A note on robustness of D-optimal block designs for two-colour microarray experiments

R. A. Bailey[a,*], Katharina Schiffl[b], Ralf-Dieter Hilgers[c]

[a] *School of Mathematical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK*
[b] *Roche Diagnostics, 82377 Penzberg, Germany*
[c] *RWTH Aachen University, Department of Medical Statistics, 52074 Aachen, Germany*

Abstract

Two-colour microarray experiments form an important tool in gene expression analysis. Due to the high risk of missing observations in microarray experiments, it is fundamental to concentrate not only on optimal designs but also on designs which are robust against missing observations. As an extension of Latif et al. (2009), we define the *optimal* breakdown number for a collection of designs to describe the robustness, and we calculate the breakdown number for various D-optimal block designs. We show that, for certain values of the numbers of treatments and arrays, the designs which are D-optimal have the highest breakdown number. Our calculations use methods from graph theory.

Keywords: Breakdown number, Design of microarray experiments, D-optimality, Graph theory, Robustness
2008 MSC: 62K05, 62K10, 05C40

1. Introduction

Microarrays play a key role in modern molecular biology since they enable simultaneous monitoring of the expression levels of thousands of genes: see, for example, Brown and Botstein (1999). The main goal of cDNA microarray experiments is to identify significantly up- or down-regulated genes.

[*] Corresponding author

Email addresses: r.a.bailey@qmul.ac.uk (R. A. Bailey), kschiffl@web.de (Katharina Schiffl), rhilgers@ukaachen.de (Ralf-Dieter Hilgers)

These genes serve as possible targets for therapy for severe diseases, such as malignant tumours. Since two samples of different treatments are coloured green and red and are applied (hybridized) onto one microarray, designs for microarray experiments can be considered as row–column designs with two rows corresponding to the two dyes. Ignoring the two different dyes, the designs can be considered as incomplete-block designs with block size two, provided that the number of treatments exceeds two.

Microarray experiments have been widely studied in the literature. For instance, Kerr et al. (2000) first recommended analysing microarray data with analysis-of-variance models. Most articles focus on the derivation of optimal designs in specific scenarios, but only a few authors address the problem of missing values. Missing values often occur in microarray experiments, for example due to insufficient resolution, image corruption, or simply dust or scratches on the slide (Latif et al., 2009). Thus, this data cannot be involved in the analysis of the experiment (Troyanskaya et al., 2001) and so it is important to use robust experimental designs, which ensure precise estimation of the treatment effects even if observations are missing. Latif et al. (2009) investigated specific robustness properties of commonly used microarray designs. They proposed two robustness criteria and calculated these criteria for the commonly used designs. However, to date no attempts have been made to investigate these robustness criteria analytically. We will derive an upper bound for the breakdown number, which enables us to define an optimal breakdown number and then investigate some published optimal designs with respect to the breakdown number.

This paper is structured as follows. Section 2 introduces the statistical model which is used to describe microarray experiments. Robustness criteria are defined in Section 3, where optimal robustness properties are derived. Section 4 shows that several families of published D-optimal designs achieve the optimal breakdown number. A short conclusion is given in Section 5.

2. Preliminaries

Suppose that there are t treatments and a arrays. The statistical analysis is based on the gene-specific model

$$\log_2(y_{ij\ell}) = \tau_i + \alpha_j + \delta_\ell + \epsilon_{ij\ell}, \tag{1}$$

where $y_{ij\ell}$ describes the intensity of treatment i coloured in dye ℓ on array j, for $i \in \{0, \ldots, t\}$, $j \in \{1, \ldots, a\}$ and $\ell \in \{\text{green}, \text{red}\}$, and $\epsilon_{ij\ell}$ are the error terms.

Suppose that array j has treatments i and k coloured green and red respectively. For analysis using intra-array information only, model (1) can be replaced by:

$$\log_2\left(\frac{y_{ij\text{green}}}{y_{kj\text{red}}}\right) = \tau_i - \tau_k + \delta_{\text{green}} - \delta_{\text{red}} + \epsilon_{ij\text{green}} - \epsilon_{kj\text{red}}. \tag{2}$$

As in Bailey (2007, Sections 2–6), we ignore the dye effect in the consideration of robustness and optimality. Then, written in matrix notation, model (2) simplifies to

$$z = X\tau + \eta, \tag{3}$$

where $z = (z_1, \ldots, z_a)$ is the a-dimensional vector of log ratios of the dye intensities, $\tau = (\tau_1, \ldots, \tau_t)$ is the t-dimensional vector of unknown treatment effects, and X is the $a \times t$ design matrix, with each row containing exactly one 1 and one -1, all other entries being equal to zero. The term η is the random error vector with independent identically distributed components having expectation zero and variance σ^2.

In most situations one is interested in estimating all linear contrasts of the parameter vector τ. A design is called *connected* if all linear contrasts in τ are estimable. If the design matrix is X then the matrix $X^\top X$ is called the *information matrix* of the design. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{t-1} \geq \lambda_t$ be the eigenvalues of $X^\top X$; these are non-negative. The entries in each row of $X^\top X$ sum to zero, so $\lambda_t = 0$. It can be shown (Shah and Sinha, 1989) that the design is connected if and only if the remaining $t-1$ eigenvalues are non-zero. In this case, the vector τ is estimable in the hyperplane $\sum_i \tau_i = 0$, and the volume of the confidence ellipsoid for τ is inversely proportional to $\sqrt{\prod_{i=1}^{t-1} \lambda_i}$. Thus, a design is called *D-optimal* if it maximizes the value of $\prod_{i=1}^{t-1} \lambda_i$.

3. Optimal breakdown number

Latif et al. (2009) introduced the *breakdown number* for microarray experiments, but they did not derive designs with optimal breakdown numbers for given values of t and a. Adapting their definition to the case where all linear contrasts are to be estimated gives the following.

Definition. Assume the model (3), with $a \times t$ design matrix X. Given any subset S of $\{1, \ldots, a\}$, let X_S be the design matrix obtained from X by deleting the rows corresponding to the arrays in S. The *breakdown number* of the design is equal to m if all contrasts are estimable with reduced design matrix X_S for all subsets S of size $m-1$ (that is, with $m-1$ missing observations) but there exists at least one subset S of size m for which not all contrasts are estimable.

Note that, for $1 \leq n \leq a$, every design matrix X with a rows gives $\binom{a}{n}$ matrices X_S.

Since designs with large breakdown numbers can be considered robust, we aim to search for designs which maximize the breakdown number. Let $\Omega_{t,a,2}$ be the collection of all designs for t treatments using a arrays of size two.

Definition. A design in $\Omega_{t,a,2}$ has *optimal breakdown number* if it maximizes the breakdown number over all designs in $\Omega_{t,a,2}$.

Each design in $\Omega_{t,a,2}$ can be considered as a graph with vertices 1, …, t. The number of edges joining distinct vertices i and k is equal to the number of arrays where treatments i and k are applied. The design is connected if and only if the graph is connected. It is convenient to extend the notation $\Omega_{t,a,2}$ to denote this set of graphs. See Bollabás (1979) for basic ideas and vocabulary for graph theory, but note that he uses the word ‘multigraph’ for what we call a graph. We shall call a graph *simple* if there is at most one edge between each pair of distinct vertices.

As noted by Bailey (2007), the breakdown number is a well-known concept in graph theory known as *edge-connectivity.* The edge-connectivity of a graph is defined to be the minimal number of edges whose removal results in a disconnected graph. Thus, the graph-theoretical expression ‘edge-connectivity’ is exactly the same as the breakdown number introduced by Latif et al. (2009). Further properties regarding edge-connectivity can be found in Bollabás (1979, Chapter 3).

If $a < t-1$ then the design is disconnected, and so its breakdown number is 0. If $a = t-1$ then the only connected graphs are *trees*; that is, graphs with no cycles. In this case, removal of any edge disconnects the graph, and so the breakdown number is 1.

The following theorems give the upper bound for the breakdown number for given numbers a and t with $a \geq t$. Here $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x, and $\deg(i)$ denotes the degree of vertex i, that is, the number of edges incident with vertex i.

Theorem 1. *If G is a connected graph, its breakdown number is less than or equal to the minimal degree of a vertex in G.*

Proof. Removal of all the edges incident with vertex i disconnects the graph, so the breakdown number of G is less than or equal to $\deg(i)$ for all vertices i.

Theorem 2. *Let G be a graph in $\Omega_{t,a,2}$, where $3 \leq t \leq a$. Then the breakdown number of G is less or equal to $\lfloor 2a/t \rfloor$. Moreover, equality holds for at least one graph in $\Omega_{t,a,2}$.*

Proof. Since $\sum_{i=1}^{t} \deg(i) = 2a$, there must be at least one vertex i for which $\deg(i) \leq \lfloor 2a/t \rfloor$. Theorem 1 shows that the breakdown number is at most $\lfloor 2a/t \rfloor$.

Now we construct a graph G whose breakdown number m achieves this maximum. Suppose that $2a = ct + u$, where $0 \leq u < t$. Then $c \geq 2$, because $a \geq t$. Since $u/2 < t/2$, there is at least one vertex of degree c or less, so $m \leq c = \lfloor 2a/t \rfloor$. Now we proceed separately in the cases that c is even or odd.

If $c = 2e$ with $e \geq 1$ then construct G from a cycle H of length t by using e copies of each edge; the remaining $u/2$ edges are arbitrary. If any $2e - 1$ edges are removed from G, this leaves at least one copy of at least $t-1$ edges of H, so the remaining graph is connected. Thus $m \geq 2e = c$ and hence $m = c = \lfloor 2a/t \rfloor$.

On the other hand, suppose that $c = 2e + 1$ with $e \geq 1$. As before, use e copies of each edge of H. If t is even, insert a further $t/2$ edges between all pairs of vertices at maximal distance in H. If t is odd then $u \geq 1$ so insert a further $(t + 1)/2$ edges between some pairs of vertices at maximal distance in H, in such a way that each vertex is incident with at least one such edge. In both cases, the remaining edges are arbitrary. If any $2e$ edges are removed, this leaves at least one copy of at least $t - 2$ edges of H. If all copies of two edges are removed then this splits the vertices of H into two components, but these are joined by at least one of the 'maximal distance' edges. Hence $m \geq 2e + 1 = c$ and so $m = c = \lfloor 2a/t \rfloor$.

A design is called *equireplicate*, and the corresponding graph *regular*, if all vertices have the same degree. This is not possible unless t divides $2a$, so we call a design in $\Omega_{t,a,2}$ *nearly equireplicate* if every vertex has degree $\lfloor 2a/t \rfloor$ or $\lfloor 2a/t \rfloor + 1$. Theorem 2 suggests that nearly equireplicate designs are good candidates for robust designs. However, Figure 3 of Bailey and Cameron (2009) shows an equireplicate design in $\Omega_{10,15,2}$ whose breakdown number is equal to 1. Moreover, there are some designs which are not nearly equireplicate but whose breakdown number achieves the upper bound: for example, when $t = 4$ and $a = 3$ the tree with edges $\{1, 2\}$, $\{1, 3\}$ and $\{1, 4\}$ has optimal breakdown number (namely, 1) but it is not nearly equireplicate.

John and Mitchell (1977) called an equireplicate design a *regular-graph design* if there is an integer λ such that every pair of distinct treatments occur together in either λ or $\lambda + 1$ blocks. To allow for the case that t does not divide $2a$, we define $\Omega^*_{t,a,2}$ to be the sub-collection of graphs in $\Omega_{t,a,2}$ for which every vertex has degree $\lfloor 2a/t \rfloor$ or $\lfloor 2a/t \rfloor + 1$ and there is some number λ such that the number of edges joining any pair of distinct vertices is either λ or $\lambda + 1$. Cheng et al. (1985) called such designs *nearly balanced.*

In what follows, we make frequent use of two results from the literature. A design in $\Omega_{t,a,2}$ is *balanced* if there is a positive integer λ such that every pair of distinct treatments are applied to exactly λ arrays. In graph-theortical terms, this is a *complete* graph with λ copies of each edge. For a balanced design, $a = \lambda t(t-1)/2$ and so $\lfloor 2a/t \rfloor = 2a/t = \lambda(t-1)$; moreover, every treatment occurs in r arrays, where $r = \lambda(t-1)$. Theorem 2 of Ghosh (1982) states that balanced incomplete-block designs are robust against the unavailability of all observations in any $r - 1$ blocks. In our notation $r = 2a/t$, so this theorem shows that the breakdown number for balanced designs is $2a/t$, which is the upper bound from Theorem 2.

The other result is the following version of Menger's Theorem (see Bollobás, 1979, Chapter 3). Two paths in a graph are *independent* if they have no edge in common. Menger's Theorem states that the maximum number of independent paths connecting distinct vertices i and k in a graph G is equal to the minimal number of edges whose removal disconnects i and k in G.

Theorem 3. *Let H be the graph with t vertices and λ edges between each pair of vertices, where $t \geq 3$ and $\lambda \geq 1$.*

- *(i) If G is obtained from H by inserting one copy of each of s mutually non-adjacent edges, where $1 \leq s < t/2$, then the breakdown number of G is $\lambda(t-1)$, and this cannot be improved upon.*

- *(ii) If t is even and G is obtained from H by inserting one copy of each of*

$t/2$ mutually non-adjacent edges, then the breakdown number of G is $\lambda(t-1)+1$, and this cannot be improved upon.

(iii) If G is obtained from H by removing one copy of each of s mutually non-adjacent edges, where $1 \leq s \leq t/2$, then the breakdown number of G is $\lambda(t-1)-1$, and this cannot be improved upon.

(iv) If G is obtained from H by inserting one copy of every edge in G_0, where G_0 has optimal breakdown number in $\Omega_{t,s,2}$, then G has optimal breakdown number in $\Omega_{t,a,2}$, where $a = \lambda t(t-1)/2 + s$.

Proof. Let m and m' be the breakdown numbers of G and H respectively. It follows from Ghosh (1982) that $m' = \lambda(t-1)$.

(i) Inserting edges cannot decrease the breakdown number, so $m \geq m' = \lambda(t-1)$. On the other hand, $s < t/2$, so G has at least one vertex of degree $\lambda(t-1)$, so Theorem 1 shows that $m \leq \lambda(t-1)$. Also, $2a = \lambda t(t-1) + 2s < \lambda t(t-1) + t$, so $\lfloor 2a/t \rfloor = \lambda(t-1) = m$, and so the upper bound from Theorem 2 is achieved.

(ii) Now $2a/t = \lambda(t-1)+1$, so Theorem 1 shows that $m \leq \lambda(t-1)+1$. Let i and k be any distinct vertices. If $\{i,k\}$ is one of the extra edges, then vertices i and k are connected by $\lambda+1$ paths of length 1 and by $\lambda(t-2)$ paths of length 2 (these have the form (i,j,k) for $j \notin \{i,k\}$). If $\{i,k\}$ is not one of the extra edges, and the extra edges through i and k are $\{i,i'\}$ and $\{k,k'\}$ respectively, then there are only λ paths of length 1 connecting i and k but there is now the path (i,i',k',k) of length three. In both cases, vertices i and k are connected by at least $\lambda(t-1)+1$ independent paths. Menger's Theorem shows that $m \geq \lambda(t-1)+1$. Hence $m = \lambda(t-1)+1$.

(iii) Now the minimal degree of G is $\lambda(t-1)-1$, and so Theorem 1 shows that $m \leq \lambda(t-1)-1$. An argument similar to the one in part (ii) shows that if one copy of $\{i,k\}$ is removed then there are $\lambda(t-1)-1$ independent paths connecting i and k. If $\{i,k\}$ is not removed but one copy of either or both of $\{i,i'\}$ and $\{k,k'\}$ is removed then the paths (i,i',k) and (i,k',k) may be lost but can be replaced by (i,k',i',k), so there are still at least $\lambda(t-1)-1$ independent paths connecting i and k. Thus Menger's Theorem shows that $m \geq \lambda(t-1)-1$, and so

$m = \lambda(t-1) - 1$, which once again achieves the upper bound from Theorem 2.

(iv) Let m'' be the breakdown number of G_0. By Theorems 1 and 2, G_0 has a vertex of degree m'', and so G has a vertex of degree $\lambda(t-1) + m''$: therefore $m \leq \lambda(t-1) + m''$. If i and k are any distinct vertices then they are connected by $\lambda(t-1)$ independent paths in H and by at least m'' independent paths in G_0: hence $m \geq \lambda(t-1) + m''$, and so $m = \lambda(t-1) + m''$. Now, $2a = \lambda t(t-1) + 2s$, and so $\lfloor 2a/t \rfloor = \lambda(t-1) + \lfloor 2s/t \rfloor = \lambda(t-1) + m'' = m$. Thus G has optimal breakdown number in $\Omega_{t,a,2}$.

4. Some classes of D-optimal designs

Many authors have derived D-optimal block designs for various different scenarios, but few attempts have been made to investigate the breakdown number of these designs when all blocks have size two. Ghosh (1982) and Bhaumik and Whittinghill (1991) considered balanced incomplete-block designs and variance-balanced block designs: when all blocks have size two then these are just one or more copies of the complete graph. Baksalary and Tabis (1987) and Godolphin and Warren (2011) give some sufficient conditions for a design to achieve the optimal breakdown number, but these conditions are rarely satisfied when all blocks have size two.

In this section we calculate the breakdown number for some known classes of D-optimal designs with block-size two, and we show that these designs have optimal breakdown number.

The relationship between graph theory and optimal design theory was described by Gaffke (1982): see also Cheng (1981) and Bailey and Cameron (2009). In particular, a subgraph H of a graph G is called a *spanning tree* for G if H is a tree which includes every vertex of G. Gaffke (1982) showed that a block design is D-optimal if and only if it maximizes the number of spanning trees.

4.1. Small number of arrays

When $a = t-1$ then the only connected block designs are the trees. Bailey (2007) pointed out that all trees are D-optimal. In Section 3 we observed that all trees have optimal breakdown number in $\Omega_{t,t-1,2}$: this number is 1.

Bailey (2007) also showed that the D-optimal designs when $a = t$ are the cycles. Theorem 2 shows that these designs have optimal breakdown number (which is 2) in $\Omega_{t,t,2}$.

We now extend this result to all designs for which $t \leq a < 3t/2$. Note that a *bridge* in a connected graph is a single edge whose removal disconnects the graph.

Theorem 4. *Let G be the graph corresponding to a D-optimal design in $\Omega_{t,a,2}$, where $a \geq t \geq 3$. Then G does not contain a bridge. In particular, no vertex of G has degree less than 2.*

Proof. Suppose that the edge $\{i, k\}$ is a bridge of G. Removing this bridge splits G into two components H and K, where $i \in H$ and $k \in K$. Every spanning tree for G consists of a spanning tree for H, the edge $\{i, k\}$, and a spanning tree for K. Hence the number of spanning trees for G is uv, where u and v are the numbers of spanning trees for H and K respectively. This number is positive, because G is connected.

Since $a \geq t$, there is at least one edge e in G which is in a cycle. Without loss of generality, e is in component H. Let u' be the number of spanning trees for H which do not contain e. Since e is in a cycle, $u' > 0$. Create a new graph H' by inserting a new vertex j into e. Every spanning tree for H which contains e gives a spanning tree for H' containing both edges at j; every spanning tree for H which does not contain e gives two spanning trees for H', one containing each edge at j. Hence the number of spanning trees for H' is $u - u' + 2u' = u + u' > u$.

Create a new graph G' by replacing H by H', removing the bridge, and identifying the vertices i and k. Then G' has t vertices and a edges. Every spanning tree for G' consists of a spanning tree for H' with a spanning tree for K. Hence G' has $(u + u')v$ spanning trees. This number is greater than uv, so G cannot be D-optimal.

Corollary 1. *If $t \leq a < 3t/2$ then every D-optimal design in $\Omega_{t,a,2}$ has breakdown number 2, which is the upper bound from Theorem 2.*

4.2. Bipartite graphs and related designs

A graph is called *bipartite* if its vertices can be partitioned into two parts such that no vertices in the same part are adjacent. Each edge is incident with one vertex from each part. A simple bipartite graph is *complete bipartite* if every vertex is adjacent to all vertices in the other part. A complete bipartite graph is regular if and only if the two parts have the same size.

Proposition 1. *Let $t = 2u$. If G is a regular complete bipartite graph with t vertices then every pair of distinct vertices in G can be connected by u independent paths.*

Proof. Label the vertices of one part $i_1, \dots i_u$ and those of the other part k_1, …, k_u. Vertices i_1 and i_2 are connected by the u independent paths (i_1, k_j, i_2) for $j = 1, \dots, u$. Vertices i_1 and k_1 are connected by the u independent paths (i_1, k_1) and (i_1, k_j, i_j, k_1) for $j = 2, \dots, u$.

Cheng (1981) found several classes of D-optimal designs. His Theorem 2.1 shows that when $t = 2u$ and $\lambda \geq 0$ then the graph formed from a complete regular bipartite graph with t vertices by inserting λ further edges between each pair of vertices is D-optimal. We now calculate the breakdown number for these designs.

Proposition 2. *Let $t = 2u$, $\lambda \geq 0$ and $a = u^2 + \lambda t(t-1)/2$. Let H be the graph with t vertices and λ edges between each pair of vertices, and let G be the graph obtained from H by inserting one further copy of each edge of a regular complete bipartite graph G_0 on the same vertices. Then the breakdown number of G is equal to $\lambda(t-1)+u$, and this is the optimal breakdown number for $\Omega_{t,a,2}$.*

Proof. All vertices of G_0 have degree u, so Theorem 2 and Proposition 1 show that the breakdown number of G_0 is u and that this is the optimal breakdown number for $\Omega_{t,u^2,2}$. Then Theorem 3(iv) completes the proof.

Multipartite graphs generalize bipartite graphs. If $t = mu$, with $m \geq 2$, then a simple graph with t vertices is regular complete m-partite if the vertices are partitioned into m parts of size u and each vertex is adjacent to every vertex in all other parts but to no vertex in the same part.

Proposition 3. *Let $t = mu$, where $m \geq 2$. If G is a regular complete m-partite graph with t vertices then every pair of distinct vertices in G can be connected by $(m-1)u$ independent paths.*

Proof. If i and k are different vertices in the same part, then they are connected by the $(m-1)u$ independent paths (i, j, k) for vertices j in the other parts. If i and k are in different parts, then Proposition 1 shows that they are connected by u independent paths lying within those two parts, and they are also connected by the further $(m-2)u$ independent paths (i, j, k) for vertices j in the remaining parts.

Theorem 3.1 of Cheng (1981) shows that if $m \geq 2$, $t = mu$ and $a = u^2m(m-1)/2$ then the regular complete m-partite graphs are D-optimal among simple graphs in $\Omega_{t,a,2}$.

Proposition 4. *Let $m \geq 2$, $t = mu$ and $a = u^2m(m-1)/2$. Regular complete m-partite graphs with t vertices have breakdown number $(m-1)u$, which is the optimal breakdown number for $\Omega_{t,a,2}$.*

Proof. All vertices in such a graph have degree $(m-1)u$, so Theorem 2 and Proposition 3 show that the breakdown number is $(m-1)u$ and that this is the optimal breakdown number for $\Omega_{t,a,2}$.

A multipartite graph on t vertices with s parts of size 2 and the remaining parts of size 1 is obtained from the complete graph on t vertices by deleting s mutually non-adjacent edges, where $1 \leq s \leq t/2$. Theorem 4.1 of Cheng (1981) shows that such graphs are D-optimal among simple graphs in $\Omega_{t,a,2}$, where $a = t(t-1)/2 - s$. Theorem 3(iii) shows that they have optimal breakdown number, which is $t-2$.

4.3. Small number of treatments

The case $a = t-1$ has been covered in Section 4.1. When a is a multiple of $t(t-1)/2$, the D-optimal designs are the balanced incomplete-block designs (Kiefer, 1975), which have optimal breakdown number by Theorem 2 of Ghosh (1982). From now on, we assume that $a = \lambda t(t-1)/2 + s$ with $a \geq t$ and $0 < s < t(t-1)/2$.

Theorem 2.2 of Gaffke (1982) shows that if $t \leq 5$ then there are some designs in $\Omega^*_{t,a,2}$ which are D-optimal in $\Omega_{t,a,2}$, while his Theorem 2.3 gives the same result for $t = 6$ if a is divisible by 3. (This does not imply that all D-optimal designs in $\Omega_{t,a,2}$ are in $\Omega^*_{t,a,2}$: we have already noted that this is not true when $t = 4$ and $a = 3$.) Gaffke (1982) used these results to find those D-optimal designs in $\Omega_{t,a,2}$ which are in $\Omega^*_{t,a,2}$ when $2 \leq t \leq 5$ and when $t = 6$ and a is a multiple of 3.

When $t = 4$ and $a = 6\lambda + s$ with $a \geq 4$ and $0 < s < 6$, Gaffke (1982) showed that those D-optimal designs which are in $\Omega^*_{4,a,2}$ consist of λ copies of all edges of the complete graph together with a collection of s edges isomorphic to those in Figure 1.

Proposition 5. *Suppose that $a \geq t = 4$ and that a is not divisible by 6. Then the D-optimal designs given by Gaffke (1982) have optimal breakdown number in $\Omega_{4,a,2}$.*

$\{1,2\}$	$\{1,2\}$	$\{1,2\}$ $\{2,3\}$	$\{1,2\}$ $\{2,3\}$	$\{1,2\}$ $\{1,3\}$ $\{1,4\}$
	$\{3,4\}$	$\{3,4\}$	$\{1,4\}$ $\{3,4\}$	$\{2,3\}$ $\{2,4\}$
(a) $s=1$	(b) $s=2$	(c) $s=3$	(d) $s=4$	(e) $s=5$

Figure 1: The extra edges in D-optimal designs which are also nearly balanced, when $t=4$ and $a=6\lambda+s$

Proof. Let $a=6\lambda+s$ with $0<s<6$. If $s=1$ or $s=2$ or $s=5$ then the result follows from Theorem 3(i), (ii) or (iii) respectively. If $s=3$ then the extra edges form a tree, so the result follows from Section 4.1 and Theorem 3(iv). If $s=4$ then the extra edges form a cycle: Section 4.1 shows that cycles have optimal breakdown number, so the result follows from Theorem 3(iv).

Likewise, Gaffke (1982) showed that, when $5 \leq a = 10\lambda+s$ with $0<s<10$, then the D-optimal designs in $\Omega_{5,a,2}$ which are in $\Omega^*_{5,a,2}$ are those where the extra edges in Figure 2 are added to λ copies of every edge in the complete graph.

Proposition 6. *If $a \geq t = 5$ and a is not divisible by 10 then the D-optimal designs given by Gaffke (1982) have optimal breakdown number in $\Omega_{5,a,2}$. If $a=10\lambda+s$ with $0<s<10$ then this breakdown number is equal to 4λ if $1 \leq s \leq 2$, to $4\lambda+1$ if $3 \leq s \leq 4$, to $4\lambda+2$ if $5 \leq s \leq 7$, and to $4\lambda+3$ if $8 \leq s \leq 9$.*

Proof. All of these designs are nearly equireplicate, so in every case the smallest degree of a vertex is equal to the bound in Theorem 2. When $1 \leq s \leq 2$, the result follows from Theorem 3(i). For $s=3$, the argument is similar to the proof of Theorem 3(ii). When $s=4$ or $s=5$, the extra edges form a tree or cycle respectively, so the result follows from Section 4.1 and Theorem 3(iv). For $6 \leq s \leq 7$ it can be verified directly that there are two independent paths between any pair of distinct vertices, using only the extra edges. For $8 \leq s \leq 9$, the result follows from Theorem 3(iii).

For $t=6$, Gaffke (1982) considered only families of designs containing equireplicate designs; that is, a is a multiple of 3. For these values of a, he showed that the regular-graph designs obtained by adding the extra edges in Figure 3 to copies of the complete graph are D-optimal.

$\{1,2\}$	$\{1,2\}$	$\{1,2\}$ $\{2,3\}$	$\{1,2\}$ $\{2,3\}$	$\{1,2\}$ $\{2,3\}$ $\{3,4\}$
	$\{3,4\}$	$\{4,5\}$	$\{2,4\}$ $\{4,5\}$	$\{4,5\}$ $\{5,1\}$
(a) $s=1$	(b) $s=2$	(c) $s=3$	(d) $s=4$	(e) $s=5$

$\{1,2\}$ $\{2,5\}$	$\{1,2\}$ $\{2,3\}$ $\{3,4\}$	$\{1,2\}$ $\{2,3\}$ $\{3,4\}$	$\{1,2\}$ $\{2,3\}$ $\{3,4\}$
$\{1,3\}$ $\{3,5\}$	$\{4,5\}$ $\{5,1\}$ $\{1,3\}$	$\{4,5\}$ $\{5,1\}$ $\{1,3\}$	$\{4,5\}$ $\{5,1\}$ $\{1,3\}$
$\{1,4\}$ $\{4,5\}$	$\{2,4\}$	$\{2,4\}$ $\{2,5\}$	$\{2,4\}$ $\{2,5\}$ $\{1,4\}$
(f) $s=6$	(g) $s=7$	(h) $s=8$	(i) $s=9$

Figure 2: The extra edges in D-optimal designs which are also nearly balanced, when $t=5$ and $a=10\lambda+s$

$\{1,2\}$	$\{1,2\}$ $\{2,3\}$	$\{1,4\}$ $\{1,5\}$ $\{1,6\}$	$\{1,3\}$ $\{1,4\}$ $\{1,5\}$ $\{1,6\}$
$\{3,4\}$	$\{3,4\}$ $\{4,5\}$	$\{2,4\}$ $\{2,5\}$ $\{2,6\}$	$\{2,3\}$ $\{2,4\}$ $\{2,5\}$ $\{2,6\}$
$\{5,6\}$	$\{5,6\}$ $\{6,1\}$	$\{3,4\}$ $\{3,5\}$ $\{3,6\}$	$\{3,5\}$ $\{3,6\}$ $\{4,5\}$ $\{4,6\}$
(a) $s'=1$	(b) $s'=2$	(c) $s'=3$	(d) $s'=4$

Figure 3: The extra edges in D-optimal designs which are also regular-graph designs, when $t=6$ and $a=15\lambda+3s'$

Proposition 7. *If* $6 \leq a = 15\lambda + 3s'$ *with* $0 < s' < 5$ *then the D-optimal designs in* $\Omega_{6,a,2}$ *given by Gaffke (1982) have optimal breakdown number, which is equal to* $5\lambda + s'$.

Proof. When $s'=1$ this follows from Theorem 3(ii). When $s'=2$, the extra edges form a cycle, and so the result follows from Section 4.1 and Theorem 3(iv). When $s'=3$, the extra edges form a complete bipartite graph, and so Proposition 1 shows that the bound in Theorem 2 is achieved for the extra edges: then Theorem 3(iv) gives the result. Finally, when $s'=4$ the result follows from Theorem 3(iii).

5. Conclusion

In Section 4 we showed that several classes of D-optimal designs have optimal breakdown number. The converse is not true. For example, Figure 3 of Bailey (2007) shows that there are four (isomorphism classes of) designs in $\Omega_{8,12,2}$ with optimal breakdown number but that only one of these is D-optimal. However, all four of them are better on the D-criterion than all designs with smaller breakdown number.

Our results lead us to conjecture that D-optimal block designs with block size 2 always achieve the optimal breakdown number. However, other plausible conjectures about optimal block designs have turned out to be either wrong or very hard to prove. For example, John and Mitchell (1977) conjectured that if $\Omega_{t,a,k}$ contains any regular-graph designs then all A-optimal designs in $\Omega_{t,a,k}$ are regular-graph designs: this is now known to be false. Cheng et al. (1985) proved that, for each value of t, there is a value $a_0(t)$ such that if $a \geq a_0(t)$ then all D-optimal designs in $\Omega_{t,a,2}$ are nearly balanced. We know that this is not true when $a = t-1$, but the values of $a_0(t)$ given by this theorem appear to be far larger than necessary. It may be similarly difficult to prove a general theorem about the breakdown number of D-optimal designs.

Many authors focus on the derivation of optimal designs for two-colour microarray experiments, but only a few have investigated optimal designs in settings with missing values. Latif et al. (2009) introduced the breakdown number to analyse the robustness of efficient microarray experiments. We have considered this number analytically, studied connections to graph theory, and investigated designs with optimal breakdown number. We showed that several D-optimal designs have optimal breakdown number. Although we have not proved that this holds in general, it seems prudent to recommend D-optimal block designs, as they appear to provide some safeguard against missing values.

Acknowledgments

The research was financially supported by the BMBF within the joint research project SKAVOE (Foerderkennzeichen: 03HIPAB4). Part of the work was done while RAB and RDH were at the 2011 programme on Design and Analysis of Experiments at the Isaac Newton Institute, Cambridge, UK.

References

Bailey, R. A., 2007. Designs for two-colour microarray experiments. Journal of the Royal Statistical Society, Series C (Applied Statistics) 56, 365–394.

Bailey, R. A., Cameron, P. J., 2009. Combinatorics of optimal designs. In Huczynska, S., Mitchell, J. D., Roney-Dougal, C. M. (editors), Surveys in Combinatorics 2009, London Mathematical Society Lecture Notes 365, Cambridge University Press, Cambridge, 19–73.

Baksalary, Jerzy K., Tabis, Zenon, 1987. Conditions for the robustness of block designs against the unavailability of data. Journal of Statistical Planning and Inference 16, 49–54.

Bhaumik, Dulal K., Whittinghill, Dexter C., 1991. Optimality and robustness to the unavailability of blocks in block designs. Journal of the Royal Statistical Society, Series B 53, 399–407.

Bollobás, Béla, 1979. Graph Theory: An Introductory Course, Springer, New York.

Brown, P. O., Botstein, D., 1999. Exploring the new world of the genome with DNA microarrays. Nature Genetics 21, 33–7.

Cheng, C.-S., 1981. Maximizing the total number of spanning trees in a graph: two related problems in graph theory and optimum design theory. Journal of Combinatorial Theory, Series B 31, 240–248.

Cheng, Ching-Shui, Masaro, Joseph C., Wong, Chi Song, 1985. Do nearly balanced multigraphs have more spanning trees? Journal of Graph Theory 9, 335–341.

Gaffke, N., 1982. D-optimal block designs with at most six varieties. Journal of Statistical Planning and Inference 6, 183–200.

Ghosh, S., 1982. Robustness of BIBD against the unavailability of data. Journal of Statistical Planning and Inference 6, 29–32.

Godolphin, J. D., Warren, H. R., 2011. Improved conditions for the robustness of binary block designs against the loss of whole blocks. Journal of Statistical Planning and Inference 141, 3498–3505.

John, J. A., Mitchell, T. J., 1977. Optimal incomplete block designs. Journal of the Royal Statistical Society, Series B 39, 39–43.

Kerr, M. K., Martin, M., Churchill, G. A., 2000. Analysis of variance for gene expression microarray data. J. Computational Biology 7, 819–837.

Kiefer, J., 1975. Construction and optimality of generalized Youden designs. In Srivastava, J. N. (editor), A Survey of Statistical Design and Linear Models, North-Holland, Amsterdam, 333–353.

Latif, A. H. M. M., Bretz, F., Brunner, E., 2009. Robustness considerations in selecting efficient two-color microarray designs. Bioinformatics 25, 2355–2361.

Shah, K., Sinha, B. K., 1989. Theory of Optimal Designs, Springer, New York.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R. B., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17, 520–525.