup that we consider, namely elevated mean detection and correlation detection.

Elevated mean detection: Here, the aim is to detect the existence of a subgraph $S \in \Lambda_r$ comprising of nodes with an elevated mean compared to the other nodes. A simple example is the Gaussian elevated mean model, where $x_v = \mu 1\{v \in S\} + z_v$ for $\mu > 0$ and $z_v \sim \mathcal{N}(0, \sigma^2)$. Another example we consider is the Poisson variant, where $x_v \sim \mathrm{Poisson}((1+\mu 1\{v \in S\})\,\lambda)$. We consider the optimization (1) with the scan statistic

$$c_1(S) = \frac{1}{\sqrt{|S|}} \sum_{i \in S} x_i, \tag{2}$$

which can be shown to correspond to the generalized likelihood ratio test (GLRT) for the Gaussian detection problem, while also encouraging graph-sparse solutions.

Correlation detection: Another example is the problem of detecting and estimating a subgraph with correlated signal values. The canonical statistical model that has been investigated for this problem is where the signals are jointly Gaussian random variables $X_1, \ldots, X_n$, where $\mathrm{cov}(X_i, X_j) = 1$ if $i = j$, ρ if $i, j \in S$ and zero otherwise (Arias-Castro et al., 2012). Note that while related work consider arbitrary k-sets or special shapes such as intervals for S, we allow arbitrary connected subgraphs. One simple test for detecting or estimating a correlated subgraph induced by set S is (1) with the scan statistic

$$c_2(S) = \frac{1}{|S|} \sum_{i,j \in S} \hat{\Sigma}_{ij}, \tag{3}$$

where $\hat{\Sigma}$ is the estimated covariance matrix which can either be defined by a single observation $xx^\top$ when $p = 1$ or multiple observations for $p > 1$.

Characterizing subgraph connectivity: Rather than focusing on exactly characterizing the connectedness of an induced subgraph G_S, we aim to enforce it by lower bounding the *conductance* of cuts within G_S. For a weighted graph $G = (V, E, w)$, the conductance of a cut S is:

$$\phi_G(S) = \frac{w(S, V \setminus S)}{\mathrm{Vol}(S)},$$

where $w(S, V \setminus S)$ is the total weight of edges connecting nodes in S to nodes in $V \setminus S$. The graph conductance is the lowest conductance among cuts containing at most half of the volume of the graph, i.e., $\phi_G = \min_{S \subset V : \mathrm{Vol}(S) \le \mathrm{Vol}(V)/2} \phi_G(S)$. Conductance is a natural graph-partitioning objective because of its intimate connection with the behavior of random walks. It is also widely used in practice in the design of clustering and segmentation algorithms. We can use conductance to ensure subgraph connectivity by imposing the following constraints[1]

[1] Note that for technical reasons, the conductance ϕ_{G_S} on the induced graph G_S still employs the definition of volume given by the larger graph G.

on the integral solution S:

$$\phi_{G_S}(T) = \frac{w(T, S \setminus T)}{\mathrm{Vol}(T)} \geq \gamma > 0, \quad \forall T \subseteq S \setminus \{r\}. \quad (4)$$

For γ sufficiently small, i.e., $\gamma < 1/\mathrm{Vol}(V)$, this requirement is equivalent to the connectivity condition on G_S. It is useful to notice at this stage that the condition in constraint (4) is stronger than a lower bound on the conductance of the induced subgraph G_S. Indeed, for an unweighted graph the constraint (4) implies

$$\phi_{G_S} = \min_{U \subseteq S : \mathrm{Vol}(U) \leq \frac{\mathrm{Vol}(S)}{2}} \frac{|E(U, S \setminus U)|}{\mathrm{Vol}(U)} \geq \gamma,$$

where $E(U, S \setminus U)$ is the set of edges between sets of nodes U and $S \setminus U$. However, our constraint is stronger than requiring the induced conductance of G_S to be γ, as the bound also holds for subsets $T \subset S$ comprising more than half the volume of G_S. In the appendix, we provide a brief comparison with other measures of connectivity.

3. Relaxation

We next consider a convex relaxation of the objectives of (2) and (3) as a linear functional of positive semidefinite matrix variable M. We remark that in the case that $x \geq 0$ (e.g. the Poisson model), maximizing the scan statistic $c_1(S)$ is equivalent to maximizing its square

$$c_1^2(S) = \frac{1}{|S|} \left(\sum_{i \in S} x_i \right)^2 = \frac{1}{|S|} \sum_{i,j \in S} x_i x_j, \quad (5)$$

which has the same form as the statistic $c_2(S)$. Defining the indicator vector $u = 1_S$ and noting that $u_i = u_i^2$ we can write the quadratic integer program (IP) as

$$\max_{u \in \{0,1\}^n, \{i : u_i = 1\} \in \Lambda_r} \frac{\sum_{i,j} x_i x_j u_i u_j}{\sum_i u_i^2}. \quad (6)$$

We relax this IP to a semidefinite program (SDP) by turning each element u_i to a vector $v_i \in \mathbb{R}^n$ such that scalar multiplication is transformed to inner product and we have $\langle v_i, v_j \rangle = 1$ if $i, j \in S$ and zero otherwise. Moreover, we also enforce non-negativity by requiring that $\langle v_i, v_j \rangle \geq 0$ for all $i, j \in V$. We then have

$$\max_{\substack{v_i \in \mathbb{R}^n, \{i : |v_i| > 0\} \in \Lambda_r \\ \langle v_i, v_j \rangle \geq 0, \ \forall i,j \in V}} \frac{\sum_{i,j} x_i x_j \langle v_i, v_j \rangle}{\sum_i |v_i|^2}. \quad (7)$$

Using the the Gram matrix $M = V^\top V \succeq 0$ instead of the vectors v_i's and fixing the trace of M, $I \cdot M$ to 1 w.l.o.g. (due to the homogeneity of the ratio in the objective), we obtain the relaxation

$$\max_{M \in \Delta_n, M \geq 0} C \cdot M \quad \text{s.t.} \quad \{i : |M_{ii}| > 0\} \in \Lambda_r, \quad (8)$$

where $C = xx^\top$ and we define Δ_n to be the spectrahedron of unit trace PSD matrices. The relaxation follows along exactly for $c_2(S)$ with $C = \hat{\Sigma}$. This linear functional formulation is very general and can be adapted to solve subgraph problems with other general cost functions.

Next, we propose a novel SDP formulation of the connected subgraph constraint $\{i : |M_{ii}| > 0\} \in \Lambda_r$, as a single linear matrix inequality based on a spectral relaxation of the integral conductance constraint of (4).

3.1. Spectral Graph Theory

We start by introducing some basic notation and concepts from spectral graph theory. For a weighted graph $H = (V_H, E_H, h)$, we denote its adjacency by A_H and its degree matrix D_H. The Laplacian of H is then defined as $L_H = D_H - A_H$. The $n \times n$ Laplacian matrix for the graph on V consisting only of edge $\{i, j\}$ is $L_{ij} = e_{ii} + e_{jj} - e_{ij} - e_{ji}$, where e_{ij} is an all-zero matrix except for a one at index (i, j). Notice that $L_H = \sum_{(i,j) \in E_H} h_{ij} L_{ij}$. We omit the subscripts for all graph matrices and sets when referring to the instance graph G. For a subset $S \subset V$, we denote by K_S the complete graph on S, i.e., the graph having an edge of weigth $d_i d_j$ between i and j for any $i, j \in S$. The spectral gap λ_S of an induced subgraph G_S of the input graph G is defined as the minimum generalized eigenvalue of L_{G_S} with respect to $\frac{1}{\mathrm{Vol}(S)} L_{K_S}$. Equivalently, the spectral gap λ_S is the largest real λ such that $L_{G_S} \succeq \frac{\lambda}{\mathrm{Vol}(S)} L_{K_S}$. The star graph $\mathrm{Star}^{(r)}$, rooted at a vertex $r \in V$, is the graph consisting of the $n-1$ edges of the form (r, i) for all $i \in V$, each with weight d_i. We associate to a solution $M \in \Delta_n$ of (8) two weighted graphs, $G[M]$ and $\mathrm{Star}^{(r)}[M]$, defined by their Laplacians:

$$L_{G[M]} = \sum_{(i,j) \in E} M_{ij} L_{ij} \text{ and } L_{\mathrm{Star}^{(r)}[M]} = \sum_{i \in V} d_i \, M_{ii} L_{ri}.$$

Cheeger's inequality: An important result in spectral graph theory is Cheeger's inequality (Chung, 1997) that relates the conductance of a graph with the spectrum of its Laplacian. An equivalent statement for the subgraph conductance that follows from Cheeger's inequality and relates the spectral gap of G_S to the conductance of G_S can be written as follows.

Theorem 3.1 (Cheeger's Inequality)**.** *For $S \subseteq V$, $\frac{\lambda_S}{2} \leq \phi_{G_S} \leq \sqrt{2\lambda_S}$.*

The right-hand side of this inequality is proved by rounding the generalized eigenvector of L_{G_S} associated with λ_S to a low-conductance cut by using the following lemma.

Lemma 3.1. *Let $y \geq 0$ and $y_r = 0$. Assume that $y^T L_{G_S} y < \lambda \sum_{i \in S} d_i y_i^2$. Then, there exists $\tau > 0$ such that the sweep cut $L_\tau = \{i \in S : y_i \geq \tau\}$ of vector y has $\phi_{G_S}(L_\tau) < \sqrt{2\lambda}$.*

3.2. Relaxing the Conductance Requirement

Our proposed relaxation of the integral conductance constraint (4) with parameter γ is the following:

$$L_{G[M]} \succeq \frac{\gamma^2}{2} L_{\mathrm{Star}^{(r)}[M]} \tag{9}$$

To see how this relaxes the integral constraint, take $M_S = \frac{1}{|S|} 1_S 1_S^\top$ to be an integral solution corresponding to a subset $S \subseteq V$. We have:

$$L_{G[M_S]} = \frac{1}{|S|} L_{G_S} \quad \text{and} \quad L_{\mathrm{Star}^{(r)}[M_S]} = \frac{1}{|S|} L_{\mathrm{Star}_S^{(r)}}.$$

Then, our proposed constraint becomes $L_{G_S} \succeq \frac{\gamma^2}{2} L_{\mathrm{Star}_S^{(r)}}$. We now show that this constitutes a relaxation of constraint (4). This can be seen as a variant of Cheeger's inequality for our relaxed notion of conductance in (9). The proof of the following theorem appears in the appendix.

Theorem 3.2. *For $S \subseteq V$, if, for all $T \subseteq S, r \notin T$ and $\phi_{G_S}(T) \geq \gamma$, then $L_{G_S} \succeq \frac{\gamma^2}{2} L_{\mathrm{Star}_S^{(r)}}$.*

Moreover, for a candidate integral solution M_S, in the same way as the integral constraint lower bounds the conductance of the induced subgraph G_S, our relaxation can be shown to lower bound the spectral gap of G_S. This result is a simple consequence of Schur's complementation.

Lemma 3.2. *For $S \subset V$ with $r \in S$, let $y = (\mathrm{Vol}(S) - d_r)\, e_r - \sum_{j \in S, j \neq r} d_j e_j$. Then, $L_{\mathrm{Star}_S^{(r)}} = \frac{1}{\mathrm{Vol}(S)} \left[L_{K_S} + yy^\top \right]$.*

Applying this lemma to (9), we observe that our constraint, applied to an integral solution M_S, implies a lower bound of $\frac{\gamma^2}{2}$ on the induced spectral gap λ_S through the inequality

$$L_{G_S} \succeq \frac{\gamma^2}{2} L_{\mathrm{Star}_S^{(r)}} \succeq \frac{\gamma^2}{2\,\mathrm{Vol}(S)} L_{K_S}.$$

Finally, we prove that if any feasible candidate M that satisfies (9) is rounded to a subset S in a certain way, then the connectivity of subgraph G_S is ensured. This result shows that the inequality constraint (9) is sufficent to ensure connectivity in our framework.

Theorem 3.3. *For any $\gamma > 0$ and $M \succeq 0$ that satisfies (9), the subgraph $G_{\hat{S}}$ for induced subset $\hat{S} = \{i \in V : M_{ii} > 0\}$ is connected.*

Proof of Theorem 3.3 is in the appendix. It follows from an alternative formulation of constraint (9) based on effective resistance in electrical networks.

3.3. Primal and Dual Formulations

In this subsection, we study the dual form of our relaxation, which will be important in designing an efficient iterative algorithm. We start by introducing some shorthand notation for its constraints. Let $Q_\gamma(M) \succeq 0$ be our relaxed connectedness constraint, i.e., $Q_\gamma(M) = L_{G[M]} - \frac{\gamma^2}{2} L_{\mathrm{Star}^{(r)}[M]}$. Then our relaxation is given by

$$\max_{\substack{M \in \Delta_n, M \geq 0, \\ Q_\gamma(M) \succeq 0}} C \cdot M. \tag{10}$$

We now write the SDP dual of our relaxation. We consider a scalar α as the Lagrange multiplier corresponding to constraint $I \cdot M = 1$, and matrices $Y, Z \in \mathbb{R}^{n \times n}$ corresponding to $Q_\gamma(M) \succeq 0$ and $M \geq 0$ respectively. Let $P_\gamma(Y) = \sum_{(i,j) \in E} (L_{ij} \cdot Y) e_{ij} - \gamma \sum_{i \in V} d_i (L_{ri} \cdot Y) e_{ii}$ be the transpose of the constraint Q_γ, i.e., $P_\gamma(Y) \cdot M = Q_\gamma(M) \cdot Y$. The dual can then be written as:

$$\begin{aligned} \min \quad & \alpha \\ & C + P_\gamma(Y) + Z \preceq \alpha I \\ & \alpha \geq 0, Y \succeq 0, Z \geq 0 \end{aligned}$$

An intuitive interpretation for this dual follows from considering $P_\gamma(Y)$ as a matrix of gains to be added to the objective C as to force the primal solution M towards feasibility. In particular, $P_\gamma(Y)$ establishes a gain of $L_{ij} \cdot Y$ for including edge $\{i, j\}$ in the primal solution and a cost of $L_{ri} \cdot Y$ for including vertex i in the primal solution. Naturally, vertices are more expensive the further they are from the root r in the dual solution and edges are more beneficial if they bridge longer distances in the dual.

Distinctive Properties of Our Relaxation We wish to highlight two important (and rare) structural properties of our relaxation. The first property relates to the form of the dual. We have $C \geq 0$, by definition for the elevated mean problem with nonnegative signal values and with high probability for correlation detection. Then the term $C + P_\gamma(Y)$ in the dual constraint is the sum of a nonnegative matrix plus a diagonal matrix. By the Perron-Frobenius Theorem, the top eigenvector of this matrix has nonnegative components, allowing us to assume that $Z = 0$ wlog. We will use the same reasoning in the next section to show that we do not need to explicitly enforce the n^2 element-wise non-negativity constraints corresponding to $M \geq 0$, as our dual formulation will automatically yield such solutions. This is a great advantage of our relaxation as enforcing the $M \geq 0$ constraints is known to be a computational roadblock to the efficient solution of SDP relaxations of $\{0, 1\}$-integral problems.

The second property has a similar flavor, but it concerns the primal optimal solution. It is captured by the following theorem, which is proved in the appendix.

Theorem 3.4. *When $C \geq 0$, the relaxation 10 always has an optimal solution of rank-1. Moreover, any higher rank solution M can be turned into a rank-1 solution $mm^\top$ such that $M_{ii} = m_i^2$.*

The fact that a rank-1 solution is a remarkable property for a SDP relaxation, making the fractional solution easier to visualize and hopefully easier to round to integral.

Future work: In this work, we did not perform a theoretically study of the approximation guarantees achievable in rounding our relaxation to an integral solution in the worst-case. The rank-1 property of the optimal solution should be useful in this pursuit. At the same time, we believe that additional constraints may be required to obtain meaningful approximation guarantees in the worst-case. This is an interesting direction for future work.

Statistical Bounds: We omit developing statistical analysis of the proposed approach for subgraph detectability in this paper for lack of space (see (Aksoylar, 2017) for analysis for simple graphs). We can derive statistical guarantees for grid graphs similar to (Qian & Saligrama, 2014) based on analyzing primal and dual values. In particular the primal provides a bound on the value of the positive hypothesis (anomaly), while a feasible solution to the dual provides an upper-bound of the value for null hypothesis. Detectability bounds for elevated mean follows by comparing the primal value with the dual.

4. Mirror Descent on SDPs

We first consider a modification of our original SDP by adding the slack variable $s \geq 0$. For some fixed margin value $\delta \geq 0$ we write

$$\max_{M\in\Delta_n, s} \quad C\cdot M - s \quad \text{s.t.} \quad Q_\gamma(M) + s\cdot\delta\cdot D \succeq 0. \quad (11)$$

Recalling that D is the degree matrix of G, the last term provides a measure of how violated the SDP constraint is. For now, we fix δ as a parameter of our algorithm. We discuss choices of δ at the end of this section.

Introducing the Lagrange multiplier $Y \succeq 0$ corresponding to the constraint $\frac{1}{\delta}\cdot Q_\gamma(M) + sD \succeq 0$, we then obtain the saddle point problem

$$\max_{M\in\Delta_n, s}\min_{Y\succeq 0} \quad C\cdot M - s + Y\cdot(\frac{1}{\delta}\cdot Q_\gamma(M) + sD),$$

from which we obtain the dual

$$\min_{Y\in\Delta_n^D} f(Y), \text{ where } f(Y) = \max_{\substack{M\in\Delta_n,\\ M\geq 0}} \left(C + \frac{1}{\delta}\cdot P_\gamma(Y)\right)\cdot M,$$

and we defined Δ_n^D to be the D-spectrahedron $\{X \succeq 0 : D\cdot X = 1\}$. For this dual optimization over Y we utilize the mirror descent method, which is the optimal optimization algorithm for non-smooth functions in the black-box model. We refer the reader to Section 5.2 of (Ben-Tal & Nemirovski, 2015) for more details on mirror descent and its application in the spectahedron setup. For the purposes of this section, we simply state the following theorem, which is a simple consequence of Theorem 5.2.1 in (Ben-Tal & Nemirovski, 2015).

Theorem 4.1. *Let f be a convex function over the spectrahedron Δ_n^D such that $\|D^{-1/2}\nabla_Y f(Y) D^{-1/2}\| \leq L$ for all $Y \in \Delta_n^D$. For a parameter $\eta = \frac{\epsilon}{L^2}$, the mirror descent update takes the following form at iteration t:*

$$Y^{(t)} = \frac{\exp\left(-\eta\cdot D^{-1/2}\left[\sum_{j=0}^{t-1}\nabla_Y f(Y^{(j)})\right]D^{-1/2}\right)}{D\cdot\exp\left(-\eta\cdot D^{-1/2}\left[\sum_{j=0}^{t-1}\nabla_Y f(Y^{(j)})\right]D^{-1/2}\right)}$$

With this update, the algorithm achieves the following performance guarantee, where f^* is the minimum of f:

$$f(Y^T) - f^* \leq L\cdot\sqrt{\frac{\log n}{T}}.$$

To apply mirror descent as described in the previous theorem, we need access to the gradient of f at $Y^{(t)}$. By Danskin's Theorem (Bertsekas et al., 2003), this is given by:

$$\nabla_Y f(Y^{(t)}) = \frac{1}{\delta}\cdot Q_\gamma(M^{(t)}),$$
$$M^{(t)} = \arg\max_{M\in\Delta_n, M\geq 0}(C + \frac{1}{\delta}\cdot P_\gamma(Y^{(t)}))\cdot M.$$

Hence, computation of the gradient requires finding $M^{(t)}$, which plays the role of the primal update at time t. However, this is just the rank-1 matrix given by the projection over the top eigenvector of $C + \frac{1}{\delta}\cdot P_\gamma(Y^{(t)})$, where $M \geq 0$ is once again ensured by Perron-Frobenius. The following lemma provides us with a bound on the Lipschitz parameter L of our objective f. Its straightforward proof appears in the appendix.

Lemma 4.1. *For all $Y \in \Delta_n^D$, we have: $\|D^{-1/2}\nabla_Y f(Y) D^{-1/2}\| \leq \frac{2}{\delta}$.*

With this setting of L, Theorem 4.1 yields the following convergence bound for our mirror descent algorithm.

Theorem 4.2. *Algorithm 1 converges to an ϵ-additive approximation of optimal in $T = O\left(\frac{\log n}{\delta^2\epsilon^2}\right)$ steps.*

Moreover, each iteration consists of computing the top eigenvector of a non-negative matrix and the matrix exponential of a the sum of a Laplacian and a rank-1 term (cf. Lemma 3.2). Thanks to recent breakthrough theoretical results, both of these objects can be approximated sufficiently closely in almost-linear-time (Orecchia et al., 2012; Cohen et al., 2016). In practice, existing iterative solvers, combined with the use of the Johnson-Lindenstrauss Lemma to keep a low-dimensional sketch of the matrix exponential,

Algorithm 1 Mirror Descent Algorithm

```
  **Input:** $C, \delta, r, \gamma, \epsilon$
  **Output:** $\hat{M}$
  $L \leftarrow \frac{2}{\delta}, \;\; \eta \leftarrow \frac{\epsilon}{L^2}$
  $Y^{(0)} \leftarrow \frac{1}{\mathrm{Tr}(D)} I_n, \;\; G^{(0)} \leftarrow 0$
  **for** $t = 1, \ldots, T$ **do**
    $v \leftarrow \mathrm{eig}\left(C + P_\gamma(Y^{(t-1)})\right)$
    $M^{(t)} \leftarrow vv^\top$
    $G^{(t)} \leftarrow G^{(t-1)} + \frac{1}{\delta} \cdot Q_\gamma(M^{(t)})$
    $Y^{(t)} \leftarrow \exp\left(-\eta D^{-1/2} G^{(t)} D^{-1/2}\right)$
    $Y^{(t)} \leftarrow \frac{1}{D \cdot Y^{(t)}} Y^{(t)}$
  **end for**
  $\hat{M} \leftarrow \frac{1}{T} \sum_{t=1}^{T} M^{(t)}$
```

already provide a very efficient computational approach to this problem, as we demonstrate in our experiments.

We formally present the resulting algorithm in Algorithm 1 for our function f, where $\mathrm{eig}(\cdot)$ operator returns the eigenvector corresponding to the largest eigenvalue.

Choosing the margin δ: If $s \geq \frac{\gamma^2}{\delta}$, it is possible to prove that the SDP constraint is trivially satisfied for any M in Δ_n at a cost of s in the objective (which follows from the fact that $L_{\mathrm{Star}^{(r)}} \preceq 2D$). To avoid such trivial solutions, we wish to set the margin δ to be sufficiently small. In particular, we should have $\delta \leq \frac{\gamma^2}{4\epsilon}$. However, from a worst-case point of view, this setting of δ may be insufficient to obtain a solution that can be rounded to a connected subgraph. The choice of δ in this case depends on the rounding procedure used and its sensitivity. As a formal study of the rounding of our relaxation is beyond of the scope of this paper, we cannot provide a definitive setting of δ. Our preliminary calculations show that, from a theoretical point of view, a setting of $\delta = O\left(\frac{\gamma^2}{K}\right)$ should be sufficient for rounding, where K is the size of the optimal set. In practice, we have found that setting δ to be order $O(\gamma^2)$ suffices for most of the examples we considered. This corresponds to a number of iterations that is $O\left(\frac{\log n}{\gamma^4 \epsilon^2}\right)$.

5. Experiments

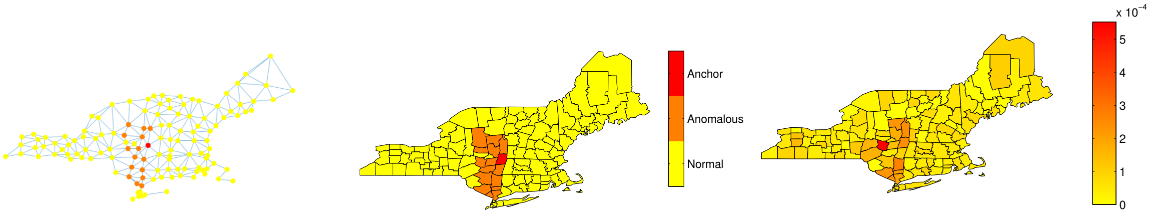
We present experiments on two datasets: a real world geographical network of disease outbreaks and elevated mean detection on very large random geometric graphs. In the former we compare the statistical detection performance of our mirror descent (MD) algorithm with subgraph detection methods from related work. For the latter we demonstrate the scalability of our method on large graphs.

5.1. Disease Outbreak Detection

We consider a geographical map and its corresponding network that are illustrated in Figures 1b and 1a respectively, with 129 nodes representing counties in the northeastern United States and average degree 4.7. The ground truth cluster of 16 nodes for the anomalous case and the chosen anchor node are also illustrated. Following (Patil et al., 2003) and (Qian et al., 2014; Qian & Saligrama, 2014), we consider an elevated mean Poisson formulation for modeling the diseased population, where the number of disease cases y_i for a county i is given by $y_i \sim \mathrm{Poisson}(N_i \lambda_0)$ where N_i is the population of the county, whereas for anomalous counties we have $y_i \sim \mathrm{Poisson}(N_i \lambda_1)$. We consider $\lambda_0 = 5 \times 10^{-5}$ for the base disease rate and different $\frac{\lambda_1}{\lambda_0}$ ratios $\{1.1, 1.3, 1.5\}$ corresponding to different SNR values. As our test statistic we consider the disease rate per person $x_i = \frac{y_i}{N_i}$. One sample realization for the anomaly case with high SNR $\frac{\lambda_1}{\lambda_0} = 4$ appears in Fig. 1c.

To compare the performance with MD as proposed in Algorithm 1, we consider several other methods in the related literature, including the LMI-test (LMIT) method of (Qian & Saligrama, 2014), simulated annealing (SA) of (Duczmal & Assuncao, 2004) and the nearest-ball test (NB), which is a parametric method that scans over nearest-neighbor balls of different sizes for all nodes. For the MD method we consider the optimization value $xx^\top \cdot M$ as the scan statistic, with $T = 100$ iterations, $\eta = 5$ and different γ values to quantify the size and conductance of the anomalous graph. For LMIT we use the same anchor node as MD, anomaly size $|S| = 16$ corresponding to the ground truth and consider scan statistic $x^\top \mathrm{diag}(M)$. We search over a range of values for parameter γ. For SA and NB we consider the test statistic $\frac{\sum_{i \in S} x_i}{\sqrt{|S|}}$. We initialize SA with the result from NB and run for 40 restarts. To quantify detection performance, we threshold the scan statistics given by the algorithms with various threshold values and compute missed detection and false positive rates over a number of samples (50 for MD and LMIT, 25 for SA and NB) generated from both H_0 and H_1. We then compute the area under the curve (AUC) generated by the pairs of missed detection and false positive rates corresponding to threshold values.

Sensitivity of MD to the choice of γ: We first investigate the sensitivity of detection performance of MD to γ, which serves the purpose of parameterizing the internal conductance of candidate subgraphs. We note that unlike (Qian et al., 2014; Qian & Saligrama, 2014), we do not explicitly specify or search over different cluster sizes, but size information is also implicitly incorporated in γ. We run MD on the range of values 0.01 to 5 in 10 logarithmic intervals. We illustrate the obtained AUC values for different SNR's in Figure 2. We observe that while optimal γ values differ slightly with different SNR levels, the 0.3–0.7 value range is mostly optimal in all cases. This is in accordance with our expectations, since the size and conductance of the ground truth anomalies do not change.



(a) Graph representation of the county graph.

(b) Ground truth anomalous cluster and anchor node.

(c) Sample realization of disease rates for the anomalous case.

Figure 1. Disease outbreak detection in northeastern United States counties.

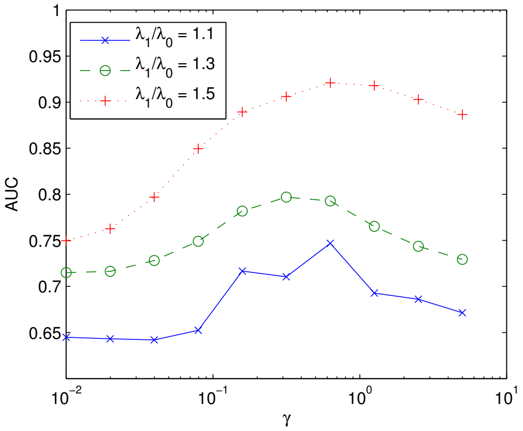


Figure 2. Performance of MD algorithm for different γ values.

Comparison to related methods: We also compare AUC performance of MD to aforementioned methods and tabulate the results in Table 1. For MD we use a γ value of 0.7 and for LMIT we use a γ value of 0.3 which we observed to perform best empirically. We see that MD performs relatively similar to LMIT, with better performance at some SNR levels. This is expected since the LMI connectivity constraints in both methods are very similar, even though the relaxation to the space of matrices M differ. On the other hand SA and NB perform worse, with SA not significantly improving upon the results of NB. It is also notable that the performance of these three methods seem low when compared to their performance in (Qian & Saligrama, 2014), which can be partially explained by the different scan statistics used: the Poisson likelihood test in contrast to the simpler linear form we specified above to be better in line with the scan statistic of MD. It is also possible that the performance of SA can be improved with a larger number of restarts. We also provide the average runtime for the recovery methods for a single set of measurements in Table 1, where the experiments were run on MATLAB on a computer with an Intel i5 4590 processor.

5.2. Random Geometric Graphs

We also conducted experiments on simulated geometric graphs to demonstrate the scalability of our method. For this, we generated n points uniformly on the hypercube $[-1,1]^D$ and created approximate k-NN graphs using the ANN library (Arya et al., 1998). We generated anomalous clusters by determining points that fall in hyperellipsoids centered at the origin of the space. We consider different hyperellipsoid axes lengths that correspond to different internal subgraph conductance.

AUC	λ_1/λ_0			Runtime
	1.1	1.3	1.5	
MD	0.74	0.79	0.92	0.8s
LMIT	0.65	0.81	0.86	3s
SA	0.57	0.67	0.72	∼3m
NB	0.57	0.67	0.68	5s

Table 1. AUC performance of algorithms with different SNRs.

Memory and run-time scalability: For very large graphs with n nodes, storing on memory and operating on non-sparse $n \times n$ matrices present major problems for computational feasibility. While for primal variable M we work directly with vectors v in Alg. 1, we also consider a lower-rank approximation scheme for representing dual variable Y and an approximate computation for the matrix exponential. We define an $n \times k$ matrix Y_k such that we have the update $Y_k^{(t)} \propto \exp\left(-\eta Q_\gamma(M^{(t)}\right) W^{(t)}$, where $W^{(t)}$ is an $n \times k$ matrix with IID elements $\mathcal{N}(0, \frac{1}{k})$. With this definition we have the approximation $Y^{(t)} \approx Y_k^{(t)} Y_k^{(t)\top}$ exploiting the Johnson-Lindenstrauss lemma, for $Y_k^{(t)}$ normalized appropriately. We then utilize the Leja method (Caliari et al., 2016) to directly compute the action of the matrix exponential on vectors. We again consider elevated mean detection with $y_i \sim \text{Poisson}(\lambda_0)$ for non-anomalous nodes and $y_i \sim \text{Poisson}(\lambda_1)$ otherwise. We specifically consider 10-NN graphs with parameters $n = 10^4$ and $D = 3$. We consider two types of anomalous clusters: "thick" cluster as a sphere with radius r and "thin" cluster as an ellipsoid with radii $(8r, r, r)$, where r is chosen such that on average the clusters would contain $K = 40$ nodes.

Performance for different conductance anomalies and comparison: We investigate the AUC performance of MD and compare to the NB scan statistic over 40 sample realizations of measurements, for different SNR ratios λ_1/λ_0 in Table 2 where we fix $\lambda_0 = 100$. Due to the memory and run-time scaling of SA and LMIT it was not feasible to apply these methods to the large graphs. For MD, we chose $\frac{\gamma^2}{2} = 10^{-3}$ for the thick cluster and 5×10^{-4} for the thin one. We chose a random node in the cluster as the anchor, $k = 10$ vectors for approximating Y and ran the algorithm for $T = 300$ iterations. From the results we observe that MD and NB perform similarly on thick clusters. This is expected since a spherical cluster is the optimal scenario for NB, whereas MD still considers different shaped and sized clusters for the given gamma. However for the thin cluster we observe that MD improves upon NB significantly as expected, especially for higher SNR values. We also note that each iteration of MD takes about 1s and empirically scales linearly with n (as we demonstrate in the next set of experiments), where we applied the method for graphs with up to 10^5 nodes.

AUC		λ_1/λ_0		
		1.1	1.3	1.5
Thick	MD	0.71	0.93	0.99
	NB	0.70	0.92	0.96
Thin	MD	0.70	0.92	0.99
	NB	0.68	0.90	0.92

Table 2. AUC performance of MD and NB with different SNR values and cluster shapes.

Performance for graph sizes and iterations: We also investigate the effect of graph size n in conjunction with the number of iterations T on the accuracy as quantified by the detection AUC. We again consider random geometric graphs generated with parameters in the previous experiments for $\frac{\lambda_1}{\lambda_0} = 1.3$, vary graph size n from 4000 to 10000 in 2000 increments and consider ellipsoidal anomalies of radii $(4r, r, r)$ encapsulating approximately $K = 40$ nodes with $\frac{\gamma^2}{2} = 10^{-3}$. We plot the AUC performance vs. number of iterations for different graph sizes in Figure 3. First, we observe that detection accuracy deteriorates for larger graph sizes as expected. Moreover, the rate of increase in the accuracy with the increasing number of iterations T does not seem to change too much for different sizes n, which lends empirical support to Theorem 4.1 regarding the sublinear relationship between accuracy and n. We also plot the average run-time per iteration vs. graph size with standard deviation error bars in Fig. 4, which illustrates the approximately linear scaling of run-time per iteration as discussed in Sec. 4.

Performance vs. Anomaly size: We investigate detection performance for different anomalous cluster sizes $K = |S|$ in Figure 5. We again consider a fixed SNR $\frac{\lambda_1}{\lambda_0} = 1.3$ for $n = 10000$ and ellipsoidal anomalies of radii $(4r, r, r)$, with varying r such that K varies between 20 and 80. We performed $T = 300$ iterations and used the same value of $\frac{\gamma^2}{2} = 10^{-3}$ for all K, as different values did not result in a significant accuracy improvement in our cross-validation experiments. This in turn confirms the robustness of the choice of γ which we also observed for the county graph dataset. As seen the detection accuracy increases rapidly with K for a fixed per-node SNR. This is in line with the theoretical scaling behavior in (Qian & Saligrama, 2014).

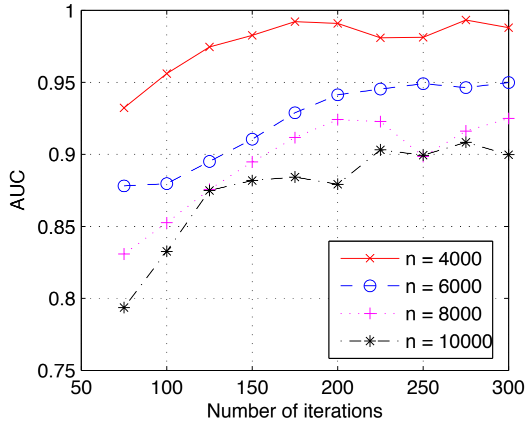


Figure 3. AUC performance for different graph sizes n for differing number of total iterations T.

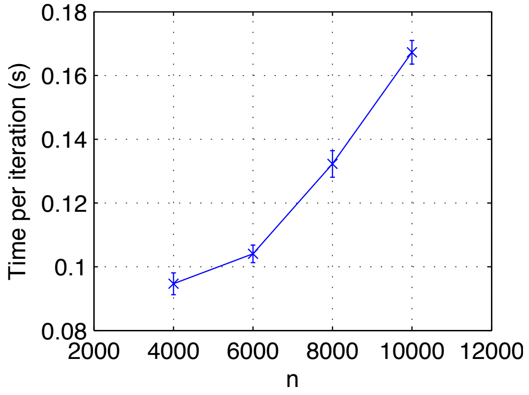


Figure 4. Run-time per iteration vs. graph size n.

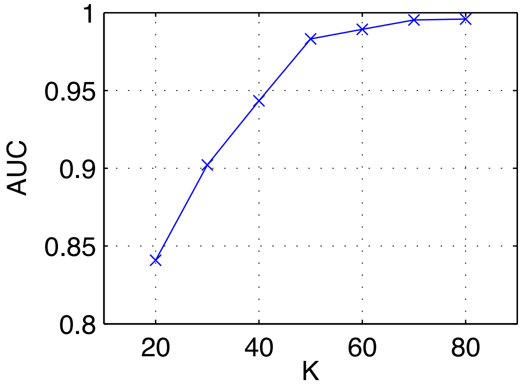


Figure 5. AUC performance for different anomaly sizes K.

References

Addario-Berry, Louigi, Broutin, Nicolas, Devroye, Luc, Lugosi, Gábor, et al. On combinatorial testing problems. *The Annals of Statistics*, 38(5):3063–3092, 2010.

Aksoylar, Cem. *Discovery of Low-Dimensional Structure in High-Dimensional Inference Problems*. PhD thesis, Boston University, Boston, MA, 2017.

Arias-Castro, Ery, Candès, Emmanuel J, Helgason, Hannes, and Zeitouni, Ofer. Searching for a trail of evidence in a maze. *The Annals of Statistics*, pp. 1726–1757, 2008.

Arias-Castro, Ery, Bubeck, Sébastien, Lugosi, Gábor, et al. Detection of correlations. *The Annals of Statistics*, 40(1):412–435, 2012.

Arya, Sunil, Mount, David M., Netanyahu, Nathan S., Silverman, Ruth, and Wu, Angela Y. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *J. ACM*, 45(6):891–923, November 1998. ISSN 0004-5411. doi: 10.1145/293347.293348.

Ben-Tal, Aharon and Nemirovski, Arkadi. Lectures on modern convex optimization. 2015.

Bertsekas, Dimitri P, Nedic, Angelia, and Ozdaglar, Asuman E. *Convex Analysis and Optimization*. Athena Scientific, 2003. ISBN 1-886529-45-0.

Caliari, Marco, Kandolf, Peter, Ostermann, Alexander, and Rainer, Stefan. The Leja method revisited: Backward error analysis for the matrix exponential. *SIAM Journal on Scientific Computing*, 38(3):A1639–A1661, 2016. doi: 10.1137/15M1027620.

Chung, F. R. K. *Spectral Graph Theory*, volume 92. American Mathematical Society, 1997.

Cohen, Michael B., Kelner, Jonathan A., Peebles, John, Peng, Richard, Rao, Anup, Sidford, Aaron, and Vladu, Adrian. Almost-linear-time algorithms for markov chains and new spectral primitives for directed graphs. *CoRR*, abs/1611.00755, 2016. URL `http://arxiv.org/abs/1611.00755`.

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

Glaz, Joseph, Naus, Joseph, and Wallenstein, Sylvan. *Scan Statistics*. Springer Science & Business Media, 2001.

Johnson, David S., Minkoff, Maria, and Phillips, Steven. The prize collecting steiner tree problem: Theory and practice. In *Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '00, pp. 760–769, Philadelphia, PA, USA, 2000. Society for Industrial and Applied Mathematics. ISBN 0-89871-453-2.

Kulldorff, Martin, Huang, Lan, Pickle, Linda, and Duczmal, Luiz. An elliptic spatial scan statistic. *Stat. Med.*, 25(22):3929–3943, 2006.

Orecchia, Lorenzo, Sachdeva, Sushant, and Vishnoi, Nisheeth K. Approximating the exponential, the Lanczos method and an Õ(m)-time spectral algorithm for balanced separator. In *Proceedings of the 44th symposium on Theory of Computing - STOC '12*, volume 2, pp. 1141–1160, New York, NY, USA, November 2012. ACM Press. doi: 10.1145/2213977.2214080.

Patil, GP, Taillie, C, et al. Geographic and network surveillance via scan statistics for critical area detection. *Statist. Sci.*, 18(4): 457–465, 2003.

Priebe, Carey E, Conroy, John M, Marchette, David J, and Park, Youngser. Scan statistics on Enron graphs. *Computational & Mathematical Organization Theory*, 11(3):229–247, 2005.

Qian, J. and Saligrama, V. Efficient minimax signal detection on graphs. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2708–2716. Curran Associates, Inc., 2014.

Qian, J., Saligrama, V., and Chen, Y. Connected sub-graph detection. In *Proc. of the Seventeenth Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, pp. 796–804, Reykjavik, Iceland, April 2014.

Sharpnack, James, Rinaldo, Alessandro, and Singh, Aarti. Detecting anomalous activity on networks with the graph Fourier scan statistic. *IEEE Transactions on Signal Processing*, 64(2): 364–379, 2016.

Vishnoi, Nisheeth K. Laplacian solvers and their algorithmic applications. *Theoretical Computer Science*, 8(1-2):1–141, 2012.

Wu, Nannan, Chen, Feng, Li, Jianxin, Zhou, Baojian, and Ramakrishnan, Naren. Efficient nonparametric subgraph detection using tree shaped priors. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Acknowledgements: This material is based upon work supported in part by NSF Grants CCF: 1320566, NSF Grant CNS: 1330008 NSF CCF: 1527618, the U.S. DHS, Science and Technology Directorate, Office of University Programs, under Grant Award 2013-ST-061-ED0001, and by ONR contract N00014-13-C-0288. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the social policies, either expressed or implied, of the NSF, U.S. DHS, ONR or AF.