Sudoku, gerechte designs, resolutions, affine space, spreads, reguli, and Hamming codes

R. A. Bailey, Peter J. Cameron
School of Mathematical Sciences, Queen Mary, University of London,
London E1 4NS, UK

Robert Connelly*
Department of Mathematics, Malott Hall, Cornell University,
Ithaca, NY 14853, USA

Abstract

The subjects of the title are all connected. The purpose of this paper is to explain the relation between *Sudoku solutions* (filled Sudoku grids) and the more general (and earlier) concept of *gerechte designs*, and to give some examples of these. Constructing gerechte designs can be viewed as finding resolutions of a block design constructed from the relevant grid. We give an upper bound for the number of mutually orthogonal Sudoku solutions, and a smaller upper bound if certain additional constraints are prescribed, and show by a geometric construction using spreads (explicitly realised by Hamming codes) that the bounds are attained. We generalise the bounds and the constructions to other situations. We explain the statistical background and construct a few Sudoku solutions with remarkable additional properties. For one of these types, which we call "symmetric Sudoku solutions", we give a construction, and an elementary proof that there are just two inequivalent solutions; the proofs are based on projective and affine geometry over GF(3) and the properties of the Hamming code of length 4 over GF(3). We also explain the statistical considerations behind gerechte designs, and construct some Sudoku solutions which have good statistical properties.

*This research partially supported by NSF Grant Number DMS-0510625.

1 Gerechte designs

1.1 Introduction

In 1956, W. U. Behrens [4] introduced a specialisation of Latin squares which he called “gerechte”. The $n \times n$ grid is partitioned into n regions $S_1, \ldots, S_n$, each containing n cells of the grid; we are required to place the symbols $1, \ldots, n$ into the cells of the grid in such a way that each symbol occurs once in each row, once in each column, and once in each region.

The row and column constraints say that the solution is a Latin square, and the last constraint restricts the possible Latin squares.

By this point, many readers will recognise that solutions to Sudoku puzzles are examples of gerechte designs, where $n = 9$ and the regions are the 3×3 subsquares. (The Sudoku puzzle was invented, with the name “number place”, by Harold Garns in 1979.)

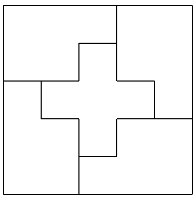
Also, let L be any Latin square of order n. Let the region S_i be the set of cells containing the symbol i in the square L. A gerechte design for this partition is precisely a Latin square orthogonal to L. (This shows that there is not always a gerechte design for a given partition. A simpler negative example is obtained by taking one region to consist of the first $n - 1$ cells of the first row and the nth cell of the second row.) We might ask: given a grid, what is the complexity of deciding whether a gerechte design exists?

For another example, consider the grid shown in Figure 1: this example was considered by Behrens in 1956. (Ignore the triples to the right of the grid for a moment.) Six solutions are shown. Up to rotations of the grid and permutations of the symbols $1, \ldots, 5$, these are all the solutions, as we will explain shortly. (The complete set of fifteen solutions is given in [3].)

1.2 Resolvable block designs

The basic data for a gerechte design is an $n \times n$ grid partitioned into n regions $S_1, \ldots, S_n$, each containing n cells. We can represent this structure by a block design as follows:

- the points (or treatments) are $3n$ objects $r_1, \ldots, r_n$, $c_1, \ldots, c_n$, $s_1, \ldots, s_n$;
- for each cell of the grid, there is a block $\{r_i, c_j, s_k\}$, if the cell lies in the ith row, the jth column, and the kth region.



$\{r_1, c_1, s_1\}$	$\{r_1, c_2, s_1\}$	$\{r_1, c_3, s_1\}$	$\{r_1, c_4, s_2\}$	$\{r_1, c_5, s_2\}$
$\{r_2, c_1, s_1\}$	$\{r_2, c_2, s_1\}$	$\{r_2, c_3, s_5\}$	$\{r_2, c_4, s_2\}$	$\{r_2, c_5, s_2\}$
$\{r_3, c_1, s_4\}$	$\{r_3, c_2, s_5\}$	$\{r_3, c_3, s_5\}$	$\{r_3, c_4, s_5\}$	$\{r_3, c_5, s_2\}$
$\{r_4, c_1, s_4\}$	$\{r_4, c_2, s_4\}$	$\{r_4, c_3, s_5\}$	$\{r_4, c_4, s_3\}$	$\{r_4, c_5, s_3\}$
$\{r_5, c_1, s_4\}$	$\{r_5, c_2, s_4\}$	$\{r_5, c_3, s_3\}$	$\{r_5, c_4, s_3\}$	$\{r_5, c_5, s_3\}$

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

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

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

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

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

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

Figure 1: Gerechte designs for a 5×5 grid

Proposition 1.1 *Gerechte designs on a given grid correspond, up to permuting the symbols* $1, \ldots, n$, *in one-to-one fashion with resolutions of the above design.*

Proof Given a gerechte design, let C_i be the set of cells containing the symbol i. By definition, the corresponding blocks contain each row, column, or region symbol exactly once, and so form a partition of the point set. Any cell contains a unique symbol, so every block occurs in just one class. Thus we have a resolution. The converse is proved in the same way.

The GAP [9] share package DESIGN [18] can find all resolutions of a block design, up to isomorphisms of the block design. In our case, isomorphisms of the block design come from symmetries of the partitioned grid, so we can use this package to compute all gerechte designs up to permutation of symbols and symmetries of the grid.

For example, the partition of the 5×5 grid discussed in the preceding section is represented as a block design with 15 points and 25 blocks of size 3, also shown in Figure 1. The automorphism group of the design is the cyclic group of order 4 consisting of the rotations of the grid through multiples of $\pi/2$. The DESIGN program quickly finds that, up to automorphisms, there are just six resolutions of this design, corresponding to six inequivalent gerechte designs; these are shown in the figure.

The same method shows that, for a 6×6 square divided into 3×2 rectangles, there are 49 solutions up to symmetries of the corresponding block design and permutations of the symbols. (The number of symmetries of the block design in this case is 3456; the group consists of all row and column permutations preserving the appropriate partitions.)

1.3 Orthogonal and multiple gerechte designs

Recall that two Latin squares L_1 and L_2 of order n are *orthogonal* if, for each ordered pair (i, j) of symbols, there is a unique cell (k, l) for which $(L_1)_{kl} = i$ and $(L_2)_{kl} = j$. A set of Latin squares is *mutually orthogonal* if each pair is orthogonal.

It is well known that the size of a set of mutually orthogonal Latin squares of order n is at most $n-1$, with equality if and only if there exists a projective plane of order n.

Similar definitions and results apply to gerechte designs. We say that two gerechte designs with the same grid are *orthogonal* if they are orthogonal as Latin squares, and a set of gerechte designs is *mutually orthogonal* if each pair is orthogonal.

Proposition 1.2 *Given a partition of the $n \times n$ grid into regions $S_1, \ldots, S_n$ each of size n, the size of a set of mutually orthogonal gerechte designs for this partition is at most $n-d$, where d is the maximum size of the intersection of a region S_i and a line (row or column) $L_j \neq S_i$.*

Proof Take a cell $c \in L_j \setminus S_i$. By permuting the letters in each square, we may assume that all the squares have entry 1 in the cell c. Now in each square, the symbol 1 occurs exactly once in the region S_i and not in the line L_j; and all these occurrences must be in different cells, since for each pair of squares, the pair $(1, 1)$ of entries already occurs in cell c. So there are at most $|S_i \setminus L_j|$ squares in the set.

This bound is not always attained. Consider the 5×5 gerechte designs given earlier. The maximum intersection size of a line and a region is clearly 3, so the bound for the number of mutually orthogonal designs is 2. But by inspection, each design has the property that the entries in cells $(2, 3)$ and $(3, 5)$ are equal. (The reader is invited to discover the simple argument to show that this must be so, independent of the classification of the designs.) Hence no two orthogonal designs are possible. Similarly, for the 6×6 square divided into 3×2 rectangles, there cannot exist two orthogonal gerechte designs, since it is well-known that there cannot exist two orthogonal Latin squares of order 6.

The Proposition gives an upper bound of 6 for the number of mutually orthogonal Sudoku solutions. In Section 3.1, we will see that this bound is attained.

The concept of a gerechte design can be generalised. Suppose that we are given a set of r partitions of the cells of an $n \times n$ grid into n regions each of size n. A *multiple gerechte design* for this partition is a Latin square which is simultaneously a gerechte design for all of the partitions.

For example, given a set of (mutually orthogonal) Latin squares, the letters in each square define a partition of the $n \times n$ array into regions. A Latin square is a multiple gerechte design for all of these partitions if and only if it is orthogonal to all the given Latin squares.

The problem of finding a multiple gerechte design can be cast into the form of finding a resolution of a block design, in the same way as for a single gerechte design. The block design has $(r + 2)n$ points, and each cell of the grid is represented by a block indexing its row, its column, and the region of each family which contain it. Again, we can use the DESIGN program to classify such designs up to symmetries of the grid.

For example, Federer [8], in a section which he attributed to G. M. Cox, called a $m_1m_2 \times m_1m_2$ Latin square *magic* if it is a gerechte design for the regions forming the obvious partition into $m_1 \times m_2$ rectangles, and *super magic* if it is simultaneously a gerechte design for the partition into $m_2 \times m_1$ rectangles, where $m_1 \neq m_2$. He considered the problem of finding multiple gerechte designs (which he called "super magic Latin squares") for the 6×6 square partitioned into 3×2 rectangles and 2×3 rectangles. The DESIGN package finds that there are 26 such designs up to symmetries.

We can also define a set of mutually orthogonal multiple gerechte designs in the obvious way, and prove a similar bound for the size of such a set.

We will see examples of these things in Section 3.1

2 Statistical considerations

2.1 Agricultural experiments in Latin squares

If a Latin square experiment is to be conducted on land that has recently been used for another Latin square experiment, it is sensible to regard the previous treatments as relevant and so to use a Latin square orthogonal to the previous one. As explained above, this is technically a sort of gerechte design, but no agricultural statistician would call it that.

The purpose of a gerechte design in agricultural experimentation is to ensure that all treatments are fairly exposed to any different conditions in the field. Rows and columns are good for capturing differences such as distance from a wood but not for marking out stony patches or other features that tend to clump in compact areas. Thus, in the statistical and agronomic literature, the regions of a gerechte design are always taken to be "spatially compact" areas.

2.2 Randomization

Before a design is used for an experiment, it is *randomized*. This means that a permutation of the cells is chosen at random from among all those that preserve the three partitions: into rows, into columns, and into regions. It is by no means common for the cells to be actually square plots on the ground; when they are, it is also possible to transpose rows and columns, if the regions are unchanged by this action. This random permutation is applied to the chosen gerechte design before it is laid out in the field.

One important statistical principle is *lack of bias*. This means that every plot in the field should be equally likely to be matched, by the randomization, to each abstract cell in the gerechte design, so that any individual plot with strange characteristics is equally likely to affect any of the treatments. To achieve this lack of bias, the set of permutations used for randomizing must form a *transitive* group, in the sense that there is such a permutation carrying any nominated cell to any other. The allowable permutations of the 5×5 grid in Figure 1 do not have this property, but those for magic Latin squares do. (Are there any others?)

For the remainder of this section we assume that $n = m_1 m_2$ and the regions are $m_1 \times m_2$ rectangles. Then the rows, columns and regions define some other areas: a *large row* is the smallest area that is simultaneously a union of regions and a union of rows; a *minirow* is the non-empty intersection of a row and region; *large columns* and *minicolumns* are defined similarly.

A pair of distinct cells in such a grid is in one of eight relationships, illustrated in Figure 2 for the 6×6 grid with 3×2 regions. For $i = 1$, …, 8, the cell labelled $*$ is in relationship i with the cell labelled i. Thus a pair of distinct cells is in relationship 1 if they are in the same minirow; relationship 2 if they are in the same minicolumn; relationship 3 if they are in the same region but in different rows and columns; relationship 4 if they are in the same row but in different regions; relationship 5 if they are in the same column but in different regions; relationship 6 if they are in the same large row but in different rows and regions; relationship 7 if they are in the same large column but in different columns and regions; relationship 8 if they are in different large rows and large columns.

$*$	1		4		
2				6	
	3				
5					8
	7				

Figure 2: Eight relationships between pairs of distinct cells in the 6×6 grid

The group of permutations used for randomization has the property that a pair of distinct cells can be mapped to another pair by one of the permutations if and only if they are in the same relationship. If, in addition, we can transpose the rows and columns (not possible in Figure 2) then relationships 1 and 2 are merged, as are 4 and 5, and 6 and 7.

The simple-minded analysis of data from an experiment in a gerechte design just fits rows, columns, regions and treatments (symbols). This is explained in [2]. However, one school of statistical thought holds that if the innate differences between rows, between columns and between regions are relevant, then so potentially are those between minirows, minicolumns, large rows and large columns. Yates took this view in his 1939 paper [22], whose discussion of a 4×4 Latin square "with balanced corners" may be the first published reference to gerechte designs. Thus the eight relationships all have to be considered when the gerechte design is chosen.

2.3 Orthogonality and the design key

Two further important statistical properties often conflict with each other. One is ease of analysis, which means not ease of performing arithmetic but ease of explaining the results to a non-statistician. So-called *orthogonal* designs, like the one in Figure 3, have this property.

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

Figure 3: An orthogonal design for the 6×6 grid with 3×2 regions

In Figure 3, differences between the two sets of three treatments $\{1, 2, 3\}$ and $\{4, 5, 6\}$ are assessed against the variability of minicolumns within regions and columns. Similarly, differences between the three sets of two treatments $\{1, 4\}$, $\{2, 5\}$ and $\{3, 6\}$ are assessed against the variability of minirows within regions and rows. Treatment comparisons orthogonal to all of those, such as $\{1, 5\}$ versus $\{2, 4\}$, are assessed against the residual variability of the plots after allowing for the variability of all the partitions.

An orthogonal design for an $m_1m_2 \times m_1m_2$ square with $m_1 \times m_2$ regions may be constructed using the *design key* method [19, 20], as recommended in [3]. The large rows are labelled by A_1, which takes values 1, ..., m_2. Within each large row, the rows are labelled by A_2, which takes values 1, ..., m_1. Similarly, the large columns are labelled by B_1, taking values 1, ..., m_1, and the columns within each large column by B_2, taking values 1, ..., m_2. Then put $N_1 = A_1 + B_2$ modulo m_2 and $N_2 = A_2 + B_1$ modulo m_1. The ordered pairs of values of N_1 and N_2 give the m_1m_2 symbols. In Figure 3, the rows are numbered from top to bottom, the columns from left to right, and the correspondence between the ordered pairs and the symbols is as follows.

N_1 \ N_2	1	2	3
1	1	2	3
2	4	5	6

(When explaining this construction to non-mathematicians we usually take the integers modulo m to be 1, ..., m rather than 0, ..., $m-1$.)

Variations on this construction are possible, especially when m_1 and m_2 are both powers of the same prime p. For example, if $m_1 = 4$ and $m_2 = 2$ then we can work modulo 2, using A_1 to label the large rows, A_2 and A_3 to label the rows within large rows, B_1 and B_2 to label the large columns, and B_3 to label the columns within large columns. Numbers can be allocated by putting $N_1 = A_1 + B_3$, $N_2 = A_2 + B_1$ and $N_3 = A_3 + B_2$. All that is required is that no non-zero linear combination (modulo 2) of N_1, N_2 and N_3 contains only A_1, B_1 and B_2, or a subset thereof.

2.4 Efficiency and concurrence

The other important statistical property is *efficiency*, which means that differences between treatments should be estimated with small variance. At one extreme, we might decide that the innate differences between minicolumns are so great that the design in Figure 3 provides no information at all about the difference between treatments 1, 2, 3 and treatments 4, 5, 6; and similarly for minirows. In this case, it can be shown (see [1, Chapter 7]) that the relevant variances can be deduced from the matrix

$$M = m_1 m_2 I - \frac{1}{m_2}\Lambda_R - \frac{1}{m_1}\Lambda_C + J.$$

Here I is the $n \times n$ identity matrix and J is the $n \times n$ all-1 matrix. The *concurrence* of symbols i and j in minirows is the number of minirows containing both i and j (which is n when $i = j$): the matrix Λ_R contains these concurrences. The matrix Λ_C is defined similarly, using concurrences in minicolumns. It is known that if the off-diagonal entries in the matrix M are all equal then the average variance is as small as possible for the given values of m_1 and m_2, so the usual heuristic is to choose a design in which the off-diagonal entries differ as little as possible.

A compromise between these two statistical properties is *general balance* [12, 14, 15], which requires that the concurrence matrices Λ_R and Λ_C commute with each other. A special case of general balance is *adjusted orthogonality* [7, 13], for which $\Lambda_R\Lambda_C = n^2 J$. The design is orthogonal if it has adjusted orthogonality and $\Lambda_R^2 = nm_2\Lambda_R$ and $\Lambda_C^2 = nm_1\Lambda_C$.

These properties are explored further in Section 3.3.

3 Some special Sudoku solutions

3.1 Sudoku and geometry over GF(3)

As we noted earlier, a completed Sudoku puzzle (for short, a Sudoku solution) is a gerechte design for the 9×9 array partitioned into nine 3×3 subsquares. Here we describe a very special class of Sudoku solutions.

First we define some further partitions of the 9×9 square. We define a *broken row* to be the union of three minirows occurring in the same position in three subsquares in a column, and a *broken column* to be the union of three minicolumns occurring in the same position in three subsquares in a row. A *location* is a set of nine cells occurring in a fixed position in all of the subsquares (for example, the centre cells of each subsquare).

Now we coordinatise the cells over GF(3), the field of integers mod 3, as described in Section 2.3. Each cell c has four coordinates (x_1, x_2, x_3, x_4), where

- x_1 is the row number of the subsquare containing c;
- x_2 is the number of the minirow of this subsquare which contains c;
- x_3 is the column number of the subsquare containing c;
- x_4 is the number of the minicolumn of this subsquare which contains c.

(In each case we start the numbering at zero, so that each coordinate takes the values $0, 1, 2$. We number rows from top to bottom and columns from left to right.)

Now the cells are identified with the points of the four-dimensional affine space $\mathrm{AG}(4, 3)$ over GF(3). Various interesting regions are cosets of 2-dimensional subspaces, as shown in the following table.

Equation	Description of cosets
$x_1 = x_2 = 0$	Rows
$x_3 = x_4 = 0$	Columns
$x_1 = x_3 = 0$	Subsquares
$x_1 = x_4 = 0$	Broken columns
$x_2 = x_3 = 0$	Broken rows
$x_2 = x_4 = 0$	Locations

In addition, the main diagonal is the subspace defined by the equations $x_1 = x_3$ and $x_2 = x_4$, and the antidiagonal is $x_1 + x_3 = x_2 + x_4 = 2$, a coset of the subspace $x_1 = -x_3$, $x_2 = -x_4$. (The other cosets of these two subspaces are not so obvious in the grid.)

Now in a Sudoku solution, each symbol occurs in nine positions forming a transversal to the cosets of the subspaces defining rows, columns, and subsquares as above. We call a Sudoku solution *linear* if these nine positions form an affine subspace (a coset of a vector subspace) in the affine space. All the Sudoku solutions constructed in this section are linear.

Theorem 3.1 *(a) There is a set of six mutually orthogonal Sudoku solutions. These squares are also gerechte designs for the partition into locations, and have the property that each symbol occurs once on the main diagonal and once on the antidiagonal.*

(b) There is a set of four mutually orthogonal multiple gerechte designs for the partitions into subsquares, locations, broken rows and broken columns; they also have the property that each symbol occurs once on the main diagonal and once on the antidiagonal.

Remarks 1. We saw already that the number 6 in part (a) is optimal. The number 4 in (b) is also optimal. For given such a set, we can as before suppose that they all have the symbol 1 in the cell in the top right corner. Now the 1s in the subsquare in the middle of the top row cannot be in its top minirow or its left-hand minicolumn, so just four positions are available; and the squares must have their ones in different positions.

2. We will use the phrase *symmetric Sudoku solution* to denote a placement of symbols in the grid satisfying the conditions that each symbol occurs once in each row, column, subsquare, broken row, broken column, and location.

Proof (a) Our six Sudoku solutions will be given by six parallel classes of planes in the affine space. Given two parallel classes of planes, each plane of the first meets each plane of the second in a single point if and only if the two vector subspaces meet just in the origin; this means that, in the projective space $\mathrm{PG}(3,3)$, these subspaces correspond to disjoint (or skew) lines.

We will translate the problem into the projective space. We refer to [6, 11] for more information about 3-dimensional projective geometry. We will use the following notions:

- A hyperbolic quadric contains two families of "rulings", each family consisting of a set of pairwise disjoint lines covering all the points (Figure 4). Such a ruling is called a *regulus*, and the other ruling is the *opposite regulus*. Any three pairwise skew lines lie in a unique regulus.

- A *spread* is a family of pairwise disjoint lines covering all the points of the projective space. A spread is *regular* if it contains the regulus through any three of its lines. If the field F is not quadratically closed, then there exists a regular spread. In particular, this holds when $F = \mathrm{GF}(3)$.

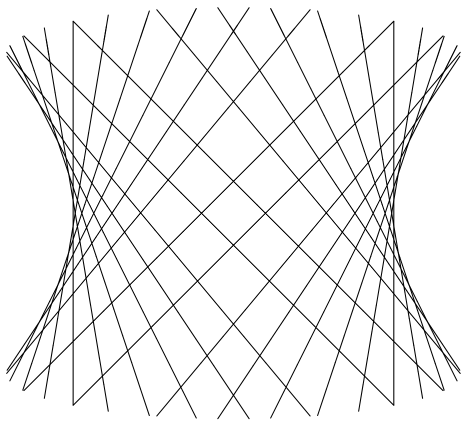


Figure 4: A ruled quadric and two reguli

In our situation, the planes $x_1 = x_2 = 0$ and $x_3 = x_4 = 0$ correspond to two skew lines L_1 and L_2 of $\mathrm{PG}(3,3)$, and $x_1 = x_3 = 0$ to a common transversal L_3. So we have to find six pairwise skew lines which are skew to the given three lines.

Since the automorphism group of PG(3, 3) is transitive on triples (L_1, L_2, L_3) of lines such that L_1 and L_2 are skew and L_3 intersects both, we can assume that L_1 and L_2 belong to a fixed regulus, and that L_3 belongs to the opposite regulus; moreover, the first regulus is contained in a regular spread. Then the six lines of the spread not in the first regulus are disjoint from L_3, and have the required property. (See Figure 5.)

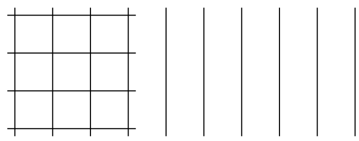


Figure 5: A regulus, the opposite regulus, and a spread

Now a simple calculation shows that the remaining lines of $\mathcal{R}$ are $x_1 - x_3 = x_2 - x_4 = 0$ and $x_1 + x_3 = x_2 + x_4 = 0$, and the other three lines of the opposite regulus are $x_2 = x_4 = 0$, $x_1 - x_2 = x_3 - x_4 = 0$, and $x_1 + x_2 = x_3 + x_4 = 0$. Since the remaining six lines of the spread are skew to these, our claim about locations and diagonals follows.

A different solution is obtained by choosing a regulus $\mathcal{R}^*$ disjoint from $\mathcal{R}$ and contained in the spread, and replacing it by the opposite regulus.

(b) For the second part, it is more convenient to work in the affine space. We want four 2-dimensional subspaces disjoint from all the subspaces $x_i = x_j = 0$ for $i \neq j$. Such a subspace has the property that each of its non-zero vectors has at most one zero coordinate. In coding theory terms, its minimum weight is at least 3. Now there is (up to equivalence) a unique linear code of length 4, dimension 2 and minimum weight 3 over GF(3), namely the *Hamming code*, which can be taken to be spanned by the rows of the matrix $\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 2 \end{bmatrix}$. (We will describe Hamming codes and their connection with Sudoku solutions more carefully in the next section.)

Explicitly, the nine words of the Hamming code coordinatise the nine cells of the grid in which the first symbol is placed. Each coset similarly coordinatises the positions of a further symbol.

The resulting square is shown in Figure 6. It is easily checked that, as expected, each symbol occurs once in each row, column, subsquare, broken row, broken column, and location, as well as on each diagonal. The square shown has the further property that each of the 3×3 subsquares is "semi-magic", that is, its row and column sums (but not necessarily its diagonal sums) are 15 (John Bray [5]).

8	1	6	2	4	9	5	7	3
3	5	7	6	8	1	9	2	4
4	9	2	7	3	5	1	6	8
7	3	5	1	6	8	4	9	2
2	4	9	5	7	3	8	1	6
6	8	1	9	2	4	3	5	7
9	2	4	3	5	7	6	8	1
1	6	8	4	9	2	7	3	5
5	7	3	8	1	6	2	4	9

Figure 6: A symmetric Sudoku solution

In order to find four pairwise disjoint subspaces, we must apply equivalences (column permutations and sign changes) to this code. The following four matrices span the required codes:

$$\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 2 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 2 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 2 \end{bmatrix}$$

Another solution is obtained by changing the sign of the last coordinate.

We can use the solution to (b) to find an explicit construction for (a). Recall that we seek six lines of the projective space disjoint from the lines $x_1 = x_2 = 0$, $x_3 = x_4 = 0$, and $x_1 = x_3 = 0$. Four of these are also disjoint from $x_2 = x_4 = 0$, $x_1 = x_4 = 0$ and $x_2 = x_3 = 0$; these are the four Hamming codes we constructed. Now there is a unique regulus $\mathcal{R}'$ containing the first two of these six lines and having the last two in the opposite regulus; the other two lines of $\mathcal{R}'$ can be added to the four disjoint lines to produce the required set of six lines. They have equations $x_1 + x_4 = x_2 + x_3 = 0$ and $x_1 - x_4 = x_2 - x_3 = 0$. See Figure 7. The resulting six mutually orthogonal Sudoku solutions are shown in Figure 8; the last four are symmetric.

Note that the four lines disjoint from the two reguli themselves form a regulus; they and the lines of the opposite reguli are the eight Hamming codes.

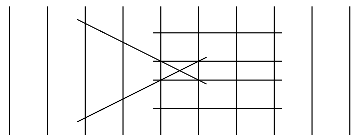


Figure 7: Two reguli

111 111	222 222	333 333	749 658	857 469	968 547	475 896	586 974	694 785
444 444	555 555	666 666	173 982	281 793	392 871	718 239	829 317	937 128
777 777	888 888	999 999	416 325	524 136	635 214	142 563	253 641	361 452
326 589	134 697	215 478	952 734	763 815	841 926	687 342	498 153	579 261
659 823	467 931	548 712	385 167	196 248	274 359	921 675	732 486	813 594
983 256	791 364	872 145	628 491	439 572	517 683	354 918	165 729	246 837
238 965	319 746	127 854	864 273	945 381	756 192	593 427	671 538	482 619
562 398	643 179	451 287	297 516	378 624	189 435	836 751	914 862	725 943
895 632	976 413	784 521	531 849	612 957	423 768	269 184	347 295	158 376

Figure 8: The six Sudoku solutions

3.2 Linear Sudoku solutions

A *code* of length n over an alphabet S is a set C of n-tuples (called *codewords*) with entries from S. The *Hamming distance* between two codewords is the number of positions where they differ. The *minimum distance* of the code is the smallest Hamming distance between distinct codewords.

If a code of length n over an alphabet of size q has minimum distance $2d+1$ or greater, then the balls of radius d centred at the codewords are pairwise distinct, from which the *sphere-packing bound*

$$|C| \leq q^n \Big/ \left(\sum_{i=0}^{d} \binom{n}{i} (q-1)^i \right)$$

holds. A code meeting this bound is called *perfect*.

Now consider the set of positions where a given symbol occurs in a symmetric Sudoku solution, regarded as a subset of $\mathrm{GF}(3)^4$: these positions form a code of length 4 containing nine codewords. The minimum distance of this code is at least 3. (For if two words agree in two positions, say the ith and jth, then they lie in the same coset of the subspace given by $x_i = x_j = 0$, contradicting the definition of a symmetric Sudoku solution. Since

$$3^4 \Big/ \left(\sum_{i=0}^{1} \binom{4}{i} 2^i \right) = 9,$$

the code C is perfect. So we have:

Proposition 3.2 *A symmetric Sudoku solution is equivalent to a partition of* $\mathrm{GF}(3)^4$ *into nine perfect codes.*

It is known that a perfect code with these parameters containing the zero word must be a linear subspace of $\mathrm{GF}(3)^4$, and is unique up to coordinate permutations and sign changes. The unique such code is called a *Hamming code* (see Hill [10]). The linearity of such codes containing zero shows that any perfect code must be an affine subspace. We give below what amounts to an elementary proof of this, together with an analysis of the possible partitions of $\mathrm{GF}(3)^4$ into perfect codes. The result is that there are just two different solutions up to equivalence.

Lemma 3.3 *If* $p_1 = (a_1, a_2, a_3, a_4)$ *and* $p_2 = (b_1, b_2, b_3, b_4)$ *are two distinct vectors in* $V = \mathrm{GF}(3)^4$ *that carry the same symbol in a symmetric Sudoku solution, then* $a_i = b_i$ *for exactly one* $i = 1, 2, 3, 4$. *That is, any two such words have Hamming distance* 3.

Proof Due to the symmetry of the constraints, for convenience, we may assume, without loss of generality, that $p_1 = (0, 0, 0, 0)$. Suppose the conclusion is false. Then each coordinate of p_2 is non-zero, and since there are only 3 elements in GF(3), some pair must be the same. By multiplying by $2 = -1$ and permuting the coordinates, if necessary, we may assume $p_2 = (1, 1, a, b)$, where $a \neq 0$ and $b \neq 0$. In the coset defined by $x_1 = 0, x_3 = a$, the symbol cannot correspond to the any of the points $x_1 = 0; x_2 \neq 0, 1; x_3 = a; x_4 \neq 0, b$. The only point left in that coset is $(0, 2, a, -b)$. In the coset defined by $x_1 = 1, x_3 = 0$, the symbol cannot correspond to the any of the points $x_1 = 1; x_2 \neq 0, 1; x_3 = 0; x_4 \neq 0, b$. The only point left in that coset is $(1, 2, 0, -b)$. But then $(0, 2, a, -b)$ and $(1, 2, 0, -b)$ must have the same symbol, but they lie in the same coset of the Location space $x_2 = x_4 = 0$, a contradiction.

Notice that the above argument is typical for Sudoku puzzles, where some of the positions for a symbol are deduced by eliminating all but one in a certain set. The following is also a similar Sudoku argument.

Lemma 3.4 *If $p_1 = (a_1, a_2, a_3, a_4)$ and $p_2 = (b_1, b_2, b_3, b_4)$ are two distinct vectors in V that carry the same symbol in a symmetric Sudoku solution, then the midpoint $(p_1 + p_2)/2 = -(p_1 + p_2)$ also carries the same symbol.*

Proof Given Lemma 3.3 we may assume, without loss of generality, that $p_1 = (0, 0, 0, 0)$, and $p_2 = (0, 1, a, b)$, where $a \neq 0$, and $b \neq 0$. In the coset defined by $x_1 = 0, x_2 = 2$, none of the points $x_1 = 0; x_2 = 2; x_3 = 0, a; x_4 = 0, b$ can carry the symbol, leaving only the midpoint $(0, 2, -a, -b)$.

Corollary 3.5 *The positions of any symbol in a symmetric Sudoku solution form a 2-dimensional affine subspace of V.*

Proof Let X be the set of points corresponding to one Sudoku symbol. By Lemma 3.4, for any two points p_1, p_2 in X the third point on the line through p_1, p_2 is also in X. So X is the affine span of its points. So it must be an affine linear subspace of V. Since X has 9 points it must be 2-dimensional.

It is easy to determine a basis for a linear 2-dimensional subspace (that intersects the origin) that corresponds to a Sudoku solution. Call such a subspace *allowable.*

Lemma 3.6 *The vectors $p_1 = (a_1, a_2, a_3, a_4)$ and $p_2 = (b_1, b_2, b_3, b_4)$ are two independent vectors in an allowable 2-dimensional subspace X of V if and only if the four ratios a_i/b_i, for $i = 1, 2, 3, 4$ are distinct, where $\pm 1/0 = \infty$ is one ratio that must appear, and the indeterminant form $0/0$ does not appear.*

Proof The vectors $p_1 = (x_1, x_2, x_3, x_4)$, $p_2 = (y_1, y_2, y_3, y_4)$ and any two of the standard basis vectors, with just one non-zero coordinate 1, must be independent. So the determinant of the corresponding matrix $x_i y_j - x_j y_i \neq 0$. Then the result follows.

Given Lemma 3.6, we see that when a basis for an allowable linear subspace is put into row-reduced echelon form, it takes one the following eight possibilities.

$$\left\{ \left\{ \begin{matrix} 1011 \\ \text{or} \\ 1022 \end{matrix} \right\} \text{and} \left\{ \begin{matrix} 0112 \\ \text{or} \\ 0121 \end{matrix} \right\} \right\} \text{or} \left\{ \left\{ \begin{matrix} 1012 \\ \text{or} \\ 1021 \end{matrix} \right\} \text{and} \left\{ \begin{matrix} 0111 \\ \text{or} \\ 0122 \end{matrix} \right\} \right\} \qquad (1)$$

These are the only allowable linear subspaces. So any set of positions of a Sudoku solution is a translated copy of one of those eight linear subspaces.

Next we come to the question of how such subsets can partition V. One simple way is just to take one of the above 2-dimensional linear subspaces and translate it to 9 distinct parallel positions. Another choice is the following. Extend an allowable linear subspace X to an appropriate 3-dimensional subspace Y of V. Three copies of Y partition V, and we can look for another allowable subspace X' of Y which can be used to partition the other translated copies of Y. For this to work, it is necessary that the linear span of X and X' be 3-dimensional. For each choice of an allowable X, it is easy to check that there are four other allowable X' such that the span of X and X' is 3-dimensional, but there is no set of three allowable linear subspaces such that the span of each pair is 3-dimensional.

Conversely, suppose that there are symmetric Sudoku solutions. We consider the corresponding partitions of V into translates of linear allowable 2-dimensional subspaces. If any pair of such linear subspaces are distinct and such that their span is all of V, then any of their translates will intersect, contradicting the Sudoku property. Thus their span must be a 3-dimensional linear subspace Y and hence correspond to two subspaces X and X' as in the previous paragraph. Thus X and X' are the only two allowable linear subspaces parallel to any set in the partition for a Sudoku solution. Furthermore, in each of the three translates of Y, only translates of one of X or X' can appear. Thus the Sudoku solutions described in the previous paragraph are the only ones possible.

Using this analysis we can see that for each choice of one of the 8 allowable planes, since there are exactly 4 choices for another such that their span is 3-dimensional, there are $8 \cdot 4/2 = 16$ possible choices of such pairs. For each pair there are 6 choices of which of the two planes to partition each of the three 3-dimensional spaces determined by the pair of planes such that a translate of each plane is chosen at least once. Thus there are $6 \cdot 16 = 96$ possible Sudoku solutions of this sort. In addition, there are 8 possible Sudoku solutions that correspond to translates of a single plane. This gives $96 + 8 = 104$ total number of symmetric Sudoku solutions, falling into just two classes up to equivalence under symmetries of the grid.

This analysis can also be used to define and count orthogonal symmetric Sudoku solutions, where each comes from translates of a single plane. Two such Sudoku solutions are orthogonal if the vectors that form a basis of the two planes, together form a basis for V. But it is easy to see from (1) that there are two sets of 4 planes such that each pair has a 4-dimensional span, i.e. they are independent.

In the spirit of the Sudoku puzzle, we give in Figure 9 a partial symmetric Sudoku which can be uniquely completed (in such a way that each row, column, subsquare, broken row, broken column or location contains each symbol exactly once). The solution is not equivalent to the one shown in Figure 6.

The fact that there are just two non-isomorphic symmetric Sudoku solutions, proved in the above analysis, can be confirmed with the DESIGN program, which also shows that, if we omit the condition on locations, there are 12 different solutions; and, if we omit both locations and broken columns, there are 31021 different solutions. The total number of Sudoku solutions up to equivalence (that is, solutions with only the conditions on rows, columns, and subsquares) is 5472730538; this number was computed by Ed Russell and Frazer Jarvis [16].

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

Figure 9: A Sudoku-type puzzle

3.3 The design in minirows and minicolumns

The cells in the minirows and minicolumns form lines of the affine space AG$(4, 3)$. In any Sudoku solution given by the construction in Section 3.1, such a line together with the subspace S defining one of the Sudoku symbols span a 3-dimensional subspace which contains three cosets of S. So all the nine lines in this subspace contain the same three symbols. This means that the 27 minirows define just three triples from $\{1, \ldots, 9\}$, each triple occuring in nine minirows. The same condition holds for the minicolumns. Thus the design is orthogonal. Moreover, the block design on $\{1, \ldots, 9\}$ formed by the minirows and minicolumns is a 3×3 grid with each grid line occurring nine times as a block. Each pair of treatments lies in either 0 or 9 blocks of the design.

Is there a better balanced Sudoku solution? Since the average number of blocks containing a pair of symbols from $\{1, \ldots, 9\}$ in this design is $2 \cdot 27 \cdot 3/\binom{9}{2} = 9/2$, we could ask whether there is a Sudoku solution in which each pair occurs in either 4 or 5 blocks. The first example of such a design was constructed by Emil Vaughan [21].

Given such a design with two pairwise concurrences, we obtain a regular graph of valency 4 on $\{1, \ldots, 9\}$ by joining two vertices if they occur in five blocks of the design. The "nicest" such graph is the 3×3 grid, the line graph of $K_{3,3}$. (This graph is strongly regular, and the resulting design would be partially balanced with respect to the Hamming association scheme consisting of the graph and its complement: see [1].) Vaughan's solution does not realise this graph, but we subsequently found one which does. An example is given in Figure 10.

1	5	2	6	8	9	7	4	3
3	8	7	1	2	4	9	6	5
9	4	6	3	5	7	1	8	2
2	1	4	8	7	6	3	5	9
6	9	5	4	1	3	2	7	8
8	7	3	5	9	2	6	1	4
5	6	1	9	3	8	4	2	7
7	3	8	2	4	1	5	9	6
4	2	9	7	6	5	8	3	1

Figure 10: Balancing minirows and minicolumns

We could ask whether even more is true: is there a Sudoku solution in which each pair of symbols occur together 2 or 3 times in a minirow, 2 or 3 times in a minicolumn, and 4 or 5 times altogether? A computation using GAP showed that such a solution cannot exist; one cannot place more than five symbols satisfying these constraints without getting stuck.

We further found that there exist Sudoku solutions in which the design in minirows and minicolumns is partially balanced with respect to the 3×3 grid with concurrences $(4, 5)$, $(3, 6)$, $(2, 7)$ or $(0, 9)$, but not $(1, 8)$ (for which at most four symbols can be placed). Our "nice" example (equivalent to the third, fourth, fifth and sixth squares in Figure 8) realises the case $(0, 9)$.

We also considered another special type of Sudoku solution based on the properties of the minirows and minicolumns: those for which the designs formed by minirows and minicolumns have adjusted orthogonality, in the sense that their concurrence matrices Λ_R and Λ_C satisfy $\Lambda_R \Lambda_C = 81J$. The special Sudoku solution of Figure 6 has this property, but it is not unique. (In this solution, all entries of each concurrence matrix are 0 or 9.) We found that there are, up to symmetry, 194 Sudoku solutions for which the minirows and minicolumns have adjusted orthogonality in this sense, of which 104 have the property that both Λ_R and Λ_C have entries different from 0 and 9. One of these solutions is shown in Figure 11.

1	2	3	4	5	6	7	8	9
7	8	9	1	3	2	6	5	4
4	5	6	7	8	9	1	3	2
3	1	2	6	4	5	9	7	8
9	7	8	2	1	3	4	6	5
6	4	5	9	7	8	2	1	3
8	9	1	5	6	4	3	2	7
2	3	7	8	9	1	5	4	6
5	6	4	3	2	7	8	9	1

Figure 11: Minirows and minicolumns form designs with adjusted orthogonality, but the overall design is not orthogonal

A word about the computations reported in this section. The strategy is to place the symbols $1, \ldots, 9$ in the grid successively to satisfy the constraints. There are 6^6 possible "Sudoku permutations" which place a symbol once in each row, column and subsquare. We say that two of these permutations are "compatible" if they place symbols in disjoint cells satisfying the appropriate conditions (for example, for concurrences 4 and 5, that they place either 4 or 5 symbols in the same minirow or minicolumn). Then we join two Sudoku permutations if they are compatible, and search randomly for a clique of size 9 in the compatibility graph.

For adjusted orthogonality of the two designs, there is no obvious compatibility condition, and we proceeded differently. Since each of the two concurrence matrices has diagonal entries 9, we see that adjusted orthogonality implies that two symbols cannot occur both in the same minirow and in the same minicolumn. Using this as the compatibility condition, we found all cliques in the graph, using the GAP package GRAPE [17]. Remarkably, it turned out that all of them actually give designs with adjusted orthogonality; we know no simple reason for this.

4 Other finite field constructions

The construction in the previous section can be generalised.

Proposition 4.1 *Let q be a prime power, and a and b positive integers. Let $n = q^{a+b}$. Partition the $n \times n$ square into $q^a \times q^b$ rectangles. Then we can find*

$$q^{a+b} - 1 - \frac{(q^a - 1)(q^b - 1)}{q - 1}.$$

mutually orthogonal gerechte designs for this grid.

Remark If $a < b$, our upper bound for the number of mutually orthogonal gerechte designs for this grid is $q^b(q^a - 1)$. If $a = 1$, this bound is equal to the number in the theorem, so our bound is attained. If $a > 1$, however, the bound is not met by the construction. For example, if $p = 2$, $a = 2$ and $b = 3$, the bound is 24 but the construction achieves 10. If a and b are not coprime, we can improve the construction by replacing q, a, b by $q^d, a/d, b/d$, where $d = \gcd(a, b)$.

Proof Represent the cells by points of the affine space $\mathrm{AG}(2(a+b), q)$ with coordinates $x_1, \ldots, x_{a+b}, y_1, \ldots, y_{a+b}$. The rows are cosets of the subspace $x_1 = \cdots = x_{a+b} = 0$, the columns are cosets of the subspace $y_1 = \cdots = y_{a+b} = 0$, and the rectangles are cosets of $x_1 = \cdots = x_a = y_1 = \cdots = y_b = 0$.

As before, we work in the projective space $\mathrm{PG}(2(a+b) - 1, q)$. The first two subspaces are disjoint, and are part of a spread of $q^{a+b} - 1$ subspaces of the same dimension. The third subspace meets the first in $(q^b - 1)/(q - 1)$ points and the second in $(q^a - 1)/(q - 1)$ points, and has $(q^a - 1)(q^b - 1)/(q - 1)$ further points. In the worst case, this subspace meets $(q^a - 1)(q^b - 1)/(q - 1)$ further spaces of the spread, each in one point. This leaves $q^{a+b} - 1 - (q^a - 1)(q^b - 1)/(q - 1)$ spread spaces disjoint from it, as required.

Our construction of symmetric Sudoku solutions also generalises:

Proposition 4.2 *Let q be a prime power, and consider the $q^2 \times q^2$ grid, partitioned into $q \times q$ subsquares, broken rows, broken columns, and locations as in the preceding section. Then there exist $(q - 1)^2$ mutually orthogonal multiple gerechte design for these partitions; this is best possible.*

Proof We follow the same method as before, working over $\mathrm{GF}(q)$. The lines of $\mathrm{PG}(3, q)$ defining rows, columns, subsquares, broken rows, broken columns, and locations lie in the union of two reguli with two common lines, which form part of a regular spread. The remaining $(q - 1)^2$ lines of the spread give the required designs. The upper bound is proved as before.

References

[1] R. A. Bailey, *Association Schemes: Designed Experiments, Algebra and Combinatorics*, Cambridge Studies in Advanced Mathematics **84**, Cambridge University Press, Cambridge, 2004.

[2] R. A. Bailey, J. Kunert and R. J. Martin, Some comments on gerechte designs. I. Analysis for uncorrelated errors. *J. Agronomy & Crop Science* **165** (1990), 121–130.

[3] R. A. Bailey, J. Kunert and R. J. Martin, Some comments on gerechte designs. II. Randomization analysis, and other methods that allow for inter-plot dependence, *J. Agronomy & Crop Science* **166** (1991), 101–111.

[4] W. U. Behrens, Feldversuchsanordnungen mit verbessertem Ausgleich der Bodenunterschiede, *Zeitschrift für Landwirtschaftliches Versuchs- und Untersuchungswesen* **2** (1956), 176–193.

[5] J. N. Bray, personal communication, February 2006.

[6] P. J. Cameron, *Projective and Polar Spaces*, QMW Maths Notes **13**, Queen Mary and Westfield College, London, 1991; available from `http://www.maths.qmul.ac.uk/~pjc/pps/`

[7] J. A. Eccleston and K. G. Russell, Connectedness and orthogonality in multi-factor designs, *Biometrika* **62** (1975), 341–345.

[8] W. T. Federer, *Experimental Design—Theory and Applications*, Macmillan, New York, 1955.

[9] The GAP Group, GAP — Groups, Algorithms, and Programming, Version 4.6; Aachen, St Andrews, 2005, `http://www.gap-system.org/`

[10] R. Hill, *A First Course in Coding Theory*, Clarendon Press, Oxford, 1986.

[11] J. W. P. Hirschfeld, *Finite Projective Spaces of Three Dimensions*, Oxford University Press, Oxford, 1985.

[12] A. M. Houtman and T. P. Speed, Balance in designed experiments with orthogonal block structure, *Ann. Statist.* **11** (1983) 1069–1085.

[13] S. M. Lewis and A. M. Dean, On general balance in row–column designs, *Biometrika* **78** (1991), 595–600.

[14] J. A. Nelder, The analysis of randomized experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance, *Proc. Roy. Soc. London A* **283** (1965) 163–178.

[15] J. A. Nelder, The combination of information in generally balanced designs, *J. Roy. Statistic. Soc. B* **30** (1968), 303–311.

[16] Ed Russell and Frazer Jarvis, There are 5472730538 essentially different Sudoku grids,
`http://www.shef.ac.uk/~pm1afj/sudoku/sudgroup.html`

[17] L. H. Soicher, `GRAPE`: a system for computing with graphs and groups. In L. Finkelstein and W. M. Kantor, editors, *Groups and Computation*, volume 11 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 287–291. American Mathematical Society, 1993. `GRAPE` homepage:
`http://www.gap-system.org/Share/grape.html`

[18] Leonard H. Soicher, The DESIGN package for GAP,
`http://designtheory.org/software/gap_design/`

[19] H. D. Patterson, Generation of factorial designs, *J. Roy. Statist. Soc. B* **38** (1976), 175–179.

[20] H. D. Patterson and R. A. Bailey, Design keys for factorial experiments, *Applied Statistics* **27** (1978), 335–343.

[21] Emil Vaughan, personal communication, November 2005.

[22] F. Yates, The comparative advantages of systematic and randomized arrangements in the design of agricultural and biological experiments, *Biometrika* **30** (1939), 440–466.