ARTICLES

The Other Map Coloring Theorem

The problem of coloring a map on a pretzel or Möbius band was solved before the Four-Color Problem.

SAUL STAHL
The University of Kansas
Lawrence, KS 66045

The Four Color Problem has probably been the most notorious mathematical problem of modern times. This problem asks whether four colors suffice to color every planar map so that adjacent countries, i.e., countries that share a border of positive length, receive different colors. This question's deceptive simplicity attracted many would-be solvers who oft times spent years on their search for a solution. Most returned empty-handed, or worse, with a false proof. Some were fortunate enough to have devised a new twist on the original problem that was sufficiently interesting to attract the attention of other aficionados. The **Heawood Conjecture**, one of the earliest of these offshoots, proved to be also one of the most fascinating and difficult. This other coloring conjecture guessed at the number of colors required by maps on other, more complicated, surfaces. Surprisingly enough, even though this later problem seems more difficult than its planar progenitor, Heawood's conjecture was actually verified a decade earlier. In my opinion this verification marks a milestone in the development of the modern combinatorial approach to geometry. It is my intention here to formulate this problem, recount its history, and discuss its relationship to the original Four Color Problem, as well as to other branches of mathematics.

Heawood's Conjecture

It is generally agreed that the Four Color Conjecture was first formulated by Francis Guthrie, a graduate student at University College, London, in 1852. Appel and Haken's proof of this conjecture is described by them in [**1**] and has also received wide publicity elsewhere, so I will not discuss it here and only refer to it when it provides an interesting parallel or contrast with the other map coloring theorem. The first false proof of the Four Color Conjecture to be published was given in 1879 by A. B. Kempe [**14**], barrister and part-time mathematician. The validity of this proof was not challenged until 1890 when P. J. Heawood published a paper [**10**] in which he accomplished several things. First, he pointed out the error in Kempe's reasoning. Next, he salvaged the remains of Kempe's fallacious proof by using its techniques to show that every planar map can be colored with *five* colors. Finally, he went on to state and solve, so he thought, the same problem in a new context. This came to be known as the Heawood Conjecture.

Before describing this new context, a minor clarification is in order. Whenever a map is mentioned above and in the sequel, it is to be understood that each country forms a single contiguous geographical unit, unlike pre-Bangladesh Pakistan. Also, in order to minimize redundancy it will be assumed that whenever a map is colored, the coloring satisfies this map-coloring constraint: adjacent countries receive different colors.

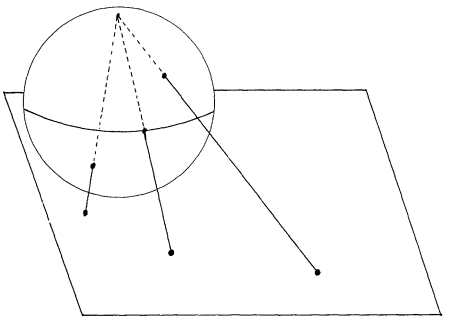


FIGURE 1. **Stereographic projection. A technique used to represent spherical maps on flat paper. The particular version portrayed here will represent regions near the south pole fairly faithfully, but will greatly distort northern countries. Nevertheless, it does preserve the adjacency pattern.**

The plane is not the only surface on which maps can be drawn; they can of course also be drawn on the surface of a sphere. However, coloring maps on spheres is really not different from coloring them on the plane. The well-known stereographic projection of FIGURE 1, often employed by map makers, can be used to convert any spherical map to a planar one, and the coloring pattern of the planar projection of a spherical map can also be used to color the original map on the sphere. Nothing is therefore to be gained by reformulating the Four Color Problem for the sphere. Maps on the surface of a cone or a pyramid do not provide any new challenges either, for these surfaces can be easily deformed into spherical ones in a manner that preserves the adjacency pattern of any maps drawn on them. So, if one is to find a surface that will yield a genuinely new coloring problem, then this surface cannot be deformable into a sphere in any "nice" way. One such surface is a torus—the surface of the doughnut. In M_5 (FIGURE 2) we have a toroidal map consisting of five countries every two of which are adjacent. Such a map clearly requires *five* colors. In fact, M_6 and M_7 (FIGURES 3, 4) are toroidal maps with six and seven countries in which every two countries are adjacent to each other. Consequently, these maps require six and seven colors, respectively. The subsequent discussion centers around maps of this type and it will therefore be convenient to have a name for them. Accordingly, a **complete m-map** is a map which consists of m countries every two of which are adjacent. Such a map clearly requires m colors. It is natural to ask at this point whether there exists a complete 8-map. The answer is negative for the torus.

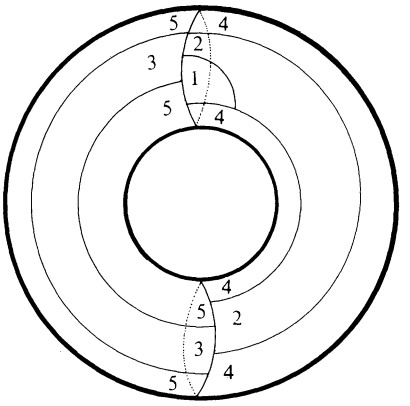


FIGURE 2. **M_5, a toroidal map of five countries every two of which touch each other.**

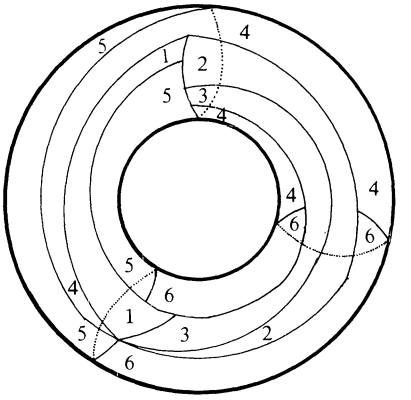


FIGURE 3. **M_6, a toroidal map of six countries every two of which touch each other.**

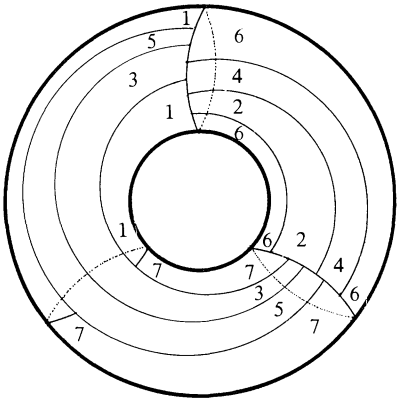


FIGURE 4. M_7, a toroidal map of seven countries every two of which are adjacent.

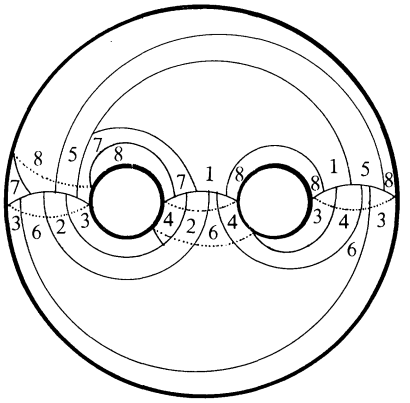


FIGURE 5. M_8, a complete 8-map on the double torus S_2.

The mathematician A. F. Möbius (see [**2**]) had already pointed out in the 1840's that it is not possible to draw a complete 5-map in the plane. In later years this observation was sometimes misconstrued as a solution of the Four Color Problem. It is, of course, no such thing. The nonexistence of a complete 5-map in the plane does not preclude the existence of an incomplete planar map with a large number of countries, whose complexity actually requires five or more colors. We now have an analogous situation on the torus, for it supports a complete 7-map and no complete 8-map can be drawn on it. Does this mean that every toroidal map can be colored with seven colors? Such indeed is the case. What is surprising is the relative ease with which this answer was obtained. So simple is this proof, that its main points deserve to be brought out here.

In any map, let n denote the number of its **nodes**, namely, the number of points at which three or more countries meet. Let b denote the number of **borderlines** in the map, in other words, curves that go from one node to another. The Euler-Poincaré formula [**15**], one of the central facts of geometry, implies that if a toroidal map has m countries, then

$$n - b + m = 0. \tag{1}$$

For the maps M_5, M_6, and M_7, this formula takes the forms $9 - 14 + 5 = 0$, $9 - 15 + 6 = 0$, and $14 - 21 + 7 = 0$, respectively. For the sake of accuracy let me point out that the Euler-Poincaré formula does not apply to maps in which one country has the shape of a ring and completely surrounds one or several other countries. Such maps, however, can be easily disposed of by certain standard reduction techniques and will be ignored in the sequel.

Let A denote the average number of borderlines that occur on the boundaries of the m countries of a given map. Then mA is the count of all the occurrences of borderlines on the boundaries of these countries. Since each of the b borderlines occurs on 2 of these boundaries, it follows that

$$b = \frac{mA}{2}. \tag{2}$$

Turning our attention to the nodes, observe that A must also be the average number of *nodes* on the boundaries of these countries, since each country has an equal number of nodes and borderlines on its boundary. However, in contrast with the borderlines, we cannot specify the number of countries on whose boundary a given node will occur. Still we can say that each node appears on the boundaries of at least 3 countries. This gives us the following weak analog of equation (2), which is nevertheless sufficient for our purpose:

$$n \leqslant \frac{mA}{3}. \tag{3}$$

Combining (1), (2), and (3) we obtain

$$\frac{mA}{3} - \frac{mA}{2} + m \geqslant n - b + m = 0,$$

$$m\left(1 - \frac{A}{6}\right) \geqslant 0.$$

Since m is a positive quantity, we may conclude that $A \leqslant 6$. In other words, the average number of borderlines on the boundary of a typical country in any toroidal map does not exceed 6. Consequently we obtain the following surprising fact: *in every toroidal map there must be a country that is adjacent to no more than six other countries*. This fact immediately points out the impossibility of drawing a complete 8-map on the torus, since in such a map every country is necessarily adjacent to *seven* other countries. This fact can also be used to define an algorithm for coloring any toroidal map with $6 + 1 = 7$ colors. Thus, every map on the torus can be colored with seven colors and some such map (the complete 7-map) actually requires seven colors. In other words, the answer to the toroidal analog of the Four Color Problem is that 7 colors suffice. Heawood was the first to formulate this analog clearly and to answer it. He did much more, however.

Why does the torus allow for more complicated maps than the sphere? (The complexity of a map is here equated with the number of colors it requires.) One might say that this additional complexity is made possible by the hole in the torus—the doughnut hole. The adjacency pattern of a spherical map is constrained by the fact that a country that lies completely in the northern hemisphere cannot possibly be adjacent to one that lies entirely in the southern hemisphere. However, if a tunnel is bored from the north pole of the sphere to its south pole, thus converting it essentially into a doughnut, it now becomes possible for a northern country to reach out and touch a southern one *without crossing the equator*, namely, along the walls of the tunnel. From this point of view, a tunnel through a surface may be thought of as a bridge that connects two parts of the surface and allows for a higher degree of complexity in the adjacency pattern of its maps. So if we wish to find a surface that supports a complete 8-map (which the torus does not), it is clear what must be done. A tunnel should be bored through the torus, or, equivalently, two nonintersecting tunnels should be bored through the sphere. The resulting surface, the double torus, does indeed support the complete 8-map M_8 of FIGURE 5.

We now have a technique for creating surfaces that would seem to allow for maps requiring arbitrarily many colors. All that needs to be done is to bore enough tunnels through the sphere. We will call the surface that results from boring g nonintersecting tunnels through the sphere the **surface of genus** g and denote it by S_g. Thus S_2 is the double torus, the torus itself is S_1, and the surface of the unperforated sphere is S_0. The great mathematician Bernard Riemann brought these surfaces into the foreground of mathematics in 1851 [**18**] when he showed that they play a focal role in the calculus of complex variables. His theory was so central to nineteenth century mathematics that it has been said [**4**, p. 121] that at one time all research mathematicians had to be familiar with it. It is therefore not surprising that shortly after the Four Color Problem was formulated for planar maps, mathematicians would ask the same question in the context of Riemann's surfaces.

The Euler-Poincaré formula states that whenever a map with m countries, b borderlines, and n nodes is drawn on the surface S_g, then these parameters are linked by the equation

$$n - b + m = 2 - 2g. \tag{4}$$

Heawood used this fact to show that if $g \geqslant 1$, *then any such map can be colored with no more than*

$$H_g = \left[\frac{1}{2}\left(7 + \sqrt{1 + 48g}\right)\right] \tag{5}$$

colors, where the brackets in (5) denote the integer part of the enclosed number. This guarantees that any map on the torus can be colored with $H_1 = [\frac{1}{2}(7 + \sqrt{1 + 48})] = 7$ colors, that every map on the double torus can be colored with $H_2 = [\frac{1}{2}(7 + \sqrt{1 + 96})] = [8.42\ldots] = 8$ colors, and that

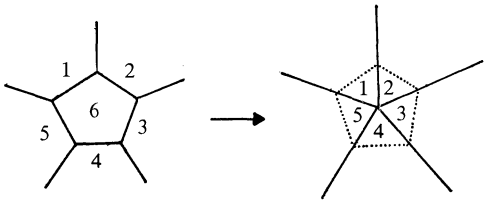


FIGURE 6. **A reduction process for maps that diminishes the number of countries without increasing the number of required colors.**

every map on S_3 can be colored with 9 colors.

Heawood's proof that the number of colors specified in (5) is actually sufficient is paraphrased as follows. Call those borderlines along which a country touches its neighbors its **contacts** (clearly every country possesses at least as many contacts as neighbors, and sometimes more). Suppose that there is an integer x such that every map on S_g has a country with fewer than x contacts. Then we claim that x colors suffice to color every map on S_g. For in any such map we can annihilate the country with less than x contacts, making those round it close up in the space which it occupied (FIGURE 6), and obviously the original map can be done in x colors if the reduced map can (for whatever the coloring of the countries around, there would be a color to spare for the annihilated country). Having thus described an induction process, Heawood now needed only to find the smallest x that satisfied this condition, and he used an averaging argument that is essentially the same as the one used above to argue that the torus does not support a complete 8-map. First, however, Heawood observed that every map on any surface can be converted by the operations described in FIGURE 7 to a map in which every node appears on the boundary of *exactly* 3 countries, this new map requiring no more colors than the original one. Consequently, with n, b, m, A denoting the same quantities as before, equation (2) still holds, whereas inequality (3) can now be replaced by the equation

$$n = \frac{mA}{3}. \tag{6}$$

If we now substitute (2) and (6) in (4) we obtain

$$A = 6\left(1 + \frac{2g-2}{m}\right). \tag{7}$$

As was argued for the torus, every map on S_g clearly has a country with no more than A adjacencies, so any x satisfying

$$x > A \tag{8}$$

should do. Expression (7), however, still contains an m which depends on the map rather than the surface. To eliminate this m Heawood observed that A diminishes as m increases, and since every map with no more than x countries clearly must have a country with fewer than x contacts, it follows that in order for (8) to hold we need only guarantee that

$$x > 6\left(1 + \frac{2g-2}{x+1}\right). \tag{9}$$

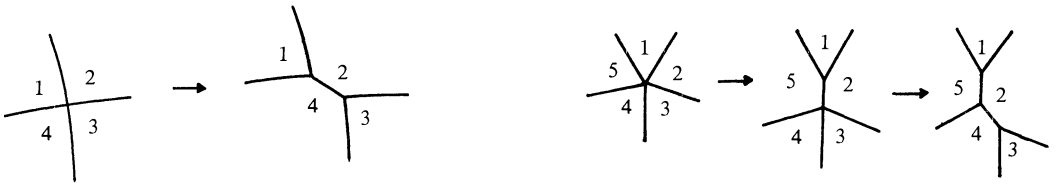


FIGURE 7. **How to transform any map to one requiring no more colors, but in which every node is on the boundary of exactly three countries.**

Thus, if we let H_g denote the smallest of all the integers x that satisfy inequality (9), then the maps on the surface S_g can be colored with H_g colors. Since for all smaller values of x the reverse inequality must hold, this can be rephrased as saying that H_g is the largest integer x that satisfies the inequality

$$x - 1 \leqslant 6\left(1 + \frac{2g-2}{x}\right).$$

This is a straightforward quadratic inequality in x, whose solution yields the value specified in (5).

The above argument is essentially the same as the one presented by Heawood himself. I have even gone so far as to leave intact some of the rougher edges and omissions in his proof, partly for the sake of brevity, and partly to prepare the reader for Heawood's real mistake, which is to be discussed shortly. Cleaner arguments can be found in [**2**] and [**20**].

To demonstrate that H_g colors are in fact required by some map on S_g, Heawood also drew a complete 7-map on the torus—essentially the same as M_7—but only *asserted* the existence of a complete H_g-map on the surface S_g for all $g \geqslant 2$. "For more highly connected surfaces," Heawood stated, "it will be observed that there are generally contacts enough and to spare for the above number of divisions [countries] each to touch each." In other words, Heawood thought it was easy to see that a complete 9-map could be drawn on S_3. Anyone who tries to draw such a map will quickly discover that not only is this quite a difficult task, but it also sheds no light on the question of how to draw a complete 10-map on S_4. Thus, what Heawood accomplished was to show that no map on the surface of genus $g \geqslant 1$ requires more than H_g colors. What he failed to show was that a map requiring this number of colors could actually be drawn on S_g. So a question he believed he had both raised and answered, in fact remained open. The number H_g defined in (5) acquired the name **Heawood number** and the assertion that *The Heawood number H_g is the largest number of colors required to color any map on the surface of genus $g \geqslant 1$* came to be known as the **Heawood Conjecture**.

The observant reader will have noted the qualification $g \geqslant 1$ that occurs in the statement of Heawood's conjecture. Were this qualification absent, the conjecture would apply to the surface S_0 (the sphere), and it would assert that any spherical map could be colored with no more than $H_0 = [\frac{1}{2}(7 + \sqrt{1 + 0})] = 4$ colors! In other words, the removal of the inequality $g \geqslant 1$ would make the Heawood Conjecture contain the Four Color Conjecture as a special case. Unfortunately, the condition $g \geqslant 1$ cannot be disregarded. In Heawood's proof, the quantity

$$A = 6\left(1 + \frac{2g-2}{m}\right)$$

was observed to be a *decreasing* function of m. This of course fails to be the case when $g = 0$, and so the proof breaks down for this value of g. This is a good example of the many near misses that proliferate in the history of the Four Color Problem.

Heffter's contributions

The deficiency in Heawood's paper was pointed out one year after its publication by Lothar Heffter [**11**]. Heffter also realized that the problem needed a new approach. Trying to draw complicated maps on 2-dimensional drawings of perforated spheres was a very cumbersome way to attack this problem. As an alternative he suggested that the adjacency pattern of a map (which, after all, is all that matters here) be recorded in the following manner. Suppose an inhabitant of some country were to inspect its borders by traveling along them in, say, the counterclockwise direction (counterclockwise from the point of view of an observer stationed right above his country). This inhabitant might then record, in order, the neighboring countries along whose border he is traveling. For instance, in the map M_4 of FIGURE 8, the inspector of country 1, if he starts his tour from a location on the border his country shares with country 2, would write 2, 4, 3 in his logbook. Had he started from a location on the border his country shares with country 3, his entry would have been 3, 2, 4. These two records are considered to be the same since it is only the cyclic ordering of the neighbors that matters to us. Now, country 2's inspector, in touring his

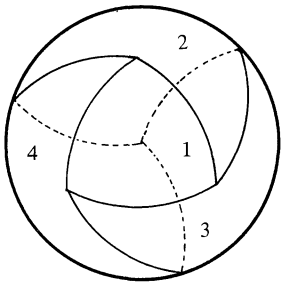


FIGURE 8. M_4, a complete 4-map on the sphere.

country	adjacency record		
1)	2	4	3
2)	1	3	4
3)	4	2	1
4)	3	1	2

FIGURE 9. A_4, the adjacency pattern for the map M_4.

borders, would write 1, 3, 4. The information obtained from these tours and those of the inspectors of countries 3 and 4, can be tabulated as the array A_4 in FIGURE 9. When applied to the toroidal complete 7-map M_7, this process yields the array A_7 in FIGURE 10. The array A_7 has a very exciting pattern. Every row can be derived from the previous one by the addition of 1 to each entry of the latter. This arithmetic is modulo 7; that is, we stipulate that $7 + 1 = 1$ and that the first row follows the seventh one.

country	adjacency record					
1)	2	4	3	7	5	6
2)	3	5	4	1	6	7
3)	4	6	5	2	7	1
4)	5	7	6	3	1	2
5)	6	1	7	4	2	3
6)	7	2	1	5	3	4
7)	1	3	2	6	4	5

FIGURE 10. A_7, the adjacency pattern for the map M_7 in Figure 4.

Before we go on to discuss these arrays in general, it should be pointed out that two miracles occurred in the passage from the complete 7-map M_7 to its array A_7. In the first place, the array displays a pattern, or a symmetry, that is totally obscured in the original map. Symmetry, of course, is one of the strongest tools of mathematics and of science. Hence, finding it in such an unexpected place is indeed an undisguised blessing. The second observation is even more surprising. The numbers $1,2,3,\ldots$ were used as a matter of convenience only. We could, and perhaps should, have used labels such as "Spain" or "Union of the Free Toroidal Republics." Nevertheless, we find that these countries, when symbolized by numbers, follow a very rigid arithmetical pattern. This phenomenon has been observed and exploited in other coloring problems, but *not* in the proof of the Four Color Theorem. Both of these miracles were crucial to the eventual resolution of Heawood's conjecture.

Let us now return to the arrays themselves. The array A_4 obtained from the complete 4-map M_4 does not quite conform to the nice pattern that was discovered in A_7. It seems that in certain columns one might need to subtract rather than add 1. Still, there is enough regularity in the rows of this array to justify some hope for a general pattern.

An obvious question at this point is: *which arrays correspond to some complete m-map*? Such an array must clearly have m rows, and the ith row must be some permutation of the numbers $1,2,\ldots,i-1,i+1,\ldots,m$. Let us call this requirement the **shape constraint**. An example of another array that satisfies the shape constraint is B_4, shown in FIGURE 11. A little experimentation, however, will quickly convince the reader that the array B_4 cannot correspond to any complete 4-map. This can be reasoned out by observing that according to the first row of this array, country 1's inspector, traveling counterclockwise, encounters country 2 just before country 3. Hence, the node formed by countries 1, 2, and 3 must form the configuration of FIGURE 12. However, according to this portion of the map, country 2's inspector must encounter country 3

country	adjacency record		
1)	2	3	4
2)	3	4	1
3)	4	1	2
4)	1	2	3

FIGURE 11. **B_4, an array that is not the adjacency pattern of any map.**

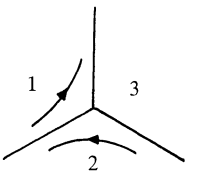


FIGURE 12

just before he encounters country 1, which contradicts the information in the second row of B_4. This observation generalizes to the following **consistency constraint** to which all arrays that describe complete m-maps must conform: *If in the ith row of the array, j is followed by k, then in the kth row, j must be preceded by i.* In other words,

if the ith row is

$i)$ ... j k ...

then the kth row is

$k)$ ... i j ... .

Surprisingly, there are no other constraints. *Any array that satisfies both the shape and consistency constraints describes a complete m-map.* Moreover, as Heffter demonstrated, this complete m-map verifies the Heawood Conjecture for the surface S_g for $g=(m-3)(m-4)/12$. Since the genus g (number of tunnels) of a surface is necessarily an integer, it follows that the product $(m-3)(m-4)$ must be divisible by 12, from which it follows that m itself, when divided by 12, must yield 0, 3, 4, or 7 as remainder. Since, in general, division by 12 yields one of 12 possible remainders, this approach can yield the appropriate complete m-map for at most one third of the possible values of m. As it happens, even this estimate is overly optimistic. Heffter was aware that an array possessing the additional cyclic structure displayed by A_7 could only be constructed when the remainder of m divided by 12 was 7. All of these limitations notwithstanding, he felt that this was an approach well worth pursuing. He translated the problem into a purely number theoretic one and then showed that the required cyclic arrays A_m exist for those values of m that satisfy conditions i)–iii) below:

i) m leaves remainder 7 when divided by 12,

ii) $(m+2)/3$ is a prime number,

iii) 2^k-1 is not divisible by $(m+2)/3$ for $k=1,2,3,\ldots,(m-7)/6$.

Among the values that satisfy these conditions we find the numbers $m=19,31,55,67,139,175,199$. Consequently, Heffter verified the Heawood Conjecture for the surfaces S_g where $g=20,63,221,336,1530,2451,3185$. Actually, he accomplished a little more than that. Since $H_{21}=[\frac{1}{2}(7+\sqrt{1+48\times 21})]=[19.38\ldots]=19=H_{20}$, and since every map that can be drawn on S_{20} can also be drawn on S_{21}, it follows that Heffter's construction of A_{19} also verifies the Heawood Conjecture for S_{21}, as it does for S_{22} as well. Similarly, the existence of the array A_{199} verifies the Heawood Conjecture for all surfaces with at least 3185 but no more than 3217 tunnels. The algebra Heffter used in his work was deep and difficult enough that he could not decide whether the class of surfaces to which his proof applies was finite or infinite. Even today it is still not known whether Heffter's solution applies to infinitely many surfaces.

Heffter's contributions can be summarized as follows. He pointed out the error in Heawood's paper and thereby attracted the attention of the mathematical world to Heawood's beautiful conjecture. He showed how this geometrical problem could be translated first to a combinatorial one of arrays, and then to an algebraic problem in the theory of numbers. He went on to solve this number theoretic problem in some special cases. Finally, a fact that was not mentioned above, he also confirmed the Heawood conjecture for all surfaces of genus at most 7 by constructing the appropriate arrays.

The resolution of Heawood's conjecture

Nothing was contributed toward the resolution of Heawood's conjecture in the sixty years that followed the publication of Heffter's work. That mathematicians were aware of it is attested to by the fact that it is mentioned in Hilbert and Cohn-Vossen's 1932 book *Geometry and the Imagination* [**12**], where it is called the problem of contiguous regions. The lack of progress on this problem during the first half of this century can be attributed to three factors. First comes its inherent difficulty. Next, some mathematicians believed that the problem had indeed been solved by Heawood. Courant and Robbins, in their 1941 book *What is Mathematics*? [**5**, p. 248] wrote: "A remarkable fact connected with the four color problem is that for surfaces more complicated than the plane or the sphere the corresponding theorems have actually been proved." (As late as 1980, I met a topologist who believed the same.) Finally, and perhaps most significantly, to many mathematicians this problem seemed to represent a blind alley. It had been shown towards the end of the 19th century that the combinatorial approach could be very fruitful in analyzing geometrical problems in all dimensions. However, there were no indications that a solution to Heawood's problem would lead anywhere else. Even the Four Color Problem, which many had discounted for the same reason, has corollaries which mathematicians consider to be interesting, although perhaps not "important."

The first breakthrough to follow Heffter's pioneering work was produced by G. A. Dirac [**7**], a relative of the famed physicist, in 1952. To understand his accomplishment, let us recall the issue. It was known that every map on the surface of genus g could be colored with no more than H_g colors. To show that this number of colors might actually be required, it would suffice to draw a complete H_g-map on this surface. But is the existence of such a complete map on the surface really necessary? Conceivably the surface S_3 might support a map that requires 9 colors even though it might not admit a complete 9-map. That, after all, was the whole point of the Four Color Problem. It was known that no complete 5-map could be drawn in the plane, but that did not preclude the possibility of some other planar map actually requiring 5 colors. Dirac showed that this situation could not arise on the other surfaces. Specifically, he demonstrated that if the surface of genus g supported a map that required H_g colors, then it would also support a complete H_g-map. Actually, he only proved this for the cases $g = 3$ and $g \geqslant 5$. His arithmetic got in the way of the cases $g = 0, 1, 2, 4$. Had his proof applied to the case $g = 0$, he would have produced a proof of the Four Color Conjecture—yet another near miss. However, the other values he missed, $g = 1, 2, 4$, were already covered by Heffter's work.

Next, in what has been called "a tour de force of combinatorial brilliance" ([**22**, p. 317]), Gerhard Ringel [**19**] constructed in 1954 a set of arrays which confirmed the existence of a complete H_g-map on the surface S_g for all those values of H_g that leave a remainder of 5 upon division by 12. At last the Heawood conjecture had been verified for an infinite number of surfaces.

The special significance that the number 12 has for this problem was pointed out earlier in the discussion of Heffter's work. It was noted there that arrays that satisfied both the shape and the consistency constraints could only be obtained when m left remainders of 0, 3, 4, or 7 upon division by 12. Since the remainder 5 is not one of these, Ringel had to relax the constraints somewhat. He chose to relax the shape constraint as is evident in the arrays $A_5^{(-1)}$ in FIGURE 13

country	adjacency record			
1)	2	4	3	5
2)	3	4	1	5
3)	1	4	2	5
4)	1	2	3	
5)	3	2	1	

FIGURE 13. **$A_5^{(-1)}$, the adjacency pattern of the map $M_5^{(-1)}$.**

1)	6	4	11	8	13	3	9	17	2	16	15	12	7	5	14	10
4)	9	7	14	11	1	6	12	17	5	16	3	15	10	8	2	13
7)	12	10	2	14	4	9	15	17	8	16	6	3	13	11	5	1
10)	15	13	5	2	7	12	3	17	11	16	9	6	1	14	8	4
13)	3	1	8	5	10	15	6	17	14	16	12	9	4	2	11	7
2)	6	14	7	10	5	9	3	16	1	17	12	15	11	13	4	8
5)	9	2	10	13	8	12	6	16	4	17	15	3	14	1	7	11
8)	12	5	13	1	11	15	9	16	7	17	3	6	2	4	10	14
11)	15	8	1	4	14	3	12	16	10	17	6	9	5	7	13	2
14)	3	11	4	7	2	6	15	16	13	17	9	12	8	10	1	5
3)	16	2	9	1	13	7	6	8	17	10	12	11	14	5	15	4
6)	16	5	12	4	1	10	9	11	17	13	15	14	2	8	3	7
9)	16	8	15	7	4	13	12	14	17	1	3	2	5	11	6	10
12)	16	11	3	10	7	1	15	2	17	4	6	5	8	14	9	13
15)	16	14	6	13	10	4	3	5	17	7	9	8	11	2	12	1
16)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
17)	14	13	6	11	10	3	8	7	15	5	4	12	2	1	9	

$A_{17}^{(-1)}$

FIGURE 14. **An array that describes the adjacency pattern of a nearly complete map on S_{15}. This map has seventeen countries of which every two, except 16 and 17, are adjacent to each other.**

and $A_{17}^{(-1)}$ in FIGURE 14. Specifically, in $A_5^{(-1)}$, row 4 does not contain the entry 5 and, vice versa, row 5 does not contain the entry 4. This means that in the associated map $M_5^{(-1)}$ of FIGURE 15 the two countries 4 and 5 are not adjacent to each other. The superscript (-1) in the notation for the array and the map records the fact that both the array and the map lack one adjacency to being complete. Similarly, in the map represented by $A_{17}^{(-1)}$, countries 16 and 17 are not adjacent to each other. Now it so happens that the array $A_{17}^{(-1)}$ actually represents an "almost complete" 17-map on the surface S_{15}. Connect the nonadjacent countries 16 and 17 by boring an additional tunnel through S_{15}. This gives us a complete 17-map on the surface S_{16}. Since $H_{16} = 17$ this array implies the validity of the Heawood Conjecture for the surface of genus 16.

Due to the special role played by the number 12, the problem was now recognized as possessing 12 cases, depending on H_g's remainder when divided by 12. By 1961 Ringel had also resolved the cases of remainder 7, 10, and 3, making use of the same technique as in the case of 5. The year 1963 saw the emergence of some new tools. William Gustin [**9**] discovered a way of encoding arrays in a form that greatly resembles electrical networks. In these, the numbers originally used to label countries are interpreted as denoting the intensity of a (fictitious) current flowing along the branches of the network in the direction indicated by the arrowheads. Kirchhoff's Current Law, which states that the total current entering a node equals the total

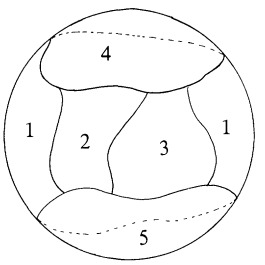


FIGURE 15. **$M_5^{(-1)}$, an almost complete 5-map on the sphere. Every two countries, except 4 and 5 touch each other.**

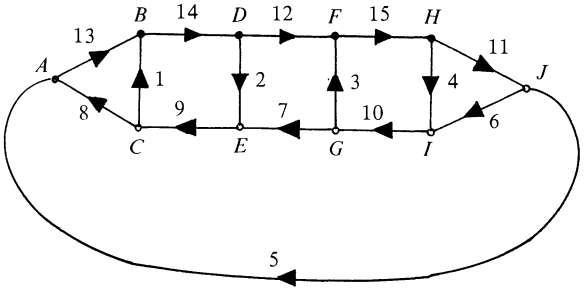


FIGURE 16. **CG_{31}, a current graph which encodes a complete 31-map on S_{63}.**

current leaving it, is also satisfied by these networks. Because of this resemblance, Gustin's networks have been dubbed "Current Graphs." To indicate the manner in which an array may be stored in a current graph, we decode CG_{31} (FIGURE 16). Choosing our departure point arbitrarily, we start out with the network branch labelled 5 and proceed to node A. From here we could choose to go to either B or C. The correct choice is indicated by the way in which the node is drawn. A solid dot indicates that at this node the traveler should choose the left fork, whereas a hollow dot dictates a choice of the right fork. Hence we go on to B and record a 13 in our logbook, to follow the previous entry 5. Node B is also a solid dot, and so we go to D from here and record a 14 in the logbook. By the time node J is reached, the logbook's entries will be $5,13,14,12,15,11$. Since the node J is represented by a hollow dot, we choose the right fork and go on to I, adding the entry 6 to the log. From I we again bear right, toward H, but this time, since we are progressing *against* the arrowhead, instead of recording 4 in the log, we enter $31 - 4 = 27$. From the solid dot of H, we choose the left fork to F and record $31 - 15 = 16$ in the log. The solid and hollow nodes are so placed that the initial current 5 will not be encountered again before all the possible currents from 1 to 30 have been recorded in the logbook. In other words, this procedure is set up so that each of the branches of the network will be traversed exactly twice, once in each direction. If we consider the above tour as terminating when 5 is reencountered, then the final log is:

$$5,13,14,12,15,11,6,27,16,28,7,29,17,30,8,$$
$$26,20,4,10,3,19,2,9,1,18,23,22,24,21,25.$$

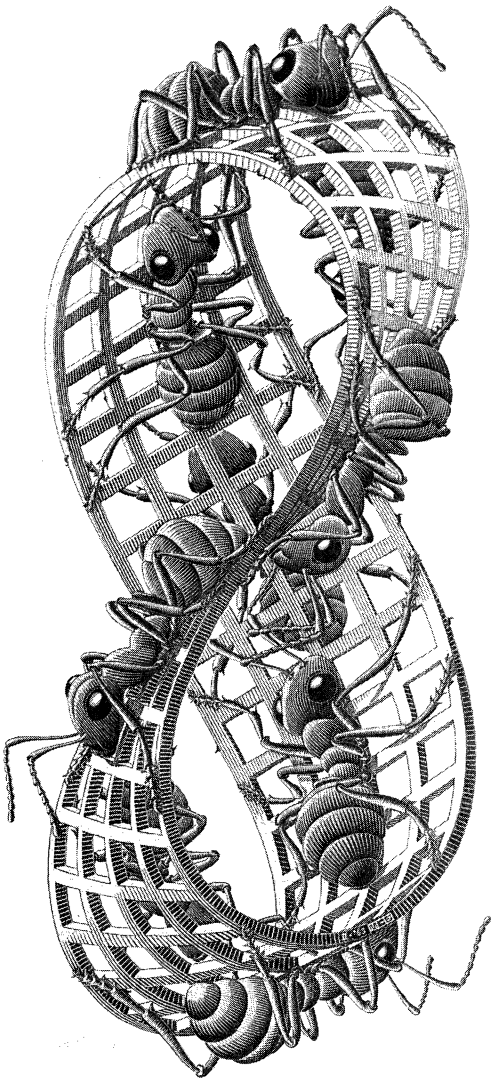
The ith row of the array A_{31} is now obtained by adding i to each entry of the above log, with the stipulation that $31 + i$ is to be replaed by i. This array satisfies both the shape and consistency constraints and so it attests to the existence of a complete 31-map on the surface S_{63}. Since $H_{63} = [\frac{1}{2}(7 + \sqrt{1 + 48 \times 63})] = 31$, the Heawood Conjecture for that surface is verified.

Mysterious as the above procedure may seem, we know why it works. The distribution of the currents among the links of the network and the relative placements of the solid dots versus the hollow dots among the nodes are such as to guarantee that the shape constraint is fulfilled. It can also be shown that Kirchhoff's Current Law is transformed in the array into the consistency constraint. On the other hand, the heuristics underlying the idea that a map can be encoded as an electrical network are still not well understood. It should be pointed out that complex function theory, within which Riemann's surfaces were first recognized, has very strong connections to potential theory. Indeed, some very deep mathematical theorems become "obvious" when translated into statements about electrons and electrical fields. Be that as it may, these current graphs are of course much more tractable than the arrays they represent, and these arrays are in turn much more tractable than the actual maps they represent. Nevertheless, even these graphs are far from easy to find, especially as for some of the cases they have to be slightly modified in order to work. It was not until 1968 that the combined efforts of W. Gustin, R. K. Guy, C. M. Terry, L. R. Welch, and, most of all, G. Ringel and J. W. T. Youngs (see [**20**] for details) resolved all of the remaining eight cases of remainders $0,1,2,4,6,8,9,11$. Their work left the cases $H_g = 18,20,$

and 23 unresolved, but these gaps were filled within one year by J. Mayer [**16**] (professor of French literature at the University of Montpellier). It took three quarters of a century to verify completely the statement that Heawood felt was too obvious to require justification.

Coloring maps on other surfaces

The work described above completely resolved the coloring problem on Riemann's surfaces, except for the sphere. An issue that has so far been sidestepped is the question of whether there are any other surfaces for which interesting coloring problems could be posed. The reader is reminded that maps of higher complexity were made possible by boring tunnels through the sphere, thus motivating the definition of the Riemann surfaces. Is there anything else that could be done to a sphere to allow for more complex maps? The answer is yes.



Mobius Strip II, woodcut by M. C. Escher, 1963.
Copyright M. C. Escher heirs c/o Cordon Art B. V.-Baarn, Holland.

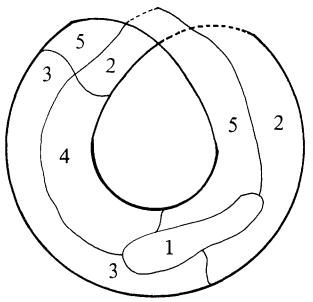


FIGURE 17. $\tilde{M}_5$, **a complete 5-map on the Möbius strip.**

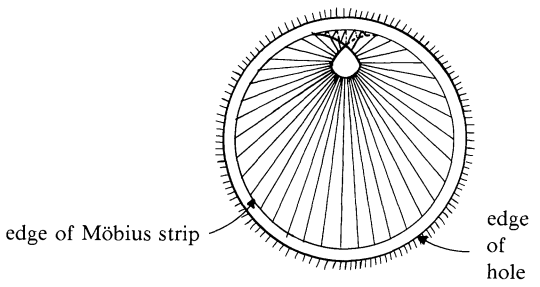


FIGURE 18. **Trying to patch a hole with a Möbius strip.**

The map $\tilde{M}_5$ (FIGURE 17) displays a complete 5-map on the Möbius strip. This strip is obtained by making a 180° twist in a long ribbon and then gluing its two ends to each other. Now this map can be transferred to the sphere in the following manner. Observe that the edge of the Möbius strip consists of a *single* closed loop, as opposed to, say, the two loops that would have formed the edge had the ribbon not been twisted before its ends were glued. If a disk is cut out of the surface of the sphere, a closed loop is left as the edge of the perforated sphere. Since all closed loops are essentially identical, it should be possible to patch the hole in the sphere by sewing its edge to the edge of the Möbius strip. Most of the readers who try to carry out this patching process will soon discover that the twist in the strip has a nasty tendency to get in the way (see FIGURE 18). However, my four-dimensional readers should experience no such difficulties. That extra dimension will provide them with all the room they need to maneuver. For this reason mathematicians have accepted the resulting hybrid as a genuine surface, even though its internal structure prevents it from being realized in our three-dimensional space. It meets the criteria that mathematicians have set for surfaces, namely, they can have no edges and at each point they must be "reasonably" flat.

Of course, once this twisted surface is accepted, one must be prepared to allow for the possibility of adding more than one twist, just as we allowed for the possibility of boring several tunnels in the sphere. The surface obtained by adding g twists to the surface of the sphere is called the **one-sided surface of genus** g, and is denoted by $\tilde{S}_g$. It is one-sided because the twisted patches would allow an ant living on the outside of the sphere to move to its inside simply by walking along a twist in the manner depicted by Escher's famous woodcut. The one-sided surface $\tilde{S}_1$ is the projective plane and $\tilde{S}_2$ is the well-known Klein bottle. By way of contrast, Riemann's surfaces are all two-sided. An ant on the outside has no way of getting into the inside. And what would happen if twists were added to the other two-sided surfaces? Would we then obtain a whole new bewildering collection of surfaces with mixed numbers of tunnels and twists? Fortunately the answer is no. It has been known since the turn of the century that when a twist is added to the two-sided surface of genus g, one simply obtains the one-sided surface of genus $2g + 1$. Similarly, when a tunnel is bored into the one-sided surface of genus g, the result is the one-sided surface of genus $g + 2$. Moreover, there are no other surfaces. The two sided and the one-sided surfaces comprise the totality of all surfaces (see, for example, [**15**]).

One-sided surfaces were still very new when Heawood formulated his problem. At the time they were probably regarded as a curiosity rather than a significant phenomenon, which may explain why Heawood, as well as others who should have known better, ignored them. They were mentioned by Heffter, but he felt, for good reasons, that once the coloring problem was resolved for the two-sided surfaces, a certain theoretical connection between them and their one-sided siblings could be utilized to solve the same problem for the latter as well. In the event, Heffter's good reasons notwithstanding, history did not bear him out.

The Euler-Poincaré formula for the one-sided surfaces states that

$$n - b + m = 2 - g$$

holds for maps on $\tilde{S}_g$. From this it follows that every map on this surface can be colored with

$$\tilde{H}_g = \left[\tfrac{1}{2}\left(7 + \sqrt{1 + 24g}\right)\right]$$

colors, and it is of course natural to conjecture that every $\tilde{S}_g$ supports a map that actually requires that many colors. This was indeed done in 1910 by the topologist H. Tietze [**21**], who also pointed out that the complete map $\tilde{M}_6$ of FIGURE 19 was implicit in some work by Möbius [**17**]. Since $\tilde{H}_1 = 6$, this map solves the coloring problem for $\tilde{S}_1$. Every map on $\tilde{S}_1$ can be colored with 6 colors, and some such map in fact requires as many as 6 colors.

In 1934, P. Franklin [**8**] uncovered a surprising fact. He showed that while $\tilde{H}_2 = 7$, every map on the Klein bottle $\tilde{S}_2$ could be colored with no more than 6 colors! This was the first known failure of the conjecture that each of the numbers H_g and $\tilde{H}_g$ is indeed required by some map on the corresponding surface. Once such an exception is found, of course, it becomes reasonable to expect others to occur. By 1943, the work of I. N. Kagno [**13**], H. S. M. Coxeter [**6**], and R. C. Bose [**3**] showed that no such failures could occur on $\tilde{S}_3$, $\tilde{S}_4$, $\tilde{S}_5$, $\tilde{S}_6$, or $\tilde{S}_7$. They did this by constructing arrays for the appropriate complete m-maps. These arrays are very similar to the ones used for the description of maps on the two-sided surfaces. The main difference is that the consistency rule is replaced by:

If the ith row is

$$i) \quad \ldots \quad j \quad k \quad \ldots$$

then the kth row is

either $\quad k) \quad \ldots \quad i \quad j \quad \ldots$

or $\quad k) \quad \ldots \quad j \quad i \quad \ldots$

and the jth row is

either $\quad j) \quad \ldots \quad k \quad i \quad \ldots$

or $\quad j) \quad \ldots \quad i \quad k \quad \ldots .$

In 1954, in the same paper that contained his first major contribution towards the resolution of the Heawood Conjecture, Ringel completely solved Tietze's coloring problem for one-sided surfaces. He accomplished this by producing Heffter-style arrays for the requisite complete maps on all of these surfaces. This was a truly formidable achievement. His proof, by the way, showed that the Klein bottle provided the only exception to the rule that, in general, the number $\tilde{H}_g$ is the solution of the coloring problem on $\tilde{S}_g$. The reason for the earlier resolution of the problem for one-sided surfaces is that the consistency constraint for these is considerably less restrictive than the one for the two-sided surfaces. This pattern recurs frequently in the theoretical study of

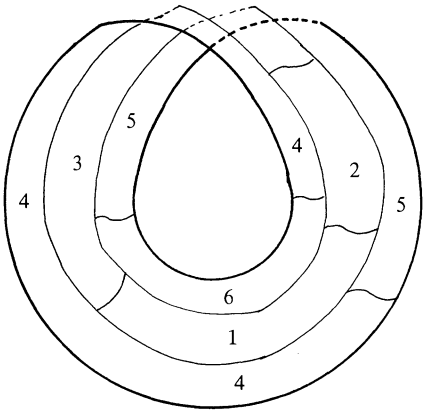


FIGURE 19. **$\tilde{M}_6$, a complete 6-map on the Möbius strip.**

surfaces. Many questions are formulated first in the context of two-sided surfaces, because they are so easily visualized. They are first resolved, however, for one-sided surfaces, despite the fact that the natural habitat of these surfaces is in four-dimensional space.

In 1967 Youngs simplified Ringel's solution by replacing his arrays with current graphs. The complete solution to the problems posed by Heawood and Tietze is now known as the **Ringel-Youngs Theorem**. In 1974 Ringel published his book *Map Color Theorem* [**20**], which contains both the complete details of the solution and an explanation of the underlying theory of current graphs.

I had several reasons in mind when I decided to recount the history of the Ringel-Youngs Theorem. It is one of my favorite theorems, both because it deals with surfaces and because its proof is so rich. Moreover, I felt that its history makes for a good story. Finally, I saw this as a way to bring the reader closer to the way mathematicians actually operate. Problems, be they solved or unsolved, give rise to more problems. We saw the Four Color Problem motivate the Heawood conjecture, and the latter eventually gave rise to Tietze's one-sided analog. The path leading to a problem's solution is often littered with mistakes, such as those made by Kempe and Heawood. It is the good mathematician who can extract useful information from his own and other people's mistakes, and use them as a basis for new investigations. We saw mere notational conventions transformed into crucial breakthroughs, while other promising technical approaches dwindled into blind alleys. The final solution came as a result of fusion of disparate mathematical disciplines and the cooperative efforts of several mathematicians. These are only some of the elements that go into the production of a mathematical proof, but they are also ones that do not appear in the final product.

The author thanks all those readers of earlier drafts of this article for their helpful critical comments.

References

[1] K. Appel and W. Haken, The four color problem, Mathematics Today, ed. L. A. Steen, Springer-Verlag, 1978, pp. 153–180.
[2] N. L. Biggs, E. K. Lloyd, and R. J. Wilson, Coloring maps on surfaces, Graph Theory, 1736–1936, Oxford University Press, 1976, pp. 109–130.
[3] R. C. Bose, On the construction of balanced incomplete block designs, Ann. of Eugenics, 9 (1939) 353–399.
[4] H. Cohn, Conformal Mapping on Riemann Surfaces, Dover, 1967.
[5] R. Courant and H. Robbins, What is Mathematics?, Oxford, 1941.
[6] H. S. M. Coxeter, The map-coloring of unorientable surfaces, Duke Math. J., 10 (1943) 293–304.
[7] G. A. Dirac, Map colour theorems, Can. J. Math., 4 (1952) 480–490.
[8] P. Franklin, A six color problem, J. Math. Physics, 13 (1934) 363–369.
[9] W. Gustin, Orientable embeddings of Cayley graphs, Bull. Amer. Math. Soc., 69 (1963) 272–275.
[10] P. J. Heawood, Map color theorem, Quart. J. Math., 24 (1890) 332–338.
[11] L. Heffter, Uber das Problem der Nachbargebiete, Math. Ann. 38 (1891) 477–508. An English translation of much of this is in reference [**2**].
[12] D. Hilbert and S. Cohn-Vossen, Geometry and the Imagination, Chelsea, 1952.
[13] I. N. Kagno, A note on the Heawood color formula, J. Math. Physics, 14 (1935) 228–231.
[14] A. B. Kempe, On the geographical problem of the four colors, Amer. J. Math., 2 (1879) 193–200.
[15] W. S. Massey, Algebraic Topology: An Introduction, Harcourt, Brace & World, 1967.
[16] J. Mayer, Le Problème des Regions Voisines sur les Surfaces Closes Orientables, J. Comb. Th., 6 (1969) 177–195.
[17] A. F. Möbius, Uber die Bestimmung des Inhaltes eines Polyeders, Ber. K. Sächs. Ges. Wiss. Leipzig Math.-Phys Cl., 17 (1865) 31–68, also Werke, vol. 2, pp. 473–512.
[18] B. Riemann, Grundlagen für eine allgemeine Theorie des Funktionen einer verandliche Grösse, Collected Works, Dover, 1953.
[19] G. Ringel, Bestimmung der Maximalzahl der Nachbargebiete auf nichtorientierbaren Flachen, Math. Ann., 127 (1954) 181–214.
[20] _____, Map Color Theorem, Springer-Verlag, 1974.
[21] H. Tietze, Eine Bemerkungen uber das Problem der Kartenfärbens auf einseitigen Flachen, Jahrsber. Deutsch. Math. Vereinigung, 19 (1910), pp. 155–159. An English translation of much of this is in reference [**2**].
[22] J. W. T. Youngs, The Heawood map-coloring conjecture, Graph Theory and Theoretical Physics, Academic Press, 1967, pp. 313–354.