A Path Algorithm for Localizing Anomalous Activity in Graphs

James Sharpnack
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA, USA
jsharpna@gmail.com

***Abstract*—The localization of anomalous activity in graphs is a statistical problem that arises in many applications, such as network surveillance, disease outbreak detection, and activity monitoring in social networks. We will address the localization of a cluster of activity in Gaussian noise in directed, weighted graphs. We develop a penalized likelihood estimator (we call the relaxed graph scan) as a relaxation of the NP-hard graph scan statistic. We review how the relaxed graph scan (RGS) can be solved using graph cuts, and outline the max-flow min-cut duality. We use this combinatorial duality to derive a path algorithm for the RGS by solving successive max flows. We demonstrate the effectiveness of the RGS on two simulations, over an undirected and directed graph.**

I. Introduction

Classically, the detection and identification of anomalies has focused on identifying rare behaviors and aberrant bursts in activity over a single data source or channel. With the advent of large surveillance projects, social networks, and mobile computing, statistics needs to comprehensively address the detection of anomalous activity in graphs. In reality, very little is known about the detection and localization of activity in graphs, despite a variety of real-world applications such as activity detection in social networks, network surveillance, disease outbreak detection, biomedical imaging, sensor network detection, gene network analysis, environmental monitoring and malware detection. Recent theoretical contributions in the statistical literature[1], [2] have detailed the inherent difficulty of such combinatorial statistical problems but have positive results only under restrictive conditions on the graph topology.

In machine learning and computer vision, Markov random fields (MRF) with Ising priors have been used to model activation patterns that are consistent with some graph structure. In this Bayesian setting, the maximum a-posteriori (MAP) estimate has dominated the research, due to its computational feasibility and success in computer vision applications. In this paper, we propose a penalized likelihood estimator that takes a similar form to the MRF MAP estimate. We develop a path algorithm for this estimator, as the regularization parameter varies, which can be solved with successive maximum flow computations.

A. Problem Setup

Consider a connected, possibly weighted, directed graph G defined by a set of vertices V ($|V| = p$) and directed edges E ($|E| = m$) which are ordered pairs of vertices. We will let $u \to v$ denote an edge from the vertices u to v. Furthermore, the edges may be assigned weights, $\{W(e)\}_{e \in E}$, that determine the relative strength of the interactions of the adjacent vertices. In the graph-structured normal means problem, we observe one realization of the random vector

$$\mathbf{y} = \mathbf{x} + \boldsymbol{\xi}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^p$, $\boldsymbol{\xi} \sim N(0, \mathbf{I}_{p\times p})$. The structure of activation pattern $\mathbf{x}$ is determined by the graph G. Specifically, we assume that there is a parameter ρ (possibly dependent on p) such that the class of graph-structured activation patterns $\mathbf{x}$ is given as follows.

$$\mathcal{C} = \left\{ C \subseteq V : \frac{\text{out}(C)}{|C|} \le \rho \right\}$$

$$\mathcal{X} = \{\mathbf{x} : \mathbf{x} = \mu \mathbf{1}_C, \mu > 0, C \in \mathcal{C}\}$$

Here $\text{out}(C) = \sum_{u \to v \in E} W(u \to v) I\{u \in C, v \in \bar{C}\}$ is the weight of edges leaving the cluster C. In other words, the set of activated vertices C have a small *cut size* in the graph G. Notice that the model (1) is equivalent to the more general model in which $\mathbb{E}\xi_i^2 = \sigma^2$ with σ known. As a notational convenience, if $\mathbf{z} \in \mathbb{R}^p$ and $C \subseteq [p]$, then we denote $\mathbf{z}(C) = \sum_{v \in C} z_v$. Throughout the study, let the edge-incidence matrix of G be $\nabla \in \mathbb{R}^{m \times p}$ such that for $v \to u \in E$, $\nabla_{v\to u, v} = -W(v \to u)$, $\nabla_{v \to u, u} = W(v \to u)$ and is 0 elsewhere.

B. Related Work

There have been several approaches to signal processing over graphs. Markov random fields (MRF) provide a succinct framework in which the underlying signal is modeled as a draw from an Ising or Potts model [3], [4]. A similar line of research is the use of kernels over graphs, which began with the development of diffusion kernels [5], and was extended through Green's functions on graphs [6]. In this study, we develop a path algorithm for a localization of structured signals, which is similar to the work of [7], [8].

The estimation of the mean of a Gaussian has served as a canonical problem in nonparametric statistics. When the mean is assumed to be sparse, asymptotic minimaxity has been established for thresholding and false-discovery rate based estimators [9], [10]. When the mean vector is assumed to lie within an ellipsoid then Pinsker's estimator has been shown to be asymptotically optimal as σ approaches 0 [11].

In spatial statistics, it is common, when searching for anomalous activity to scan over regions in the spatial domain testing for elevated activity[12], [13]. There have been scan statistics proposed for graphs, most notably the work of [14] in which the authors scan over neighborhoods of the graphs defined by the graph distance. Other work has been done on the theory and algorithms for scan statistics over specific graph models, but are not easily generalizable to arbitrary graphs [15], [1]. More recently, it has been found that scanning over all well connected regions of a graph can be computationally intractable, and so approximations to the intractable likelihood-based procedure have been studied [16], [17]. We follow in this line of work, with a relaxation to the intractable restricted likelihood maximization.

II. Method: The Flow Path Algorithm

A. Relaxed Graph Scan

The fundamental difficulty of obtaining an accurate estimate of $C^\star$ is that the true cluster of activation may be any of $\mathcal{C}$. It is known by Karger's cut counting theorem[18] that $|\mathcal{C}|$ can be exponential in ρ and p. Hence, a brute-force scan of all of the elements of $\mathcal{C}$ is infeasible.

It is instructive, when faced with a class of probability distributions, indexed by subsets $\mathcal{C} \subseteq 2^{[p]}$, to think about what techniques we would use if we knew the correct cluster $C = C^\star \in \mathcal{C}$ (which is often called oracle information). The maximum likelihood estimator for $\mathbf{x}$ would thus be $\frac{(\mathbf{1}_C^\top \mathbf{y})_+}{|C|}\mathbf{1}_C$ and the log-likelihood would be proportionate to $(\mathbf{y}^\top \mathbf{1}_C)_+^2/|C|$. Because, of course, we would not be supplied with the true cluster $C^\star$, we must choose among $\mathcal{C}$. We may then maximize the likelihood under the constraints to obtain the *graph scan statistic*,

$$\hat{s} = \max \frac{\mathbf{y}^\top \mathbf{1}_C}{\sqrt{|C|}} \text{ s.t. } \frac{\text{out}(C)}{|C|} \leq \rho$$

Notice that there is no guarantee that the program above is computationally feasible. If we were to add the constraint that $|C| \leq n/2$ then determining feasibility would be NP-hard because the constraint $\text{out}(C)/|C|$ corresponds to the sparsest cut (a known NP-hard combinatorial program). With this in mind we consider the following relaxation,

$$\begin{aligned}\hat{s} &\leq \max_{k\in[p]} \frac{1}{\sqrt{k}} \max_{C\subseteq V} \mathbf{y}^\top \mathbf{1}_C \text{ s.t. } \text{out}(C) \leq \rho k, |C| \leq k \\ &\leq \max_{k\in[p]} \frac{1}{\sqrt{k}} \max_{C\subseteq V} \mathbf{y}^\top \mathbf{1}_C \text{ s.t. } \text{out}(C) + \rho|C| \leq 2\rho k \\ &\leq \max_{k\in[p]} \frac{1}{\sqrt{k}} \max_{C\subseteq V} \mathbf{y}^\top \mathbf{1}_C - \nu^{-1}(\text{out}(C) + \rho|C|) + \nu^{-1} 2\rho k\end{aligned}$$

by weak duality with dual parameter $\nu^{-1} > 0$. We have thus established that we can relax the graph scan statistic program to a program with a submodular objective (the above display is a cut size with a modular term). Moreover, because we would like to reconstruct $C^\star$ and not in the value of $\hat{s}$, it suffices to solve the following penalized negative log-likelihood minimization,

$$\min(\rho\mathbf{1} - \nu\mathbf{y})(C) + \text{out}(C) \text{ s.t. } C \subseteq V \tag{2}$$

The result of this primal problem will be our estimator $\hat{C}$, the *relaxed graph scan*. Notice that the objective is a modular term $(\rho\mathbf{1} - \nu\mathbf{y})(C)$ and a out-degree $\text{out}(C)$, which is solvable by s-t graph cuts, as we will see below. This is similar to the MRF MAP estimators for binary activation patterns, which take the same form. (Such programs are called graph-representable in [19].) Thus, by the min-cut max-flow theorem the MRF MAP estimate, and our estimator (2), can be obtained by computing a maximum flow.

B. The Flow Path Algorithm

Dual to our primal program (2), is a max-flow linear program, due to the min-cut max-flow theorem [20]. First we form an augmented graph over the vertices $V \cup \{s,t\}$, where s and t are the source and sink respectively. The edges are contained within $E \cup \{s \to v\}_{v\in V} \cup \{v \to t\}_{v\in V}$ with capacitances

$$\mathbf{c}(\nu) = \begin{cases} c(s \to v) = (\rho - \nu y_v)_-, & v \in V \\ c(v \to t) = (\rho - \nu y_v)_+, & v \in V \\ c(u \to v) = W(u \to v), & u \to v \in E \end{cases}$$

A flow over this graph, is a function mapping edges to $\mathbb{R}$ that satisfies the feasibility conditions below. The dual max-flow program is

$$\max \sum_{v\in V} f(s \to v) \quad \text{s.t.} \quad \mathbf{f} \succeq \mathbf{0}, \mathbf{f} \preceq \mathbf{c}(\nu), \nabla^\top \mathbf{f} = \mathbf{0} \tag{3}$$

where ∇ is the edge incidence matrix for the graph G. One can form the residual flow graph from a feasible $\mathbf{f}$ and $\mathbf{c}$ by making a new graph with capacitances $\mathbf{c} - \mathbf{f}$. Algorithms such as Ford-Fulkerson and Edmonds-Karp iteratively find paths from s to t in the residual graph and increase $\mathbf{f}$ along that path.

We can then construct a set C from $\mathbf{f}$ by including any vertex that is connected to s by a path in the residual flow graph. Specifically, $\mathbf{f}$ is a solution to the max-flow program (3), if and only if

(a) $\mathbf{f}$ is feasible, i.e. $\mathbf{f} \succeq \mathbf{0}, \mathbf{f} \preceq \mathbf{c}(\nu), \nabla^\top \mathbf{f} = \mathbf{0}$
(b) $\sum_{v\in V} f(s \to v) - (\rho - \nu y_v)_- = (\rho\mathbf{1} - \nu\mathbf{y})(C) + \text{out}(C)$

If (a) and (b) hold then C is a solution to (2).

The intuition behind the FlowPath algorithm is that we can construct a solution path for (3) that is piecewise linear, where the slope is calculated from max-flow over what we call the gradient graph resulting in the gradient flow, $\partial\mathbf{f}$. This gradient flow calculation occurs at knots where the slope of the solution $\mathbf{f}$ changes. We calculate the gradient graph in two stages: calculating a negative gradient flow to accommodate capacitances that are decreasing, and calculating a positive gradient flow in order to maintain that the solution indeed is the maximizer of (3). For some edges that enter the sink, t, the capacitances are decreasing as ν is increasing. If the current flow $\mathbf{f}$ meets these capacity constraints then the flow across these edges must decrease to maintain feasibility. Hence, we construct a negative gradient flow such that when we subtract this flow it compensates for the decrease in capacity. The capacities for the negative gradient flow are given below,

$$\begin{array}{ll} c_-(u \to t) = y_u, & u \in V, 0 < \rho - \nu y_u = f(u \to t) \\ c_-(s \to u) = \infty, & u \in V, 0 < f(u \to t) \\ c_-(u \to v) = \infty, & u, v \in V, 0 < f(u \to v) \\ c_-(u \to v) = 0, & \text{otherwise} \end{array} \tag{4}$$

We will thus solve the max flow over $\mathbf{c}_-$, obtaining $\partial \mathbf{f}_-$, and then impose that for each increase of ν by $\partial\nu$ we will decrease the flow by $\partial \mathbf{f}_-$. We now add a positive gradient flow to make these increments maximize the objective.

$$\begin{array}{ll} c_+(s \to u) = y_u + \partial f_-(s \to u), & 0 < \nu y_u - \rho = f(s \to u) \\ c_+(s \to u) = \infty, & 0 < \nu y_u - \rho < f(s \to u) \\ c_+(u \to t) = \infty, & f(u \to t) < \rho - \nu y_u \\ c_+(u \to v) = \infty, & f(u \to v) < W(u \to v) \\ c_+(u \to v) = \partial f_-(u \to v), & f(u \to v) = W(u \to v) \\ c_+(u \to v) = 0, & \text{otherwise} \end{array} \tag{5}$$

The intuition behind these choices of capacitances is that if at a specific edge the capacitance constraint is not met then it provides no constraint on the infinitesimal change in flow (hence ∞ capacitances). We are guaranteed to not have a path with all infinite capacitances because then there would be a path in the residual graph at the knot and by induction this will not happen.

Algorithm 1 Flow Path

```
Initialize $\nu = \rho / \max\{y_i\}_{i=1}^p$, $\mathbf{f} = \mathbf{0}$, $\mathcal{P} = \emptyset$
while $\nu < \infty$ do
    Form the negative derivative graph with the vertices $V \cup \{s, t\}$ and capacitances according to (4).
    Let $\partial f_-$ be a solution to the max flow over $\mathbf{c}_-$.
    Form the positive derivative graph with the vertices $V \cup \{s, t\}$ and capacitances according to (5).
    Let $\partial f_+$ be a solution to the max flow over $\mathbf{c}_+$.
    Set $\partial\nu = \max\{\partial\nu > 0 : \mathbf{f} + \partial\nu(\partial\mathbf{f}_+ - \partial\mathbf{f}_-) \text{ is feasible}\}$
    $\nu = \nu + \partial\nu$
    $\mathbf{f} = \mathbf{f} + \partial\nu(\partial\mathbf{f}_+ - \partial\mathbf{f}_-)$
    Add $\mathcal{P} \leftarrow (\nu, \mathbf{f})$.
end while
```

In order to find $\partial\nu$, we look for the first constraint that is violated as we add the gradient flow to $\mathbf{f}$. By its construction, $\mathbf{c}$ is non-negative, piecewise linear, and continuous in ν with slope (derivative calculated from positive ν) at ν of $\partial\mathbf{c}$. Then let $\partial\mathbf{f}$ be the solution to the program,

$$\max \sum_{v \in V} \partial\mathbf{f}(s \to v) \text{ s.t. } \partial\mathbf{f} \in \partial\mathcal{F}(\nu)$$

where $\partial\mathcal{F}(\nu) = \{\partial\mathbf{f} : \partial f(e) \geq 0 \text{ if } f(e) = 0 \text{ and } \partial f(e) \leq \partial c(e) \text{ if } f(e) = c(e)\}$. We argue (in the supplementary material [21]) that for $\partial\nu > 0$ small enough $\mathbf{f} + \partial\nu\partial\mathbf{f}$ is a solution to (3) at $\nu + \partial\nu$. Furthermore, Algorithm 1 solves the gradient flow program in the above display.

Theorem 1. *Let $\mathcal{P}$ be the result of the Algorithm 1. Then for each $\nu, \mathbf{f} \in \mathcal{P}$, $\mathbf{f}$ is the solution to the dual program* (3) *for ν.*

The proof can be found in [21]. Hence, the path algorithm returns maximal flows according to (3), which by analyzing the residual flow graph gives us a solution path for the RGS.

III. Experiments

We will now conclude with an empirical study of the effectiveness of Algorithm 1 on some simulations. Notice that a specific max flow algorithm was not prescribed when computing the gradient flow. In our implementation, we use the Edmonds-Karp algorithm, in which residual flows are found using breadth-first search. More recently, a dual-decomposition algorithm has been developed in order to parallelize the computation of the MAP estimator for binary MRFs [22], [23].

We construct an undirected, unweighted lattice graph by identifying each vertex with a square in a 15×15 grid, and adjoining vertices that share an side of the square with weight 1. A 4×4 rectangle was constructed to be $C^\star$ and the signal size $\mu = 2.35$ with noise level $\sigma = 1$. (Figure 1 depicts the cluster, noisy observations and reconstruction.) The smallest MSE (Hamming distance between $C^\star$ and $\hat{C}$) in the regularization path was 2. (The MSE throughout the regularization path is given in Figure 3 (left)).

We form a weighted directed graph by associating the vertices to squares in a grid ($\{(i, j) : i, j \in [15]\}$) such that $(i_1, j_1), (i_2, j_2)$ have an edge between them if $|i_1 - i_2| \leq 1$ and $|j_1 - j_2| \leq 1$. Moreover, the edge weight is equal to $e^{(i_2 - i_1 + j_2 - i_1)/3}$ so that the weight is larger in the direction of the arrows of Figure 2. $C^\star$ contains 16 vertices depicted in Figure 2 (left), the signal size is $\mu = 1.75$ and $\sigma = 1$. Again the MSE is given in Figure 3 (right). It has been demonstrated that the RGS can successfully reconstruct the true cluster of activation $C^\star$. This can also be done efficiently with the Flow Path algorithm, which gives us the entire regularization path for the RGS. The algorithm above is computationally slower than the spanning tree wavelet [17], but it experimentally dominates it and other pre-existing methods [24].

References

[1] E. Arias-Castro, E.J. Candes, and A. Durand. Detection of an anomalous cluster in a network. *The Annals of Statistics*, 39(1):278–304, 2011.

[2] L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi. On combinatorial testing problems. *The Annals of Statistics*, 38(5):3063–3092, 2010.

[3] V. Cevher, C. Hegde, M.F. Duarte, and R.G. Baraniuk. Sparse signal recovery using markov random fields. Technical report, DTIC Document, 2009.

[4] P. Ravikumar and J.D. Lafferty. Quadratic programming relaxations for metric labeling and markov random field map estimation. 2006.

[5] R.I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 315–322. Citeseer, 2002.

[6] A. Smola and R. Kondor. Kernels and regularization on graphs. *Learning theory and kernel machines*, pages 144–158, 2003.

[7] Ryan Joseph Tibshirani. *The solution path of the generalized lasso*. Stanford University, 2011.

[8] Holger Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010.

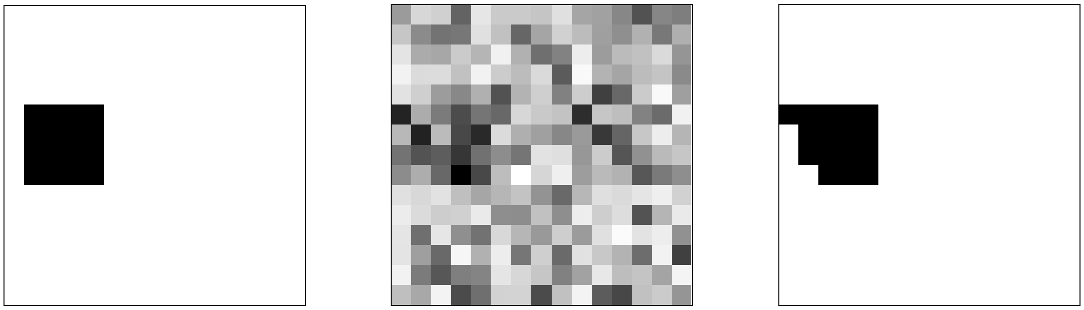

Fig. 1. The undirected 25×25 lattice graph was used with a 4×4 rectangular region $C^\star$ (left). The signal size used was $\mu = 2.35$ with $\sigma = 1$ ($\mathbf{y}$ is depicted in the middle). The best reconstruction is displayed (right).

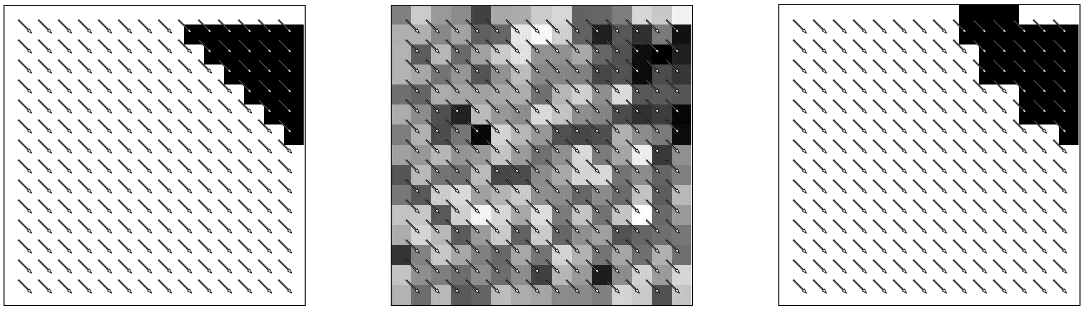

Fig. 2. A directed graph (depicted by the arrows) was constructed with edges between all 8 neighboring blocks in which more weight is on edges that descend to the right. A region was chosen that has a low cut size (left). The signal size is $\mu = 1.75$ and $\sigma = 1$ ($\mathbf{y}$ is depicted in the middle). The best reconstruction is displayed (right).

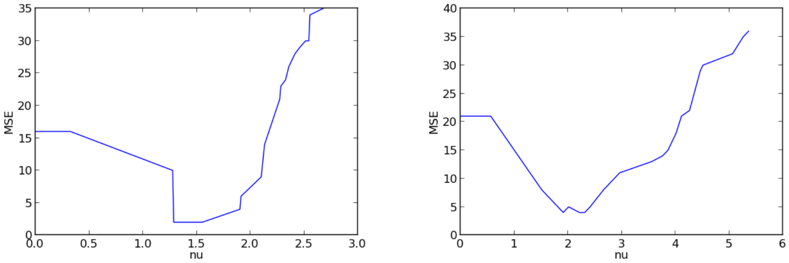

Fig. 3. MSE as a function of regularization parameter ν. The lattice graph example is left and the directed graph example is right.

[9] David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 301–369, 1995.

[10] Felix Abramovich, Yoav Benjamini, David L Donoho, and Iain M Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584–653, 2006.

[11] IM Johnstone. Function estimation and gaussian sequence models. *Unpublished manuscript*, 2002.

[12] Daniel B Neill and Andrew W Moore. Rapid detection of significant spatial clusters. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 256–265. ACM, 2004.

[13] Deepak Agarwal, Andrew McGregor, Jeff M Phillips, Suresh Venkatasubramanian, and Zhengyuan Zhu. Spatial scan statistics: approximations and performance study. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 24–33. ACM, 2006.

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

[15] Chih-Wei Yi. A unified analytic framework based on minimum scan statistics for wireless ad hoc and sensor networks. *Parallel and Distributed Systems, IEEE Transactions on*, 20(9):1233–1245, 2009.

[16] J. Sharpnack, A. Rinaldo, and A. Singh. Changepoint detection over graphs with the spectral scan statistic. *Arxiv preprint arXiv:1206.0773*, 2012.

[17] James Sharpnack, Akshay Krishnamurthy, and Aarti Singh. Detecting activations over graphs using spanning tree wavelet bases. *arXiv preprint arXiv:1206.0937*, 2012.

[18] David R Karger and Clifford Stein. A new approach to the minimum cut problem. *Journal of the ACM (JACM)*, 43(4):601–640, 1996.

[19] Vladimir Kolmogorov and Ramin Zabin. What energy functions can be minimized via graph cuts? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(2):147–159, 2004.

[20] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2001.

[21] James Sharpnack. A path algorithm for localizing anomalous activity in graphs: supplementary material. *www.stat.cmu.edu/ jsharpna/*, 2013.

[22] Petter Strandmark and Fredrik Kahl. Parallel and distributed graph cuts by dual decomposition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2085–2092. IEEE, 2010.

[23] David Sontag, Amir Globerson, and Tommi Jaakkola. Introduction to dual decomposition for inference. *Optimization for Machine Learning*, 1, 2011.

[24] James Sharpnack, Akshay Krishnamurthy, and Aarti Singh. Near-optimal anomaly detection in graphs using lovasz extended scan statistic. *Neural Information Processing Systems*, 2013.