Article type: Focus Article

Scan Statistics on Graphs **Article ID**

David Marchette

Naval Surface Warfare Center

Keywords

Discrete Mathematics, Anomaly Detection, Scan Statistics, Graphs

Abstract

Scan statistics are used in spatial statistics and image analysis to detect regions of unusual or anomalous activity. A scan statistic is a maximum (or minimum) of a local statistic – one computed on a local region of the data. This is sometimes called "moving window analysis" in the Engineering literature. The idea is to "slide" a window around the image (or map or whatever spatial structure the data have), compute a statistic within each window, and look for outliers – anomalously high (or low) statistics. We discuss extending this idea to graphs, in which case the local region is defined in terms of the connectivity of the graph – the neighborhoods of vertices.

The basic parameter of a traditional scan statistic is the definition of the "local region" in which the local statistic will be calculated. In spatial statistics and image analysis there is a natural definition of "local region": generally a rectangle (or ball) centered on a particular coordinate. One computes a statistic on all observations contained in the region, and looks for regions with unusually large or small values of the statistic.

In recent years the analysis of graphs (such as communications graphs, social networks, sensor networks and the like) has become of considerable interest in a wide range of scientific and engineering disciplines. Graphs also have a natural definition of "local": each vertex in the graph is connected to its neighbors, and this defines a local neighborhood of the vertex. One then defines scan statistics on graphs using these local neighborhoods. In this article we discuss the extension of the basic ideas of scan statistics to graphs, and in particular to anomaly detection in a time series of graphs.

Scan Statistics

Traditional scan statistics are used to detect anomalies in a random field, spatial point process, image, or geographical data. Given data X and a window w, which may be an interval in a time series, a square in an image, etc., a statistic L_w is computed on

the data contained in w. $L_w = L_w(X)$ is referred to as a **local statistic**. So, given windows $w_1, \ldots, w_n$ such that $X \subset \cup w_i$, we define the scan statistic as

$$S(X) = \max_i L_{w_i}(X). \tag{1}$$

Under standard hypothesis testing, given a null hypothesis H_0 defining "homogeneity", one specifies an α level and critical value c_α such that $P_{H_0}[S(X) \leq c_\alpha] = \alpha$. If the observed value of $S(X)$ exceeds the critical value, one rejects the null in favor of "inhomogeneity". Further, the window for which the locality statistic attains the maximum provides information about where the anomaly occur ed within the time series, random field, or image. The computation of c_α requires knowledge about the distribution of the statistic under the null, which may be complicated by the fact that dependencies may be introduced if the windows overlap. See Naiman and Priebe (2001) for some discussion on methods for computing p-values for scan statistics, and Loader (1991) for some theory of approximations.

The genesis of scan statistics is Fisher's "quadrat counts", Fisher et al. (1922), in which disjoint windows were used. In the absence of prior knowledge of the position of the anomaly, disjoint windows can have poor performance due to "splitting" the anomaly between multiple windows. By using overlapping windows, modern scan statistics solves this problem, with the downside of increasing the difficulty of exact analysis, as mentioned above. For information on point process techniques the interested reader is encouraged to consider Diggle (1983). For more information on scan statistics, see Cressie (1977, 1980); Chen and Glaz (1996); Kulldorff (1997); Glaz et al. (2001).

Scan Statistics on Graphs

Graphs have been used in the scan statistics literature primarily to define the scan regions of spatial data (Patil and Taillie (2003)). The first definition of a true scan statistic on graphs of which we are aware is Priebe (2004); Priebe et al. (2005). The main focus of this work was the detection of anomalies in a time series of graphs. The goal was to detect that a small region of vertices had an unusually high number of connections amongst themselves as compared to their previous activity and that of other groups. More recently, Arias-Castro et al. (2011) considered a similar problem, but on a single graph. Scan statistics have also been applied to graphs which are defined in terms of their relationships in space. Yi (2009) provides results for sensor and wireless networks in which the graph is defined according to the relative positions of the sensors or devices.

Consider first a graph $G = (V(G), E(G))$ where $V(G) = [n] = \{1, \ldots, n\}$ is the set of vertices (or nodes) of the graph, and $E(G) \subset V(G)^{(2)}$ is the set of edges. Here $V^{(2)}$ is the set of unordered pairs of elements of V (for simplicity we will assume the graph is undirected, but directed graphs can be analyzed similarly). We will write V and E for the vertices and edges of a graph if this is unambiguous. The number of edges in a graph, $|E|$ is called the **size** of the graph. These are examples of **graph invariants**.

A graph invariant is a function from graphs to real numbers (or any given range set) that is invariant on isomorphism classes. Informally, it is a number calculated from the graph that does not depend on how the graph is presented – relabeling the vertices, laying out the graph in a figure, or any other manipulation of the graph that does not change the graph structure.

Priebe et al. (2005) defines a local statistic as a graph invariant applied to the induced subgraph of a neighborhood of a vertex. Thus, given a graph invariant $\psi : G \to \mathbb{R}$, we define the k-neighborhood of a vertex $v \in V$ as

$$N_k(v) = \{u \in V : d(u, v) \leq k\},$$

where d is the graph distance – the number of edges that must be traversed between u and v (the convention is to set $d = \infty$ for vertices in different connected components). Let $\Omega(U)$ denote the induced subgraph of the subset $U \subset V$. This is the graph $(U, E(U))$, where $E(U) \subset E$ consisting of all edges in E between vertices in U. The locality statistic is then $\Psi_k(v) = \psi(\Omega(N_k(v)))$. The scan statistic at scale k is then

$$M_k(G) = \max_{v \in V} \Psi_k(v).$$

Comparisons of this approach for detection of an anomaly in a random graph is provided in Pao et al. (2011).

Arias-Castro et al. (2011) take a slightly different approach to the question. In their formulation, each vertex v has associated with it a random variable X_v (in their simple case $X_v \sim N(0, 1)$ or $X_v \sim N(\mu, 1)$.) Their test is then for:

$$\begin{aligned} H_0 : & X_v \sim N(0, 1) && \forall v \in V \\ H_A : & X_v \sim N(0, 1) && v \notin K \subset V, \\ & X_v \sim N(\mu, 1) && v \in K. \end{aligned}$$

They also differ from the graph-invariant approach in that their graph is assumed to be embedded in $\mathbb{R}^d$, and they take the neighborhood to be a ball in $\mathbb{R}^d$ centered at a vertex.

Taking more general hypotheses (which Arias-Castro et al. (2011) do) and utilizing graph neighborhoods rather than Euclidean neighborhoods should enable one to reconcile the differences between these two views of scan statistics on graphs. However in this paper we will concentrate primarily on the first, in which the scan statistic is defined directly from the connectivity of the graph, and not from some embedding of the graph in another space. As we discuss below, the issue of attributes on the vertices (X_v) as well as on the edges can be incorporated in the scan statistics approach of Priebe et al. (2005). See Grothendieck et al. (2010) (Priebe et al. (2010) is also relevant to the problem of anomaly detection in attributed graphs).

Time Series of Graphs

A time series of graphs is a collection of graphs $\{G_t\}$ indexed by time. We assume that the graphs all share the same vertex set, and so $G_t = (V, E_t)$. For a given graph G_t,

we define the locality statistic to be size of the neighborhood of a vertex:

$$\psi_k^t(v) = s(\Omega(N_k(v))), \tag{2}$$

where $s(G) = |E(G)|$.

In real graphs, particularly communications graphs where the edges correspond to communications between the entities represented by the vertices, a homogeneous graph model is inappropriate. We are interested not in finding regions in one graph that are anomalous, but rather in regions which are anomalous when compared to their past history. Thus we "standardize" the locality statistic. Set

$$\mu_k^t(v) = \frac{1}{w} \sum_{\ell \in \{t-w,\ldots,t-1\}} \psi_k^\ell(v) \tag{3}$$

$$\sigma_k^t(v)^2 = \frac{1}{w-1} \sum_{\ell \in \{t-w,\ldots,t-1\}} (\psi_k^\ell(v) - \mu_k^t(v))^2 \tag{4}$$

$$\Psi_k^t(v) = \frac{\psi_k^t(v) - \mu_k^t(v)}{\max\{1, \sigma_k^t(v)\}}. \tag{5}$$

The maximum in Equation (5) is to avoid instabilities due to unchanging neighborhoods. Large values of $\Psi_k^t(v)$ are indicative of regions that are anomalous – they have a larger than expected number of communications amongst their neighbors.

The scan statistic at time t then is

$$\widetilde{\Psi_k^t} = \max \Psi_k^t(v), \tag{6}$$

and the $\arg\max \Psi_k^t(v)$ is the vertex of interest – the "center" of the detected anomaly.

A further normalization may be necessary for nonstationary graphs, as discussed in Priebe et al. (2005). Here the scan statistic $\widetilde{\Psi_k^t}$ is itself standardized in the same manner as above.

In time series of graphs, the question arises as to how one should define the neighborhood. Priebe et al. (2005) used the neighborhood of each graph:

$$\psi_k^t(v) = s(\Omega(N_k(v; G_t); G_t)).$$

However, if we are interested in a detection at time T, perhaps we should use the neighborhood at time T:

$$\psi_k^t(v) = s(\Omega(N_k(v; G_T); G_t)).$$

This is one of the extensions discussed in Wan et al. (2006a,b).

The graph invariant we use throughout this paper (which the one used primarily in the literature to date) is the size of the induced subgraph. Other invariants could certainly be used. In particular, if something is known about the expected change in behavior it might be possible to design an invariant to have more power to detect the change.

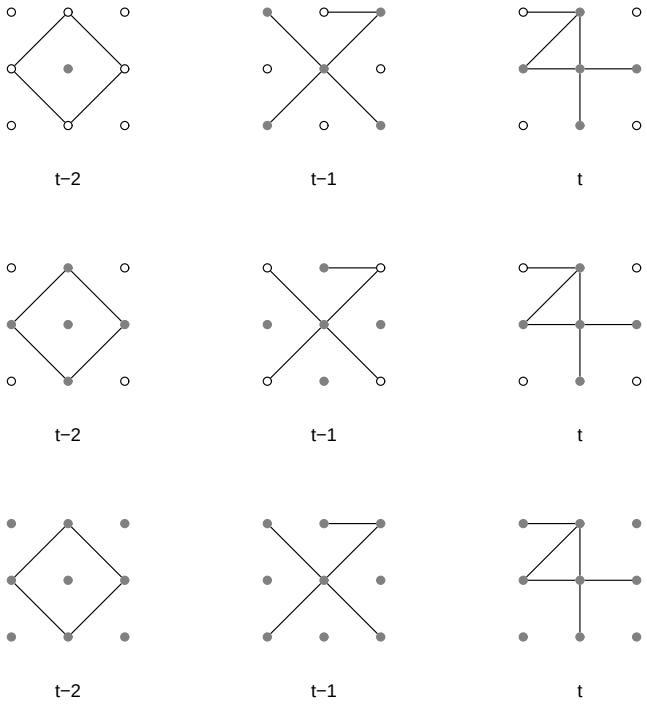

Figure 1: Three different definitions of local regions that could be used in computing the locality statistic. In all of these, the region is centered on the central vertex, and indicated by the solid gray dots. The top plot uses $N_1(v; G_s)$ for $s \in \{t-2, t-1, t\}$. The middle plot uses $N_1(v; G_t)$ for all the graphs. The bottom plot uses $\cup_s N_1(v; G_s)$.

Johannsen (2012) suggests that using the Betti numbers of a graph might be of value, as they count the numbers of particular induced subgraphs.

Figure 1 illustrates the use of different definitions of a local region. In the top we use the method of Priebe et al. (2005) wherein each region is defined by the current graph. Here the locality statistics are $(\psi^{t_2}(v), \psi^{t_1}(v), \psi^{t}(v)) = (0, 4, 5)$. In the second we impose the neighborhood at time t upon the previous graphs, and the locality statistics are: $(4, 0, 5)$. Finally, we use the union of all neighborhoods within the window, and the locality statistics are $(4, 5, 6)$. Which definition one uses for the local region should be driven by the type of anomaly one is trying to identify, and whatever specific details about the application are relevant.

Choosing the right value of k, the "radius" of the scan region, is also important, just as it is in the traditional scan statistic. It can be particularly important in graphs – the "six degrees of Kevin Bacon" effect[1] shows that many real-world networks have relatively small diameters, and so large values of k quickly become very non-local. Better ways of designing the neighborhoods for real-world graphs, and for detecting particular types of anomalies, are important areas of research.

A Monte Carlo Experiment

To illustrate the procedure, consider a sequence of graphs in which at a given point in time a small group have an increased number of within-group edges. Figure 2 depicts a typical such situation, showing that the scan statistic easily detects this case. Here we have generated a time series of 100 independent graphs, all except graph 50 are Erdös-Renyí graphs: the edges are drawn independently with a fixed probability p (in this case $p = 0.05$). Graph 50 is the same, with the exception of a small group ($m = 10$) of vertices with a higher within-group edge probability ($q = 0.75$). The vertical line indicates time $t = 50$, and so as we can see, the scan statistic easily detects the anomalous graph.

Of course, we have set the problem up so that it is easy to detect the anomaly. In Figure 3 we run a more extensive simulation. Here, for each value of q, we run 100 Monte Carlo experiments and consider the difference between the test statistic $\hat{\Psi}$ at the true anomaly and the maximum of the $\hat{\Psi}$ at the other times. As can see, at $q = 0.7$, the scan statistic has a high detection rate. See Pao et al. (2011), Lee and Priebe (2011) and Rukhin and Priebe (2012) for various discussions about the problem of detection of anomalies in time series of random graphs, and the scan statistic approach in particular.

In real-world situations the graphs are not independent and they are not Erdös-Renyí. The scan statistic is designed to detect changes amongst small groups (the neighbors of a vertex) in these graphs, and in the next section we illustrate this with a real dataset.

[1] `en.wikipedia.org/wiki/Six_Degrees_of_Kevin_Bacon`, see also Watts (2003).

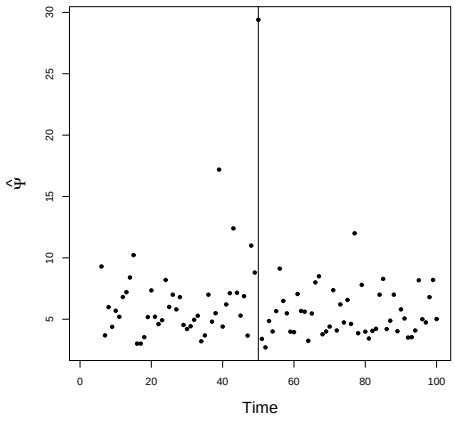

Figure 2: Scan statistics for a series of random graphs. These are independent, and for time $t \neq 1$ these are independent Erdös-Renyí with $n = 100$ vertices and $p = .05$. At time $t = 50$, a small group ($m = 10$) of vertices have the probability of within group edges increased to $q = 0.85$.

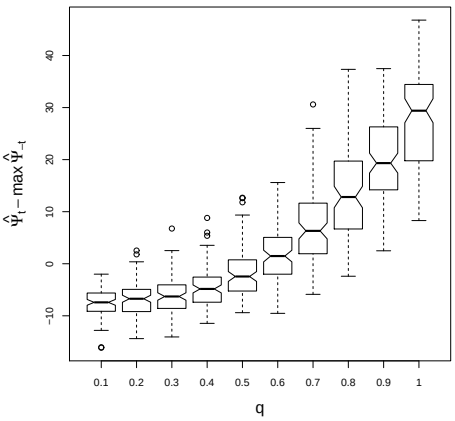

Figure 3: The results from a 100-sample Monte Carlo experiment in which each run is an experiment as depicted in Figure 2, with varying q as indicted by the horizontal axis. The vertical axis corresponds to the difference between the scan statistic at time $t = 50$, when the anomaly occurs, and the maximum scan statistic over all other times.

Enron

In 2005 the email data extracted from the mailboxes of 150 Enron executives was made available by the Federal Energy Regulatory Commission (see `www-2.cs.cmu.edu/~enron` and `cis.jhu.edu/~parky/Enron/enron.html`). This dataset consists of 184 distinct email addresses used by the 150 executives, covering 187 weeks. For each week, a graph is constructed on the 184 email addresses with an edge between any two addresses if one of them sent an email to the other (the sender-recipient pair results in a directed edge in this experiment; only the existence of a communication is kept, not the number of such communications). Figure 4 shows some statistics on these graphs. Note that the overall amount of communications grows through time (left plot) until near the end of the company's existence.

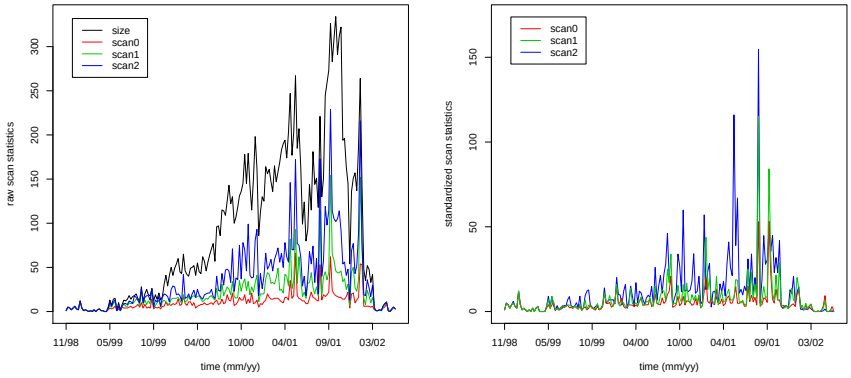

Figure 4: The size and scan statistics for the Enron data. By convention, scan_0 refers to using the degree of the vertex rather than the size of an induced neighborhood. On the left is the un-normalized version, also showing the size of the graph, while on the right the statistics have been normalized as discussed in the text (Equation (5)).

In the original Enron scan statistic paper, Priebe et al. (2005), two main detections were found. The first was a case of aliasing: an executive started using a new email address that he had not previously used. This is a trivial detection, in the sense that the local neighborhood went from a single vertex for all previous time to suddenly having a non-trivial neighborhood. Although the vertex was inactive prior to the time of the detection, it is necessary to go two steps out from the vertex in order to detect it. This is because the new vertex had a small number of neighbors at the time it was created: a change from 0 to 3 neighbors is not large, in the grand scheme of things.[2]

A more interesting detection is illustrated in Figure 5. In this case the detection was a result of individuals discussing the California energy crisis, and their complicity in

[2]Note however that this is in part an artifact of using 1 for the minimum value for σ in Equation (5).

it – in particular they were discussing ways to convince regulators and the public that they were not complicit and that the crisis was the result of market forces. As discussed above, only the existence of communications was used to find the anomaly, not the specific content. For discussion of ways to bring content into the analysis, see Grothendieck et al. (2010); Priebe et al. (2010). Figure 5 depicts the local region corresponding to the detection.

In the plots of Figure 5 we see one of the arguments for considering other definitions of "locality" for our scan statistic. In the bottom plot we see the neighborhood defined for G_t (with one extra vertex – kenneth.lay). One could argue that considering this full neighborhood, rather than the induced one, would be less sensitive to instabilities caused by the fact that small changes in connections can cause large changes in the overall neighborhood. In a sense, using time T to define the local region is equivalent to defining the current neighborhood as a community of individuals, and the scan statistic is then a measure of how this community has or has not changed over time. This is in contrast to the individual-based information obtained when one uses the induced neighborhood at each time t, as was done in Priebe et al. (2005).

Open Questions and Future Research

There are several areas that should be investigated further. In Borges et al (2011) the scan statistic approach is compared to a number of local and global graph invariants. In this work the scan statistic has higher power than other approaches for several random graph models. However, it is still not clear how to determine the best approach for a given class of random graphs or type of anomaly. Certainly some anomalies do not lend themselves well to using the neighborhoods of vertices as the scan region. For example, Neil (2011) looks for anomalies built from paths. Tailoring the scan region, and the statistic, to a particular application is an area that needs further investigation.

In unpublished work, Priebe has investigated the issue of which of the definitions of local regions (such as those depicted in Figure 1) have more power for detecting certain types of anomalies in certain types of random graphs, but much work still needs to be done. As shown in Wan et al. (2006a), in the Enron example, the different locality regions can produce different detections, and understanding the properties and tradeoffs of these is an important area of research.

The question of the scale of the anomaly is also an important one: what value of k should be used to compute N_k? How should one perform a multi-scale analysis, utilizing several values of k to find anomalies at different scales? More generally, how can one combine the results for regions at different scales, however the regions are defined?

As mentioned above, using the neighborhood of a vertex is essentially taking the position that the anomaly is a result of the actions of a single entity, or at least is focused in the region of a single entity. One could mitigate this by defining a scan region as the union of neighborhoods of a small collection of vertices (using larger values of k is one way to do this), but this can result in a combinatorial explosion of regions unless

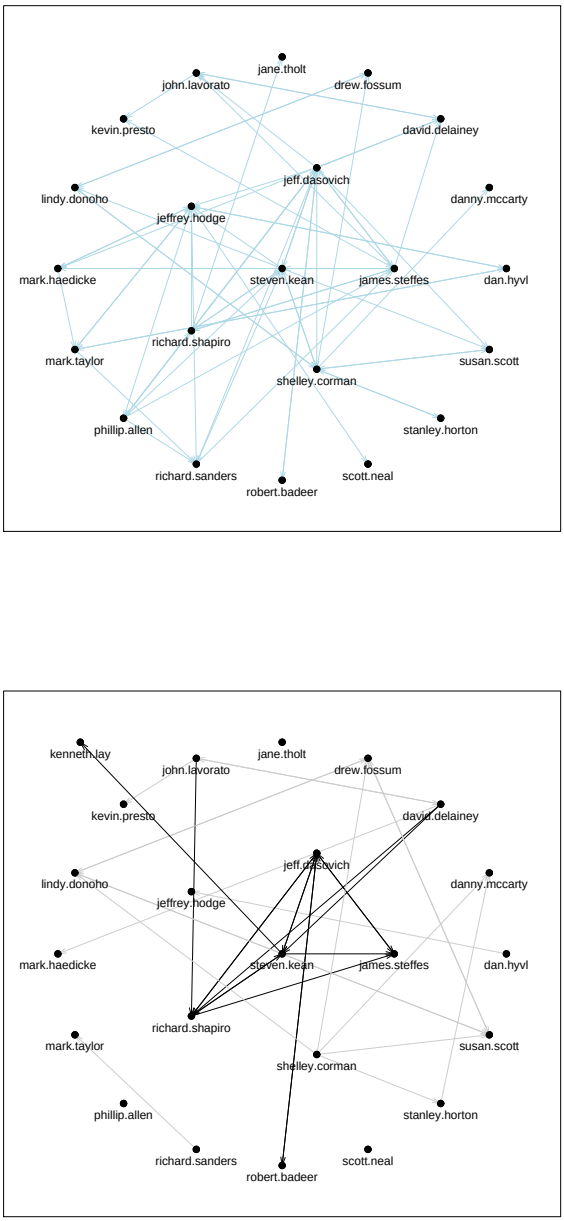

Figure 5: The central vertex v, it's neighbors $N_1(v)$ and their neighbors $N_2(v)$ at the time of the detection. The bottom plot shows the neighborhood for the prior week (dark edges); the light edges in this plot are the other edges between these individuals in this week.

care is taken. Other approaches to defining the local region, analogous to using non-rectangular regions in image analysis, would be of interest. The work of Arias-Castro et al. (2011) and Neil (2011) is relevant here.

Given that one has defined the locality region to be used in an application, how does one design the graph invariant to optimally detect the desired type of anomaly? How can one ensure robustness of this invariant (or the locality region) to misspecifications of the random graph model or anomaly? In Johannsen (2012) a very different graph invariant is introduced (actually, a matrix of invariants is computed),[3] however, it is not clear how one could go about determining the situations in which this class of invariants (or any other given invariant) is optimal, except through Monte Carlo simulation. Clearly more analytical methods are needed.

Conclusion

This article discusses an extension of the scan statistic to graphs. The basic idea is to use the graph structure to define the local region, just as spatial scan statistics use the spatial structure of the data to define the window. The main difference is the use of graph invariants (computed on the induced subgraph of the local region) rather than a variable attached to the positions. This is analogous to point process analysis, where one counts the number of points within the region. We discussed the extension of the scan statistic to time series of graphs, and illustrated this on the Enron emails graphs. This idea that an anomalous activity at a particular time within a particular group of individuals has wide applicability in fraud detection, law enforcement, homeland security, and similar domains. An area of important research is to better design the regions and the statistic for the detection of particular types of anomalies, particularly in graphs that are attributed with information about the edges or the vertices. Some of the references touch on attributed graphs, and the interested reader is encouraged to seek these out.

Acknowledgments

This was work funded by the Naval Surface Warfare Center In-House Laboratory Independent Research Program.

References

Ery Arias-Castro, Emmanuel J. Candès, and Arnaud Durand. Detection of an anomalous cluster in a network. *Annals of Statistics*, 39(1):278–304, 2011.

Nash Borges, Glen A. Coppersmith, Gerard G. L. Meyer and Carey E. Priebe. Anomaly Detection for Random Graphs using Distributions of Vertex Invariants.

[3] See the mfr package on CRAN (http://cran.r-project.org/) for source code to compute this invariant.

*P*roceedings of the 45th Annual Conference on Information Sciences and Systems, 1–6, 2011.

Jie Chen and Joseph Glaz. Two-dimensional discrete scan statistics. *Statistics and Probability Letters*, 31:59–68, 1996.

Noel A. C. Cressie. On some properties of the scan statistic on the circle and the line. *Journal of Applied Probability*, 14:272–283, 1977.

Noel A. C. Cressie. The asymptotic distribution of the scan statistic under uniformity. *Annals of Probability*, 8:828–840, 1980.

Peter J. Diggle. *Statistical Analysis of Spatial Point Patterns*. Academic Press, New York, 1983.

R. A. Fisher, H. G. Thornton, and W. A. Mackenzie. The accuracy of the plating method of estimating the density of bacterial populations, with particular reference to the use of Thornton's agar medium with soil samples. *Annals of Applied Biology*, 9:325–359, 1922.

Joseph Glaz, Joseph Naus, and Sylvan Wallenstein. *Scan Statistics*. Springer, New York, 2001.

John Grothendieck, Carey E. Priebe, and Allen L. Gorin. Statistical inference on attributed random graphs: Fusion of graph features and content. *Computational Statistics and Data Analysis*, 54:1777–1790, 2010.

David A. Johannsen and David J. Marchette. Betti numbers of graphs with an application to anomaly detection. *Statistical Analysis and Data Mining*, to appear, 2012.

Martin Kulldorff. A spatial scan statistic. *Communications in Statistics – Theory and Methods*, 26(6):1481–1496, 1997.

Nam H. Lee and Carey E. Priebe. A latent process model for time series of attributed random graphs. *Statistical Inference for Stochastic Processes*, 14: 231–253, 2011.

Clive R. Loader. Large-deviation approximations to the distribution of scan statistics. *Advances in Applied Probability*, 23:751–771, 1991.

Daniel Q. Naiman and Carey E. Priebe. Computing scan statistic p–values using importance sampling, with applications to genetics and medical image analysis. *Journal of Computational and Graphical Statistics*, 10:296–328, 2001.

Joshua Neil. Scan statistics for the online detection of locally anomalous subgraphs. Dissertation, Dept. of Mathematics and Statistics, University of New Mexico, July, 2011.

Henry Pao, Glen Coppersmith, and Carey E. Priebe. Statistical inference on random graphs: Comparative power analysis via monte carlo. *Journal of Computational and Graphical Statistics*, 20:395–416, 2011.

G. P. Patil and C. Taillie. Geographic and network surveillance via scan statistics for critical area detection. *Statistical Science*, 18(4):457 – 465, 2003.

Carey E. Priebe. Scan statistics on graphs. Technical Report 650, Johns Hopkins University, Baltimore, MD 21218-2682, 2004.

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

Carey E. Priebe, Youngser Park, David J. Marchette, John M. Conroy, John Grothendieck, and Allen L. Gorin. Statistical inference on attributed random graphs: Fusion of graph features and content: An experiment on time series of Enron graphs. *Computational Statistics and Data Analysis*, 54:1766–1776, 2010.

Andrey Rukhin and Carey E. Priebe. On the limiting distribution of a graph scan statistic. *Communications in Statistics: Theory and Methods*, 41:1151–1170, 2012.

Xiaomeng Wan, Jeannette Janssen, and Nauser Kalyaniwalla. Statistical analysis of dynamic graphs. In *Proceedings of AISB06: Adaptation in Artificial and Biological Systems*, volume 3, pages 176–179, 2006a.

Xiaomeng Wan, Jeannette Janssen, Nauser Kalyaniwalla, and Evangelos Milios. Statistical analysis of dynamic communications graphs. In *Network Analysis in Natural Sciences and Engineering*, 2006b. URL `http://www.aisb.org.uk/convention/aisb06/`.

Duncan J. Watts. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, 2003.

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

Cross-References

Spatial statistics, Anomaly detection, Social network analysis