SIGNAL DETECTION ON GRAPHS: BERNOULLI NOISE MODEL

Siheng Chen[1,2]*, Yaoqing Yang*[1]*, Aarti Singh*[2]*, Jelena Kovačević*[1,3]

[1]Dept. of ECE, [2] Dept. of ML, [3]Dept. of BME,
Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

We consider detecting localized binary attributes on a graph. A localized binary attribute means that the nodes activated by the attribute form form a subgraph that can be easily separated from the other nodes. We formulate a statistical hypothesis test to decide whether a given attribute is localized or not. We propose two statistics: graph wavelet statistic and graph scan statistic. Both are shown to be efficient and statistically effective. We further apply the proposed methods to rank research keywords as attributes in a coauthorship network collected from IEEE Xplore. The experimental results show that the proposed graph wavelet statistic and graph scan statistic are effective and efficient.[1]

1. INTRODUCTION

In this paper, we focus on detecting localized attributes corrupted by the Bernoulli noise. We model such an attribute by a binary graph signal: when a signal coefficient is one, the corresponding node is activated by this attribute. When the activated nodes form a subgraph that can be easily separated from the other nodes, this attribute is localized. This task is relevant to many real-world applications: identifying brain activity in the brain connectivity networks, viruses in cyber-physical systems and clustered attributes in social networks [2, 3, 4].

We formulate a hypothesis test to decide whether a given attribute is localized or not. We propose two statistics: graph wavelet statistic and graph scan statistic. Similarly to detect transient changes in time-series signals by using wavelet techniques [5], we design a graph wavelet statistic based on a Haar-like graph wavelet basis. Since the graph wavelet basis is preconstructed, the computational cost is linear with the number of nodes. We also formulate the generalized likelihood test and propose a graph scan statistic, which can be efficiently solved by the standard graph-cut algorithm. The intuition behind the proposed statistics is to localize the underlying localized pattern activated by the given attributes, which is equivalent to denoise the given attribute based on the graph structure, and then compute the statistic values based on the denoised attribute.

The authors gratefully acknowledge support from the NSF through awards 1130616 and 1017278, and CMU Carnegie Institute of Technology Infrastructure Award.

[1]The full version of this paper is in [1].

2. PROBLEM FORMULATION

We consider a weighted, undirected graph $G = (\mathcal{V}, \mathcal{E}, \mathrm{A})$, where $\mathcal{V} = \{v_1, \ldots, v_N\}$ is the set of nodes, $\mathcal{E} = \{e_1, \ldots, e_M\}$ is the set of edges and $\mathrm{A} \in \mathbb{R}^{N\times N}$ is a weighted adjacency matrix[2]. The ith signal coefficient corresponds to node v_i. Let $\Delta \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{V}|}$ be the *graph incidence matrix*, whose row correspond to edges [6]. For example, if e_i is an edge that connects the jth node to the kth node ($j < k$), the ith row of Δ is $\Delta_{i,\ell} = 1$, for $\ell = j$; 1, for $\ell = k$; and 0, otherwise. A *graph signal* is defined as the map on the graph nodes that assigns the signal coefficient $x_n \in \mathbb{R}$ to the node v_n; a vector form is $\mathbf{x} = \left[x_1, x_2, \ldots, x_N\right]^T \in \mathbb{R}^N$. Let $C \subseteq \mathcal{V}$ be a set of nodes. We can represent this set by using an indicator function, $\mathbf{1}_C \in \mathbb{R}^N$. That is, the signal coefficients are ones in the node set C and zeros in the complement node set $\bar{C} = \mathcal{V} \setminus C$. We call the node set C *activated*. In many applications, such as community detection, we want to find a node set C with a small number of connections between C and $\bar{C}$. To quantify the cut cost, we use the ℓ_p-norm based total variation. For a graph signal $\mathbf{x} \in \mathbb{R}^N$, the ℓ_p-norm based total variation is $\mathrm{TV}_p(\mathbf{x}) = \|\Delta\mathbf{x}\|_p$. The ℓ_0-norm based total variation $\mathrm{TV}_0(\mathbf{1}_C)$ counts the number of edges connecting C and $\bar{C}$. The ℓ_1-norm based total variation $\mathrm{TV}_1(\mathbf{1}_C) = \sum_{(i,j)\in\mathcal{E}} \mathrm{A}_{i,j}\, \mathrm{I}\{i \in C, j \in \bar{C}\}$ takes the edge weights into account, where $\mathrm{A}_{i,j}$ is the edge weight between the ith and jth nodes. When edges are unweighted, the ℓ_0-norm based total variation and the ℓ_1-norm based total variation are the same. In general, a smaller total variation means that it is easier to separate the node set C from $\bar{C}$.

In this paper, we are particularly concerned with the following testing problem:

$$\begin{aligned} H_0 &: \mathbf{y} \sim \mathrm{Bernoulli}(\epsilon_0 \mathbf{1}_{\mathcal{V}}), \\ H_1 &: \mathbf{y} \sim \mathrm{Bernoulli}(\mu\mathbf{1}_C + \epsilon\mathbf{1}_{\bar{C}}) \text{ for some } \mathrm{TV}_p(\mathbf{1}_C) \leq \rho, \end{aligned} \tag{1}$$

where Bernoulli(c) is 1 with probability c and 0 with probability $(1-c)$, μ is the signal strength and $\epsilon(<\mu)$ is the noise level, and $\epsilon_0 = \epsilon + |C|\,(\mu-\epsilon)\,/N$, such that the signals under both H_0 and H_1 have the same mean. The constraint $\mathrm{TV}_p(\mathbf{1}_C) \leq \rho$ with p either 0 or 1 defines a class of node sets that are easy to be separated from its compliment. The null hypothesis H_0 describes a noisy attribute and the alternative hypothesis H_1 describes a localized attribute.

[2]both the node order and the edge order are fixed

The motivation behind this setting is the community detection with attributes [7, 8]. For example, we want to identify whether users who are graduated from Carnegie Mellon University form a social circle on Facebook. The binary value "Is a student from Carnegie Mellon University?" is a binary attribute on Facebook. When this attribute is relevant to a social circle, we should find a subgraph such that (1) most nodes are activated in the subgraph and few nodes are activated out of the subgraph; (2) the connection between this subgraph and its complement is weak. In (1), we describe a binary attribute by the Bernoulli noise model and the localized patterns by a class of graph signals that have small total-variations.

3. METHODOLOGY

In this section, we propose two statistics for (1). The first statistic is based on a graph wavelet basis. The second statistic is based on searching all the feasible node sets that satisfy the constraint in (1) by solving an optimization problem.

3.1. Graph Wavelet Statistic

In classical signal processing, we detect transient changes in a time-series signal by projecting the signal on the wavelet basis [5]. When a coefficient in a high-frequency band is large, there exists a transient change. Similarly, we construct a Haar-like graph wavelet basis to detect the boundary, which connects the activated node set and the inactivated node set. When the boundary exists, we reject the null hypothesis. The main idea of constructing the graph wavelet basis is to recursively partition a node set into two disjoint child node sets with similar sizes. Each partition generates a basis that spans the resulting two children node sets. We start from the entire node set, corresponding to the coarsest resolution of the graph vertex domain. Finally, each node set is either an individual node or an empty set, corresponding to the finest resolution of the graph vertex domain. Please see more details in [9, 10].

We use the graph wavelet basis to detect localized patterns. A graph wavelet basis vector captures the transient change (boundary) in a graph signal. For a localized graph signal, the ℓ_0 norm based total variation is small and the energy of the original graph signal is concentrated in a few graph wavelet coefficients, corresponding to the boundary; for a noisy graph signal, the ℓ_0 norm based total variation is large and the energy of the original graph signal is spread over all the graph wavelet coefficients. Mathematically, given an noisy attribute $\mathbf{y} \in \mathbb{R}^N$, the *graph wavelet statistic* is the maximum absolute value of the graph wavelet coefficients, $\widehat{w} = \left\| \mathrm{W}_{(-1)}^T \mathbf{y} \right\|_\infty$, where the infinity norm $\|\cdot\|_\infty$ takes the maximum value, $\mathrm{W}_{(-1)} \in \mathbb{R}^{N\times(N-1)}$ is the graph wavelet basis $\mathrm{W} \in \mathbb{R}^{N\times N}$ without the first constant column. We remove the first constant column as it does not provide discriminative information. When $\widehat{w}$ is larger than some threshold, we detect the activated node set and reject the null hypothesis. To analyze the detection error, we have the following theorem.

Theorem 1. Perform the following statistical test (1) for $p = 0$. We reject the null if $\widehat{w} = \left\| \mathrm{W}_{(-1)}^T \mathbf{y} \right\|_\infty > \tau$. Set $\tau = \sqrt{\log N} + \sqrt{2\log(\frac{2}{\delta})}$. When

$$\mu - \epsilon \geq \frac{\sqrt{1+\rho\log N}(\sqrt{\log N} + 2\sqrt{2\log(\frac{2}{\delta})})}{\sqrt{|C|\left(1-\frac{|C|}{N}\right)}}, \quad (2)$$

under H_0, $\mathbb{P}\{\text{reject}\} \leq \delta$; under H_1, $\mathbb{P}\{\text{reject}\} \geq (1-\delta)^4$.

The main idea of the proof is to show that under the null hypothesis, each graph wavelet coefficient is a subgaussian random variable, which means that its statistical performance is similar to a gaussian random variable [11]; and under the alternative hypothesis, the maximum value of the graph wavelet coefficients is large because the energy of the original graph signal concentrates in a few graph wavelet coefficients. The assumption (2) shows that when the size of the ground-truth activated node set is larger and the ground-truth activated node set has a small ℓ_0-norm based total variation, it is easier for graph wavelet statistic to detect the activation.

The computational bottleneck of constructing the graph wavelet basis is the graph partition algorithm. When a graph partition algorithm takes $O(N\log N)$, the total construction costs $O(N\log^2 N)$. Since the graph wavelet basis is constructed only based on the graph structure, the construction only computes once. The total computational cost to obtain the graph wavelet statistic only involves a matrix-vector multiplication and a search for the maximum value. The graph wavelet statistic is thus scalable to large-scale graphs.

3.2. Graph Scan Statistic

The graph wavelet statistic is efficient, but the construction of the graph wavelet basis is independent from the given attribute. Now we aim to propose an data-adaptive approach, which scans all feasible node sets based on the given graph signal. For the simplicity of derivation, we assume that $\epsilon_0 = \epsilon$ in (1). This implies that the average value under H_1 is larger than the average value under H_0. A naive approach is then to use the average of the observation as the statistic. Here we propose a graph scan statistic that works better than this naive approach. The intuition behind the proposed statistics is that given the noisy observation $\mathbf{y}$, we search for the activated node set C. If we can find such C, we reject the null, and vice versa. To make a fair comparison, we still set $\epsilon_0 = \epsilon + |C|\,(\mu-\epsilon)\,/N$ when validating this statistic in the experiments in Section 4.

When we knew the true activated node set $C \in \mathcal{C}$, an intuitive idea is to test the null hypothesis $H_0 : \mathbf{x} = 0$ against the alternative $H_1 : \mathbf{x} = \mu\mathbf{1}_C$ by using the likelihood ratio test. Given the observation $\mathbf{y}$, based on the Bernoulli noise model, the likelihood is $\mathbb{P}(\mathbf{y}|H_1) = \prod_{i\in C}\mu^{y_i}(1-\mu)^{1-y_i}\prod_{i\in\bar{C}}\epsilon^{y_i}(1-\epsilon)^{1-y_i}$. The maximum likelihood esti-

mator is $\widehat{\mu} = \mathbf{1}_C^T\mathbf{y}/|C|$ and the generalized likelihood ratio

$$\begin{aligned}\widehat{g} &= \max_C |C|\,\mathrm{KL}\left(\frac{\mathbf{1}_C^T\mathbf{y}}{|C|}||\epsilon\right) \\ &\quad \text{subject to } \mathrm{TV}_p(\mathbf{1}_C) \le \rho.\end{aligned} \tag{3}$$

When the average value in the activated node set $\mathbf{1}_C^T\mathbf{y}/|C|$ is much larger than the background randomness ϵ, $\widehat{g}$ is large and the we detect the activated node set C. We call $\widehat{g}$ *graph scan statistic* (this is the Bernoulli version [12, 13]). The intuition behind this statistic is that we want to maximize both the size of the the activated node set and the average value in the activated node set. When $\mathbf{y}$ is noise belonging to H_0, it has a small probability to have the activated nodes in $\mathbf{y}$ that are clustered together and have small cut cost.

When $\widehat{g}$ is larger than some threshold, we detect the activated node set and reject the null hypothesis. To analyze the detection error, we have the following theorem.

Theorem 2. Perform the following statistical test (1) for $p = 1$. We reject the null if $\widehat{g} > \tau$. Set

$$\begin{aligned}\tau &= 8\Bigg(\left(\sqrt{\rho} + \sqrt{\frac{1}{2}\log N}\right)\sqrt{2\log(N-1)} \\ &\quad + \sqrt{2\log 2} + \sqrt{2\log(1/\delta)}\Bigg)^2.\end{aligned}$$

$$\begin{aligned}\text{when } \mu - \epsilon &\ge \frac{2\sqrt{2}}{\sqrt{|C|}}\Bigg(\left(\sqrt{\rho} + \sqrt{\frac{1}{2}\log N}\right)\sqrt{2\log(N-1)} \\ &\quad + \sqrt{2\log 2} + \sqrt{2\log(1/\delta)}\Bigg) + \sqrt{\frac{\log(2/\delta)}{2|C|}},\end{aligned} \tag{4}$$

under H_0, $\mathbb{P}\{\text{reject}\} \le \delta$; and under H_1, $\mathbb{P}\{\text{reject}\} \ge 1 - \delta$.

The main idea of the proof is to show that under the null hypothesis, $\mathbf{1}_C^T(\mathbf{y} - \epsilon)/\sqrt{|C|}$ is a subgaussian random variable; and under the alternative hypothesis, the maximum likelihood estimator $\mathbf{1}_C^T\mathbf{y}/|C|$ is close to μ with high probability. Similarly to the graph wavelet statistic, the assumption (4) shows that when the size of the ground-truth activated node set is larger and the ground-truth activated node set has a smaller ℓ_1-norm based total variation, it is easier for graph scan statistic to detect the activation.

The advantages of graph scan statistic over graph wavelet statistic are: (1) graph scan statistic is data adaptive. When an attribute is given, graph scan statistic actively searches the activated node set, instead of using a pre-constructed basis; (2) graph scan statistic is a generative approach searching an activated node set. Thus, graph scan statistic not only detects the whether the localized activated node set exists, but also localize these regions; and (3) graph scan statistic takes care of the edge weights by using the ℓ_1-norm based total variation. The ℓ_1-norm based total variation is more general because when we only consider binary edge weights, the ℓ_1-norm based total variation and the ℓ_0-norm based total variation are the same. Thus, all the results based on the ℓ_1-norm can be directly applied to the ℓ_0-norm.

In the previous analysis, we use the global optimum of (3), however, this global optimum is hard to obtain because the optimization problem is not convex. To make the graph scan statistic practical, we reformulate (3) and solve

$$\begin{aligned}\widehat{g} &= \max_t \max_{\mathbf{x}} t\,\mathrm{KL}\left(\frac{\mathbf{x}^T\mathbf{y}}{t}||\epsilon\right) \\ &\quad \text{subject to } \mathbf{x} \in \{0,1\}^N, \mathrm{TV}_1(\mathbf{x}) \le \rho, \mathbf{1}^T\mathbf{x} \le t,\end{aligned} \tag{5}$$

where $\mathbf{x}$ is equivalent to $\mathbf{1}_C$ and t is equivalent to $|C|$. Since ϵ is a small constant, for each t, we optimize over $\mathbf{x}$ to make $\mathbf{x}^T\mathbf{y}/t$ as far away from ϵ as possible, which is equivalent to maximizing $\mathbf{x}^T\mathbf{y}$ within the feasible region[3]. The Lagrange dual function is

$$\begin{aligned}h(\eta_1,\eta_2) &= \min_{\mathbf{x}\in\{0,1\}^N} L(\eta_1,\eta_2,\mathbf{x}) \\ &= \min_{\mathbf{x}\in\{0,1\}^N} \left(-\mathbf{x}^T\mathbf{y} + \eta_1\mathbf{1}^T\mathbf{x} + \eta_2\left\|\Delta\mathbf{x}\right\|_1\right) - \eta_1 t - \eta_2\rho \\ &= f(\eta_1,\eta_2) - \eta_1 t - \eta_2\rho.\end{aligned}$$

For given η_1, η_2, the function $f(\eta_1,\eta_2)$ can be efficiently solved by s-t graph cuts [14, 15]. We then maximize $h(\eta_1,\eta_2)$ by using the simulated annealing. Since $\mathbf{x}$ takes only binary values, the optimization problem (5) is not convex. However, because optimizing $f(\eta_1,\eta_2)$ is a standard graph-cut problem, even the local minimum provides decent results [15] and the computation is remarkably efficient.

4. EXPERIMENTS: RANKING ATTRIBUTES

Relevant attributes improve the accuracy in community detection and interpreting the detected communities, but irrelevant attributes may harm the accuracy and cause computational inefficiency [7]. Here we aim to use the proposed statistics to find the most useful attributes to benefit the community detection. We use the IEEE Xplore database to find working collaborations between scholars. The raw files were downloaded from [16] and we focus on papers in 10 IEEE Transactions journals. The collected dataset includes the coauthorship network with $7,330$ authors and $108,719$ coauthorships, the author-journal matrix with 10 journals, and the author-keyword matrix with $3,596$ keywords. In other words, we have a graph with $7,330$ nodes and $108,719$ edges and $3,596$ corresponding attributes. Here we want to detect different academia communities based on graph structure and attributes. The academia communities are defined based on the journals. Thus, ten journals correspond to ten groundtruth communities: when authors publish at least ten papers in a same journal, we assign them into a community. For example, authors who publish papers in IEEE Transactions on Signal Processing form

[3] we implicitly assume that $\mathbf{1}_C^T\mathbf{y}/|C| > \epsilon$.

the signal processing community. In this community, some keywords are frequently used, such as 'filtering' and 'Fourier transform'. By using those relevant keywords, the quality of community detection may be improved.

The goal is to rank all the keywords based on their contribution to community detection, where the magnitude or value (if negative is meaningful) of the corresponding statistic is used to determine the rank. Here we consider four ranking methods: graph wavelet statistic based ranking (Wavelet), graph scan statistic based ranking (LGSS), modularity-based ranking (Modularity) and cut-based ranking (Cuts). For the first two ranking methods, we compute the statistical value and rank the keywords according to the statistical value in a descending order. This is because a larger statistic means a larger probability that this keyword forms a community. For the modularity-based ranking, we compute the modularity of each keyword and rank the keywords according to modularity in a descending order [6]. For the cut-based ranking, we compute the cut cost of each keyword and rank the keywords according to the number of cuts in a ascending order [6].

To quantify the real community detection power of keywords, we compare each keyword to the ground-truth community and compute the correspondence by using the average F1 score, which is suggested in [17, 7]. Mathematically, let C^* be a set of the ground-truth communities and $\widehat{C}$ be a set of the activated node sets provided by the node attributes. Each node set $\widehat{C}_i \in \widehat{C}$ collects the nodes that have the same attribute. The average F1 score is $\frac{1}{2|C^*|}\sum_{C_i\in C^*} F1(C_i, \widehat{C}_{g(i)}) + \frac{1}{2|\widehat{C}|}\sum_{\widehat{C}_i\in\widehat{C}} F1(\widehat{C}_{g'(i)}, \widehat{C}_i)$, where the best matching g and g' is defined as follows: $g(i) = \arg\max_j F1(C_i, \widehat{C}_j)$ and $g'(i) = \arg\max_j F1(C_j, \widehat{C}_i)$, with $F1(C_i, \widehat{C}_j)$ is the harmonic mean of precision and recall. A large average F1 score means that the community induced by a keyword agrees with the community induced by journal papers. We also compute the average F1 score of each keyword and rank the keywords according to the average F1 scores in a descending order, which is the ground-truth ranking. We compare the four estimated rankings with the ground-truth ranking by using the Spearman's rank correlation coefficient [18]. The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables: $\rho = 1 - \frac{6\sum |p_i - q_i|^2}{N(N^2-1)}$, where $p_i - q_i$ is the difference between two rankings. The Spearman correlation coefficients of modularity-based ranking, cut-based ranking, graph wavelet statistic based ranking and local graph scan statistic based ranking are shown in Figure 1. We see that graph wavelet statistic and graph scan statistic outperform the other methods. Cut-based ranking performs bad because it may rank infrequent keywords higher. To normalize this factor, we also consider the average modularity, which is the modularity divided the number of activated authors, and the average cuts, which are the number of cuts divided the number of activated authors. Average cuts perform much better than the total cuts.

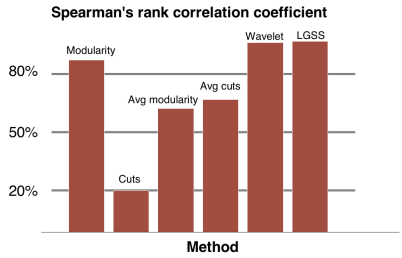

Fig. 1: Comparison of Spearman's rank correlation coefficient. A larger Spearman's rank correlation coefficient means higher correlation to the ground-truth ranking.

Figure 2 compares the average F1 scores as a function of individual keywords ordered by the six ranking methods. The x-axis is the ranking provided by the proposed ranking methods and the y-axis is the average F1 score of the corresponding keyword. For example, graph scan statistic ranks 'Maximum likelihood detection' in the first place, we then put the corresponding average F1 score as the first element in the red curve (most left). We expect that the curve goes down as the rank grows because a good ranking method ranks the important keywords higher. We also use cluster affiliation model for big networks (BIGCLAM, shown in a black, horizontal line), a large-scale community detection algorithm to provide a baseline [17]. We see that graph scan statistic is slightly better than graph wavelet statistic and both of them outperform the other methods, which is consistent to the results given by the Spearman correlation coefficients in Figure 1. Surprisingly, using high ranking keywords to detect communities works even better than BIGCLAM. The reason may be that in this coauthorship network, the keywords are informative and strongly related to the journals, which are the ground-truth communities.

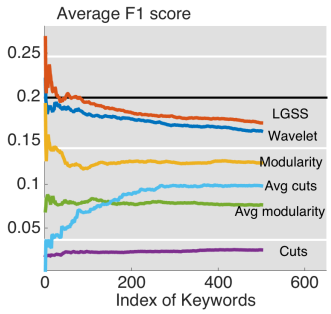

Fig. 2: Comparison of average F1 score as a function of the top k ranked keywords. A higher average F1 score means better detection performance by using each individual keyword.

5. CONCLUSIONS

This paper focused detecting localized attributes on a graph. We formulated hypothesis tests to decide whether the attributes activate a community in a graph. We proposed two statistics: graph wavelet statistic and graph scan statistic. Both were shown to be efficient and statistically effective for the detection. We further applied the proposed method to rank keywords in a coauthorship network.

6. REFERENCES

[1] S. Chen, Y. Yang, S. Zong, A. Singh, and J. Kovačević, "Detecting structure-correlated attributes on graphs," *arXiv:1604.00657*, 2016.

[2] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data Mining and Knowledge Discovery*, vol. 29, no. 3, pp. 626–688, 2015.

[3] A. Krishnamurthy, "Minimaxity in structured normal means inference," *http://arxiv.org/pdf/1506.07902*, 2015.

[4] J. Sharpnack, A. Krishnamurthy, and A. Singh, "Detecting activations over graphs using spanning tree wavelet bases," in *AISTATS*, Scottsdale, AZ, Apr. 2013.

[5] M. Vetterli, J. Kovačević, and V. K. Goyal, *Foundations of Signal Processing*, Cambridge University Press, Cambridge, 2014, http://www.fourierandwavelets.org/.

[6] M. Newman, *Networks: An Introduction*, Oxford University Press, 2010.

[7] J. Yang, J. J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *ICDM*, Dallas, TX, Dec. 2013, pp. 1151–1156.

[8] J. Zhu Y. Zhang, E. Levina, "Community detection in networks with node features," *http://arXiv:1509.01173*, 2015.

[9] S. Chen, R. Varma, A. Singh, and J. Kovačević, "Signal representations on graphs: Tools and applications," *arXiv:1512.05406*, 2015.

[10] S. Chen, R. Varma, A. Singh, and J. Kovačević, "Representations of piecewise smooth signals on graphs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016.

[11] M. Ledoux, *The concentration of measure phenomenon*, vol. 89, American Mathematical Soc., 2001.

[12] J. Sharpnack, A. Rinaldo, and A. Singh, "Changepoint detection over graphs with the spectral scan statistic," in *Artifical Intelligence and Statistics (AISTATS)*, 2013.

[13] J. Sharpnack, A. Krishnamurthy, and A. Singh, "Near-optimal anomaly detection in graphs using lovasz extended scan statistic," in *Neural Information Processing Systems (NIPS)*, 2013.

[14] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1222–1239, Nov. 2001.

[15] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.

[16] "IEEE Xplore digital library," http://ieeexplore.ieee.org/search/advsearch.jsp.

[17] J. Yang and J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach," in *WSDM*, Rome, Italy, Feb. 2013, pp. 587–596.

[18] J. L. Myers, A. D. Well, and R. F. Lorch Jr, *Research Design and Statistical Analysis (3nd ed.)*, Routledge, 2010.