Using graphs and Laplacian eigenvalues to evaluate block designs

R. A. Bailey
University of St Andrews / QMUL (emerita)

Ongoing joint work with Peter J. Cameron

Modern Trends in Algebraic Graph Theory,
Villanova, June 2014

Abstract

Consider an experiment to compare v treatments in b blocks of size k.

Statisticians use various criteria to decide which design is best.

It turns out that most of these criteria are defined by the Laplacian eigenvalues of one of the two graphs defined by the block design:

- the Levi graph, which has $v + b$ vertices,
- or the concurrence graph, which has v vertices.

The algebraic approach shows that sometimes all the criteria prefer highly symmetric designs but sometimes they favour very different ones.

An experiment on detergents

In a consumer experiment, twelve housewives volunteer to test new detergents. There are 16 new detergents to compare, but it is not realistic to ask any one volunteer to compare this many detergents.

Each housewife tests one detergent per washload for each of four washloads, and assesses the cleanliness of each washload.

The experimental units are the 48 washloads.
The housewives form 12 blocks of size 4.

The treatments are the 16 new detergents.

Experiments in blocks

I have v treatments that I want to compare.
I have b blocks, with k plots in each block.

blocks	b	k	treatments	v
housewives	12	4	detergents	16

How should I choose a block design
for these values of b, v and k?

What makes a block design good?

Two designs with $v = 5$, $b = 7$, $k = 3$: which is better?

Conventions: columns are blocks;
order of treatments within each block is irrelevant;
order of blocks is irrelevant.

1	1	1	1	2	2	2
2	3	3	4	3	3	4
3	4	5	5	4	5	5

binary

1	1	1	1	2	2	2
1	3	3	4	3	3	4
2	4	5	5	4	5	5

non-binary

A design is binary if no treatment occurs more than once in any block.

Two designs with $v = 15$, $b = 7$, $k = 3$: which is better?

1	1	2	3	4	5	6
2	4	5	6	10	11	12
3	7	8	9	13	14	15

replications differ by ≤ 1

1	1	1	1	1	1	1
2	4	6	8	10	12	14
3	5	7	9	11	13	15

queen-bee design

The replication of a treatment is its number of occurrences.

A design is a queen-bee design if there is a treatment that occurs in every block.

Two designs with $v = 7$, $b = 7$, $k = 3$: which is better?

1	2	3	4	5	6	7
2	3	4	5	6	7	1
4	5	6	7	1	2	3

balanced (2-design)

1	2	3	4	5	6	7
2	3	4	5	6	7	1
3	4	5	6	7	1	2

non-balanced

A binary design is balanced if every pair of distinct treaments occurs together in the same number of blocks.

Experimental units and incidence matrix

There are bk experimental units.

If ω is an experimental unit, put

$$\begin{aligned} f(\omega) &= \text{treatment on } \omega \\ g(\omega) &= \text{block containing } \omega. \end{aligned}$$

For $i = 1, \ldots, v$ put

$$r_i = |\{\omega : f(\omega) = i\}| = \text{replication of treatment } i.$$

For $i = 1, \ldots, v$ and $j = 1, \ldots, b$, let

$$n_{ij} = |\{\omega : f(\omega) = i \text{ and } g(\omega) = j\}|$$

$=$ number of experimental units in block j which have treatment i.

The $v \times b$ incidence matrix N has entries n_{ij}.

Levi graph

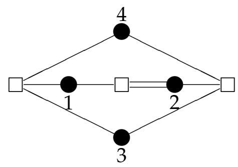
The Levi graph $\tilde{G}$ of a block design Δ has

- one vertex for each treatment,
- one vertex for each block,
- one edge for each experimental unit, with edge ω joining vertex $f(\omega)$ (the treatment on ω) to vertex $g(\omega)$ (the block containing ω).

It is a bipartite graph,
with n_{ij} edges between treatment-vertex i and block-vertex j.

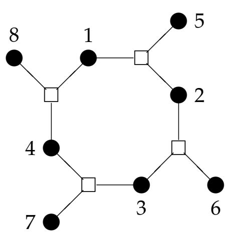
Example 1: $v = 4$, $b = k = 3$

1	2	1
3	3	2
4	4	2



Example 2: $v = 8$, $b = 4$, $k = 3$

1	2	3	4
2	3	4	1
5	6	7	8



Concurrence graph

The concurrence graph G of a block design Δ has

- one vertex for each treatment,
- one edge for each unordered pair α, ω, with $\alpha \neq \omega$, $g(\alpha) = g(\omega)$ (in the same block) and $f(\alpha) \neq f(\omega)$: this edge joins vertices $f(\alpha)$ and $f(\omega)$.

There are no loops.

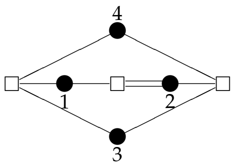
If $i \neq j$ then the number of edges between vertices i and j is

$$\lambda_{ij} = \sum_{s=1}^{b} n_{is} n_{js};$$

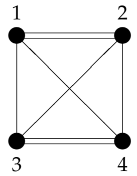
this is called the concurrence of i and j,
and is the (i, j)-entry of $\Lambda = NN^\top$.

Example 1: $v = 4$, $b = k = 3$

1	2	1
3	3	2
4	4	2



Levi graph
can recover design
more vertices
more edges if $k = 2$

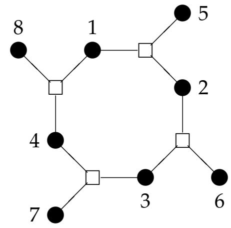


concurrence graph
may have more symmetry

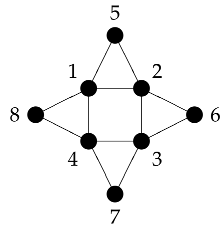
more edges if $k \geq 4$

Example 2: $v = 8$, $b = 4$, $k = 3$

1	2	3	4
2	3	4	1
5	6	7	8



Levi graph

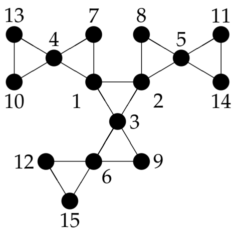


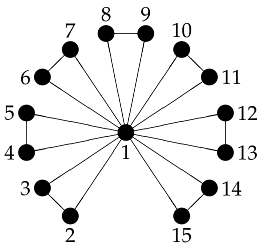
concurrence graph

Example 3: $v = 15$, $b = 7$, $k = 3$

1	1	2	3	4	5	6
2	4	5	6	10	11	12
3	7	8	9	13	14	15

1	1	1	1	1	1	1
2	4	6	8	10	12	14
3	5	7	9	11	13	15





Laplacian matrices

The Laplacian matrix L of the concurrence graph G is a $v \times v$ matrix with (i, j)-entry as follows:

- if $i \neq j$ then
  $L_{ij} = -$(number of edges between i and j) $= -\lambda_{ij}$;
- $L_{ii} =$ valency of $i = \sum_{j \neq i} \lambda_{ij}$.

The Laplacian matrix $\tilde{L}$ of the Levi graph $\tilde{G}$ is a $(v + b) \times (v + b)$ matrix with (i, j)-entry as follows:

- $\tilde{L}_{ii} =$ valency of i
  $$= \begin{cases} k & \text{if } i \text{ is a block} \\ \text{replication } r_i \text{ of } i & \text{if } i \text{ is a treatment} \end{cases}$$
- if $i \neq j$ then $L_{ij} = -$(number of edges between i and j)
  $$= \begin{cases} 0 & \text{if } i \text{ and } j \text{ are both treatments} \\ 0 & \text{if } i \text{ and } j \text{ are both blocks} \\ -n_{ij} & \text{if } i \text{ is a treatment and } j \text{ is a block, or vice versa.} \end{cases}$$

Connectivity

All row-sums of L and of $\tilde{L}$ are zero,
so both matrices have 0 as eigenvalue
on the appropriate all-1 vector.

Theorem
The following are equivalent.

1. *0 is a simple eigenvalue of L;*
2. *G is a connected graph;*
3. *$\tilde{G}$ is a connected graph;*
4. *0 is a simple eigenvalue of $\tilde{L}$;*
5. *the design Δ is connected in the sense that all differences between treatments can be estimated.*

From now on, assume connectivity.

Call the remaining eigenvalues *non-trivial*.
They are all non-negative.

Generalized inverse

Under the assumption of connectivity,
the Moore–Penrose generalized inverse L^- of L is defined by

$$L^- = \left(L + \frac{1}{v}J_v\right)^{-1} - \frac{1}{v}J_v,$$

where J_v is the $v \times v$ all-1 matrix.

(The matrix $\frac{1}{v}J_v$ is the orthogonal projector onto the null space of L.)

The Moore–Penrose generalized inverse $\tilde{L}^-$ of $\tilde{L}$ is defined similarly.

Estimation

We measure the response Y_ω on each experimental unit ω.

If experimental unit ω has treatment i and is in block m ($f(\omega) = i$ and $g(\omega) = m$), then we assume that

$$Y_\omega = \tau_i + \beta_m + \text{random noise}.$$

We will do an experiment, collect data y_ω on each experimental unit ω, then want to estimate certain functions of the treatment parameters using functions of the data.

We want to estimate contrasts $\sum_i x_i \tau_i$ with $\sum_i x_i = 0$.

In particular, we want to estimate all the simple differences $\tau_i - \tau_j$.

Variance: why does it matter?

We want to estimate all the simple differences $\tau_i - \tau_j$.

Put $V_{ij} =$ variance of the best linear unbiased estimator for $\tau_i - \tau_j$.

The length of the 95% confidence interval for $\tau_i - \tau_j$ is proportional to $\sqrt{V_{ij}}$. (If we always present results using a 95% confidence interval, then our interval will contain the true value in 19 cases out of 20.)

The smaller the value of V_{ij}, the smaller is the confidence interval, the closer is the estimate to the true value (on average), and the more likely are we to detect correctly which of τ_i and τ_j is bigger.

We can make better decisions about new drugs, about new varieties of wheat, about new engineering materials … if we make all the V_{ij} small.

How do we calculate variance?

Theorem

Assume that all the noise is independent, with variance σ^2.
If $\sum_i x_i = 0$, then the variance of the best linear unbiased estimator of $\sum_i x_i \tau_i$ is equal to

$$(x^\top L^- x)k\sigma^2.$$

In particular, the variance of the best linear unbiased estimator of the simple difference $\tau_i - \tau_j$ is

$$V_{ij} = \left(L^-_{ii} + L^-_{jj} - 2L^-_{ij}\right)k\sigma^2.$$

(This follows from assumption

$$Y_\omega = \tau_i + \beta_m + \text{random noise}.$$

by using standard theory of linear models.)

… Or we can use the Levi graph

Theorem

The variance of the best linear unbiased estimator of the simple difference $\tau_i - \tau_j$ is

$$V_{ij} = \left(\tilde{L}^-_{ii} + \tilde{L}^-_{jj} - 2\tilde{L}^-_{ij}\right)\sigma^2.$$

(Or $\beta_i - \beta_j$, appropriately labelled.)

(This follows from assumption

$$Y_\omega = \tau_i - \tilde{\beta}_m + \text{random noise}.$$

by using standard theory of linear models.)

How do we calculate these generalized inverses?

We need L^- or $\tilde{L}^-$.

- Add a suitable multiple of J,
  use GAP to find the inverse with exact rational coefficients,
  subtract that multiple of J.
- If the matrix is highly patterned,
  guess the eigenspaces,
  then invert each non-zero eigenvalue.
- Direct use of the graph: coming up.

Not all of these methods are suitable for generic designs with a variable number of treatments.

Electrical networks

We can consider the concurrence graph G as an electrical network with a 1-ohm resistance in each edge.
Connect a 1-volt battery between vertices i and j.
Current flows in the network, according to these rules.

1. Ohm's Law:
   In every edge, voltage drop = current $\times$ resistance = current.
2. Kirchhoff's Voltage Law:
   The total voltage drop from one vertex to any other vertex is the same no matter which path we take from one to the other.
3. Kirchhoff's Current Law:
   At every vertex which is not connected to the battery, the total current coming in is equal to the total current going out.

Find the total current I from i to j, then use Ohm's Law to define the effective resistance R_{ij} between i and j as $1/I$.

Electrical networks: variance

Theorem
The effective resistance R_{ij} between vertices i and j in G is

$$R_{ij} = \left(L^-_{ii} + L^-_{jj} - 2L^-_{ij} \right).$$

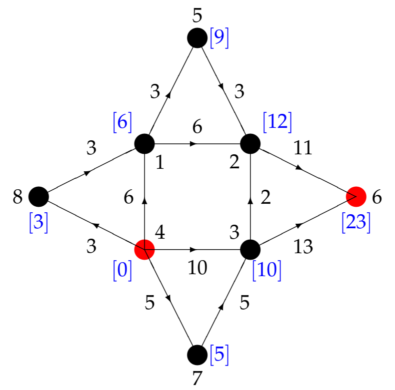
So

$$V_{ij} = R_{ij} \times k\sigma^2.$$

Effective resistances are easy to calculate without matrix inversion if the graph is sparse.

Example 2 calculation: $v = 8$, $b = 4$, $k = 3$

$$V = 23 \quad I = 24 \quad R = \frac{23}{24}$$



... Or we can use the Levi graph

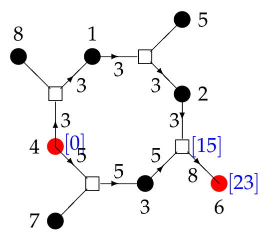
If i and j are treatment vertices in the Levi graph $\tilde{G}$ and $\tilde{R}_{ij}$ is the effective resistance between them in $\tilde{G}$ then

$$V_{ij} = \tilde{R}_{ij} \times \sigma^2.$$

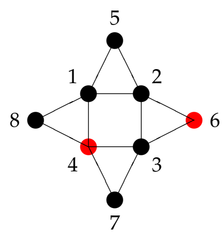
Example 2 yet again: $v = 8$, $b = 4$, $k = 3$

$$V = 23 \quad I = 8 \quad \tilde{R} = \frac{23}{8}$$

1	2	3	4
2	3	4	1
5	6	7	8



Levi graph



concurrence graph

Optimality: Average pairwise variance

The variance of the best linear unbiased estimator of the simple difference $\tau_i - \tau_j$ is

$$V_{ij} = \left(L^-_{ii} + L^-_{jj} - 2L^-_{ij} \right) k\sigma^2 = R_{ij} k\sigma^2.$$

We want all of the V_{ij} to be small.

Put $\bar{V} =$ average value of the V_{ij}. Then

$$\bar{V} = \frac{2k\sigma^2 \operatorname{Tr}(L^-)}{v-1} = 2k\sigma^2 \times \frac{1}{\text{harmonic mean of } \theta_1, \ldots, \theta_{v-1}},$$

where $\theta_1, \ldots, \theta_{v-1}$ are the nontrivial eigenvalues of L.

A-Optimality

A block design is called A-optimal if it minimizes the average of the variances V_{ij};
—equivalently, it maximizes the harmonic mean of the non-trivial eigenvalues of the Laplacian matrix L;
over all block designs with block size k and the given v and b.

Optimality: Confidence region

When $v > 2$ the generalization of confidence interval is the confidence ellipsoid around the point $(\hat{\tau}_1, \ldots, \hat{\tau}_v)$ in the hyperplane in $\mathbb{R}^v$ with $\sum_i \tau_i = 0$. The volume of this confidence ellipsoid is proportional to

$$\begin{aligned} \sqrt{\prod_{i=1}^{v-1} \frac{1}{\theta_i}} &= (\text{geometric mean of } \theta_1, \ldots, \theta_{v-1})^{-(v-1)/2} \\ &= \frac{1}{\sqrt{v \times \text{number of spanning trees for } G}}. \end{aligned}$$

D-Optimality

A block design is called D-optimal if it minimizes the volume of the confidence ellipsoid for $(\hat{\tau}_1, \ldots, \hat{\tau}_v)$;
—equivalently, it maximizes the geometric mean of the non-trivial eigenvalues of the Laplacian matrix L;
—equivalently, it maximizes the number of spanning trees for the concurrence graph G;
—equivalently, it maximizes the number of spanning trees for the Levi graph $\tilde{G}$;
over all block designs with block size k and the given v and b.

Optimality: Worst case

If x is a contrast in $\mathbb{R}^v$ then the variance of the estimator of $x^\top \tau$ is $(x^\top L^- x)k\sigma^2$.

If we multiply every entry in x by a constant c then this variance is multiplied by c^2; and so is $x^\top x$.

The worst case is for contrasts x giving the maximum value of

$$\frac{x^\top L^- x}{x^\top x}.$$

These are precisely the eigenvectors corresponding to θ_1, where θ_1 is the smallest non-trivial eigenvalue of L.

E-Optimality

A block design is called E-optimal if it maximizes the smallest non-trivial eigenvalue of the Laplacian matrix L; over all block designs with block size k and the given v and b.

BIBDs are optimal

Theorem (Kshirsagar, 1958; Kiefer, 1975)

If there is a balanced incomplete-block design (BIBD) (2-design) for v treatments in b blocks of size k,
then it is A-, D- and E-optimal.
Moreover, no non-BIBD is A-, D- or E-optimal.

D-optimality: spanning trees

A spanning tree for the graph is a collection of edges of the graph which form a tree (connected graph with no cycles) and which include every vertex.

Cheng (1981), after Gaffke (1978), after Kirchhoff (1847):

$$\begin{aligned} &\text{product of non-trivial eigenvalues of } L \\ &= v \times \text{number of spanning trees.} \end{aligned}$$

So a design is D-optimal if and only if its concurrence graph G has the maximal number of spanning trees.

This is easy to calculate by hand when the graph is sparse.

What about the Levi graph?

Theorem (Gaffke, 1982)

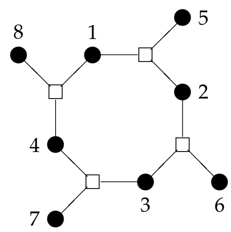
Let G and $\tilde{G}$ be the concurrence graph and Levi graph for a connected incomplete-block design for v treatments in b blocks of size k. Then the number of spanning trees for $\tilde{G}$ is equal to k^{b-v+1} times the number of spanning trees for G.

So a block design is D-optimal if and only if its Levi graph maximizes the number of spanning trees.

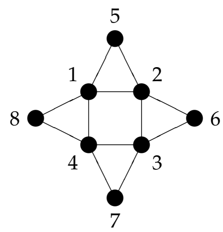
If $v \geq b+2$ it is easier to count spanning trees in the Levi graph than in the concurrence graph.

Example 2 one last time: $v = 8$, $b = 4$, $k = 3$

1	2	3	4
2	3	4	1
5	6	7	8



Levi graph
8 spanning trees



concurrence graph
216 spanning trees

E-optimality: the edge-cutset lemma

A design is E-optimal if it maximizes the smallest non-trivial eigenvalue θ_1 of the Laplacian L of the concurrence graph G.

Lemma

Let G have an edge-cutset of size c (set of c edges whose removal disconnects the graph) whose removal separates the graph into components of sizes m and n. Then

$$\theta_1 \leq c\left(\frac{1}{m}+\frac{1}{n}\right).$$

If c is small but m and n are both large, then θ_1 is small.

E-optimality: the vertex-cutset lemma

A design is E-optimal if it maximizes the smallest non-trivial eigenvalue θ_1 of the Laplacian L of the concurrence graph G.

Lemma

Let G have a vertex-cutset of size c (set of c vertices whose removal disconnects the graph) whose removal separates the graph into components of sizes m and n, with m' and n' edges between them and the vertices in the cutset. Then

$$\theta_1 \leq \frac{m'n^2 + n'm^2}{mn(n+m)},$$

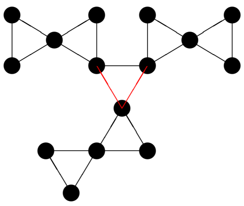
which is at most c if no multiple edges are involved.

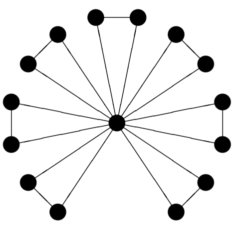
If $m' << m$ and $n' << n$ then θ_1 is small.

Minimal connectivity

If the block design is connected then $bk \geq b+v-1$.

If the block design is connected and $b(k-1) = v-1$ then the Levi graph is a tree and the concurrence graph is a b-tree of k-cliques.





Optimality of minimally connected designs

The Levi graph is a tree, so all connected designs are equally good under the D-criterion.

The Levi graph is a tree, so effective resistance = graph distance, so the only A-optimal designs are the queen-bee designs.

The concurrence graph is a b-tree of k-cliques, so the Cutset Lemmas show that the only E-optimal designs are the queen-bee designs.

Can we use the Levi graph to find E-optimal designs?

For binary designs with equal replication,
$\theta_1(L)$ is a monotonic increasing function of $\theta_1(\tilde{L})$.

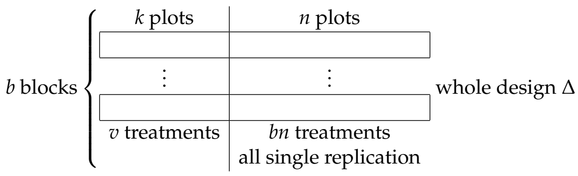
But, queen-bee designs are E-optimal under minimal connectivity,
and some non-binary designs are E-optimal.

For general block designs, we do not know if we can use the Levi graph to investigate E-optimality.

Large blocks; many unreplicated treatments

Suppose that $\bar{r} = \frac{\sum_i r_i}{v} < 2$.

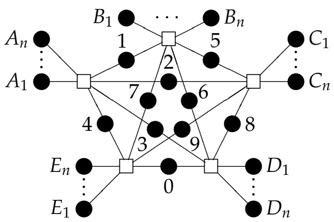
New conventions: blocks are rows, and block size $= k + n$.



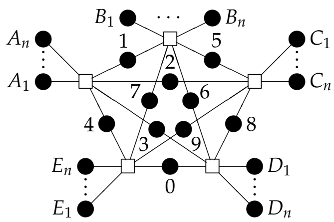
Whole design Δ has $v + bn$ treatments in b blocks of size $k + n$;
the subdesign Γ has v core treatments in b blocks of size k;
call the remaining treatments orphans.

Levi graph: $10 + 5n$ treatments in 5 blocks of $4 + n$ plots

1	2	3	4	A_1	$\cdots$	A_n
1	5	6	7	B_1	$\cdots$	B_n
2	5	8	9	C_1	$\cdots$	C_n
3	6	8	0	D_1	$\cdots$	D_n
4	7	9	0	E_1	$\cdots$	E_n



Pairwise resistance



$$\begin{aligned}
\text{Resistance}(A_1, A_2) &= 2 \\
\text{Resistance}(A_1, B_1) &= 2 + \text{Resistance}(\text{block } A, \text{block } B) \text{ in } \Gamma \\
\text{Resistance}(A_1, 8) &= 1 + \text{Resistance}(\text{block } A, 8) \text{ in } \Gamma \\
\text{Resistance}(1, 8) &= \text{Resistance}(1, 8) \text{ in } \Gamma
\end{aligned}$$

Sum of the pairwise variances

Theorem (cf Herzberg and Jarrett, 2007)
The sum of the variances of treatment differences in Δ

$$= \textit{constant} + V_1 + nV_3 + n^2 V_2,$$

where

$$\begin{aligned}
V_1 &= \textit{the sum of the variances of treatment differences in } \Gamma \\
V_2 &= \textit{the sum of the variances of block differences in } \Gamma \\
V_3 &= \textit{the sum of the variances of sums of} \\
&\quad\ \textit{one treatment and one block in } \Gamma.
\end{aligned}$$

(If Γ is equi-replicate then V_2 and V_3 are both increasing functions of V_1.)

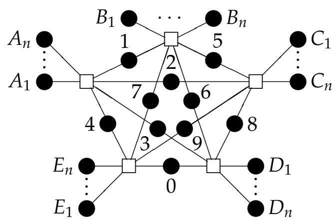
Consequence
For a given choice of k, make Γ as efficient as possible.

A less obvious consequence

Consequence
If n or b is large,
and we want an A-optimal design,
it may be best to make Γ a complete block design for k' controls,
even if there is no interest in
comparisons between new treatments and controls,
or between controls.

Spanning trees

A spanning tree for the Levi graph is a collection edges which provides a unique path between every pair of vertices.



The orphans make no difference to the number of spanning trees for the Levi graph.

D-optimality under very low replication

Consequence

The whole design Δ is D-optimal
for $v + bn$ treatments in b blocks of size $k + n$
if and only if the core design Γ is D-optimal
for v treatments in b blocks of size k.

Consequence

Even when n or b is large,
D-optimal designs do not include uninteresting controls.

Conjectures

Conjecture (Underpinned by theoretical work by C.-S. Cheng)

If the A-optimal design is very different from the D-optimal design, then the E-optimal design is (almost) the same as the A-optimal design.

Conjecture (Underpinned by theoretical work by C.-S. Cheng)

If the connectivity is more than minimal, then all D-optimal designs have (almost) equal replication.

Conjecture (Underpinned by theoretical work by J. R. Johnson and M. Walters)

If $\bar{r} > 3.5$ *then designs optimal under one criterion are (almost) optimal under the other criteria.*

Main References

- R. A. Bailey and Peter J. Cameron:
  Combinatorics of optimal designs.
  In *Surveys in Combinatorics 2009*
  (eds. S. Huczynska, J. D. Mitchell and
  C. M. Roney-Dougal),
  London Mathematical Society Lecture Note Series, 365,
  Cambridge University Press, Cambridge, 2009, pp. 19–73.
- R. A. Bailey and Peter J. Cameron:
  Using graphs to find the best block designs.
  In *Topics in Structural Graph Theory*
  (eds. L. W. Beineke and R. J. Wilson),
  Cambridge University Press, Cambridge, 2013,
  pp. 282–317.