Combinatorics of optimal designs

R. A. Bailey and Peter J. Cameron

Queen Mary
University of London

p.j.cameron@qmul.ac.uk

British Combinatorial Conference, St Andrews, July 2009

Mathematicians and statisticians

> There is a very famous joke about Bose's work in Giridh. Professor Mahalanobis wanted Bose to visit the paddy fields and advise him on sampling problems for the estimation of yield of paddy. Bose did not very much like the idea, and he used to spend most of the time at home working on combinatorial problems using Galois fields. The workers of the ISI used to make a joke about this. Whenever Professor Mahalanobis asked about Bose, his secretary would say that Bose is working in fields, which kept the Professor happy.

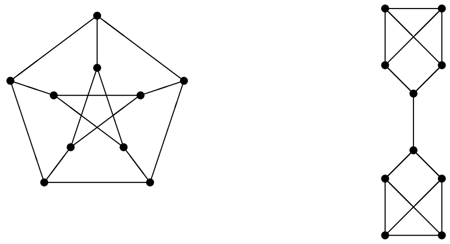
Bose memorial session, in *Sankhyā* **54** (1992) (special issue devoted to the memory of Raj Chandra Bose), i–viii.

First topic

A block design with block size 2 is just a (multi)graph.

What graph-theoretic properties make it a "good" block design (in the sense that the information obtained from an experiment is as accurate as possible given the resources?

Which graph is best?



Of course the question is not well defined. But which would you choose for a network, if you were concerned about connectivity, reliability, etc.?

Which graph is best connected?

Here are some ways of measuring the "connectivity" of a graph.

- How many spanning trees does it have? The more spanning trees, the better connected. The first graph has 2000 spanning trees, the second has 576.
- Electrical resistance. Imagine that the graph is an electrical network with each edge being a 1-ohm resistor. Now calculate the resistance between each pair of terminals, and sum over all pairs; the lower the total, the better connected. In the first graph, the sum is 33; in the second, it is 206/3.

Which graph is best connected?

- Isoperimetric number. This is defined to be

$$i(G) = \min\left\{\frac{|\partial S|}{|S|} : S \subseteq V(G),\ 0 < |S| \leq v/2\right\},$$

where, for a set S of vertices, ∂S is the set of edges from S to its complement. Large isoperimetric number means that there are many edges out of any set of vertices. The isoperimetric number for the first graph is 1 (there are just five edges between the inner and outer pentagons), that of the second graph is $1/5$ (there is just one edge between the top and bottom components).

Laplacian eigenvalues

Let G be a graph on n vertices. (Multiple edges are allowed but loops are not.)

The *Laplacian matrix* of G is the $n \times n$ matrix $L(G)$ whose (i,i) entry is the number of edges containing vertex i, while for $i \neq j$ the (i,j) entry is the negative of the number of edges joining vertices i and j.

This is a real symmetric matrix; its eigenvalues are the *Laplacian eigenvalues* of G. Note that its row sums are zero.

Laplacian eigenvalues

- $L(G)$ is positive semi-definite, so its eigenvalues are non-negative.
- The multiplicity of 0 as an eigenvalue of G is equal to the number of connected components of G. In particular, if G is connected, then 0 is a simple eigenvalue (called the *trivial eigenvalue*) corresponding to the all-1 eigenvector, and the non-trivial eigenvalues are positive.
- The number of spanning trees of G is the product of the non-trivial Laplacian eigenvalues, divided by v: this is *Kirchhoff's Matrix-Tree Theorem*.

Laplacian eigenvalues

- The sum of resistances between all pairs of vertices is the sum of reciprocals of the non-trivial Laplacian eigenvalues, multiplied by v.
- The isoperimetric number is at least half of the smallest non-trivial eigenvalue of G.

There is also an upper bound for $i(G)$ in terms of μ_1, an *inequality of Cheeger type*, which is crucial for other applications (to random walks etc.)

Recently, Krivelevich and Sudakov have shown that, in a k-regular graph G on v vertices, if μ_1 is large enough in terms of v and k, then G is Hamiltonian. Pyber used this to show that all but finitely many strongly regular graphs are Hamiltonian.

Graphs as block designs

Suppose that we have ten "treatments" that we want to compare. We have enough resources to perform fifteen trials, each one of which compares two of the treatments.

The trials can be regarded as the edges of a graph with 10 vertices and 15 edges. So our two examples are among the graphs we could use. Which will give the best possible information about treatment differences?

We model the result of each trial as giving a number for each of the two treatments in the trial, which is the sum of an effect due to a treatment, an effect due to the trial, and some random variation.

Treatment contrasts

We cannot estimate treatment effects directly, because adding the same quantity to each treatment effect and subtracting it from each trial effect will not change the results.

We can estimate *treatment differences*, or more generally *treatment contrasts*, linear combinations of treatment effects where the coefficients sum to zero.

Each treatment contrast estimator is a random variable, and the smaller its variance, the more accurate the estimate. Accurate estimates are important to reduce the risk that we rate one treatment better than another just because of random variation.

Optimality criteria

Among desirable criteria we might ask for an experimental design to do one of the following:

- minimize the average variance of the treatment differences (such a design is called *A-optimal*);

- minimize the volume of a confidence ellipsoid containing the estimated treatment contrasts (such a design is called *D-optimal*;
- minimize the maximum variance of any normalised treatment contrast (such a design is called *E-optimal*).

There are other types of optimality too, but these will do for now! (For D-optimality, we need to assume the errors are independent normal.)

Optimality and graph properties

Theorem 1. *In any given class of graphs,*

- *the A-optimal graph mimimizes the sum of resistances between all pairs of vertices;*
- *the D-optimal graph maximizes the number of spanning trees in the graph;*
- *the E-optimal graph maximizes the minimum non-trivial Laplacian eigenvalue of the graph.*

So E-optimal graphs will tend to have large isoperimetric numbers.

Second topic

A block design with block size greater than 2 is not a graph. Perhaps we should regard it as a hypergraph of some kind?

It will turn out that optimality properties of such a block design are determined by a graph, the *concurrence graph* of the block design, no matter what the block size. So we do not need a new theory!

What is a block design?

We wish to do an experiment to test v different treatments. We have available bk experimental units, divided into b "blocks" of k; there are systematic but unknown differences between the blocks. We model the response of an experimental unit as the sum of a treatment effect, a block effect, and random variation, and we want to estimate treatment differences, or more generally, treatment contrasts.

For example, we may be testing varieties of seed, and have k plots available for planting the seed on each of b farms in different geographic and climatic areas.

Mathematicians tend to represent a block design by a family of subsets of the treatment set, where each block corresponds to a set of k treatments. There are different schools of thought about whether "repeated blocks" should be allowed.

In fact there is a much more serious problem ...

An example

We have five treatments numbered $1, \ldots, 5$, and 21 experimental units, divided into seven blocks of three.

The design is given in the following table:

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

A combinatorialist wanting to represent this block design in the "traditional" way, with blocks as subsets of the set of treatments, has a problem: the first block is a *multiset* $[1,1,2]$.

Nevertheless, to a statistician there is no problem with this; indeed, it can be shown that this design is E-optimal among all designs for 5 treatments and 7 blocks of size 3.

An example, continued

Look at the example again:

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

1 and 2 occur together twice in the first block.

With this convention, you can easily check that the block design is *balanced*, that is, the equivalent of a 2-design: every pair of treatments lie together in exactly two blocks.

We have called these designs "variance-balanced designs" or *VB-designs* in the paper; some statisticians call them "completely symmetric designs" (a term unlikely to appeal to mathematicians)!

It is known that VB-designs are E-optimal as long as they don't have too much "badness" (multiple occurrences of treatments in blocks). See the paper for details.

The concurrence graph

The *concurrence graph* of a block design is defined as follows. The vertex set is the set of v treatments. There are no loops. For every occurrence of treatments i and j together in a block, we put an edge from i to j. (For example, if a block contains p occurrences of treatment i and q of treatment j, then it contributes pq edges from i to j.)

In our example, the concurrence graph is the complete multigraph on 5 vertices, where every edge has multiplicity 2.

We form the Laplacian matrix of this graph in the usual way: the (i, i) entry is the valency of vertex i; and for $i \neq j$, the (i, j) entry is the negative of the number of edges from i to j.

Estimation and variance

This topic is covered in detail in the paper. The upshot is that, in order to extract information about treatment differences from the experimental results, we require a matrix called the *information matrix* of the design, and we require its non-trivial eigenvalues to be "large".

Now in the case of a block design with v treatments and b blocks of size k, we have the following result:

Theorem 2. *The information matrix of a block design with block size k is equal to the Laplacian matrix of its concurrence graph divided by k.*

So optimality criteria can be expressed in terms of the Laplacian eigenvalues …

Optimality and Laplace eigenvalues

Let $\mathcal{D}$ be a class of connected block designs (with fixed v, b, k), and $\mathcal{G}$ the set of concurrence graphs of designs in $\mathcal{D}$.

- A design in $\mathcal{D}$ is A-optimal if and only if its concurrence graph maximizes the harmonic mean of the non-trivial Laplace eigenvalues over the class $\mathcal{G}$.
- A design in $\mathcal{D}$ is D-optimal if and only if its concurrence graph maximizes the geometric mean of the non-trivial Laplace eigenvalues over the class $\mathcal{G}$.
- A design in $\mathcal{D}$ is E-optimal if and only if its concurrence graph maximizes the minimum non-trivial Laplace eigenvalue over the class $\mathcal{G}$.

The interpretation of A- and D-optimality in terms of resistances and spanning trees is exactly as before.

Which graphs are concurrence graphs?

Let $w_1, \dots, w_m$ be positive integers with sum k. Define a *weighted clique* with weights $w_1, \dots, w_m$ in a graph to be a clique of m vertices, numbered $1, \dots, m$, such that the number of edges joining i to j is $w_i w_j$.

Theorem 3. *A graph is the concurrence graph of a block design with block size k if and only if the edge set of G can be partitioned into weighted cliques with total weight k.*

Our example corresponds to a partition of $2K_5$ into six triangles and one double edge (with weights 1 and 2).

Third topic

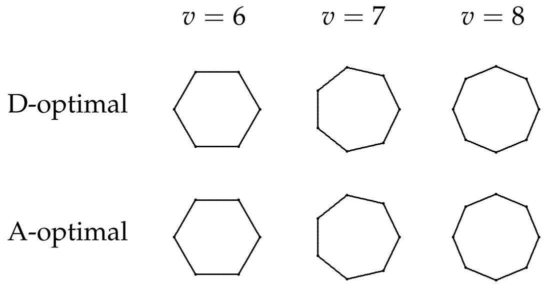
Different optimality criteria do not always agree on what is the best design to use.

We will see an example shortly, but first, here is Kiefer's theorem:

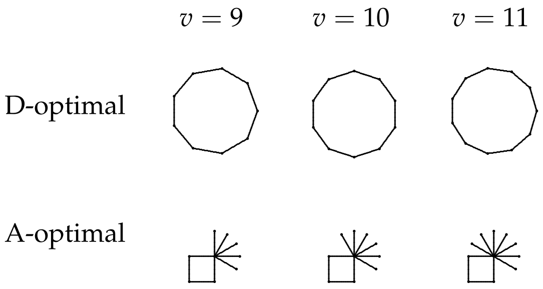
Theorem 4. *A 2-design (that is, a balanced incomplete-block design in which treatments are not repeated in blocks) is optimal with respect to the A-, D- and E-criteria (and indeed all other proposed criteria).*

Now we look at the case where $k = 2$ and $b = v$ (so the design is a unicyclic graph). What is the "nicest" unicyclic graph?

Optimal designs when $b = v$, $k = 2$



Optimal designs when $b = v, k = 2$



More generally . . .

Let us just consider the set $\mathcal{G}$ of designs with block size 2 (that is, graphs), having v vertices and e edges, where $e \geq v$.

Theorem 5.

- *A graph having a leaf cannot be D-optimal in $\mathcal{G}$.*
- *On the other hand, if* $20 \leq v \leq e < 5v/4$*, then any E-optimal graph in $\mathcal{G}$ has a leaf.*

You can find the proof in the paper.

Things to do (a short list)

- Develop an existence theory for VB-designs similar to Wilson's existence theory for 2-designs. (The number of blocks is not determined by the parameters v, k, λ; the theory should also take account of possible numbers of blocks.)
- For designs with block size 2, is there a "threshold" for edge density below which the A- and E-optimal designs look very different? What about larger block size?