Available online at www.sciencedirect.com
ScienceDirect
Discrete Mathematics 308 (2008) 2034–2037
DISCRETE MATHEMATICS
www.elsevier.com/locate/disc

Note

On the anomalous behaviour of a class of locality statistics

C.E. Priebe[a], W.D. Wallis[b]

[a] *Johns Hopkins University, USA*
[b] *Southern Illinois University, USA*

Received 18 July 2006; received in revised form 5 April 2007; accepted 19 April 2007
Available online 6 May 2007

Abstract

A scan statistic methodology for detecting anomalies has been developed for application to graphs, where "anomalies" are equated with vertices that exhibit distinctive local connectivity properties. We present an "anomaly graph" construction that illustrates the capabilities of these scan statistics via the behaviour of their associated locality statistics on our anomaly graphs.

Keywords: Scan statistic; Anomaly graph

1. Introduction

1.1. Scan statistics

Scan statistics (also known as "moving window analysis") is a statistical inference methodology in which a window is scanned about a data field, a locality statistic is calculated based on the data in each window—e.g, the mean for an image or a time series, or the number of events for a point pattern—and the maximum of these locality statistics is compared against some appropriate extreme value null distribution. This approach has long been used to detect anomalies—local regions of excessive activity—in spatial or temporal data. There is a vast literature on this methodology; see, for instance, the survey book [1] for historical context, development and applications.

Recently, an analogous methodology for detecting anomalies has been developed for application to graphs, where "anomalies" are equated with vertices that exhibit distinctive local connectivity properties [3,4].

We assume the standard ideas of graph theory. The *distance* between two vertices v, u in a graph is defined to be the number of edges in the shortest path from v to u. The *closed k-neighbourhood* of a vertex v is defined as

$$N_k[v] = \{u \in V : d(v, u) \leqslant k\}.$$

We define the collection of scale-k locality statistics $\{\Psi_k(v)\}_{v\in V}$ to be given by the size (number of edges, $|\cdot|$) of the subgraph induced by the closed k-neighbourhood of v

$$\Psi_k(v) = |\Omega(N_k[v])|.$$

E-mail address: wdwallis@math.siu.edu (W. Wallis).

0012-365X/$ - see front matter
doi:10.1016/j.disc.2007.04.037

The scale-k scan statistic $M_k(G)$ is then defined to be the maximum over v of the scale-k locality statistics

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

In an abuse of notation, we define $\Psi_0(v)$ to be the degree of vertex v in G, and $M_0(G)$ to be the maximum degree in G.

Large values of $M_k(G)$, with "large" dictated by the distribution of M_k under some appropriate homogeneous random graph null hypothesis, are used to detect anomalies, i.e. the existence of local regions of excessive activity, or more local connectivity than would be expected under the null hypothesis.

The vertices associated with these anomalies, elements of the set

$$V_k^*(G) = \arg\max_{v \in V} \Psi_k(v),$$

are potentially operating under some alternative model H_A and may be candidates for further investigation by subsequent processes. More generally, outliers amongst the $\{\Psi_k(v)\}_{v \in V}$ are anomalies. However, outliers with unusually *small* locality statistics would need to be investigated by other methods, and are not the subject of this study.

1.2. Anomaly graphs

The purpose of this article is to present an "anomaly graph" construction that illustrates the capabilities of the scan statistics $\{M_k(G): k = 0, 1, 2, 3, \ldots\}$ via the behaviour of their associated locality statistics $\{\Psi_k(v): v \in V, k = 0, 1, 2, 3, \ldots\}$.

That is, we construct anomaly graphs G such that, for some integer $K \geqslant 2$, G has the properties:

(P1) *Locality homogeneity* for all scales $k < K$: for $k < K$, there exists a constant c_k such that $\Psi_k(v) = c_k$ for all $v \in V$—that is, these scale-specific locality statistics are constant across vertices;

(P2) *Unique and dramatic champion* for scale K: there exist a constant c_K and a distinguished vertex v^* such that $\Psi_K(v) = c_K$ for all $v \neq v^*$ and $\Psi_K(v^*) \gg c_K$—that is, the scale-K locality statistic is constant across vertices except for v^* and is dramatically larger for the distinguished vertex v^*.

Graphs satisfying properties (P1) and (P2) are graphs for which there is a clear outlier—the unique and dramatic champion v^*—amongst the scale-K locality statistics $\{\Psi_K(v)\}_{v \in V}$ and no outliers amongst locality statistics for any smaller scale; thus the scale-K scan statistic M_K will detect the anomaly while no other scale-specific scan statistic M_k with $k < K$ will. These anomaly graphs, then, provide examples for which the higher-scale scan statistics are a necessary detection tool.

In the case $K = 1$, it is not known whether graphs with the properties (P1) and (P2) exist. One interesting family is the following set of graphs $G(1, r)$. For integer $r \geqslant 2$, $G(1, r)$ has $n = 4r + 1$ vertices labelled $0, 1, 2, \ldots, 4r$. Vertex 0 is adjacent to vertices $1, 2, \ldots, r, 2r + 1, 2r + 2, \ldots, 3r$. Vertex i is adjacent to vertex $j + 2r$ when $1 \leqslant i, j \leqslant 2r$, *except* i is not adjacent to $i + 2r$ when $1 \leqslant i \leqslant r$. It is easily seen that $G(1, r)$ is a regular graph of degree $2r$. Moreover $\Psi_1(0, G(1, r)) = r^2 + r$, $\Psi_1(i, G(1, r)) = 3r - 1$ when $1 \leqslant i \leqslant r$ or $2r + 1 \leqslant i \leqslant 3r$, and $\Psi_1(i, G(1, r)) = 2r$ otherwise. So vertex 0 is a unique outlier; the scale 1 locality statistics of the other vertices are of the order n, while $\Psi_1(0, G(1, r))$ is of order n^2. Considerably more work on the case $K = 1$ will appear in a forthcoming paper [2].

2. Construction of anomaly graphs

We present now the construction of the class of anomaly graphs $G_{K,r}$ for integers $K \geqslant 2$ and $r \geqslant 1$. $G_{K,r}$ is constructed from $2r + 1$ depth-K $2r$-ary trees T_i, where the subscripts K and r are integers mod $2r + 1$, another vertex v^*, and the following additional edges: the root of each tree is joined to v^*, and the $(2r)^{K-1}$ leaves of tree T_i are connected to the $(2r)^{K-1}$ leaves of the trees $T_{(i-1)}$ and $T_{(i+1)}$ in r-regular bipartite fashion. This can be done in many ways: for example, the $2r$ leaves with a common parent could arbitrarily be partitioned into two r-sets, and the members of each such set in T_i could be joined to the members of one of the sets in $T_{(i-1)}$ and one of the sets in $T_{(i+1)}$. See Figs. 1 and 2.

Theorem 1. *Graph $G_{K,r}$ has properties* (P1) *and* (P2).

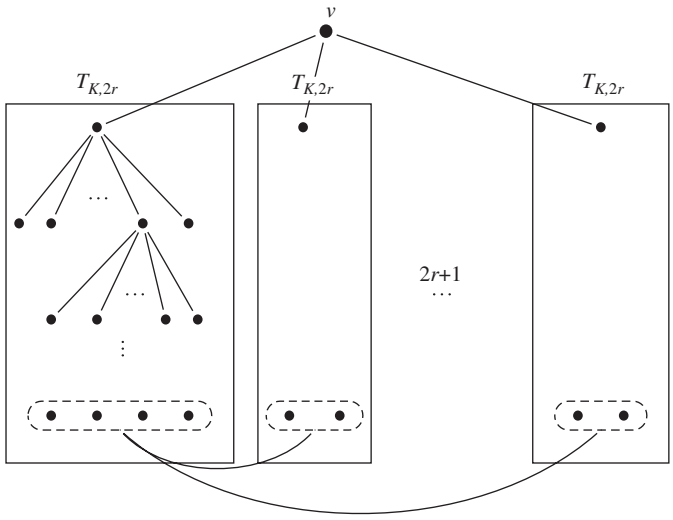

Fig. 1. Illustration of the construction of anomaly graph $G_{K,r}$.

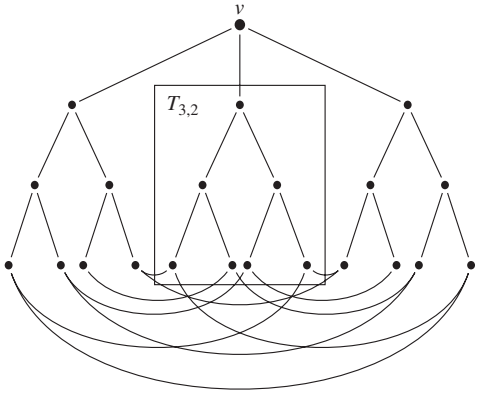

Fig. 2. Illustration of anomaly graph $G_{K,r}$ for $K = 3$, $r = 1$.

Proof. Graph $G_{K,r}$ has order $n = |V| = 1 + (2r + 1)\sum_{k=0}^{K-1} (2r)^k$ and size $|E| = n(2r + 1)/2$, and has the property that, for all v,

$$\Psi_0(v) = \Psi_1(v) = 2r + 1$$

(that is, the graph is $(2r + 1)$-regular and triangle free) and for $k = 2, \ldots, K - 1$

$$\Psi_k(v) = \Psi_{k-1}(v) + (2r + 1)(2r)^{k-1}.$$

Thus $G_{K,r}$ has the locality homogeneity property P1 for all scales $k < K$.

For $v \neq v^*$,

$$\Psi_K(v) = \Psi_{K-1}(v) + (2r + 1)(2r)^{K-1},$$

whereas

$$\Psi_K(v^*) = \Psi_{K-1}(v^*) + (2r + 1)(2r)^{K-1} + (2r + 1)(2r)^K/2 = |E| = n(2r + 1)/2.$$

Thus $G_{K,r}$ has the unique and dramatic champion property P2 for scale K. □

3. Discussion

The construction we have presented herein demonstrates the utility of higher-scale scan statistics for the detection of anomalies. Our anomaly graph $G_{K,r}$ provides an example for which the collection of scale-K locality statistics has a unique and dramatic champion—$\Psi_K(v^*)$ is a clear outlier amongst the $\{\Psi_K(v)\}_{v\in V}$—while all other scale-specific locality statistics Ψ_k with $k < K$ are constant. Thus the scan statistic $M_K(G)$ will detect the anomaly while no $M_k(G)$ for $k < K$ will do so. The statistical inference implication is that investigation at larger scales is necessary—a collection of scale-specific scan statistics $\{M_k\}_{k\in\mathcal{K}}$ is required in order to provide satisfactory performance in applications of anomaly detection in graphs.

References

[1] J. Glaz, J. Naus, S. Wallenstein, Scan Statistics, Springer, Berlin, 2001.
[2] J. McSorley, C.E. Priebe, W.D. Wallis, Neighborhood Champions, in preparation.
[3] C.E. Priebe, Scan statistics on graphs, Technical Report #650, The Johns Hopkins University, Baltimore, MD, 2004.
[4] C.E. Priebe, J.M. Conroy, D.J. Marchette, Y. Park, Scan statistics on Enron graphs, Comput. Math. Organization Theory 11 (2005) 229–247.