# Anomaly Graphs and Champions

J. P. McSorley*, C. E. Priebe†, W. D. Wallis*

**Abstract**

A scan statistic methodology for detecting anomalies has been developed for application to graphs. We equate *anomalies* with vertices that exhibit high local connectivity properties. In particular we look for cases where all vertices have similar local connectivity, except for one vertex (a *champion*) that has much higher connectivity at a certain level. For example, a *neighborhood champion* is a vertex whose closed neighborhood is larger than those of other vertices; a scale $k$ champion is a vertex whose distance $k$ closed neighborhood is larger than those of other vertices. An *anomaly graph* is a graph with a scale $k$ champion, in which all neighborhoods are the same size at distance $h$ when $h < k$, and the distance $k$ closed neighborhoods of the non-champions are of equal size.

We shall survey the constructions of anomaly graphs and more general results on neighborhood champions.

# 1 Introduction

## 1.1 Scan Statistics

Scan statistics provide 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

*Southern Illinois University
†Johns Hopkins University

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 [4, 5].

We assume the standard ideas of graph theory. We sometimes specify the vertex-set $V$ and edge-set $E$ of a graph $G$ by denoting the graph $V(G, E)$. $|V|$ and $|E|$ are respectively the *order* and *size* of $G$. The complete graph on $n$ vertices is $K_n$, while $K_{m,n}$ denotes the complete bipartite graph with vertex-sets of sizes $m$ and $n$. The *distance* $d(v, u)$ between two vertices $v, u$ in a graph is defined to be the number of edges in a shortest path from $v$ to $u$. The *closed $k$-neighborhood* of a vertex $v$ is defined as

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

When $k = 1$ we simply use the term "closed neighborhood." We sometimes say a vertex *sees* the edges in its closed neighborhood (or *sees at level $k$* the vertices in its closed $k$-neighborhood).

The scale-$k$ locality statistic $\{\Psi_k(v)\}_{v \in V}$ of a graph $G(V, E)$ was defined in [6] to be the size of the subgraph induced by the closed $k$-neighborhood of $v$:

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

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

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

In a mild 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.

An *anomaly graph* (for scale $K$) is a graph $G$ such that, for some integer $K \geq 2$:
(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 $\Psi_K(v^*) >> c_K$ — that is, the scale-$K$ locality statistic is constant across vertices except for $v^*$ and is dramatically larger for the distinguished vertex $v^*$. An anomaly graph possesses a unique and dramatic champion $v^*$, a clear outlier 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 an anomaly while no other scale-specific scan statistic $M_k$ with $k < K$ will do so.

# 2 Families of anomaly graphs

## 2.1 Anomaly graphs with $K > 1$

The following construction, from [6], yields graphs $G_{K,r}$ for integers $K \geq 2$ and $r \geq 1$; $G_{K,r}$ is an anomaly graph for scale $K$, except for $G_{3,1}$. $G_{K,r}$ 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 $T_{(i+1)}$. The construction is shown in Figure 1. It may be shown that $G_{K,r}$ has properties (P1) and (P2) unless $K = 3, r = 1$ (in that case, the vertices at level 2 are all equal "champions").

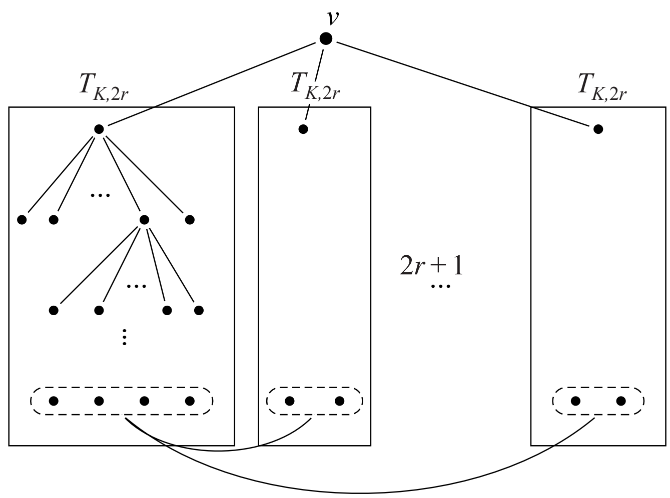


Figure 1: Illustration of the construction of anomaly graph $G_{K,r}$.

## 2.2 Some anomaly graphs with $K = 1$

The existence of anomaly graphs for $K = 1$ appears to be a difficult problem. We shall present some infinite families of graphs that have a unique vertex $v^*$ for which $\Psi_1(v^*) >> \Psi_1(v)$ for all vertices $v \neq v^*$. We do not satisfy (P2), because $\Psi_K(v)$ is not constant, although (P1) is satisfied in some cases (in this case (P1) is simply regularity). These results will appear in [2].

## 2.3 First construction

In our first construction, the degree is approximately half the order of the graph.

We first form a graph with $4n + 1$ vertices, regular of degree $2n$, for positive integer $n$.

We start with a complete bipartite graph $K_{2n,2n}$ on $4n$ vertices, say 1, 2, $\cdots$, $4n$, where each of vertices 1, 2, $\cdots$, $2n$ is joined to each of $2n+1$, $2n+2$, $\cdots$, $4n$ (and no others). We add a new vertex, 0, join it to each of 1, 2, $\cdots$, $n$, $2n+1$, $2n+2$, $\cdots$, $3n$. We then delete edges $(1, 2n+1)$, $(1, 2n+1)$, $\ldots$, $(n, 3n)$. Formally:

(i) vertices are 0, 1, 2, $\cdots$, $4n$;

(ii) 0 is adjacent to 1, 2, $\cdots$, $n$, $2n+1$, $2n+2$, $\cdots$, $3n$;

(iii) $i$ is adjacent to $j+2n$ for $1 \le i, j \le 2n$, *except* $i$ is not adjacent to $i+2n$ when $\le i \le n$.

$\Psi_1(0) = n^2 + n$, $\Psi_1(i) = 3n - 1$ when $1 \le i \le n$ or $2n+1 \le i \le 3n$, and $\Psi_1(i) = 2n$ otherwise. So 0 is a neighborhood champion.

This construction is useful when $n \ge 2$, that is the number of vertices is at least 9. It is of course more important for large $n$.

If the number of vertices is congruent to 3 modulo 4, say $4n+3$, a similar construction is available. The vertices are 0, 1, 2, $\cdots$, $4n+2$. Vertex 0 is adjacent to $1, 2, \cdots, n$, $2n+2$, $2n+2$, $\cdots$, $3n+1$. The other vertices form a $K_{2n+1,2n+1}$ with vertices 1, 2, $\cdots$, $2n+1$ joined to each of $2n+2$, $2n+3$, $\cdots$, $4n+2$, except each edge $(i, i+2n+1)$ is deleted, as are the edges $(1, 2n+3)$, $(2, 2n+4)$, $\ldots$, $(n-1, 3n+1)$, $(n, 2n+2)$. Then $\Psi_1(0) = n^2$, $\Psi_1(i) = 3n-2$ when $1 \le i \le n$ or $2n+2 \le i \le 3n+1$, and $\Psi_1(i) = 2n$ otherwise.

If the number of vertices is even, interesting graphs can be constructed by deleting one vertex (other than 0). Although not regular, these graphs have neighborhood champions and are almost regular. We shall refer to them as approximation graphs.

Another interesting construction is available when there are $4n+3$ vertices; again, call them 0, 1, 2, $\cdots$, $4n+2$. Vertex 0 is adjacent to $1, 2, \cdots, n$, $2n+2$, $2n+2$, $\cdots$, $3n+1$. The other vertices form a $K_{2n+1,2n+1}$ with vertices 1, 2, $\cdots$, $2n+1$ joined to each of $2n+2$, $2n+3$, $\cdots$, $4n+2$, except each edge $(i, i+2n+1)$ is deleted for $1 \le i \le n$. Then $\Psi_1(0) = n^2+n$, $\Psi_1(i) = 3n$ when $1 \le i \le n$ or $2n+2 \le i \le 3n+1$, and $\Psi_1(i) = 2n$ otherwise.

This graph would be regular except for the fact that the *champion* vertex, vertex 0, has degree **one smaller** than all the others.

### 2.4 Second Construction

We now present an infinite family of regular graphs with neighborhood champions, in which the degree is relatively small. We form a graph with $tn + 1$ vertices, regular of degree $2n$. Clearly $2n$ can be arbitrarily small compared to $tn + 1$.

We start with $t$ sets $S_1, S_2, \ldots, S_t$ (where these subscripts are integers modulo $t$), each containing $n$ vertices: write

$$S_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}.$$

We add another vertex $x_0$. Then:

(i) $x_0$ is adjacent to all members of $S_1 \cup S_2$;

(ii) Each member of $S_i$ is adjacent to each member of $S_{i-1} \cup S_{i+1}$ *except*

(iii) $x_{1j}$ is *not* adjacent to $x_{2j}$ for $1 \leq j \leq n$.

$\Psi_1(x_0) = n^2 + n$, $\Psi_1(x_{ij}) = 3n - 2$ when $i = 1$ or 2, and $\Psi_1(x_{ij}) = 2n$ otherwise.

This construction is useful when $t \geq 4$. It is of course more important for large $n$.

If the required number of vertices is not congruent to 1 modulo $n$, approximation graphs can be constructed by adding one vertex to one, two, ... or all of the sets $S_i$. The degree will be $2n+1$ or $2n+2$ for some vertices.

## 3 More theoretical results

In this section, which presents results from [3], we focus on the pure graph theory of the situation, and consider the existence of neighborhood champions for scale 1 in connected regular graphs. We shall sometimes discuss graphs in which more than one vertex attains the maximum value $M_1(G)$. We shall use the word “co-champion” to denote these vertices.

**Theorem 1** *For $d = 1, 2,$ and 3 there are no $d$-regular graphs with a neighborhood champion.*

**Proof** Clearly regular graphs of degrees $d = 1$ or 2 have no champions.

Now suppose $G$ is a cubic graph: If $M_1(G) = 3$ then *every* vertex attains $M_1(G)$. If $M_1(G) = 4$, then any vertex $x$ with $\Psi_1(x) = 4$ lies in exactly one triangle, and the other vertices of the triangle are co-champions; so $G$ contains at least three co-champions. If $M_1(G) = 5$ and vertex $x$ sees five edges, the configuration must be as shown in Figure 2, where $y$ is a co-champion. And if $M_1(G) = 6$ (the maximum) we have $G = K_4$, and every vertex is a co-champion. Thus no cubic graph has a champion. $\square$

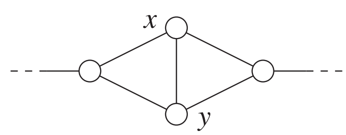


Figure 2: A cubic configuration

However, one can construct cubic graphs with precisely two co-champions, or *twin champions*, for every even number $n \geq 10$ of vertices. From above we must have $M_1(G) = 5$.

A short exhaustive search shows that this is impossible for fewer than 10 vertices (the graphs may be found on page 127 of [7]).

For every even $n \geq 10$ we construct a cubic graph $G$ on $n$ vertices for which $\Psi_1(x) = M_1(G) = 5$ for precisely two vertices. Our technique is to implant the graph shown in Figure 3 as a subgraph of a host graph. The implant graph $H$ has six vertices $a, b, p, q, y, z$ and adjacencies $ap$, $bq$, $yp$, $yq$, $zp$, $zq$, $yz$.

**Construction** Select any triangle-free cubic graph on $n - 4$ vertices and choose any edge $ab$ in that graph. Delete this edge. Then identify vertices $a, b$ with the vertices $a, b$ of $H$. See Figure 4 for an example of this; the value of $\Psi_1(x)$ is shown on each vertex and the champion is emphasized.

$\Psi_1(y) = \Psi_1(z) = 5$, $\Psi_1(p) = \Psi_1(q) = 4$ and $\Psi_1(x) = 3$ otherwise.

To show that the construction is always possible for $n \geq 10$, we observe

**Lemma 1** *If $n - 4 = 2s \geq 6$, there is a triangle-free cubic graph on $n - 4$ vertices.*

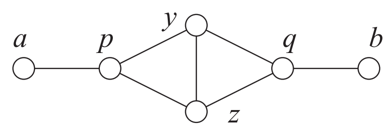


Figure 3: The graph $H$ to be implanted

**Proof** Take the integers modulo $2s$ as vertices. For each $i$, let vertex $i$ be adjacent to vertices $i-1$, $i+1$, and $i+s$ (modulo $2s$). □

(This graph is called a *Möbius ladder* [7, p263].)

So we have

**Theorem 2** *For every even $n \geq 10$ there exists a cubic graph on $n$ vertices with precisely two co-champions.*

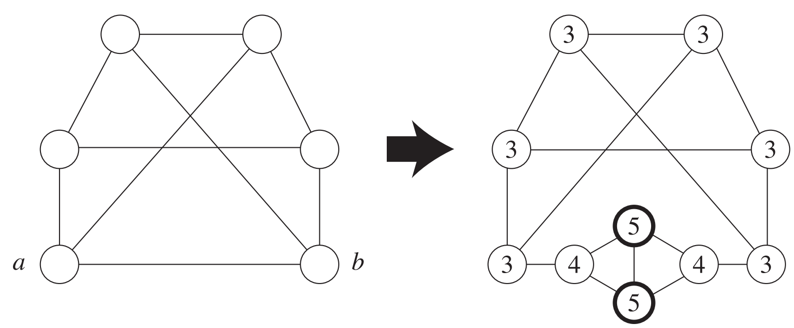


Figure 4: Example for 10 vertices

## 3.1 General Constructions: $d \geq 4$

For even $d \geq 4$ let $n_0(c, d)$ be the smallest number such that for *every* $n \geq n_0(c, d)$ there exists a $n$ vertex $d$-regular graph with precisely $c$ neighborhood co-champions; for odd $d$ we require existence for even $n \geq n_0(c, d)$ only.

In this section we discuss $n_0(1, d)$.

A *one-factor* is a graph consisting of disjoint edges; in particular, given two ordered sets of vertices $Y = \{y_0, y_1, \ldots, y_{n-1}\}$ and $Z = \{z_0, z_1, \ldots, z_{n-1}\}$, we define the one-factor $F_j^n(Y, Z)$ to consist of the edges

$$y_0 z_j, y_1 z_{1+j}, , \ldots, y_{n-1} z_{n-1+j},$$

where subscripts are reduced modulo $n$. Then $K_{n,n}$ can be represented as

$$F_0^n(Y, Z) \cup F_1^n(Y, Z) \cup \ldots F_{n-1}^n(Y, Z).$$

**Lemma 2** *Suppose $d \geq 4$. For every $t \geq 0$ there exists a $d$-regular graph, with a neighborhood champion, on $n = 3d + 2t + 1$ vertices.*

**Proof** Let $H$ represent the complete graph on the $d+1$ vertices $x_0, x_1, \ldots,$ $x_d$ with the $d$ edges of the cycle $x_0 x_1 x_2 \ldots x_{d-1}$ deleted. Take a copy of $F_0^n(Y, Z) \cup F_1^n(Y, Z) \cup \ldots F_{d-1}^n(Y, Z)$, where $n = d + t$, and delete the edges $y_0 z_0, y_1 z_1, \ldots,$ $y_{d-1} z_{d-1}$. Adjoin this to $H$ by adding edges $x_0 y_0, x_0 z_0, x_1 y_1,$ $x_1 z_1, \ldots, x_{d-1} y_{d-1}, x_{d-1} z_{d-1}$. In this graph,

$$\begin{aligned} \Psi_1(x_d) &= d(d-1)/2, \\ \Psi_1(x_i) &= (d^2 - 5d + 14)/2, \text{ for } 0 \leq i \leq d-1, \\ \Psi_1(y_j) &= \Psi_1(z_j) \; = \; d, \text{ for } 0 \leq j \leq n-1. \end{aligned}$$

Then $x_d$ is a champion provided $d(d-1)/2 > (d^2 - 5d + 14)/2$, that is $d \geq 4$. $\square$

The above construction gives graphs whose order is of opposite parity to $d$. When $d$ is odd, this provides all possible orders from some point on, because regular graphs of odd degree must have even order. However, for even degree, another construction is needed for even orders.

Suppose $G$ is the graph of Lemma 2 in the case where $d \geq 4$ is even. We modify $G$ to form $\hat{G}$ as follows: Add a vertex $\hat{x}$. Delete the $d/2$ edges $x_0 y_0$, $x_2 y_2$, $x_4 y_4$, $\ldots, x_{d-2} y_{d-2}$, and add the $d$ edges $\hat{x} x_0, \hat{x} y_0, \hat{x} x_2, \hat{x} y_2, \hat{x} x_4, \hat{x} y_4,$ $\ldots, \hat{x} x_{d-2}, \hat{x} y_{d-2}$.

The $\Psi_1$ values of all vertices of $G$ are unchanged. We have $\Psi_1(\hat{x}) =$ $d + \binom{d/2}{2} = (d^2 + 6d)/8$. Thus vertex $x_d$ is still the champion, and we have

**Lemma 3** *Suppose $d \geq 4$ is even. For every $t \geq 0$ there exists a $d$-regular graph, with a neighborhood champion, on $n = 3d + 2t + 2$ vertices.*

**Theorem 3** *Suppose $d \geq 4$. Then*

$$d + 3 \leq n_0(1, d) \leq 3d + 1.$$

**Proof** For any $d \geq 4$ the only $d$-regular graph with $d+1$ vertices is $K_{d+1}$, which clearly does not have a unique champion. And for odd $d \geq 4$ there is no $d$-regular graph with $d+2$ vertices, so $n_0(1, d) \geq d+3$. And for even $d \geq 4$ the only $d$-regular graph with $d+2$ vertices is $K_{d+2}$ minus a one-factor, which again doesn't have a unique champion; so $n_0(1, d) \geq d + 3$ here also. Hence, for any $d \geq 4$, we have $n_0(1, d) \geq d + 3$. The upper bound $n_0(1, d) \leq 3d + 1$ comes from Lemmas 2 and 3. □

## 3.2 Small degrees, $d = 4, 5$

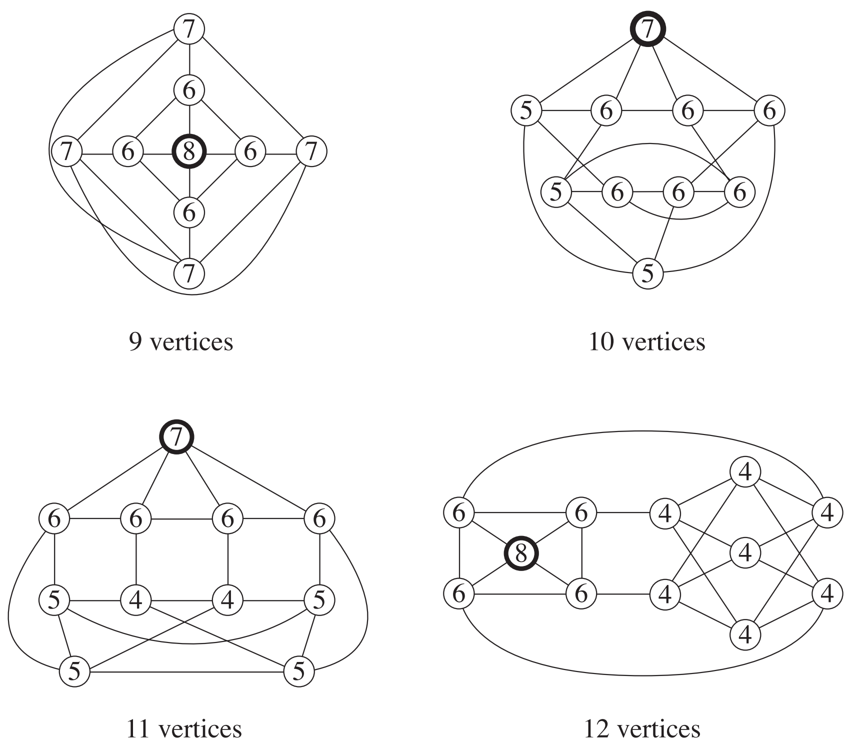


Figure 5: Small quartic graphs, each with a champion

By inspection, there are no 4-regular (quartic) graphs on $n = 7$ or 8 vertices with a unique champion (see [7, p145]). From Theorem 3 and the examples for orders $n = 9, 10, 11$ and 12 shown in Figure 5 we see that $n_0(4) = 9$, *i.e.,* there is a 4-regular graph, with a neighborhood champion, on $n$ vertices whenever $n \geq 9$.

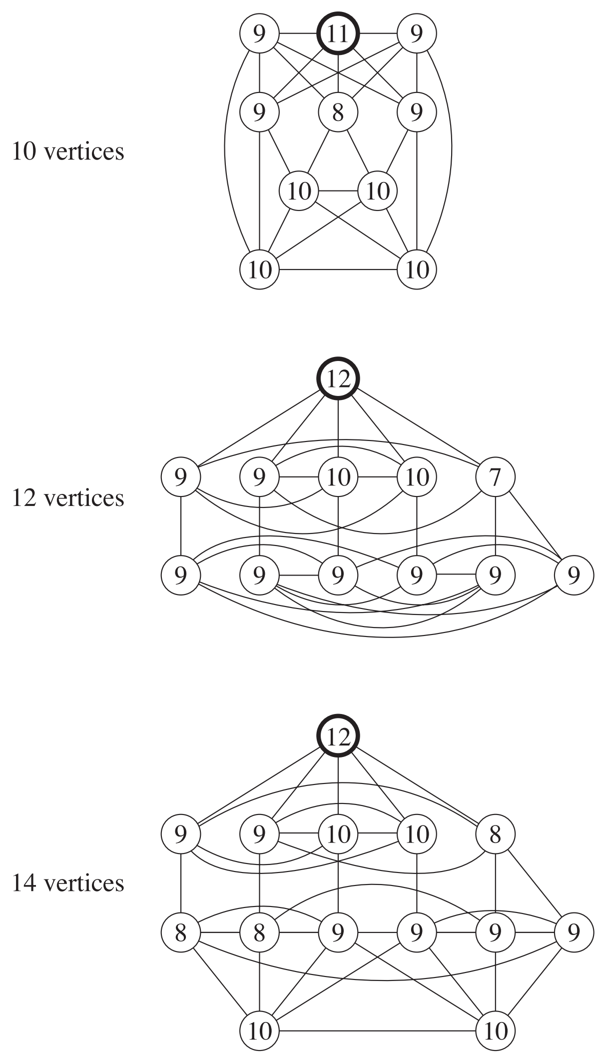


Figure 6: Small quintic graphs, each with a champion

Similarly, at degree 5, inspection shows (see [7, p154]) there are no quintic graphs on $n = 6$ or 8 vertices with a unique champion. We present examples on 10, 12 and 14 vertices in Figure 6, showing that $n_0(5) = 10$. Thus there is a 5-regular graph, with a neighborhood champion, on $n$ vertices for every even $n \geq 10$. So the cases of $d = 4$ or 5 are completely solved.

# References


[1] J. Glaz, J. Naus, and S. Wallenstein. *Scan Statistics* (Springer, 2001).

[2] J. P. McSorley, C. E. Priebe, and W. D. Wallis, Neighborhood Champions. (In preparation).

[3] J. McSorley and W. D. Wallis. Neighborhood champions in regular graphs. *J. Combin. Math. Combin. Comput.* (to appear).

[4] C. E. Priebe. *Scan statistics on graphs*. Technical report #650, The Johns Hopkins University (Baltimore, MD, 2004).

[5] C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park. Scan statistics on Enron graphs. *Comp. Math. Organization Theory*, **11**(2005), 229–247.

[6] C. E. Priebe and W. D. Wallis. On the Anomalous Behaviour of a Class of Locality Statistics. *Discrete Math.* **308**(2008), 2034–2037.

[7] R. C. Read and R. J. Wilson, *An Atlas of Graphs* (Oxford U.P., 1999).