MINI-SUDOKUS AND GROUPS

CARLOS ARCOS, GARY BROOKFIELD, AND MIKE KREBS

By now you probably have at least a passing acquaintance with Sudoku, the pencil-and-paper puzzle that has, for the past few years, been displacing advice columns and word jumbles from the back pages of newspapers all over the world.

The rules are simple. One is given a 9×9 grid. Each cell in the grid is to be filled in with one of the digits from 1 to 9. Some of the cells have been filled in already, as in the example below.

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

The puzzler may not fill in the empty cells willy-nilly; he or she must obey the Rule of One, which requires that each row, each column, and each block (the 3×3 subgrids with thick borders) must contain every digit from 1 to 9 exactly once. To simplify our discussion we say that a 9×9 grid completely filled with the digits 1 to 9 such that the Rule of One holds is a *Sudoku*. (So the grid above, then, is *not* a Sudoku according to our definition, because not all of the cells have been filled in. Once all the cells have been filled in, *then* it's a Sudoku.)

The rules of Sudoku suggest many natural mathematical questions: *How do you construct these puzzles? How do you solve these puzzles? How many different Sudokus are there? How many of these are essentially different?* We call two Sudokus *essentially the same*, or *equivalent*, if you can get from one to the other in finitely many steps where a single step might be switching the first two columns, or rotating the grid ninety degrees, or relabeling entries (replacing every 2 with a 7 and every 7 with a 2, for example). We will make this notion of equivalence more precise in the sections that follow.

The answers to the questions above are known. Felgenhauer and Jarvis [4] found that there are $6,670,903,752,021,072,936,960$ Sudokus. That's a big number. Also, Jarvis and Russell [8] found that there are $5,472,730,538$ essentially different Sudokus. That's a smaller number. But it's still pretty darn big. Both of these numbers were calculated using computers.

Date: November 3, 2009.

There is no need to limit oneself to 9×9 grids; any grid of size $n^2 \times n^2$ will do. Herzberg and Murty [7] use graph-theoretic techniques to provide an asymptotic estimate for the number of $n^2 \times n^2$ Sudokus.

In this article, we wish to be accessible to those without a background in graph theory, and we also wish to keep things on an order of magnitude that a human can more readily comprehend. We therefore answer these questions about the much simpler, but still interesting, case of 4×4 Sudokus. We call these *mini-Sudokus*. Thus a mini-Sudoku is a 4×4 grid, for example,

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

,

such that the Rule of One holds: Each row, each column, and each block (the 2×2 subgrids with thick borders) contains every digit from 1 to 4 exactly once.

Our main tools come from group theory. In particular, the notion of groups acting on sets will enable us to define precisely what it means for two mini-Sudokus to be essentially the same. Undergraduate math majors will have seen these concepts in a first abstract algebra class and may find that applying newly-learned group theory methods to a familiar, concrete example brings the abstract theory to life.

Various sources discuss the mathematics of Sudoku in general [2, 3, 6, 7].

Counting mini-Sudokus

How many mini-Sudokus are there? They can be enumerated in many ways. One method is to consider first the four entries in the upper left 2×2 block. These entries must be 1, 2, 3, and 4, but they can be put in any order. This gives $4! = 24$ ways of filling this block. The reader should confirm that, once this block has been filled, for example,

1	2		
3	4		
		*	
			*

then all the other entries are determined by the Rule of One and the choice of the two entries marked *. These two entries are arbitrary except that they must be different, so there are $4 \cdot 3 = 12$ ways of choosing them.

Here are the 12 possible mini-Sudokus obtained by filling in the empty cells in the above example:

$\mathtt{A}_1 =$

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

$\mathtt{A}_2 =$

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

$\mathtt{A}_3 =$

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

$\mathtt{A}_4 =$

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

$\mathtt{B}_1 =$

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

$\mathtt{B}_2 =$

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

$\mathtt{B}_3 =$

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

$\mathtt{B}_4 =$

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

$\mathtt{C}_1 =$

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

$\mathtt{C}_2 =$

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

$\mathtt{C}_3 =$

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

$\mathtt{C}_4 =$

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

We have labeled these $\mathtt{A}_1, \mathtt{A}_2, \ldots, \mathtt{C}_3, \mathtt{C}_4$ for future reference.

Now we can calculate the number of mini-Sudokus. There are 24 ways of filling in the upper left 2×2 block, and, once that is done, there are 12 ways of filling in the rest of the grid. This gives a total of $24 \cdot 12 = 288$ different mini-Sudokus. (So, while the number of different 9×9 Sudokus—$6,670,903,752,021,072,936,960$—is excessively disgusting, the number of different mini-Sudokus is merely two gross!)

Row and column symmetries

Are the 12 mini-Sudokus listed above really that different from one another? After all, interchanging the last two columns of $\mathtt{A}_1$ gives $\mathtt{A}_2$. Similarly, interchanging the bottom two rows of $\mathtt{A}_2$ gives $\mathtt{A}_3$. Indeed, the mini-Sudokus $\mathtt{A}_1$, $\mathtt{A}_2$, $\mathtt{A}_3$, and $\mathtt{A}_4$ differ only by switching columns and/or rows. We would like to say that these mini-Sudokus are essentially the same or, using a more standard nomenclature, that they are equivalent. Similarly, we would like to say that the mini-Sudokus $\mathtt{B}_1$, $\mathtt{B}_2$, $\mathtt{B}_3$, and $\mathtt{B}_4$ are all equivalent (as are $\mathtt{C}_1$, $\mathtt{C}_2$, $\mathtt{C}_3$, and $\mathtt{C}_4$).

But are $\mathtt{A}_1$ and $\mathtt{B}_1$ equivalent? How about $\mathtt{A}_1$ and $\mathtt{C}_1$? Are *all* mini-Sudokus equivalent in some sense?

To answer these questions, we need to be precise about what *equivalent* means. And to do that, we have to understand the set of mini-Sudoku symmetries. We have already discovered some of these symmetries. For example, interchanging the bottom two rows in any given mini-Sudoku always yields another mini-Sudoku. So the operation of interchanging these two rows is a mini-Sudoku symmetry. If we give this symmetry the symbol ρ then $\rho(\mathtt{A}_1) = \mathtt{A}_4$, $\rho(\mathtt{A}_4) = \mathtt{A}_1$, $\rho(\mathtt{A}_2) = \mathtt{A}_3$, etc. Interchanging the last two columns is also mini-Sudoku symmetry—call it σ. Note that a mini-Sudoku symmetry is, among other things, a one-to-one onto function from the set of all mini-Sudokus to itself.

Composing any two symmetries yields another symmetry. For example, interchanging the bottom two rows, followed by interchanging the last two columns, is also a mini-Sudoku symmetry which we would write as $\sigma\rho$. The symmetry that leaves all mini-Sudokus unchanged is called the *identity symmetry* and denoted id. Every symmetry γ has an *inverse symmetry* γ^{-1} which undoes whatever the symmetry does, that is, $\gamma\gamma^{-1} = \gamma^{-1}\gamma = \mathrm{id}$. For example, $\rho^{-1} = \rho$ since switching the bottom two rows of a mini-Sudoku twice gives the original mini-Sudoku back. For any three mini-Sudoku symmetries ξ, γ, ρ, we have $(\xi\gamma)\rho = \xi(\gamma\rho)$, since function composition is associative. In short, these properties mean that the set of mini-Sudoku symmetries is a group.

Now we can explain equivalence. If K is a group of mini-Sudoku symmetries (that is, a subgroup of the set of all mini-Sudoku symmetries), then two mini-Sudokus $\mathtt{X}$ and $\mathtt{Y}$ are *K-equivalent* if one can be obtained from the other by applying some symmetry in K, that is, $\mathtt{Y} = \gamma(\mathtt{X})$ for some $\gamma \in K$. Since K is a group, this is, in fact, an equivalence relation. The set of all mini-Sudokus which are K-equivalent to $\mathtt{X}$ is called the *K-equivalence class* containing $\mathtt{X}$. Every mini-Sudoku is contained

in a unique K-equivalence class. We say that the group K *acts* on the set of mini-Sudokus. An introduction to the theory of groups acting on sets can be found in, for example, [5] or [9].

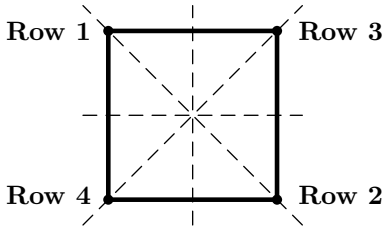
As we have already seen, given a mini-Sudoku, there are some easy ways to make a new mini-Sudoku from it. For example, we could switch the first row with the second row, and leave the bottom two rows alone. Another example would be to send Row 1 to Row 3, Row 3 to Row 2, Row 2 to Row 4, and Row 4 to Row 1.

Since there are four rows in a mini-Sudoku, we can regard the set of such row symmetries as a subgroup R of the symmetric group S_4. However, not all row permutations are symmetries of mini-Sudokus. For example, the permutation taking Row 1 to Row 2, Row 2 to Row 3, and Row 3 to Row 1, leaving Row 4 unchanged, takes the mini-Sudoku $\mathtt{A}_1$ to

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

which is *not* a mini-Sudoku. Thus R is isomorphic to a proper subgroup of S_4. Which subgroup?

The answer is that R is isomorphic to the dihedral group D_4, the group of symmetries of a square. One way to see this is to draw a square and to label its vertices with the rows of the mini-Sudoku, as below. (Do not confuse this square with the mini-Sudoku grid itself—that comes later!)



The group D_4 consists of 8 symmetries: four rotations, by 0°, 90°, 180° and 270°, and four reflections through the axes indicated by the dotted lines in the diagram.

Thus, switching the top two rows of a mini-Sudoku corresponds to reflecting the square about the diagonal axis through the vertices labeled Row 3 and Row 4. Rotation of the square by 90° clockwise corresponds to the mini-Sudoku symmetry which sends Row 1 to Row 3, Row 3 to Row 2, Row 2 to Row 4, and Row 4 to Row 1. The reader should check that each symmetry of the square corresponds to a mini-Sudoku row symmetry and vice versa. The isomorphism between R and D_4 is then transparent.

By replacing the word *row* with the word *column* in the above discussion, we get a new group C of mini-Sudoku column symmetries, again isomorphic to D_4. If $\mu \in R$ and $\nu \in C$, then applying μ and then ν to a mini-Sudoku gives the same result as applying first ν then μ. In other words, row symmetries commute with column symmetries. This means that, combining the 8 row symmetries with the 8 column symmetries, we get 64 different symmetries forming a group $R \times C$ isomorphic to $D_4 \times D_4$.

The reader should check that $\mathtt{A}_1$ and $\mathtt{A}_2$ are C-equivalent but not R-equivalent, and that $\mathtt{A}_1$ and $\mathtt{A}_3$ are R-equivalent but not C-equivalent. The mini-Sudokus $\mathtt{A}_1$, $\mathtt{A}_2$, $\mathtt{A}_3$ and $\mathtt{A}_4$ are all in the same $R \times C$-equivalence class. Similar statements hold for $\mathtt{B}_1$, $\mathtt{B}_2$, $\mathtt{B}_3$ and $\mathtt{B}_4$, as well as for $\mathtt{C}_1$, $\mathtt{C}_2$, $\mathtt{C}_3$ and $\mathtt{C}_4$.

Is $\mathtt{A}_1$ $R \times C$-equivalent to $\mathtt{B}_1$ or $\mathtt{C}_1$? That is, is there some combination of row and column symmetries which applied to $\mathtt{A}_1$ yields $\mathtt{B}_1$ or $\mathtt{C}_1$?

To show that the answer to these questions is no, we associate with each column and row of a mini-Sudoku a partition of the set $\{1, 2, 3, 4\}$—specifically, one of the three partitions

$$\alpha = \{\{1, 2\}, \{3, 4\}\} \qquad \beta = \{\{1, 3\}, \{2, 4\}\} \qquad \gamma = \{\{1, 4\}, \{2, 3\}\}$$

The notation { } means that order does not matter. For example, $\{\{1, 2\}, \{3, 4\}\}$, $\{\{2, 1\}, \{3, 4\}\}$, $\{\{3, 4\}, \{1, 2\}\}$, and $\{\{4, 3\}, \{2, 1\}\}$ are all different ways of writing α.

The partition associated with a row or a column is fairly obvious—just take the entries and put them in order into $\{\{*, *\}, \{*, *\}\}$ in place of the asterisks. For example, all the row partitions of $\mathtt{A}_1$ are α, and all the column partitions are β. All the row partitions of $\mathtt{B}_1$ are α, but the column partitions are β, β, γ and γ from left to right. The row partitions of $\mathtt{C}_1$ are α, α, γ, γ from top to bottom, but all the column partitions are β.

For an arbitrary mini-Sudoku $\mathtt{X}$, it is not hard to see that the Rule of One applied to the top two blocks implies that the partitions associated with Row 1 and Row 2 are the same. Of course, the same holds for Row 3 and Row 4, Column 1 and Column 2, and Column 3 and Column 4. So $\mathtt{X}$ is associated with two row partitions and two column partitions. We will record all this information as an ordered pair $[\mathtt{X}]$ of (unordered) pairs of partitions that we call the *partition type* of $\mathtt{X}$. For example,

$$[\mathtt{A}_1] = (\{\alpha, \alpha\}, \{\beta, \beta\}) \qquad [\mathtt{B}_1] = (\{\alpha, \alpha\}, \{\beta, \gamma\}) \qquad [\mathtt{C}_1] = (\{\alpha, \gamma\}, \{\beta, \beta\})$$

The first entry contains the two row partitions, and the second entry contains the two column partitions. Note that $(\{\alpha, \alpha\}, \{\beta, \gamma\})$ and $(\{\alpha, \alpha\}, \{\gamma, \beta\})$ are equal, but $(\{\alpha, \alpha\}, \{\beta, \gamma\})$ is not equal to $(\{\beta, \gamma\}, \{\alpha, \alpha\})$). In particular, $[\mathtt{A}_1] = [\mathtt{A}_2] = [\mathtt{A}_3] = [\mathtt{A}_4]$, $[\mathtt{B}_1] = [\mathtt{B}_2] = [\mathtt{B}_3] = [\mathtt{B}_4]$ and $[\mathtt{C}_1] = [\mathtt{C}_2] = [\mathtt{C}_3] = [\mathtt{C}_4]$.

What makes these partitions useful is how they change under the mini-Sudoku symmetries we have discussed. For example, applying the eight row symmetries in R to the first column of $\mathtt{A}_1$ yields eight different columns:

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

However, each of these columns is associated with the same partition, namely β.

Thus column partitions are invariant under the row symmetries, and, similarly, row partitions are invariant under the column symmetries. Of course, the column partitions are simply permuted by column symmetries, and row partitions are permuted by row symmetries. Thus we have the following rule:

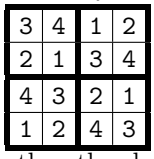
> **Rule 1:** If mini-Sudokus $\mathtt{X}$ and $\mathtt{Y}$ are $R \times C$-equivalent, then their partition types are the same—that is, $[\mathtt{X}] = [\mathtt{Y}]$.

Since the partition types of $\mathtt{A}_1$, $\mathtt{B}_1$, and $\mathtt{C}_1$ are distinct, no pair of these mini-Sudokus is $R \times C$-equivalent.

The situation we have been considering is typical in mathematics. One has a collection of objects (in our case, mini-Sudokus) and a notion of equivalence, often from a group action (in our case, the group is $R \times C$). The goal is to determine which objects are equivalent. In algebra, the objects might be groups, rings, or fields, and *equivalent* means *isomorphic*. In topology, the objects might be topological spaces or manifolds, and *equivalent* means *homeomorphic*. In linear algebra, the objects might be square matrices, two of which are equivalent if they are similar.

The general strategy for such problems is to attach an *invariant* to each object—something which is the same for equivalent objects. The partition type is an invariant for mini-Sudokus; that is what Rule 1 says.

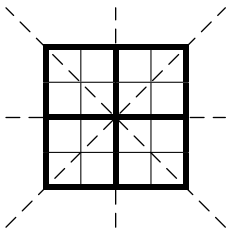
Other examples of invariants are the order of a group, the elementary divisors of a finite abelian group, the characteristic of a field, the fundamental group of a topological space, the genus of a compact surface, the determinant of a square matrix, and the Jordan canonical form of a square matrix with complex entries. The ideal invariant is easy to compute and completely determines whether or not two objects are equivalent (in which case, we say the invariant is *complete*). The set of elementary divisors is a complete invariant for the set of finite abelian groups. The order of a group is not a complete invariant, however, as nonisomorphic groups can have the same order. The determinant is not a complete invariant for square matrices, but the Jordan canonical form is.

When we defined the notion of *partition type*, we did so carefully, to ensure that it would be an invariant. In particular, it was necessary to define $[\mathtt{X}]$ as an ordered pair of *unordered* pairs. For example, let $\mathtt{D} =$  . Notice that $\mathtt{D}$ and $\mathtt{B}_1$ are $R \times C$-equivalent; you can obtain one from the other by swapping the two left columns with the two right columns. Rule 1 assures us that they have the same partition type, and indeed we can verify directly that $[\mathtt{D}] = (\{\alpha, \alpha\}, \{\gamma, \beta\}) = [\mathtt{B}_1]$. Had we defined $[\mathtt{X}]$ as an ordered pair of *ordered* pairs, however, we would not have had the desired equality $[\mathtt{D}] = [\mathtt{B}_1]$.

Is partition type a complete invariant for mini-Sudokus with respect to $R \times C$-equivalence? In other words, can we find two mini-Sudokus which have the same partition type, but which are not $R \times C$-equivalent? We will see the answer to this question later.

Geometric symmetries

Have we now found all mini-Sudoku symmetries? Definitely not! After all, a mini-Sudoku is itself a square, and it is not hard to see that any symmetry of a square is also a symmetry of mini-Sudokus.



For example, reflecting a mini-Sudoku across its horizontal axis produces a new mini-Sudoku. But this symmetry is just the row symmetry that reverses the order of the rows—it interchanges Row 1 and Row 4, and Row 2 and Row 3. Similarly, reflecting across the vertical axis is a column symmetry.

What about reflections across a diagonal? For concreteness, let τ be the symmetry which reflects mini-Sudokus across the main diagonal (from top left to bottom right). For example,

$$\tau(\mathtt{A}_1) = \begin{array}{|cc|cc|}\hline 1&3&2&4\\ 2&4&1&3\\ \hline 3&1&4&2\\ 4&2&3&1\\ \hline\end{array}.$$

Since $[\tau(\mathtt{A}_1)] = (\{\beta,\beta\},\{\alpha,\alpha\})$, this mini-Sudoku is not $R\times C$-equivalent to $\mathtt{A}_1$. In other words, the symmetry τ cannot be in $R\times C$. This particular symmetry will have an important role in our discussion. Note that $\tau^2 = \mathrm{id}$. This means that applying τ twice to any mini-Sudoku leaves it unchanged. Hence $\tau^{-1} = \tau$ and $Z = \{\mathrm{id}, \tau\}$ is a group of mini-Sudoku symmetries isomorphic to $\mathbb{Z}_2$.

Rotating 90° clockwise is the result of first reflecting across the main diagonal and then reflecting across the vertical axis. Applying these two symmetries in the other order results in a rotation of 270°. Thus the rotations by 90° and 270° are compositions of τ and symmetries in $R\times C$. Rotation by 180° is the composition of the reflections across the horizontal and vertical axes (in either order). So this rotation is in $R\times C$. Specifically, it is the result of reversing the orders of the rows and the columns.

For example, rotating $\mathtt{B}_1$ by 90°, 180° and 270° clockwise we get the mini-Sudokus

$$\mathtt{S} = \begin{array}{|cc|cc|}\hline 4&2&3&1\\ 3&1&4&2\\ \hline 1&4&2&3\\ 2&3&1&4\\ \hline\end{array} \qquad \mathtt{T} = \begin{array}{|cc|cc|}\hline 2&1&3&4\\ 3&4&1&2\\ \hline 1&2&4&3\\ 4&3&2&1\\ \hline\end{array} \qquad \mathtt{U} = \begin{array}{|cc|cc|}\hline 4&1&3&2\\ 3&2&4&1\\ \hline 2&4&1&3\\ 1&3&2&4\\ \hline\end{array}$$

with partition types $[\mathtt{S}] = [\mathtt{U}] = (\{\beta,\gamma\},\{\alpha,\alpha\})$ and $[\mathtt{T}] = (\{\alpha,\alpha\},\{\beta,\gamma\})$. Since $[\mathtt{B}_1] = (\{\alpha,\alpha\},\{\beta,\gamma\})$, neither $\mathtt{S}$ nor $\mathtt{U}$ is $R\times C$-equivalent to $\mathtt{B}_1$. This means that the corresponding symmetries, rotation by 90° and 270°, are not in $R\times C$.

What about $\mathtt{T}$? As suggested above, $\mathtt{T}$ can be obtained from $\mathtt{B}_1$ by reversing the orders of both the rows and the columns:

$$\mathtt{B}_1 = \begin{array}{|cc|cc|}\hline 1&2&3&4\\ 3&4&2&1\\ \hline 2&1&4&3\\ 4&3&1&2\\ \hline\end{array} \longrightarrow \begin{array}{|cc|cc|}\hline 4&3&1&2\\ 2&1&4&3\\ \hline 3&4&2&1\\ 1&2&3&4\\ \hline\end{array} \longrightarrow \begin{array}{|cc|cc|}\hline 2&1&3&4\\ 3&4&1&2\\ \hline 1&2&4&3\\ 4&3&2&1\\ \hline\end{array} = \mathtt{T}$$

There is just one symmetry in D_4 which we have not yet discussed, namely reflection across the other diagonal (from top right to bottom left). We leave the reader the task of showing that this symmetry is not in $R \times C$, but is, nonetheless, τ composed with a symmetry in $R \times C$.

Note that τ interchanges the rows and the columns of any mini-Sudoku. It sends Row 1 to Column 1, Row 2 to Column 2, etc. This implies also that τ interchanges row symmetries and column symmetries. For example, if $\rho \in R$ is the row symmetry which interchanges Row 1 and Row 2, then $\sigma = \tau\rho\tau$ is the column symmetry which interchanges Column 1 and Column 2. This equation can be written as $\sigma\tau = \tau\rho$, which shows that, even though τ does not commute with elements of $R \times C$, it does so at the cost of interchanging rows and columns. As a consequence, any symmetry which can be obtained by composing τ and elements of $R \times C$ in any order can be written in the form $\tau\mu$ with $\mu \in R \times C$. (The same symmetry can also be written in the form $\nu\tau$ with $\nu \in R \times C$ where ν and μ are the same except for the interchange of rows and columns.)

We now have 64 symmetries in $R \times C$, and 64 more symmetries of the form $\tau\mu$ with $\mu \in R \times C$. In the second category are the rotations by 90° and 270°, as well as the reflections across the diagonal axes. Together, these symmetries form a group H of order 128. Since τ does not commute with all elements of $R \times C$, we know that H is not the direct product of $R \times C$ and $Z = \{\mathrm{id}, \tau\}$. Instead, H is a semi-direct product [9] of these groups:

$$H = (R \times C) \rtimes Z.$$

In other words, $R \times C$ is normal in H and has trivial intersection with Z.

Naturally, we will say that two mini-Sudokus $\mathtt{X}$ and $\mathtt{Y}$ are *H-equivalent* if one can be obtained from the other by applying one of the symmetries in H.

Is $\mathtt{A}_1$ H-equivalent to $\mathtt{B}_1$ or $\mathtt{C}_1$? That is, is there some symmetry in H which applied to $\mathtt{A}_1$ yields $\mathtt{B}_1$ or $\mathtt{C}_1$? Once again, using partition types, we can show that the answer is no.

Since the group H acts on mini-Sudokus, it also acts on partition types of mini-Sudokus. By Rule 1, symmetries in H which are also in $R \times C$ leave partition types unchanged. Because τ interchanges the rows and columns of mini-Sudokus, this symmetry interchanges the associated row and column partitions of any partition type. For example, since $[\mathtt{C}_1] = (\{\alpha, \gamma\}, \{\beta, \beta\})$, we have $[\tau(\mathtt{C}_1)] = (\{\beta, \beta\}, \{\alpha, \gamma\})$.

In view of the nature of the symmetry τ, it is natural to call $\tau(\mathtt{X})$ the *transpose* of $\mathtt{X}$ and $[\tau(\mathtt{X})]$ the *transpose* of $[\mathtt{X}]$. So we use the notation $[\mathtt{X}]^T = [\tau(\mathtt{X})]$. Thus $[\mathtt{X}]^T$ is obtained from $[\mathtt{X}]$ by switching its two entries. Now, if mini-Sudokus $\mathtt{X}$ and $\mathtt{Y}$ are H-equivalent, then $\mathtt{X}$ is $R \times C$-equivalent to $\mathtt{Y}$ or to $\tau(\mathtt{Y})$. We therefore say that $[\mathtt{X}]$ and $[\mathtt{Y}]$ are *H-equivalent* if $[\mathtt{X}] = [\mathtt{Y}]$ or $[\mathtt{X}] = [\mathtt{Y}]^T$. Define $[\mathtt{X}]_H$ to be the H-equivalence class of $[\mathtt{X}]$. Note that we now have two H-equivalences: H-equivalence of mini-Sudokus and H-equivalence of partition types.

With Rule 1, we have the following:

Rule 2: If mini-Sudokus $\mathtt{X}$ and $\mathtt{Y}$ are H-equivalent, then $[\mathtt{X}]_H = [\mathtt{Y}]_H$.

Since we have $[\mathtt{A}_1]^T = (\{\beta,\beta\},\{\alpha,\alpha\})$, $[\mathtt{B}_1]^T = (\{\beta,\gamma\},\{\alpha,\alpha\})$ and $[\mathtt{C}_1]^T = (\{\beta,\beta\},\{\alpha,\gamma\})$, no pair of the mini-Sudokus $\mathtt{A}_1$, $\mathtt{B}_1$ and $\mathtt{C}_1$ is H-equivalent.

Is $[\cdot]_H$ a complete invariant with respect to H-equivalence? In other words, are there mini-Sudokus $\mathtt{X}$ and $\mathtt{Y}$ which have H-equivalent partition types, but which are not H-equivalent? We will see the answer to this question shortly.

Relabeling symmetries

There is yet one other way of creating a new mini-Sudoku from a given mini-Sudoku—simply relabel it, that is, apply a permutation of the set $\{\mathtt{1},\mathtt{2},\mathtt{3},\mathtt{4}\}$ to its entries. For example, starting with $\mathtt{A}_1$, we could interchange $\mathtt{1}$ and $\mathtt{2}$ to get

$$\mathtt{V} = \begin{array}{|cc|cc|}\hline 2&1&3&4\\ 3&4&2&1\\ \hline 1&2&4&3\\ 4&3&1&2\\ \hline\end{array}.$$

Since $[\mathtt{V}] = (\{\alpha,\alpha\},\{\gamma,\gamma\})$, $\mathtt{V}$ is not H-equivalent to $\mathtt{A}_1$, and so this relabeling symmetry, switching $\mathtt{1}$ and $\mathtt{2}$, is not in H. We have found a new symmetry! Since there are $4! = 24$ different permutations of $\{\mathtt{1},\mathtt{2},\mathtt{3},\mathtt{4}\}$ forming the group S_4, there is a corresponding group $L \cong S_4$ of relabeling symmetries of mini-Sudokus.

What do the relabeling symmetries do to the partitions α, β, and γ? The answer is that these symmetries permute them. For example, interchanging $\mathtt{1}$ and $\mathtt{2}$ takes α to α, β to γ, and γ to β. This is why this particular labelling symmetry takes $[\mathtt{A}_1] = (\{\alpha,\alpha\},\{\beta,\beta\})$ to $[\mathtt{V}] = (\{\alpha,\alpha\},\{\gamma,\gamma\})$.

For another example, consider the relabeling symmetry $\lambda \in L$ which maps $\mathtt{2}$ to $\mathtt{3}$, $\mathtt{3}$ to $\mathtt{4}$, and $\mathtt{4}$ to $\mathtt{2}$, leaving $\mathtt{1}$ fixed. In cycle notation, we would write $\lambda = (\mathtt{2},\mathtt{3},\mathtt{4})$. This symmetry takes α to β, β to γ, and γ back to α. It is not hard to see that each of the six permutations of α, β, and γ comes from exactly 4 relabeling symmetries in L. Hence we say that two partition types are *L-equivalent* if one can be obtained from the other by a permutation of α, β and γ.

We claim that there is no element of H that has the same effect on all mini-Sudokus as the relabeling symmetry $\lambda = (\mathtt{2},\mathtt{3},\mathtt{4})$. To show that this is so, we pick a mini-Sudoku with partition type $(\{\alpha,\alpha\},\{\beta,\gamma\})$—for example, $\mathtt{B}_1$ will do. Then λ acting on this mini-Sudoku produces a mini-Sudoku with partition type $(\{\beta,\beta\},\{\gamma,\alpha\})$. By Rule 2, the new mini-Sudoku is not H-equivalent to the original one, and so λ cannot be in H. (Another way to see that $\lambda \notin H$ is to use Lagrange's theorem [5, 9]. No element of order 3 can be in H, since $|H| = 128$.) Are *any* of the symmetries in L also in H? It turns out that only the identity symmetry is in both groups. We sketch a proof of this fact, leaving the details to the reader. First observe that if $\lambda \in L$ and $\sigma \in H$, then applying first λ and then σ has the same effect as applying first σ and then λ. In other words, $\lambda\sigma = \sigma\lambda$. So if $\sigma \in L \cap H$, then σ is in the center of L. But L is isomorphic to S_4, whose center contains only the identity.

How many symmetries do we now have? We know that relabeling symmetries commute with symmetries in H, so combining all of these symmetries, we get a group isomorphic to the direct product of H and L. Therefore we define

The mini-Sudoku Symmetry Group $= G \cong H \times L$.

This group has order $|G| = |H| \cdot |L| = 128 \cdot 24 = 3072$. Since G contains all mini-Sudoku symmetries that we wish to consider, instead of saying that mini-Sudokus $\mathtt{X}$ and $\mathtt{Y}$ are G-equivalent, we will just say that they are *equivalent*. Equivalence of this type is what is meant by "essentially the same" in the introduction and in [7, 8].

Why should G be *the* group of symmetries that we, and others, have chosen to consider? Here's why. Let M be the set of all sixteen cells in a 4×4 grid. Then a mini-Sudoku is nothing more and nothing less than a function $f : M \to \{1, 2, 3, 4\}$ which obeys the Rule of One. An element $\lambda \in L$ acts on a mini-Sudoku f by pre-composition, sending f to $\lambda \circ f$. Likewise, H is a subgroup of the group S_M of all bijective functions from M to itself, and an element $\sigma \in H$ acts on a mini-Sudoku f by post-composition, sending f to $f \circ \sigma$. We have chosen H with care, so that $f \circ \sigma$ necessarily still obeys the Rule of One; hence we shall say that every element of H is *mini-Sudoku-preserving*. It is tedious but straightforward to verify the converse, that every mini-Sudoku-preserving element of S_M is in H. So we have not chosen G arbitrarily at all—it is the set of all mini-Sudoku symmetries that can be obtained by permuting cells and permuting labels.

(For 9×9 Sudokus, the group generated by the row and column symmetries together with the rotations and reflections has order $3,359,232$, and there are $9!$ relabeling symmetries. Hence the Sudoku symmetry group has order $3,359,232 \cdot 9! = 1,218,998,108,160$. We remark that the row symmetry group for an $n^2 \times n^2$ Sudoku is an n-fold wreath product [8].)

Note that, like the groups H and L, the group G acts on mini-Sudokus and also on partition types. Define $[\mathtt{X}]_G$ to be the G-equivalence class of $[\mathtt{X}]$. In other words, $[\mathtt{X}]_G = [\mathtt{Y}]_G$ if and only if $[\mathtt{X}]$ is L-equivalent to $[\mathtt{Y}]$ or to $[\mathtt{Y}]^T$.

Rule 3: If mini-Sudokus $\mathtt{X}$ and $\mathtt{Y}$ are equivalent, then $[\mathtt{X}]_G = [\mathtt{Y}]_G$.

Now we can see whether $\mathtt{A}_1$, $\mathtt{B}_1$, and $\mathtt{C}_1$ are equivalent. If a mini-Sudoku $\mathtt{X}$ is equivalent to $\mathtt{A}_1$, then, by Rule 3, $[\mathtt{X}]$ is $(\{\alpha,\alpha\},\{\beta,\beta\})$, $(\{\beta,\beta\},\{\alpha,\alpha\})$, $(\{\alpha,\alpha\},\{\gamma,\gamma\})$, $(\{\gamma,\gamma\},\{\alpha,\alpha\})$, $(\{\beta,\beta\},\{\gamma,\gamma\})$ or $(\{\gamma,\gamma\},\{\beta,\beta\})$. So, in particular, $\mathtt{X}$ cannot be $\mathtt{B}_1$ or $\mathtt{C}_1$, and so $\mathtt{A}_1$ is not equivalent to either of these mini-Sudokus.

What about the equivalence of $\mathtt{B}_1$ and $\mathtt{C}_1$? If $\mathtt{X}$ is equivalent to $\mathtt{B}_1$, then, by Rule 3, $[\mathtt{X}]$ is $(\{\alpha,\alpha\},\{\beta,\gamma\})$, $(\{\beta,\gamma\},\{\alpha,\alpha\})$, $(\{\beta,\beta\},\{\alpha,\gamma\})$, $(\{\alpha,\gamma\},\{\beta,\beta\})$, $(\{\gamma,\gamma\},\{\alpha,\beta\})$ or $(\{\alpha,\beta\},\{\gamma,\gamma\})$. Because $[\mathtt{C}_1] = (\{\alpha,\gamma\},\{\beta,\beta\})$, it is possible that $\mathtt{B}_1$ and $\mathtt{C}_1$ are equivalent. Since we have not (yet) shown that the converse of Rule 3 holds, we do not yet know whether $\mathtt{B}_1$ is equivalent to $\mathtt{C}_1$ or not.

If these mini-Sudokus are equivalent, then the partition types of $\mathtt{B}_1$ and $\mathtt{C}_1$ suggest how to construct a symmetry that takes one to the other. There will have to be a relabeling symmetry which interchanges α and β, composed with the transposition τ, composed (perhaps) with some row and column symmetry:

$$[\mathtt{B}_1] = (\{\alpha,\alpha\},\{\beta,\gamma\}) \longrightarrow (\{\beta,\beta\},\{\alpha,\gamma\}) \longrightarrow (\{\alpha,\gamma\},\{\beta,\beta\}) = [\mathtt{C}_1].$$

In fact, by choosing the relabeling symmetry, $\lambda \in L$, which interchanges $\mathtt{2}$ and $\mathtt{3}$, no row or column symmetry is needed:

$$\mathtt{B}_1 = \begin{array}{|cc|cc|}\hline 1&2&3&4\\ 3&4&2&1\\ \hline 2&1&4&3\\ 4&3&1&2\\ \hline\end{array} \xrightarrow{\lambda} \begin{array}{|cc|cc|}\hline 1&3&2&4\\ 2&4&3&1\\ \hline 3&1&4&2\\ 4&2&1&3\\ \hline\end{array} \xrightarrow{\tau} \begin{array}{|cc|cc|}\hline 1&2&3&4\\ 3&4&1&2\\ \hline 2&3&4&1\\ 4&1&2&3\\ \hline\end{array} = \mathtt{C}_1.$$

This shows that $\mathtt{B}_1$ and $\mathtt{C}_1$ are equivalent, and hence that $\mathtt{B}_1$, $\mathtt{B}_2$, $\mathtt{B}_3$, $\mathtt{B}_4$, $\mathtt{C}_1$, $\mathtt{C}_2$, $\mathtt{C}_3$, and $\mathtt{C}_4$ are all in the same equivalence class. Since $\mathtt{A}_1$ and $\mathtt{B}_1$ are not equivalent, there must be a second equivalence class containing $\mathtt{A}_1$, $\mathtt{A}_2$, $\mathtt{A}_3$, and $\mathtt{A}_4$.

Now we are ready for the main (and only) theorem of this article.

Theorem. There are exactly two equivalence classes of mini-Sudokus:

$\mathcal{C}_1$: All mini-Sudokus with the following partition types:

$$\begin{array}{lll} (\{\alpha,\alpha\},\{\beta,\beta\}) & (\{\alpha,\alpha\},\{\gamma,\gamma\}) & (\{\beta,\beta\},\{\gamma,\gamma\}) \\ (\{\beta,\beta\},\{\alpha,\alpha\}) & (\{\gamma,\gamma\},\{\alpha,\alpha\}) & (\{\gamma,\gamma\},\{\beta,\beta\}) \end{array}$$

$\mathcal{C}_2$: All mini-Sudokus with the following partition types:

$$\begin{array}{lll} (\{\alpha,\alpha\},\{\beta,\gamma\}) & (\{\beta,\beta\},\{\alpha,\gamma\}) & (\{\gamma,\gamma\},\{\alpha,\beta\}) \\ (\{\beta,\gamma\},\{\alpha,\alpha\}) & (\{\alpha,\gamma\},\{\beta,\beta\}) & (\{\alpha,\beta\},\{\gamma,\gamma\}) \end{array}$$

Proof. We know already that there are at least two distinct equivalence classes.

Let $\mathtt{X}$ be a mini-Sudoku. By applying a suitable relabeling symmetry, the top left box of $\mathtt{X}$ can be put in the form $\begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 & 4 \\ \hline \end{array}$, and so $\mathtt{X}$ is equivalent to one of the 12 mini-Sudokus $\mathtt{A}_1, \mathtt{A}_2, \ldots, \mathtt{C}_3, \mathtt{C}_4$. From the above discussion, $\mathtt{X}$ is equivalent to either $\mathtt{A}_1$ or $\mathtt{B}_1$. But we have seen already that, if $\mathtt{X}$ is equivalent to $\mathtt{A}_1$, then its partition type is as described in $\mathcal{C}_1$, and if $\mathtt{X}$ is equivalent to $\mathtt{B}_1$, then its partition type is as described in $\mathcal{C}_2$. □

(A similar argument in [7] purports to demonstrate the same result; in fact, the line of reasoning in that article shows only that there are *at most* two equivalence classes. What is missing is an invariant to distinguish the two classes.)

There are exactly 24 mini-Sudokus L-equivalent to each of $\mathtt{A}_1, \mathtt{A}_2, \ldots, \mathtt{C}_3, \mathtt{C}_4$, and hence there are $4 \cdot 24 = 96$ mini-Sudokus in $\mathcal{C}_1$ and $8 \cdot 24 = 192$ mini-Sudokus in $\mathcal{C}_2$.

Since we now know the partition types that are in each of the equivalence classes, it is easy to see that the converse of Rule 3 holds, that is, mini-Sudokus $\mathtt{X}$ and $\mathtt{Y}$ are equivalent if and only if $[\mathtt{X}]$ and $[\mathtt{Y}]$ are equivalent.

Notice that the mini-Sudokus $\mathtt{A}_1$, $\mathtt{A}_2$, $\mathtt{A}_3$, and $\mathtt{A}_4$ have either two or four distinct entries on the main diagonal, whereas the mini-Sudokus $\mathtt{B}_1, \mathtt{B}_2, \ldots, \mathtt{C}_3, \mathtt{C}_4$ have exactly three distinct entries on the main diagonal. Since any mini-Sudoku is L-equivalent to one of these 12, and the number of distinct entries on the main diagonal is unchanged by relabeling symmetries, it now quite easy to tell which equivalence class a mini-Sudoku $\mathtt{X}$ belongs to. If $\mathtt{X}$ has two or four distinct entries on the main diagonal it must be L-equivalent to $\mathtt{A}_1$, $\mathtt{A}_2$, $\mathtt{A}_3$, or $\mathtt{A}_4$, and so $\mathtt{X}$ is in $\mathcal{C}_1$. If $\mathtt{X}$ has three distinct entries on the main diagonal it must be L-equivalent to one of $\mathtt{B}_1, \mathtt{B}_2, \ldots, \mathtt{C}_3, \mathtt{C}_4$, and so $\mathtt{X}$ is in $\mathcal{C}_2$. Hence the diagonal entries of $\mathtt{X}$ suffice to determine its equivalence class.

It is, of course, easier to count entries along the main diagonal of a mini-Sudoku than to write down its partition type. So why bother with partition types at all? The reason is that they are better suited for weaker forms of equivalence, such as $R \times C$-equivalence and H-equivalence. In fact, the following converses to Rules 1 and 2 assert that $[\cdot]$ is a complete invariant of mini-Sudokus, modulo $R \times C$-equivalence and that $[\cdot]_H$ is a complete invariant with respect to H-equivalence:

Proposition. Let $\mathtt{X}$ and $\mathtt{Y}$ be mini-Sudokus.

(1) If $[\mathtt{X}] = [\mathtt{Y}]$, then $\mathtt{X}$ and $\mathtt{Y}$ are $R \times C$-equivalent.

(2) If $[\mathtt{X}]_H = [\mathtt{Y}]_H$, then $\mathtt{X}$ and $\mathtt{Y}$ are H-equivalent.

Proof. (1) Suppose $[\mathtt{X}] = [\mathtt{Y}]$. Recall that Row 1 and Row 2 of $\mathtt{X}$ are associated with the same partition—either α, β, or γ—and that the same is true of Rows 3 and 4. For convenience, we will call these four partitions, in order, the *row partitions* of $\mathtt{X}$.

It may be that the row partitions of $\mathtt{X}$ match (row for row) the row partitions of $\mathtt{Y}$, in which case let $\mathtt{Y}_1 = \mathtt{Y}$. If this is not the case, then, since $[\mathtt{X}] = [\mathtt{Y}]$, applying the *blockwise* row symmetry that switches Rows 1 and 2 with Rows 3 and 4 to $\mathtt{Y}$, yields a mini-Sudoku $\mathtt{Y}_1$ whose row partitions match those of $\mathtt{X}$. Similarly, by applying a blockwise column symmetry to $\mathtt{Y}_1$ if necessary, we obtain a mini-Sudoku $\mathtt{Y}_2$ such that both the row and column partitions of $\mathtt{X}$ and $\mathtt{Y}_2$ match up, and such that $\mathtt{Y}_2$ is $R \times C$-equivalent to $\mathtt{Y}$.

The two row symmetries that switch Rows 1 and 2, and Rows 3 and 4, and the two column symmetries that switch Columns 1 and 2, and Columns 3 and 4, can be used to put any mini-Sudoku into the form

1	*	*	*
*	*	*	*
*	*	1	*
*	*	*	*

.

Moreover, this can be done without changing row and column partitions. So by applying this procedure to $\mathtt{Y}_2$, we obtain a mini-Sudoku $\mathtt{Y}_3$, which is $R \times C$-equivalent to $\mathtt{Y}_2$, such that $\mathtt{X}$ and $\mathtt{Y}_3$ have this special form, in addition to having matching row and column partitions.

Since $\mathtt{X}$ and $\mathtt{Y}_3$ have the same top row partition, it follows that they have the same entries in Row 1, Column 2. Similarly, using the leftmost column partition, we see they have the same entries in Row 2, Column 1. Thus, they have the same upper-left block. Likewise, we find that they have the same lower-right block. We can then use the Rule of One to fill in the remaining entries in the grid and conclude that $\mathtt{X} = \mathtt{Y}_3$. The result follows.

(2) We know that $[\mathtt{X}] = [\mathtt{Y}]$ or $[\mathtt{X}] = [\mathtt{Y}]^{\mathrm{T}}$. So by (1), $\mathtt{X}$ is $R \times C$-equivalent to $\mathtt{Y}$ or to $\mathtt{Y}^{\mathrm{T}}$. In either case, $\mathtt{X}$ is H-equivalent to $\mathtt{Y}$. □

More mini-Sudoku puzzles

Though we now know how many mini-Sudokus there are and how many of them are essentially different, many mini-Sudokus puzzles remain for the reader to solve. Here are some suggestions:

(1) According to the theorem, if $\mathtt{X}$ is a mini-Sudoku, then $[\mathtt{X}] = (\{\alpha, \alpha\}, \{\alpha, \alpha\})$, $[\mathtt{X}] = (\{\alpha, \beta\}, \{\alpha, \beta\})$, $[\mathtt{X}] = (\{\alpha, \alpha\}, \{\alpha, \beta\})$, and $[\mathtt{X}] = (\{\alpha, \beta\}, \{\alpha, \gamma\})$ are not possible. Show directly that it is not possible for a row partition to equal a column partition.

(2) For a mini-Sudoku $\mathtt{X}$, let $\det \mathtt{X}$ be the determinant of $\mathtt{X}$ thought of as a 4×4 matrix. Then $\det \mathtt{X}$ is unchanged or changes sign under the symmetries in H. Hence, if $\mathtt{X}$ and $\mathtt{Y}$ are H-equivalent, then $|\det \mathtt{X}| = |\det \mathtt{Y}|$. Is the converse true?

It might be useful to replace the entries, 1, 2, 3 and 4, by variables, w, x, y and z, so that $\det \mathtt{X}$ is a polynomial in the four variables. For example,

$$\det \mathtt{A}_1 = \det \begin{bmatrix} w & x & y & z \\ y & z & w & x \\ x & w & z & y \\ z & y & x & w \end{bmatrix}$$
$$= -(w+x-y-z)(w+y-x-z)(w+z-x-y)(w+x+y+z)$$

How do these determinants change under relabeling symmetries? Can such determinants be used to determine whether mini-Sudokus are H-equivalent or equivalent?

(3) Prove that $R \cap C = (R \times C) \cap Z = \{\mathrm{id}\}$.

References

[1] S.F. Bammel and J. Rothstein, The Number of 9×9 Latin squares, *Discrete Mathematics* **11** (1975), 93–95.

[2] T. Davis, *The Mathematics of Sudoku*, http://www.geometer.org/mathcircles/sudoku.pdf, October 2008.

[3] J-P. Delahaye, The Science behind Sudoku, *Sci. American*, June 2006.

[4] B. Felgenhauer and F. Jarvis, Mathematics of Sudoku I, *Mathematical Spectrum* **39** (2006), 15–22.

[5] J. Gallian, *Contemporary Abstract Algebra, 6th edition*, Houghton-Mifflin, 2006.

[6] B. Hayes, Unwed Numbers, *American Scientist* **94** (2006), 12–15.

[7] A. Herzberg and M. Murty, Sudoku Squares and Chromatic Polynomials, *Notices of the AMS*, **54** (2007), 708–717.

[8] F. Jarvis and E. Russell Mathematics of Sudoku II, *Mathematical Spectrum* **39** (2006), 54–58.

[9] J. Rotman, *An Introduction to the Theory of Groups*, Springer Verlag, 1995.

California State University, Los Angeles CA 90032-8204

E-mail address: gbrookf@calstatela.edu, mkrebs@calstatela.edu