The Mathematics of Sudoku

Tom Davis
tomrdavis@earthlink.net
http://www.geometer.org/mathcircles
(Preliminary)
October 12, 2008

1 Introduction

Sudoku is a (sometimes addictive) puzzle presented on a square grid that is usually 9×9, but is sometimes 16×16 or other sizes. We will consider here only the 9×9 case, although most of what follows can be extended to larger puzzles. Sudoku puzzles can be found in many daily newspapers, and there are thousands of references to it on the internet. Most puzzles are ranked as to difficulty, but the rankings vary from puzzle designer to puzzle designer.

Sudoku is an abbreviation for a Japanese phrase meaning ”the digits must remain single”, and it was in Japan that the puzzle first became popular. The puzzle is also known as ”Number Place”. Sudoku (although it was not originally called that) was apparently invented by Howard Garns in 1979. It was first published by Dell Magazines (which continues to do so), but now is available in hundreds of publications.

At the time of publication of this article, Sudoku is very popular, but it is of course difficult to predict whether it will remain so. It does have many features of puzzles that remain popular: puzzles are available of all degrees of difficulty, the rules are very simple, your ability to solve them improves with time, and it is the sort of puzzle where the person solving it makes continuous progress toward a solution, as is the case with crossword puzzles.

The original grid has some of the squares filled with the digits from 1 to 9 and the goal is to complete the grid so that every row, column and outlined 3×3 sub-grid contains each of the digits exactly once. A valid puzzle admits exactly one solution.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
a			4	8					
b		9		4	6			7	
c		5					6	1	4
d	2	1		6			5		
e	5	8		7		9		4	1
f			7			8		6	9
g	3	4	5					9	
h		6			3	7		2	
i						4	1		

Figure 1: An easy sudoku puzzle

Figure 1 is a relatively easy sudoku puzzle. If you have never tried to solve one, attempt this one (using a pencil!) before you continue, and see what strategies you can find. It will probably take more time than you think, but you will get much better with practice. The solution appears in section 20.

Sudoku is mathematically interesting in a variety of ways. Both simple and intricate logic can be applied to solve a puzzle, it can be viewed as a graph coloring problem and it certainly has some interesting combinatorial aspects.

We will begin by examining some logical and mathematical approaches to solving sudoku puzzles beginning with the most obvious and we will continue to more and more sophisticated techniques (see, for example, multi-coloring, described in section 9.2). Later in this article we will look at a few other mathematical aspects of sudoku.

A large literature on sudoku exists on the internet with a fairly standardized terminology, which we will use here:

- A "square" refers to one of the 81 boxes in the sudoku grid, each of which is to be filled eventually with a digit from 1 to 9.
- A "block" refers to a 3×3 sub-grid of the main puzzle in which all of the numbers must appear exactly once in a solution. We will refer to a block by its columns and rows. Thus block $ghi456$ includes the squares $g4$, $g5$, $g6$, $h4$, $h5$, $h6$, $i4$, $i5$ and $i6$.
- A "candidate" is a number that could possibly go into a square in the grid. Each candidate that we can eliminate from a square brings us closer to a solution.
- Many arguments apply equally well to a row, column or block, and to keep from having to write "row, column or block" over and over, we may refer to it as a "virtual line". A typical use of "virtual line" might be this: "If you know the values of 8 of the 9 squares in a virtual line, you can always deduce the value of the missing one." In the 9×9 sudoku puzzles there are 27 such virtual lines.
- Sometimes you would like to talk about all of the squares that cannot contain the same number as a given square since they share a row or column or block. These are sometimes called the "buddies" of that square. For example, you might say something like, "If two buddies of a square have only the same two possible candidates, then you can eliminate those as candidates for the square." Each square has 20 buddies.

2 Obvious Strategies

Strategies in this section are mathematically obvious, although searching for them in a puzzle may sometimes be difficult, simply because there are a lot of things to look for. Most puzzles ranked as "easy" and even some ranked "intermediate" can be completely solved using only techniques discussed in this section. The methods are presented roughly in order of increasing difficulty for a human. For a computer, a completely different approach is often simpler.

2.1 Unique Missing Candidate

If eight of the nine elements in any virtual line (row, column or block) are already determined, the final element has to be the one that is missing. When most of the squares are already filled in this technique is heavily used. Similarly: If eight of the nine values are impossible in a given square, that square's value must be the ninth.

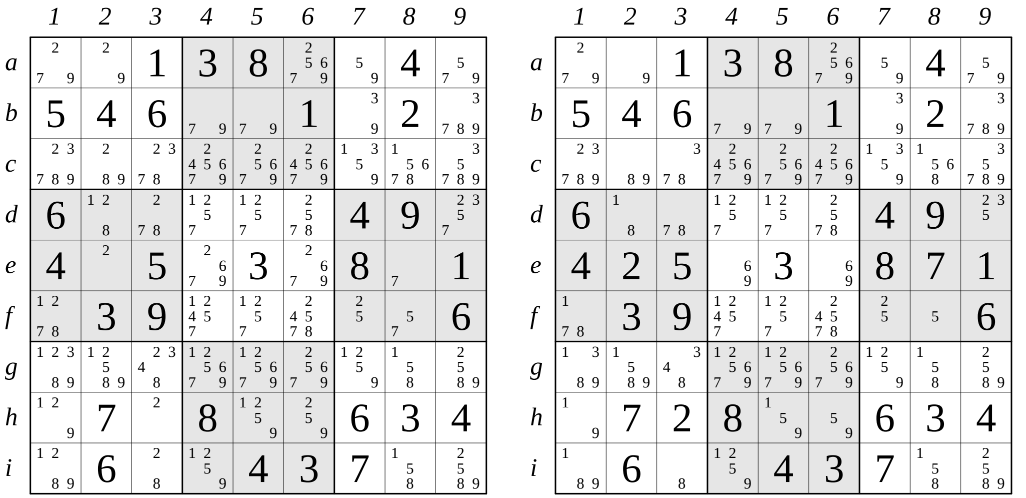

Figure 2: Candidate Elimination and Naked Singles

2.2 Naked Singles

For any given sudoku position, imagine listing all the possible candidates from 1 to 9 in each unfilled square. Next, for every square S whose value is v, erase v as a possible candidate in every square that is a buddy of S. The remaining values in each square are candidates for that square. When this is done, if only a single candidate v remains in square S, we can assign the value v to S. This situation is referred to as a “naked single”.

In the example on the left in figure 2 the larger numbers in the squares represent determined values. All other squares contain a list of possible candidates, where the elimination in the previous paragraph has been performed. In this example, the puzzle contains three naked singles at $e2$ and $h3$ (where a 2 must be inserted), and at $e8$ (where a 7 must be inserted).

Notice that once you have assigned these values to the three squares, other naked singles will appear. For example, as soon as the 2 is inserted at $h3$, you can eliminate the 2’s as candidates in $h3$’s buddies, and when this is done, $i3$ will become a naked single that must be filled with 8. The position on the right side of figure 2 shows the same puzzle after the three squares have been assigned values and the obvious candidates have been eliminated from the buddies of those squares.

2.3 Hidden Singles

Sometimes there are cells whose values are easily assigned, but a simple elimination of candidates as described in the last section does not make it obvious. If you reexamine the situation on the left side of figure 2, there is a hidden single in square $g2$ whose value must be 5. Although at first glance there are five possible candiates for $g2$ (1, 2, 5, 8 and 9), if you look in column 2 it is the unique square that can contain a 5. (The square $g2$ is also a hidden single in the block $ghi123$.) Thus 5 can be placed in square $g2$. The 5 in square $g2$ is “hidden” in the sense that without further examination, it appears that there are 5 possible candidates for that square.

To find hidden singles look in every virtual line for a candidate that appears in only one of the squares making up that virtual line. When that occurs, you’ve found a hidden single, and you can immediately assign that candidate to the square.

To check your understanding, make sure you see why there is another hidden single in square $d9$ in figure 2.

The techniques in this section immediately assign a value to a square. Most puzzles that are ranked "easy" and many that are ranked "intermediate" can be completely solved using only these methods.

The remainder of the methods that we will consider usually do not directly allow you to fill in a square. Instead, they allow you to eliminate candidates from certain squares. When all but one of the candidates have been eliminated, the square's value is determined.

3 Locked Candidates

Locked candidates are forced to be within a certain part of a row, column or block.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
a	1	(4 5)	8	6	7	2	(4 5)	9	3
b	(3 5 7)	(3 5 7)	9	8	1	4	6	(5 7)	2
c	(2 4 7)	6	(2 7)	9	5	3	8	(4 7)	1
d	(3 4 5 9)	(3 4 5)	6	(2 4 5)	(2 3)	7	1	8	(5 9)
e	(4 5)	2	1	(4 5)	9	8	7	3	6
f	(3 4 5 7 9)	8	(3 5 7)	1	(2 3)	6	(2 4 5)	(2 4 5)	(5 9)
g	(2 5)	1	4	3	8	9	(2 5)	6	7
h	6	(3 5 7)	(2 3 5 7)	(2 7)	4	1	9	(2 5)	8
i	8	9	(2 7)	(2 7)	6	5	3	1	4

Figure 3: Locked Candidates

Sometimes you can find a block where the only possible positions for a candidate are in one row or column within that block. Since the block must contain the candidate, the candidate must appear in that row or column within the block. This means that you can eliminate the candidate as a possibility in the intersection of that row or column with other blocks.

A similar situation can occur when a number missing from a row or column can occur only within one of the blocks that intersect that row or column. Thus the candidate must lie on the intersection of the row/column and block and hence cannot be a candidate in any of the other squares that make up the block.

Both of these situations are illustrated in figure 3. The block $def789$ must contain a 2, and the only places this can occur are in squares $f7$ and $f8$: both in row f. Therefore 2 cannot be a candidate in any other squares in row f, including square $f5$ (so $f5$ must contain a 3). Similarly, the 2 in block $ghi456$ must lie in column 4 so 2 cannot be a candidate in any other squares of that column, including $d4$.

Finally, the 5 that must occur in column 9 has to fall within the block $def789$ so 5 cannot be a candidate in any of the other squares in block $def789$, including $f7$ and $f8$.

4 Naked and Hidden Pairs, Triplets, Quads, . . .

These are similar to naked singles, discussed in section 2.2, except that instead of having only one candidate in a cell, you have the same two candidates in two cells (or, in the case of naked triplets, the same three candidates in three cells, et cetera).

A naked pair, triplet or quad must be in the same virtual line. A naked triplet's three values must be the only values that occur in three squares (and similarly, a naked quad's four values must be the only ones occurring in four squares). When this occurs, those n squares must contain all and only those n values, where $n = 1, 2$ or 3. Those values can be eliminated as candidates from any other square in that virtual line.

Figure 4 shows how to use a naked pair. In squares $a2$ and $a8$ the only candidates that appear are a 2 and a 7. That means that 7 must be in one, and 2 in the other. But then the 2 and 7 cannot appear in any of the other squares in that row, so 2 can be eliminated as a candidate in $a3$ and both 2 and 7 can be eliminated as candidates in $a9$.

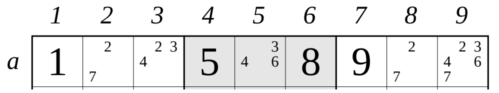

Figure 4: A Naked Pair

For a naked pair, both squares must have exactly the same two candidates, but for naked triplets, quads, et cetera, the only requirement is that the three (or four) values be the *only* values appearing in those squares in some virtual line. For example, if three entries in a row admit the following sets of candidates: $\{1, 3\}$, $\{3, 7\}$ and $\{1, 7\}$ then it is impossible for a 1, 3 or 7 to appear in any other square of that row.

Figure 5 contains a naked triple. In row a squares $a2$, $a8$ and $a9$ contain the naked triple consisting of the numbers 1, 3 and 7. Thus those numbers must appear in those squares in some order. For that reason, 1 and 3 can be eliminated as candidates from squares $a4$ and $a5$.

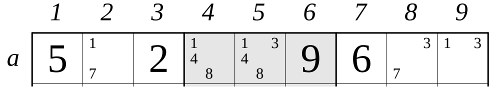

Figure 5: A Naked Triple

Hidden pairs, triples and quads are related to naked pairs, triples and quads in the same way that hidden singles are related to naked singles. In figure 6 consider row i. The only squares in row i in which the values 1, 4 and 8 appear are in squares $i1$, $i5$ and $i6$. Therefore we can eliminate candidates 2 and 6 from square $i1$ and candidate 3 from $i5$.

Figure 6: A Hidden Triple

Remember, of course, that although the three examples above illustrate the naked and hidden sets in a row, these sets can appear in any virtual line: a row, column, or block. There is also no reason that there could not be a naked or hidden quintet, sextet, and so on, especially for versions of sudoku on grids that are larger than 9×9.

There is also the possibility of something called a remote naked pair, but we will discuss that later, in section 10.

5 Tuleja's (or Mr. T's) Theorem

A few months ago Alan Lipp communicated to me a very interesting observation made by his friend, Greg Tuleja, about Sudoku that can sometimes be very useful in solving a puzzle. Since many people have trouble pronouncing "Tuleja", Alan called it "Mr. T's Theorem". I'll call it "Tuleja's Theorem" here.

This is different from all the other strategies in this paper in that the theorem does not usually yield information directly that allows us to fill in a square or to eliminate a candidate, but rather provides an observation that allows us to do so after a small (or sometimes large) number of logical conclusions. In this section we will illustrate and prove Tuleja's Theorem, and will then go through a series of examples where deductions from the theorem allow us to advance our solution. It is hard to know exactly where to place this section in the article, since some applications are trivial and some deep.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
a	1	2	3	4	5	6	7	8	9
b	6	4	5	7	9	8	3	2	1
c	9	8	7	3	1	2	5	4	6
d	8	5	9	6	7	3	4	1	2
e	4	1	6	2	8	5	9	3	7
f	7	3	2	9	4	1	8	6	5
g	5	6	4	8	2	9	1	7	3
h	3	7	1	5	6	4	2	9	8
i	2	9	8	1	3	7	6	5	4

Figure 7: Tuleja's Theorem

Consider the completed Sudoku puzzle illustrated in Figure 7. The theorem applies to any row of blocks or any column of blocks and states that there are only two situations that can occur.

The first is illustrated in the row of blocks $abc123$, $abc345$ and $abc789$. In this case the numbers 1, 2 and 3 occur in row 1 of the first block, row 3 of the second, and row 2 of the third. If this situation occurs where the same three numbers appear in different rows of all three blocks, then all the other sets of three numbers in rows will also repeat. So in this case, we've got 645, 456 and 546 as rows in those blocks, and also 987, 798 and 789. The order of the numbers within the blocks is arbitrary. In the example, the same thing occurs in the row of blocks labeled with g, h and i. If you reflect this puzzle across the main diagonal, it is obvious that the same thing can occur in columns.

If a set of three numbers is repeated in a row of any two blocks it is clear that the same set must occur in the third block in the other row. Then however an additional row is filled in, it forces those same numbers to appear as sets of three in rows in the other blocks. In the example, suppose that the 123 rows are filled in but everything else is blank. When we choose 645 in the $b123$ row, it is clear that those numbers then must go in the $a456$ and thus in the $c789$ rows. That leaves the numbers 9, 8 and 7 to fill the remaining rows.

That's pretty simple, but Tuleja noticed that there is *only one other thing that can happen*. Look at the blocks in the def row. Instead of all three numbers being repeated, there are three pairs of two numbers that repeat: 85, 41 and 73 in this example. It will always happen that if the first situation does not occur, there will be three sets of two numbers that sit in the rows (or columns, obviously) as in the example.

In the column of three blocks under 123, those pairs are 69, 48 and 35. As an exercise, make sure you can find the numbers that do this in the columns of blocks headed by 456 and by 789.

Since the exact numbers don't matter, we'll use nice sets like 123 to make the proof easier to follow. We will also state the proof in terms of rows of blocks, but the same reasoning can be used for columns of blocks.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
a	1	2	3						
b				1	2		3		
c				3			1	2	

Figure 8: Proof of Tuleja's Theorem

Suppose 123 appears in the top row of the first block. If the 123 appears in another row of another block then all three sets of numbers repeat in all the blocks' columns. If not, the 1, 2 and 3 must be split between two rows of the next block. If the split is not three and zero, it has to be two and one, so one of the other rows in the second block contains two of the three numbers, say 1 and 2. Then the 12 pair has to be in the remaining row of the third block. Figure 8 illustrates what might occur in this situation. In the figure, the numbers 1, 2 and 3 are placed arbitrarily in the rows, but once the rows for 1 and 2 are determined, the rows where the 3 must go are fixed as in the figure. There are no vertical divider lines within the blocks to emphasize the fact that the horizonal order of the entries makes no difference in these arguments.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
a	1	2	3						
b				1	2	4	3		
c				3			1	2	5

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
a	1	2	3	5			4		
b	5			1	2	4	3		
c	4			3			1	2	5

Figure 9: Proof of Tuleja's Theorem

Next, we observe that the number completing the 12 set in the other rows cannot be 3, and it cannot be the same in both rows or it would force that same number in the top left row. Let us assume that the numbers filling the rows are 4 and 5 as in the left side of Figure 9. This, of course, forces rows for 4 and 5 in the other blocks as illustrated in the example on the right of the same figure.

From that, it is clear that whatever pair goes into $b23$ also has to go into $a89$, and that, in turn, forces the numbers that go into the remaining slots. This completes the proof of Tuleja's Theorem.

Notice also that the other three numbers, the ones that are not in the repeated pairs, have to cycle through the remaining slots. In this article I will refer to those unmatched numbers as "floaters". In the examples in Figure 9, the floaters are 3, 4 and 5.

5.1 Applications of Tuleja's Theroem

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
a	8			3		6			
b	1					4	7	3	
c		3	9	7	2	1	8	5	
d		8			7		9		
e	7				6		5	8	3
f			5		3			6	7
g	5	4	8	6	1	7	3	9	2
h	3	6	1	2			4	7	5
i						3	6	1	8

Figure 10: Example 1: Tuleja's Theorem

In the examples that follow, in most cases, at least, other standard techniques can be used to fill in numbers or eliminate candidates. The examples simply illustrate how Tuleja's theorem might be used. Generally it is most useful when a fair number of squares have been filled since some data is required to figure out which numbers form the pairs and which numbers are the floaters.

We will begin with a simple example, illustrated in Figure 10.

Consider the second column of blocks. In that column, 73 is a pair and it is obvious that we have only pairs and not triples repeating. The number 6 is obviously a floater (since it is matched with 73 in the center block), making 41 a pair. That means there must be a 4 in either $h5$ or $i5$, but it cannot go in $h5$, so a 4 must go in the circled $i5$. There is a faster way to get this, of course: the 4 in column 6 forces a 4 in column 1 of the central block, forcing a 4 in $h5$ or $i5$.

	1	2	3	4	5	6	7	8	9
a	9		1	2		4		5	
b	4		8				9	1	
c			5	1	9			6	4
d			4	3	6	7			9
e			3	5	8	9	7	4	x
f	7		9	4	1	2	5	3	x
g	5	9			4				
h	3	1					4		5
i	8	4			5	1			3

Figure 11: Example 2: Tuleja's Theorem

Figure 11 illustrates another use of the theorem.

This time we consider the column of blocks under 789. Since we have the numbers 615 in column 8 of the top block and we have a 75 in column 7 of the middle block, we've got sets of pairs and not triples in that column.

One of the pairs must be either 61, 65 or 15 from the 615 set, and those two numbers must go into the squares marked with an "x". Since there is already a 5 in the third block under column 9, that eliminates 65 and 15, so the two "x" slots contain a 1 and a 6. Thus a 6 goes in the circled square and a 1 in the other.

When you first try to apply Tuleja's theorem, you will probably find that it is easy to make mistakes. A good way to practice is to do easy Sudoku puzzles on the internet or other place where you have easy access to the answers and as you work them, search for applications of Tuleja's theorem before you try anything else. When you think you've got one, check the answer to make sure your reasoning was correct. When your answer is wrong, review your reasoning to learn where you made your mistake.

	1	2	3	4	5	6	7	8	9
a			5	8	4	6	7	9	
b				7	5	9			
c		9	7	2	3	1			5
d	1		4	6		5	3		
e				3		4			
f			3	1	2	7	8		9
g	7			5	6	3	9	2	4
h				4	1	8			
i		4	6	9	7	2	5		

	1	2	3	4	5	6	7	8	9
a			5	8	4	6	7	9	
b				7	5	9			
c		9	7	2	3	1			5
d	1		4	6		5	3		
e				3	x	4			
f			3	1	2	7	8		9
g	7			5	6	3	9	2	4
h				4	1	8			
i	3	4	6	9	7	2	5		

Figure 12: Examples 3 and 4: Tuleja's Theorem

The next two examples come from the same Sudoku puzzle. In the puzzle on the left in Figure 12 consider the lower row of blocks. 92 is a pair, so 4 and 7 are floaters. In the middle lower block we have 563, so 56, 53 or 63 must be a pair. But the 4 in the lower left block tells us that 6 is part of the pair, so it is either 56 or 63. Since there is a 5 in row i, the pair must be 63, so a 3 belongs in the circled position $i1$.

After placing the 3 in $i1$, we now have the configuration shown on the right in Figure 12. We are missing 8 and 9 in the center block, so position x must contain an 8 or a 9. In the center row of blocks, 89 cannot be a pair since 8 and 9 go in two rows in the center block. In the center row of the center block, we can conclude that the pair is one of: 38, 39, 48 or 49. It can't be 38 or 39 because of the right block in the central row of blocks. It also can't be 49 since there is a 9 in position $c2$. Thus there is an 8 in the circled $d2$ square.

The final example displayed in Figure 13. Consider the top row of blocks.

Since there is a 2 in column 9 there must be a 2 in one of the squares marked with an x. Therefore one of the squares marked y must also contain a 2. Because we then have 21 in two block rows, 21 must be a pair, and 7 a floater. (Clearly there are only pairs and not triples in the row since we have a 721 and a 73 in different block rows.)

Since 7 is a floater and appears with a 3 in the rightmost block, 3 is not a floater. The 439 in the leftmost block tells us that the pair is either 34 or 39. Since 3 appears with 6 in the top row of the middle block, 6 is a floater, so 85 is a pair and the only place the 5 can go is in the circled square, 12.

(Also note that since 34 or 39 is a pair, and there is a 4 elsewhere in the middle block, that $a5$ must be 9.)

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
a	8			3		6	x	x	
b	1	y	y			4	7	3	
c	4	3	9	7	2	1	8	5	6
d		8			7	5	9		
e	7				6	2	5	8	3
f			5		3			6	7
g	5	4	8	6	1	7	3	9	2
h	3	6	1	2			4	7	5
i				5	4	3	6	1	8

Figure 13: Example 5: Tuleja's Theorem

6 X-Wings and Swordfish

An X-wing configuration occurs when the same candidate occurs exactly twice in two rows and in the same columns of those two rows. (Or you can swap the words "rows" and "columns" in the previous sentence.) In the configuration on the left in figure 14 the candidate 3 occurs exactly twice in rows c and h and in those two rows, it appears in columns 2 and 7. It does not matter that the candidate 3 occurs in other places in the puzzle.

The squares where the X-wing candidate (3, in this case) can go form a rectangle, so a pair of opposite corners of that rectangle must contain the candidate. In the example, this means that the 3's are either in both $c2$ and $h7$ or they are in both $c7$ and $h2$. Perhaps the fact that connecting the possible pairs would form an 'X', like the X-wing fighters in the Star Wars movies gives this strategy its name.

In any case, since one pair of two corners must both contain the candidate, no other squares in the columns or rows that contain the corners of the rectangle can contain that candidate. In the example, we can thus conclude that 3 cannot be a candidate in squares $a7$, $f7$ or $i2$.

A swordfish is like an X-wing except that there must be three rows with the three candidates appearing in only three columns. As was the case with naked and hidden triples, for a swordfish there is no requirement that the candidate be in all three positions. The candidate must occur three times, once in each row, but since the occurrences in those rows are restricted to exactly three columns, all the

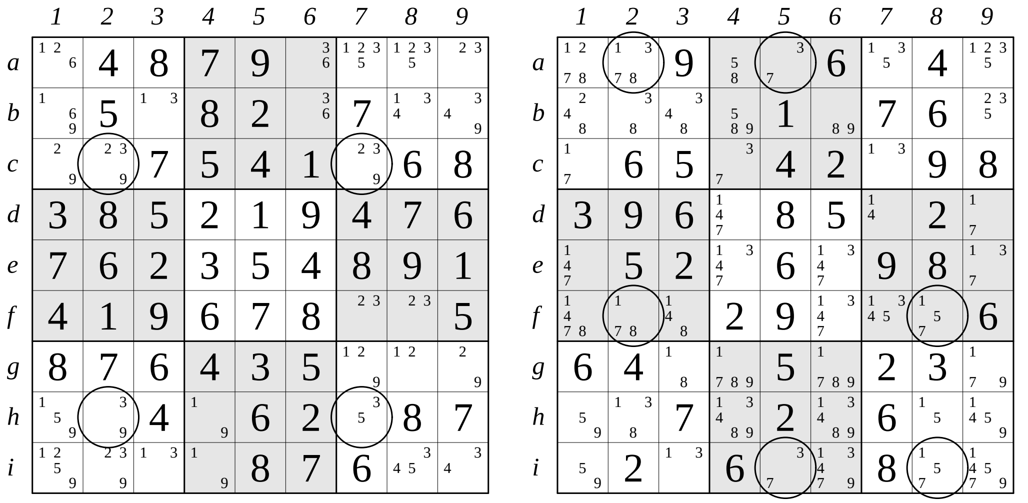

Figure 14: X-Wing (left) and Swordfish

columns must be used as well. The reasoning is similar to that used for the x-wing: once you find a swordfish configuration, the candidate cannot appear in any other squares of the three columns and rows. Of course you can again swap the words “rows” and “columns” in the description above.

A sample swordfish configuration appears on the right in figure 14. In this case, the candidate is 7, and the columns that form the swordfish are 2, 5 and 8. In these columns the value 7 appears only in rows a, f and i. One 7 must appear in each of these rows and each of the columns, so no other squares in those rows and columns can contain a 7. Thus the candidate 7 can be eliminated from squares $a1$, $f1$, $f6$, $i6$ and $i9$.

Of course there is nothing special about a swordfish configuration; “super-swordfish” with 4, 5, or 6 candidates might be possible. They are rare but not particularly difficult to spot. The “super-swordfish” with 4 rows and 4 columns is sometimes called a “jellyfish”. If you are playing on a standard 9×9 grid, the most complex situation you would need to look for would be a jellyfish, since if there were a 5×5 super-swordfish, there would have to be in addition a 4×4 or smaller swordfish in the remaining rows or columns. It’s too bad that there’s no real need for the 5×5 super-swordfish, since the usual name in the internet literature is so nice: it is called a “squirmbag”.

But let us see why this is so, for a particular situation. It will be clear how the argument can be extended to others.

Suppose that the candidate 1 has been assigned to two squares. Then there are 7 rows and columns in which a 1 has *not* been assigned. If we find a (4-row) jellyfish, we would like to show that there must be a 3-column (or simpler) swordfish. Assume that in each of rows w, x, y and z the number 1 is a possible candidate only in columns α, β, γ and δ. It may not be a candidate in all those columns, but in those four rows, it will never be outside those columns.

But that means that in the *other* three columns, the candidate 1 will be missing from the rows w, x, y and z, so it must appear only in the other three rows. That means there is at most a (3-column) swordfish.

Obviously there is nothing special about the 7, 4 and 3 in the argument above. If there are n available rows and columns, and you find a k-row “swordfish”, there must be an $(n-k)$-column “swordfish”.

7 The XY-Wing and XYZ-Wing

Sometimes a square has exactly two candidates and we are logically led to the same conclusion no matter which of the two we assume to be the correct one. An "XY-wing" represents such a situation. This is a sort of "guess and check" strategy, but it only looks ahead one step so it is easily done by a human.

In the configuration in figure 15, suppose that there are two possible candidates in squares $b2$, $b5$ and $e2$. In the figure, the candidates are just called X, Y and Z. Consider the contents of square $b2$. If X belongs in $b2$, then there must be a Z in $e2$ and therefore Z cannot be a candidate in $e5$. But the other possibility is that $b2$ contains a Y. In this case, $b5$ must be Z and again, $e5$ cannot be Z. Thus no matter which of the two values goes in $b2$, we can deduce that Z is impossible in square $e5$.

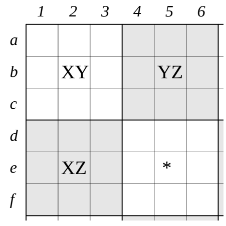

Figure 15: XY-Wing

Now consider the configuration on the left in figure 16. If either X or Y belongs in square $a2$, Z cannot be a candidate in the three squares indicated by asterisks. Similarly, in the configuration in the center in the same figure, Z cannot be a candidate in two more squares indicated by asterisks.

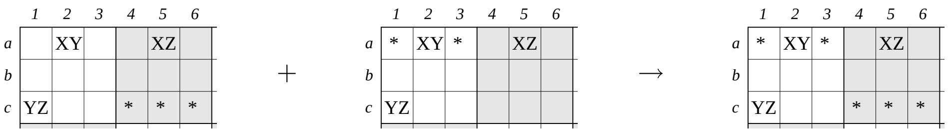

Figure 16: XY-Wing

Obviously, the two configurations on the left in figure 16 can be combined to make the configuration on the right in the same figure where Z can be eliminated as a candidate in any of the squares marked with an asterisk.

An example of an XY-wing in an actual puzzle appears in figure 18. Notice that in squares $d8$ and $f7$ (both in the same block, $def789$) and in square $d1$ we have candidates $\{8, 9\}$, $\{3, 9\}$ and $\{8, 3\}$, respectively. No matter which of the two values appear in $d8$, a 3 must appear in either $d1$ or $f7$. Because of this, we can eliminate 3 as a candidate from squares $d7$, $f1$ and $f2$.

The XYZ-wing is a slight variation on the XY-wing. If you can find a square that contains exactly the candidates X, Y and Z and it has two buddies, one of which has only candidates X and Z while the other has only candidates Y and Z, then any square that is buddies with *all three* of those squares cannot admit the candidate Z. On the left side of figure 17 this situation occurs, and candidate Z can be eliminated from squares $b2$ and $b3$.

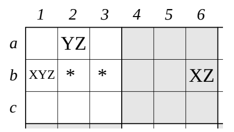

Figure 17: XYZ-Wing

8 XY-Chains

There is another way to look at the XY-wing. We can think of the example in figure 15 as a sort of chain from $e2$ to $b2$ to $b5$. If the value in $e2$ is not Z, then it is X. Since $b2$ is a buddy of $e2$ this would force $b2$ to be Y, and since $b5$ is a buddy of $b2$, then $b5$ would be forced to be Z. The reasoning can be reversed if we assume $b5$ is not Z and we can conclude that $b2$ *must* be Z. Thus exactly one of $b2$ or $b5$ must be Z and the other is not. Any squares that are buddies of both $b2$ and $b5$ (only $e5$ in the example in figure 15) cannot possibly be Z.

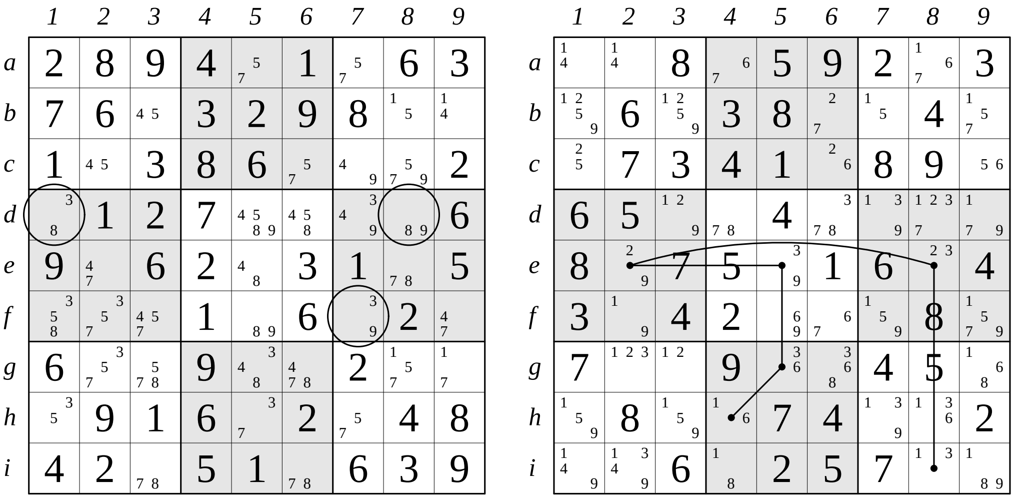

Figure 18: XY-Wing Example (left), XY-Chain (right)

The interesting observation we can make from this is that there is no need for such a chain to be only two steps long: it can be as long as we want, as long as the same candidate appears at both ends. When this occurs, we can eliminate that candidate from any of the squares that are buddies of both squares that are the endpoints of such a chain. We will call these XY-chains.

Consider the situation on the right in figure 18. Look at the following chain of squares linked in exactly the same way that the three squares in an XY-wing are linked: $i8 - e8 - e2 - e5 - g5 - h4$. Each of the squares is a buddy of the next; each square contains only two possible candidates, and finally, those two candidates match with one of the two candidates of the squares on either side of it in the chain. Finally, the left-over candidate (1 in this example) is the same in squares $i8$ and $h4$. By stepping through the chain we can conclude that if $i8$ is not 1 then $h4$ is, and if $h4$ is not 1 then $i4$ is. Thus either $i8$ or $h4$ *must* be 1, so squares that are buddies of both $i8$ and $h4$ cannot be 1 and we can eliminate 1 as a candidate from squares $h7$, $h8$ and $i4$.

We can also note that naked pairs (discussed in section 4) are a simpler version of this idea, but while an XY-wing is an XY-chain with three links, a naked pair is a chain with only two links. But the naked pair that contains a 1 and a 2 in each square is like two of these chains: one with 1 at the endpoints and one with 2 at the endpoints, so all 1's and 2's can be eliminated from squares that are buddies of the two that make up the square. Similar reasoning can be applied to see that a naked triple like 12, 23 and 31 can be thought of as three different XY-chains where each pair is a different set of endpoints.

9 Coloring and Multi-Coloring

Coloring and multi-coloring are techniques that eliminate candidates based on logical chains of deduction. The coloring method, especially, is simple enough that it can be done by hand.

9.1 Simple Coloring

Consider the example in figure 19 where we consider a few squares that admit the candidate 1. Let's assume for now that these are the only possible locations for the candidate 1 in the puzzle. Certain virtual lines contain exactly two places where the candidate 1 can go: rows b and i, columns 3 and 6, and block $def123$. In each of these virtual lines, exactly one of the possible squares can contain a 1 and once it is selected, the other cannot.

But this creates a sort of "chain" if $f1$ contains 1, then $e3$ must not, and since $e3$ must not, $b3$ must, so $b6$ must not, $i6$ must, and $i9$ must not. If, on the other hand, $f1$ does not contain a one, the same series of virtual line interactions will force an alternating set of conclusions and every square in the chain will be forced to have the opposite value.

In the figure we've marked the squares with $+$ and $-$ according to the assumption that $f1$ does contain a 1, but of course it may be the case that $f1$ does not contain 1, and all the $+$ and $-$ signs would be interchanged. Rather than using the "$+$" and "$-$" characters that could imply presence or absence of a value it is better simply to imagine coloring each square in the chain black or white, and either all the black squares have a 1 and all the white squares do not, or the opposite.

In most situations, not all of the squares in a puzzle that admit a candidate can be colored: only those squares where the candidate appears exactly twice in some virtual line can be part of a chain. If there are three candidates in a row, for example, and one of them is colored, we cannot immediately assign colors to the two others in that row although we may be able to do so later, based on other links in the chain.

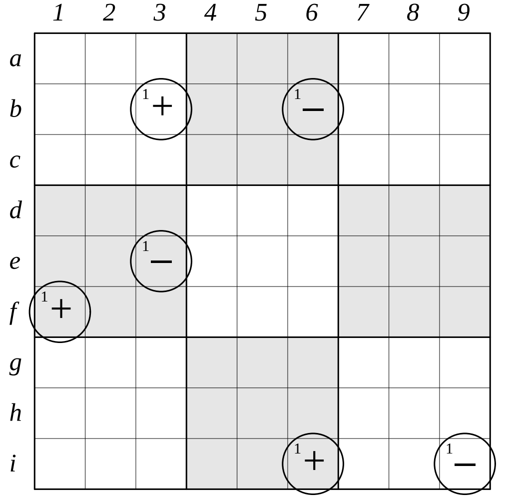

Figure 19: Simple Coloring

Suppose now that for some candidate you have discovered such a chain[1] and have colored it in this alternating manner.

It may be that there are additional squares where the candidate could possibly occur that do not happen to lie in the colored chain. In figure 19, suppose square $f1$ is colored black and so square $i9$ must be colored white. Consider the square $f9$ that lies at the intersection of $f1$'s row and $i9$'s column. Since $f1$ and $i9$ have opposite colors, exactly one of them *will* contain a 1, and therefore it is impossible for the square $f9$ to contain a 1, so 1 can be eliminated as a possible candidate in square $f9$.

This is probably easier to see with the concrete example displayed on the left in figure 20 where we consider 1 as a possible candidate. In row d, $d1$ and $d5$ are the only occurrences of candidate 1, so we color $d5$ black and $d1$ white. But $d1$ and $f3$ are the only possibilities for 1 in block $def123$, so since $d1$ is white, $f3$ is black. By similar reasoning, since $f3$ is black, $g3$ and $f8$ are white. Since $f8$ is white, $e7$ is black, and since $e7$ is black, $c7$ is white. That's a pretty complicated chain, but here's what we've got: black: $\{d5, f3, e7\}$ white: $\{d1, g3, f8, c7\}$. A grid that displays just the colored squares appears on the right in figure 20.

[1] For astute readers, it may not really be a chain, but it could be a tree, or even have loops, as long as the black/white

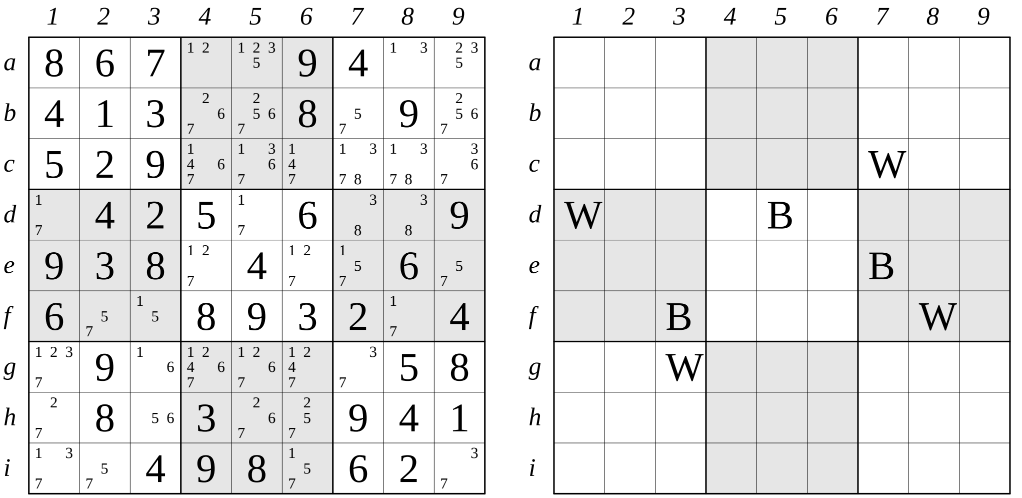

Figure 20: Simple Coloring

Square $c5$ is at the intersection of $c7$'s row and $d5$'s column, but $c7$ is white and $d5$ is black, so 1 cannot be a candidate in square $c5$. Similarly, square $g5$ is in the same row as $g3$ and same column as $d5$ which are white and black, respectively, so 1 also cannot be a candidate in $g5$.

Note that after making eliminations like this, it may be possible to extend the coloring to additional squares although that is not the case in figure 20.

There is nothing special about a row-column intersection. Any time two oppositely-colored squares "intersect" via virtual lines of any sort in another square, the candidate can be eliminated as a possibility in that square. Sometimes a candidate can be assigned immediately to a square on the basis of coloring. Suppose that two squares of a chain are the same color, but lie in the same virtual line. If squares of that color contained the candidate, then two squares in the same virtual line would contain it which is impossible, so the candidate can be immediately assigned to all squares of the other color. This situation is shown in figure 21. In that figure, the board on the left is colored for candidate 8 as shown on the right. Note that the black squares conflict in a few places: in column 1, row c, and in block $abc123$. This means that the candidate 8 can be eliminated from every square colored black.

9.2 Multi-coloring

Sometimes a position can be colored for a particular candidate and multiple coloring chains exist, but none of them are usable to eliminate that candidate from other squares. If there are multiple chains, it is worth looking for a multi-coloring situation.

Consider the puzzle in figure 22. Assume that in the parts of the puzzle that are not shown there are no other places that the candidate 1 can occur. When this diagram is colored, there are two coloring chains. Instead of using words like "black" and "white" we will used letters, like A, B, C, a, b and c where the A and a represent opposite colors, as do the B and b, C and c and so on. In figure 22 rows a and c and in column 3 there are only two possible locations for candidate 1.

alternation is preserved.

	1	2	3	4	5	6	7	8	9
a	1	5	4 8	9	4 8	3	7	6	2
b	9	3 8	2	7	1	6	3 8	5	4
c	3 4 8	7	6	5	4 8	2	3 8	1	9
d	3 6	1	5	8	9	4	2	7	3 6
e	2	3 8	9	6	7	5	4	3 8	1
f	6 8	4	7	3	2	1	5	9	6 8
g	7	6	3 8	4	5	9	1	2	3 8
h	5	2	3 4	1	6	8	9	3 4	7
i	4 8	9	1	2	3	7	6	4 8	5

	1	2	3	4	5	6	7	8	9
a			B		W				
b		B					W		
c					B		B		
d									
e		W						B	
f	B								W
g			W						B
h									
i	B							W	

Figure 21: Simple Coloring: Chain Conflict

When this grid is colored, it will look something like this: squares $c1$ and $f3$ have color A and square $c3$ has color a. Square $a2$ has color b and square $a5$ has color B. (Note that the colors assigned are arbitrary. All that matters is that squares $c1$ and $f3$ have the same color that is the opposite of $c3$ and that $a2$ and $a5$ have opposite colors that are different from the other assigned colors. Note that none of the other squares with 1 as a candidate can be colored, since all are in virtual lines with more than two squares that potentially could contain the candidate 1.

	1	2	3	4	5	6
a		1 b			1 B	
b						
c	1 A		1 a			
d						
e	1	1			1	
f	1	1	1 A		1	1

Figure 22: Multi-coloring

If we consider the color "a" as standing for the sentence: "Every square containing the color a contains a 1," and so on, then we can write little logical expressions indicating the relationships among the various colors when they are interpreted as sentences. The obvious ones are of the form: "a $=\neg$ A" or "A $=\neg$ a" (where the logical symbol "$\neg$" means "not" and the symbol "$=$" means "is logically equivalent to"). In other words, if a is true then A is false, and vice-versa.

In this section, we will be performing what is known as boolean algebra[2] on expressions involving "sentences" such as a, A, b, B and so on.

Although the values of non-opposite colors do not necessarily have anything to do with each other, in figure 22, the pair a and b, for example, are linked, since they occur in the same block. If a is true, then b cannot be, and vice-versa, but it may be true that both a and b are false. We will express this relationship as "a$\barwedge$b" and read it as "a excludes b". Obviously, if a$\barwedge$b then b$\barwedge$a[3]. Also, it is obvious in the configuration in figure 22 that b$\barwedge$A.

Another way to think of a$\barwedge$b is as "If a is true then so is B." and at the same time, "If b is true then so is A." If a$\barwedge$b then at least one of a or b must be false. Equivalently, if a$\barwedge$b os true then at least

[2]See section 19 for a text on boolean algebra.

[3]The truth table for a$\barwedge$b is equivalent to the "nand" ("A nand B" is the same as "not(A and B)") logical operator that is heavily used in computer hardware logic designs and is sometimes represented by the symbol "$\barwedge$".

one of A or B must be true. This means that any square that is a buddy of two squares colored A and B must not allow the candidate since one of the two squares colored A or B *must* contain the candidate. In figure 22, this means that 1 cannot be a candidate in square $f5$.

To condense all of the above into a single statement, we know that if a$\overline{\wedge}$b for some candidate then any square that is buddies of both A and B cannot contain that candidate.

Let us begin with a simple example of multicoloring displayed in figure 23. On the left is the complete situation, and on the right is a simplified version where only squares having the number 6 as a possible candidate are displayed.

	1	2	3	4	5	6	7	8	9
a	3 6	1 3 4 6	1 3	9	2	5	6 7	8	4 7
b	8	5	9	1 6	7	4	2	1 6	3
c	7	4 6	2	1 3 6	1 3	8	9	1 4 6	5
d	4	7	6	8	5	3	1	9	2
e	2	9	1 3	1 4	6	7	3 4	5	8
f	5	1 3	8	2	1 4	9	3 6 7	4 6 7	4 6 7
g	3 6 9	8	4	3 7	3 9	2	5	6 7	1
h	3 6 9	3 6	7	5	3 4 9	1	8	2	4 6
i	1	2	5	4 7	8	6	4 7	3	9

	1	2	3	4	5	6	7	8	9
a	6	6					B6		
b				C6				c6	
c		6		c6				6	
d									
e									
f							b6	6	a6
g	A6							a6	
h	6	6							A6
i									

Figure 23: Simple Multi-coloring Example: Coloring on Right

In row b and column 4 there are only two squares that admit candidate 6, so we have colored all those squares with C and c. In the same way, the two squares in column 6 are colored with B and b, while A and a are used to color four squares that share, in pairs, row g, column 9 and block $ghi789$.

In this example, we note that a$\overline{\wedge}$b because instances of them lie in squares $f7$ and $f9$ which are buddies. Because of this, any square that is a simultaneous buddy of a square colored A and of one colored B cannot allow 6 as a candidate. In the figure, square $a1$ is buddies of both $g1$ and $a7$, so 6 cannot be a candidate in square $a1$, so we can see in full puzzle on the left that 3 can be assigned to square $a1$.

In the figure, B$\overline{\wedge}$c (since they lie in the block $abc789$) and c$\overline{\wedge}$a (since they lie in column a) as well, but there are no squares that are simultaneous buddies of squares colored b and C or of C and A so we cannot use those exclusion relationships to help solve this puzzle.

Next, we will look at a multicoloring situation that is quite complex because much more can be done with multicoloring. In complex situations, there may be many independent color chains with colors A and a, B and b, C and c, and so on. When that occurs, we need to look for consequences of the following inference:

If a$\overline{\wedge}$b and B$\overline{\wedge}$c then a$\overline{\wedge}$c.

It's not hard to see why: If a is true, b is not, so B is true, and the second exclusion implies that c is not. The reasoning is trivially reversed to show that if c is true then a is not, so we obtain a$\overline{\wedge}$c.

Thus to do multi-coloring for a particular candidate, proceed as follows:

- Construct all possible color chains for the diagram.
- Find all exclusionary relationships from pairs of colored squares that are buddies.
- Take the collection of relationships and complete it to its transitive closure using the idea that if ($a \overline{\wedge} b$ and $B \overline{\wedge} c$) then $a \overline{\wedge} c$.
- For every exclusionary pair in the transitive closure, find buddies of squares colored with colors opposite to those in the pair, and eliminate the candidate as a possibilty from all of them.

	1	2	3	4	5	6	7	8	9
a	4	8	$_9$	$_9$	6	$_9$	2	7	5
b	2	5			7		9		6
c	$_9$	7	6		$_9$	$_9$	4	3	
d	5	2	$_9$	8	4	$_9$	6	$_9$	3
e	$_9$	6	8	3	5	2	7	4	$_9$
f	3	$_9$	$_9$	$_9$	1	6	8	5	2
g	8	3	5		$_9$	$_9$	1	6	7
h	6	1	$_9$		8		3	$_9$	4
i	7	$_9$	$_9$	6	3	1	5	$_9$	$_9$

	1	2	3	4	5	6	7	8	9
a			A_9	E_9		$_9$			
b									
c	a_9				b_9	$_9$			
d			$_9$			$_9$		A_9	
e	A_9								a_9
f		D_9	$_9$	e_9					
g					B_9	b_9			
h			C_9					c_9	
i		d_9	$_9$					$_9$	A_9

Figure 24: Complex Multi-coloring Example: Coloring on Right

Let's look at a very complex multi-coloring application. See figure 24 where only the presence of squares that admit the candidate 9 are marked (all, of course, must admit other candidates). On the left is the complete grid and on the right is a simplified version where only the squares admitting candidate 9 are shown, and all of the color chains are displayed. It is an excellent exercise to look at the diagram on the right to make certain that you understand exactly how all the color chains are constructed.

The next step in the application of multi-color is to find all the exclusionary pairs, and the initial list is displayed in table 1. Note that the "$\overline{\wedge}$" operation is commutative, so if you think $a \overline{\wedge} b$ should be in the list and it is not, be sure to look for $b \overline{\wedge} a$ as well.

Table 1: Direct Exclusions

$A \overline{\wedge} E$	$a \overline{\wedge} b$	$D \overline{\wedge} e$	$A \overline{\wedge} d$	$A \overline{\wedge} C$
$A \overline{\wedge} c$	$b \overline{\wedge} E$	$A \overline{\wedge} D$	$C \overline{\wedge} d$	

From these initial exclusions, a number of others can be deduced. For example, from $a \overline{\wedge} b$ and $A \overline{\wedge} d$ we can conclude that $b \overline{\wedge} d$. Note that to make this implication, we are implicitly using the fact that $a \overline{\wedge} b$ and $b \overline{\wedge} a$ are equivalent. Similarly, since $A \overline{\wedge} E$ and $D \overline{\wedge} e$ we can conlude that $A \overline{\wedge} D$. See if you can discover others before reading on.

In fact, if we make all such deductions, and then all deductions from those, and so on, there are twelve additional exclusions that we find, and they are displayed in table 2

Table 2: Derived Exclusions

b$\overline{\wedge}$d	b$\overline{\wedge}$C	b$\overline{\wedge}$c	b$\overline{\wedge}$D	C$\overline{\wedge}$e
A$\overline{\wedge}$A	A$\overline{\wedge}$b	b$\overline{\wedge}$e	b$\overline{\wedge}$b	A$\overline{\wedge}$e
E$\overline{\wedge}$b	A$\overline{\wedge}$D			

For most of them, we need to look for generalized intersections of the opposites of the exclusionary values. For example, since A$\overline{\wedge}$e and there is an a in $c1$ and a E in $a4$, then 9 cannot be a candidate in squares $c5$ or $c6$. Also, since we've got A$\overline{\wedge}$A and b$\overline{\wedge}$b we can conclude that a and B are true since if it is impossible for a statement and itself to be true, the statement must be false.

10 Remote Naked Pairs

This technique is related both to naked pairs, simple coloring, and XY-chains. Sometimes there will be a series of squares with the same two candidates, and only those two candidates that form a chain in the same way that we considered chains of single candidates in section 9.

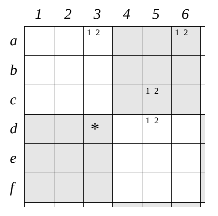

Figure 25: Remote Naked Pairs

Consider figure 25. In it, we have naked pairs $a3$ and $a6$, $a6$ and $c5$, and $c5$ and $d5$, where each pair shares a row, column or block. By an alternating coloring argument, it should be clear that $a3$ and $d5$ are effectively a naked pair: one of the two must contain the candidate 1 and the other, 2. Thus the square marked with an asterisk, $d3$ cannot contain either a 1 or a 2, and they can be eliminated as possible candidates for square $d3$.

11 Unique Solution Constraints

If you know that the puzzle has a unique solution, which any reasonable puzzle should, sometimes that information can eliminate some candidates. For example, let's examine the example in figure 26.

In row c, columns 4 and 6, the only possible candidates are 1 and 2. But in row g, columns 4 and 6, the candidates are 1, 2 and 8. We claim that 8 must appear in $g4$ or $g6$. If it does not, then the four corners of the square $c4$, $c6$, $g4$ and $g6$ will all have exactly the same two candidates, 1 and 2, so we could assign the value 1 to either pair of opposite corners, and both must yield valid solutions. If there is a unique solution, this cannot occur, so one of $g4$ or $g6$ must contain the value 8. But if that's the case, square $i4$ cannot be 8, so the candidate 8 can be eliminated from square $i4$. In addition, since either $g4$ or $g6$ must be 8, $g8$ cannot be 8 since it is in the same row as the other two.

In the same figure, a similar situation appears in another place. See if you can find it. Hint: it column-oriented instead of row-oriented.

Let's go back and see exactly what is going on, and from that, we'll be able to find a number of techniques that are based on the same general idea. Figure 27 shows a basic illegal block. Anything at all can occur in the squares that are not circled, but note that an assignment of a 2 or a 7 to any of the circled squares forces the values of the others in an alternating pattern. But any of the squares can be assigned a 2 or a 7 and the resulting pattern will be legal, and this means there are two valid solutions to the puzzle.

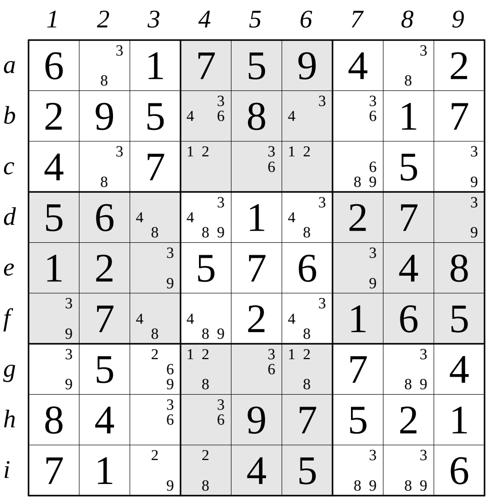

Figure 26: Uniqueness Constraint

This means that if some assignment causes an illegal block to be formed, that assignment is impossible, and we can use that fact to eliminate certain possibilities, as we did in the example in figure 26. Note that the four corners must not only form a rectangle, but they must be arranged so that two pairs of adjacent corners must lie within the same blocks. If the four corners lie in four different blocks, then constraints from those different blocks can force the values one way or the other.

Now let us examine some variations of this theme. In the rest of the examples in this section, we'll assume nothing about the empty squares. They may have values assigned to them or may be empty. On the left in figure 28 we see something that is almost the same as what we saw in figure 27 and the only thing that makes it legal is the presence of the possibility of a 3 in square $b1$. If it is not a 3, then we would have the illegal block, so there must be a 3 in square $b1$. Note that if, in the figure, square $b1$ had contained the possibilities $2, 3, 4$, and 7, at least the two possibilities 2 and 7 could still be eliminated as candidates, so only a 3 or a 4 could be entered in that square.

The example in the middle of figure 28 is similar to the original example in this section except that the additional number occurs in two different blocks instead of one. As before, at least one of those squares must contain the number (3 in this case), so the value 3 can be eliminated from any of the other squares in that row (row b, in this case), but not in either of the blocks, since the one that is forced to be 3 might be in the other block.

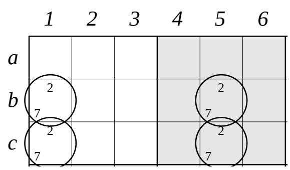

Figure 27: An illegal block

The example on the right in figure 28 illustrates another sort of deduction that could be made. We know that at least one of $b1$ and $c1$ must contain a number other than a 2 or a 7, but we don't know which one. If we think of the combination of the two squares as a sort of unit, we *do* know that this unit will contain either a 3 or a 4. This two-square unit, together with square $a3$ (which has 3 and 4 as its unique possibilities) means that no other square in the block $abc123$ can contain a 3 or a 4. If the 34 square had been in $a1$ we could in addition eliminate 3 and 4 as candidates from any of the other squares in column 1 outside the first block.

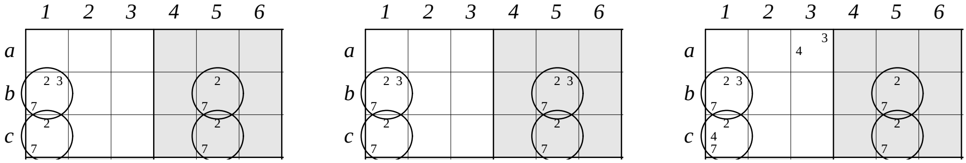

Figure 28: Uniqueness Considerations

Note that we can have both a 3 and a 4 in either or both squares $b1$ and $c1$ in this example on the right. As long as both occur, the argument holds. Also note that if the 3's and 4's appeared in row b and the entries in row c were both 27, and the 34 square were in row b we could eliminate any more 3's and 4's in that row.

12 Forcing Chains

This method is almost like guessing, but it is a form of guessing that is not too hard for a human to do. There are various types of forcing chains, but the easiest to understand works only with cells that contain two candidates.

The idea is this: for each of the two-candidate cells, tentatively set the value of that cell to the first value and see if that forces any other two-candidate cells to take on a value. If so, find additional two-candidate cells whose values are forced and so on until there are no more forcing moves. Then repeat the same operation assuming that the original cell had the other value.

If, after making all possible forced moves with one assumption and with the other, there exists a cell that is forced to the same value, no matter what, then that must be the value for that cell.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
a	1 2 6	4	8	7	9	3 6	1 2 3 5	1 2 3 5	2 3
b	1 6 9	5	1 3	8	2	3 6	7	1 3 4	3 4 9
c	2 9	2 3 9	7	5	4	1	2 3 9	6	8
d	3	8	5	2	1	9	4	7	6
e	7	6	2	3	5	4	8	9	1
f	4	1	9	6	7	8	2 3	2 3	5
g	8	7	6	4	3	5	1 2 9	1 2	2 9
h	1 5 9	3 9	4	1 9	6	2	3 5	8	7
i	1 2 5 9	2 3 9	1 3	1 9	8	7	6	3 4 5	3 4

Figure 29: Forcing Chains

As an example, consider the example in figure 29, and let's begin with cell $b3$ which can contain either a 1 or a 3. If $b3 = 1$, then $i3 = 3$, so $h2 = 9$, so $h4 = 1$. On the other hand, if $b3 = 3$ then $i3 = 1$ so $i4 = 9$ so $h4 = 1$. In other words, it doesn't matter which value we assume that $b3$ takes; either assumption leads to the conclusion that $h4 = 1$, so we can go ahead and assign 1 to cell $h4$.

Note that XY-wings that we considered in section 7 are basically very short forcing chains.

13 Guessing

The methods above will solve almost every sudoku puzzle that you will find in newspapers, and in fact, you will probably hardly ever need to use anything as complex as multi-coloring to solve such puzzles. But there do exist puzzles that do have a unique solution, but cannot be solved using all the methods above.

One method that will always work, although from time to time it needs to be applied recursively, is simply making a guess and examining the consequences of the guess. In a situation that seems impossible, choose a square that has more than one possible candidate, remember the situation, make a guess at the value for that square and solve the resulting puzzle. If you can solve it, great—you're done. If that puzzle cannot be solved, then the guess you made must be incorrect, it can be eliminated as a candidate for that square, and you can return to the saved puzzle and try to solve it with one candidate eliminated.

Obviously, when you try to solve the puzzle after having made a guess, you may arrive at another situation where another guess is required, in which case a second level of guess must be made, and so on. But since the method always eventually eliminates candidates, you must arrive at the solution, if there is one. In computer science, this technique is known as a recursive search. Figure 30 is an example of such a puzzle that cannot be cracked with any of the methods discussed so far except for guessing. The solution to this puzzle can be found in section 20.

Guessing is a direct logical approach in the sense that we assume something is true, such as "square $i3$ is a 3" and then we follow the consequences of that to see if it results in a contradiction. A purely formal approach can also be taken. For every one of the 81 squares, there is a set of 18 statements about the square that can be either true or false. These have the form "square x has the value i" or "square x does not have the value i", where i is a number between 1 and 9. There are $18 \times 81 = 1458$ of these assertions.

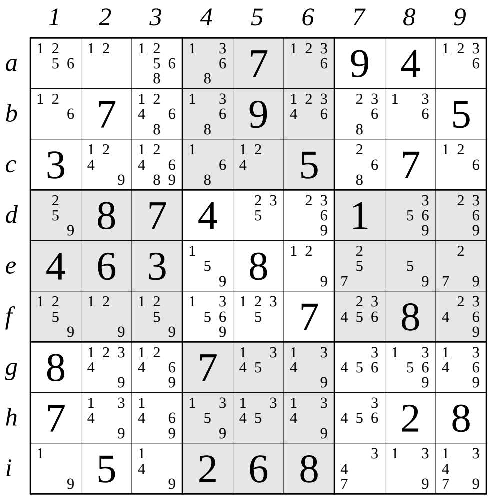

Figure 30: A Very Hard Sudoku Puzzle

From the initial configuration, we know some are true and some are false, and it is possible to make logical deductions from them. If we know that square x *is* 4 then we know that if $i \neq 4$ the statements "x is i" are false, and the statements "x is not i" are all true except when $i = 4$. Similarly, we can make logical conclusions about the buddies of x if we know that square x has a certain value. If all the buddies of x cannot have a value, then x must have that value, and there are a few other similar rules of inference we can use to assign truth values to the 1458 propositions. In principle, a person can search the list over and over and see if any of these rules of inference can be applied, and if so, apply them to assign yet more truth values. Repeating this over and over will almost always succeed in solving the puzzle.

The method described above does not involve guessing, and works directly forward using only logical consequences, but it is not a reasonable way for a human to solve the puzzle. Computer programs are great at this sort of analysis, but they may need to apply thousands of such inference rules to take each step forward.

14 Equivalent Puzzles

There is no reason that the numbers 1 through 9 need to be used for a sudoku problem. We never do any arithmetic with them: they simply represent 9 different symbols and solving the puzzle consists of trying to place these symbols in a grid subject to various constraints.

In fact, the construction of a valid completed sudoku grid is equivalent to a graph-theoretic coloring problem in the following sense. Imagine that every one of the 81 squares is a vertex in a graph, and there is an edge connecting every pair of vertices whose squares are buddies. Each vertex will be connected to 20 other vertices, so the sudoku graph will consist of $81 \cdot 20/2 = 810$ edges. Finding a valid sudoku grid amounts to finding a way to color the vertices of the graph with nine different colors such that no two adjacent vertices share the same color.

Since the symbols do not matter, we could use the letters A through I or any other set of nine distinct symbols to represent what is essentially the same sudoku puzzle. If we take a valid grid and exchange the numbers 1 and 2, this is also essentially the same puzzle. In fact, any permutation of the values 1 through 9 will also yield an equivalent puzzle, so there are $9! = 362880$ versions of every puzzle available simply by rearranging the digits.

If you were trying to calculate how many grids there are, a good approach would be to assume that the top row consists of the numbers 1 through 9 in order, to count the number of grids of that type there are, and then to multiply that result by $9! = 362880$.

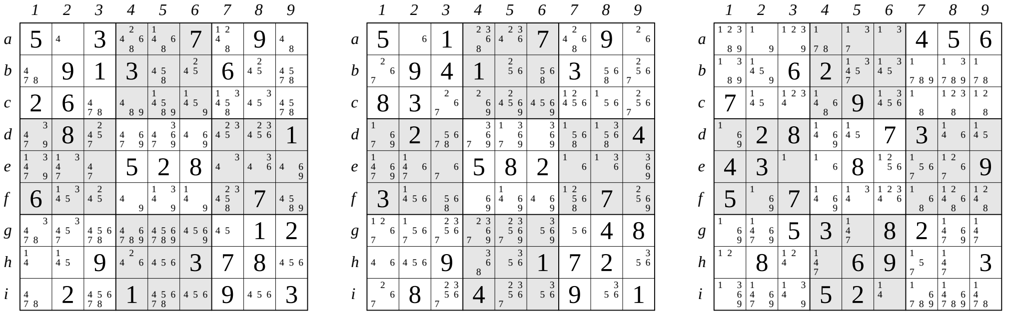

Figure 31: Essentially Equivalent Puzzles

In addition to simply rearranging the numbers, there are other things you could do to a puzzle that would effectively leave it the same. For example, you could exchange any two columns (or rows) of numbers, as long as the columns (or rows) pass through the same blocks. You can exchange any column (or row) of blocks with another column (or row) of blocks. Finally, you can rotate the entries in a grid by any number of quarter-turns, or you could mirror the grid across a diagonal.

Figure 31 shows some examples. If the puzzle on the left is the original one, the one in the center shows what is obtained with a trivial rearrangement of the digits 1 through 9 (the entries 1, 2, $\ldots$, 9 are replaced in the center version by 4, 8, 1, 6, 5, 3, 7, 2 and 9, respectively). The version on the right is also equivalent, but it is very difficult to see how it is related to the puzzle on the left.

One obvious mathematical question is then, how many equivalent puzzles are there of each sudoku grid in the sense above?

Another interesting mathematical question arises, and that is the following: given two puzzles that are equivalent in the sense above, and given a sequence of steps toward the solution of one that are selected from among those explained in earlier chapters, will those same steps work to solve the other puzzle. In other words, if there is a swordfish position in one, will we arrive at a different swordfish in the other? The answer is yes, but how would you go about proving it?

Notice that the puzzle on the left (and in the center) in figure 31 is symmetric in the sense that if you mark the squares where clues appear, they remain the same if the puzzle is rotated by 180 degrees about the center. Other versions of symmetric puzzles could be obtained by mirroring the clue squares horizontally or vertically. Most published puzzles have this form. This doesn't necessarily make them easier or harder, but it makes them look aesthetically better, in the same way that most crossword puzzles published in the United States are also symmetric.

Another interesting question is this: given a symmetric puzzle, how many equivalent versions of it are there?

15 Counting Sudoku Grids

A sudoku grid is a special case of a 9×9 latin square[4] with the additional constraint barring duplicates in the blocks. There are a lot of 9×9 latin squares:

$$5524751496156892842531225600.$$

Bertram Felgenhauer (see the first article referenced in section 19) has counted the number of unique sudoku grids using a computer, and his result has been verified by a number of other people, and that number turns out to be much smaller, but also huge:

$$6670903752021072936960 = 2^{20}3^{8}5^{1}7^{1}27704267971^{1}.$$

The number above includes all permutations of the numbers 1 through 9 in each valid grid, so if we divide it by $9!$ we obtain:

$$18383222420692992 = 2^{13}3^{4}27704267971^{1},$$

which will be the number of inequivalent grids.

16 Magic Sudoku Grids

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
a					4	7			2
b			9				8		
c				1					6
d	6	5	2					7	8
e			8	9					
f				7	8			2	3
g						8			1
h	2	9	4				6		
i	8								

Figure 32: Magic Sudoku Puzzle

A latin square has all the digits in each row and column. A "magic square" is a latin square where each diagonal also contains all the digits. Is there such a thing as a magic sudoku grid? The answer is yes, and in fact there are a lot of them: 4752, in fact, if we assume that the main diagonal contains the digits in a fixed order. All 4752 of the grids can be completed, and all of them in multiple ways. The puzzle presented in figure 32 is a standard sudoku puzzle, except that it is easier since it requires that each of the diagonals contains all the digits from 1 to 9.

[4] A latin square is a grid where the only constraint is that there be no duplicate entries in any row or any column.

17 Minimal Sudoku Puzzles

What is the minimal number of locations must be filled in an otherwise empty grid that will guarantee that there is a unique solution? As of the time this paper was written, the answer to that question is still unknown, but examples exist of puzzles that have only 17 locations filled and do have a unique solution. Figure 33 shows such a puzzle on the left. Although this puzzle contains the minimum amount of information in terms of initial clues, it is not, in fact, a difficult puzzle. The puzzle to the right in the same figure contains 18 clues, and is symmetric. This is the smallest known size for a symmetric sudoku puzzle.

	1	2	3	4	5	6	7	8	9
a						6	1		
b					9	1	6		8
c	7								
d									2
e	3	7							
f						4			6
g				7	3				
h		8	1						
i				5					

	1	2	3	4	5	6	7	8	9
a			8					3	2
b					6	1			
c			5						
d	6					3			
e	1								7
f				2					8
g							6		
h				8	2				
i	5	3					9		

Figure 33: Minimal Puzzles

18 Constructing Puzzles and Measuring Their Difficulty

The difficulty of a sudoku puzzle has very little to do with the number of clues given initially. Usually, the difficulty ratings are given to indicate how hard it would be for a human to solve the puzzle. A computer program to solve sudoku puzzles is almost trivial to write: it merely needs to check if the current situation is solved, and if not, make a guess in one of the squares that is not yet filled, remembering the situation before the guess. If that guess leads to a solution, great; otherwise, restore the grid to the state before the guess was made and make another guess.

The problem with the guessing scheme is that the stack of guesses may get to be twenty or thirty deep and it is impossible for a human to keep track of this, but trivial for a computer. A much more typical method to evaluate the difficulty of a puzzle is relative to the sorts of solution techniques that were presented in the earlier sections of this article.

In this article, the techniques were introduced in an order that roughly corresponds to their difficulty for a human. Any human can look at a row, column or block and see if there is just one missing number and if so, fill it in, et cetera.

So to test the difficulty of a problem, a reasonable method might be this. Try, in order of increasing difficulty, the various techniques presented in this article. As soon as one succeeds, make that move, and return to the beginning of the list of techniques. As the solution proceeds, keep track of the number of times each technique was used. At the end, you'll have a list of counts, and the more times difficult techniques (like swordfish, coloring, or multi-coloring) were used, the more difficult the puzzle was.

The rankings seen in newspapers generally require that the first couple of rankings (say beginning and intermediate) don't use any technique other than those that yield a value to assign to a square on each move. In other words, they require only obvious candidates, naked and hidden singles to solve.

Published puzzles almost never require guessing and backtracking, but the methods used to assign a degree of difficulty vary from puzzle-maker to puzzle-maker.

With a computer, it is easy to generate sudoku puzzles. First find a valid solution, which can be done easily by assigning a few random numbers to a grid and finding any solution. Next start removing numbers (or pairs of numbers, if a symmetric puzzle is desired) and try to solve the resulting puzzle. If it has a unique solution, remove more numbers and continue. If not, replace the previously-removed numbers and try again until a sufficient number of squares are empty. The puzzle's difficulty can then be determined using the techniques described above. The entire process will take only a fraction of a second, so one would not need to wait long to obtain a puzzle of any desired degree of difficulty.

19 References

At the time of writing this article, the following are good resources for sudoku on the internet and in books:

- `http://www.afjarvis.staff.shef.ac.uk/sudoku/`: Felgenhauer's paper that counts possible sudoku grids.
- `http://www.geometer.org/puzzles`: You can download the source code for the author's program that solves sudoku puzzles and can generate the graphics used in this article.
- `http://www.websudoku.com/`: This page by Gideon Greenspan and Rachel Lee generates sudoku games of varying degrees of difficulty and allows you to solve the problem online.
- `http://angusj.com/sudoku/`: From this page you can download a program written by Angus Johnson that runs under Windows that will help you construct and solve sudoku problems. In addition, the page points to a step-by-step guide for solving sudoku, similar to what appears in this document.
- `http://www.sadmansoftware.com/sudoku/index.html`: This site by Simon Armstrong points to some nice descriptions of solution techniques, most of which are discussed in this article.
- `http://www.setbb.com/phpbb/index.php`: This page is a forum for people who want to solve and construct sudoku puzzles as well as for people who want to write computer programs to solve sudoku automatically.
- `http://www.madoverlord.com/projects/sudoku.t`: A downloadable program called Sudoku Susser by Robert Woodhead for the Mac, Windows and Linux that will solve almost any puzzle using logic alone. The distribution comes with great documentation as well, that describes many of the techniques presented here and others besides.
- **How to solve sudoku: A step-by-step guide** by Robin Wilson, published by The Infinite Ideas Company, Limited, Oxford, 2005.
- At the time of publication, there are literally hundreds of books filled with sudoku puzzles of varying degrees of difficulty available in any bookstore.
- **Introduction to Boolean Algebras** by Philip Dwinger, published by Physica-Verlag, 1971.

20 Solutions

	1	2	3	4	5	6	7	8	9
a	6	2	4	8	7	1	9	5	3
b	1	9	3	4	6	5	8	7	2
c	7	5	8	3	9	2	6	1	4
d	2	1	9	6	4	3	5	8	7
e	5	8	6	7	2	9	3	4	1
f	4	3	7	1	5	8	2	6	9
g	3	4	5	2	1	6	7	9	8
h	8	6	1	9	3	7	4	2	5
i	9	7	2	5	8	4	1	3	6

	1	2	3	4	5	6	7	8	9
a	2	1	5	8	7	6	9	4	3
b	6	7	8	3	9	4	2	1	5
c	3	4	9	1	2	5	8	7	6
d	5	8	7	4	3	2	1	6	9
e	4	6	3	9	8	1	7	5	2
f	1	9	2	6	5	7	3	8	4
g	8	2	6	7	4	3	5	9	1
h	7	3	4	5	1	9	6	2	8
i	9	5	1	2	6	8	4	3	7

Figure 34: Solutions to Puzzles