

ON THE FOUR COLOR PROBLEM

BY

DANIEL I. A. COHEN AND VICTOR S. MILLER

SUMMARY

I. — Prologue	5
II. — A quick trip through the elementary results	7
III. — Introduction to discharging	17
IV. — The classical method of reducibility	19
V. — Block count consistency	40
VI. — Reduction without Kemp Chains; <i>V</i> -Reducibility	53

I. Prologue

As with most of the other topics in Combinatorial Theory, research on the Four Color Problem historically developed outside the mainstream of classical mathematics. This is not to say that the problem was ignored by serious mathematicians. On the contrary, the solution presented here relies heavily on the monumental contributions of such giants as Cayley, Birkhoff and Lebesgue. These few men were daring enough to divert their attention from their more conventional pursuits to a problem whose allure outweighed its questionable social reputation. However, it is quite true that, more than any other major modern mathematical development, this subject is particularly indebted to the invaluable insights of talented amateurs. Its very existence springs from this source. Unlike the Riemann Zeta Conjecture, the Poincaré Conjecture, Fermat's Last Theorem, or the Weil Conjectures, the Four Color Conjecture does not bare the name of its originator, solely because the conjecturer is a man of no other mathematical achievement. This gives the problem a false aura of antiquity like Trisecting an Angle or Squaring the Circle. As it fits into no great man's global program for mathematics and its applications are few and

disperse, the persistent interest in this problem is a measure of its inherent charm.

The dilettante connection has been both a blessing and a predicament for the Four Color Problem. On the one hand the amateurs' freedom from traditional methodological constraints enabled them to be particularly adventurous in their use of novel devices. Unfortunately the unfamiliarity of the techniques along with the unprofessionalism of the investigators occasionally lead to unsound conclusions. The machinery employed was *sui generis*, *ad hoc* and bootstrapped. The notation and terminology were a hodge-podge of miscommunication. The types of arguments presented were so unusual that appropriate levels of scrutiny had not yet been developed. How could one be sure all the relevant cases had been considered? How much detail is required before sufficient rigor is attained. Inaccurate published "proofs" besmirched the entire escutcheon of the emerging discipline of Graph Theory. Bright young mathematicians seeking to build careers in the academic establishment were warded off this problem by conservative mentors not for its difficulty (for a dozen more difficult conjectures have been proven in as many years by aspiring neophytes) but primarily for its tarnished image.

Against this background, the announced solution by Kenneth Appel and Wolfgang Haken, employing a computer calculation beyond man's capability to survey, caused less celebration than grumbling, generated less new research than bogus philosophical essays and contributed to Chromatic Graph Theory less praise than disrepute. A book on this subject by Thomas Saaty and Paul Kainen¹ published just after the announcement of the proof was careful to claim only that "it appears that the problem has been solved" despite the fact this very book itself presents the mathematical portion of the work of Appel and Haken. There is a continual circulation of rumors that errors have been found in the "solution" and an equally unfounded spate of rumors to the effect that somebody somewhere has performed a totally independent check of the Appel and Haken work. The general impression in the mathematical community is that the Four Color Problem is dead — though nobody knows whether four colors are sufficient to color all planar maps. This is not the way mathematics should be.

Even if the Appel and Haken solution is perfectly correct in all the claims that it makes, it is painfully obvious that it is not an *adequate* solution to the problem. For one thing it fails to *explain* why four is the final answer. In mathematics a decent proof is at least an explanation of why the result is true. Appel and Haken leave us with the understanding that all maps can be colored in four colors because the smallest map

¹ *The Four Color Problem : Assaults and Conquest*, McGraw-Hill, 1977.

ON THE FOUR COLOR PROBLEM

that would require five colors must contain some configuration from a certain long and complicated list, each of which possibilities leads to an independent and different contradiction. It is hard to imagine a proof technique that is more unsatisfactory.

It thus behooves the community of Combinatorialists to publish a comprehensive account of a proof of this theorem which can be considered and discussed knowledgeably by all mathematicians and which will enable others to figure what, after all, it is that goes on in map coloring.

A reporter from the New York Times visiting the Institute for Advanced Study when it was newly opened asked Hermann Weyl for an example of the type of advanced study performed in the mathematics department. To illustrate the non-calculational nature of mathematics and to avoid a long and unsuccessful description his own research, Weyl explained the Four Color Problem to the reporter. "I think I see," said the latter, "Red, blue, green and yellow."

Most today would not offer the Four Color Problem as an example of the elegance of mathematics. It is towards the rehabilitation of this small gem that we offer our contribution.

II. A quick trip through the elementary results

Where did the Four Color Conjecture come from? The chain of events which brought it to the attention of mathematicians is that one Francis Guthrie (1831-1899) asked the question of his brother Frederick who brought it to his teacher Augustus De Morgan who discussed it in a letter to Sir William Rowan Hamilton dated October 23, 1852. Arthur Cayley brought it before the London Mathematical Society in 1878² and it soon became of interest to the British mathematical community. Francis Guthrie became a professor of mathematics at the South African University in Cape Town where he presumably found other coloring problems.

It is possible that cartographers actually deduced this result empirically a long time prior to this date. The interest in the problem in the 1850's can then be seen as a sociological phenomenon. The argument goes that before the British school of De Morgan, Hamilton and Cayley no mathematician would consider this type of question to be within his purview of expertise. Euclidean Geometry, Algebra, Number Theory and Analysis comprised all of mathematics. Any problem not approachable by these means was suspect. Graph Theory was in an extended infancy. The

² cf. May, K.O., *The Origin of the Four-Color Conjecture*, 56, *Isis*, 346, 1965 and Ore, Oysten *The Four-Color Problem*, Academic Press, 1967 at xi.

lonesome examples of Euler's solution of the Königsberg Bridge Problem and the definition of the Euler characteristic failed to excite interest in graphical questions as appropriate for mathematicians.

With the development of logical, discrete and applied mathematics, what had heretofore been classical rules-of-thumb from other disciplines, could now be fitted with appropriate mathematical proofs. Hence the question of the possibility of proof of the Four Color Theorem became Mathematics for the first time. By 1840 A.F. Möbius was challenging his students to prove that K_5 is not planar. This they could not do.³ The great weakness in the presumption that the fact that four colors suffice was known to antiquity, is that ancient map-makers did not color their maps by the rules of the problem. There was a reformation of cartography during the 18th century before which maps were often adorned with monsters, lions and swash lines; and neighboring geo-political regions were not generally colored differently. If the Four Color Conjecture was known at all before its articulation by Guthrie it couldn't have been too much earlier.

From 1870 to 1940 progress was made on the conjecture along several divergent lines. The results we summarize below are only those which contributed towards the recent proofs. It may yet turn out to be the case that better solutions can be found by developing some of the old lines of reasoning not included here.

Our presentation is not strictly chronological and the terminology and statements of the theorems have been retroactively modernized.

By a map we shall mean a finite planar graph embedded on the surface of the sphere. The regions defined by the minimal circuits we shall call countries or faces. The graph partitions the entire surface of the sphere into finitely many simply connected countries each of which is to be assigned a color pursuant to the requirement that no two countries that share a common boundary edge can be labeled with the same color. If such a labeling is done using n or fewer colors it is called a legal n -coloring.

For simplicity we shall assume that the graphs that induce the maps are connected,⁴ that no vertex in the graph has degree 1 or 2,⁵ that no one country shares a border edge with itself,⁶ that all countries have at least three neighbors (have three or more bounding edges),⁷ and that no two countries have more than one edge in common.⁸

³ Barnette, David, *Map Coloring, Polyhedra, and the Four-Color Problem*, Dolciani Mathematical Expositions No. 8, M.A.A., 1983.

⁴ Although a simple argument shows that the four-color conjecture has the same truth status for either the connected or the non-connected versions.

⁵ Also a removable stipulation.

⁶ This requirement is crucial.

⁷ Removable.

⁸ Again removable.

ON THE FOUR COLOR PROBLEM

The four-color conjecture then states that every such map is four-colorable. That four colors are required by some maps can be demonstrated by the example of Fig. 1.

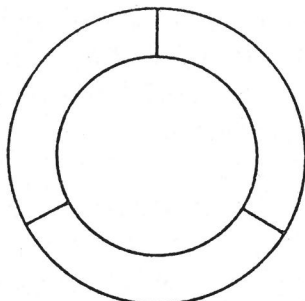


Fig. 1

Definition. — A map is called *regular* if it is formed by a connected planar graph as described above all of whose vertices have degree 3.

THEOREM (Alfred Bray Kempe & William E. Story).⁹ — *If the Four Color Conjecture is true for all regular maps, then it is true for all maps.*

Proof. — If in any map the countries around some vertex were to come together as in Fig. 2, then the map could be made “harder” to color by readjusting the boundaries slightly so as to make more countries border on each other while creating vertices of degree three and decreasing the degree of the target vertex to three. One way of doing this is illustrated in Fig. 3.

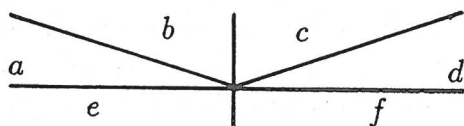


Fig. 2

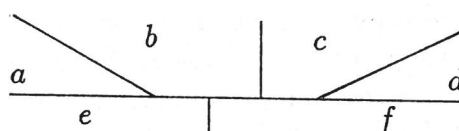


Fig. 3

The rest of the map remains the same. We have only made an adjustment in one small neighborhood. In the first map b did not border on e or f , but it does in the modified version. The same with c and f .

⁹ Kempe was an English barrister who thought that he had solved the problem and had sent it to be published, which it was in “On the Geographical Problem of the Four-Colors,” 2, *Amer. J. Math.*, 193, 1879. Story was the referee to whom the paper was sent. He realized that there were some gaps in the argument which he felt he could remedy. His paper appears as “Note on Mr. Kempe’s Paper on the Geographical Problem of the Four-Colors,” 2, *Amer. J. Math.*, 201, 1879.

We say that the resulting map is “harder” to color because any legal four-coloring of the altered map can be pulled back to give a legal four-coloring of the original map while not all the possible four-colorings of the original map necessarily induce legal four-colorings of the adjusted map.

In such a way we can, vertex by vertex, change an arbitrary map into one where each vertex has degree three. Any legal four-coloring of the map that results from all of this adjustment will correspond to a legal four-coloring of the original map.

If we had a theorem that said that all regular maps could be four-colored this theorem could be applied to the altered map and hence would should show that the original map was four-colorable. \square

We shall make use of two observations which were known before the four-color conjecture was articulated.

THEOREM (Euler, 1752). — *In any map the number of vertices V , minus the number of edges E , plus the number of faces F equals 2 :*

$$V - E + F = 2. \quad \square$$

THEOREM. — *In a regular map $3V = 2E$; therefore*

$$F - E/3 = 2 = F - V/2. \quad \square$$

If the four-color conjecture is false then there are some maps that require more than four colors to color them. Let us place these non-four-colorable maps into classes depending on the number of countries they have (counting the whole surface of the sphere). One of these classes contains the maps with the fewest number of countries. Let us call this the critical class; let us call the common number of countries in each of these maps the critical number; and let us call each of these maps a *critical map*. By definition, any map with fewer than the critical number of countries is four-colorable.

It is traditional in this problem to call the graph edges that bound countries “sides.” Similarly we call a country with three neighbors a triangle, one with four neighbors a square, one with five neighbors a pentagon etc.

THEOREM (Kempe & Percy John Heawood).¹⁰ — *All critical maps contain some countries with five or fewer sides.*

¹⁰ Kempe’s false “proof” remained in the literature unchallenged for eleven years until Heawood pointed out the flaw in “Map Color Theorems,” 24, *Quart. J. Math. Oxford*, ser. 332, 1890. What we present here is what Heawood could salvage from Kempe’s work.

ON THE FOUR COLOR PROBLEM

Proof. — Let f_i be the number of faces in the map with exactly i sides. Since all faces have three or more sides

$$F = f_3 + f_4 + f_5 + \dots$$

Since each edge belongs to exactly two faces we also have

$$2E = 3f_3 + 4f_4 + 5f_5 + 6f_6 + \dots$$

Therefore

$$\begin{aligned} 2 &= F - E/3 \\ &= \left(1 - \frac{3}{6}\right)f_3 + \left(1 - \frac{4}{6}\right)f_4 + \left(1 - \frac{5}{6}\right)f_5 + \dots \\ &= \sum_{i \geq 3} \left(1 - \frac{i}{6}\right)f_i. \end{aligned}$$

Since the left side of the equation is positive the right side must be too. However, the only positive terms on the right side are the first three. Therefore, f_3 , f_4 and f_5 cannot all be zero. \square

Definition. — Any configuration of countries (i.e. a part of a map) that can be shown never to exist in any critical map is called *reducible*.

THEOREM (Kempe). — *No critical map can contain a 3-sided country or a 4-sided country, i.e. the triangle and the square are reducible.*

Proof. — If the triangle ABC (cf. Fig.4) is part of a critical map remove the edge BC and amalgamate the two countries. The map now has fewer countries than a critical map and can therefore be four-colored. Replacing the edge BC returns the original map but it is now improperly 4-colored because the two countries on either side of BC have the same color. This can be remedied by recoloring the country ABC with a color different from any of its 3 neighbors. This gives us a four-coloring of the original map contradicting the hypothesis that it was critical.

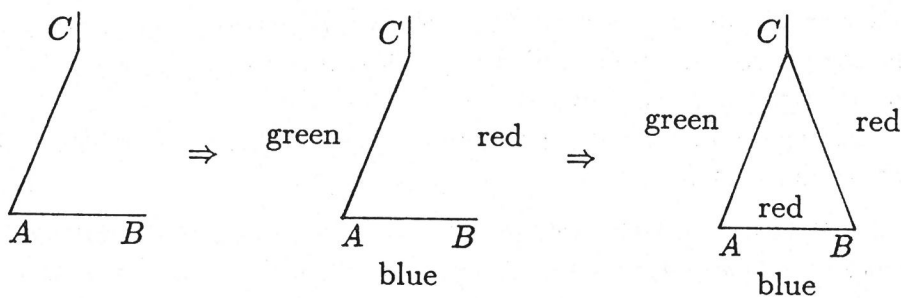


Fig. 4

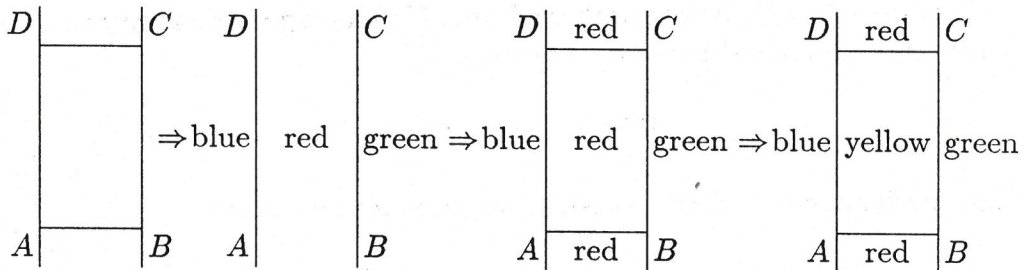


Fig. 5

If we have a square $ABCD$, we do the same by removing sides AB and CD amalgamating three countries, coloring the resulting map, replacing the edges and recoloring the center (cf. Fig. 5).

A problem may arise when we try to amalgamate the three countries. It could be that the top country and the bottom country bordered on each other somewhere else in the map. The amalgamation would then create a country that borders on itself, which is illegal. If the top country and the bottom country did share an edge then the right country and the left country couldn't. We could then amalgamate the three horizontal countries and the argument goes through as before. \square

This last challenge that had to be answered for the amalgamation argument to be made complete is a component of the insecurity mathematicians have with such proofs. How are we to know when all such difficulties have been considered?

Definition. — Any configuration which must exist in all critical maps is called *unavoidable*.

THEOREM (Kempe). — *The pentagon is unavoidable.*

Proof. — If f_3, f_4 and f_5 are not all zero but $f_3 = f_4 = 0$, then $f_5 \neq 0$. In fact we know that $(1 - 5/6)f_5$ is at least two. Therefore f_5 is greater than or equal to 12 in any critical map. \square

If we could show that the pentagon is also reducible, this would provide a contradiction which would prove the non-existence of critical maps and hence prove the whole conjecture. This direct short proof of the theorem has not yet been discovered.

Let us define the dual graph of a map to be the dual graph to its graph of borders (cf. Fig. 6).

THEOREM (Hassler Whitney)¹¹. — *The dual graph of every regular map is a triangulation i.e. every face in the dual has three sides.*

¹¹ Whitney, H., A Theorem on Graphs, **32**, *Ann. Math.*, 378, 1931

ON THE FOUR COLOR PROBLEM

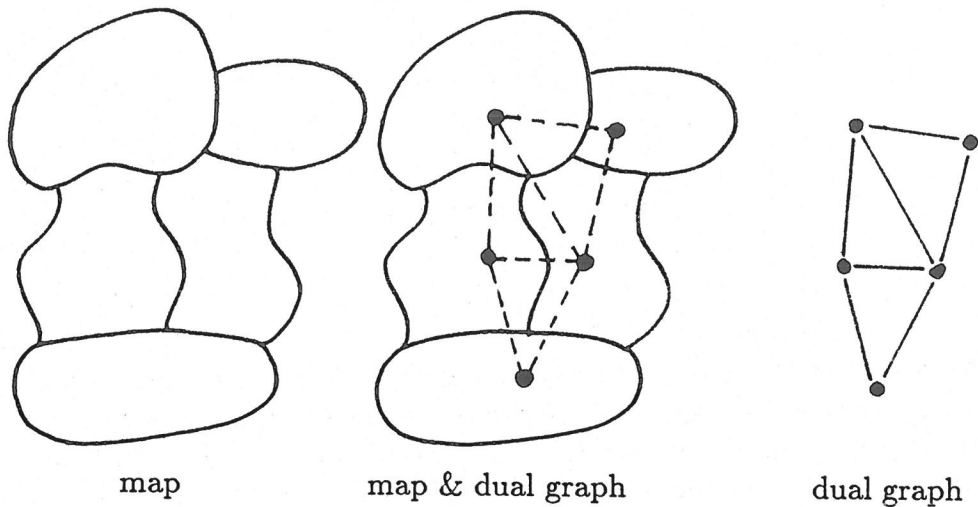


Fig. 6

Proof. — If G is the dual graph of a map M , then every vertex of G corresponds to a face of M , and every face of G to a vertex of M . As the vertices of M all have degree three so the faces of G have three sides. \square

In the dual graph coloring the vertices corresponds to coloring of the faces in M .

THEOREM (Paul Wernicke & Henri Lebesgue).¹² — *Every critical graph must either contain the configuration of two neighboring pentagons, 5 : 5 below, or the configuration of a neighboring pentagon and hexagon, 5 : 6 below.*



Fig. 7

¹² This fact was first published in Wernicke, P., Über den Kartographischen Vierfarbensatz, 58, *Math. Ann.*, 413, 1904. The method of proof we follow is due to Lebesgue, H., Quelques conséquences simples de la formule d'Euler, 9, *J. de Math. Ser. 19*, 27, 1940

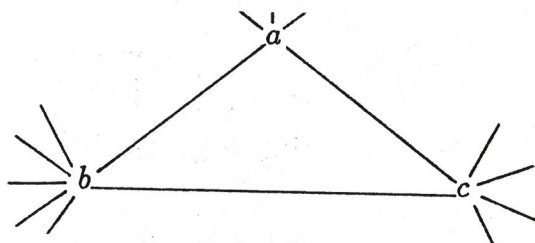


Fig. 8

Proof. — Let us consider the dual graph G of some critical map M . Let us assign to each triangular face f the weight $w(f) = (1/a) + (1/b) + (1/c) - (1/2)$, where a , b and c are the degrees of the vertices of f (cf. Fig. 8).

Let us consider the sum of all the weights of all the faces in the dual graph. Each vertex of degree n will be counted n times (once for each face it is part of) and each time with the weight $1/n$. Therefore, the contribution of each vertex to the total is 1.

Other than the contribution of the vertices, each face contributes an amount of $-1/2$. Therefore,

$$\begin{aligned} \sum_{\text{all } f} w(f) &= \text{number of vertices of } G - (\text{number of faces of } G)/2 \\ &= \text{number of faces of } M - (\text{number of vertices of } M)/2 \\ &= F - (V/2) = 2. \end{aligned}$$

In particular, $\sum_{\text{all } f} w(f) > 0$. Thus G must contain some faces of positive weight, but $((1/a) + (1/b) + (1/c) - (1/2))$ is not often positive when a , b and c are greater than or equal to 5, as is necessary in the dual of a critical map. In fact there are only seven possibilities for faces of positive weight.

Degrees of vertices in G	Weight of face in G
5,5,5	.1
5,5,6	.0667
5,5,7	.0429
5,5,8	.025
5,5,9	.0111
5,6,6	.0333
5,6,7	.0095

The face 6,6,6 has weight 0 and all others have negative weight. For the total weight of all faces to be positive, some (actually many) of the faces of G must be on this list.

Every face on this list contains either 5,5 or 5,6 which means M must contain one of these configurations. \square

ON THE FOUR COLOR PROBLEM

Definition. — A set of configurations such that every critical map must contain some configurations from the set is called an *unavoidable set*.

The last theorem proved that 5 : 5 and 5 : 6 formed an unavoidable set. Lebesgue's proof actually provides a larger unavoidable set, i.e. the set 5 - 5 - 5, 5 - 5 - 6, 5 - 5 - 7, 5 - 5 - 8, 5 - 5 - 9, 5 - 6 - 6, 5 - 6 - 7 is an unavoidable set of configurations.

If we can ever show that *all* the configurations in some unavoidable set are reducible we would then have the contradiction which proves the theorem. Every critical map would have to contain some configuration from the set but no critical maps could contain any of those configurations, therefore, no critical maps could exist.

The method used to show that the triangle and the square are reducible fails to show that the pentagon, or 5 : 5 or 5 : 6 are reducible. A better method for demonstrating reducibility and perhaps a different unavoidable set are required to reach the desired contradiction.

Some unavoidable sets had actually been found earlier, such as

THEOREM (Ph. Franklin).¹³ — *Every critical map contains a pentagon touching at least two faces which are each pentagons or hexagons.* □

The advantage of the method of Lebesgue is that it can be used to produce unavoidable sets of greater cardinality and with larger configurations, and surprisingly this is just what will be needed.

If we look at one of the last configurations in Lebesgue's unavoidable set, the 5 - 6 - 6, we may ask what country borders on both 6's in the triangulation of the dual graph shown in Fig. 9.

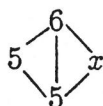


Fig. 9



Fig. 10a



Fig. 10b



Fig. 10c

The total weight of the two triangles in Fig. 9 is

$$\frac{1}{5} + \frac{1}{6} + \frac{1}{6} - \frac{1}{2} + \frac{1}{6} + \frac{1}{6} + \frac{1}{x} - \frac{1}{2} = \frac{1}{x} - \frac{2}{15}$$

In order for the total contribution to the weight-sum of this pair of faces to be positive, x must be 5, 6 or 7.

If a certain critical map only met the unavoidable set of seven configurations at the configuration 5 - 6 - 6 then it must actually contain one of the graphs represented in Fig. 10a, b and c.

¹³ Franklin, Ph., The Four Color Problem, 44, *Amer. J. Math.*, 225, 1922.

So we have the new unavoidable set

- 5 : 5 - 5 5 : 6 - (5) - 6 5 : 6 - 7
- 5 : 5 - 5 5 : 6 - (6) - 6
- 5 : 5 - 7 5 : 6 - (7) - 6
- 5 : 5 - 8
- 5 : 5 - 9

The notation above indicates the following. The first number is the number of sides on the base country. This is followed by a colon. The neighbors of this base country are written after the colon, in (clockwise, though it doesn't matter) order. If the size of any country is unknown a variable must be used in its place. Countries in the second neighborhood of the base country are indicated by being inserted, in parentheses, between the two first-order neighbors they border. Unknown countries not needed for further specification are usually omitted. This means that instead of writing $5 : 5 - 5 - x - x - x$ we could write only $5 : 5 - 5$.

This is a very useful notation. By correct choice of a base country we can denote some configurations five countries thick. For example, the graph represented in Fig. 11 can be denoted by the formula $5 : 7 - (5) - 6 - (6) - 6 - (7) - 5 - (6) - x - (5)$.

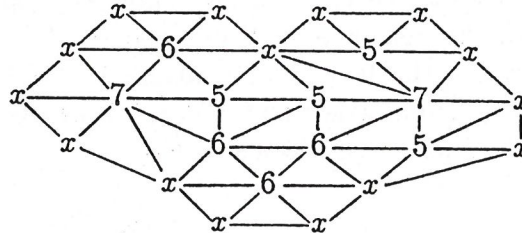


Fig. 11

This notation is not unique.


	$= 5 : 6 - (6) - 5$ $= 5 : 5 - 6 - 6$ $= 6 : 6 - 5 - 5$
--	---

By repeatedly using the same method that we illustrated above on the example $5 : 6 - 6$, the addition of another vertex of the triangulation and the substitution of all of its possibilities, we can produce indefinitely larger unavoidable sets (more configurations) with indefinitely larger configurations (more countries in each). In growing these we may employ considerable discretion in our choice of where to extend each configuration. We could seek to incorporate certain structures and we could avoid certain structures. The reasons for doing so will be made clear below.

III. Introduction to discharging

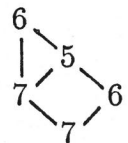
From Euler's formula and Lebesgue's formula we see that in a critical map there is a war between the positive influence of the pentagons and the negative influence of the larger (more sided) countries.

In the last section we indicated that unavoidable sets were made of exhaustive lists of configurations of positive weight. In exercising this argument we must be very careful that the negative effect of one large country isn't balanced off by a collection of different partially positive faces. In the figure below we have positive total weight even though it is comprised of overlapping negative regions. For example, the weight of the configuration 5 : 6 - 7 - 7 is negative,



$$\frac{2}{5} + \frac{3}{7} + \frac{1}{6} - 1 = -\frac{1}{210}.$$

while the weight of the configuration 5 : 6 - 7 - 7 - 6, which can be viewed as two of them overlapping, is positive.



$$\frac{3}{5} + \frac{2}{6} + \frac{4}{7} - \frac{3}{2} = \frac{43}{210}.$$

In order to keep straight which negative influence is counted against which positive influence methods were developed for spreading the negative effect of the majors (countries with more than six neighbors) to negate the positive effects of the pentagons. These redistributions are called discharging. The name comes from the analogous allocations of potential energy in electric circuit diagrams.

Let us rewrite Euler's formula as follows.

$$12 = \sum_{i \geq 5} (6 - i) f_i.$$

And let us assign a charge of $(6 - i)$ to every i -gon. The sum of the charges is 12 but the only positive charge comes from the pentagons. A discharging method is a redistribution of the negative charge of the majors to the pentagons.

The most important set of such methods is due to Heinrich Heesch (1969)¹⁴ who generalized the work of Franklin (1922, 1938) and C.E. Winn

¹⁴ Heesch, H., *Untersuchungen zum Vierfarbenproblem*, Bibliog. Institut, AG, Mannheim, 1969.

(1937, 1940). Heesch's aim was to show how to grow unavoidable sets in which each configuration had properties considered favorable for the possibility of proving reducibility. Discharging is a method of showing that a particular set is unavoidable.

Let us give a discharging proof that $\{5 : 5, 5 : 6\}$ is an unavoidable set.

Proof. — Assuming that a map has neither a $5 : 5$ nor a $5 : 6$ all the neighbors of every pentagon are major. Every 7-gon can have at most three pentagon neighbors without those pentagons themselves being neighbors. The 7-gon has a charge of -1 , distribute $-1/3$ of this to each of the pentagon neighbors it does have. Similarly the 8-gon has a charge of -2 and at most 4 pentagon neighbors. Discharge $-1/2$ from this charge to each of its pentagon neighbors. Discharge the other majors similarly.

Every pentagon is surrounded by majors. From each of its five neighbors it gets a charge of at most $-1/3$. Therefore the most its charge can be after discharging is $1 - (5/3)$, which is negative. Discharging thus leaves the total charge in the map constant, but results in all countries becoming negative. This is impossible. Therefore, there are no critical maps that avoid both $5 : 5$ and $5 : 6$. \square

One of the complexities which renders the Appel and Haken proof mysterious comes from its complicated discharging method. They use more than one stage of discharging with the charges traveling in circuitous directions.

Basically, there is no algorithm known at the moment which can input a set of configurations and decide whether the presented list is or isn't an unavoidable set. For each set we consider we must somehow produce a discharging method which works for it. The Heesch-Appel Haken approach is to stick to one particular discharging method and then to keep modifying the unavoidable set. That is, until we produce one in which all the configurations are reducible. If we have a configuration that is in the unavoidable set, yet which is not reducible by our methodology (a methodology we have not yet presented) we may replace it as above, with another set of configurations which still can be proven to be unavoidable by the discharging argument. If some of these new configurations cannot be shown to be reducible we reiterate. We stop only when we have arrived at an unavoidable set all of whose configurations are reducible. If we carefully follow the reasoning about how a particular set of configurations was produced, tracing through the generating tree, keeping in mind the well defined discharging procedure underlying the process, we may be convinced that the resultant set of configurations is unavoidable and reducible. Q.E.D.

ON THE FOUR COLOR PROBLEM

The difficulty with attacking the Four-Color Problem by this approach is that this constant growing of the unavoidable set may never terminate (either because the conjecture is false or because it somehow cannot be proven this way) or it may terminate theoretically but in some fantastically large number of steps, a number far beyond human capabilities to understand. We have not yet begun investigating the method of proving reducibility for large configurations. We shall see that the complexity of this method grows doubly exponentially with the size of the configurations. Still, in the words of Saaty and Kainen, it "appears" that this has been done.

IV. The classical method of reducibility

In this section, as before, we have paraphrased and restated older results to conform to the definitions and terminology which will be productive for us later.

The amalgamation method invented by Kempe to prove reducibility of the triangle and square has a limited range of applicability. The erroneous attempt at proof published by Kempe contained another idea which when coupled with a brilliant insight of George David Birkhoff¹⁵ (and later developed by Franklin and Winn) became a powerful method for determining reducibility. This method we shall call the classical method of reducibility as it was (essentially) universally adopted by all researchers and was the basis for the first two solutions to the Four Color Problem, those of Appel and Haken¹⁶ and of Frank Allaire.¹⁷

After describing this method in our own terminology we will demonstrate the advantages of the improved method which we have designed that is presented in the next section.

The invention of Birkhoff begins by considering a map on a sphere as having an equator or a ring of countries such that each country in the ring neighbors exactly two other countries in the ring. This ring will divide the map into two hemispheres (figuratively, not geometrically). Let us call the

¹⁵ Birkhoff, G. D., The reducibility of maps, *35, Amer. J. Math.*, 115, 1913.

¹⁶ Appel, K.I. and Haken, W., The existence of unavoidable sets of geographically good configurations, *20, Ill. J. Math.*, 218, 1976; Every planar map is four colorable, Part I : discharging, *21, Ill. J. Math.*, 429, 1977 and with Koch, J., Part II : reducibility, *id.* at 491. For the purposes of this paper we shall assume that these works are substantially correct.

¹⁷ Allaire, F. and Swart, E.R., A systematic Approach to the determination of reducible configurations in the Four-Color conjecture, *25, J. Comb. Th. Ser. B*, 339, 1978, and Allaire, F., Another Proof of the four color theorem, Part I, *Proc. Seventh Manitoba Conf. on Numerical Math. and Computing*, 1977 at 3. And also personal communication. We shall credit this work also as a solution.

M'

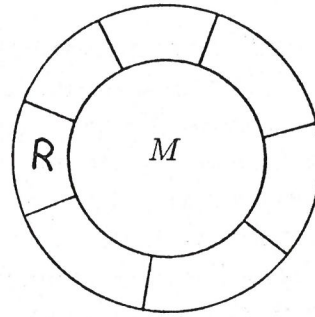


Fig. 12

ring R and call the configuration of countries on one side M and on the other side M' . (cf. Fig. 12).

Speaking heuristically, the influence which the particular details of the coloring of configuration M exerts on the colorability of M' is felt through the ring R . If we may somehow find a different configuration N which exerts the same influence on R but which has fewer countries than M we can “unplug” M and replace it with N creating a new smaller map with the same colorability properties. If $M + R + M'$ is not 4-colorable then neither will be $N + R + M'$ and so $M + R + M'$ can be seen not to have been a critical map since it had more countries than necessary. This will then show the reducibility of M . We will now be very specific about what this all means.

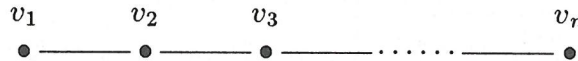
THEOREM (Whitney).¹⁸ — *Let R be a ring of n countries. Consider two 4-colorings of R isomorphic if one can be transformed into the other by a relabeling of the names of the colors. There are exactly*

$$\frac{3^{n-1} + 2 + (-1)^n 3}{8}$$

non-isomorphic 4-colorings of R .

The number of non-isomorphic 3-colorings of R is $\frac{2^{n-1} + (-1)^n}{3}$.

Proof. — Let us discuss the possibility of coloring a path of n labeled vertices with λ colors.



The first vertex can be colored in λ ways. The next in all but one, i.e. $(\lambda - 1)$ ways. The third also in $(\lambda - 1)$ ways, etc... In total the number of total colorings (not worrying about isomorphic colorings) is $\lambda(\lambda - 1)^{n-1}$.

Now let $f(n, \lambda)$ be the number of ways of coloring (again ignoring isomorphisms) the cycle of n labeled vertices in λ colors. If we consider a

¹⁸ Whitney, H., A logical expansion in mathematics, **38**, *Bull. Amer. Math. Soc.*, 572, 1932.

ON THE FOUR COLOR PROBLEM

coloring of the path above that has $\text{color}(v_1) \neq \text{color}(v_n)$ we can connect them by an edge and obtain a colored n -cycle. If we have a colored path such that $\text{color}(v_1) = \text{color}(v_n)$ we can identify these two vertices and produce a colored $(n - 1)$ -cycle. This process produces a bijection between all colored n - and $(n - 1)$ -cycles and all colored paths. Therefore,

$$\lambda(\lambda - 1)^{n-1} = f(n, \lambda) + f(n - 1, \lambda).$$

Reiterating this observation we get

$$-\lambda(\lambda - 1)^{n-2} = -f(n - 1, \lambda) - f(n - 2, \lambda),$$

$$+\lambda(\lambda - 1)^{n-3} = f(n - 2, \lambda) - f(n - 3, \lambda),$$

$$\dots = \dots$$

$$\pm\lambda(\lambda - 1) = f(2, \lambda) - f(1, \lambda),$$

$$= f(2, \lambda) - 0.$$

Adding these gives

$$\begin{aligned} f(n, \lambda) &= \lambda[(\lambda - 1)^{n-1} - (\lambda - 1)^{n-2} + \dots + \pm(\lambda - 1)] \\ &= \lambda \left[\frac{(\lambda - 1)^n + (-1)^n(\lambda - 1)}{1 + (\lambda - 1)} \right] \\ &= (\lambda - 1)^n + (-1)^n(\lambda - 1). \end{aligned}$$

Now we must consider now much duplication we have if we are to count only non-isomorphic colorings. Let us concentrate on the case $\lambda = 4$. If n is odd each coloring uses 3 or 4 colors : if 4, it is counted $4!$ times; if 3, it is counted $\binom{4}{3}3! = 24$ times. Therefore, for n odd the number of non-isomorphic colorings is $f(n, 4)/24$. If n is even, the special coloring 121212...12 is counted $\binom{4}{2}2! = 12$ times, so the number of non-isomorphic colorings is $[f(n, 4)/24 + 1/2]$. The number of non-isomorphic 4-colorings of all n -cycles is then

$$\frac{3^n + (-1)^3}{24} + \frac{1 + (-1)^n}{4} = \frac{3^{n-1} + 2 + (-1)^n 3}{8}.$$

The case for 3-colorings is similar. \square

We will find it very useful to refer to this table of values given by ring size (number of countries). The ring of n countries is referred to as the n -ring (*cf.* Table 1).

Let us denote the set of 4-colorings of the n -ring by C_n . When we depict them we will always start numbering the colors with 1,2... and then employing the lowest unused digit for each new color to appear clockwise around the ring.

For example $C_4 = \{1212, 1213, 1232, 1234\}$ or geometrically, as shown in Fig. 13.

Ring Size	No. of Colorings	Ring Size	No. of Colorings
3	1	11	7,381
4	4	12	22,144
5	10	13	66,430
6	31	14	199,291
7	91	15	597,871
8	274	16	1,793,614
9	820	17	5,380,840
10	2,461		

Table 1

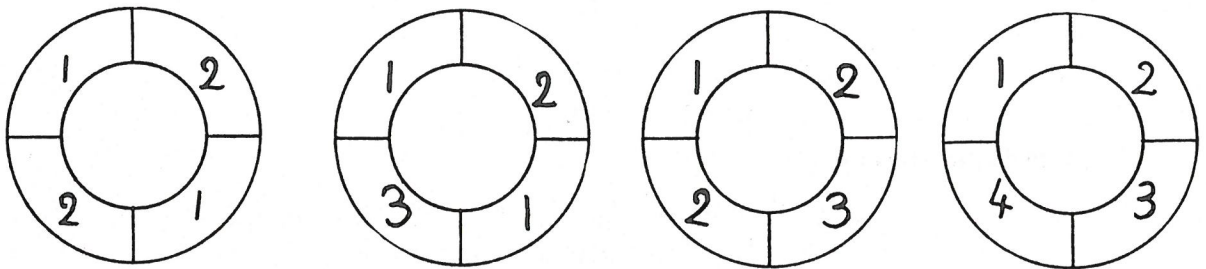


Fig. 13

Let us number these colorings lexicographically $C_1 = 1212$, $C_2 = 1213$, $C_3 = 1232$, $C_4 = 1234$. Then C_5 is

- | | |
|---------------|------------------|
| $C_1 = 12123$ | $C_6 = 12314$ |
| $C_2 = 12132$ | $C_7 = 12323$ |
| $C_3 = 12134$ | $C_8 = 12324$ |
| $C_4 = 12312$ | $C_9 = 12342$ |
| $C_5 = 12313$ | $C_{10} = 12343$ |

Any given specific configuration M is bounded by some determinable ring of countries R . For example $5 : 5$ lives inside a 6-ring and $5 : 6$ inside a 7-ring (see Fig. 14). The Lebesgue configuration $5 : 5 - 5$ also lives inside a 6-ring (cf. Fig. 15).

Given a configuration M inside an n -ring R some of the 4-colorings in C_n extend inward to become consistent colorings of $M + R$ and some do not. For example consider the double pentagon inside a six ring and the two ring colorings $C = 121213$ and $C = 121232$ (cf. Fig 16). Then C_1 can extend inside. Pentagon a borders countries colored 1, 2 and 3 so it must receive color 4. Thus pentagon b borders countries colored 1, 2 and 4 so it must receive color 3. On the other hand C_2 does not extend inside the ring. Both pentagons border countries colored 1, 2 and 3 so both must receive color 4. But since they border each other this is impossible.

ON THE FOUR COLOR PROBLEM

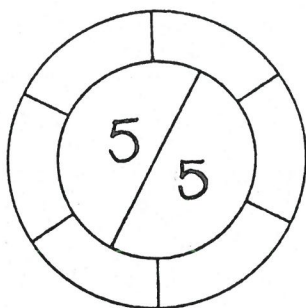


Fig. 14

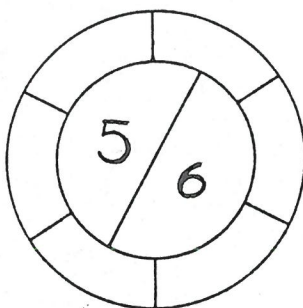


Fig. 15

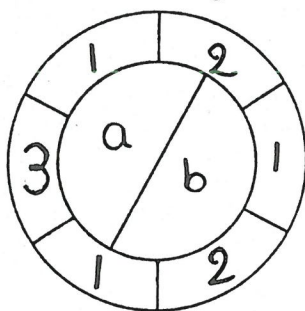
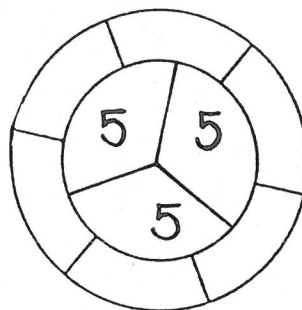
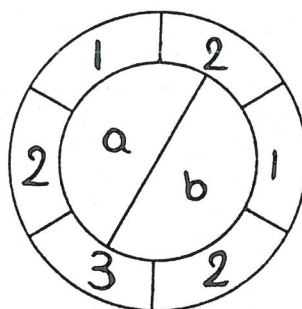


Fig. 16



Definition. — The set of 4-colorings of the ring R which extend inward to the configuration M is called the *scheme* of M , written $S(M)$.

Example. — Consider the square. Of the four colorings of C_4 , three extend to the square and one ($C_4 = 1234$) does not.

$$S(\text{square}) = \{C_1, C_2, C_3\}.$$

Example. — The configuration $5 : 4$ (pentagon-square) lives inside a 5-ring. Six of the 10 colorings in C_5 extend to the $5 : 4$ when imbedded in the ring as shown in Fig. 17.

$$S(5 : 4) = \{C_2, C_3, C_5, C_7, C_9, C_{10}\}.$$

If a configuration M sits inside a ring R in a critical map and if we call the other side of R , M' , then no 4-coloring of R can both extend to M successfully and extend to M' successfully since they could then be tied together to produce a four-coloring of the whole critical map. This means that the scheme of M' must be disjoint from the scheme of M :

$$S(M) \cap S(M') = \emptyset.$$

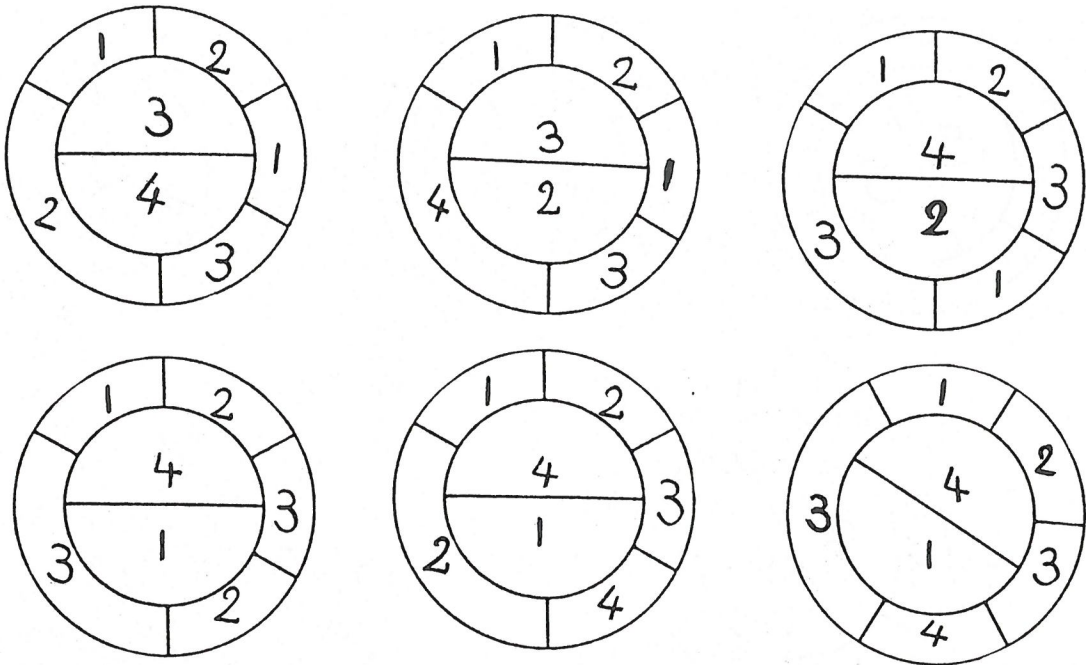


Fig. 17

For example if the 5 : 4 were a part (called M) of a critical map the scheme of its complement (M') would have to satisfy

$$S(5 : 4') \subset \{C_1, C_4, C_6, C_8\}.$$

(Of course, we already know that the configuration 5 : 4 is reducible because it contains a known reducible configuration, namely the square, but the general idea holds for all configurations.) As a consequence of this we have a new method for showing that a configuration is reducible.

THEOREM (Birkhoff). — *If $S(M) = C_n$, then M is reducible.*

Proof. — Let us assume that M sits inside some critical map $M + R + M'$. Since the scheme of the complement must be disjoint from the scheme of the configuration itself we can conclude $S(M') = \emptyset$, i.e. no four coloring of the ring R extends to M' . But in this case $R + M'$ is already impossible to 4-color and M is not needed. Let us construct a new map by taking one country of R and amalgamating all of M into it. The new map is smaller than the one we started with but it cannot be four-colored. Therefore the original map was not critical and, therefore, M is reducible. \square

Another way to use schemes to show that some configuration M is reducible is to find a smaller configuration N which lives inside the same

ON THE FOUR COLOR PROBLEM

size ring and which has a scheme which is a subset of that of M . If M belonged to a critical map $M + R + M'$ and there exists an N such that

$$\text{ring size of } (N) = \text{ring size of } (M)$$

and number of countries of $(N) = \text{number of countries of } (M)$

and $S(N) \subset S(M)$.

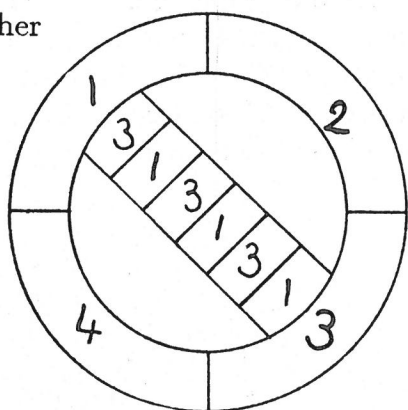
then when we remove M and replace it with N we have the smaller map $N + R + M'$ which is also not four-colorable since $S(N) \cap S(M') = \emptyset$. We will consider this method further shortly. First we must introduce the idea from Kempe's original paper which Birkhoff used in conjunction with ring-coloring. We rephrase this as follows :

THEOREM (Kempe & Birkhoff). — *Not every subset of colorings of C_n can be the scheme of some configuration.*

Proof. — Consider a configuration inscribed in a 4-ring which has coloring $C_4 = 1234$ in its scheme. Let us consider colors 1 and 3 the red colors and colors 2 and 4 the blue colors. Let us fix our attention on one particular extension of C_4 to M .

Without knowing anything about the structure of M we can conclude that there is either a sequence of neighboring red countries through M stretching from the ring country colored 1 to the ring country colored 3 or else there is a sequence of neighboring blue countries through M stretching from the ring country colored 2 to the ring country colored 4 (*cf.* Fig. 18).

either



or

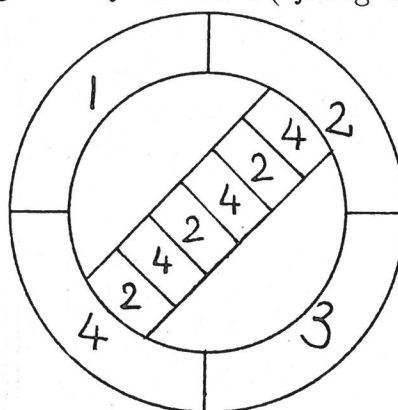


Fig. 18

The reason for this is that if we consider all the red countries attached (through red neighbors) to 1 this component either reaches to the ring country colored 3 or else it is cut off by a string of blues going all around it from 2 to 4 (*cf.* Fig. 19). Clearly both diagonals cannot coexist in the same colored figure.

Let us consider the case in which there exist a red chain from 1 to 3. Here we can reverse all the blue colors below the chain, that is, in the half containing the 4, reverse meaning to change 2's into 4's and 4's into

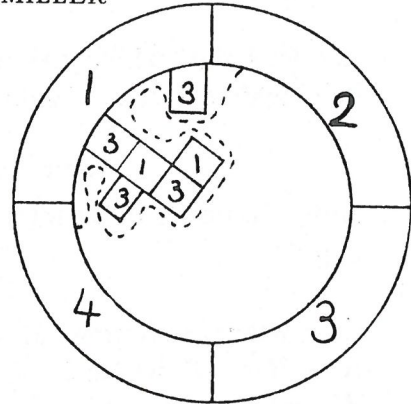
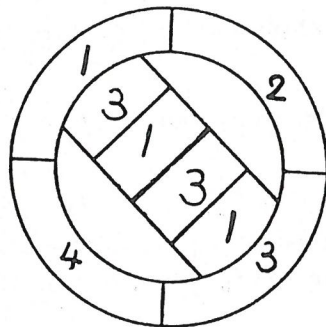


Fig. 19



becomes

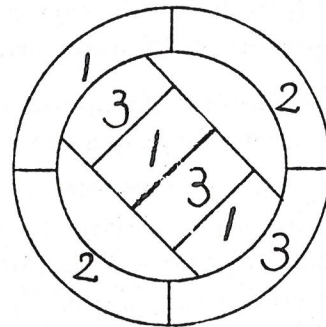
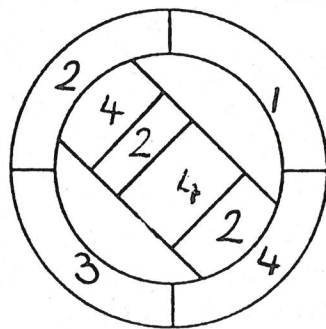


Fig. 20

2's. This will not disturb the blue colors above the red chain and the ring country colored 2 will stay 2 (*cf.* Fig. 20).

What we obtain is a perfect 4-coloring of the configuration and ring in which the ring is colored 1232. Therefore if $C_4 \in S(M)$ and some coloring of $R+M$ has the above red chain then also $C_3 \in S(M)$. On the other hand if there is a blue chain from ring country 2 to ring country 4 then we can reverse 1's and 3's below this chain to obtain a new coloring (*cf.* Fig. 21).



becomes

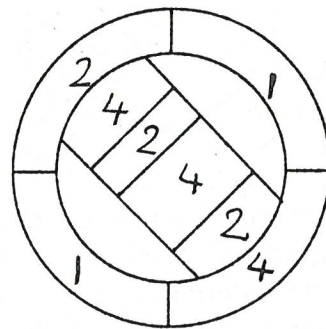


Fig. 21

ON THE FOUR COLOR PROBLEM

If this situation obtains then we can see that the ring coloring 1214 (which is isomorphic to 1213 = C_2) can extend into the ring in a legal four-color fashion. Therefore if $C_4 \in S(M)$ and some coloring of $R + M$ has the blue chain described above then we also know that $C_2 \in S(M)$.

Since $C_4 \in S(M)$ implies that there are some extensions of C_4 to M and since each of these extensions must contain either a red or a blue chain we can conclude that

$$C_4 \in S(M) \Rightarrow [C_2 \in S(M) \text{ or } C_3 \in S(M)]$$

regardless of any properties of M . In particular the set $\{C_1, C_4\}$ which is a subset of C_4 cannot be the scheme of any existing configuration. \square

When put in this fashion we have a whole slew of conditions which must be satisfied by the schemes of existing configurations each in the form of an implication called Kempe implication. These red or blue chains are called Kempe-chains. It is not necessary to group the colors as 1 & 3 vs. 2 & 4. The two other pairings which give us similar implication consequences are 1 & 2 vs. 3 & 4 and 1 & 4 vs. 2 & 3.

For example, the 8-ring shown below can exhibit any of these three structures (cf. Fig. 22).

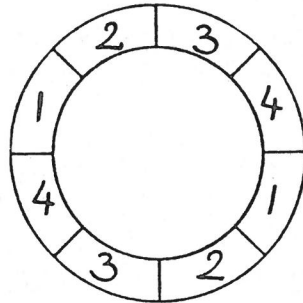


Fig. 22

The codifying of these implications was not done by Kempe but was named for him by Birkhoff since the fundamental idea of reversing two colors in one isolated component of a map to produce a new legal four-coloring comes from Kempe's published false proof. The full set of Kempe implications for the 4-ring can be stated as follows.

Let S be the scheme of some configuration :

$$C_1 \in S \Rightarrow C_2 \text{ or } C_3 \in S$$

$$C_2 \in S \Rightarrow C_1 \text{ or } C_4 \in S$$

$$C_3 \in S \Rightarrow C_1 \text{ or } C_4 \in S$$

$$C_4 \in S \Rightarrow C_2 \text{ or } C_3 \in S$$

This analysis of the 4-ring is essentially the work of Paul Wernicke.¹⁹

It is important to realize that these implications must be satisfied no matter what configuration M sits inside the ring. In particular, these implications must be satisfied by the scheme of the complementary hemisphere of any configuration in a critical map.

Let us use this requirement to provide another proof of the fact that the square is reducible.

Proof. — Suppose there is a critical map which contains a four sided country. We can decompose this map into

$$\text{Square} + 4\text{-ring} + \text{Square}'.$$

Since the colors C_1, C_2 and C_3 all extend to the Square the scheme of the complement of the square, Square' , must satisfy

$$S(\text{Square}') \subset C_4 - S(\text{Square}) = \{C_4\}.$$

Since $S(\text{Square}')$ cannot be C_4 alone, as we know from the Kempe-implications, it must be empty. But this would mean that the square is reducible by the theorem before last. \square

The Kempe-implications for the 4-ring are simple; not so the analogous set of implication for larger rings. For example consider the six-ring : C_6 has 31 colorings

$C_1 = 121212$	$C_9 = 121324$	$C_{17} = 123143$	$C_{25} = 123412$
$C_2 = 121213$	$C_{10} = 121342$	$C_{18} = 123212$	$C_{26} = 123413$
$C_3 = 121232$	$C_{11} = 121343$	$C_{19} = 123213$	$C_{27} = 123414$
$C_4 = 121234$	$C_{12} = 123123$	$C_{20} = 123214$	$C_{28} = 123423$
$C_5 = 121312$	$C_{13} = 123124$	$C_{21} = 123232$	$C_{29} = 123424$
$C_6 = 121313$	$C_{14} = 123132$	$C_{22} = 123234$	$C_{30} = 123432$
$C_7 = 121314$	$C_{15} = 123134$	$C_{23} = 123242$	$C_{31} = 123434$
$C_8 = 121323$	$C_{16} = 123142$	$C_{24} = 123243$	

Let us consider a situation in which C_{20} extends to a configuration inside the ring. Considering colors 1 & 2, red and colors 3 & 4 blue we have two only alternatives, just as with the situation in 4-rings (*cf.* Fig. 23). Therefore $C_{20} \Rightarrow C_{19}$ or C_{13} . We have dropped the repetitive symbol " $\in S$ ".

Now suppose that we call colors 1 & 3 red and colors 2 & 4 blue. There are then five possible distinct internal Kempe-chain structures (*cf.* Fig. 24).

Let us consider the second of these in detail. We can reverse the 1 & 3 component below the 2 & 4 line without changing any other country's

¹⁹ Id. 1904.

ON THE FOUR COLOR PROBLEM

either

or

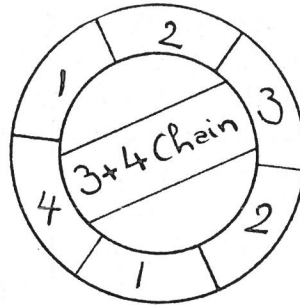
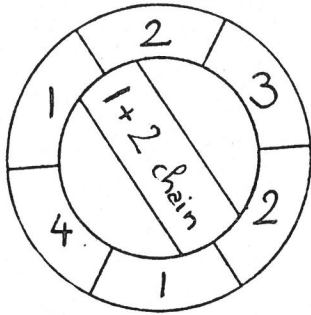


Fig. 23

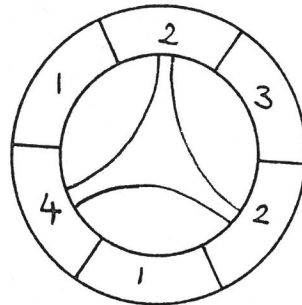
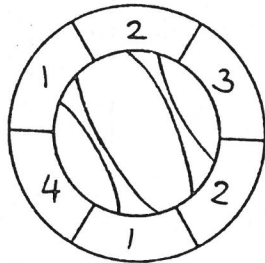
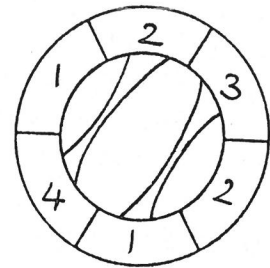
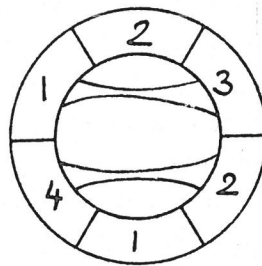
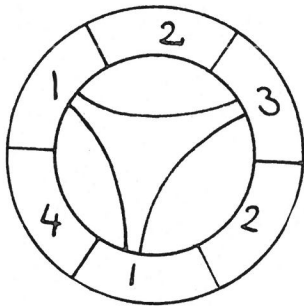


Fig. 24

color. That will produce Fig. 25 a. Alternately we could leave the bottom 1 & 3 component as in the original coloring but reverse the 2 & 4 chain itself. This would produce the legal coloring shown in Fig. 25b. As a third possibility we can simultaneously perform both reversals. This too will produce a legal four-coloring as in Fig. 25 c.

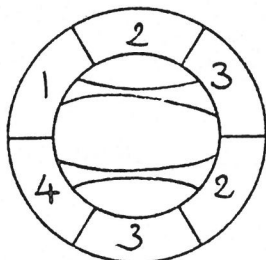


Fig. 25a

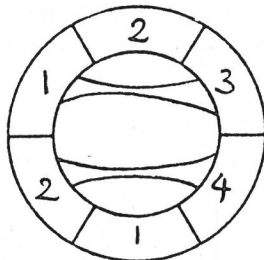


Fig. 25b

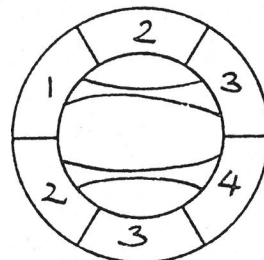


Fig. 25c

In summary we can then say that if coloring C_{20} extends to the inside configuration in a way that has the Kempe-chain structure shown in the second case of the five above then all three new colorings must be included in the scheme of the configuration.

The fact that there are five possible internal structures for Kempe-chain gives us this full implication.

$$C_{20} \Rightarrow [C_1 \text{ and } C_2 \text{ and } C_{18}] \text{ or } [C_2 \text{ and } C_4 \text{ and } C_{22}] \\ \text{or } [C_4 \text{ and } C_{11} \text{ and } C_{27}] \text{ or } [C_{18} \text{ and } C_{25} \text{ and } C_{27}] \\ \text{or } [C_{22} \text{ and } C_{25} \text{ and } C_{30}].$$

If we consider the pairing of colors 1 & 4 vs. 2 & 3 we see that C_{20} cannot yield any internal chains applicable to reversing. Of the three Kempe analyses of this ring coloring we can say that the first was like the 4-ring, the second had the most complexity a 6-ring can have while the third was immutable like a 2-ring.

The thirty-one colorings of the 6-ring fall into three classes: Class 1 (maximal alternating) = the three Kempe analyses are like C_{20} above, Class 2 = all three Kempe analyses reduce to 4-ring-like situations (such as in C_7 in Fig. 26) or Class 3 = the special coloring C_1 alone.

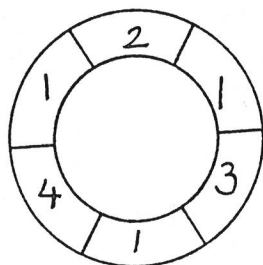


Fig. 26

Coloring C_1 is special because if we interchange the colors 3 & 4 throughout the interior we do not produce any change in the ring coloring itself. A 1 & 3 Kempe analysis is identical to a 1 & 4 Kempe analysis.

Let us examine the Kempe implications for the maximal alternating colorings in more detail. If we have some coloring C_x which has the Kempe implication

$$C_x \Rightarrow \dots \text{ or } [C_p \text{ and } C_q \text{ and } C_r] \text{ or } \dots$$

then when we look at the Kempe implication for C_p we will find

$$C_p \Rightarrow \dots \text{ or } [C_x \text{ and } C_q \text{ and } C_r] \text{ or } \dots$$

and part of the Kempe implication for C_q will be

$$C_q \Rightarrow \dots \text{ or } [C_x \text{ and } C_p \text{ and } C_r] \text{ or } \dots$$

and, similarly, part of the Kempe implication for C_r will be

ON THE FOUR COLOR PROBLEM

$C_r \Rightarrow \dots$ or $[C_x$ and C_p and $C_q]$ or \dots

The reason for this is that all four of these ring colorings correspond to the same red vs. blue Kempe chain structure in the six-ring shown in Fig. 27.

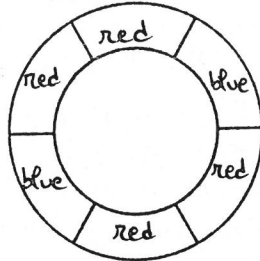


Fig. 27

A maximal alternating 6-ring red vs. blue Kempe structure can be colored in four ways: the two red sections could be coherent or one could be reversed and the two blue sections could be coherent or one could be reversed (if there are three red sections and one blue then the three reds could be coherent or reversed in four ways).

Let us say then that what we are coloring is not the ring but the Kempe chain structure within the ring.

Definition. — We shall say that the set of ways of coloring the chain structure within the ring is a *block*.

A simple list of the blocks will then provide all the maximal alternating Kempe implications. For any maximal alternating coloring, ring-countries one, three and five get one pair of colors (this can happen in four ways since ring-country one must get color 1 and three either agrees or has the other and five either agrees or has the other) and ring-countries two, four and six can be colored with the other pair in four ways too. This means that there will be 16 maximal alternating four-coloring of the 6-ring. Each coloring will be in five blocks (the five Kempe chain patterns). Each block will have four colors (the possible colorings of the chains of the block).

Now,

$$\begin{aligned} &(\text{no. of blocks})(\text{no. colorings in a block}) \\ &= (\text{no. max. alt. colorings})(\text{no. of blocks each is in}). \end{aligned}$$

Therefore, for the 6-ring, the number of blocks is equal to $(16)(5)/(4) = 20$. The twenty blocks for the 6-ring are listed below, where the notation $x : abcd$ is to be read “block number x is composed of colorings C_a, C_b, C_c, C_d .”

1 : 1 2 5 6	2 : 1 2 18 20	3 : 1 3 6 11
4 : 1 3 18 21	5 : 1 5 21 30	6 : 2 4 5 10
7 : 2 4 20 22	8 : 2 6 22 31	9 : 3 4 10 11
10 : 3 4 21 22	11 : 3 10 18 25	12 : 4 11 20 27
13 : 5 6 25 27	14 : 5 10 25 30	15 : 6 11 27 31
16 : 10 11 30 31	17 : 18 20 25 27	18 : 18 21 27 31
19 : 20 22 25 30	20 : 21 22 30 31	

This list then contains all the information in the more complicated Kempe implications. This formulation of the Kempe implications is due to Cohen.²⁰ At the moment this is simply a notation, the true power of defining the blocks will be seen below. The first complete analysis of the Kempe implications of the 6-ring was not done by Birkhoff but by Arthur Bernhart²¹ who developed a substantially different compact schematic representation.

When the Kempe chain analysis makes the 6-ring reduce to a 4-ring-like situation we also see a pattern analogous to blocks in the 4-ring. If we have the Kempe implication

$$C_x \Rightarrow C_y \text{ or } C_z,$$

then there will be a unique other ring coloring C_w with the property that it has the Kempe implication

$$C_w \Rightarrow C_y \text{ or } C_z.$$

Furthermore, we shall then expect to find both

$$C_y \Rightarrow C_x \text{ or } C_w \quad \text{and} \quad C_z \Rightarrow C_x \text{ or } C_w,$$

as Kempe implications.

The reason for this is that given C_x if we have a red chain we can reverse to C_y while a blue chain reverses to C_z while if we made both reverses (even though we could not have both chains) we would produce C_w . This new C_w then, can red or blue reverse back into the same things that C_x does. We shall denote the relationship these four colorings have to one another by writing $x | w = y | z$.

How many of these relationships should we expect to find in the 6-ring? We can turn a 6-ring into a virtual 4-ring by ignoring any two of the six inner-ring-country-borders. We can choose the 2 out of 6 in 15 ways. Once we have the grouping of the ring-countries there is only one way they can

²⁰ Cohen, D. I. A., *Small Rings in Critical Maps*, Ph.D. Thesis Harvard Univ., 1975.

²¹ Bernhart, A.F., Six-Rings in Minimal Five-Color Maps, 69, *Amer. J. Math.*, 391, 1947.

ON THE FOUR COLOR PROBLEM

be colored and one way they can be reversed. Therefore there are exactly 15 4-ring-type equations in the Kempe implications of the 6-ring. They are:

2 26 = 7 19	3 29 = 8 23	4 28 = 9 24
5 15 = 7 14	6 9 = 7 8	10 13 = 9 16
11 12 = 8 17	18 24 = 19 23	20 12 = 13 19
21 16 = 14 23	22 17 = 15 24	25 17 = 16 26
27 28 = 26 29	30 12 = 14 28	31 13 = 15 29

This table along with the previous one summarizes all the information in the many Kempe implications for the 6-ring. We are now ready to duplicate, by our more modern approach, something which Birkhoff did in 1913.

THEOREM (Birkhoff). — *The diamond shaped configuration of four pentagons (shown in Fig. 28) is reducible.*

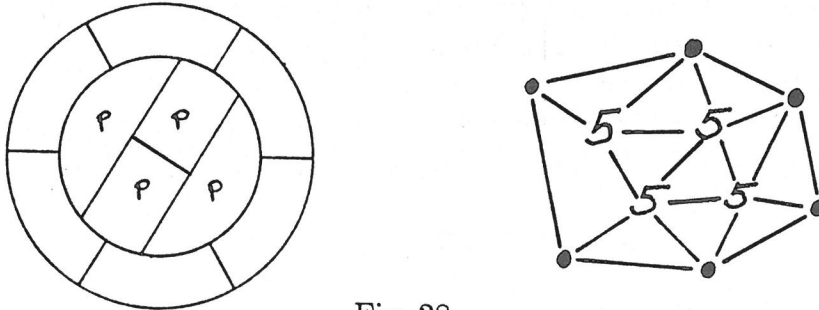


Fig. 28

Proof. — The scheme of this configuration can be calculated to be

$$\{C_2, C_3, C_6, C_8, C_{10}, C_{11}, C_{14}, C_{15}, C_{16}, C_{18}, C_{19}, C_{20}, C_{22}, C_{25}, C_{30}, C_{31}\}.$$

If the diamond is part of a critical map what is the scheme of the configuration on the other side of the 6-ring?

$$\begin{aligned} S(\text{diamond}') &\subset C_6 - S(\text{diamond}) \\ &= \{C_1, C_4, C_5, C_7, C_9, C_{12}, C_{13}, C_{17}, C_{21}, C_{23}, C_{24}, C_{26}, C_{27}, C_{28}, C_{29}\}. \end{aligned}$$

From our table of blocks we see that coloring C_5 is a member of blocks 1, 5, 6, 13 and 14. None of those blocks, however, are totally contained in the list of colorings above. This means that C_5 cannot be in the scheme of diamond' for reasons of Kempe chain analysis.

We now note that a Kempe implication for C_7 says $7 | 14 = 5 | 15$. This means that if coloring 7 is in the scheme then so is coloring 5 or coloring 15. Coloring 15 is not on the list of possibilities because it extends to

the diamond, and coloring 5, while on the list of possibilities, has been eliminated from scheme of diamond' by Kempe considerations of its own.

Next we recall $6 | 9 = 7 | 8$. Since we now know that we cannot have 7 or 8, 9 becomes impossible too. Recalling $10 | 13 = 9 | 16$ eliminating 9 means 13 is also out.

Now $20 | 12 = 13 | 19$ so 12 is out. Since $12 | 11 = 8 | 17$, 17 is out.

Since $17 | 22 = 15 | 24$, 24 is out. Since $18 | 24 = 19 | 23$, 23 is out.

Since $4 | 28 = 9 | 24$, 4 is out. Since $21 | 16 = 14 | 23$, 21 is out.

Since $25 | 17 = 16 | 26$, 26 is out. Since $27 | 28 = 26 | 29$,
both 27 and 28 are out.

This leaves only 1, without enough for any block, therefore 1 is out. Therefore $S(\text{diamond}') = \emptyset$. From which we conclude that the diamond is reducible. \square

When a configuration is determined to be reducible by showing that no subset of the colorings left for its complement can satisfy any Kempe-implication we call this D -reducible. This terminology is due to Heesch though the concept dates from Birkhoff.

There is another possibility. We may begin with a configuration M , and assuming that it is part of a critical map $M + R + M'$, we calculate

$$S(M') \subset C_n - S(M).$$

Now we eliminate colorings out of $C_n - S(M)$ by reason of Kempe-chain implications until we find its maximal Kempe-consistent subset, that is the set of colorings left which among themselves satisfy their Kempe implications. Let us call this the Max-Kempe of M' . If this Max-Kempe is \emptyset we are finished as above, but even if Max-Kempe doesn't vanish entirely it may be so small that we can find a real life configuration N , having the same ring-size as M but with fewer countries, whose scheme is disjoint from Max-Kempe M' . This would imply that if M' existed then we could produce a smaller uncolorable map $N + R + M'$. Such an N is called a *reducer*.

The Kempe-implications are important here to reduce the possible size of $S(M')$ which if it were larger might intersect $S(N)$ for too many N 's.

Notice that $S(N)$ does not have to be contained in $S(M)$ as we required previously. It is possible that $S(N)$ is larger than $S(M)$ but still is disjoint from Max-Kempe.

A configuration for which there exists such a reducer is called by Heesch C -*Reducible*.

Actually Birkhoff's original proof of the reducibility of the diamond did not follow our methodology but showed rather that the diamond is C -reducible. It is valuable for us to examine Birkhoff's approach both because

ON THE FOUR COLOR PROBLEM

it shows how the Kempe implications grew out of the amalgamation argument and because C -reducibility (and reducers) are indispensable to our project.

Proof. — Let us begin by assuming to the contrary that $5 : 5-5-5$ does exist in some critical map, M . Let us look carefully at the configuration and its bounding ring of six countries as shown in Fig. 29.

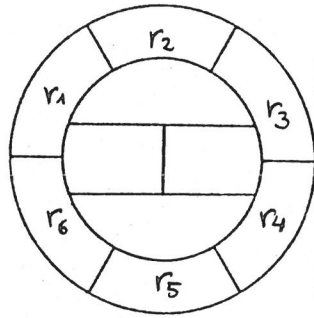


Fig. 29

Let us leave the rest of the map M alone and modify this section as follows. Amalgamate the four pentagons, country r_1 and country r_3 . Let us call the map that is then produced A . The map A has four fewer countries than M and so it must be four-colorable. Color it. Now if we replace the old borders that used to exist in map M and erase the colors left over on the pentagons, we find that the six ring countries have been colored in one of only six possible ways. Each coloring must assign color 1 to r_1 and r_3 but to none of the other r 's. The only colorings that do this are 121232, 121234, 121323, 121324, 121342 and 121343. If all of these colorings of the 6-ring could be extended to colorings of the diamond the from four-coloring A we could produce a four-coloring of M , which would contradict the existence of M .

As we can see from the picture below all of these 6-ring colorings do extend except the first, 121232. This is because pentagons a and b must be colored 3 and 4 since they both border a 1 and a 2. This, however, means that the top pentagon has neighbors of four different colors (*cf.* Fig. 30).

Let us suppose then, that the coloring that M inherits from A is 121232. Everything in M is legally four-colored except the four pentagons. Let us look through the colored diamond' for a 1 & 3 chain from r_1 to r_3 . If such a chain exists then we can reverse colors 2 and 4 in the component of the picture that contains r_2 . This would then induce on the partially colored M the 6-ring coloring 141232. But this, we see below, extends to the diamond as well, providing a four-coloring of the whole map. Therefore M could not be critical. Similarly, if there existed a 1 & 3 chain from r_1

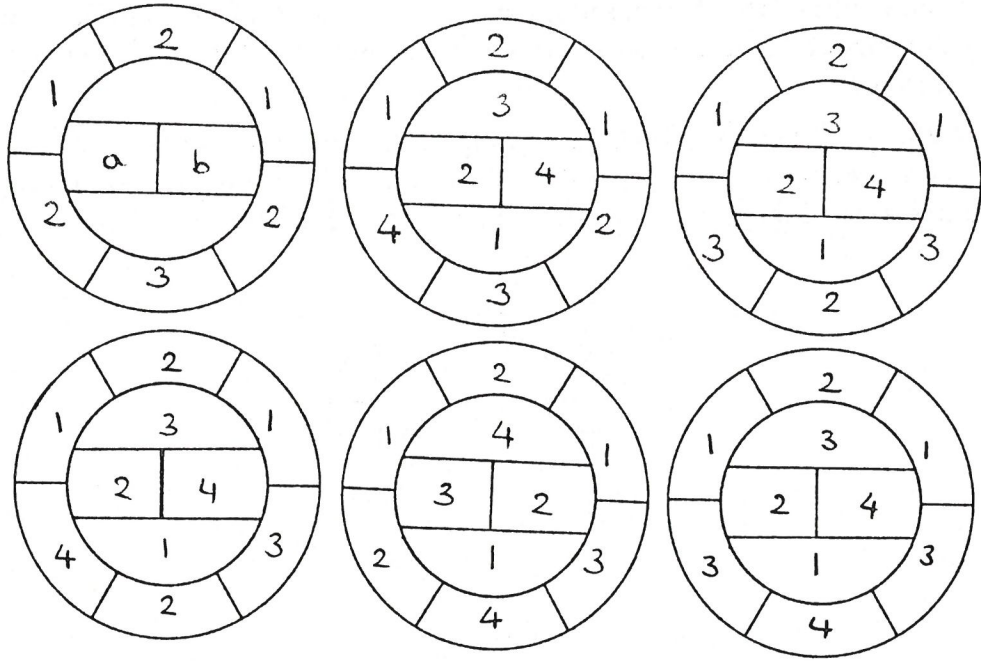


Fig. 30

to r_5 we could reverse 2 and 4 in the component containing r_6 . The ring would then be colored 121234 which we saw above extends to the diamond. If r_1 is not connected to r_3 or r_5 by 1 & 3 chains we can reverse colors 1 and 3 in the r_1 or r_3 component and leave the rest of the ring alone. This would induce the ring coloring 321232 which we see below also extends to the diamond (*cf.* Fig. 31).

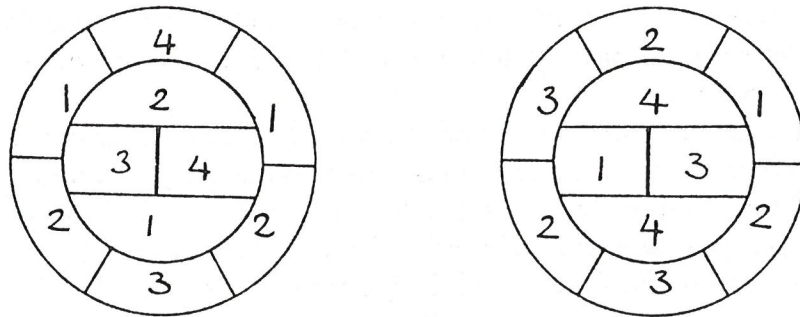


Fig. 31

This covers all possible cases and so the diamond is reducible. \square

In the terminology of all the researchers since Birkhoff this amalgamation formation that limits the cases to be examined by Kempe chains is

ON THE FOUR COLOR PROBLEM

also called a reducer. The particular reducer used above can be depicted by the diagram below where the dotted line means "is amalgamated with" and the solid lines mean "is incident to." (cf. Fig. 32). We will cover some general theory of reducers later.

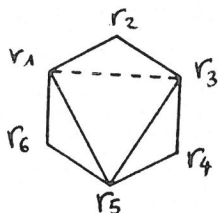


Fig. 32

It may not at first be obvious how our earlier proof is similar to Birkhoff's proof. The analogy is this. In the first proof we showed that the scheme of diamond' (the complement of the diamond) is so small that it does not contain any Kempe-consistent subset. The second (though chronologically prior) proof depended on showing that the scheme of diamond' was so small that it did not contain a Kempe-consistent subset including the coloring 121232. Once this coloring was Kempe-eliminated from the scheme the scheme no longer had any intersection with the scheme of the reducer. This meant that we had found an N that could replace the diamond in the critical map and leave it uncolorable. This means that Birkhoff's proof is like our second paradigm, the C -reducibility model.

C -reducibility is easier to prove than D -reducibility even when D -reducibility is true. That is, if it is easy to put one's hand on the correct reducer. But if one could easily put one's hand on the correct reducer, why not put one's hand on the critical map itself?

The general classical strategy of C -reducibility is this. Start with a configuration M that is bounded by an n -ring R . Check the schemes of all known reducers for the n -ring. This includes those, such as the one above, that involve amalgamating countries from the ring itself, and those that are simply other configurations that live inside the same size ring. If the scheme of M contains the scheme of any reducer then we can replace M by N in the critical map and so M is reducible. If the scheme of M completely misses the scheme of N then $N + R + M$ is not four-colorable and the theorem has been disproved. If the scheme of N is mostly contained in the scheme of M then check to see if what is left of the scheme of N meets what is left of the ring coloring set in any collection of colorings big enough to be Kempe-consistent.

In the above proof, the scheme of diamond' met the scheme of the reducer in only one coloring and none of the Kempe chain consequences of that coloring could be satisfied by $S(\text{diamond}')$. In this particular case this fact followed from the stronger result that no subset of $C_6 - S(\text{diamond})$ could be Kempe-consistent. But there are numerous examples of configurations that are C -reducible but not D -reducible. For example, using practically the same method Franklin²² showed that $6 : 5 - 5 - 5$ is reducible and Winn²³ showed that $7 : 5 - 5 - 5 - 5$ is reducible. The latter configuration sits in an 8-ring as we see in Fig. 33.

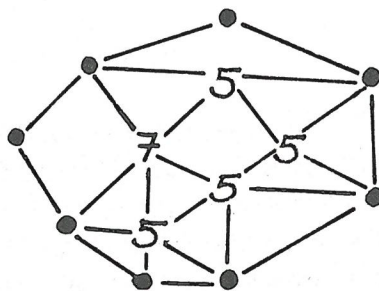


Fig. 33

If we were to write out all the Kempe implications for the 8-ring the calculation would be enormous. It is much easier to locate the correct reducer and then investigate the Kempe implications for the few colorings of the reducer not contained in the scheme of $M = 7 : 5 - 5 - 5 - 5$. To do this we do not even have to know the entire scheme of the configuration M . All we have to do is to check the extendability to M of the few colorings in the scheme of the reducer.

Despite this diminution the task is gargantuan. Work essentially stopped after all the D - and C -reducible configurations inside rings of size ≤ 10 had been discovered until Heesch introduced the possibility of computation by machine. With considerable experience in the frustrating effort to determine reducible configurations Heesch described certain characteristics a configuration might possess which he claimed made it a likely candidate for D - or C -reducibility. This was not a theorem but the pragmatic evidence supported his belief.

Heesch proposed the following project. Start with an unavoidable set. Test each configuration by computer for D - or C -reducibility. If any of them are not reducible replace them in the set by supersets (determined by discharging) which possess as best as possible the properties most

²² Franklin, P., The Four Color Problem, 44, *Amer. J. Math.*, 225, 1922.

²³ Winn, C. E., On Certain Reductions in the Four Color Problem, 16, *J. Math. Phys.*, 159, 1938.

conducive for reducibility. Now go back to the step of testing each new configuration in the unavoidable set. Repeating this process may produce an unavoidable set of reducible configurations.

The properties Heesch recommended be avoided are these : a configuration in which one country borders four ring countries, a configuration in which a cut-point country (one whose removal would disconnect the configuration into disjoint parts) which borders three ring countries, and a configuration with a pentagon that borders on only one other configuration country which itself is a pentagon that also borders on only one other configuration country (called a hanging 5 - 5). These three unfavorable cases are illustrated in Fig. 34.

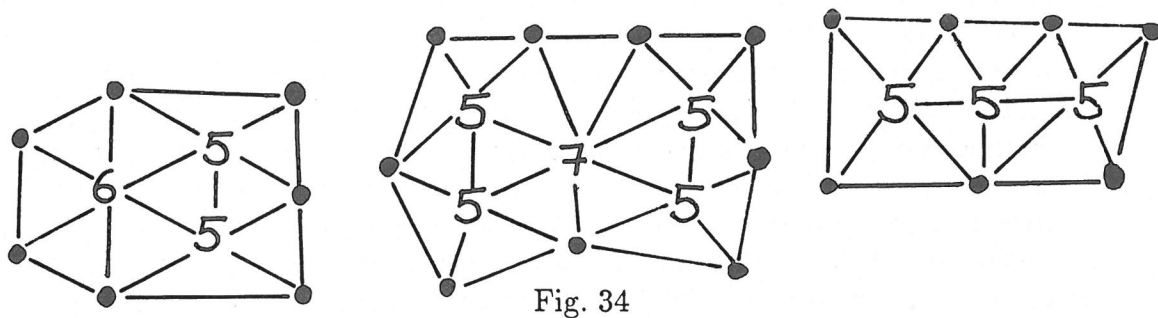


Fig. 34

Haken added to these caveats the stipulation that no configuration should sit in a ring of more than four countries more than itself in size. No such things had been found to C-reduce. And this is for a good reason. Whitney and Tutte²⁴ and more completely Stromquist²⁵ showed how to find Kempe-consistent subsets for the scheme of the complement of such configurations based on uniting copies of the hypothetical Kempe consistent critical complement of the pentagon.

Appel and Haken announced in 1976 that the Heesch plan was realized. Using a modified version of discharging they grew their tree until it had 1936 terminal configurations (the number has since been reduced to 1482) most of which were of such a size that they required 13- and 14-ring boundaries.

Let us now mention briefly on what basis the claims were made that any proof of the Four Color Theorem must be essentially beyond the scope of human surveillance and that the decisions of a computer program must be accepted on faith or at best checked experimentally by other computer programs. (We give no references for this claim because it is so ill-considered as to be beneath citation). A computer is (when operating

²⁴ Whitney, H. and W. T. Tutte, Kempe Chains and the Four Color Problem, *2*, *Utilitas Math.*, 241, 1972.

²⁵ Stromquist, W., Ph.D. Thesis Harvard Univ., 1975.

perfectly) only a Turing machine and a human can duplicate any of its operations, why then should the Appel and Haken proof be beyond surveillance? Simply because the number of steps required to check the calculation and the size of the bookkeeping space needed are astronomical. This is a quantitative matter not a qualitative distinction.

If one were to design, from scratch, a machine to check rapidly whether sets of one hundred thousand elements (the usual size of schemes of configurations inside 14-rings) do or do not contain certain subsets we would re-build the current version of the binary digital computer. These machines can compare bit strings of enormous length in parallel operations at virtually the speed of light. They determine whether one set of ones in a bit string is a subset of another set of ones in another bit string while the same task would take a man hours.

In the Appel and Haken works the number of these comparisons alone ran into the billions. Man, as we know and love him, is not only incapable of duplicating this calculation, he cannot even meaningfully watch it pass before him on a TV screen, since displaying this calculation would slow down the time to greatly exceed the longevity of Methuselah.

Of course, all claims that the calculation could never be shortened, and that new and very different proofs could never be found, are just plain silly and could not be made by competent mathematicians. What we present in the following sections is just such a modification in the method for determining reducibility which renders a huge but surveillable solution along the Birkhoff-Heesch lines. What is more, is that this method was formulated before Appel and Haken made their announcement. Considerable algebraic machinery will be introduced which should rekindle hope for the existence of a short understandable mathematical proof.

V. Block count consistency

Let us return to a consideration of the 4-ring, this time by the diagram shown in Fig. 35.

Let us recall that we are presuming that there is some configuration inside the ring to which the ring coloring is being extended. Every extension of a 4-ring coloring flips into a different extension of a 4-ring coloring. Which coloring it flips into depends on which colored chain was found inside the configuration. If coloring x with a blue chain flips to coloring y then coloring y with a blue chain will flip back to coloring x .

It is also true that not every extension of ring coloring x when extended to the inside must have a blue chain. Some extensions of x may have blue chains and some extensions may have red chains. The red chain extension flips into some other coloring. The number of extensions of coloring 1 with

ON THE FOUR COLOR PROBLEM

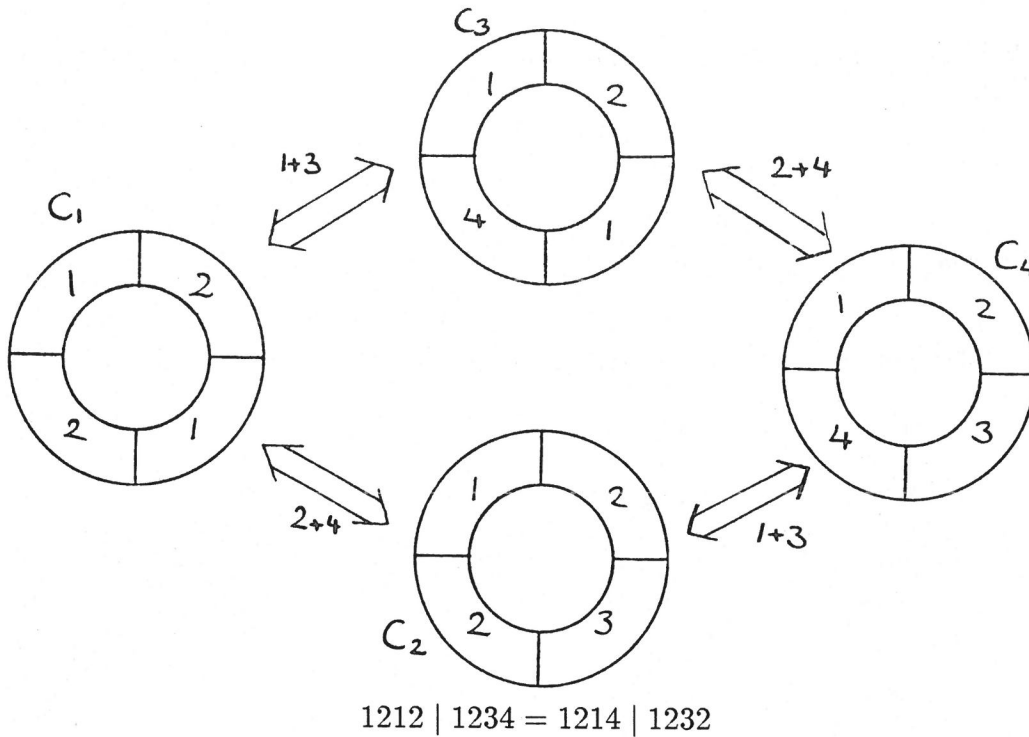


Fig. 35

a 1 & 3 chain all flip into extensions of coloring 3 with a 1 & 3 chain. All the extensions of coloring 1 with a two four chain flip into extensions of coloring 2 with a 2 & 4 chain. All the extensions of coloring 2 with a 1 & 3 chain flip into extensions of coloring 4 with a 1 & 3 chain.

Definition. — Considering the extensions of colorings of a given ring R onto the configuration M inside the ring let us denote the *number of extensions* of coloring C_i by the symbol x_i .

THEOREM (Birkhoff and Lewis,²⁶ the others²⁷ and Cohen²⁸). — *For any configuration inside a 4-ring*

$$x_1 + x_4 = x_2 + x_3.$$

Proof. — Let B_1 be the number of extensions of coloring 1 with a 1 & 3 chain. Let B_2 be the number of extensions of coloring 1 with a 2 & 4 chain. Let B_3 be the number of extensions of coloring 4 with a 1 & 3

²⁶ Birkhoff, G. D. and Lewis, D., Chromatic Polynomials, 60, *Trans. Amer. Math. Soc.*, 355, 1946.

²⁷ As described below.

²⁸ *Ibid.*, 1975.

chain. Let B_4 be the number of extensions of coloring 4 with a 2 & 4 chain. Then

$$\begin{aligned} x_1 &= B_1 + B_2; & x_2 &= B_2 + B_3; \\ x_3 &= B_1 + B_4; & x_4 &= B_3 + B_4. \end{aligned}$$

Therefore, $x_1 + x_4 = B_1 + B_2 + B_3 + B_4 = x_2 + x_3$. □

Example. — Let us consider the double square 4 : 4 (cf. Fig. 36).

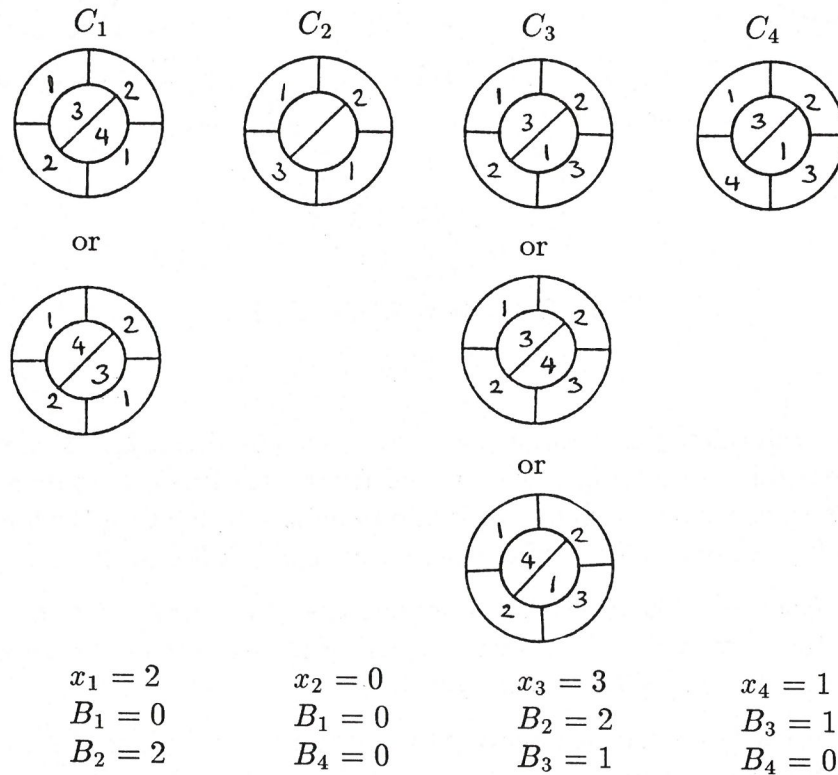


Fig. 36

Let us notice that in this one equation we have summarized all of the information encoded in the four Kempe-implications. If C_1 is in the scheme of M then $x_1 \neq 0$ which means then that $x_2 + x_3 \neq 0$ which means that either C_2 or C_3 must extend to M . Similarly the implications for C_2 , C_3 and C_4 are also superceded by this one equation. This is a powerful equation.

The proof that we have given above of this equation is due to the treatment by Cohen. The characteristic of that approach is that the x 's are not seen as the basic units to be discussed but are themselves the sum of blocks analogous to those introduced in the discussion of the 6-ring.

ON THE FOUR COLOR PROBLEM

Certain sets of x 's, when added together, equal the grand total of all the blocks.

We have prepared this transformation by previously showing how to avoid using the Kempe implications in their raw extensive disjunctive/conjunctive form by concentrating on block structures.

As was indicated above the correct attribution of this discovery must include the work of Birkhoff and Lewis on chromatic polynomials. They realized that the Kempe implications were much too cumbersome to work with. They, however, did not identify the block structures. They completely abandoned the unavoidable set search in favor of investigating more global properties of map coloring (even in more than four colors). They saw that what was necessary was to ask the quintessential combinatorial question. It is not just important to know *whether* a given configuration can be colored but also to know *in how many ways*.

It is this consideration that extracts this problem from the wilds of propositional calculus and embeds it firmly in algebraic structures wherein it can be approached by previously developed mathematical tools.

This idea is inherent in the papers of Birkhoff from 1912 onward. It is nearly expressed in the cited reference (at 416 and at 432), but the authors do not incorporate this equation into a body of machinery that can completely replace the analysis of Kempe implications. It has been said that the replacement of Kempe implications by equations involving extension numbers was considered and rejected by Lewis and then considered and rejected by Arthur Bernhart.²⁹ In his extensive analysis of the 6-ring [1947] and his reduction of the Bernhart diamond $5 : 6 - 5 - 6$ [1948] he does not mention equations at all. He did mention the possibility of using equations for analysis of higher rings at the International Congress of Mathematicians in 1950.

Dick Wick Hall and Lewis³⁰ gave a set of chromatic polynomial formulas for the 6-ring which can, with a little extra work, be specialized to some equations. The Ph.D. theses of Robert Wyman Rector³¹ and Frank R. Bernhart³² include some equations for the 7-ring and indicate how to find more equations for larger rings. It seems that the early 1970's were a time of great action on this problem, especially as far as thinking about extension numbers and equations relating them. This research was done independently and in semi-isolation (Stromquist was in touch with both the Harvard and Kansas groups).

²⁹ This was told to us by his son Frank Bernhart.

³⁰ Hall, D. W. and Lewis, D., Coloring Six Rings, 64, *Trans. Amer. Math. Soc.*, 184, 1948.

³¹ Rector, R.W., Fundamental Linear Relations for the Seven-ring, Ph.D. Thesis, 1973.

³² Bernhart, F.R. in Topics in Graph Theory Related to the Five Color Conjecture, Ph.D. Thesis, Kansas State Univ., 1974.

However, we must note one thing about these other researchers. They all failed to develop a method of equations that could be used to *completely* replace Kempe analysis. Their equations did carry with them much of the information that was in the Kempe-chain implications but they did not capture all of it. Their investigations had therefore to be a mixture of equation analysis and then Kempe analysis to kill the last unattached colorings in the scheme of the complement. They could not prove that their equations carried the full information of the Kempe chains because they did not begin by attaching extension numbers to the block numbers themselves.

It is a primary advance of the work of Cohen that he introduced the block-count numbers, the B 's. The Cohen equations do not only relate x 's, but x 's and B 's. Instead of simply equating sets of x 's, by using the B 's it is possible to prove that all of the information in the Kempe chain implications can be incorporated into a system of equations of the form

$$\begin{aligned} x &= B + B + \cdots + B; \\ x &= B + B + \cdots + B; \\ &\dots\dots\dots \end{aligned}$$

We shall prove this presently.

We have seen that the 4-ring admits only two possible Kempe chain structures: a chain from ring country 1 to ring country 3 or else a chain from ring country 2 to ring country 4. We have also seen that the 6-ring can sometimes look like a 4-ring or else (in its maximal alternating form) it allows five different internal chain structures. It is interesting to note that the 5-ring cannot look like a 6-ring in this respect but only like a 4-ring (or an immutable 2-ring).

The 7-ring also has at most 5 possible structures for internal chaining. But the 8-ring has more (*cf.* Fig. 37).

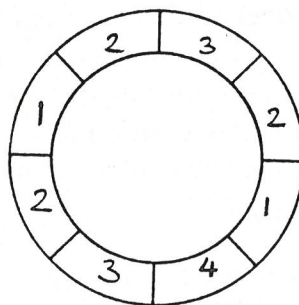


Fig. 37

ON THE FOUR COLOR PROBLEM

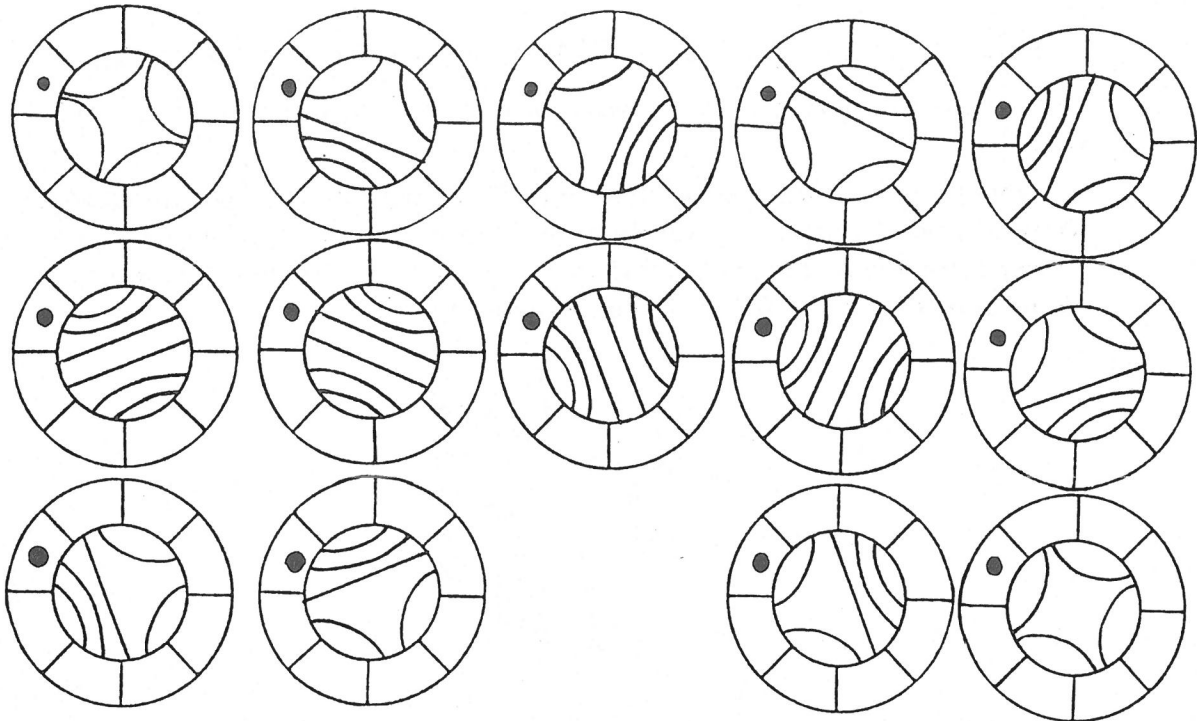


Fig. 38

There are 14 possible ways we can find 1,3-chains (cf. Fig. 38) versus 2,4-chains inside the ring in Fig. 37.

It is easy to see that the $2n$ and the $(2n + 1)$ -ring admit the same possible structures since only with rings of even size can we have a strict alternation around the ring between the feet of the chains.

Definition. — A pattern of coloring an n -ring where all the odd numbered countries are colored (1 or x) while all the even numbered countries are colored (2 or y) will be called an *alternating coloring*.

Note that if a coloring is alternating with respect to 1 & x versus 2 & y then it is not alternating with respect to the other Kempe analyses. Some colorings are not alternating with respect to any chains. One simple case of this is colorings in which 1 neighbors 3 and 4 somewhere around the ring.

Obviously the $2n$ -ring admits 4^{n-1} such alternating colorings since of the odd numbered countries the first gets a 1 and the rest a (1 or x) in 2^{n-1} ways, while of the even numbered countries the first gets a 2 and the rest get (2 or y) in the same number of ways.

THEOREM (F.R. Bernhart). — *There are $\frac{1}{n+1} \binom{2n}{n}$ different structures possible for the Kempe chains in any alternating coloring of a $2n$ -ring.*

Proof. — The Catalan sequence is well known to count the number of ways $2n$ points on a circle can be connected with n non-intersecting arcs. If we consider the arcs to be the borders of the Kempe-chains we see they count all such structures. See, for example, Fig. 39.

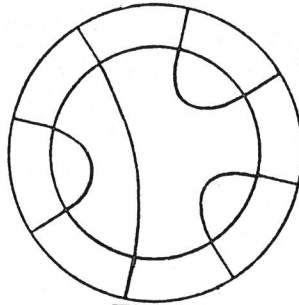


Fig. 39

As we have seen a Kempe-chain need not be a simple sequence of neighboring countries but might have several feet on the ring (i.e. the connected component of those two colors might include more than one of the ring countries). Even so there are always the same numbers of reversals possible with a given alternating coloring and a given Kempe chain structure, whether the chains are simple arcs or a complicated patterns of nested many legged regions. This is because the Kempe chain boundary lines dissect the pie into the same number of regions no matter how they are drawn.

Using the same analogy to the arcs above we can show

THEOREM (Whitney and Tutte). — *Any alternating coloring in a $2n$ -ring has (counting isolated countries as reversible chains) $(n+1)$ chains and 2^{n-1} colorings in the equivalence class under Kempe inversion.*

Bernhart attributes some of his results to unpublished work by his father, A. F. Bernhart. This last theorem was probably known to Birkhoff and Lewis.

Definition. — Let us call an equivalence class of colorings of a configuration M and its bounding ring R a *block* if they can be flipped into each other by reversing the Kempe chains.

Although we have defined blocks on the colorings of the ring and the configuration together we can restrict our attention to the ring coloring alone. A set of ring colorings then belong to the same block if the same

ON THE FOUR COLOR PROBLEM

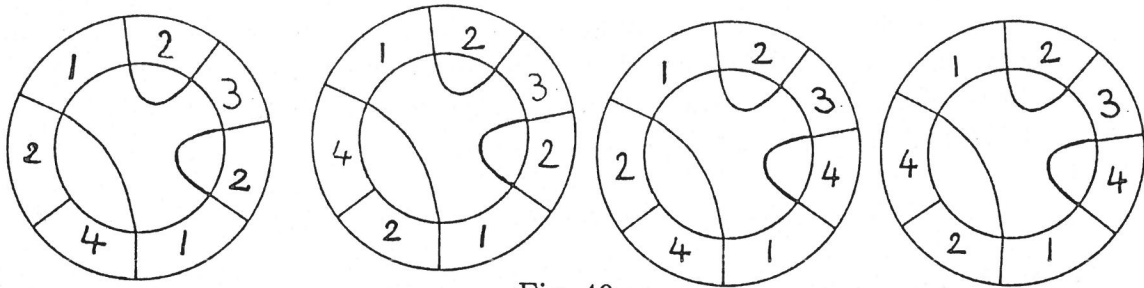


Fig. 40

one inner Kempe chain structure will allow us to flip among them. Fig. 40 shows a block in a 7-ring.

Note again that the $(2n + 1)$ -rings must be treated like $2n$ -rings by pairing two consecutive countries (in any of $(2n + 1)$ ways). The concept of blocks is analogous to the chromodendron of Whitney and Tutte. However, the latter was never related to problems of enumeration.

As a result of the last two theorems we have

THEOREM (new).

(a) There are exactly $\frac{2^{n-1} \binom{2n}{n}}{n+1}$ blocks for the $2n$ -ring, and,

(b) There are exactly $\frac{(2n+1)2^{n-1} \binom{2n}{n}}{n+1}$ blocks for the $(2n+1)$ -ring.

Proof. — This result follows the same way we calculated the number of blocks for the 6-ring. The number of blocks is equal to the number of alternating colorings for the ring, times the number of Kempe chain structures per coloring, divided by the number of colorings per block, i.e.

$$\frac{4^{n-1} \left(\frac{1}{n+1} \right) \binom{2n}{n}}{2^{n-1}} = \frac{2^{n-1}}{n+1} \binom{2n}{n}. \quad \square$$

Remember that every alternating coloring in a $2n$ -ring belongs to exactly $\left(\frac{1}{n+1} \right) \binom{2n}{n}$ blocks, and each of its extensions reflects one of these blocks.

Let us define the function $f(n)$ as follows

$$f(2n) = \frac{2^{n-1} \binom{2n}{n}}{n+1}; \quad f(2n+1) = \frac{2^{n-1}(2n+1) \binom{2n}{n}}{n+1}.$$

Definition. — Given a configuration M in a ring R . Let C_i be any coloring of R . Let the blocks that C_i belongs to be $B_1, B_2, \dots, B_{f(n)}$. Let the numbers $b_i(M)$ be the number of extensions of C_i on R which extend to M and belong to block B_i . Clearly

$$x_i = b_1(M) + b_2(M) + \dots + b_{f(n)}(M).$$

If C_1 and C_2 are two alternating colorings belonging to block B_4 then the number of extensions of C_1 which belong to B_4 is the same as the number of extensions of C_2 which belong to B_4 . Therefore there is only one *block count* variable b_i for each block B_i common to all of its associated colorings.

Therefore there are exactly $f(n)$ block count variables and all extension numbers of alternating colorings can be written as sums of these.

$$b_i(M) = b_i \text{ for all } M.$$

Definition. — The block-count equations for the alternating colorings from the $2n$ -ring are the 4^{n-1} equations of the form

$$x_i = b_{j_1} + b_{j_2} + \dots$$

in which each b reflects the number of extensions of each block applicable to C_i . There are $f(n)$ terms in the right hand side of each equation.

Example. — The 4 four-colorings of the 4-ring include $4^{2-1} = 4$ alternating colorings C_1, C_2, C_3, C_4 . There are

$$f(4) = \frac{2^{2-1} \binom{4}{2}}{3} = 4$$

blocks each with $2^{2-1} = 2$ colorings in them.

$$B_1 = \{C_1, C_3\}; \quad B_2 = \{C_1, C_2\}; \quad B_3 = \{C_2, C_4\}; \quad B_4 = \{C_3, C_4\}.$$

The block count equations are

$$x_1 = b_1 + b_2; \quad x_2 = b_1 + b_4; \quad x_3 = b_2 + b_3; \quad x_4 = b_3 + b_4.$$

From this we can deduce the equation

$$x_1 + x_4 = x_3 + x_2,$$

but little else.

ON THE FOUR COLOR PROBLEM

Whenever a coloring is reduced, through some Kempe analysis to a 4-ring we leave the one equation alone, we do not add block count variables.

Example. — The 31 four-colorings of the 6-ring include $4^2 = 16$ alternating colorings :

There are $2^{3-1} \binom{6}{3} = 20$ blocks, each with $2^{3-1} = 4$ colorings in them. We have listed the blocks for the 6-ring above. The 6-ring block-count equations are

$$\begin{array}{ll}
 x_1 = b_1 + b_2 + b_3 + b_4 + b_5 & x_{18} = b_2 + b_4 + b_{11} + b_{17} + b_{18} \\
 x_2 = b_1 + b_2 + b_6 + b_7 + b_8 & x_{20} = b_2 + b_7 + b_{12} + b_{17} + b_{19} \\
 x_3 = b_3 + b_4 + b_9 + b_{10} + b_{11} & x_{21} = b_4 + b_5 + b_{10} + b_{18} + b_{20} \\
 x_4 = b_6 + b_7 + b_9 + b_{10} + b_{11} & x_{22} = b_7 + b_8 + b_{10} + b_{19} + b_{20} \\
 x_5 = b_1 + b_5 + b_6 + b_{13} + b_{14} & x_{25} = b_{11} + b_{13} + b_{14} + b_{17} + b_{19} \\
 x_6 = b_1 + b_3 + b_8 + b_{13} + b_{15} & x_{27} = b_{12} + b_{13} + b_{15} + b_{17} + b_{18} \\
 x_{10} = b_6 + b_9 + b_{11} + b_{14} + b_{16} & x_{30} = b_5 + b_{14} + b_{16} + b_{19} + b_{20} \\
 x_{11} = b_3 + b_9 + b_{12} + b_{15} + b_{16} & x_{31} = b_8 + b_{15} + b_{16} + b_{18} + b_{20}
 \end{array}$$

The 15 other colorings of the 6-ring all reduce to 4-ring structures under Kempe analysis as do the alternating colorings when we consider different pairs of colors.

The number of block count equations which come from 4-ring structures in the 6-ring is :

$$\begin{aligned}
 & 15 \text{ (one from each alternating coloring except 121212)} \\
 & \quad + 45 \text{ (3 from each non-alternating coloring)} \\
 & \quad \text{divided by 4 (we have counted each equation 4 times)} \\
 & = 15.
 \end{aligned}$$

They are

$$\begin{array}{lll}
 x_2 + x_{26} = x_7 + x_{19} & x_{10} + x_{13} = x_9 + x_{16} & x_{22} + x_{17} = x_{15} + x_{24} \\
 x_3 + x_{29} = x_8 + x_{23} & x_{11} + x_{12} = x_8 + x_{17} & x_{25} + x_{17} = x_{16} + x_{26} \\
 x_4 + x_{28} = x_9 + x_{14} & x_{18} + x_{24} = x_{19} + x_{23} & x_{27} + x_{28} = x_{26} + x_{29} \\
 x_5 + x_{15} = x_7 + x_{14} & x_{20} + x_{12} = x_{13} + x_{19} & x_{30} + x_{12} = x_{14} + x_{28} \\
 x_6 + x_8 = x_7 + x_8 & x_{21} + x_{16} = x_{14} + x_{23} & x_{31} + x_{13} = x_{15} + x_{29}
 \end{array}$$

Cohen [1975] describes the complete system of block-count equations for the 6- and 7-ring. The 7-ring has 91 color variables x_i , 140 block variables b_i , 91 equations of the form

$$x = b + b + \dots +$$

and 35 equations of the form

$$x + x = x + x.$$

THEOREM (Cohen).

1) *The block count equations have a solution in non-negative integers corresponding to every existing configuration $M + R$.*

2) *Given any solution to these equations in non-negative integers the extendable colorings (those C_i for which $x_i = 0$) form a set satisfying all Kempe implications.*

3) *There exist sets of colorings C_i , which are consistent with all Kempe-implications but which cannot lead to a solution of the block-count equations in which all corresponding x 's are non-zero.*

To prove the third point an example is given³³ of a Kempe-consistent set of colorings of the 7-ring which when approached by equations can only have solutions in which some x 's must be 0 even though the corresponding C 's are assumed to extend. In other words :

All Block-count consistent sets are Kempe consistent but not all Kempe-consistent sets are Block-count consistent. We paraphrase this by saying block-count consistency dominates Kempe-consistency. Using x 's alone it is not possible to prove this theorem. This is why the independent discovery of the x 's and the invention of the b 's by Cohen is vitally important.

The other authors who have set about to find sets of x 's with equal sum (similar to the 4-ring equation) have in essence found sets of x 's such that their total represents all the blocks once each.

Definition. — Let us define $B = \sum b_i$ summed over all i .

Let us call a set of x 's *exhaustive* if their sum totals B .

For example for the 6-ring we have

$$x_1 + x_3 + x_4 + x_{25} = b_1 + b_2 + \dots + b_{20}$$

and

$$x_1 + x_{10} + x_{22} + x_{27} = B.$$

Therefore

$$x_3 + x_4 + x_{25} = x_{10} + x_{22} + x_{27}.$$

One important point to be said for block-count consistency is that it dominates Kempe-consistency. The set of equations that we get by equating all exhaustive sets does not.

THEOREM (Donald Coppersmith).³⁴ — *The block variables b_i cannot be recovered from equating exhaustive sets of x 's, i.e. block-count consistency dominates the set of equations formed from exhaustive sets alone.*

³³ Thesis at 156.

³⁴ Personal communication, 1979.

ON THE FOUR COLOR PROBLEM

To use block-count consistency as a method for demonstrating reducibility we proceed as follows.

- 1) Beginning with the configuration M in the ring R calculate $S(M)$.
- 2) In the general block-count equations for the n -ring set $x_i = 0$ if $C_i \in S(M)$. What we are going to study are the x 's representing the scheme of the possible complement, M' , in a critical map.
- 3) Find all solutions to the block-count equations. If there is only the trivial solution, all $x = 0$, then M is D -reducible.
- 4) If there is a non-trivial solution but it is so small that a reducer N can be found such that what is left for $S(M')$ is disjoint $S(N)$, then M is C -reducible.

All that is required here is linear algebra to solve systems of homogeneous linear equations in non-negative integers. This is a very standard process and does not require ad hoc computer programming since there exist many standard packages. However the calculation is still lengthy. It is surveyable but just so.

One philosophical advantage is that once a computer has determined that a configuration is, say, D -reducible it can then print out a mathematical proof of this fact which can be checked by anyone. The steps of this proof are of the following types

Equation 1234 now reads

$$0 = x_7 + x_{18}, \text{ therefore } x_7 = x_{18} = 0.$$

Equation 3783 now reads

$$x_4 = x_4 + x_{21}, \text{ therefore } x_{21} = 0,$$

and so on.

At this stage it would be possible for the computer to print out a mathematical proof of the fact that the system of equations left for the complement is degenerate. Children could then check this proof simple step by simple step. However, even this much exertion is unnecessary.

THEOREM (Nering).³⁵ — *Given a system of homogeneous linear equations that has no solution in non-negative integers except the trivial one there exists a set of constants k_i such that if the i -th equation is multiplied by k_i and the equations added together the result will be one equation saying that the total of all the variables is 0.*

This result is exactly what we need since it reduces the playing around with the equations to one step. Once we start with our system of equations and reduce them to one total equaling 0 we can conclude that each x

³⁵ Personal communication based on notes for a forthcoming book, 1981.

individually is 0 (they are all non-negative since they count something). The fact that the coefficients come to us from a computer program is irrelevant to the proof. As far as mathematical validity is concerned we may have just guessed to add the equations up in this fashion. The computer is totally invisible to the process.

Let us illustrate this method of Block-count reducibility on the Birkhoff diamond. We saw above that the scheme of the diamond includes the following colorings :

2, 3, 6, 8, 10, 11, 14, 15, 16, 18, 19, 20, 22, 25, 30, 31.

All of the corresponding x -numbers must be 0 when we count the extensions of these colorings to diamond'. If the sum of five block numbers is 0 then all five block numbers are 0. The b 's that must be 0 are :

1, 2, 3, ... , up to 20.

These are all the blocks that are. When these are zeroed out the system of block-count equations becomes :

$$x_1 = x_4 = x_5 = x_{21} = x_{27} = 0.$$

This eliminates all colorings except

7, 9, 12, 13, 17, 23, 24, 26, 28, 29.

The 4-ring equations quickly kill this list. The scheme of the complement of the diamond is seen to be empty. Rather than discuss in detail the algorithms used for testing Block-count consistency and the complexity of these algorithms and the quantity of their output, we will instead describe an even further improvement, the superior technique called V-Reducibility.

This new method is orders of magnitude faster and much easier to understand since it does not require Kempe-chain analysis. The purpose for the previous discussion is to place the method of V-reducibility in perspective and to compare it to the classical method and block-count consistency.

VI. Reduction without Kemp Chains : V-Reducibility

What is presented in this section is the joint work of authors. It developed from certain ideas that occurred to Miller [1979] when he undertook to implement the method of Block count consistency described above to produce a shorter (surveyable) proof that the unavoidable set of Appel and Haken is indeed reducible. Our investigation leads us to the discovery that Kemp analysis is not essential in proofs of reducibility. We can even hope to eliminate the tedious case by case calculation which characterizes the proofs of Appel-Haken and Allaire. We shall even be able to shed some light on the ultimate question : Why are four colors enough ?

The first public presentation of this work was by Cohen at the NATO conference on Higher Combinatorics, in Berlin 1976, although the full consequences of block count consistency were not then realized.

Definition. — Given a configuration M in an n -ring R let the color vector $V(M)$ be defined as

$$(x_1, x_2, \dots, x_\rho)$$

ranging over all extension numbers for the n -ring.

We have seen that not all vectors of non-negative integers can be the color vector of some real configuration M because the x 's must satisfy some equations arising from Kempe-chain type arguments. Unfortunately these equations do not provide a necessary and sufficient condition. A vector may satisfy all the equations and still not be realizable as the color vector of a configuration. (Consider for example the fact that the complement for the pentagon can be described by a vector satisfying all Kempe-implications but since, if the four color conjecture is true, there are no critical maps so no such complement exists.)

The Kempe implications so far have been our only use of the property of planarity of the graphs in question, we will now show how we can use planarity in a more profitable way.

Birkhoff [1934, p. 90] made the following observation. Let us start with a dual graph of a configuration M in a ring R which we are vertex coloring and let this graph contain the edge AB . We now form two new graphs from M , the deletion graph M_d which contains all vertices and edges of G except the edge AB , and the contraction graph M_c in which the vertices A and B are identified and all edges which used to lead to either vertex now lead to the concatenated vertex.

Every coloring of R which extends to M colors A and B differently. Every coloring of R which extends to M_c colors A and B the same. If we consider any specific coloring of R , the number of ways it extends to M

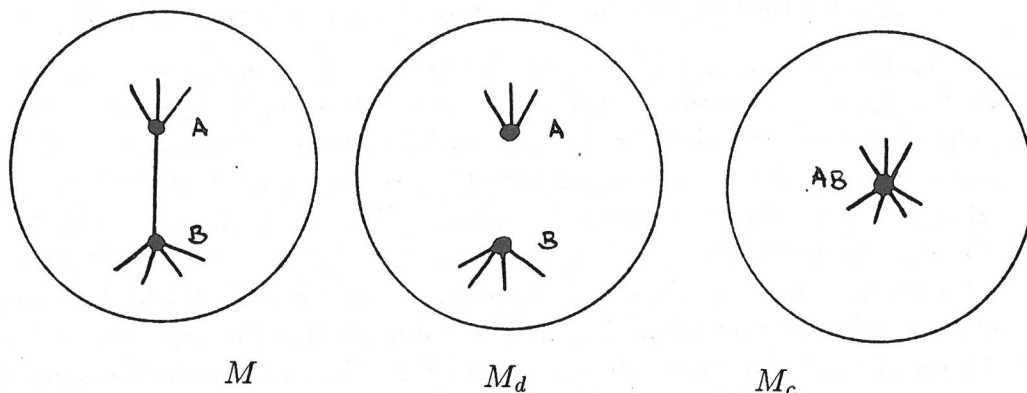


Fig. 41

plus the number of ways it extends to M_c is equal to the number of ways it extends to M_d . We may write this as the law of Deletion and Contraction.

THEOREM [essentially due to Birkhoff].

$$V(M) = V(M_d) - V(M_c).$$

Notice that if M is planar then so are M_d and M_c . The process of Deletion and Contraction has the possibility of isolating a component of the configuration and producing a disconnected graph. If we are careful we can arrange it so that the only components to be isolated are singleton vertices, floating inside the configuration. Any isolated vertex can be colored in 4 ways no matter what extension of the ring coloring is being considered.

THEOREM. — *Let M be a dual graph in a ring R and M^* be the same but with an additional isolated vertex, then*

$$V(M^*) = 4V(M).$$

The use of Deletion and Contraction by Birkhoff and later Birkhoff and Lewis was for the purpose of writing all graphs in term of complete graphs. We will use it in the opposite direction, i.e. to disassemble graphs instead of construct them. Birkhoff and Lewis were not interested in maintaining planarity — we are. An even more important distinction between our two treatments is that we demand that the outer ring be preserved intact.

We can keep repeating the process until we arrive at a graph with no contractible edges. This means that there are no vertices other than those of R and all edges are diagonals. Let us call these configurations primitives.

ON THE FOUR COLOR PROBLEM

THEOREM. — *There are exactly*

$$\frac{1}{2} \sum_{k \geq n/2}^n (-1)^{n+k} \frac{3^{2k} (2k-2)!}{6^n (k-1)! (2k-n)! (n-k)!}$$

primitives for the labeled n -ring.

The sequence begins

1 3 11 45 197 903 4279 ...

The eleven diagonalizations for the 5-ring are described in Fig. 41.

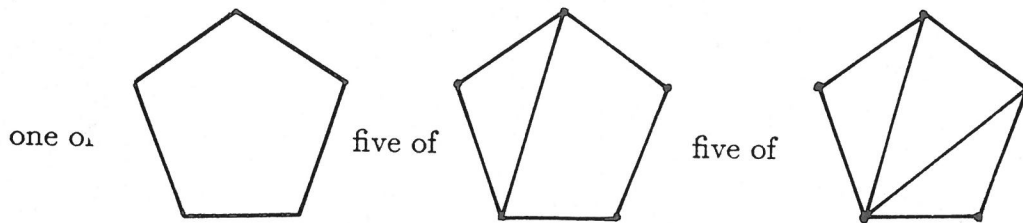


Fig. 42

All configurations inside of 5-rings can be written as linear combinations of these eleven. More importantly, all possible $V(M)$'s of any configuration inside the 5-ring are linear combinations of the eleven color vectors for these figures. All the entries in the color vectors of primitives are either 0 or 1 since any particular ring coloring is either consistent with the diagonals or not, there is no further extension possible since there are no inside countries.

These vectors are not linearly independent. To obtain the basis for the space they span we need a way of continuing the process of Deletion and Contraction beyond the stage of primitives. We introduce the notation of a dotted-line diagonal to mean that the end point vertices must be colored the same (analogous to the contraction of the two ring vertices). Deletion and Contraction now says that we take a primitive with a diagonal and form two new structures, one the primitive with the diagonal deleted and the other a figure with the diagonal replaced by a dotted line (*cf.* Fig. 43).

We continue this process until all we have are dotted diagonals. There are as many of these as solid-line primitives. Most of them, however, have the color vector of all zeros (*cf.* Fig. 44).

The configuration above requires A and B to have the same color and also for B and C to have the same color. However, A and C are adjacent ring countries and so no element in C satisfies these conditions. Of the

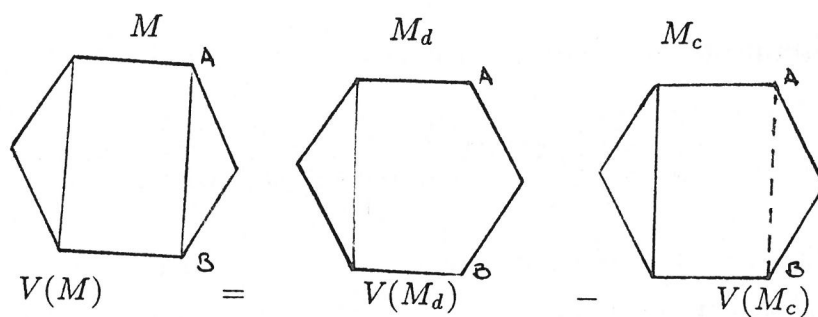
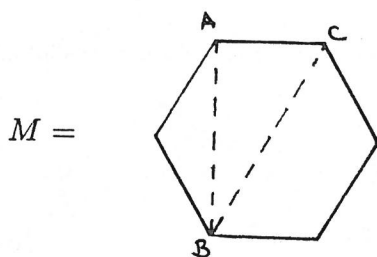


Fig. 43



$$V(M) = (0, 0, \dots, 0)$$

Fig. 44

ring countries and so no element in C satisfies these conditions. Of the eleven primitives listed for the 5-ring only the first 6 correspond to dotted diagonalization with non-zero vectors.

Definition. — A *prime* is any non-zero color vector for a dotted diagonalization of the n -ring.

Some different dotted diagonalizations give the same color vector (cf. Fig. 45)

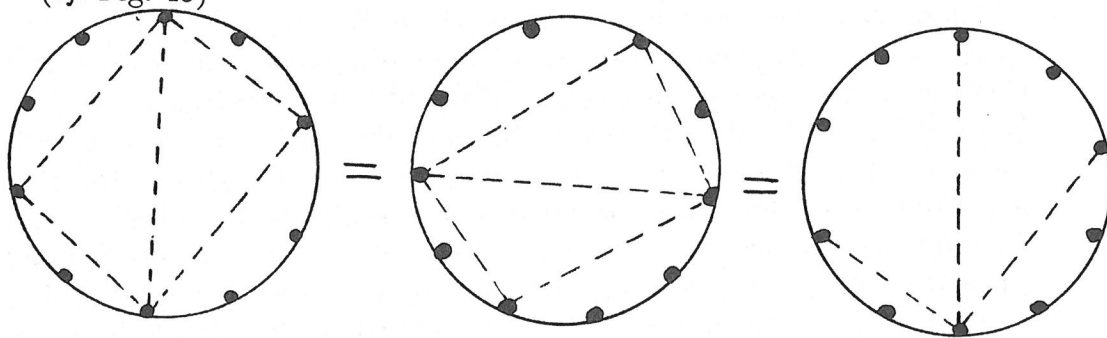


Fig. 45

ON THE FOUR COLOR PROBLEM

We have avoided this problem by defining the primes to be the color vectors not the configurations themselves. Let

$$B_n = \left(\frac{3}{2}\right)^n \sum_{n \geq j \geq n/2} \frac{1}{j3^j} \binom{2j-2}{j-1} \binom{j}{n-j};$$

and let

$$A_n = B_n - B_{n-1} + B_{n-2} - \dots \pm B_1.$$

THEOREM [new]. — *There are exactly A_n primes for the n -ring.*

We already know that the color vector of any configuration is a linear combinations of the dotted diagonalizations, we further believe that the primes are in fact the basis for this vector space. This has been calculated to be true through dimension (ring size) 14. It would be interesting to prove this though it is not essential to the proof of the four color theorem. One thing we do know is that the dimension of the space of primitives is at most the numbers of primes.

Let us compare the advantage of looking only at the primes instead of at the full set of primitives.

Ring Size	Number of primitives	Number of primes	Number of colorings
3	1	1	1
4	3	3	4
5	11	6	10
6	45	15	31
7	197	36	91
8	903	91	274
9	4,279	232	820
10	20,793	603	2,461
11	103,049	1,540	7,381
12	518,859	4,005	22,144
13	2,646,723	10,440	66,430
14	13,648,869	27,261	199,291

Example. — The 232 primes for the 9-ring are pictured in Fig. 46. The number of different variants possible through the action of the dihedral group are indicated under each figure.

In order to get the full use out of the primes we convert them into diagonal form which means we form linear combinations which have distinct first non-zero component. We call these *reduced primes*. From

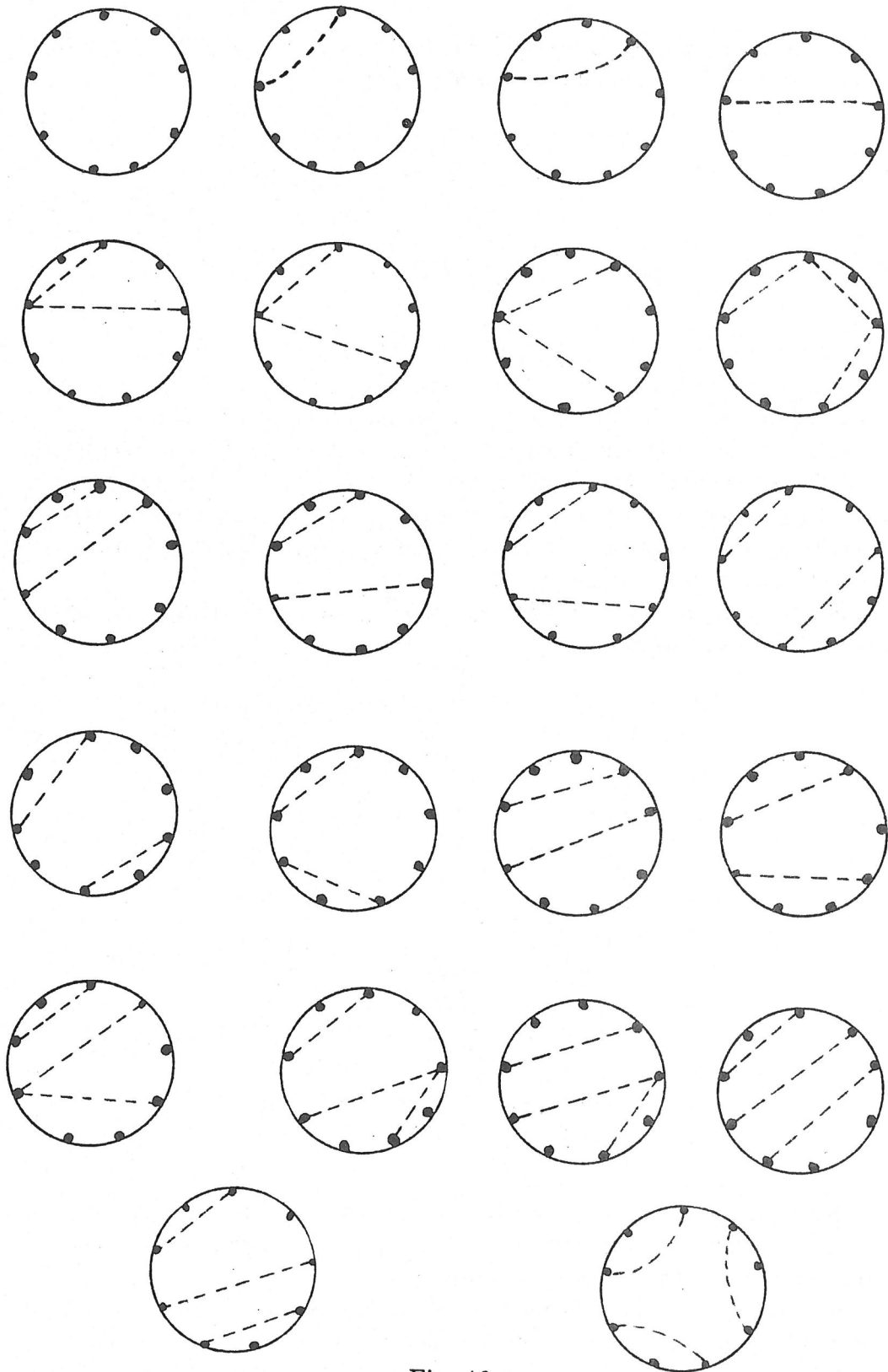


Fig. 46

ON THE FOUR COLOR PROBLEM

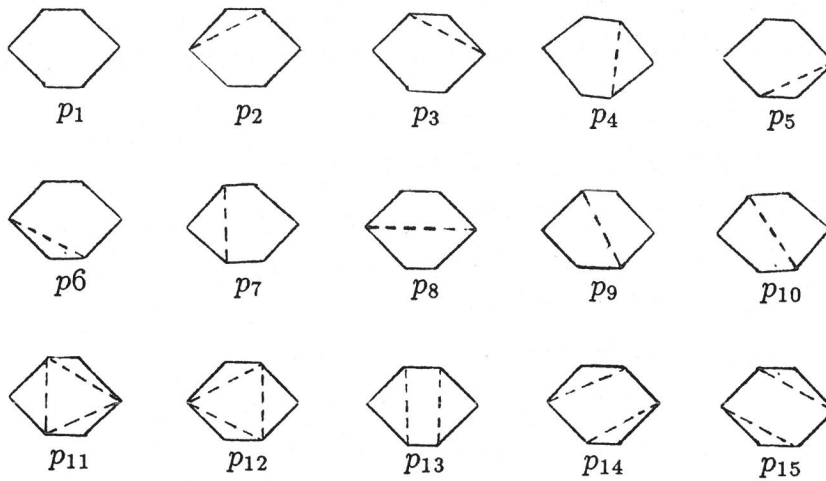


Fig. 47

these we can evaluate the whole scheme of a configuration by calculating the extension numbers of the much smaller set of reduced primes.

We shall now illustrate this whole process for the 6-ring. The primes for the 6-ring are shown in Fig. 47.

Their color vectors (sans parentheses and commas) are:

$p_1 = 11111111111111111111111111111111$
 $p_2 = 11111111111000000000000000000000$
 $p_3 = 11110000000000000011111110000000$
 $p_4 = 110011100000011000001100000011$
 $p_5 = 1010010100100000010010100010101$
 $p_6 = 1100111000000000011100001110000$
 $p_7 = 1010100001000101010010101000010$
 $p_8 = 00000000000111111000000000000000$
 $p_9 = 0000000110011000000000000001100$
 $p_{10} = 0000000000010000101000010101000$
 $p_{11} = 1010000000000000010010100000000$
 $p_{12} = 11001110000000000000000000000000$
 $p_{13} = 1000100000000100000010000000010$
 $p_{14} = 10100101001000000000000000000000$
 $p_{15} = 11000000000000000111000000000000$

It is easier to do our calculations if we replace these with linear combinations which have unique first non-zero components, such as

$$\begin{aligned}
 d_1 &= p_1 - p_2 - p_3 - p_4 - p_6 - p_9 + p_{11} + p_{12} + p_{13} + p_{15} \\
 d_2 &= -p_1 + p_3 + p_4 + p_7 + p_9 - p_{11} - p_{13} \\
 d_3 &= -p_1 + p_3 + p_4 - p_5 + p_6 + p_7 + p_9 - p_{12} - p_{13} + p_{14} - p_{15} \\
 d_4 &= p_2 - p_4 + p_5 - p_7 + p_8 - p_9 + p_{13} - p_{14} \\
 d_5 &= -p_1 + p_3 + p_6 + p_7 + p_9 - p_{11} - p_{15} \\
 d_6 &= p_1 - p_2 - p_4 + p_5 - p_6 - p_9 - p_{11} + 2p_{12} \\
 d_7 &= p_2 + p_3 + p_4 - p_5 + p_6 - p_7 + 2p_{11} - 2p_{12} \\
 d_8 &= -p_1 + p_2 + p_4 + p_6 + p_9 - 2p_{12} \\
 d_{10} &= 2p_1 - p_2 - p_3 - p_4 - 2p_6 - p_8 - p_9 + 2p_{12} \\
 d_{12} &= -p_1 + p_2 + p_3 + p_8 + p_{10} - p_{14} \\
 d_{13} &= 2p_1 - 2p_2 - p_4 - p_5 - p_6 - p_8 - p_{10} + 2p_{12} + p_{14} \\
 d_{14} &= -p_1 + p_2 + p_4 + p_6 + p_8 - 2p_{12} \\
 d_{18} &= -p_1 + p_2 + p_5 + p_6 + p_8 - p_{12} - p_{14} \\
 d_{19} &= p_1 - p_2 - p_5 - p_8 + p_{14} \\
 d_{21} &= p_1 - p_2 - p_6 - p_8 + p_{12}
 \end{aligned}$$

We have subscripted the d 's such that d_i is the only one of these vectors to have a non-zero entry in column i (the entry is a 1 or -1). That we can do so proves that the p 's are independent. Let us define the set BASIS.

$$\text{BASIS} = \{1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 10 \ 12 \ 13 \ 14 \ 18 \ 19 \ 21 \}$$

We are now ready to describe the Vector Vector Method of demonstrating the reducibility of a configuration M inside a 6-ring. For every i in the set BASIS count how many times coloring C_i extends to M , call this as before x_i . Since the vector $V(M)$ is a linear combination of the p 's and the p 's are a linear combination of the d 's, we have :

$$V(M) = \sum_{i \in \text{BASIS}} x_i d_i.$$

Therefore by counting the extension numbers of 15 colorings of the 6-ring instead of the full 31 we can still evaluate the whole scheme. For the 14-ring we need to extend only 27,000 instead of 199,000; which is a considerable saving.

From this point the reducibility method continues as the block count consistency method described above, but with one new wrinkle. C -reducibility is easier to prove than ever. The colorings whose subscripts

ON THE FOUR COLOR PROBLEM

are in the set BASIS act as a reducer in that if any configuration includes all such colorings, or misses all such, then it must be reducible since either its vector or its complement's vector is all zeroes. In fact, a particular C -reducer need only be checked by noting whether those vectors in BASIS that span it are either missed or contained. In other words, once we have chosen the vectors in BASIS all the problems of intersecting large sets of colorings can be reduced to problems of intersecting much smaller sets of colorings.

The best part of all of this is that the choice of the colorings in BASIS was not unique. There are many spanning sets possible for the set of primitives diagonalized from the set of primes as above. There are many ways of choosing the 27 thousand independent indices out of 199 thousand colorings for the 14-ring. Each selection has the possibility of acting like a reducer.

If the existing (and exhausting) Kempe-based proofs of the four color problem are correct, this method can show the same result in a tiny fraction of the time and printout space. Not only does V -reducibility mean that we do not have to check the extendability of so many colorings in the first place, but it also means finding reducers is easier. Using block count consistency means that once a configuration has been shown to be reducible the proof can be printed out by the computer in one long line. All it needs to do is to specify the constants each equation must be multiplied by to be added up to say "the sum of the x 's is 0." The number of such constants is the number of equations for the ring, so though this step is automatic and surveyable it is not trivial to perform.

What we have presented so far is a methodology which can take the final decision of reducibility out of the hands (circuits) of a computer and return it to humans. This, even when it leads to a disproof of the existence of critical maps, does not answer the important question: Why four? Perhaps the machinery we have set up can shed some light on this point. The generating function for the number of primes (A_n above) is

$$y(z) = \frac{1 - z - \sqrt{(1+z)(1-3z)}}{2z}.$$

Following the methods that P. Flajolet expounds in this very volume we observe that there are algebraic singularities at (-1) and $1/3$. The one at $1/3$ dominates giving

$$A_n \approx \frac{1}{1 + 1/3} \left[\frac{2/3 - \sqrt{(1-3z)(4/3)}}{2/3} \right].$$

Therefore :

THEOREM [Flajolet, 1987].

$$A_n = \frac{9}{8\sqrt{3\pi}} 3^n n^{-3/2} (1 + \mathcal{O}(1/n)).$$

Let us be careful to remember that the definition of the primes had nothing to do with the number of colors, it is simply a property of planar maps. Recall also that the number of λ colorings of the n -ring was shown to grow like $(\lambda - 1)^n/n!$. This means that λ less than 4 has no chance of providing enough colorings to surpass the dimension of the space spanned by the geometry, but that $\lambda = 4$ will eventually catch and exceed this quantity. For λ four or greater, the ratio of the dimension of the primitives to the number of colorings will become so large that "random" configurations will be likely to contain a spanning set of colorings and therefore be D -reducible. Note that we observed that they need only contain a spanning set of colorings (like BASIS above) not have their scheme equal to all colorings, in order to be reducible.

We therefore suggest that it is the coincidence that both the dimension of the primitives and the number of colorings grow as powers of three that lies at the heart of why four colors are enough to color any planar map.

Daniel I. A. COHEN,*
 Department of Computer Science,
 Hunter College,
 City University of New York,
 695 Park Avenue,
 New York, N.Y. 10021, U.S.A.

Victor S. MILLER,
 I.B.M. T. J. Watson Research Center,
 P.O. Box 218,
 Yorktown Heights, N.Y. 10598, U.S.A.

* This paper was prepared while the first author was visiting the Computer Science department at Columbia University.