A HISTORY OF THE FOUR-COLOR PROBLEM

By

CHARLES J. SEARCY

Bachelor of Science
Panhandle A and M College
Goodwell, Oklahoma
1957

Master of Science
Oklahoma State University
Stillwater, Oklahoma
1964

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the degree of
DOCTOR OF EDUCATION
May, 1967

A HISTORY OF THE FOUR COLOR PROBLEM

Thesis Approved:

Robert W. Gibson
Thesis Adviser

Milton E. Berg

Vernon Troxel

William Marsden

D. D. Durham
Dean of the Graduate College

PREFACE

I am extremely grateful for the many considerations, time and patience given by Professor R. W. Gibson in the preparation of this thesis. This is especially meaningful since Professor Gibson was not made a member of my advisory committee until after the final draft of this thesis had been prepared. His efforts must have come from the goodness of his heart. Also, I wish to express my thanks to the other members of my advisory committee.

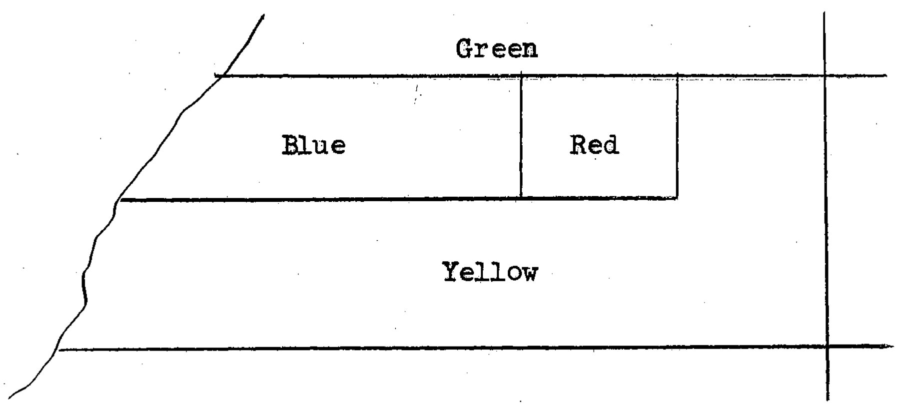
Additionally, I am deeply aware of and grateful for the many sacrifices which my family has made in order that I might have the opportunity to prepare this thesis.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Geographical maps consisting of regions bounded by simple closed curves on a plane or sphere are most often colored in such a manner that no two states or subdivisions of a country or continent which border are of the same color. The figure below is an example of a map in which four colors are necessary to accomplish this.



The question of whether four colors is sufficient to color a planar or spherical map has never been successfully answered. This question was formulated as a mathematical question as early as 1850 by Augustus DeMorgan and was put before a wide mathematical public in 1878, when Arthur Cayley proposed it to the London Mathematical Society. The next year in Volume 11 of the *American Journal of Mathematics*, a solution was published by Alfred B. Kempe. However, the problem was definitely unsolved again eleven years later, when Percy John Heawood pointed out an error in Kempe's reasoning. He did show by a revision of Kempe's proof, that five colors are always sufficient to color a planar or spherical map. However, there has never been an example given of a map where five are necessary.

The problem is today probably one of the most simply stated of the unsolved problems in mathematics. It has held the interest of many famous and capable mathematicians, not the least of whom are G. D. Birkhoff, Kempe, Veblen, Brahana, Balantine, Reynolds, P. J. Heawood, C. E. Winn, and Philip Franklin. The endurance record of interest undoubtedly goes to Heawood, who became interested in the problem in the eighties while a student under Cayley. He published his first paper on the problem in 1890, and several others including one as recent as 1935. Some of the research of F. Harary and G. Prins received financial support from the National Science Foundation and their results were published in the Canadian Journal of Mathematics. G. A. Dirac published results related to the problem as late as 1957, and G. Ringle published a solution to a related problem in 1959. Even today the problem is a popular topic for magazines and books on mathematics.

In general the mathematics student is not well acquainted with mathematics outside the textbook and he frequently considers even the statement of unsolved problems as being outside the realm of his understanding. However, here is a classical unsolved problem whose statement he can read and understand. In fact, many of the attempts to solve the problem utilized only techniques or concepts that would be within the level of understanding of many high school students.

A teacher well versed in the four-color problem could readily introduce an interested mathematics student to that section of the world of sophisticated mathematics apart from the pages of the textbook. The primary function of this paper is to provide a means for a teacher to interest such students in the problem and many of its related areas.

A second function is to provide the serious reader with an interesting inroad on the study of graph theory. Graph theory is an important mathematical tool that can be universally applied. The study of graphs came about simultaneously in a number of widely diversified disciplines and only recently has been treated as a subject independent of any specific application. The four-color problem has been a very large contributor to graph theory.

The study of the four-color problem has led to many interesting related areas. Many of these will be identified in this thesis and some of them will be emphasized.

Chapter two presents the definitions of terms used in the rest of the paper, as well as Euler's Theorem. The third chapter is intended for the lay reader. It contains a nontechnical account of the problem. Chapter four is intended for the interested but inexperienced mathematics reader; it should be readable to the above-average high school student. The remainder of the paper is not above the undergraduate mathematics major.

CHAPTER II

DEFINITIONS

A finite map on the plane is a subdivision of the Euclidean plane into regions by a finite number of finite arcs such that no region lies on both sides of any arc, and no two distinct arcs have more than their endpoints in common. A division of the plane by arcs such as is illustrated in Figure 2.1,d is not a finite map. The combination of two or more arcs can form a loop, as is illustrated by arcs α_1 and α_2 in Figure 2.1,b and 2.1,c; hence, the condition that the endpoints of an arc are distinct does not exclude the possibility of a loop.

An arc is anything topologically equivalent to the closed interval [0,1].

If S is the plane and A the union of the set of arcs forming a map M, then S - A is a collection of disjoint, connected open sets each of which is called a region of the map M.

A boundary between any two regions of a map is the union of that collection of arcs that separate those two and only those two regions. An arc is said to separate two regions if it is adjacent to both regions (every neighborhood of an interior point of the arc contains a point of each region). Hence, the union of the interior of an arc with the two regions it separates is a connected open set. Since two arcs can have only their endpoints in common, no region is separated from two other regions by the same arc. Since no region lies on both sides of any arc, every arc must be part of some boundary, that is, arcs such as α_3 and α_4 in Figures 2.1,d are not admissible as part of a proper map.

A vertex of a map, referred to hereafter as a vertex, is a point common to at least two boundaries. The existence of two boundaries at a point implies the presence of at least three regions at that point. Refer to a vertex as an end of a boundary as often as it appears as an end of an arc in the boundary. It follows that a vertex is an end of at least three, not necessarily, distinct boundaries. Notice that since a boundary can have more than two ends, Figure 2.1,a, there may be more than two vertices lying on one boundary. Also, two ends of a boundary may be the same point thus allowing the possibility of only one vertex on a boundary, Figure 2.1,b or, a boundary may have no vertices of the map (as contrasted with vertices of arcs), Figure 2.1,c. These possibilities are illustrated.

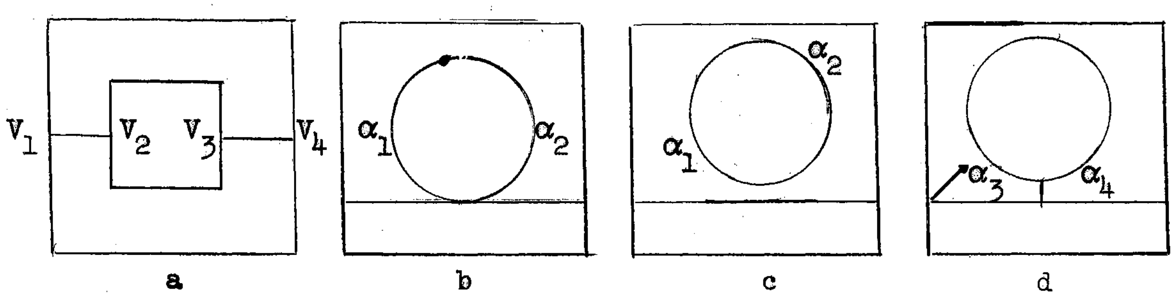


Figure 2.1

It can be shown that a boundary has no more than two ends (of its arcs) at a vertex.

Even though we have defined a map in terms of a set of arcs on the plane, the map as derived is rather independent of the set of arcs used to establish it. That is, for any particular map, the set of arcs from which it could be derived, is not unique. This is easily seen by considering the numerous ways in which a single boundary could be broken up into a set of arcs. Therefore, we shall lose the need for reference to the arcs making up a map rather early in our discussion.

Mathematicians like to make general statements and do not believe that exceptions prove the rule. The assertion that a quadratic equation has two roots is not true if we restrict ourselves to the real numbers. This was a powerful reason for introducing the complex numbers; in the enlarged number field the statement is true, counting multiple roots. The function $f(x,y) = \frac{x}{x^2 + y^2}, \frac{-y}{x^2 + y^2}$ maps the entire plane (except (0,0)) on the entire plane (except (0,0)), in a one-to-one fashion. These exceptions are found distasteful, and for this reason we consider the extended plane. Our extended plane consists of all the Euclidean plane with the annexation of one more point, the ideal point. We extend the function f to contain the ideal point and (0,0) as images of each other, preserving one-to-oneness.

A set of points consisting of the ideal point and a Euclidean line through the origin maps onto a Euclidean line through the origin and the ideal point. The set of points $\{(x,y) \mid y = ax + b\}$, $b \neq 0$, maps into a circle through the origin with the exception of the origin itself. If we define a line in the extended plane to contain the ideal point, as well as a Euclidean line, the image of a line is a circle or a line through the origin. According to our definition the ideal point lies on every line in the extended plane. This changes the concept of parallel lines as well as many of the intuitive concepts of geometry. On the extended plane two lines are considered parallel if their only point of intersection is the ideal point.

To consider maps on the extended plane it is appropriate to consider boundaries that are not finite; let us visualize what could happen to our concept of a map: Infinite boundaries contain the ideal point and could not possibly be made up of a finite number of finite arcs of the Euclidean plane. Such a boundary contains a subset which is topologically equivalent to $\{x | x = 1/t, 0 < t \leq 1\}$; we shall call such a subset, with the ideal point for the end point, an <u>infinite arc</u>. If two or more boundaries contain infinite arcs then the ideal point is common to them and is a vertex of the map. If only one boundary contains an infinite arc then this boundary will contain two, and the ideal point will be a point of that boundary without being an end.

When we consider only those geometric properties that are related to the problem, there is no distinction between a map on the extended plane and a map on a sphere. To see this, consider a sphere tangent to the plane at any point, T, on the finite plane. Let N be the point on the sphere diametrically opposite T. For every point p on the finite plane, there exists a line $\overline{Np}$ which intersects the sphere at some point p', other than N. Likewise, each p' other than N determines a point p.

This gives us a one-to-one correspondence between the points on the finite plane and the points on the sphere other than N. Then N is made to correspond to the ideal point of the extended plane. Such a transformation is called a stereographic projection. Each region in the plane corresponds under this transformation to one and only one region on the sphere. Continuity and one-to-oneness of the transformation require that the boundaries maintain the same relative configuration under this transformation. If a boundary of the map on the sphere contains the point N as an interior point then that boundary is the projection of two infinite arcs on the plane. These two arcs have the ideal point as an end in common. Since they separate the same two unbounded regions they are part of the same boundary. If several boundaries have the point N in common then it is a vertex of the map on the sphere, corresponding to the situation in which the ideal point is a vertex of a map on the extended plane. With our new concept of an infinite arc we can consider maps on the extended plane.

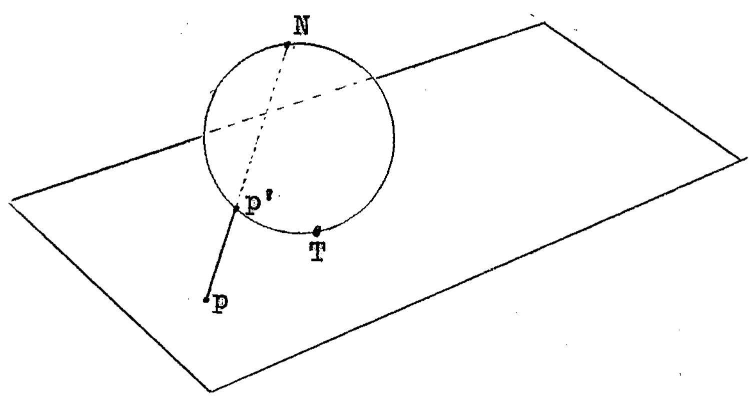


Figure 2.2

A map on the extended plane or sphere is a subdivision of the extended plane or sphere into regions by a finite number of arcs (finite or infinite), having only their endpoints in common, such that no region lies on both sides of any arc. With arc meaning either finite or infinite arc, our definitions remain unaffected. That is, a region is still a connected open set. End of boundary and vertex have picked up no new concepts, except that they may be the ideal point.

Actually we could avoid the problem of considering maps on the extended plane since each such map is homeomorphic to a map on the finite plane. Such a homeomorphism could be obtained by the composition of two stereographic projections. Let M be a map on the extended plane such that the ideal point P is on some boundary. Project M on the sphere by a projection f_1 with point of tangency T_1 and let $f_1(M) = M$. Let T_2 be a point on the sphere such that N_2, the point diametrically opposite T_2, does not lie on any boundary of M. Let f_2 be the stereographic projection of M_1 onto the plane with N_2 as the pole of projection, and let $f_2(M_1) = M_2$. Now the ideal point is not on a boundary of M_2. M_2 is the homeomorphic image of M under the transformation $f_3 = f_2 \circ f_1$.

Euler's Theorem. If M is a simply connected map on a sphere with $v > 0$ vertices, E edges and F faces then $V - E + F = 2$. We shall defer proof of this result until the end of the chapter.

The remainder of this chapter may prove to be tiresome reading since there is very little need for discussion of the definitions that remain. Continuity will not be lost if the reader goes on to chapter three and uses this chapter for reference when necessary.

The *multiplicity of a vertex* is the number of ends of boundaries at that vertex. By definition the multiplicity of a vertex of a map is at least three. However, it will sometimes be advantageous to speak of a "vertex of multiplicity two" as a particular point which is not an end-point of a boundary but is on a boundary.

Two vertices are said to be *neighboring* if there exists an arc which is a combination of the arcs that make up the map, with those two vertices as ends and with no vertex of the map an interior point of that arc. For example, V_1 and V_2 in Figure 2.1,a are neighboring.

An edge of a region is that portion of a boundary that lies between two neighboring vertices. Note that a boundary is not necessarily the union of a set of edges. See Figures 1.1,b and 1.1,c.

Two regions are contiguous (neighboring) if they have a boundary in common. It follows from the definition of boundary and edge that if two regions have an edge in common, they have a boundary in common, but not conversely.

A map is properly colored if contiguous regions are assigned different colors. This term will sometimes be shortened to saying a map is colored or colorable.

A connected map is one whose regions are simply connected. The union of a simply connected region and its boundaries may not be simply connected, for example Figure 2.1,b. It will be shown in the third chapter that for such simply connected maps there always exists a sequence of edges forming a path between any two vertices. With this we can see that the vertices and boundaries of a connected map form a connected graph. It also follows, although we will not prove it, that a connected graph which forms a map on a plane or sphere forms a connected map, thus the two are equivalent on a plane or sphere.

A regular map is a connected map such that each vertex has multiplicity exactly three.

A connected map is a polyhedral map if and only if every region has more than two boundary edges and every boundary edge has two distinct ends. Loops such as those in Figure 2.1,b and c are not present, by definition.

This polyhedral map as defined here is a generalization of the concept of a polyhedral graph as defined in [13]. A polyhedral graph does not allow two nodes to be connected by more than one edge whereas the graph composed of the boundaries and vertices of our map may have two or more edges between the same two vertices. As a result, everything that is done with polyhedral maps is applicable to polyhedral graphs.

If a convex polyhedron is enclosed in a sphere so that the center of the sphere is inside the polyhedron then it is apparent that the projection from the center of the sphere, of the vertices and edges of the polyhedron onto the sphere forms the boundaries and vertices of a polyhedral map. It is in fact a polyhedral graph on the sphere. However, a convex polyhedron cannot have two faces with more than one edge in common. Figure 2.3 shows a polyhedral map on the sphere and an equivalent non-convex polyhedron.

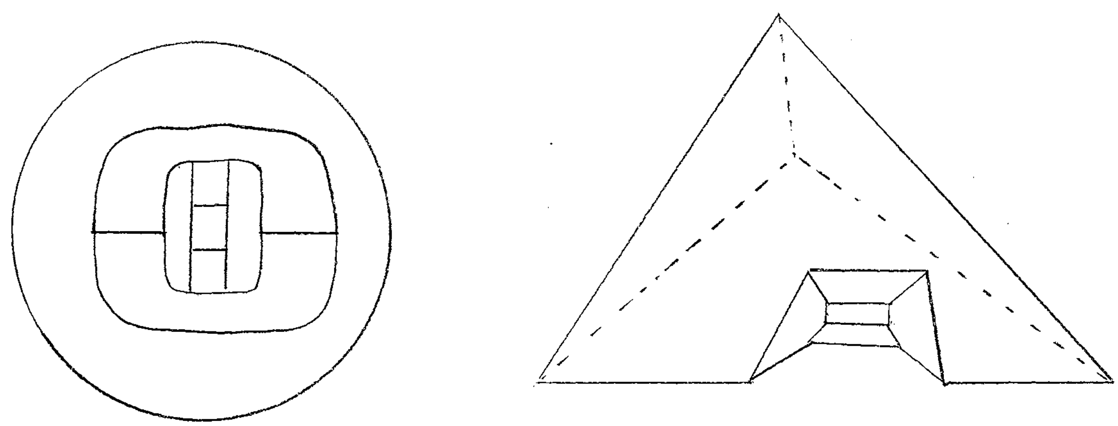


Figure 2.3

Certain polyhedra cannot be realized as a polyhedral map on the sphere. Figure 2.4 is an example of such a polyhedron.

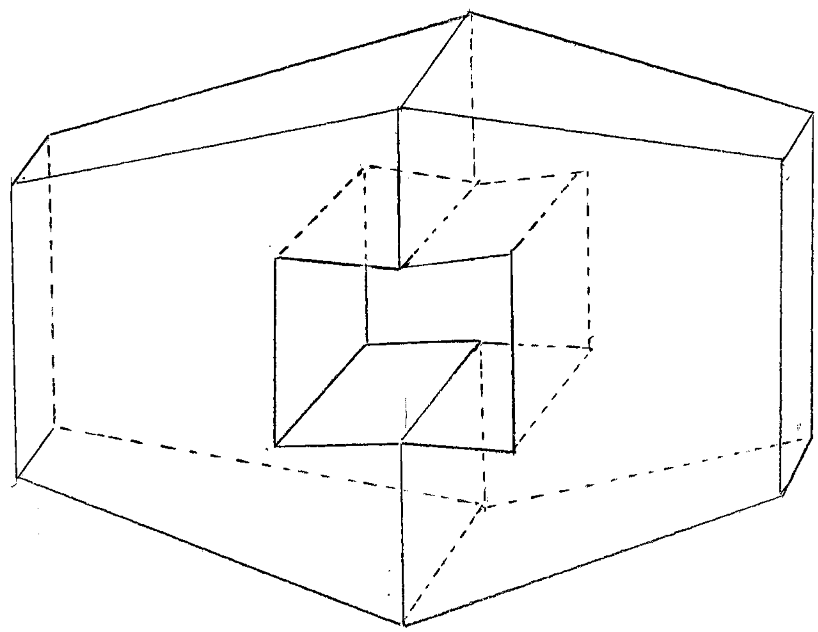


Figure 2.4

An isthmus is a boundary line which separates a region from itself. This situation cannot occur in a proper map. However, we use the term in discussing certain reduction processes which must be avoided since they can result in an isthmus.

A map is reducible for a specified number of colors if a coloration may be made to depend on a coloration, in the same number of colors, of a regular map with fewer regions.

A set of regions in a map is said to be a reducible configuration if its presence in the map renders it reducible.

The chromatic number of a map on any surface is the minimum number of colors in which the map can be properly colored.

A graph consists of a set of discrete points called nodes or zero-cells with a set of arcs called edges or one-cells such that each endpoint of an arc is a node and two arcs intersect only at nodes. A zero-cell is also called a vertex or a zero-simplex.

The number of nodes in a graph is called the order of the graph. A graph of order k is called a k-graph.

If the nodes and edges of a graph G are also nodes and edges of a graph G' then G is said to be a subgraph of G'.

If the number of edges of a graph meeting at a vertex is the same for each vertex, counting each loop twice, then the graph is called regular. If that number is N, say, then it is called a regular graph of degree N.

A regular graph is said to be factorable if it can be obtained by superimposing the vertices but not the edges of two regular graphs of the same order but each of lower degree. A regular graph of third degree may be factorable into two factors, one of first degree and the other of second degree. If the second degree factor is also factorable then the third degree graph is factorable into three factors of first degree. A non-factorable graph is said to be primative [4]. Figure 2.5 shows a factorable graph with a set of factors and a primative graph.

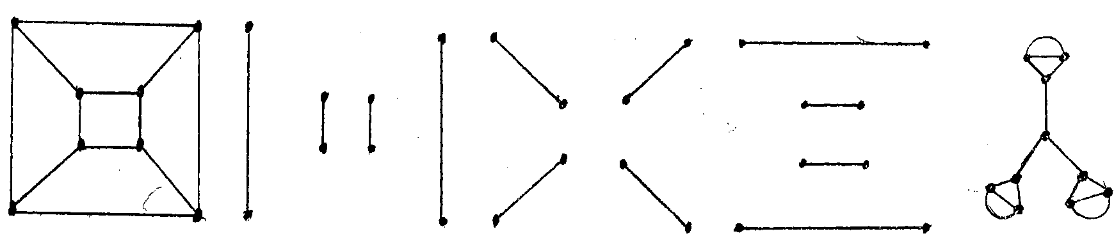


Figure 2.5

If from one vertex of a graph to another, both vertices of order greater than two, there exists only one path, and that along a single edge, then this edge is called a bridge. If this edge is removed, the graph is separated.

A graph is of genus zero if it is homeomorphic to a graph on a sphere.

A path that includes every edge of the graph once and only once is called an Euler path.

In a graph a path that includes each vertex of the graph once and only once is called a Hamiltonian path.

A leaf is a portion of a graph which is connected to the rest of the graph only by a single 1-cell such that no proper part of the leaf has this property [36]. A node of degree one is a leaf; in a tree the leaves are all nodes.

A graph is said to be k-chromatic if its nodes can be properly colored (nodes on the same edge having different colors) in k colors and if for $j < k$ there exists no proper coloration of the graph in j colors.

A graph with finite chromatic number is called critical if it has no subgraph of smaller order with the same chromatic number.

A surface is said to have connectivity h if $h - 1$, but not h, arcs can be found on it in a certain order that do not separate the surface, where it is stipulated that the first arc is actually a simple closed curve and that every subsequent arc connects two points lying on the preceding arc [47]. This means that for the sphere $h = 1$, for the Kline bottle $h = 3$, torus $h = 3$, and the projective plane $h = 2$.

Three-dimentional models of surfaces of odd connectivity can be obtained, whereas surfaces of even connectivity can be realized only in a space of higher dimension. Also, there are surfaces of odd connectivity that cannot be realized in three space.

Consider a sphere with two holes in it. Suppose the boundaries (the contours a, a') of the holes are oriented in opposite senses, Figure 2.6,a. If we consider several sets of points all as the same set of points, for example by stretching the figure homeomorphically until they coincide, we say we have made an identification of these sets. We shall call such an identification of a with a' as an identification of the first kind. If the sense of each contour is the same, e.g. Figure 2.6,b. We shall call this an identification of the second kind. We can then say that the identification of the first kind is the fitting of a and a' with a handle of the first kind; and an identification of the second kind is the fitting of a and a' with a handle of the second kind. This can be illustrated for the handle of the first kind by a cylinder

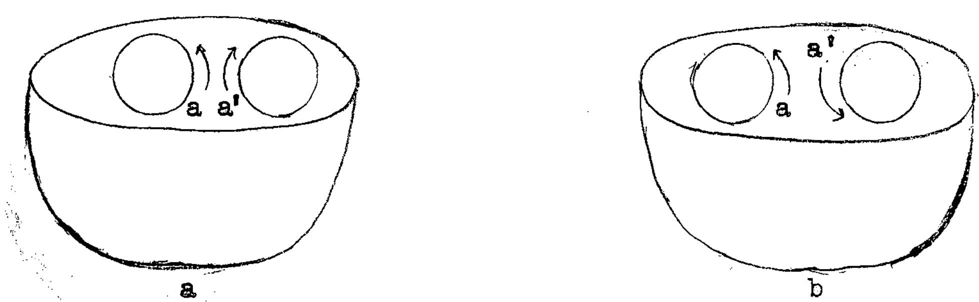


Figure 2.6

bent like an elbow macaroni and its ends fitted on the contours a and a' (see Figure 8.2). The torus is equivalent to a sphere with one handle of the first kind. A handle of the second kind cannot be realized in three space since the cylinder would have to pass through the surface of the sphere in order to link up with the contour a' with the proper orientation. This may be illustrated as in Figure 2.7; however, one must keep in mind the apparent intersection of the handle with the surface does not actually occur.

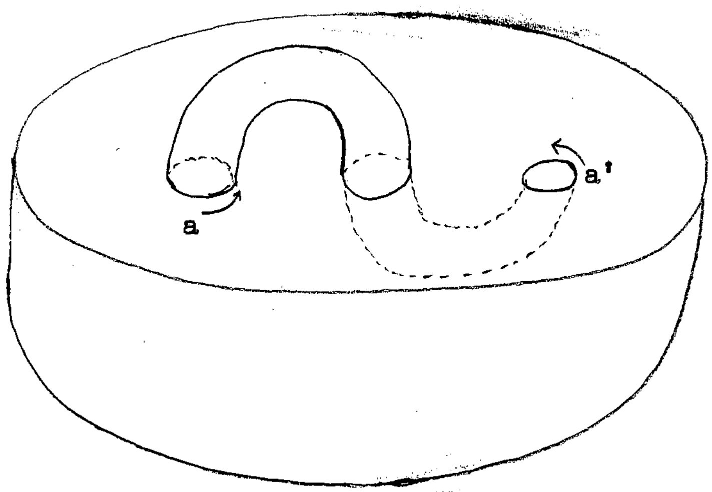


Figure 2.7

It seems that a bug walking along the surface would pass through the handle and be on the inside. It is for this reason the surface is said to be one-sided. If the bug were hitched to a triangle originally at ABC such that, as the bug moved, the triangle slid along the surface, the bug could pass through the handle and return to where it started but it would be impossible to rotate the triangle in the surface so that the vertices would fall in the same place they were before. That is, the triangle would now be oriented ACB. It is for this reason the surface is called non-orientable. A sphere with handles of the first kind is two-sided and is called orientable since sliding the triangle around gives to the surface at each point a unique orientation.

A Mobius band is a one-sided, one-edged surface. A model of one can be made by putting a half twist in a rectangular piece of paper and pasteing the ends together. The Mobius band is also non-orientable.

If a one-to-one continuous correspondence is made between the points on the edge of a Mobius band and the points on the edge of a hole in a sphere, the hole is said to be fitted with a crosscap, thus closing the hole. This of course cannot be done in three space. Another way of regarding a crosscap that is sometimes more meaningful, is to identify the diametrically opposite points of a hole. To visualize this, stretch the edge of the hole as indicated in Figure 2.8.

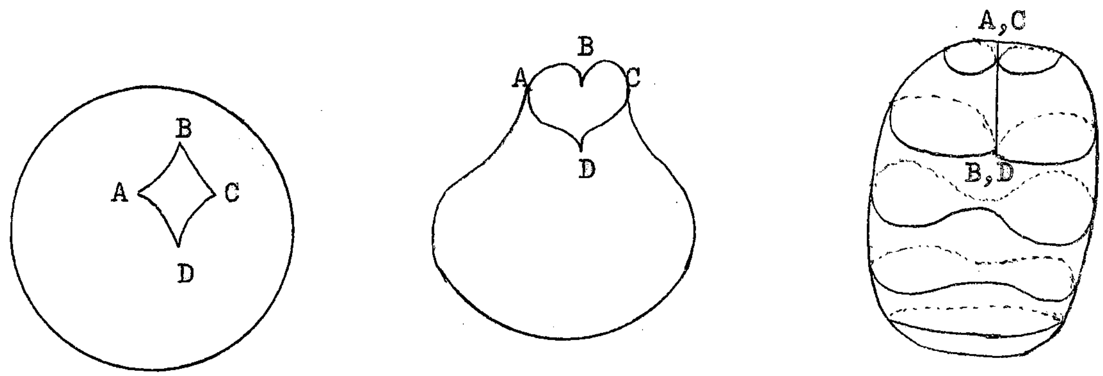


Figure 2.8

In the words of H. S. M. Coxeter, "A crosscap is a hole in a surface with the magical property that a bug approaching the hole on one side would suddenly find himself at the diametrically opposite point and on the opposite side of the surface."

It can be shown that a sphere with two crosscaps is topologically equivalent to a sphere with a handle of the second kind. To see this consider two crosscaps as indicated by the two circles in Figure 2.9, with opposite points identified. If we pull a portion of the plane through the crosscap on the right until B and D fall on the edge of that crosscap we get Figure 2.9,b. Note that points H, G and F are no longer

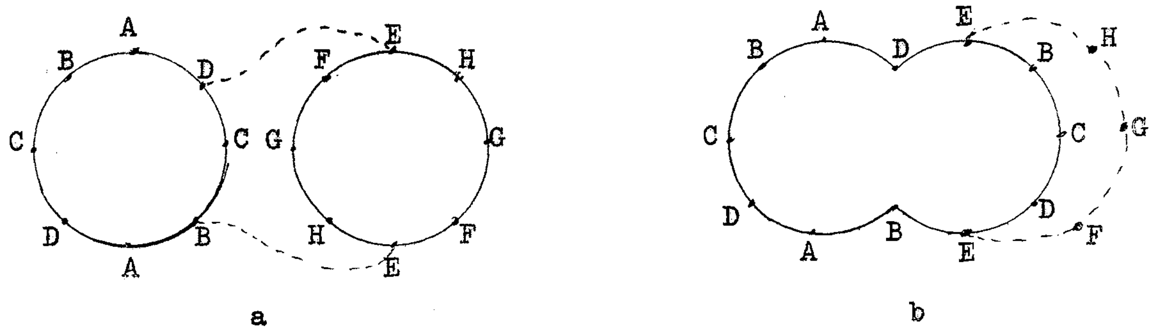


Figure 2.9

on a crosscap and the arc B, D is reversed upon passing through a crosscap. By stretching the surface we can obtain the situation illustrated in Figure 2.9,c, which is equivalent to that in Figure 2.9,d. The situation illustrated in Figure 2.9,d is recognized as a handle of the second kind.

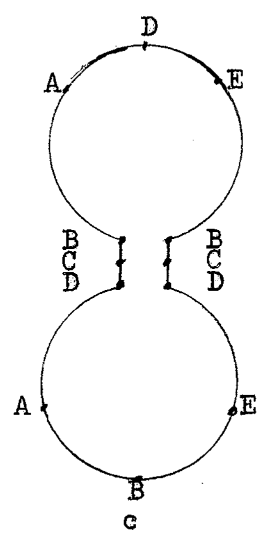
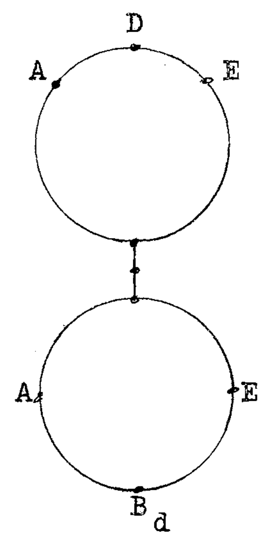


Figure 2.9



A handle of the first kind and a crosscap is equivalent to a handle of the second kind and a crosscap. This can be seen by passing one of the circles of the handle through the crosscap, thus reversing its sense and obtaining two circles of the same sense.

The Kline bottle is a sphere with two crosscaps, though it is usually pictured (Figure 2.7) as a sphere with a handle of the second kind.

The projective plane is a sphere with one crosscap.

The connectivity of the Kline bottle is the same as the torus. Note also that a surface of connectivity $k = 2s + r$ may be obtained from a sphere with $s - i$ handles and $r + 2i$, crosscaps where i is any integer such that these values are both nonnegative. The equivalence of these surfaces is stated in the following theorem, the proof of which can be found in [1].

Two closed surfaces are homeomorphic if and only if they are both orientable or both non-orientable and if they have the same connectivity.

For our purposes, we shall consider a closed surface, usually referred to as the surface, to be a sphere with m handles of the first kind and n crosscaps where m and n are nonnegative integers.

A region has been defined as a connected open set and in a connected map these regions must be simply connected. A region is said to be simply-connected if every simple closed curve lying entirely within the region can be continuously shrunk to a point. A simply-connected region is homeomorphic to the interior of the unit circle, $\{(x,y) \mid x^2 + y^2 = 1\}$ on the coordinate plane. Such a region cannot pass completely around a handle (nor, equivalently, around the hole formed by a handle) nor can it contain a crosscap. If a region contains a crosscap, a simple closed curve can be in the region and encircle the crosscap; this curve cannot be continuously shrunk to a point. The presence of a crosscap also allows the existance of a simple closed curve lying entirely in the region and not separating it, Figure 2.10. This is contrary to the Jordan Curve Theorem if the region is homeomorphic to the unit disk.

In Figure 2.8 one can also see that a closed curve cutting the line AB; CD, does not separate the surface. Because of the Jordan curve

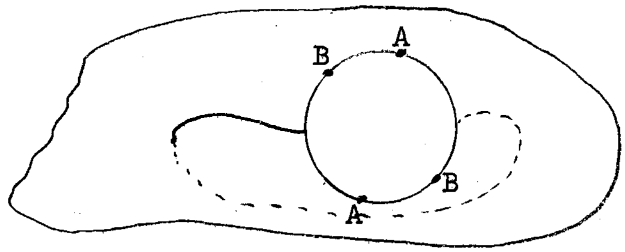


Figure 2.10

property of a disk, for a connected map, any arc connecting two points (or a loop from a single point) or the boundary of a region and lying in the region, separates the region into two simply-connected regions and forms a new connected map.

These same properties of a simply-connected region give us the following theorem:

Theorem 2.1. The edges and vertices of a connected map form a connected graph.

This result is restated and proved as a lemma in the next chapter. Although the context restricts the readers attention to the sphere, the proof being based upon the simply-connectedness of the regions holds for more general surfaces.

We shall now consider a result proved by Leonhard Euler: If V is the number of vertices of a convex polyhedron (in three space), such as a cube or tetrahedron, and E the number of edges and F the number of faces, then $V - E + F = 2$.

Since the surface of a convex polyhedron is homeomorphic to the sphere with a map on its surface, with the same number of vertices, edge and faces, we will state the theorem as follows: If M is a connected map on the sphere with $V > 0$ vertices, E edges and F regions then $V - E + F = 2$. The result has been generalized by more recent authors into the following:

Theorem 2.2. If M is a connected map on a surface of connectivity with $V > 0$ vertices, E edges, and F regions then $V - E + F = \chi$. **Chi** is called the Euler-poincare characteristic of the surface and is equal to $3 - h$, where h is the connectivity of the surface. Furthermore, $\chi = 2 - 2p - q$ where p is the number of handles and q the number of crosscaps placed on a sphere to obtain the surface.

We shall use the following lemma to prove this result:

Lemma: If edges (and vertices) are added to a connected map M with E edges, V vertices, and F regions, so as to form a connected map M' of V' vertices, E' edges and F' regions, then $V - E + F = V' - E' + F'$.

The edges can be added singly so as to obtain a connected map at each stage. In doing this one must have the edges of the map form a connected graph with each addition.

Consider first the addition of an edge between two existing vertices. This edge must lie entirely within a single region and therefore separates that region into two simply-connected regions. As a result $V' = V$, $E' = E + 1$ and $F' = F + 1$ so that $V - E + F = V' - E' + F'$.

If a point of a boundary, which is not a vertex, is considered a vertex of multiplicity two, we see that one vertex is gained and one edge is created thus not changing the value of $V - E + F$. Also if a point interior to a region is connected by two arcs to two vertices (multiplicity ≥ 2) of the map, the region is separated giving us a new map with V + 1 vertices and E + 2 edges and F + 1 regions. However, $(V + 1) - (E + 2) + (F + 1) = V - E + F$ and the lemma follows.

Let M and M' be two connected maps on a surface. Let M have V vertices, E edges, and F regions and let M' have V' vertices, E' edges, and F' regions. Form a new map M" from M' by adding new edges and vertices so that M" contains a submap that is homeomorphic to M, allowing vertices of order two (which may be removed). To do this consider a submap of M' that is homeomorphic to a submap of M; it may be quite simple. (If such a submap is not readily apparent, edges and vertices may be added to M and M' before constructing M".) Add all the edges of M that are not in this submap (or their homeomorphic equivalents) to M' as new edges forming M". Since M" was obtained from M' by the addition of new edges and since M" can be obtained from M by the addition of new edges, $V' - E' + F' = V'' - E'' + F'' = V - E + F = C$ where C is a constant peculiar to the surface in question.

On a surface of connectivity $h > 1$, consider a set of $h - 1$ arcs meeting at one vertex and forming the boundary of one simply-connected region on the surface. Now if each arc is split, say, so that we have $2(h - 1)$ arcs giving us a connected map of $1 + (h - 1)$ regions and one vertex, then we have $1 - 2(h - 1) + 1 + (h - 1) = C$. We see that for each surface of connectivity $h > 1$ the constant is $3 - h$.

In the case where $h = 1$ we consider a map of 2 vertices and three edges, forming 3 simply-connected regions. This map yields a constant of 2 which is again $3 - h$.

CHAPTER III

A HISTORY OF THE PROBLEM

It was noticed by English cartographers that it had never been necessary to use more than four colors to properly color any map. In 1850, it occurred to Francis Guthrie, a student of mathematics at Edinburgh, that if this were really so, it would be an interesting mathematical theorem. He discussed this idea with his brother Frederick (later to become professor of chemistry and physics at the newly created School of Science, South Kensington), who communicated it to Augustus De Morgan, his teacher. De Morgan, Professor of Mathematics at University College, London, and founder of the London Mathematical Society, was in a position to make many capable mathematicians aware of the problem and did so when a solution proved evasive. It was through De Morgan that the famous mathematician Arthur Cayley learned of the problem.

Cayley is an important figure in the history of mathematics, and deserves more than a casual reference. The extensive and fruitful research of this indefatigable worker (whose collected mathematical papers comprise 13 large quarto volumes) attests to his versatility and energy. This is to be wondered at, because for many years his professional interest was directed elsewhere. Arthur Cayley was born in 1821 in Surrey. He grew up in St. Petersburg, where his father was a merchant. In 1838 he went to Cambridge, where he took Firsts in Old English Usage; by the time he was 20 he began to publish in the field of mathematics.

From 1843-1863, he practiced law in London, and it is notable that it was during this time that he published his most significant mathematical papers. In addition to his papers on algebraic geometry, Cayley also published works on mechanics, astronomy, and many other subjects. His research on the theory of topological graphs deserves mention here because of its connection with the map-coloring problem. The theory of graphs is also important in other fields: in determining the number of possible isomers in organic chemistry, in Kirchoff's theory of networks, etc.

It was through Cayley that the bulk of the mathematical world became aware of the problem. He proposed it to the London Mathematical Society in an address published in the Proceedings of that Society in 1879. When the renowned Cayley confessed that with all his efforts he had been unable to prove the conjecture, he stimulated other mathematicians to attempt a solution. One of these was A. B. Kempe.

Kempe was born in Kensington on July 6, 1849. He was the third son of the Reverend John E. Kempe, Rector of St. James's Cathedral, Piccadilly. He received the M.A. degree with honors in mathematics from Trinity College, Cambridge. Although much of his interest and efforts were given to his church, he found time to cultivate his recreative interests in mathematics and music. In 1879 he published his first proof of the four-color theorem in the <u>American Journal of Mathematics</u>. W. E. Story, an editor for the journal, noted several errors in this proof that could be corrected, and he did so in a paper immediately following Kempe's. In Story's words, "Mr. Kempe has substantially proved the fundamental theorem which has been so long a desideratum, by a very ingenious method, but it seems desirable to make the proof absolutely rigorous, and

I have endeavoured to do this." [49] In 1880 Kempe published another proof of the theorem in Nature magazine, [50]. This proof used a different technique and contained none of the errors Story had noted in his first proof. In Chapter Four, these proofs will be given in reverse order; all the corrections in Kempe's first proof that were made by Story are made as the proof is presented. These proofs appear to have been accepted as valid by all concerned (including Felix Kline) until 1890, when they were refuted by Percy John Heawood in his first paper which was published in the Quarterly Journal of Pure and Applied Mathematics [39].

Heawood was born in September, 1861 at Newport, Shropshire, the eldest of four sons of the Reverend J. R. Heawood, who was rector of a church near Ipswick. In 1880 he went up to Oxford with an Open Scholarship from Exeter College. He stayed in Oxford until 1887 when he became lecturer in mathematics at the Durham Colleges, later Durham University.

Heawood's mathematical career at Oxford was extremely distinguished. He obtained a First class in Mathematical Moderations in 1881 and a First class in Mathematical Finals in 1883. He was awarded the Junior Mathematical Scholarship of the University in 1882 and the Senior Mathematical Scholarship of the University and the Astronomical award in 1886. In addition, he obtained a Second Class in Classics in 1885. He became a B.A. in 1883 and an M.A. in 1887. It was in this year he went to Durham.

His most prominent contributions to mathematics were those concerned with the coloring of maps; he was the chief architect of this branch, the central subject of which is the four-color problem.

Heawood's first paper was not just destructive in nature. It is undoubtedly the greatest contribution so far made to the mathematical theory of the coloring of maps. The paper gave several remarkable generalizations of the problem, as well as their rigorous proofs. The most noteworthy of these generalizations was that for $h > 1$ the chromatic number of a map on a surface of connectivity h is at most N_h, where $N_h = [1/2(7 + \sqrt{24h - 23})]$; ([x] denotes the integral part of x.)[27]

In March, 1879, just after the publication of Kempe's proof, P. G. Tait gave a proof to the assembly of the Royal Society of Edinburgh; however, later that same year he published a retraction for he had noted his proof was not complete [55]. He briefly described how the proof could be corrected. He based his proof on the conjecture, which he felt he had shown, that all the edges of any convex polyhedron with triple vertices could be traversed by one circuit. Tait's conjecture, as it is now called, would imply the four-color conjecture. Unfortunately, the four-color conjecture does not imply Tait's conjecture. In 1940, Franklin showed by non-convex example that if Tait's conjecture were true, then some condition implied by convexity is necessary. In 1946, W. T. Tutte of Cambridge gave a convex counter-example to Tait's conjecture; hence, it is false.

There is a bit of confusion about Tait's conjecture. The following statement, also proposed by Tait, is a result of Tait's conjecture, and has come to be known by the same name. Every bridge-less regular graph of degree three and genus zero separates into three factors. This statement is equivalent to the four-color conjecture and implies Petersen's theorem which we shall consider in Chapter Nine.

The general approach to the four-color problem has been that of trying to determine the character of a minimum irreducible map. Such a map, if it exists, is not colorable in four colors; and any other map not colorable in four colors has at least as many regions. These requirements mean the map cannot contain any reducible configurations. If one were to discover enough of the nature of minimum irreducible maps he might be able to conclude that they do not exist and thus conclude that the four-color conjecture is true.

With this in mind George David Birkhoff, Professor of Mathematics at Harvard, published an article in 1913 [6] in which he gave several reducible configurations. He showed that a minimum irreducible map had no regions of less than five sides, that each vertex had multiplicity three, and that each region was simply-connected.

In 1920 Philip Franklin, Professor of Mathematics at Massachusetts Institute of Technology, showed several more configurations were reducible and with these reductions managed to prove that every irreducible map had at least 26 regions [32]. Five years later C. N. Reynolds succeeded in replacing the number of 26 by 28, and gave an example which showed that the number could not be raised further without additional reductions [53]. He was able to simplify his argument by employing certain reductions due to A. Errera.

In 1938 both Franklin and C. E. Winn came up with several new reductions which enabled Winn to show that any map of less than 36 regions was colorable in four colors [34, 65].

Franklin used a much simpler argument to show that any map of less than 32 regions was colorable in four colors [34].

In the September 1960 issue of Scientific American, Martin Gardner stated that the four-color conjecture had been proven for all maps of 37 or fewer regions. Unfortunately he quoted no reference and an extensive search of the literature fails to turn up any verification of this statement.

There have been a multitude of fallacious proofs of the four-color conjecture, and many more will undoubtedly follow. The bulk of these have been honest attempts and have proved a result that is often confused with the four-color conjecture. This result states that there is possible no more than four mutually adjacent regions on the surface of a sphere. For an extensive discussion of this problem see Chapter 14 of [56].

In the past two decades there have been many contributions to the mathematical theory of coloring maps. Most of these were results in the more general field of graph theory. A highly significant result was published in 1956 by G. A. Dirac [26]. In this paper, he gave proofs of the following three theorems, the first of which we shall consider in detail in Chapter Seven.

For $h = 3$ and for $h \geq 5$ a map on a surface of connectivity h with chromatic number N_h always contains N_h mutually adjacent countries. As before N_h is the number obtained from Heawood's color formula.

A 6-chromatic map on a surface of connectivity 2 (projective plane) contains 6 mutually adjacent countries, or a map containing 6 mutually adjacent countries can be obtained from it by deleting suitably chosen boundary lines and uniting those countries which they separate.

If a 7-chromatic map is, on a surface of connectivity 4, either it contains 7 mutually adjacent countries or a map containing 7 mutually adjacent countries can be obtained from it by deleting suitably chosen boundary lines and uniting those countries which they separate.

This relates the problem, on any surface but the plane or sphere, to the problem of determining the maximum number of neighboring domains possible for that surface. These numbers have been obtained for $h = 3,5,7,9,11,13,15$ by L. Heffter [46], for $h = 2$ by H. Tietze [56], for $h = 4$ by I. N. Kagno [48], for $h = 6$ by H. S. M. Coxeter [17], for $h = 8$ by R. C. Bose [10] and for $h = 10,12,14$ by Heffter [26]. In each of these cases the numbers agree with the number N_h, and the proofs in some cases are of a general enough nature to hint that the number N_h is not only an upper bound for the chromatic numbers of maps on a surface of connectivity h but that it is in fact the least upper bound. However, one exception is noted in Chapter Nine.

Since the four-color conjecture has not been established, in a mathematical sense it is not known whether it is true or not. However, practically all those who have worked on the problem are inclined to guess that it is true.

An admittedly crude probability argument was given by Heawood [35]. The argument indicates that the probability of being unable to color a map of F regions and V vertices with four colors is of the order of $(1 - 3^{-F})^{2^V}$, or when F is fairly large $e^{-(4/3)F}$, approximately. This is less than 1 in $10^{10,000}$ when F exceeds 35, so that if the probability argument were valid, and uncolorable maps exist, they should not be easy to find.

We have called the map coloring problem a topological problem. Although topology has developed enormously during the past few decades, when the original four-color problem was first proposed there was no field of topology. There were isolated problems, but no discipline as such. Since then an extensive theory has been developed by many respected representatives of the field from all nations. Many difficult and far-reaching problems have been proposed and solved. Yet the modest problem of map-coloring has withstood all efforts at solution.

We cannot tell at this time whether, when a solution is found, the methods used will have wider mathematical significance. If this should turn out to be so, then the significance of the map-coloring problem in the history of mathematics will be greater than it already is.

CHAPTER IV

TWO AND THREE COLOR MAPS

The question of what maps can and cannot be colored in two or three colors has been successfully answered. This chapter will present some of the more general theorems concerning maps in two or three colors.

Theorem 4.1. Let n straight lines be drawn in the plane. The map of simple connected regions formed by these lines can be properly colored in two colors.

This can be shown by induction. It follows from plane geometry that one line drawn in the plane divides it into exactly two distinct regions, so such a map can be properly colored in two colors. Now suppose every map formed by k straight lines in the plane can be properly colored in two colors and consider any map M formed by k + 1 straight lines. Choose one of these lines, l, and erase it. By hypothesis, the resulting map, M*, can be properly colored in two colors. Now replace l. The line l divides the plane into two half planes, each of which is properly colored in two colors. If the colors of the regions in one of these half planes are interchanged, the half plane is still properly colored in two colors.

If two neighboring regions of map M are in different half planes, they border each other along a segment of the line l, but of no other line. The new regions are formed by the disection of some region of the

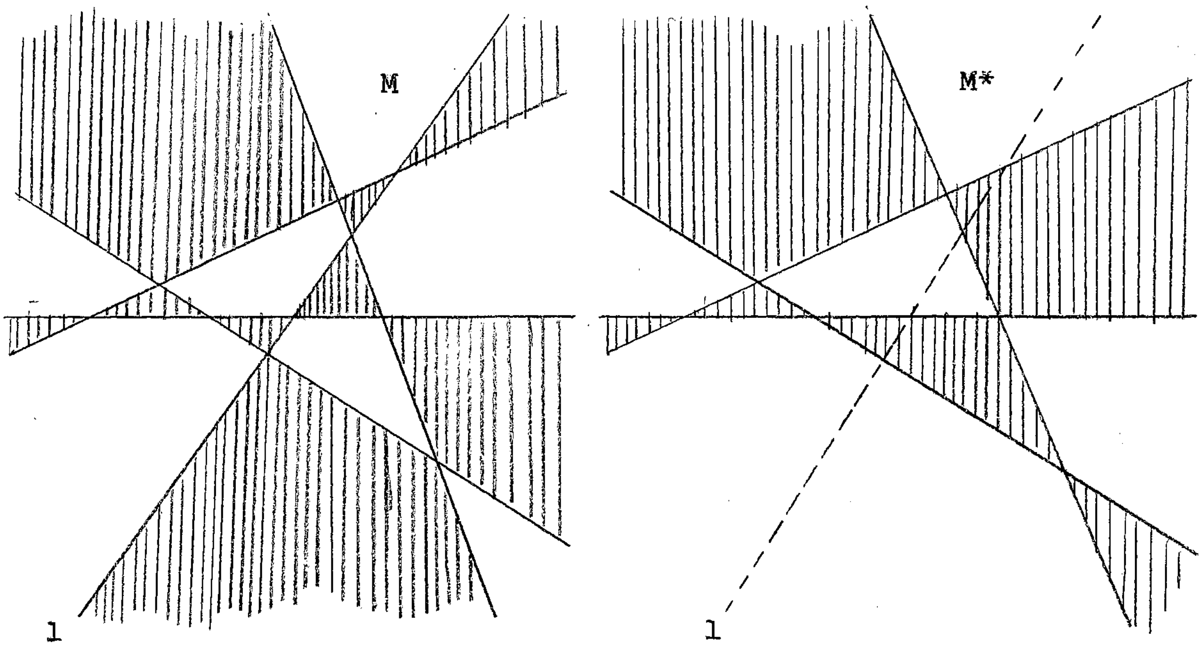


Figure 4.1

map M* by the line l, and are made different colors. If two neighboring regions of map M are in the same half plane, they must have different colors since the half planes are each properly colored. Thus, if every map formed by k straight lines in the plane is properly colorable in two colors then every map formed by k + 1 straight lines in the plane can be properly colored in two colors.

The following theorem is equivalent to the one just proven.

Theorem 4.2. Let n circles be drawn on the sphere such that they all have a common point, P, of intersection. The map of simply connected regions thus formed is properly colorable in two colors.

The equivalence of this theorem to Theorem 4.1 can be seen if P is used as the center of a stereographic projection of the sphere onto a plane.

Theorem 4.3. Let n circles be drawn in the plane. The map formed by these circles can be properly colored with two colors.

This theorem can be proved exactly as the preceeding theorem, since no two circles can have an arc in common. However, consider the following proof:

Let g be a function defined on the regions of the map, such that g(x) is the number of circles which region x lies interior to. Now if x_1 and x_2 are neighboring regions, $|g(x_1) - g(x_2)| = 1$. This follows since the boundary of x_1 and x_2 is part of some circle and if the circle were removed, then $g(x_1) = g(x_2)$ since x_1 and x_2 would be parts of the same region. (Note that there are at least four regions at a point of tangency of circles.)

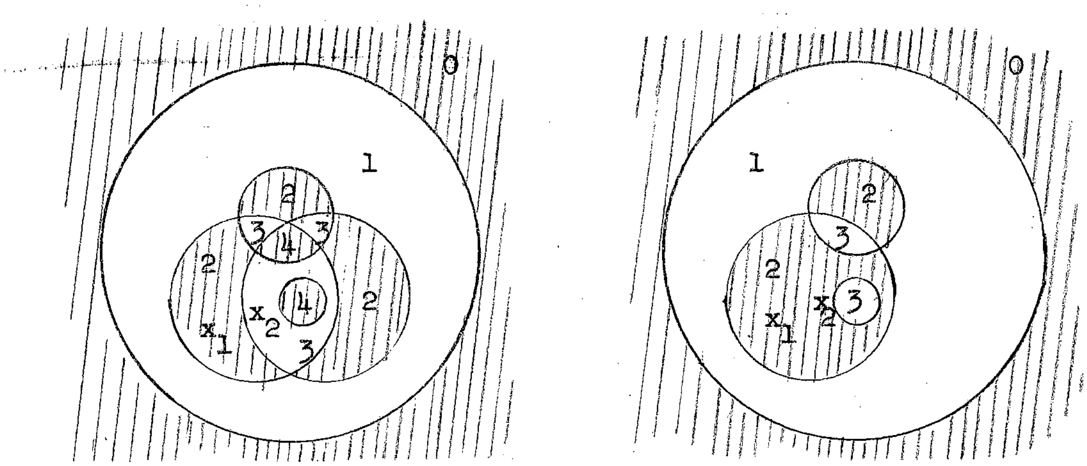


Figure 4.2

Now, if the circle is replaced, exactly one of the regions x_1 or x_2 lies on its interior and one on its exterior.[1] Thus, one of these

[1]Appendix: Jordan Curve Theorem

regions will lie on the interior of exactly one more circle than does the other. Thus, it is sufficient to color regions whose g value is even, one color, and those whose g value is odd, another color. In this way no two neighboring regions have the same color and the map is properly colored.

Lemma: If all but one of the vertices of a map are of even multiplicity, then that one is also of even myltiplicity.

Suppose that vertices V_1, V_2, V_3, . . ., V_n have multiplicities K_1, K_2, K_3, . . ., K_n respectively. Each edge contributes 1 to the multiplicity of two, not necessarily distinct, vertices. Since there is an even number of ends of boundaries, (twice the number edges), $K_1 + K_2 + K_3 + \ldots + K_n$ is even. If we know all but one of the K_i's are even then it follows that that one is also even.

This lemma will be used in the next theorem to allow us to ignore the problem of the multiplicity of the ideal vertex. That is, the multiplicity of the ideal vertex is even if all the vertices of the finite plane are of even multiplicity.

The following is a considerably more general theorem concerning maps of two colors.

Theorem 4.4. A map of simply connected regions on the plane or sphere can be properly colored with two colors if and only if all of its vertices have an even multiplicity.

The "only if" part follows quite readily from the contrapositive and an argument typified by the following figure. That is, if a vertex is not of even multiplicity, the map cannot be properly colored in two colors.

The converse, however, is not so simple. To prove this, pick an arbitrary vertex and suppose we travel over the boundaries and vertices

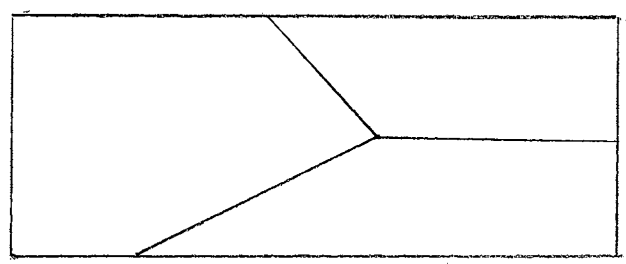


Figure 4.3

of our map. In doing so we shall leave a vertex along a boundary different from that on which we arrived. Thus, having arrived at a vertex (not already included), we can always leave since each vertex is of even multiplicity. We continue this process until we reach, for the first time, a vertex, A, which we have met before. The boundaries traversed between these two meetings of the vertex A, constitute a simple closed curve or contour. Delete this curve from the map. Now the vertices that do not belong to the closed contour are of their previous multiplicities, and those that do are reduced in multiplicity by exactly two. It now becomes necessary to consider vertices of multiplicity 2, in the graph of remaining edges. In any case, vertices are still even and the process can be repeated. In this way, it can be seen that the map is made up of overlapping simple closed curves. Define a function g as in Theorem 4.3. This function will assign a number to each region. This number will be the number of simple closed curves (of our construction) on whose interior that region lies. Just as before, the function will dictate a proper coloration of the map in two colors.

Once the map is colored, it can be seen that the evenness or oddness of each region is independent of the way the simple closed paths are chosen and is related to the color assigned a region whose function value is zero. However, the actual number assigned to each region does depend on the closed curves chosen as the following example illustrates.

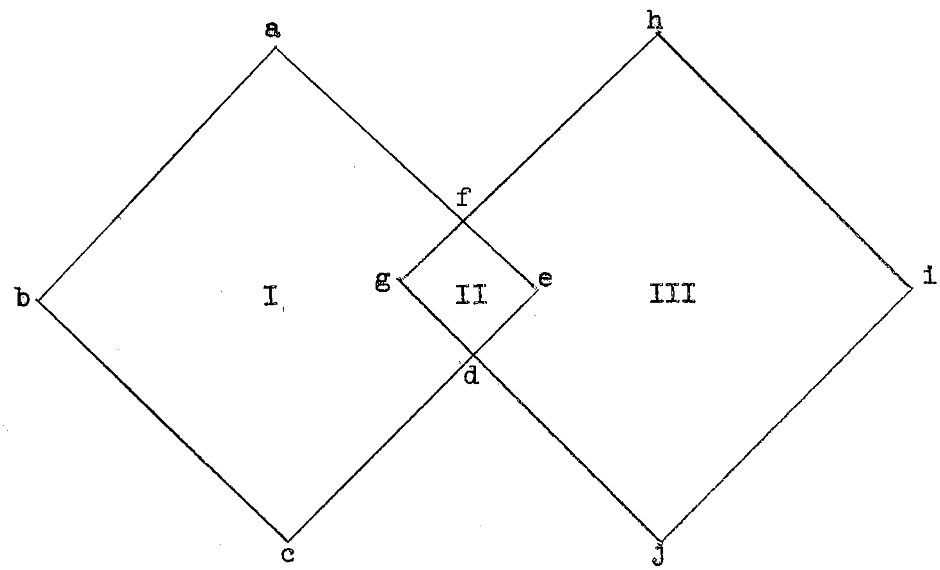


Figure 4.4

Region II may be given numbers 0 or 2 depending on whether we choose the two paths afgdcb and hfedji or the two paths afedcb and hfgdji.

Corollary 4.5. A map on the plane or sphere can be properly colored with two colors if and only if all of its verticies are of even multiplicity.

To prove this, consider a map containing a multiply-connected region

R, such that all other regions of the map are simply connected and each vertex has even multiplicity.

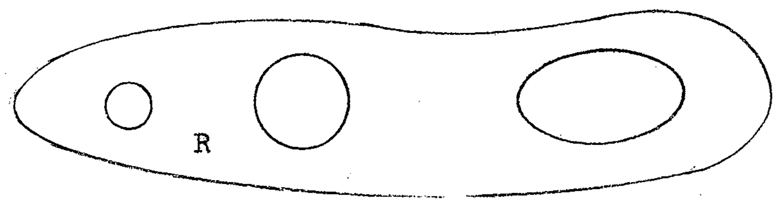


Figure 4.5

The complement of this region is several disconnected portions of the plane. If all but one piece is erased, that piece, on the plane, forms a new map which satisfies the hypothesis of the theorem just proved, since no sides or vertices or multiplicity of vertices of this piece are affected by the erasure. Thus, it is colorable in two colors. The new map for each such piece is colorable in two colors, and, conversely, being colorable in two colors implies that each new map has vertices of even multiplicity. By merely interchanging the colors if necessary, we can make the exterior region (that corresponding to the multiply-connected region) of each new map the same color. Now, if these maps are all placed with the original multiply-connected region, we obtain a proper coloration of the original map. By induction on the number of multiply-connected regions in the map, the corollary is proved.

The next theorem is based upon the concept of dual polyhedral maps, which we now define. A polyhedral map M* is said to be the dual of a polyhedral map M, on the same surface, if the two maps play completely symmetric roles, one to the other under the following conditions:

1) Each edge of one map intersects exactly one edge of its dual in precisely one point.

2) In the interior of every region of one map, there is exactly one vertex of its dual.

Immediately following from the definition we get;

3) There is a one-to-one correspondence between the vertices of one map and the regions of its dual such that neighboring vertices correspond to contiguous regions.

4) If one of the vertices of one map has multiplicity, K, then the region in the dual map corresponding to this vertex has K edges.

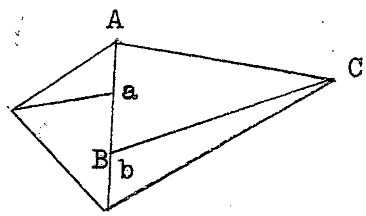
For every polyhedral map M there exists a polyhedral map M*, on the same surface, that is its dual. We must show that the graph formed by the dualizing process forms a polyhedral map. The process divides each region of M into quadrilaterals. Regions of M* are formed by regrouping the quadrilaterials such that each group completely surrounds and shares a common vertex of M. The union of the quadrilaterials of each group yields a simply connected region. M* contains no isthmus and each boundary edge has two ends, since each boundary edge of M has two distinct ends and M has no isthmus.

Let us also make note of the following definition:

Definition: If a polygon in the plane is partitioned into triangles in such a way that any two triangles either have no point in common, or have a common vertex, or have a common side, such a partition is called a <u>triangulation</u> of that polygon.

It is possible to triangulate the entire plane by considering infinite boundaries all meeting at an infinite vertex.

The following figures show that not all subdivisions of a polygon into triangles are triangulations of the polygon. The boundary a,b in each case is not a side of the triangle ABC but only a portion of a side.



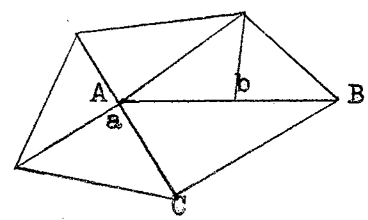


Figure 4.6

That is, there are two triangles in each case that have more than a vertex (i.e., have a boundary) in common but do not have a side in common. However, in a map ABC would be considered a quadrilateral instead of a triangle.

Suppose that a map consists of a triangulated polygon in the plane and that the map is properly colored in two colors. We shall show that the dual map can be properly colored in three colors.

Lemma: There always exists a sequence of edges or boundaries forming a path between any two vertices in a connected map.

For suppose there exist two vertices A and B such that there is no such path from A to B. Consider S_v to be the set of all vertices reachable from A and S_e to be the set of all edges that have an end in S_v. Let S_r be the set of regions of the map such that at least one edge of each is in S_e. Let U be the union of all the regions in S_r. Consider U' (the complement of U). If U' is non-empty, U borders U' along the border of some region in S_r, and some edge of this region is in S_e. Thus some vertex of this same region is in S_v. Since all the vertices of this simply connected region are reachable from this vertex, then all are in S_v. This implies that each edge of this region is in S_e and is a border of U'. That is, a border of U' is in S_e. But this implies one region of U' is in U. Now (from the finiteness of the map) $U' \subset U$, which implies $U' = \emptyset$. Thus every region is in U and each edge is in S_e, which implies all vertices are in S_v. Consequently, no such pair of vertices exists and there will always exist a path from any A to any B in a connected map.

Now mark each boundary of the properly colored triangulated-plane-polygon map with an arrow, such that there will be one color (let us say black) always to the right of the arrow and the other color (white) to the left. It is possible to go from any one vertex to any other and always travel the boundaries in the direction of the arrows. To see this, take two arbitrary vertices A and B. There exists a path from A to B. If any segment of this path is traversed against the direction of the arrows, the path may be altered by going about the other two edges of either of the two regions that have that edge as a boundary. Such alteration yields a path meeting our requirements.

(Note: It has not been required that the path not intersect or reuse a side; there is such a path possible by removing loops consisting of circuits of edges.)

Consider a path with n sides that returns to where it started and does not retrace itself. If the path crosses itself at a vertex, it forms at least two loops, having K_1 and K_2 sides, respectively. Let us show that K_1 (and similarly K_2) is divisible by three. The loop has p

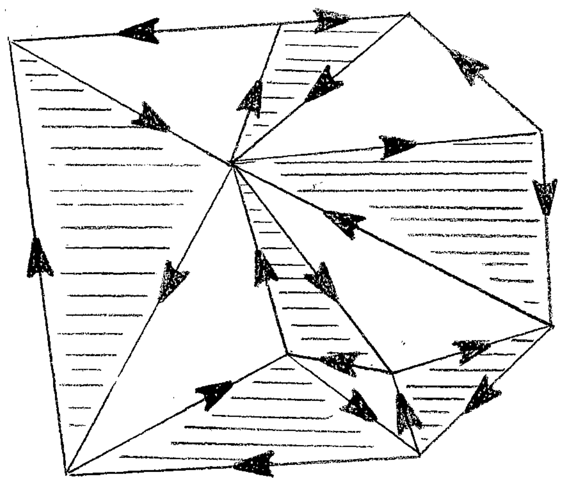


Figure 4.7

triangular regions that lie on its interior. Not more than K_1 of these will have a side of our path as one of its boundaries. These regions will all be of the same color since they all lie to the right or the left of our path depending on whether our path is directed clockwise or counterclockwise. Now disregard all boundaries exterior to our path. The one new exterior region is bordered by regions of one color, and may be colored the opposite color. Now we have a new map with several triangular regions and one exterior region with K, sides. Suppose we have S triangular regions with the same color as the exterior region with K_1 edges; these regions have three boundaries each. Suppose also that we have T regions of the color opposite of that of the exterior region; each of these regions has three boundaries. Each boundary belongs to one black and one white region, so $K_1 + 3S = 3T$. Hence, K_1 is divisible by three. A similar argument holds for any K_i. Since n is a sum of loops with K_1, K_2, ... boundaries with no boundary used more than once, n is also divisible by 3.

If a path does retrace itself, we merely need to count the boundary each time it is used. In this manner the foregoing proof still holds and a path of n edges (not necessarily distinct) that returns to where it started will necessarily have n divisible by three.

Now let A be an arbitrary vertex of our original map, assign it the number zero. Let B be any other vertex in the map. There exists a path from A to B such that each boundary is traversed in the direction of the arrows. If this path contains k boundaries, assign to B the number k (mod 3).

Now let us show the number assigned to B is independent of the path chosen. Let p and q be two distinct paths from A to B, each following the arrows. Construct a path r from B to A, following the arrows. Now p,r starts at A and returns to A. The same for q,r. If p has n_1 boundaries, if q has n_2 boundaries and if r has n_3 boundaries then $n_1 + n_3$ as well as $n_2 + n_3$ is divisible by three. Hence $(n_1 + n_3) - (n_2 + n_3)$ is divisible by three and $n_1 - n_2 \equiv 0 \pmod 3$. Hence both paths will assign the same label to the point B. It is also clear that no two neighboring vertices of our map will have the same label.

To each vertex there is assigned one of the three numbers, 0, 1, 2; and these numbers are thus assigned to the corresponding regions of the dual map. Since no two neighboring regions of the dual map will have the same label, it can be colored in three colors.

The converse follows by reversing the above argument. Our original map can have its vertices properly labeled 0, 1 or 2 corresponding to the three colors of its dual. The two possible orientations of the triangles either 0, 1, 2 or 0, 2, 1 give us the map colored in two colors. Since a boundary shared by two triangles and traversed clockwise with respect to one triangle and counterclockwise with respect to the other, the map is properly colored in two colors.

The dual of a regular polyhedral map on a sphere is a triangulated polygon (triangle) so that the forgoing with corollary 4.5 is a proof of the following theorem.

Theorem 4.6. A regular polyhedral map on a sphere can be colored in three colors if and only if each region has an even number of edges.

The restriction of Theorem 4.6 to polyhedral maps can be removed. Consider a regular map M, on a sphere. A regular map on a sphere can have a region of two boundary edges, otherwise it is polyhedral. If M has three regions it is colorable in three colors. If M has more than three regions then a region of two boundary edges can be shrunk to a point (no longer a vertex) forming a new regular map M*. M* is colorable in three colors if and only if M is colorable in three colors. If M* has more than three regions then it is either polyhedral or it has a region of two edges which can be shrunk to a point producing another new map. Continuing in this fashion we will obtain a three-region map or a polyhedral map which is colorable in three colors if and only if M is colorable in three colors. Note that shrinking a two edged region to a point changes the number of edges of each contiguous region by two. That is, each region of M has an even number of edges if and only if each region of M* has an even number of edges.

An alternate procedure would be to divide each two-edged region of M into four quadralaterals, Figure 4.8. This forms a new map M* that is colorable in three colors if and only if M is colorable in three colors.

Thus we see that Theorem 4.6 holds for regular maps in general.

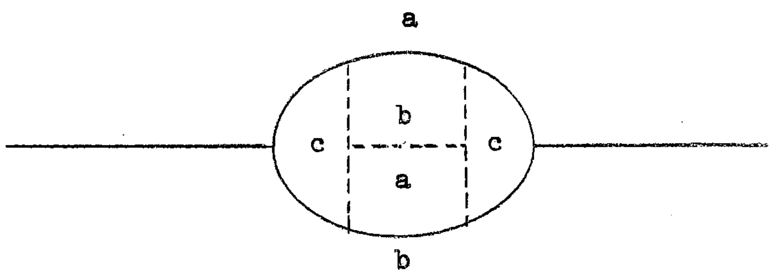


Figure 4.8

The following question is yet to be considered: When are non-regular maps colorable in three colors?

If each vertex is trivalent (of multiplicity three) the hypothesis of simply-connected is unnecessary. For suppose a map M, with each region having an even number of edges, contained a multiply connected region R, of order k. More than one such region could be handled by induction. Consider the k + 1 different maps m_1, m_2, . . . m_{k+1} each obtained from M by erasing all but one of the disjoint sections of the complement of R. Now each of the maps m_1, m_2, m_3, . . . m_{k+1} is simply connected and, being regular, is colorable in three colors. By proper permutation of the colors in each of the k + 1 maps we can have, in each case, the region that corresponds to R the same color. By superimposing these maps on one another we obtain map M properly colored in three colors. In obtaining the converse we need note that the original theorem holds for each map m_i and thus each region of m_i has an even number of edges. Hence each region of M has an even number of edges.

Thus the theorem may be restated as follows:

Theorem 4.7. A map on a sphere with each vertex of multiplicity three can be colored in three colors if and only if each region has an even number of edges.

That we cannot remove the restriction on the multiplicity of the vertices is apparent in the following examples:

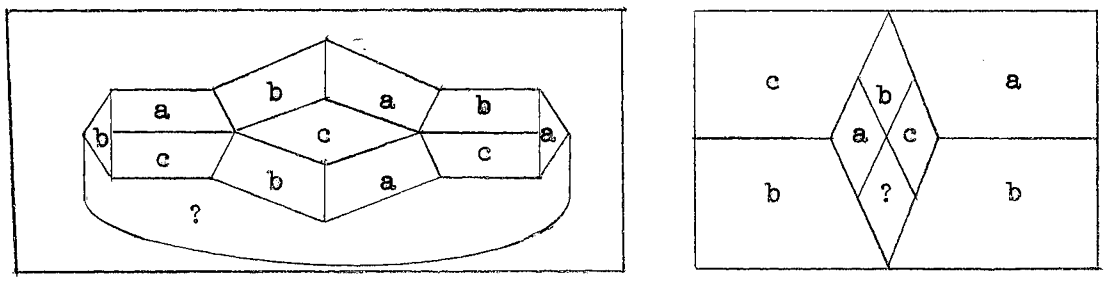


Figure 4.9

CHAPTER V

KEMPE'S PROOFS AND THE FIVE COLOR PROBLEM

Although Kempe's proofs are in error, it is worthwhile to consider them for their approach to the problem. The basic attack on the problem that has been used by many others was first outlined by Kempe in these proofs. Let us consider his second proof first since it has none of the curable errors found in the first proof. This proof was published in Nature magazine, London, in 1880 shortly after the first proof was published in the American Journal of Mathematics in 1879. We shall then consider his first proof for sake of comparison of technique.

Lemma I. Every map of more than one region on a plane or sphere must have at least one pair of adjacent regions which have only one common edge.

Suppose this were not so, then consider a pair of adjacent regions a_1 and a'_1. These two regions surround a set, S_1, of regions. If there were just one region in S_1 then it would have a simple boundary with a_1. Consider a pair of adjacent regions a_2, a'_2 of S_1. These regions surround a set, S_2, of regions. In fact we can consider an infinite sequence of pairs of regions each surrounding the next. But this is contrary to the finiteness of the map. Hence our supposition is false and the lemma is true.

Lemma II. In any map on a plane or sphere there exists at least one region of less than six edges.

If an edge whose ends lie at two different vertices is rubbed out, the multiplicity of each vertex is reduced by one, or if either vertex had multiplicity three it would no longer be a vertex of the map, further reducing the number of edges. The result is that rubbing out a boundary consisting of a single edge may reduce the number (E) of edges by three. It can; however, never cause a greater reduction, and may cause a smaller. This could happen when a loop is rubbed out or when the multiplicity of either vertex is greater than three.

Now, the obliteration of simple boundary, B, causes the two regions it separates to coalesce, thus reducing the number of regions R in the map by one. This newly formed map has a pair of regions with common boundary consisting of one edge, according to our lemma. By obliterating edges that are complete boundaries we finally get a single region, no boundaries and no vertices. Each such reduction of R by one cannot involve a reduction of E by more than three; thus $R - 1 \geq 1/3\ E$, which implies $6R > 2E$. Since $6R > 2E$ some region has less than six edges, and our lemma follows.

Suppose we have a map properly colored in four colors, blue, yellow, red, green. Consider those regions colored red and green. They form one or more noncontiguous connected sections of the map, each containing one or more regions colored red or green. These sections will surround and be surrounded by sections containing blue and yellow regions. If the colors of the regions in any arbitrary section are interchanged, it will not affect the colors of other sections of the same two colors and the map is still properly colored.

We can reduce a map to a single region by successive operations of throwing two regions into one by rubbing out the edge or edges between two regions one of which has less than six boundary edges. Conversely, we can develop a specified map, starting from a single region and adding boundaries, at each stage dividing a region into two, one of which has less than six boundary edges. Suppose at some stage of its development by this process a map can be colored with four colors (red, green, blue, and yellow). Let these colors be indicated by colored wafers placed on the regions. Proceed to the next operation; this divides a wafered region into two regions. Shift its wafer onto the region of these two which is not the one which has less than six boundaries. If both have less than six boundaries, shift the wafer onto either. If the region (w) which is left without a wafer is touched by only three colors it can be colored the fourth, but if it is touched by four colors we must take another step. This can be necessary only if w has four or five adjacent regions.

Consider the first case, in which four regions are adjacent to w. If w surrounds at least one but not all, one of these regions it surrounds can surrender its wafer to w and receive a wafer the color of one that w does not surround. It may be necessary to permute the colors of the regions surrounded by w before making the shift. To show this can be done, an argument may be used that is similar to the one in Chapter IV that shows the sufficiency of a proof for simply connected maps. So consider the four regions to surround or to be surrounded by w. Label them clockwise a,b,c,d and let a be red, b blue, c green, and d yellow. If the regions are not in a cyclical arrangement then a coloration can be obtained by a simpler argument. If starting from a, we can get to c, going through red and green regions, and not passing through any vertices, we cannot, starting from b, get to d, going similarly only through blue and yellow regions, for otherwise two tracks which pass through different regions would cross. Thus b belongs to a set, G, of blue and yellow regions which are cut off from the rest of the map by a chain (or chains) of red and green ones. We can accordingly interchange the blue and yellow wafers in G without changing any other and the map remains properly colored. This makes b yellow and we can put a blue wafer on w. Similarly, if we cannot pass from a to c, a belongs to a set of red and green regions not containing c. Interchanging the wafers in this set makes a green, b, c, and d remain unchanged; and a red wafer can be put on w. The map is now properly colored.

Similar reasoning applies in the case of five surrounding or surrounded regions: Figure 5.1. Label the regions e,f,g,h,k in a clockwise manner, making e green, f blue, g red, h blue, (two must of course have the same color) and k yellow. If the regions are not in a cyclical arrangement then one of them has at least two edges in contact with w. Such a situation would simplify the problem of finding a coloration which would permit a proper coloration of w.

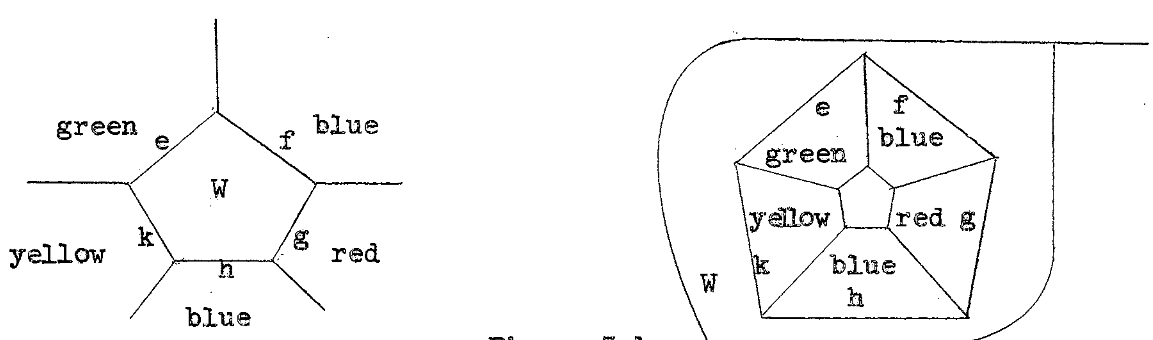


Figure 5.1

If e belongs to a different red-green section than does g, we can interchange the colors in that section to which e belongs, allowing a green wafer for w. Similarly, if k belongs to a different yellow-red section than does g, we can interchange the color in that section allowing a yellow wafer for w. If neither of these is the case, then the red-green section to which e and g belong contains a red-green chain which separates the blue-yellow sections containing k and f; i.e. they are in different blue-yellow sections. Also, the red-yellow section containing k and g separates the blue-green sections containing e and h. Now interchanging the colors in the blue-yellow section to which f belongs and in the blue-green section to which h belongs, makes f become yellow and h green, e, k, and g remaining unchanged. In any case, the number of colors adjacent to w is reduced to three and the map remains properly colored. We can place a wafer for the remaining color on w. Thus, if the map can be colored as developed at any stage, it can be colored at the next. Hence, since it can obviously be colored at the stage where it contains four regions, it can be colored at the last stage where we have our original map.

Since the map considered was arbitrary, it follows that any map on the plane or sphere can be colored in four colors.

Now let us examine Kempe's first proof.

First, consider a vertex of multiplicity three. The regions meeting there must be colored in exactly three colors.

Next consider a vertex, p, of multiplicity four with three regions. There can be only three colors at such a vertex. Next consider a vertex of multiplicity four with four distinct regions, labeled clockwise a,b, c,d, meeting there. These regions may be colored in two, three, or four colors. If they are colored in four colors then two of the regions must belong to different sections in their colors; that is, if b and d belong to the same red and green section but are not adjacent regions at the vertex, then that section surrounds a blue and yellow section containing either a or c but not both; thus a and c are in different blue and yellow sections. If we interchanged the colors in either one of these sections, we would have just three colors at the vertex p and the map would be properly colored in four colors.

Next, consider the case of a vertex of multiplicity five. The regions meeting at this vertex may happen to be colored with three colors, but they may happen to be colored with four. Figure 5.2 shows the only form which the coloring can take place in that case, one color of course occuring twice.

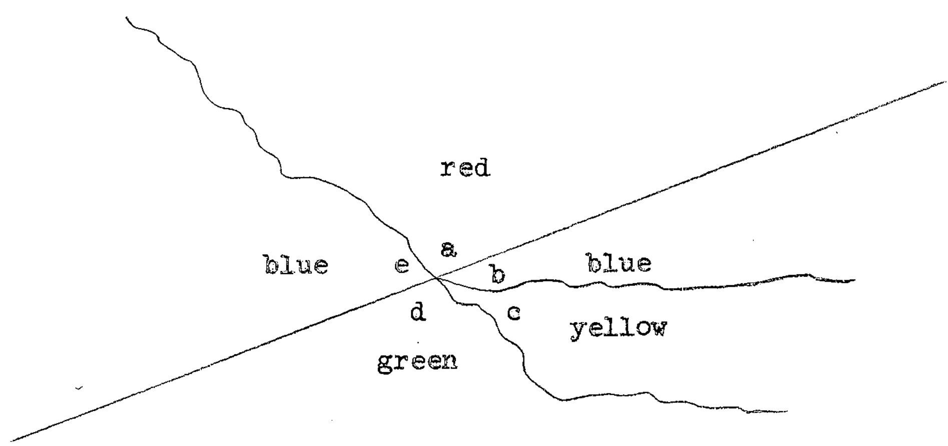


Figure 5.2

If a and c belong to different yellow and red sections, interchanging the colors in either makes a and c become the same color. If a and c belong to the same yellow and red section, see if a and d belong to the same green and red section. The two sections cut off b from e, so that the blue and green section to which b belongs is different from that to which d and e belong, and the blue and yellow section to which e belongs is different from that to which b and c belong. Thus, interchanging the colors in the blue and yellow section to which e belongs, and in the blue and green section to which b belongs makes b become green and e yellow, a, c, and d remaining unchanged. In each of the three cases, the number of colors at the vertex under consideration is reduced to three and the map remains properly colored.

What has been shown is that in any map properly colored in four colors, for any selected vertex of multiplicity less than six there exists a proper coloration of the map such that only three colors appear at that vertex.

Now consider the four-color theorem: Let M be any map on the plane (or sphere). Let r be a region of M with less than six edges. Take a piece of paper and cut it out the same shape as r, but rather larger, so as just to overlap the boundaries when laid on r. Fasten this patch to the surface and produce the boundaries which meet the patch, if there are any, to meet at a point, p, within the patch, making sure that any two boundaries that had met the region in the same vertex are joined in a vertex on the patch before they are joined to p. (If r is not simply connected the patching process is still valid. We could, of course, avoid the consideration of multiply connected regions as we did in Chapter IV.) If there are more than two boundaries so produced then the point will be a vertex of the newly formed map M_1, Figure 5.3. M_1 has one less region than M, (or fewer if we consider a multiply connected region). M_1 has a region r_1, of less than six edges and that region

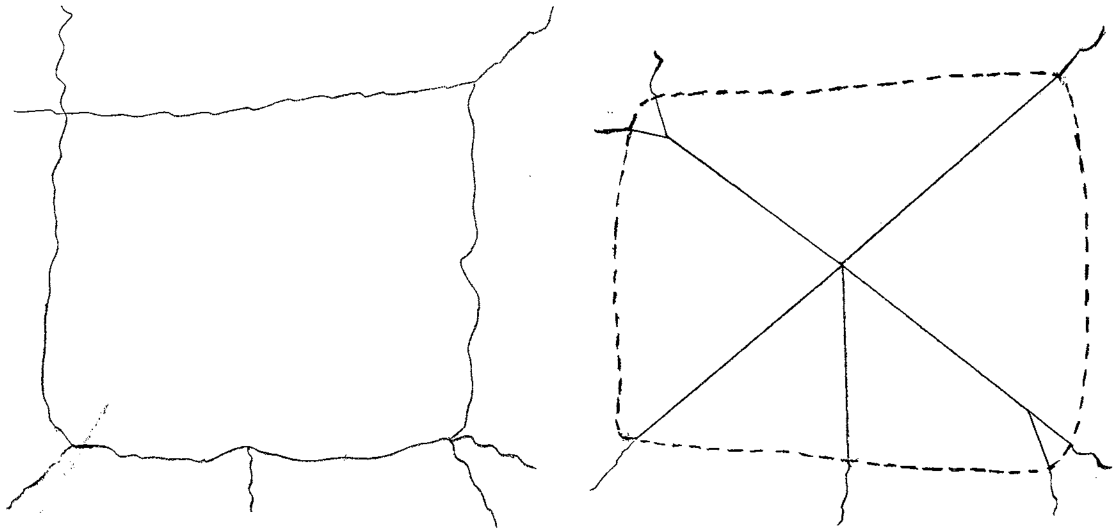


Figure 5.3

can be patched forming a map M_2. This patching process can be repeated as long as there is a region left to operate upon, the patches being in some cases stuck partially over others. The process will eventually result in a map of only one region, devoid of vertices and boundaries. This map can be properly colored in not more than four colors. Now reverse the patching process, and strip off the patches in reverse order, taking off first that which was put on last. As each patch is stripped off it disclosed a district of less than six edges. If map M_k can be properly colored in four colors then it can be properly colored in four colors such that only three colors meet on the patch next to come off. When that patch is then stripped off, the region it reveals will be surrounded by three colors and can be properly colored the fourth color. Thus, map M_{k-1} can be properly colored in four colors. Consequently, if a map M_k is colorable in four colors then map M_{k-1} is colorable in four colors; and, by induction, map M is colorable in four colors.

As presented here, the corrections in this proof due to W. E. Story have been made.

Let us consider a third proof, which involves a technique used by Heawood in proving another theorem, making a revision of the proof we have just seen.

Suppose the four-color conjecture is false. Then there exists a map (or maps) that is not properly colorable in four colors. If this be the case then there is such a map, M, with a minimum number of regions, k. That is, any map not colorable in four colors has as many regions as M. Now, there exists a region r of M with less than six edges. If r is patched, a new map M' is produced with k - 1 (or fewer) regions. This map, according to hypothesis, is colorable in four colors. Furthermore, a coloration exists, according to the foregoing proof, such that only three colors occur at the vertex on the patch. Now, if every region of M is assigned the color that is assigned to the corresponding region of M' it is properly colored except for r. But only three colors surround r so the fourth color can be assigned to r producing a proper coloration of M. Hence, M is colorable, contradicting our assumption. Thus, our assumption is false and the four color conjecture is true.

The region r and its surrounding regions are referred to as a reducible configuration making M reducible. Thus we see that the minimum irreducible map (if it exists) does not contain a reducible configuration.

It should be evident by now that any error common to these proofs would have to lie in either our lemmas or in the proof that a region of less than six edges and its surrounding regions form a reducible configuration. Each of these statements has been proven here essentially as Kempe proved them; but as Heawood pointed out there is an error in the proof of the statement that a region of less than six sides and its surrounding regions form a reducible configuration.

The error lies in the proof for the case of a vertex of multiplicity five. Suppose, Figure 5.4, there is a green region x in the blue and green section to which b belongs which borders a yellow region y in the blue-yellow section to which e belongs and, furthermore, two yellow regions bordering x and e are both contiguous to the same green region.

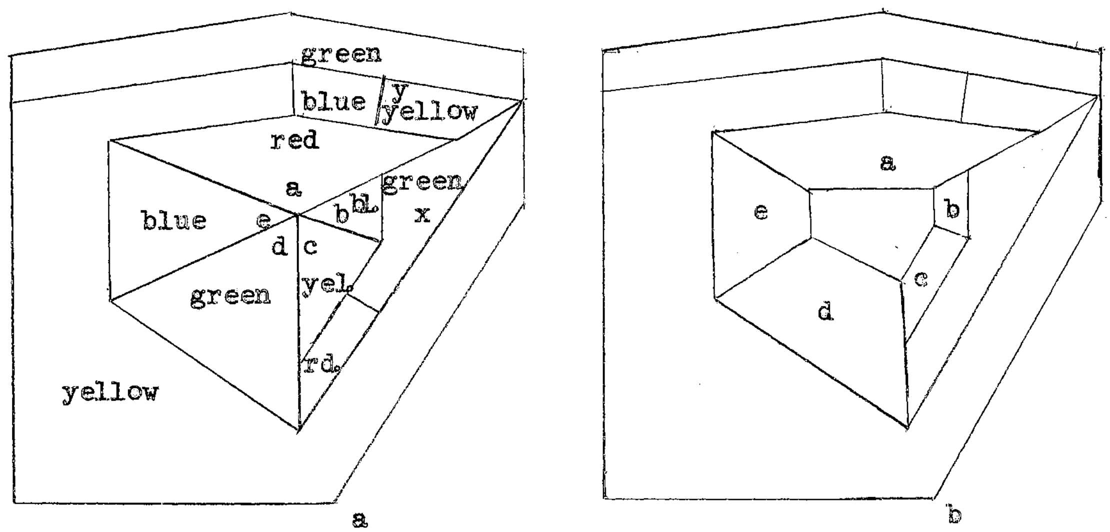


Figure 5.4

Interchanging the colors in the blue-yellow section to which e belongs will bring d into the same blue-green chain to which b belongs. As a result, interchanging the colors in the blue-green chain to which b belongs will make d blue and we still have four colors. As Heawood put it, "It is conceivable that though either transposition would remove a blue, both may not remove both blues."

Although the proof does not hold, the statement that for any vertex of multiplicity less than six there exists a proper coloration of the map such that only three colors appear at that vertex, still may be true.

The Figure 5.4a is not a counterexample of the statement for there does exist a coloration of the figure such that no more than three colors meet at the vertex in question. If there did not exist such a coloration for Figure 5.4a, then it would be impossible to color Figure 5.4b in four colors and so the four color conjecture would be proven false.

Let us now consider the five-color theorem due to Heawood.

Lemma: A region of less than six edges and its surrounding regions form a reducible configuration in five colors.

If the region, call it r, has less than five edges, the map produced by patching it has a vertex on the patch of multiplicity less than five. Coloring the reduced map dictates a coloration of the original map since at most four colors are adjacent to r, thus allowing a fifth color for it, producing a proper coloration of the map.

If the region r has five edges, the map produced by patching r has a vertex p of multiplicity five on the patch. Suppose a coloration of this map. If the five regions around r are not distinct, then a color for r is available and the lemma follows. Furthermore, if all five colors do not appear at p, a color for r is available and the lemma follows. Label the regions contiguous to r as a,b,c,d,e in a clockwise fashion, and suppose their colors are red, blue, green, yellow, and orange respectively.

If a and c belong to different red-green sections, we can interchange the colors in the red-green section containing a to yield a proper coloration with only four colors adjacent to r, thus showing the configuration reducible.

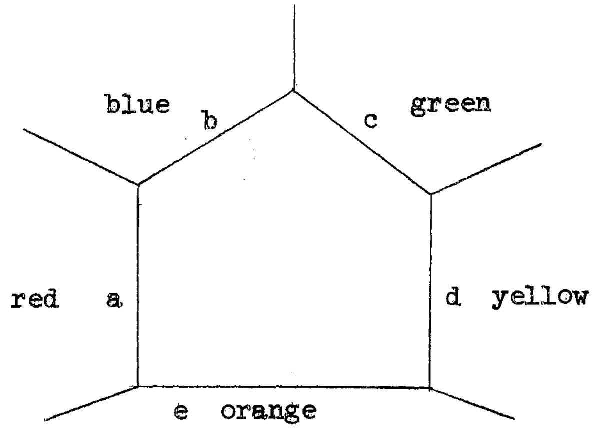


Figure 5.5

If a and c belong to the same red-green chain, the chain cuts off the blue-orange section containing b from the blue-orange section containing e. Hence we can interchange the colors in the blue-orange section containing e, making e blue and leaving a, b, c, and d unchanged, to yield a proper coloration of the map outside r. This transposition yields only four colors contiguous to r, making r and its neighbors a reducible configuration. In any case, the lemma is true.

Theorem 5.2. Every map on the plane or sphere can be colored in five colors.

Suppose there exists a map not colorable in five colors. Then there is one with a minimum number of regions; call it M. M must be irreducible in five colors. But M contains a region of less than six edges. So M is reducible, this contradicts our assumption. Thus our assumption is false and the theorem is true.

We see that the problem of proving the four-color conjecture is equivalent to showing that every map on the plane or sphere is reducible in four colors. That is, there exists no minimum irreducible map for four colors.

CHAPTER VI

REDUCTIONS

In this chapter as in the last we shall confine ourselves to maps on the plane or sphere. If a map on a plane or sphere exists that is not colorable in four colors then there is one with as few regions as any other. Such a map must be free of any configurations reducible in four colors. We shall call such a map a minimum irreducible map. De Bruijn [26] showed that the number of regions in a minimum irreducible map would be finite. Reductions are usually considered with the hope of limiting the classes of maps that could contain a minimum irreducible map to the point of concluding that such a map could not exist.

Theorem 6.1. If more than three edges meet at any vertex of a map, the coloring of the map may be reduced to the coloring of a map of fewer regions.

From the Jordan Curve theorem, not every pair of regions meeting at such a vertex have an edge in common, as is the case when three edges meet at a vertex. Thus we may join two of these regions without a common edge by opening the vertex so as to form a map of one less region. When this map is colored, the original map is obtained, properly colored, by merely restoring the vertex to its original form.

Theorem 6.2. If any region of a map is multiply-connected, the coloring of the map may be reduced to the coloring of maps of fewer regions.

For, we may color the partial maps which arise when all but one of the parts into which such a multiply-connected region separates the surface are erased, and afterward give the same color to this region in all the partial maps by a permutation of the colors. By assigning the colors of the regions of the partial maps to their corresponding regions of the original map, we obtain a coloring of the original map.

It is now evident that if the four-color conjecture holds for regular maps then it holds for all maps on a sphere and we shall henceforth consider only regular maps.

The following new terms will prove useful. A major polygon in a map is one with more than six boundary edges whereas a minor polygon refers to a pentagon or a hexagon. A triad of regions is a set of three mutually contiguous regions, all appearing at the same vertex. A cap on a ring is a region in triad with two regions of the ring (See discussion following Theorem 6.4). We shall usually refer to a cap as being next to two particular regions of a ring by the following notation: 5(5)765 means a ring of a pentagon, a heptagon, a hexagon and another pentagon with a pentagon in triad with the first pentagon and the heptagon.

The following is a list of all the known significant reductions, giving credit to their discoverers, along with several results concerning irreducible maps.

Those marked * are generalized by a later result. Those marked ** involve a combination or generalization of several different previous results.

1. (Kempe) More than three regions at a vertex form a reducible configuration.

2. (Kempe) A multiply connected region forms reducible configuration.

3. (Kempe by Birkhoff) A region of less than five sides forms a reducible configuration.

4. (Birkhoff) If two or three regions form a multiply connected region the map is reducible.

5. (Birkhoff) Four or five regions surrounding more than one region form a reducible configuration.

6. (Birkhoff) An edge having only pentagons at each vertex forms a reducible configuration.

7*. (Birkhoff) A region completely surrounded by pentagons is a reducible configuration.

8. (Birkhoff) An even edged region completely surrounded by hexagons is reducible.

9. (Kempe by Franklin) Every map containing no triangles or quadrilaterals and having three regions abutting on each vertex contains at least 12 pentagons.

10. (Franklin) A minimum irreducible map must contain a pentagon adjacent to two other minor polygons (a polygon of 5 or 6 sides).

11. (Franklin) An edge of a hexagon surrounded by this hexagon and three pentagons is a reducible configuration.

12**. (Franklin) A pentagon in contact with three pentagons and a hexagon is a reducible configuration.

13. (Franklin) A pentagon surrounded by two pentagons and three hexagons is a reducible configuration.

14. (Franklin) An even-edged region completely surrounded by hexagons and pairs of pentagons, the two of each pair being adjacent is

a reducible configuration.

15. (Franklin) A hexagon surrounded by four pentagons and two hexagons is a reducible configuration.

16*. (Franklin) A region of 2n edges surrounded by 2n - 2 pentagons and the two remaining regions adjacent is a reducible configuration.

17*. (Franklin) A region of 2n - 1 edges surrounded by 2n - 2 pentagons and any other region is a reducible configuration.

18*. (Franklin) Every map of 25 or fewer regions can be colored in four colors.

19*. (Errera) Any map containing no major polygons is reducible.

20. (Winn) A heptagon flanked by four consecutive pentagons is a reducible configuration.

21. (Winn) An even polygon bounded by two even sequences of pentagons, a hexagon and another polygon is a reducible configuration.

22**. (Winn) Any configuration bounded by a ring of an even number of pentagons together with a) two other adjacent polygons or b) one other polygon is reducible.

23. (Ratib) A hexagon surrounded by a ring n55665 is a reducible configuration.

24. (Errera) An irreducible map contains at least 6 major polygons.

25. (Winn) A hexagon surrounded by the ring mn5565 is reducible.

26. (Winn) A heptagon touching 4 pentagons and 3 hexagons in any order is reducible.

27. (Winn) An even polygon enclosed by the ring in 14 in which one hexagon is replaced by a pentagon, e.g. 55665666, is a reducible configuration.

28. (Winn) A pentagon flanked by 4 minor polygons of which the first and last minor polygons of the sequence are pentagons is a reducible configuration.

29. (Winn) A hexagon in contact with 5 minor polygons of which the first and last are pentagons is a reducible configuration.

30. (Winn) A hexagon touching two separate pairs of pentagons when each pair is in triad with another pentagon is a reducible configuration.

31. (Winn) In an irreducible map a minor polygon must touch a major polygon (a polygon of more than 6 sides).

32. (Winn) In an irreducible map a major polygon must touch either another major polygon or else 3 or more hexagons.

33. (Winn) A heptagon in contact with the chain 55655 is a reducible configuration. (The heptagon is contiguous to each region of the chain.)

34. (Winn) A pair of pentagons in triad with a heptagon and touching no other major polygon is reducible.

35. (Winn) A pair of hexagons in contact with 55655 or 556655 is reducible. (One of the pair is contiguous to each region of the chain.)

36. (Franklin) An odd-edged region surrounded by hexagons and one external pentagon 6(5)66 . . . 66, is a reducible configuration. The (5) denotes a pentagonal cap on the ring.

37. (Franklin) Two adjacent pentagons, each touching the same heptagon which together with five hexagons surrounds the pentagons, is a reducible configuration.

38. (Franklin) An odd-edged region surrounded by 5(5)66 . . . 6 is reducible.

39. (Franklin) A polygon surrounded by 5(5)76 . . . 6 is reducible.

40. (Franklin) In 39 one or more pairs of consecutive hexagons may be replaced by pairs of pentagons.

41. (Franklin) Every irreducible map must contain at least 32 regions.

42. (Franklin) Any map containing at most one major polygon is colorable in four colors.

43**. (Winn) Every irreducible map must contain at least 36 regions.

44. (S. M. De Backer) A pentagon touching two other non-consecutive minor polygons is a reducible configuration.

The most recent results found in this area are Winn's 43 in 1941 and De Backer's 44 in 1946. De Backer's results supercedes reductions 6, 7 and 11 through 40 except 18, 19, 24, 31, 32 and 36. Although his result is quite powerful, it came considerably later and was not significant in establishing the best result in connection with the number of regions in an irreducible map.

To consider here the proofs of all the reductions of our list would tend to distract from rather than add to the main theme of this paper. Most of the proofs are quite long and involve considerable case making. The proofs of the first eleven are presented, not for the sake of rigor, but rather with the intent of giving the reader examples of the way in which these proofs are made.

Theorem 6.3. If a map contains any 1, 2, 3, or 4 edged regions, the coloring of the map may be reduced to the coloring of a map of fewer regions.

If a 1- or 2-edged region be present, the map would be reducible by Theorem 6.1 or 6.2. If a 3-edged region be present, we may shrink this region to a point, color the resulting map, then introduce the region again in a color different from the colors of the regions to which it is adjacent. If a 4-edged region be present, at least one of the pairs of regions which abut on opposite edges are separated by other regions, i.e., have no common edge. Let the two opposite regions of this kind coalesce with the 4-edged region. If the resulting map in two fewer regions is colored, the original map may be colored by inserting the 4-edged region in a color different from that of the two or three colors of the regions adjacent to it.

Theorem 6.4. If two or three regions of a map form a multiply-connected region, the coloring of the map may be reduced to the coloring of maps of fewer regions.

In fact, these regions separate the surface into two or more parts each with fewer regions (counting the two or three regions in common). If we color the partial maps which arise when all but one of these parts are erased, and by a permutation of the colors then make the colors of the two or three regions the same on each of the partial maps, the original map may be colored by a natural correspondence to the partial maps.

Our quest for an irreducible map has now narrowed to the consideration of regular maps in which no two or three regions form a multiply-connected region, and every region has at least five edges. Any map now under consideration is a polyhedral map.

Consider now in a regular map a cyclical arrangement of n regions $r_1, r_2, \ldots, r_n$ such that each of these regions has a boundary in common with the one preceeding and the one following it in the cyclical order, but with no other region of the set. We shall call such a set of regions a ring of regions.

A ring R of regions of this kind divides the maps into two sets of regions M_1 and M_2 which together with R make up all the regions of the map M.

The partial map $M_1 + R$ consists of the regions in M_1 and the regions in R. $M_1 + R$ and $M_2 + R$ are bordered by the ring R of n regions. If it is possible to color $M_1 + R$ and $M_2 + R$ so that the arrangement of colors on R is the same in both cases, it is clearly possible through a natural correspondence to assign the colors to the map $M = M_1 + M_2 + R$. If the arrangement of colors on R is the same except for a permutation of the colors, then the scheme of colors on R is the same and through a permutation of colors it is again possible to color M.

Two colorations of $M_1 + R$ are said to have the same scheme on R if the regions of R have the same colors or if the colors of the regions of R can be made the same by permuting the colors of one of the colorations.

Now consider any pair of regions on the ring R in the partial map $M_1 + R$. Suppose one is colored a and the other colored either a or b. If these two regions belong to the same a-b "Kempe Chain" in M + R, this fact may be indicated by joining the two regions by a a-b line, marked a-b, lying within the chain. Note that being on the same a-b line is a symmetric and transitive relation (an equivalence relation for this pair of colors), and that either of two regions of R on the same a-b line determine the same equivalence class of regions.

From a given coloring on $M_1 + R$ we can derive a second coloring by transposing a pair of colors on the entire partial map or on one or more chains in those two colors. In the first case the scheme on R is unchanged. In the second case the scheme on R may be changed.

Suppose that in M we replace $M_2 + R$ by a set of regions M_2' of not more than a certain number k of regions but with the same peripheral boundary edge as R has with M_1, such that $M_1 + M_2'$ is regular. Two edges of this boundary will be adjacent boundaries of the same region on the periphery of M_2' if and only if they are adjacent edges on the same region of R. As a result of this construction, a coloration for M_2' will dictate a coloration for R. In this manner we form a map $M_1 + M_2'$, which has fewer regions than M if k is smaller than the number of regions in $M_2 + R$. We assume $M_1 + M_2'$ to be colorable. For a particular choice of M_2' we get a finite number of choices of colors on R which are consistent with colorations for M_1, and hence a set of schemes for $M_1 + R$.

Likewise, under similar assumptions, by forming a partial map M_1' and then $M_2 + M_1'$ we get a set of schemes for $M_2 + R$.

If it can be shown that for suitably small k there is a scheme in common to both sets, the configuration is reducible. The coloring of $M_1 + R$ and $M_2 + R$, such that the ring R has the same colors in each coloring, dictates a coloring of M in a natural way. Since both $M_1 + M_2'$ and $M_2 + M_1'$ each have fewer regions than M we are justified in the word reducible.

We might have a similar situation in which we can show that a certain ring R of n regions is such that any possible set of schemes deduced for $M_2 + M_1'$, where M_1' has fewer regions than $M_1 + R$, always contains at least one coloring for R suitable for $M_1 + R$. In such a case the ring R will be called reducible with respect to M_1.

Since the partial map $M_1 + R$ is regular, only three colors appear at a vertex and two regions are either adjacent or are separated by a region or regions. Thus two regions x and y of R in one or both of any pair of colors are either joined by a line in these colors or else separated by a chain of regions in the complementary pair of colors. That is, a pair of regions u and v of R, such that x,u,y,v occur in the cyclical order on R, are joined by a line in their colors or are separated by a line in the complementary pair of colors.

Theorem 6.5. If a map contains a ring of four or five regions about more than one region, it is reducible.

A ring of four regions about a single region is a reducible configuration since that region would be a quadrilateral.

Let us consider a ring of four regions A_1, A_2, A_3, A_4, about more than one region in a map not subject to any of the previously proved reductions. Let R be the ring of regions A_1, A_2, A_3, and A_4. Consider the maps $M_1 + M_2'$ and $M_2 + M_1'$ where M_2' and M_1' are formed from $M_2 + R$ and $M_1 + R$ by shrinking M_2 and M_1, respectively, to a point and by joining A_1 and A_3. The possibilities for the colors are essentially either a,b,a,b or a,b,a,c for A_1, A_2, A_3, A_4 respectively. If either scheme appears for both, we have a coloring. Otherwise we have a,b,a,b for $M_1 + R$, and a,b,a,c for $M_2 + R$, say. Now form a second choice for M_1' by shrinking M_1 to a point and joining A_2 to A_4, giving a,b,a,b or a,b,c,b for $M_2 + R$.

The only case necessary for us to consider is the second of these two since the first is listed for $M_1 + R$. Thus we have the scheme a,b,a,b for $M_1 + R$, and the schemes a,b,a,c and a,b,c,b for $M_2 + R$.

In the scheme a,b,a,b for $M_1 + R$ either a chain of regions in colors a,d connects the regions α_1, α_3 in which case we obtain a,b,a,c by interchanging c and b in the Kempe Chain containing α_4; or else a chain of colors b,c connects the regions colored b, and we get a,b,d,b by interchanging a and d in the Kempe Chain containing α_3. In either case we get a scheme found in the set of schemes for $M_2 + R$. Hence, the ring of four regions is reducible.

The proof for a ring of five regions uses partial maps M_1' and M_2' consisting of R with a pentagon replacing M_1 and M_2 respectively. The scheme c,a,b,a,b is then shown to be in both sets (that is, the odd-colored region c at the same place on the ring). The odd-colored regions are first shown to be adjacent, by showing that otherwise the scheme a,b,c,d,b is in both sets. Using transpositions on colorations obtained from different partial maps M_1'' and M_2'', which are formed by shrinking M_1 and M_2 respectively to a point and joining the two b-colored regions of R, the scheme in one set is rotated two regions on R. By three repititions, four such rotations are accomplished bringing the odd-colored regions to the same place.

Theorem 6.6. An edge surrounded by four pentagons is reducible.

We may assume M is not subject to any previously proved reductions. Let B_1, B_2, B_3, B_4 be the pentagonal regions, surrounded by the ring of regions α_1, α_2, α_3, α_4, α_5, α_6 as indicated in Figure 6.1. The cyclical order of R is assured by trivalence, and reduction 6.5 assures no two nonconsecutive regions of R are adjacent.

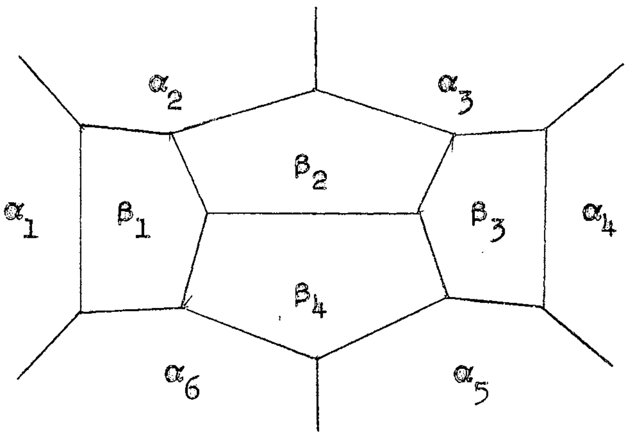


Figure 6.1.

Let B_1, B_2, B_3, B_4, α_3, α_5 of M_1 + R coalesce to form M_1'. With this choice of M_1' we obtain a coloring b,*,a,*,a,* of R in a scheme for M_2 + R. The only essentially different cases are;

b,c,a,b,a,c	b,c,a,c,a,c	b,c,a,d,a,c
b,c,a,b,a,d	b,c,a,c,a,d	b,c,a,d,a,d

Direct trial shows it is possible to color M_1 + R starting with any one of these except b,c,a,c,a,c which must be considered.

A c-d line in M_2 + R must connect all of the regions colored c in b,c,a,c,a,c for if not we could transpose c and d in some Kempe Chain and get a scheme previously acceptable. Hence any interchange of the colors a,b in R is permissible; one of these gives b,c,b,c,a,c which is a suitable arrangement for M_1 + R.

Thus, in every case, we are led to a scheme for M_2 + R having a coloring on R suitable for M_1 + R, and the theorem is true.

Theorem 6.7. A region completely surrounded by pentagons is a reducible configuration.

Let M be a map not subject to any of the previously proved reductions. Let M_1 consist of the given region V and the ring of pentagons around it. As before the set of regions R surrounding M_1 is a ring by trivalence and Theorem 6.4. Let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be the n regions of R in cyclical order, and let $B_1, B_2, \ldots, B_n$ be the pentagonal regions of M_1 that abut on $(\alpha_1,\alpha_2)(\alpha_2,\alpha_3), \ldots, (\alpha_n, \alpha_1)$ respectively. To form M_1', let alternate regions $\alpha_1, \alpha_3, \ldots, \alpha_{n-1}$ coalesce with a single region replacing M_1 if n is even, and alternate regions $\alpha_1, \alpha_3, \ldots, \alpha_{n-2}$ if n is odd. We thus obtain in a scheme for M_2 + R a coloring a,*,a,*,...,a,b or a,*,a,...,a,b,c.

In the first case if the *-colors are all one color b we can color $B_1, \ldots, B_n$ in c and d and then V in a. If this is not so, by taking account of the circular symmetry we can assume that we have a,c,a,*, . . .,a,b, i.e., we may assume α_2 and α_n are not the same color. Now color B_1 which abuts on α_1, α_2 in b, B_2 in d, B_3 in a color different from that of α_3, B_2, α_4 and so on, until we reach B_n, which may be colored in a different color than that of α_n, B_{n-1} and α_1. But α_n and B_1 are both colored b, so that this is acceptable. Having colored $B_1, \ldots, B_n$ in the three colors different from a, we may color V in a.

In the second case we can employ a similar process as follows: First color B_n in b. If then α_2 is also colored in b, we can without using color a color $B_{n-1}, \ldots, B_1$ and use a for V as in the first case. Otherwise α_2 is in c or d, and B_1 may be colored in d or c, and B_2 in b. Here again, if α_4 is colored in b, we may color $B_{n-1}, \ldots, B_3$ as in the first case. Otherwise α_4 is in c or d, so B_3 may be colored in d or c, and B_4 in b. Repeat until b is encountered as a color for α_i, where i is even; this occurs with α_{n-1} if not before.

In this way we color $B_1, \ldots, B_n$ in b,c,d and V in a. Thus we have appropriate schemes.

Theorem 6.8. An even-edged region completely surrounded by hexagons is reducible.

Let M_1 consist of the set of 2n hexagons about the single region V. As before we are assured that the regions surrounding M_1 form a ring. Let R be that ring of 4n regions. Let $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{4n}$ be the regions of R, and $B_1, B_2, \ldots, B_{2n}$ the hexagons, so that $\alpha_1, \ldots, \alpha_{4n}$ abut on B_1, (B_1,B_2), B_2, (B_2,B_3), $\ldots$, B_{2n}, (B_{2n},B_1), respectively. To form M_1' let $\alpha_2, \alpha_4, \ldots, \alpha_{4n}$ coalesce with a single region replacing M_1 and obtain a scheme *,a,*,a,*,...,a, for R + M_2. The regions B_i each touch two regions α_{2i} and one *-region. If the *-regions are all in one color b, we may color the regions B_i in c and d, alternately, and the n-sided region in a or b. If this is not the case, we may assume that we have c,a,...,b,a for $\alpha_1, \alpha_2, \ldots, \alpha_{4n}$ and proceed to color B_1(in b), $B_2, \ldots, B_{2n}$, successively, and then V just as we did in the preceding reduction.

The simplest application is to a hexagon surrounded by six hexagons.

Theorem 6.9. Every regular map of more than two regions containing no triangles, quadrilaterals, or two-edged regions contains at least twelve pentagons.

According to Euler's formula V - E + F = 2, but since exactly three regions are at a vertex with no two-edged regions, triangles, or quadrilaterals, $2E = 3V = \sum_{n \geq 5} n A_n$ where A_n is the number of regions in the map with n edges. From this we get $2E = 6(F - 2) = \Sigma n A_n$. But since $F = \Sigma A_n$, we have $6(\Sigma A_n - 2) = \sum_{n \geq 5} n A_n$ which may be written: $A_5 = 12 + \sum_{n \geq 7} (n - 6) A_n$ and since $\sum_{n \geq 7} (n - 6) A_n \geq 0$, A_5 must be at least 12.

Corollary. Every irreducible map contains at least 12 more pentagons than major polygons.

The next theorem does not identify a reducible configuration. However, it does tell us something of the nature of the minimum irreducible map. The credit for it goes to P. Wernichke and Philip Franklin.

Theorem 6.10. A minimum irreducible map must contain a pentagon adjacent to two other minor polygons.

Suppose this were not so. Let us count the number of vertices in the map which belong to hexagons and pentagons. We find that the number of vertices contributed by hexagons nowhere in contact with pentagons will be greater than twice the number of such hexagons since each hexagon has six vertices and no vertex belongs to more than three hexagons. In a cluster of hexagons, the two vertices contributed by an exterior hexagon may be chosen so that only one is not in common with another hexagon. Pentagons isolated from hexagons or other pentagons will give five vertices each. Two pentagons adjacent to each other but to no other pentagons or hexagons give eight vertices together, and hence average four each. A pentagon adjacent to a hexagon gives at least four vertices while allowing the hexagon to account for two. A pentagon adjacent to a cluster of hexagons can be adjacent to only one of the cluster; thus it contributes four vertices and allows the hexagon to contribute two, one of which it may have in common with the pentagon. Thus the number of vertices would be at least $4A_5 + 2A_6$, where A_i is the number of regions with i edges, and we would have to have: $V \geq 4A_5 + 2A_6$. From the preceding proof and the obvious inequality $\sum_{n\geq 7} (7 - n) A_n \leq 0$ we get $\sum_{n\geq 7} A_n + 12 \leq A_5$ so $\sum_{n\geq 5} A_n + 12 \leq 2A_5 + A_6$. From Euler's formula, for a regular map, $\sum_{n\geq 5} A_n = F = \frac{V}{2} + 2$ we have $\frac{V}{2} + 14 \leq 2A_5 + A_6$, or $V + 28 \leq 4A_5 + 2A_6$ which contradicts that $4A_5 + 2A_6 \leq V$, and proves the theorem.

Theorem 6.11. An edge of a hexagon surrounded by this hexagon and three pentagons is a reducible configuration.

As before, the surrounding regions form a ring of seven regions. If we erase the lines that are dotted in Figure 6.2, we would obtain a new map which would contain fewer regions than an irreducible map and hence be colorable.

From the way we selected the lines which were erased, regions 1 and 4 would have the same color, say a; while regions 5 and 7 would have a different common color, say b. Of the five essentially distinct colorations of M (allowing for symmetry of the configuration), the three cases shown in (a), (b), and (c) permit immediate coloration, as indicated. In the case shown in (d), if 5 is joined to 7 by a b-d chain in $M_2 + R$, 6 may be changed to c reducing the problem to case (a), while if 5 is joined to 3 by a b-d chain, 4 may be changed to c, and the map colored as shown in (e). If neither of these chains exists, 5 may be changed to d, and the map colored as shown in (f). Finally in the

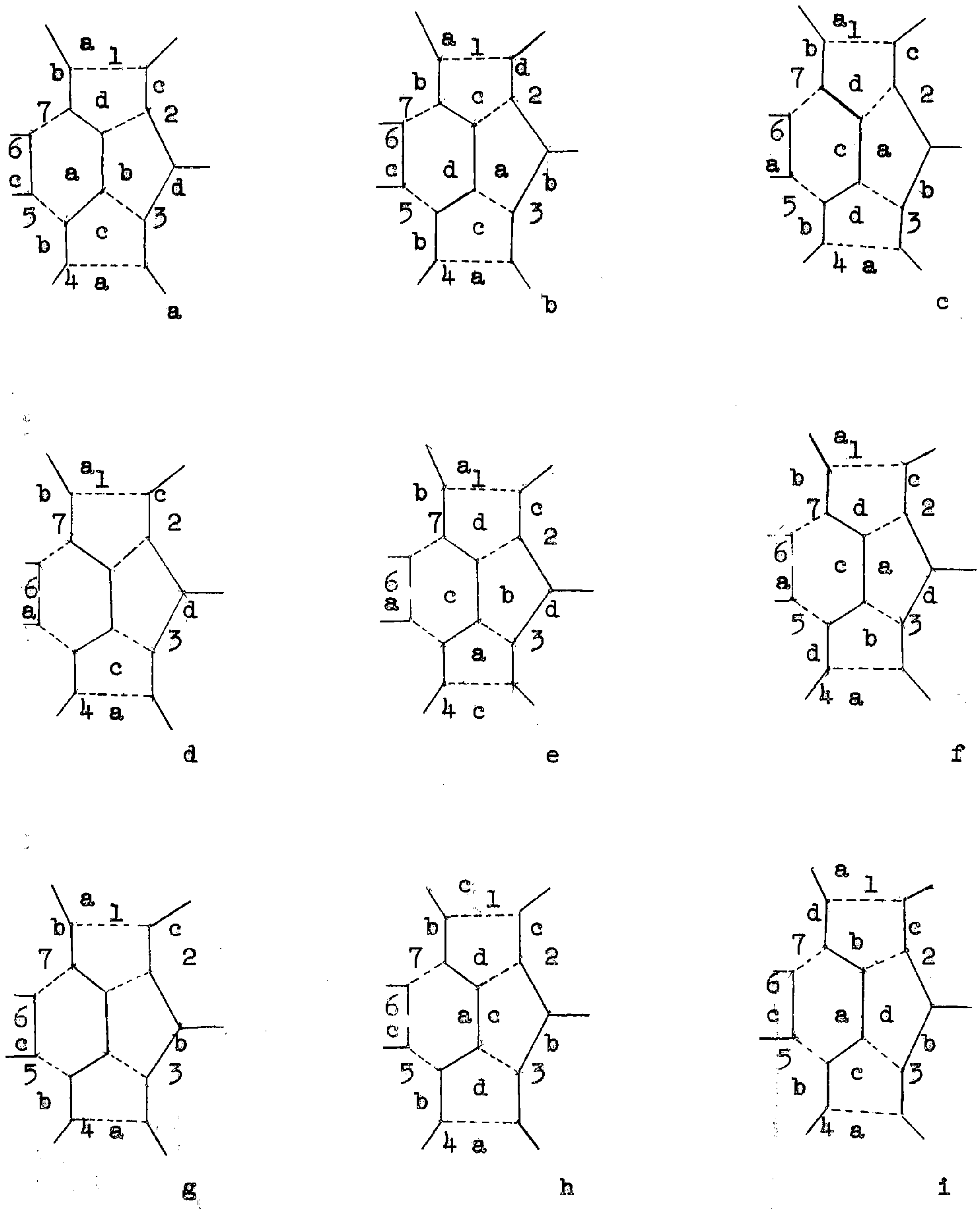


Figure 6.2

case given in (g), either a b-d chain joins 7 with 5, and we reduce to (c) by changing 6 to a; or a chain joins 7 with 3, and we color as in (h) after interchanging a and c in the a-c chain including 1 and 2; or 7 may be changed to d and we color as in (i).

Theorem 6.12 follows from 6.6 and 6.11.

Theorem 6.13 through 6.40 (except 18) use essentially the approaches which have been illustrated. Our next consideration will be a theorem on the minimum number of regions in an irreducible map.

Putting together reductions 6, 11, 16, 17, and 20 of our list we have:

A) A polygon of 5, 6, 7, or $n > 7$ edges is reducible when in contact with 3, 3, 4, or $n-1$ adjacent pentagons.

For our next theorem we will have need of some new notation.

A shall denote a region of any map, M.

A_n shall denote a region of n edges.

a_5 shall denote the number of pentagons, A_5, in M.

j_{5n} shall denote the number of contacts of the pentagons of M with polygons A_n where $n \geq 6$. Making use of this notation we have the following relation:

$$10\, a_5 \geq j_{56} + 2 \sum_{n \geq 7} j_{5n} \geq 4a_5.$$

The left inequality is obvious. The right follows since the contribution of an A_5 to the center member when it touches one or no other pentagons is not less than four; when it touches exactly two other pentagons, it touches three other polygons at least one of which is major, thereby contributing at least four; when it touches three pentagons, by 28, the other contacts cannot be minor polygons, so that it contributes four.

$A_5(t)$ shall denote an A_5 contributing $4 + t$ to the center expression of the relation just discussed. Let $a_5(t)$ be the number of such pentagons and we have:

$$j_{56} + 2 \sum_{n\geq 7} j_{5n} = \sum_{t=0}^{6}(4 + t)a_5(t) = 4a_5 + \sum_{t=1}^{6} ta_5(t).$$

Further, for $n > 5$ let $A_n(r)$ be an A_n rouching r pentagons and $a_n(r)$ their number. It follows from A) that $j_{5n} = \sum_{r=1}^{n-2} ra_n(r)$. Combining these two relations we have: $\sum_{r=1}^{4} ra_6(r) + 2 \sum_{n\geq 7} \sum_{r=1}^{n-2} ra_n(r) = 4a_5 + \sum_{t=1}^{6} ta_5(t)$ and noting that $a_n \geq \sum_{r=1}^{n-2} a_n(r)$ we get

$$2\sum_{n\geq 6}(3n-17)a_n + a_6(3) + 2a_6(4) \geq 4a_5 + \sum_{t=1}^{6} ta_5(t) + 2\sum_{n\geq 7} \sum_{r=1}^{n-2}(3n-r-17)a_n(r).$$

Consequently if we show

B) $a_6(3) + 2a_6(4) \leq \sum_{t=1}^{6} ta_5(t) + 2 \sum_{n\geq 7} \sum_{r=1}^{n-2} (3n-r-17)a_n(r)$

it will follow that $\sum_{n\geq 6}(3n-17)a_n \geq 2a_5$. This can be reduced to $a_5 + \sum_{n\geq 6} a_n \geq 3a_5 - 3 \sum_{n\geq 7} (n-6)a_n$ and using Euler's relation we get $3a_5 - 3\sum_{n\geq 7}(n-6)a_n = -3\sum_{n\geq 5}(n-6)a_n = 3(6 \Sigma a_n - \Sigma na_n) = 3(6F - 2E)$. But for regular maps $3V = 3E$ and $6F - 2E = 6F - 6E + 6V = 12$. Hence $a_5 + \sum_{n\geq 6} a_n \geq 36$. Then establishing B) proves Theorem 43.

Theorem 6.43. Every irreducible map must have at least 36 regions.

The remainder of this chapter will be devoted to the tedious chore of establishing B), above. It is here one can note the tremendous effort that has gone into obtaining these results.

Before doing this we shall alter the statement by removing the only negative term, given by $n = 7$, $r = 5$, from the double sum on the right and transposing it to the left member so as to give the following statement, where $n \neq 7$ when $r = 5$:

$$a_6(3) + 2a_6(4) + 2a_7(5) \leq \sum_{t=1}^{6} ta_5(t) + 2\sum_{n\geq 7} \sum_{r=1}^{n-2} (3n-r-17)\, a_n(r).$$

To establish B), we shall set against $A_6(3)$, $A_6(4)$ respectively one or two polygons $A_5(t)$, $A_n(r)$ adjacent to them as compensating elements which contribute to the right-hand side; and against $A_7(5)$ one or two elements $A_5(t)$. It will then be necessary to verify that the number of sources of a given element, after considering twice the source $A_7(5)$ yielding a single element, is at most equal to the corresponding coefficient on the right of B).

A hexagon in contact with the chain 5565 or 55665 is reducible by 25 and 29. From 29 we conclude that a hexagon of M that makes non-adjacent contacts with 3 pentagons must touch at least two major polygons. Thus $A_6(3)$ is bounded by either 5n5N5N, 55N5Nn, or 5N5nN5, where now and hereafter $N \geq 7$ and $n \geq 6$. In the former two cases we take as our element a the last pentagon adjacent to $A_6(3)$, which is bounded by 6NxxN.

In the last case let bcde be the last four polygons about $A_6(3)$, and let f be the outside polygon touching de. Then, if c is a major polygon, we choose the element b(6NxxN), (i.e., b bounded by 6NxxN). If c = 6 and f = 5, we choose e, which, on account of A) or 11, is bounded by 65N5N. If c = 6 and f is a major polygon, we choose e(65xNN). Finally, if c = f = 6, our element is d(66...65); if d is an $A_7(r)$, we infer by 26 that $r \leq 3$, so that neither $A_7(4)$ nor $A_7(5)$ is chosen as an element.

The contacts of $A_6(4)$ are 55N55N in view of A) and 25. Moreover, we conclude from 30 that one of these pairs of pentagons g,g' are not in triad with a third pentagon nor, by A), with another hexagon when g,g' are in chain with a third pentagon. We here select two elements, namely g,g' both (65NxN) or both (656nN).

On account of A) and 33 the ring around $A_7(5)$ is 555N55N. If the fourth (or last) polygon is also an $A_7(5)$, we take the two pentagons h, h' touching both $A_7(5)$'s. These are both $A_5(2)$'s, their contacts being 75N57 by A). But, if there is no adjacent $A_7(5)$, we consider the center pentagon of the first three; by virtue of 28 it touches an outside major polygon; thus another of the first three, call it i, is also adjacent to this major polygon. We have then a single element bounded by 75NxN', where N' is not an $A_7(5)$.

In each of the above rings about an element we have placed first its source. A polygon with such contacts might occur as an element as often as one of its adjacent polygons fits into the first place of the ring (allowing for a reversal). We have thus to examine the possible occurrences of $A_5(t)$, where $1 \leq t \leq 5$, and of $A_n(r)$, where $1 \leq r \leq n - 2$ or 3, according as n is greater than or equal to 7.

The incidence of $A_5(1)$ is at a, b(6N55N); e, g, g' (65N5N); e(655NN); g, g' (6566N).

There is no repetition here since the two adjacent hexagons touching g or g' cannot come first in any of the four rings.

The incidence of $A_5(2)$ is at a, b(6N56N); e, g, g' (656NN); g, g' (65N6N); h, h' (75N57); i(75N5N').

We observe that this element occurs at most twice in the first three rings, which contain only two hexagons. These rings are also distinct from the last two. Now, by supposition, the last polygon in the fourth ring is an $A_7(5)$, whereas the last in the fifth is not.

Hence an element $A_5(2)$ can only appear twice in the fourth ring and once in the last, but not in both. This yields altogether a maximum of 2 occurrences, counting that at i twice because only one element i was selected to account for a coefficient 2 on the left side of B).

The incidence of $A_5(3)$ is at a, b(6N5NN) or (6N66N); e, g, g'(65NNN); i(75N6N').

This element occurs only once in the last ring since the third polygon being next to a hexagon is not an $A_7(5)$ with this as its element. It can then fall but once elsewhere, namely in the first ring. Thus the maximum amounts to 3 (twice for i), since none of the first three rings contain more than 3 hexagons.

The incidence of $A_5(4)$ is at a, b(6NN6N); i(75NNN').

The N between N and N' not being an $A_7(5)$ with this as its element, we infer that this element can only fall twice in the last ring, and not then in the first. Hence, as the first ring contains but two hexagons, the maximum here is four.

Lastly, an $A_5(5)$ is only to be found once, at a, b(6NNNN), while $A_5(6)$ does not occur at all. So altogether the number of pentagons compensating $A_6(3)$, $A_6(4)$ and $A_7(5)$ is not in excess of the first sum on the right of B).

The number of occurrences of $A_n(r)$ at d(66...65), as an element for an $A_6(3)$, cannot exceed the number of hexagons touching d, which is at most $n - r$. But the coefficient of $a_n(r)$ in B) is $2(3n - r - 17)$, which is at least $n - r$ unless $n = 7$, $r = 2$ or 3. Moreover, in the last two cases the largest number of hexagons coming third in the sequence 6566 (or second in 6656) is found by inspection to be 4 or 2 respectively, i.e., not more than the coefficient of $a_7(r)$. This concludes the demonstration of B) and of Theorem 6.43.

CHAPTER VII

COLORING GRAPHS

It is the modern trend to consider the nodes of a graph rather than the regions of a map. Since we can limit the types of maps under consideration to those that are regular, have no two-region rings, and have no region of less than five edges, we are assured that these maps have unique duals. As a matter of fact all polyhedral maps have unique duals (see Chapter Four). Since there is a one-to-one correspondence between the regions of a map and the nodes of its dual and since the nodes corresponding to contiguous regions are joined by an edge, we have the equivalent problem of coloring the nodes of a graph in a surface such that no two nodes joined by an edge are the same color. Of course the graphs considered will have no node joined to itself by an edge; two nodes are joined by at most one edge (since a map containing a ring of two regions is not under consideration); two edges meet in a node or not at all.

The term colored graph (coloring of graphs) is somewhat ambiguous since it is also used to describe the situation in which the 1-cells of a first degree factor of a third degree graph are colored one color (usually red) and those of the second degree factor blue. Note there are two blue edges and one red edge at each vertex. However, the ambiguity is removed by the context. In this chapter we shall consider both types of coloring, but our first almost trivial theorem deals with simple graphs.

A simple graph is a connected regular graph of third degree without leaves. It is easy to see that a connected third degree graph is simple when and only when every 1-cell is on a closed circuit, a circuit being a finite path in which the initial node coincides with the terminal node and all other nodes are distinct. The order of such a graph is necessarily even as seen in Chapter Four. Note also that in a third degree graph a loop with a vertex is a leaf.

Theorem 7.1. Given a simple graph of order greater than two, we can always obtain from it a simple graph of lower order by removing any 1-cell and properly joining the four incident 1-cells in pairs.

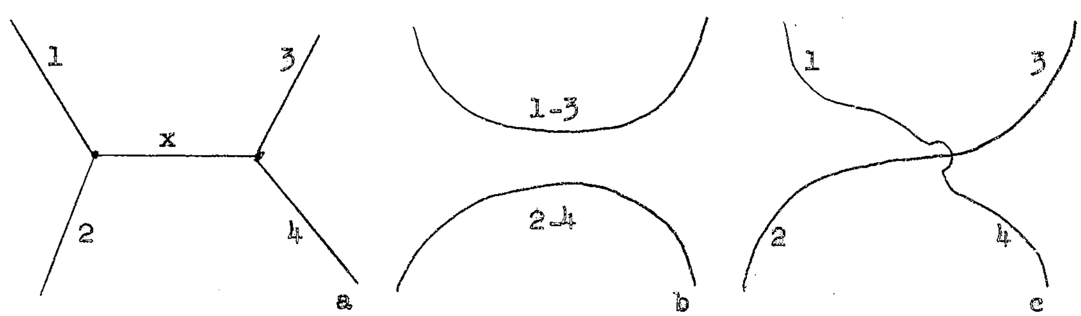


Figure 7.1.

One of two alterations will give us the desired graph. The first shown in Figure 7.1b, is obtained by deleting x and joining 1 to 3 and 2 to 4. The second, shown in Figure 7.1c, is obtained by deleting x and joining 1 to 4 and 2 to 3. Note it is not necessary to consider the graph on a surface.

The results of the first alteration will be a simple graph unless a bridge results or the graph becomes disconnected. If either of these be the case, there must be in the original graph a path from 1 to 3 and a path from 2 to 4, besides x; furthermore, these two paths do not intersect. For example, if there were no path from 1 to 3 besides x then either 1 or 3 would be the stem of a leaf, or a bridge. If the first alteration did not produce a simple graph, the second one will. To prove this it is sufficient to show that the closed circuits through 1-cells, 1, 2, 3 and 4 have been (in the second alteration) replaced by new ones containing all their arcs.

The circuits through 1 and 3 have been lengthened (except for the obvious combination of arcs), the 1-cell x having been replaced by the path we proved to exist from 2 to 4. Note, this path is contained entirely in that part of the graph containing 2 and 4 that is connected by at most one 1-cell to the part containing 1 and 3. Circuits through 2 and 4 are replaced in a similar manner. As for circuits through 1 and 2, if the first alteration disconnected the graph these must contain 3 and 4 and hence the second alteration would restore them as circuits. If the first alteration resulted in a bridge, however, a circuit through 1 and 2 need not contain 3 and 4 if it contained the bridge. In this case, the second alteration produces two new circuits through the bridge which together will contain all the arcs of the original circuit through 1 and 2. The same argument applies in the case of circuits through 3 and 4. The circuits through 1 and 4 and through 2 and 3 have not been disturbed by the second alteration.

If there exists in a colored graph a closed path consisting of an even number of 1-cells colored alternately red and blue (a red-blue path) we can interchange colors along this path and obtain a new coloration and hence a new factorization of the graph.

Theorem 7.2. Every 1-cell of a colorable simple graph is on a closed red-blue path and hence can have its color changed.

For the simple graph of order two, Theorem 7.1 (which we will use) would not apply, but the conclusion is then obvious. Suppose the theorem were false. Then, there would exist a colored simple graph of lowest order containing a 1-cell not on a closed red-blue path. If this 1-cell is red then none of the four blue 1-cells adjacent to it is on a red-blue path. Consider such a blue 1-cell, call it x, not on any closed red-blue path. Take a blue 1-cell adjacent to x, call it y, and delete the red 1-cell incident with the other end of y joining loose ends as allowed by Theorem 7.1 to obtain a new simple graph of lower order. In the new graph which is also colored, there must exist a closed red-blue path through x, which cannot contain y. But such a path exists in the original graph, because its red-blue character is not destroyed by restoring the deleted red 1-cell. It is possible that a closed red-blue path containing y might be disturbed by this restoration, but x and y, being adjacent blue 1-cells, cannot be on the same red-blue path. Thus, we obtain a contradiction and the theorem follows.

Theorem 7.3. Every simple graph is colorable.

Here again, the theorem is obvious for a graph of order two. If the theorem were false there would exist a non-colorable simple graph of lowest order. Lower its order as allowed by Theorem 7.1 and color it. Three cases arise. Case I. If both new 1-cells are colored blue, restore the deleted 1-cell colored red. Case II. If one is colored blue and the other red, restore the deleted 1-cell colored blue.

Case III. If both new 1-cells are colored red, change the color of one of them as allowed by Theorem 7.2. If this changes the color of the other also, we have Case I; if not, Case II. In any case we have a coloration of the original graph.

The following proof of Heawood's formula is due to G. A. Dirac, [24].

Suppose G is a graph drawn on a surface of connectivity h which divides the surface into polygonal regions (simply connected regions of 3 or more sides). Let N denote the number of nodes, E the number of edges and F the number of polygonal regions into which the surface is divided, then Euler's Theorem states that $N + F - E = 3 - h$.

If there are regions on the surface which are bounded by more than three edges, it is possible to add new edges until a graph is obtained which divides the surface into polygons bounded by three edges. The number of nodes of the new graph is still N. Let the number of edges be E' and the number of triangular faces F', then $E' \geq E$ and $F' \geq F$. Now every triangle is bounded by three edges and every edge separates two triangles, hence $3F' = 2E'$. By Euler's Theorem, $3N + 3F' - 3E' = 9 - 3h$; hence $3N - E' = 9 - 3h$ and so $2E'/N = 6(1 + (h - 3)/N)$. Since $E' \geq E$ for the original graph $2E/N \leq 6(1 + (h - 3)/N)$. Let us call this relation A) so we can easily refer to it.

A connected graph drawn on the surface which does not divide it into polygonal regions can be drawn on a surface with smaller connectivity, or can be made to divide the surface into polygonal regions by the addition of edges. Thus, A) holds for all connected graphs drawn on a surface of connectivity h.

Let k be the chromatic number of a graph drawn on a surface of connectivity h. If this graph is infinite, De Bruijn showed that it contains a finite k-chromatic subgraph. If this subgraph is not critical, it has a node a, which is not critical, and the deletion of this node yields a k-chromatic graph which if not critical has a non-critical node b. By successively deleting non-critical nodes in this way, after a finite number of steps we obtain a critical k-chromatic subgraph. Let this subgraph have N nodes and E edges. Then clearly $N \geq k$. The degree of each node is at least $k - 1$, so that $2E/N \geq k - 1$. Since A) holds, if $h \geq 3$, $k - 1 \leq 6(1+(h-3)/k)$. It follows that: if $h \geq 3$, every graph drawn on a surface of connectivity h can be colored using at most N_h colors, where N_n is the greatest integer satisfying $N_h - 1 \leq 6(1+(h-3)/N_h)$. If $h = 2$ we have from A) that $k - 1 < 6$, that is, $k \leq 6$. The value of N_h explicitly is $[\frac{1}{2}(7 + \sqrt{24h - 23})]$; where [ ] stands for the greatest integer in the enclosed real number. When $h = 2$, this expression gives the correct value 6. Thus we have a proof of the theorem by Heawood:

Theorem 7.4. The chromatic number of a map on a surface of connectivity $h \leq 2$ is at most N_h, where $N_h = [\frac{1}{2}(7 + \sqrt{24h - 23})]$.

If N denotes the number of nodes and E the number of edges of a critical k-chromatic graph, the degree of every node is at least $k - 1$, and $2E/N \geq k - 1$. Brooks established (for $k \leq 4$) that a k-chromatic graph, either contains a complete k-graph as a subgraph or contains a node of degree k. Hence, if $N > k$ then it does not contain a complete k-graph; so it contains a node of degree $\geq k$ and $2E/N > k - 1$. Let us call this relation B). Using the same notation we shall establish our next result in the form of a lemma.

Lemma I. If $k \geq 5$ and $N = k + 2$ for a critical graph, then $2E/N >$ $k + 1 - 12/k+2$.

From a result established by Dirac, see [24], it follows that a critical k-chromatic graph of order $k + 2$ contains a complete $(k - 1)$-graph as a subgraph. Let $k \geq 5$ and let the nodes of such a graph be denoted by $a_1, a_2, \ldots, a_{k-1}, b_1, b_2, b_3$, where every pair of $a_1, a_2, \ldots, a_{k-1}$ is joined by an edge. Because the graph is critical, b_1, b_2, and b_3 are each joined to at least $k - 1$ nodes. The number of edges in the graph, consistent with this requirement, would be least (i.e., the most economical distribution of edges is obtained) if each of b_1, b_2, and b_3 is joined to the other two and to $k - 3$ of the nodes $a_1, a_2, \ldots, a_{k-1}$. In this case the graph would contain $\frac{1}{2}(k - 1)(k - 2) + 3(k - 3) + 3$ edges; with any other distribution of edges it contains more. Hence, for such a graph, $2E/N \geq k + 1 - 12/(k+2)$. But with the distribution described above, unless b_1, b_2, and b_3 are all joined to the same $k - 3$ nodes from among $a_1, a_2, \ldots, a_{k-1}$, the graph can be colored with $k - 1$ colors. If b_1, b_2, and b_3 are all joined to the same $k - 3$ nodes from among $a_1, a_2, \ldots, a_{k-1}$ then these nodes together with b_1, b_2, and b_3 form a set of k nodes of which each pair is joined by an edge, so that the graph contains a complete k-graph as a subgraph and is not critical. This most economical distribution is therefore not permissible, and so, for a critical k-chromatic graph of order $k + 2$, $2E/N > k + 1 - 12/k+2$.

Lemma II. If $0 \leq n \leq k - 1$, a k-chromatic graph either contains a complete $(k - n)$-graph as a subgraph or it has at least $k + n + 2$ nodes.

If we let $n = 0$, we get the special statement, a k-chromatic graph which does not contain a complete k-graph as a subgraph contains at least $k + 2$ nodes. For proof of Lemma II, see [23].

Lemma III. In the notation of B), if $k \geq 5$ and $N = k + 3$ for a critical graph, then $2E/N > k + 2 - 24/(k+3)$.

The proof of this lemma is similar to the proof of Lemma I. These proofs were presented by Dirac see [24]. The following theorem is also from the same article.

Theorem 7.5. For $h = 3$ and $h \geq 5$, a graph of chromatic number N_h on a surface of connectivity h, when it exists, always contains N_h mutually adjacent nodes.

Existance will be considered in the next chapter.

To prove Theorem 7.5 it is to be shown that for $h = 3$ and $h \geq 5$ the only critical N_h-chromatic graph which can be drawn on a surface of connectivity h is the complete N_n-graph. To do this we shall prove it for $h = 3$ and for $h \geq 5$ first prove that no critical N_h-chromatic graph of order $N_h + 4$ can be drawn on a surface of connectivity h. Then it will be proved that no critical N_h-chromatic graph of order, $N_h + 2$ or $N + 3$ can be drawn on such a surface. Theorem 7.5 will then follow from the special case of Lemma II.

Case $h = 3$. In this case $N_h = 7$. By B) for $N > k$ with $k = 7$, $2E/N > 6$ and by A) with $h = 3$, $2E/N \leq 6$. This is a contradiction. So $N = k$ and we have completed the proof of Theorem 7.5 for the case $h = 3$ and excluded $N_h = 7$ from further consideration.

Suppose a critical N_h-chromatic graph of order $\geq N_h + 4$ is drawn on a surface of connectivity h. If E denotes the number of edges and

N the number of nodes then, by B), $2E/N > N_h - 1$; and by A), since $h \geq 5$, $2E/N \leq 6\ (1+(h-3)/(N_h+4))$. Hence, $N_h - 1 < 6\ (1+(h-3)/(N_h+4))$ while N_h satisfies the inequalities: $N_h - 1 \leq 6\ (1+(h-3)/N_h)$, $N_h > 6\ (1+(h-3)/(N_h+1))$. From these we get the following two relations. $N_h^2 - 3N_h \leq 6h + 9$ and $N_n^{\ 2} - 5N_h \geq 6h - 11$; hence, $N_h \leq 10$. It remains to examine those cases where $N_h \leq 10$ and $h \geq 5$.

Case $N_h = 8$. If $N_h = 8$ then $h = 5$. By B) with $k = 8$, $2E/N > 7$. By A) with $h = 5$ and $N \geq N_h + 4 = 12$, $2E/N \leq 6\ (1+2/12) = 7$. This is a contradiction.

Case $N_h = 9$. If $N_h = 9$ then $h = 6$ or $h = 7$. Consider first the case $h = 7$. By B) with $k = 9$, $2E/N > 8$. By A) with $h = 7$ and $N \geq N_h + 4 = 13$, $2E/N \leq 6\ (1 + 4/13) = 7\ 11/13$. This is a contradiction. In a similar fashion the case $h = 6$ would lead to a contradiction.

Case $N_h = 10$. If $N_h = 10$ then $h = 8$ or $h = 9$ or $h = 10$. Consider first the case $h = 10$. By B) with $k = 10$, $2E/N > 9$. By A) with $h = 10$ and $N \geq N_h + 4 = 14$, $2E/N \leq 6\ (1 + 7/14) = 9$. This is a contradiction and cases 8 and 9 also lead to a contradiction.

These contradictions prove that for $h \geq 5$ no critical N_h-chromatic graph of order $\geq N_h + 4$ can be drawn on a surface of connectivity h. It remains to be seen whether an N_h-chromatic graph of order $N_h + 2$ or $N_h + 3$ can be drawn. These cases will be considered in turn.

Graphs of order $N_h + 2$. Suppose a critical N_h-chromatic graph of order $N_h + 2$ is drawn on a surface of connectivity $h \geq 5$. By Lemma I for such a graph $2E/N > N_h + 1 - 12/(N_h+2)$. By A), $2E/N \leq 6\ (1 + (h-3)/N_h + 2)$ hence $N_h + 1 - 12/N_h+2 < 6\ (1+(h-3)/(N_h+2))$ that is, $N_n^{\ 2} - 3N_h \leq 6h+3$. But from the definition of N_h for $h \geq 5$, $N_h > 6\ (1+(h-3)/(N_h+1))$, that is $N_h^{\ 2} - 5N_h \geq 6h - 11$, and so $2N_h \leq 14$, or $N_h \leq 7$. But we have already disposed of the case $N_h = 7$.

Graph of order $N_h + 3$. Suppose a critical N_h-chromatic graph of order $N_h + 3$ is drawn on a surface of connectivity h. By Lemma III, for such a graph, $2E/N > N_h + 2 - 24/N_h+3$. By A), $2E/N \leq 6\ (1+(h-3)/(N_h+3))$; hence $N_h+2-24/N_h+3 < 6\ (1+(h-3)/(N_h+3))$, that is, $N_h^2 - N_h \leq 6h + 17$. But by the definition of N_h, $N_h > 6\ (1+(h-3)/(N_h+1))$, that is, $N_h^2 - 5N_h \geq 6h - 11$, and so $4N_h \leq 28$, or $N_h \leq 7$. But we have already disposed of this one. This completes the proof of Theorem 7.5.

By this theorem one can see that the problem of determining the least upper bound of the chromatic numbers of maps on surfaces is directly related to the problem of determining the maximum number of mutually adjacent regions. Certainly, the maximum number of mutually adjacent regions, A_h could not exceed N_h for the chromatic number of any map must be at least as large as the largest number of mutually adjacent regions in the map. However, if $N_h > A_h$ for any surface, where $h = 3$ or $h \geq 5$, this theorem shows that Heawood's number is not the least upper bound. There has been one example of this given. Franklin showed, [33] that any map on a Kline bottle can be colored in six colors whereas $N_3 = 7$.

Theorem 7.5 was supplemented by Dirac [24] with two weaker theorems, one for surfaces of connectivity 2 and one for surfaces of connectivity 4. For their statement see Chapter Three. Surfaces of connectivity 2 and 4 are exceptions to the proof of Theorem 6.5 but are not necessarily exceptions to the theorem itself. Dirac felt that the theorem held even for 2 and 4.

CHAPTER VIII

GENERALIZATIONS

It would seem, since the chromatic number of the sphere has remained undetermined, that the problem of determining the chromatic number for surfaces of higher genus would be difficult indeed. However, this is not the case.

The following table shows the values of N_h corresponding to the values of h (up to 16) and χ(the Euler-Poincare characteristic).

χ	2	1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13
h	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
N_h	4	6	7	7	8	9	9	10	10	10	11	11	12	12	12	13

Heawood's color formula (Theorem 7.4) places an upper bound on the chromatic numbers of maps on a surface of connectivity h. When Heawood advanced the theorem, he was thinking only of orientable surfaces. More recent authors have shown that his theorem applies equally well to non-orientable surfaces.

Heawood's color formula also applies to surfaces that are not closed such as the Mobius strip. The Mobius strip being homeomorphic to a crosscap will have connectivity 2 and therefore any map on its surface can be colored in 6 colors. An example demanding 6 colors is easy to find. This is of course assuming that coloring a region on a surface is analogous to dying a portion of a cloth. That is, the color penetrates and both sides in the local sense are the same color. If on the contrary, you painted the region on a surface (that is to say the color does not penetrate the surface) then the questions involving chromatic numbers seem quite different, at least in the case of non-orientable surfaces.

There is a noteworthy construction by deRham [51] associated with the question of orientability. He shows that for each non-orientable surface there is an orientable surface that can be mapped onto that non-orientable surface in a two-to-one fashion. This means that our pending question on painting a non-orientable surface is closely related to the question of dying the regions. That is, with the establishment of the proper orientable manifold for each non-orientable one, we have avoided the question of whether we are dying or painting.

To give this more meaning, consider the Mobius strip where two regions back to back along an edge could be considered contiguous along that edge. Considering the map as a graph, we have merely added the segments of the edge of the strip to the graph as edges of the graph. If we imagine a sort of inflation to separate the two surfaces lying back to back we would in effect obtain a torus.

Heawood [39] himself established by example that for an orientable surface where $h = 3$ his formula actually gave the best value. That is, N_3 colors are necessary and sufficient to color every map on the surface of a torus, see Figure 8.1. In 1891, Heffter [46] showed for orientable surfaces where $h = 5, 7, 9, 11, 13$, and 15, that the number determined by Heawood's formula was best. For $h = 5$ see Figure 8.2. H. Tietze [17] showed that the projective plane (non-orientable with $h = 2$) needed the full number of colors. I. N. Kagne [48] gave

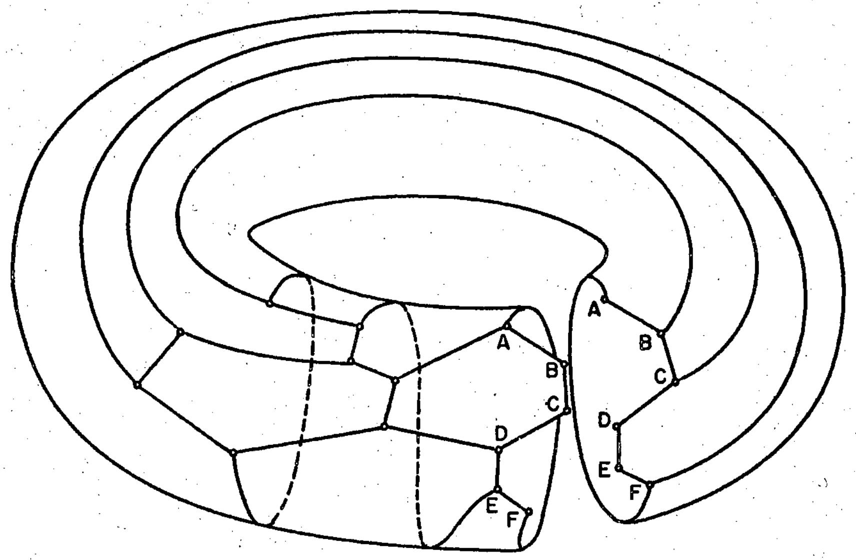


Figure 8.1

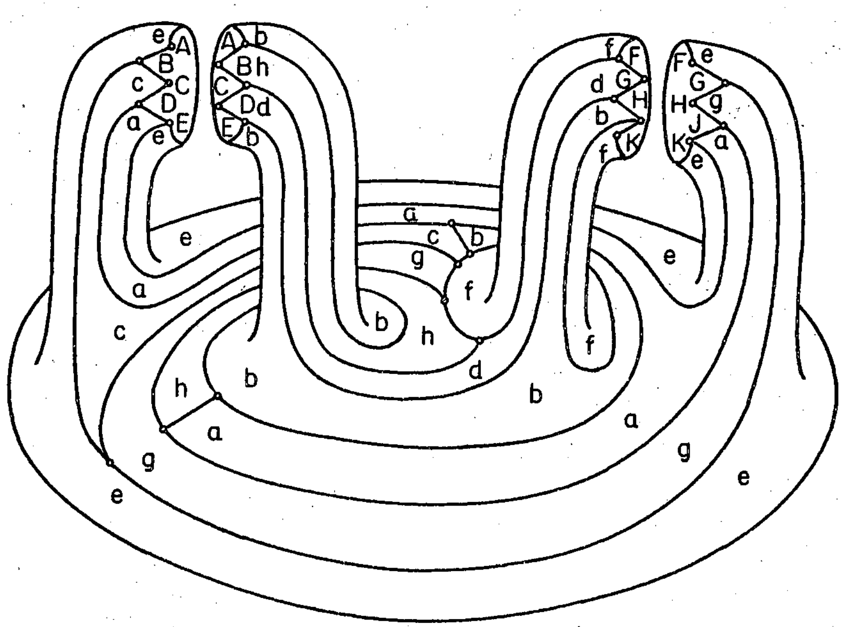


Figure 8.2

examples of maps on non-orientable surfaces where h = 4, 5 and 7 that needed the full number of colors. Coxeter [17] filled the gap by showing a non-orientable surface where h = 6 needed the full number of colors. R. C. Bose [17, 9] extended it to h = 8 in 1939. Cases where h = 10, 12, and 14 follow [24] since the connectivities of the surfaces obtained from the sphere and from the projective plane by attaching n handles are 2n + 1 and 2n + 2 respectively. Any map drawn on a sphere with n handles attached can also be drawn on a projective plane with n handles attached. It follows that $K_{2n+2} \geq K_{2n+1}$ where K_h is the maximum chromatic number of a map on a surface of connectivity h. Hence, by Heffter's results, quoted above, and by the table of values of h and N_h, Heawood's result is best for h = 10, 12, and 14.

It should be noted that although the proof of Heawood's formula does not hold for h = 1 (the sphere) the formula still yields the conjectured value 4. This is thought-provoking in itself without considering all the forgoing cases where Heawood's formula gave the best result. However, amidst all this evidence there is the result, given by Franklin in an excellent article published in 1934, which established that six colors were necessary and sufficient to color any map on a non-orientable surface of connectivity 3, whereas $N_3 = 7$. Hence, if the four-color theorem is false, it is not a unique exceptional case.

In 1959 G. Ringel proved that Klein bottle (h = 3) is the only non-orientable surface not needing as many as N_h colors. He also showed that the chromatic number of a map on an orientable surface could not differ from N_h by more than 2.

Theorem 8.1. Seven colors are always sufficient and sometimes necessary to color any map on a torus

The proof follows from Figure 8.1 since $N_3 = 7$.

A simple map was given by John Leach in 1953. It is shown in Figure 8.3 drawn first on a rectangle and then as it would look if the rectangle were made into a torus by first joining the top and bottom edges to form a cylinder and then bending the cylinder around to form a torus.

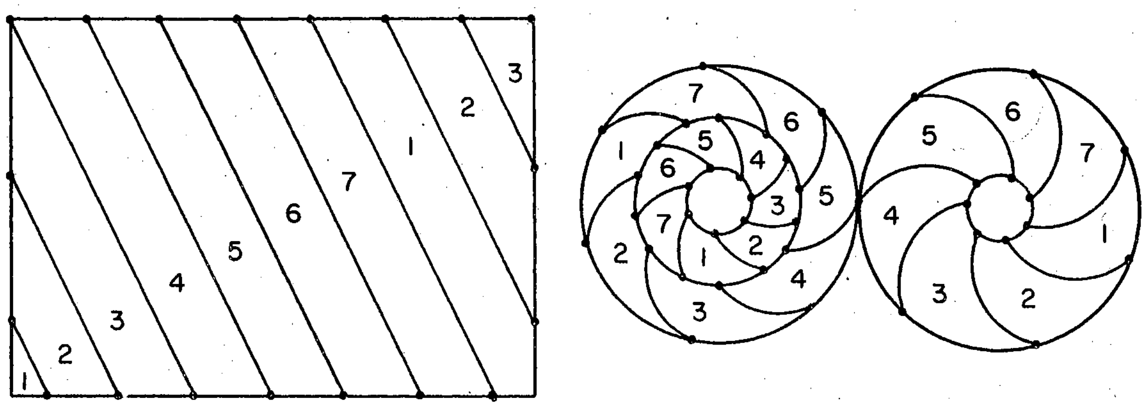


Figure 8.3

Theorem 8.2. Eight colors are always sufficient and sometimes necessary to color any map on a sphere with two handles of the first kind.

The proof follows from Figure 8.2 since $N_5 = 8$.

Up to now our attention has been focused on surfaces that existed in three space, whereas the Klein bottle and the projective plane do not. These surfaces do exist in four space as well as N space where $N \geq 4$. The question naturally arises whether these results hold for surfaces even though they are located in a space of dimension higher than that necessary for their existence. That is, does the embedding of these surfaces in a space of higher dimension add properties to them that would disturb the results of our theorems. Although no rigorous answer will be given here, it seems that since the property of "being a neighbor of" is an intrinsic one that the results would be undisturbed by the surfaces' extending into another dimension, just as they are already two-dimensional extending into the third dimension.

Theorem 8.3. Six colors are necessary and sufficient to color any map on the projective plane.

We could argue the sufficiency from an extension of the five-color theorem for surfaces of connectivity two, that is, a region of five sides is reducible in six colors and every map has a region of less than six sides. However, Heawood's formula gives us the upper bound of 6 and it is attained by the map in Figure 8.4. The map is obtained by

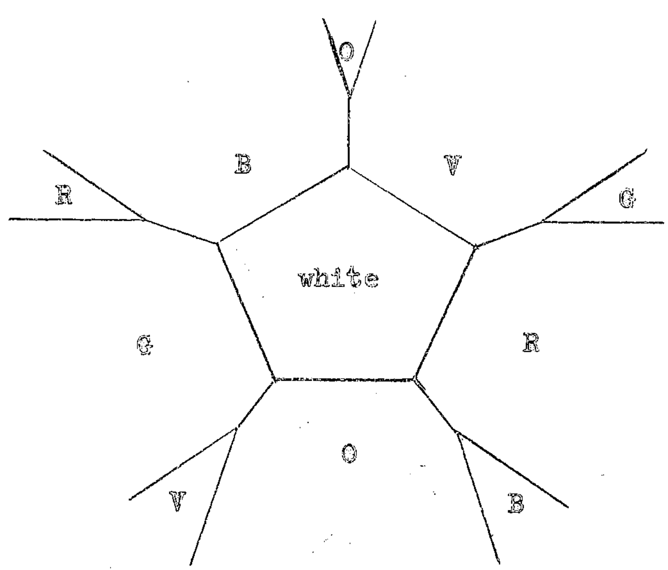


Figure 8.4

projecting the dodechedron lying on the projective plane from its center. This produces six pentagonal regions such that each region touches every other region, necessitating six colors to properly color the map.

Theorem 8.4. Six is the least number of colors needed to color all maps on a non-orientable surface of connectivity three.

Suppose there exists a map on the Klein bottle that cannot be colored in 6 colors; then there exists one with fewest regions. If this map contains multiply-connected regions then make it simply-connected by altering one of the boundaries of each multiply-connected region as shown in Figure 8.5. The altered map has the same number of regions, with corresponding pairs of contiguous regions, thus it is colorable if and only if the original map is colorable; call the simply-connected minimum irreducible map M.

Now M contains no region of less than six edges since it cannot be reducible in six colors. Now $3V \leq 2E = aF$, where V is the number of vertices of M, E the number of edges, F the number of regions and a the average number of edges per region; and since $V - E + F = \chi$ we have $E \leq 3(E - V) = 3(F - \chi)$. Thus $a = \frac{2E}{F} \leq 6(1 - \frac{\chi}{F})$ and when $\chi = 0$, $a \leq 6$. It follows that all regions of M must be hexagonal, with $a = 6$. Now since $V - E + F = 0$, $aV - aF = 0$ but $aF = 2E$ and $a = 6$ so $6V - 6E + 2E = 0$. Consequently, $3V = 2E$ and it follows that each vertex is trivalent (M is regular). Hence, M is a network of hexagons with three hexagons at each vertex. Let one of the hexagons be labeled 7, surrounded by hexagons labeled 1, 2, 3, 4, 5, 6 as shown in Figure 8.6. The regions 1, 2, 3, 4, 5, 6 must be distinct since a contraction of 7 to a point would show M to be reducible.

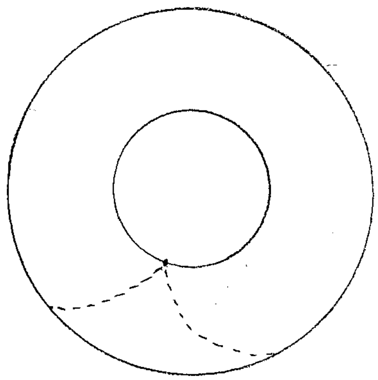


Figure 8.5

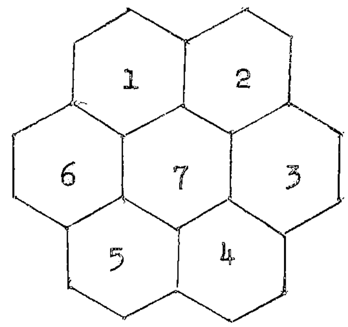


Figure 8.6

Again, region 1 must have a common edge with 3. For if they do not the configuration is reducible in six colors by a coalition of regions 1, 7, and 3. By a similar argument, the regions 1, 2, 3, 4, 5, 6 are mutually contiguous.

We shall now show that such a map M is possible only on an orientable surface. This will mean that a minimum irreducible map (in six colors) cannot exist on the Klein bottle and that each map on the Klein bottle is colorable in 6 colors.

If, in Figure 8.7, the center hexagon is 7 and those adjacent to it are 1, 2, 3, 4, 5, 6, the outer hexagons, which are the last six repeated, must each be one of the numbers placed in the top of the respective hexagons. Since hexagon Y must be 2, 3, or 4 (because it touches 6 which touches 1, 7, 5, and also 3, 4, or 5 (because it touches 1), we see that Y must be 3 or 4. If Y is 3, then Z must be 4 or 5. Suppose Z is 4; then U must be 5; and V must be 6 (not 4 since Z also contacts U); but then W cannot be 5 or 6, as it should be. It follows

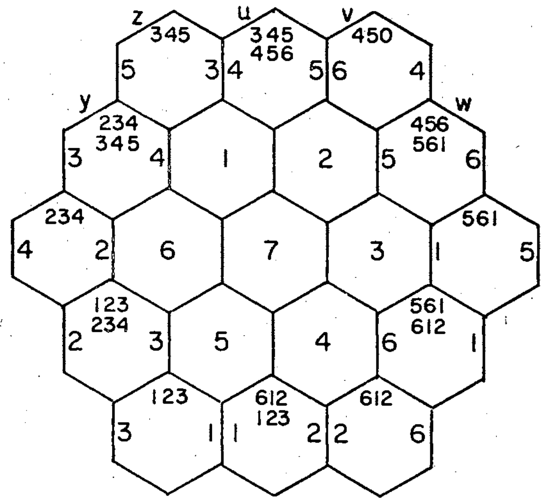


Figure 8.7

that Z cannot be 4 and must therefore be 5 when Y is 3. But with Z 5, we get a unique representation given by the numbers on the left edges of the outer hexagons. Similarly, if Y = 4, one finds that we get the unique representation given by the numbers on the right edges of the outer hexagons. We have reached the conclusion that if M exists on the Klein bottle then it must be one of the two represented by Figure 8.8.

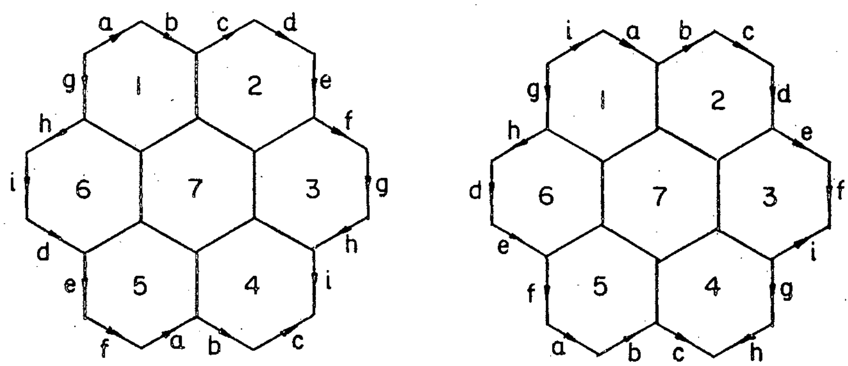


Figure 8.8

But it cannot be these for each of these forms a torus, instead of a Klein bottle. Consequently M does not exist on the Klein bottle and six colors are sufficient to color any map on the Klein bottle.

We shall show by example that six may be necessary. Consider the map in Figure 8.9 to be on the projective plane.

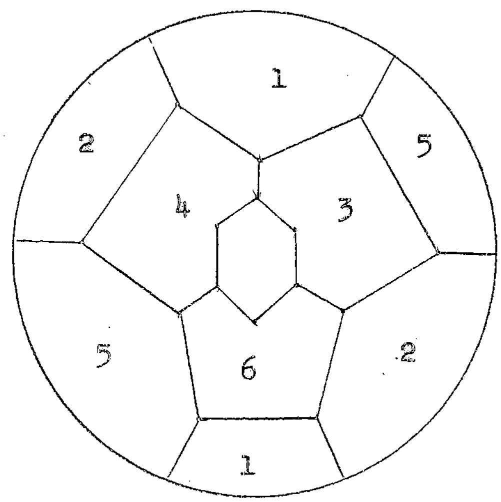


Figure 8.9

This map can be colored in no less than 6 colors. But this map can be drawn on the Klein bottle. To see this, insert a cross cap in the hexagonal region in the center of Figure 8.9. This changes the projective plane into the Klein bottle without destroying the pairs of contiguous regions. Thus 6 colors are necessary and the theorem follows.

Theorem 8.5. The maximum chromatic number of a map on a non-orientable surface of connectivity 4 is 7.

Here again, by Heawood's color formula, the 7-point graph in

Figure 8.10a, proves the theorem. The free edges of the figure are identified as indicated. Figure 8.10b is the same graph using the illustrative technique of Coxeter, by constructing the dual of this graph we obtain a map of seven mutually contiguous hexagons, Figure 8.10c. The dual map given by Kagno is shown in Figure 7.10d, the free edges being identified as indicated. X and X' are joined through a crosscap, similarly for Y and Y' and for Z and Z'. The three cross-caps raise the connectivity to 4.

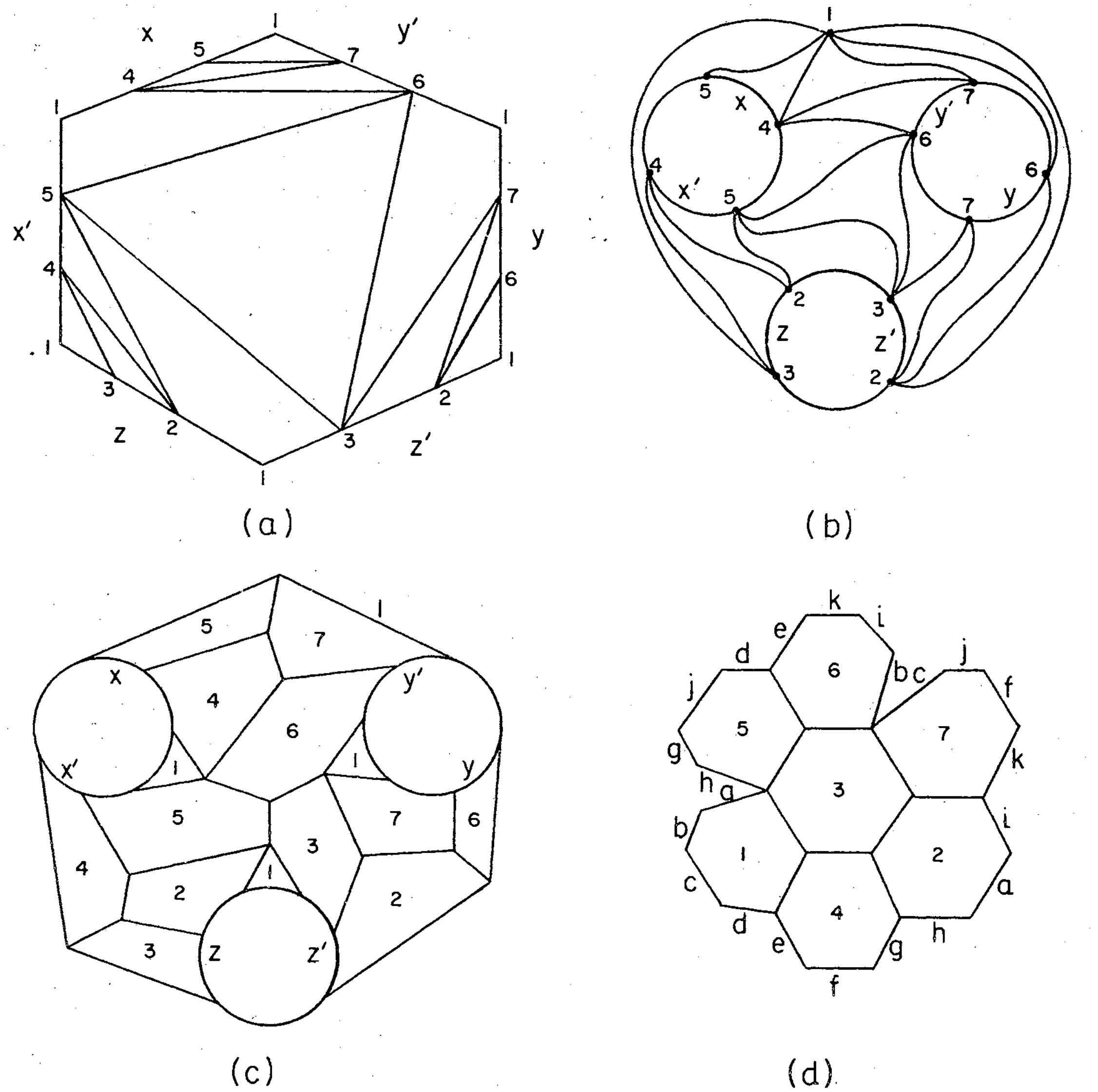


Figure 8.10

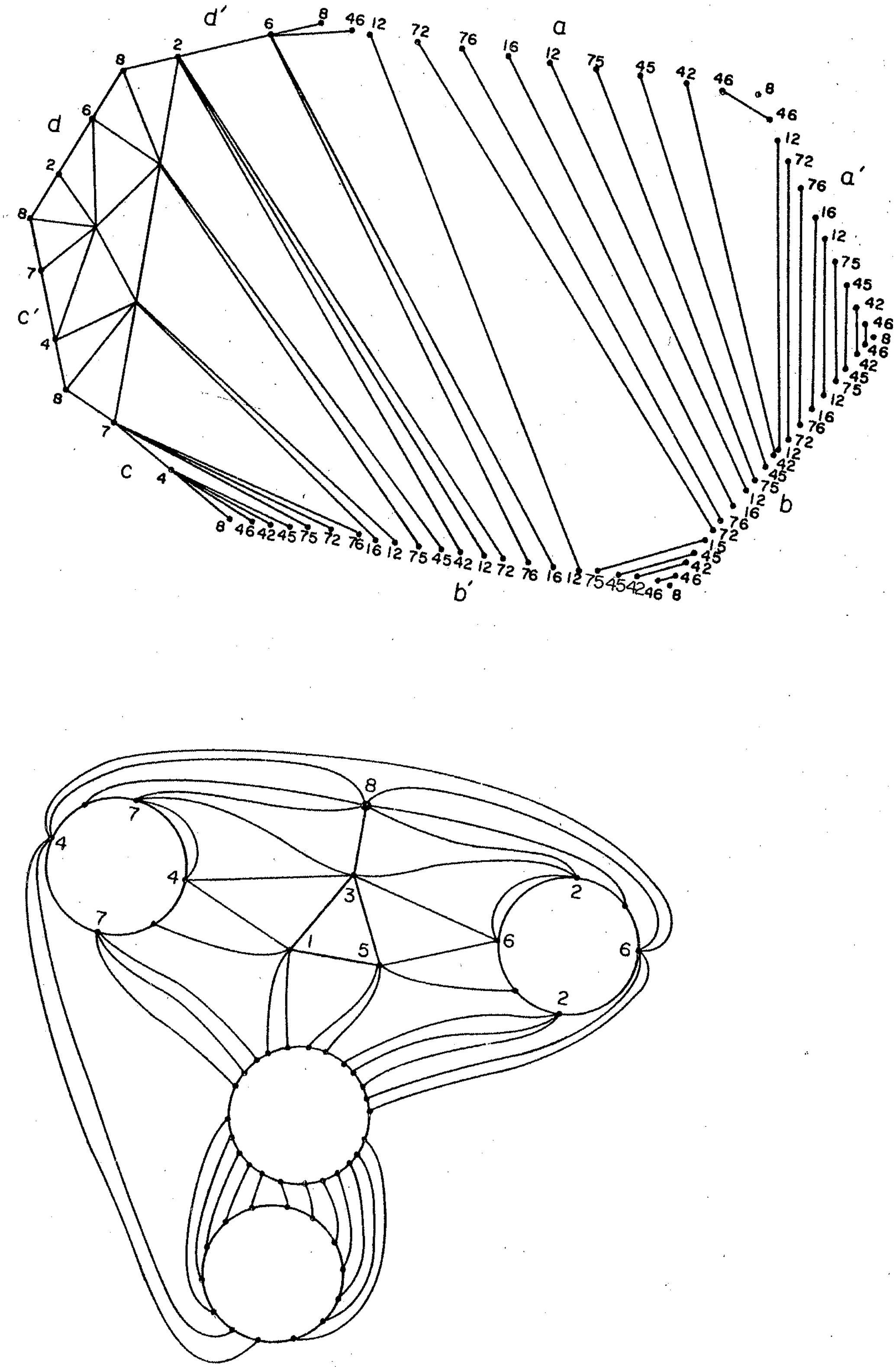


Figure 8.11

Theorem 8.6. The maximum chromatic number of a map on a surface of connectivity 5 is 8.

The graph of a complete eight point on such a surface is shown in Figure 8.11. The free edges, identified as indicated, have the order (a,a',b,b',c,c',d,d') from which we see that the surface is non-orientable and of connectivity 5. By constructing the dual of this graph we obtain a map of eight mutually touching heptagons. This map is shown in Figure 8.12, the free edges being identified as indicated. Theorem 8.2 provides for the orientable case.

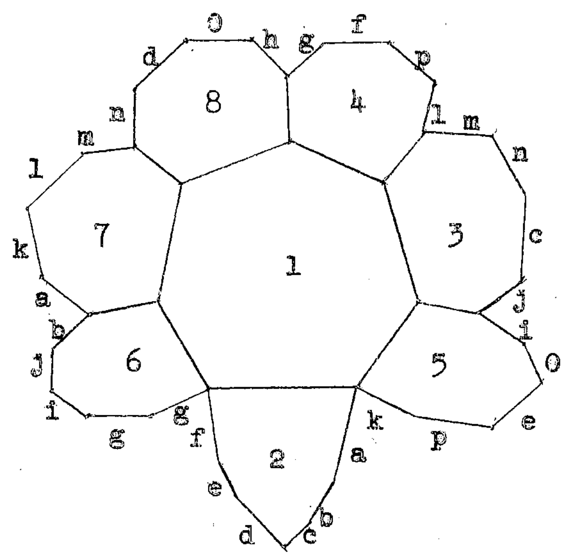


Figure 8.12

Theorem 8.7. The maximum chromatic number for maps on a surface of connectivity 6 is 9.

A complete 9 point graph and a map of 9 mutually touching octagons on a surface of connectivity 6 are given in Figure 8.13.

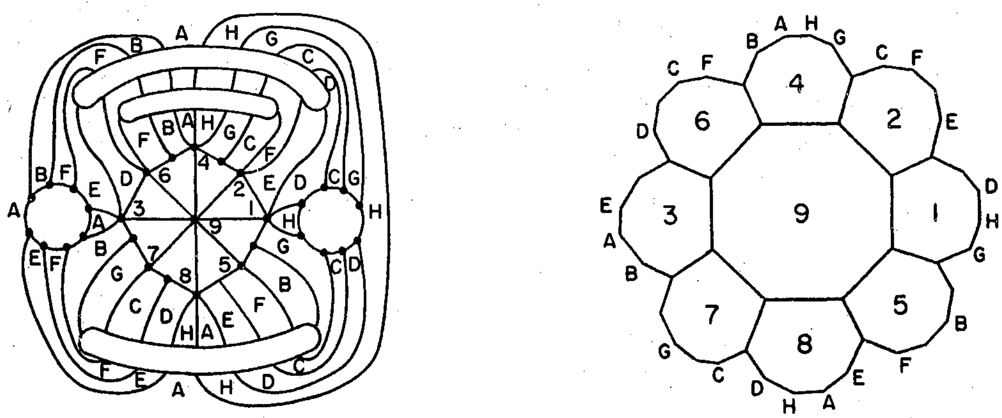


Figure 8.13

Theorem 8.8. The maximum chromatic number of a map on a non-orientable surface of connectivity 7 is 9.

This theorem follows from Theorem 7.7 since any map on a surface of connectivity 6 can be drawn on a surface of connectivity 7. This can be seen by placing a crosscap in the interior of regions 9 of Figure 8.13b, raising the connectivity of the surface to 7. Kagno; however, proved this theorem by giving the example in Figure 8.14, several years before Coxeter proved Theorem 8.7.

Theorem 8.9. The maximum chromatic number of maps drawn on a surface of connectivity 8 is 10.

This map is given in Figure 8.15.

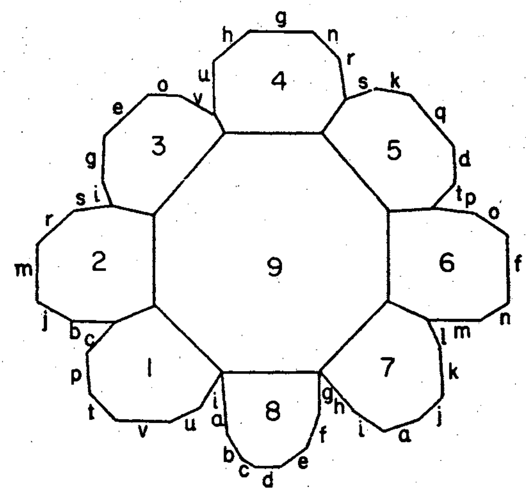


Figure 8.14

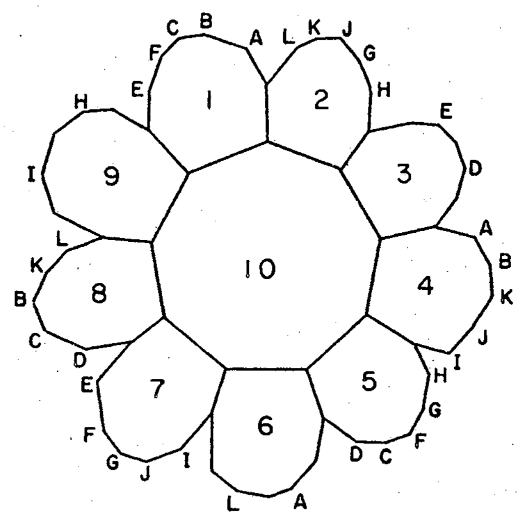


Figure 8.15

Theorem 8.10. The maximum chromatic number of maps drawn on a non-orientable surface of connectivity 9 is 10.

Any map that can be drawn on a surface of non-orientable surface of connectivity 8 can be drawn on a non-orientable surface of connectivity 9. Thus the map used to establish the necessity of 10 colors in Theorem 8.9 accomplishes the same purpose here.

From remarks made earlier in the chapter and Heffter's examples we have:

Theorem 8.11. The maximum chromatic number for maps on non-orientable surfaces of connectivity 10, 12, and 14 are 10, 11, and 12 respectively.

Similarly, any map drawn on a surface of connectivity 13 can be drawn on a non-orientable surface of connectivity 14 or 15 by the addition of 1 or 2 cross caps interior to any region of the map. With the establishment of Heffter's result we have the following:

The maximum chromatic number of maps on a non-orientable surface of connectivity 14 or 15 is 12.

There is a gap in these theorems presented thus far. For a non-orientable surface of connectivity eleven no map requiring eleven colors is known to have been presented. However, Ringel proved that the Heawood number is correct for all non-orientable surfaces, except the Klien bottle, thus, filling the gap and solving the remaining problems of a similar nature.

Theorem 8.12. The maximum chromatic number of maps on a non-orientable surface of connectivity $h > 3$ is N_h.

The question of whether the Heawood number is correct for all orientable surfaces of connectivity $h > 15$ has not yet been determined. However, Ringel proved the following.

Theorem 8.13. If K_h is the maximum chromatic number of maps on an orientable surface of connectivity h then $K_h \geq N_h - 2$.

The problem may be extended in another direction. In actual maps the countries sometimes consist of several detached portions forming an empire, with the further stipulation that a mother country and its colonies be similarly colored. If there is no limit to the number of colonies in each empire, the number of colors needed may be arbitrarily large. If no empire consists of more than r separate pieces, not only is the problem tangible but in 1891 Heawood gave a formula for an upper bound on the number of colors necessary to color such maps.

In the case of regular maps Euler's formula says $V - E + F = \chi$ and $3V = 2E$ so that $\frac{2E}{F} = 6(1 - \frac{\chi}{F})$. If we consider a regular map of n regions $(n = F)$ we have the average number of contacts per region as $A_n = \frac{2E}{F} = 6(1 - \frac{\chi}{n})$. If a map is not trivalent the value of A_n is lower since a trivalent map can be obtained by adding new edges and vertices at each of the original vertices with multiplicity greater than three, and this map has all the contacts of the old map with new contacts added; consequently, the average for the new trivalent map is higher. If the map contains a multiply-connected region, form a new map by altering the boundaries on each such region indicated in Figure 8.5. The new map is simply-connected and the average will again be raised by making it trivalent. It follows that if A_n' is the maximum of A_n over all maps of n regions then the maps to which A_n' corresponds must be regular.

Consider first maps of empires on a sphere and further suppose that each empire consists of exactly two distinct portions. Then the average number of contacts per empire in a given map of $\frac{n}{2}$ empires is $C = 2A_n \leq 2A'_n$. If we let $X = [2A'_n] + 1$ then X colors are sufficient to color a map of $\frac{n}{2}$ empires. For, some empire of that map does not have more than X - 1 contacts and hence is a reducible configuration in X colors. That is, if that empire were removed and the surrounding regions closed up in any fashion we would have a map of $n' < n$ regions and since $2A_{n'} \leq 2A'_{n'} = 12(1 - \frac{2}{n'}) < 12(1 - \frac{2}{n}) = 2A'_n < X$ we again are assured of an empire of less than X contacts; and the process can be repeated eventually giving us a map of X or fewer empires, which can certainly be colored in X colors. This in turn dictates a coloring of our original map. Hence, $[12(1 - \frac{2}{n})] + 1$ colors are sufficient to color any map of $\frac{n}{2}$ two-region empires on the sphere.

Now if n is as large as 24 then this number is 12. If n is smaller, 12 colors are still sufficient so that 12 colors are sufficient to color any map of two-regions empires on the sphere. For a map with each empire consisting of at most two regions, 12 colors are still sufficient since for each single region empire, a new region can be added as a circular patch on an edge that is not an edge of that empire. The resulting map can be colored in 12 colors and the patches then removed revealing the original map properly colored in 12 colors.

At least as many colors are necessary as the largest number of mutually contiguous empires, T. In a map of k mutually contiguous empires, C is not less than k - 1 since each empire touches k - 1 others at least once. T could not be greater than Y, the greatest integer y such that in a map of precisely y two-region empires

$C \le 12(1 - \frac{2}{2y})$ so $y - 1 \le 12 - \frac{12}{y}$ and $y \le 12$. Therefore T could not be more than 12, which we hope can be attained. However an example must be found showing that 12 can be realized.

Since $12 - \frac{12}{y} = y - 1$ when $y = 12$ there are just enough contacts available for 12 mutually contiguous empires. With each of 12 empires in two regions the chance of duplicate contacts between some pair of empires is greatly increased, and this would necessitate that some other pair of empires would fail to touch. This difficulty seems to occur with any symmetrical arrangement of 24 regions, and no general principle seems to lead to the discovery of a suitable arrangement. However, one example is sufficient to complete the proof; and that given in Figure 8.16 is an instance of 12 pairs of regions, which have the desired contacts, once and only once with a region of every other pair. Figure 8.16 then proves that 12 is the number of colors

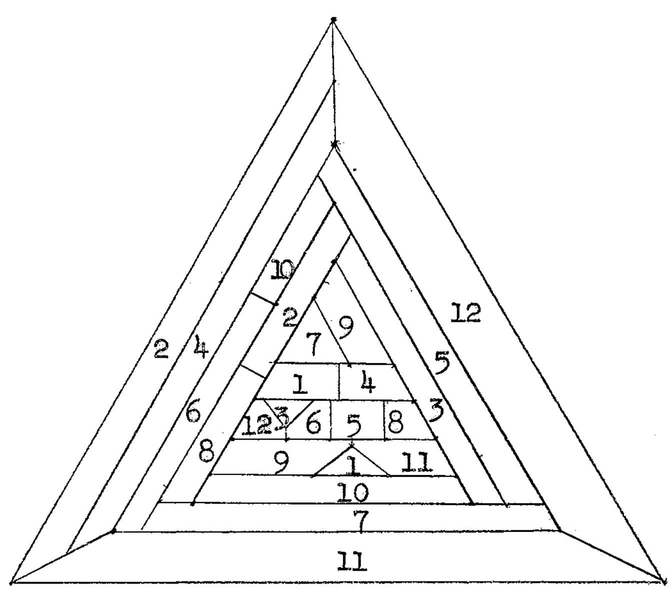


Figure 8.16

necessary and sufficient to color any arrangement of empires of at most two regions on the sphere. When we proceed to empires which may consist of 3 or more colonies, up to r say, we have $C \leq r A_n \leq 6r\ (1 - \frac{2}{n})$. If the map is regular, $rA_n = 6r\ (1 - \frac{2}{n})$, so it follows that $rA'_n = 6r\ (1 - \frac{2}{n})$ and the largest $X = [rA'_n] + 1$ could be is 6r. As before, 6r colors are sufficient. Also $rA'_{rY} = 6r\ (1 - \frac{2}{rY}) = \frac{6r}{Y}(Y - \frac{2}{r})$ which is greater than $Y - 1$ when $Y = 6r$, provided $r > 2$. In this case then there are more than enough contacts for 6r empires. The verification examples, to show that proper distribution of contacts could be realized, would then determine 6r as the number of colors necessary and sufficient to color maps of empires, consisting of no more than r regions, on the sphere. The only such map found in the literature is the one given in Figure 8.16. Although Heawood was not concerned with non-orientable surfaces, we can note that the map is also an example on the projective plane if we place a crosscap interior to some region of the map on the sphere. Later we shall see that 12 is also an upper bound on the projective plane.

Moreover, extending the proposition to the case of the torus, where $A_n \leq 6(1 - \frac{0}{n}) = 6$, we have $C \leq rA_n \leq rA'_n = 6r$, and the largest X could be is $6r + 1$. Also, $rA'_{rY} = 6r$, which is exactly $Y - 1$ when $Y = 6r + 1$; and here again $X = Y$. There are just enough contacts for $6r + 1$ empires and an example showing proper distribution of these contacts would complete the proof. The example for the case where $r = 1$ was given earlier in this chapter.

Consider next maps with empires of exactly r regions on a surface of characteristic $\chi < 0$. We have $A_n \leq A'_n = 6(1 - \frac{\chi}{n})$. Let Z be the greatest value of w such that $6r(1 - \frac{\chi}{rw}) \geq w - 1$; then for $w > Z$,

$6r(1 - \frac{\chi}{rw}) < w - 1$. Now if we let $X = [Z]$ then $6r(1 - \frac{\chi}{rZ}) = Z - 1 < X$. Any map of not more than X empires is colorable in X colors. A map of $m > X$ empires is also colorable in X colors since the average number of contacts per empire, $C_m \leq rA_{mr} \leq rA'_{mr} = 6r(1 - \frac{\chi}{mr}) \leq 6r(1 - \frac{\chi}{r(X+1)}) <$ $(X + 1) - 1 = X$.

Now since X is the greatest integer such that $6r(1 - \frac{\chi}{rX}) \geq X - 1$ we have enough contacts for X empires (but not for more). From $Z^2 - (6r + 1)Z = -6\chi$ we get $Z = \frac{1}{2}(6r + 1 + \sqrt{(6r + 1)^2 - 24\chi})$ so that $X = [\frac{1}{2}(6r + 1 + \sqrt{(6r + 1)^2 - 24\chi})]$.

For maps with empires of not more than r regions, we could show X colors sufficient by adding regions as before.

If $\chi = 0$ then $\frac{1}{2}(6r + 1 + \sqrt{(6r + 1)^2}) = 6r + 1 = X$ as we previously showed. The reader should show that for $r \geq 2$, $6r \leq \frac{1}{2}(6r + 1 + \sqrt{(6r + 1)^2 - 24(2)}) < 6r + 1$ to establish the more general expression for X in the case of the sphere. Heawood was concerned with orientable surfaces only, and our development thus far has been essentially his [39]. However, his result applies directly to non-orientable surfaces with one exception, the projective plane. This exception can be included with the verification that a similar argument applies for $\chi = 1$ and gives the value 6r. We see then that the examples for the sphere also work for the projective plane.

The more general expression for X then gives an upper bound for the number of colors necessary to color a map of empires each containing not more than r regions on a surface of characteristic χ, except where $\chi = 2$ and $r = 1$ which is of course the four-problem problem.

We have thus far considered only regions which are two-dimensional on a surface in a space of N dimensions where N is sufficiently large to allow the actual existence of the surface under consideration. According to the following, [63], five-space is sufficiently large for any surface thus far considered: Suppose X is a compact space and the dimension of X is ≤ n, n finite; then X is homeomorphic to a subset of the cross product of [0,1] with itself 2n + 1 times. Since all of the surfaces thus far considered are compact 2-manifolds they can be embedded in 5-space.

Let us now consider solid regions in three space. That is, what is the minimum number of colors necessary to color an arbitrary number of solids such that no two contiguous solids are of the same color. It can be seen that this number can be made arbitrarily large. For given any positive integer N we shall construct a situation in which N colors are needed. Consider 2N rectangular prisms of dimension 1 by 1 by N. Lay N of these side by side and number them 1 through N such that numbers i and i + 1 touch along a side of dimension 1 by N. This forms a constructed solid, 1 by N by N. Similarly lay the remaining N, numbered 1 through N, rectangular prisms on top of these with their edges of N length perpendicular to those of the first N. This is illustrated for N = 6 in Figure 8.17. This forms a constructed solid

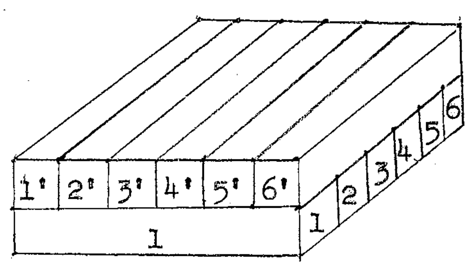


Figure 8.17

of dimension 2 by N by N. Now consider solid 1 and 1' to be the same (weld them together). Do the same for solid i and i', i ranging from 1 to N. Thus we see that solid 1 touches all the rest as does each of the others. As a result the minimum number of colors necessary to color all such configurations is infinite.

Paul Stacked, University of Heidelberg, gave the same result using only convex domains, [56 page 78] a particularly complicated and significant result.

We would certainly expect the problem of coloring N dimensional regions in N space to have the same solution as three in three-space. Also we could express the same confidence for M-dimensional regions $M \geq 3$ in N space since three-dimensional regions in N space yield the same result as they do in 3-space.

CHAPTER IX

EQUIVALENT PROBLEMS

In this chapter we shall consider some problems that are implied by the four-color proble, some that imply the four-color problem, and some that do both. We shall restrict our attention to the plane or sphere; however, some of the results generalize readily to other surfaces.

Theorem 9.1. A regular map can be properly colored in four colors if and only if its boundaries can be properly numbered with three numbers.

First let us consider the "only if" part -- if a regular map can be properly colored in four colors, then its boundaries can be properly numbered with three numbers: Given a regular map colored with four colors, call them (0, 0); (0, 1); (1, 0); (1, 1). Then assign to each boundary the sum of the labels of the two regions that it borders, (by the sum of (m, n) and (p, q) we mean ((m + p) mod 2, (n + q) mod 2). Now there are only three possible sums; (0, 1); (1, 0); (1, 1) since $a + b = (0, 0) \Longrightarrow a = b$.

Since exactly three regions are mutually contiguous at each vertex we get exactly three boundaries and the labels of these boundaries must be different. For, consider one of the regions to have color a, and the other two b and c. Now, a + b, a + c and b + c must all be different since $a + b = a + c \Longrightarrow b = c$. Thus, the two boundaries of the region of color a, that are at this vertex, must have different numbers.

This holds for any two boundaries with a common vertex. Thus, the numbering is proper.

Conversely, if the edges of a regular map can be properly numbered with three digits then the map can be properly colored in four colors: First let us show that any closed path on the surface of the map that intersects each of any number of edges exactly once in a point that is not a vertex, will do so such that the sum of the labels of these edges is (0, 0). It is sufficient to show this for a simple closed path. For, suppose a path can be broken into k loops. Then, the sum of the labels of the edges crossed in each loop is (0, 0) and k(0, 0) = (0, 0), thus the entire path yields (0, 0). Consider an arbitrary simple closed path U, that intersects several edges of the map each at a point that is not a vertex. Consider the vertices $V_1, V_2, V_3, V_4, \ldots, V_n$ that lie inside the simple closed curve. At each vertex we find the three edges (0, 1), (1, 0), (1, 1) and the sum of these three is (0, 0). Now, the edges may be placed in two classes, those that have both end-points interior to our simple closed path U and those with only one inside U. Let x be the sum of those in the first class and y the sum of those in the second. Now the sum of the labels of the edges that lie at each of these vertices is 2x + y since each edge in the first class is counted twice and each in the second, just once. But $2x + y = (0, 0)$ and since $2x = (0, 0)$ it follows that $y = (0, 0)$. Thus the sum of the labels of the edges crossed by our simple closed curve is (0, 0).

In a closed path that crosses some edge more than once, a multiple crossing can be reduced to one or no crossing since an odd number of crossings contributes the edge label to the sum and an even number of crossings contributes (0, 0) to the sum.

Now, to show that a map with its boundaries properly numbered in three digits can be properly colored in four colors, consider an arbitrary region, label it (0, 0). Consider a path that intersects some edge of each region at least once, and does not pass through any vertex. Assign to each region that label that is the sum of the labels of the edges crossed by the path while reaching the region. This labeling process is consistent for suppose a region can be reached by two paths. The label given by the first is a and the second is b. The two paths form a closed curve, so a + b = (0, 0). But this implies a = b. Furthermore, two neighboring regions have different labels since the label of the edge between them is not (0, 0) and the sum of either region label and the edge must be the label for the other region.

This result may also be stated in the following form: The coloring of the regions of a regular map in four colors is equivalent to the coloring of its edges in three colors. In fact, in place of the edges, we may use an edge map , all of whose regions are four-sided. Such a map may be obtained from a regular map by joining a point, or capital, of each region to the vertices of that region, and then erasing all the edges of the original map. We note that such edge maps are not regular. For this reason their coloration does not follow from Theorem 4.7. Every map with all its regions four-sided is not an edge map, and it is not easy to characterize those maps which are edge maps. Theorem 9.1 is not restricted to the sphere, however you will note that our proof of it is.

Since each map has a region of less than 6 edges, a regular map with each region of 3N edges must contain a traingle, and hence, is reducible. In fact, such a map is colorable in four colors. According to the previous theorem it is sufficient to show that the boundaries can be numbered with three numbers. To achieve this, consider a path along the edges of a regular map that includes each edge at least once. Choose an arbitrary edge of the map, which we number 1, and number from there on as follows. After an edge is traversed and assigned the number a we assign a + 1 to the next edge if we went to the left at the vertex and a - 1 if we went to the right at the vertex, Figure 9.1.

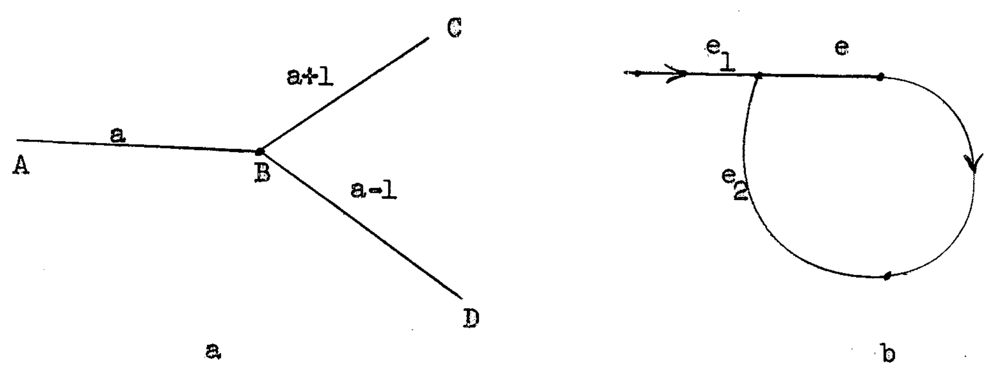


Figure 9.1

In general, an edge may receive several distinct integers as labels. However, it is our intention to show that the labels any one edge receives are all congruent modulo three. In this sense we have only three distinct labels, and no two neighboring labels are alike.

The first edge to be retraced by a path is retraced in either the same or opposite direction.

First suppose that an edge e, Figure 9.1b, is the first to be retraced and that it is traversed in the same direction. Suppose the first label e receives is p and the second is q. Suppose further that the simple closed path is traversed in a clockwise fashion and contains

b' edges, where it goes to the left b_1 times and to the right b_2 times; $b' = b_1 + b_2$ and $q = p + b_1 - b_2$. All edges not belonging to the path but having vertices on the path belong to two classes, those lying on the right with respect to the direction of the path (there are b_1 of them) and those lying on the left (there are b_2 of them). An edge, both of whose ends lie on the path, is counted twice. Let G be the domain inside the path. The edges of the first type lie inside G and those of the second type lie outside G. Let V be the number of vertices that lie inside G, and $N_1, N_2, \ldots, N_r$ the number of edges of the regions lying inside G. Since the map is regular, we have $3V = 2(b_3 + b') + b_2$, also $N_1 + N_2 + \ldots N_r = 2b_3 + b'$. Together these imply $b' + b_2 = 3V - (N_1 + N_2 + \ldots N_r)$. By hypothesis N_i is divisible by 3 so $b' + b_2 \equiv 0 \pmod 3$. But $b' = b_1 + b_2$; hence $b_1 + 2b_2 - 3b_2 = b_1 - b_2 \equiv 0 \pmod 3$ and $p \equiv q \pmod 3$. If the rotation were counterclockwise, we need to interchange roles of b_1 and b_2 to have a proof for that case.

Next suppose that the first edge to be traversed a second time is e_1, as in Figure 9.1b. In the case of a clockwise rotation e_1 would have label $p - 1$ and e_2 would have label $(p + 1) \bmod 3$.

Then the second label applied to e_1 would be $p + 2 \equiv p - 1 \pmod 3$. Also note that if several consecutive edges were retraced in the opposite direction the labeling stays consistent since right becomes left and left becomes right reversing the addition and subtraction processes.

In the second case, if the edge e_1 is inside the simple closed curve, the labels will change accordingly but the argument is essentially the same. Counter-clockwise rotation of the simple closed curve can also be similarly handled.

Since any closed path can be broken up into sections consisting of simple closed paths and retraced (either direction) sections, it follows that the second label given after traversing a closed path is congruent mod 3 to the first label.

Thus, we can see that if we reduce our labels mod 3, each edge will receive a unique label and no two edges with a vertex in common have the same label. Consequently, if each region of a regular map has 3N edges then the boundaries can be properly labeled with three digits or equivalently:

Theorem 9.2. A regular map with each region having its number of edges a multiple of 3 can be colored in four colors.

Once the edge marks used in the proof of this theorem have been established they can be replaced by the edge marks of Theorem 9.1. Note that all three labels will occur at each vertex in the same clockwise (counterclockwise) order about the vertex.

If we draw a small circle about a vertex of a regular map, and regard it as an added region, we obtain a new regular map. We refer to this process as triangulating a vertex. Now, suppose that, for a given regular map, we select certain vertices, V, which we leave unchanged, and that we triangulate the remaining vertices, V'. If the triangulated map is colorable, so is the original map, since we have merely to omit the triangles about vertices V', and leave the coloring of the other regions as they were.

In particular, if the map can be triangulated in such a way that each region of the new map has 3n edges, it will be colorable by the last theorem. The added triangular regions have three edges. Any other region of the triangulated map will correspond to a region of the original map, and will have one vertex for each vertex V and two vertices for each vertex V' in the original region. Thus, our problem is to select the vertices V' such that, when we put 1 at vertices V and -1 (or 2 modulo 3) at vertices V', the sum taken around any region will be divisible by 3, (that is, 0 (mod 3)).

On the other hand, if our map is colorable in four colors, then edge marks exist according to Theorem 9.1. Each vertex has its edges labeled (0,1) (1,0) (1,1) in a clockwise or counterclockwise fashion. Labeling the first set 1 and the second set -1 we can see that triangulation of one of the sets, -1 say, will give a new map with all clockwise vertices. Substituting the labels of Theorem 9.2 and reversing the argument used there, we can see that each region of our triangulated map has 3n edges so that the sum of the vertex labels about each region is zero (mod 3). Since, by triangulation each -1 vertex contributes two to the sum of each region to which it belongs and since $-1 \equiv 2 \pmod 3$, the sum taken about each region of our original map is 0 (mod 3), as prescribed, and we have the following theorem.

Theorem 9.3. The problem of coloring regular maps in four colors is equivalent to the problem of placing 1 or -1 at each vertex in such a way that the sum, taken about each region, is divisible by 3.

It is interesting to note that a set of vertex marks and its dual set (the set obtained by interchanging 1 and -1) each correspond to three sets of edge marks that are all related through the permutations of the three edge marks (there being 6 permutations). Also, each set of edge marks corresponds to a set of 4 colorings (e.g. from the choices of (0,0) (0,1) (1,0) or (1,1) for an arbitrary region); these colorings are all related through a permutation of colors. Each set of vertex marks (and its dual set) then corresponds to a unique coloration (24 colorings that are related through permutation of colors). In a non-simply-connected trivalent map the partition of the vertices gives a unique coloration in each partial map but not in the entire map.

Theorem 9.3 gives a theoretical method of coloring any colorable map. The marks for the vertices may be taken as variables, and the conditions then lead to a system of homogeneous linear congruences modulo 3. Any solution with no variable zero gives a coloration, and if no such solution exists, there is no coloration.

This method is not practical even for simple maps. For example, the dodecahedron may be easily colored empirically but the present method leads to a system of 12 dependent equations in 20 variables. Nevertheless, Heawood spent considerable time and effort in the investigation of these equations, which he called map-congruences. His publication of five papers, totaling 69 pages, dealing primarily with map-congruences, attests to his efforts. As he put it, "the analytical method of treatment seems to bring the problem still more into connection with other mathematical questions, and to give a clearer grasp of the conditions on which it depends."

In 1891 Petersen showed that a third degree linear graph, containing fewer than three leaves, contained a set of edges having exactly one end point at each vertex. His proof was simplified by Brahana and Errera. The proof we shall consider is due to O. Frink and is simpler still. The restriction of fewer than three leaves in a regular map is implied by the restriction that no region border itself, that is, the graph of a map has no leaves. Under this condition, the theorem of Petersen asserts that it is always possible to color the edges in two colors such that one end of the first color and two ends of the second color abut at each vertex.

Theorem 9.4. (Petersen's Theorem) A regular graph of the third degree with fewer than three leaves is colorable (in two colors).

According to Theorem 7.3, a simple graph is colorable. A graph with one leaf is impossible. To color one with two leaves create a new vertex in a 1-cell of each leaf, and join the new vertices by a new 1-cell. The resulting graph is simple and may be colored. If the new 1-cell is red, delete it. If it is blue, change its color according to Theorem 7.2 and delete it. This restores the original graph, which is now colored.

If on decomposing a graph of third degree by Petersen's theorem, the circuits (blue) are all even, they may be broken up further to give a coloring of the edges of the graph in three colors. The difficulty comes when there is one or more pairs of odd circuits, as in the case of the pair of pentagons in Figure 9.2.

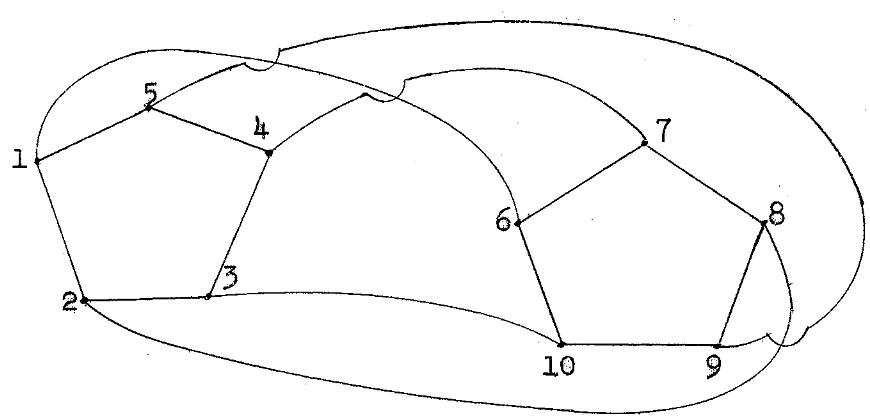


Figure 9.2

However, if the four-color conjecture is true then Petersen's theorem on a sphere is a consequence since the edges would be colorable in three colors.

Sir William Hamilton noticed that the vertices of a dodecahedron could all be traversed by one circuit (along the edges) and made a puzzle based on this fact. Tait conjectured that this held for any convex polyhedron with triple vertices (his "true polyhedron"). This result, if correct, would imply the four-color theorem. However, W. T. Tutte gave the following counterexample.

Consider a pentagonal prism. The edges and vertices constitute a cubical network N (regular third degree graph). Let the five edges which join a vertex of one pentagon to a vertex of the orther be, in their cyclic order, AF, BG, CH, DI, and EJ.

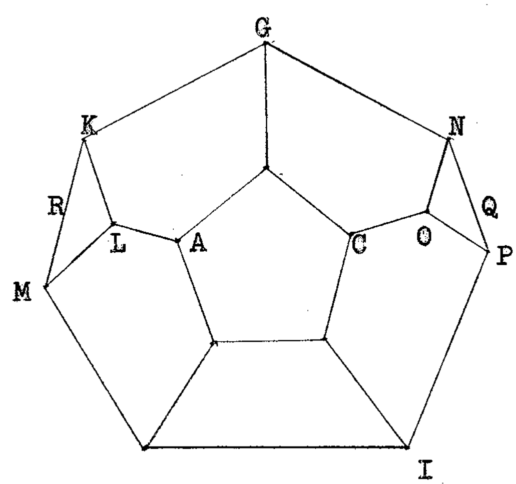


Figure 9.3

In the network N_1 of Figure 9.3, the operation of shrinking each triangle to a point and taking these points as new 0-simplexes preserves Hamiltonian circuits since, for example, a Hamiltonian circuit in this network must contain just two of the 1-simplexes NG, OC, and PI. As the operation gives a network equivalent to that of the pentagonal prism, it follows that no Hamiltonian circuit in Figure 9.3 contains OC and LA since there is no Hamiltonian circuit in N containing AF and HC.

Now if a Hamiltonian circuit in N_1 contains NP, then it contains one but not both of ON and OP, and therefore, it contains OC. Similarly, if it contains KM, it contains LA. Hence, no Hamiltonian circuit in N contains both KM and NP.

Let Q be the mid-point of NP and R the mid-point of KM. Let us now treat Q and R as 0-simplexes, replacing NP by 1-simplexes NQ and QP and replacing KM by 1-simplexes KR and RM. Let us introduce a new 1-simplex joining Q and R. The cubical network N_2 thus constructed has the same structure as the network N_3 which is obtained by taking the part of N_4 (Figure 9.4) contained in the triangle UVX with its interior, and adding three other 1-simplexes joining U, V, and X to some point S. This correspondence is most easily traced by noting the three quadrilaterals in each figure. Let M, K, Q and R correspond to U, V, X, and S.

Clearly any Hamiltonian circuit in N_2 must contain QR. Otherwise, it would be a Hamiltonian circuit of N_1 containing both NP and KM, which we have seen to be impossible. Hence, by the correspondence between N_2 and N_3 any Hamiltonian circuit in N_3 must contain XS. Now any Hamiltonian circuit in N_4 defines one in N_3 in an obvious way; thus, we can deduce that any Hamiltonian circuit in N_4 must contain XW. Hence, by symmetry, it must contain all three of the 1-simplexes WX, WY, and WZ, which is absurd.

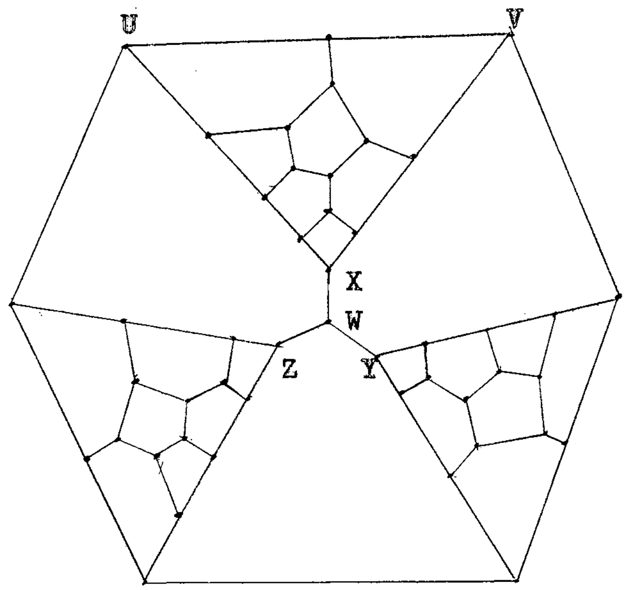


Figure 9.4

Tait also proposed the following which remains unproven: Every bridgeless regular graph of degree three and genus zero can be factored into three factors. The edges of a trivalent map on a sphere form a collection of disjoint simple graphs and is just such a graph. Such a graph also forms a trivalent map. If we generalize Theorem 9.1 to apply to all trivalent maps, the problem of numbering the boundaries is that of factoring the corresponding graph into three factors.

A SELECTED BIBLIOGRAPHY

1. Aleksandrov, P. S., Combinatorial Topology, Graylock Press, Rochester, N. Y. (1956).

2. Arnold, B. H., Intuitive Concepts in Elementary Topology, Englewood Cliffs, N. Y.: Prentice-Hall Inc., (1962).

3. Ball, W. W. R., Math Recreations and Essays, (11 ed.) MacMillan, London, (1954).

4. Barr, S., Experiments in Topology; Crowell Co., N. Y., (1964).

5. Berge, C., The Theory of Graphs, Wiley and Sons, N. Y., (1964).

6. Birkhoff, G. D., "The Reducibility of Maps," Amer. J. of Math., 35 (1913) pp. 115-128.

7. _____, "A Determinant Formula for the Number of Ways of Coloring a Map," Annals of Math., 14 (1912) p. 42.

8. _____, "On the Number of Ways of Coloring a Map," Proc. Edinburgh Math. Soc., (2); 2, (1930) pp. 83-91.

9. Birkhoff, G. D., and Lewis, D. C., "Chromatic Polynomials," Amer. Math. Soc. Trans., 60 (1946) pp. 355-451.

10. Bose, R. C., "On the Construction of Balanced Incomplete Block Designs," Annals of Eugenics, 9 (1939), p. 353.

11. Bourgin, D. G., Modern Algebraic Topology, The Macmillan Co., N. Y., (1963).

12. Brooks, R. L., "On Coloring the Nodes of a Network," Proc. Camb. Phil. Soc., 37 (1941) p. 194.

13. Bruyr, D. L., "An Introduction to Graph Theory," Thesis, Oklahoma State University, (1965).

14. Cairns, S. S., Introductory Topology, The Ronald Press Co., N. Y., (1961).

15. Cayley, A., "On the Coloring of Maps," Proc. London Math. Soc., 9 (1878) p. 148.

16. Coxeter, H. S. M., "Regular Skew Polyhedra in Three and Four Dimensions, and Their Topological Analogues," Proc. of the London Math. Soc., (2) 43 (1937) pp. 33-62.

17. _____, "The Map Coloring of Unorientable Surfaces," Duke Math. J., 10 (1943) p. 293.

18. _____, "Coloring Maps," The Math Teacher, 50 (Dec. 1957) pp. 546-550.

19. _____, "The Four Color Map Problem," Math. Teacher, 52 (1959) p. 283.

20. _____, Introduction to Geometry, N. Y., John Wiley and Sons, Inc., (1961).

21. Dirac, G. A., "Note on the Colouring of Graphs," Math Z., 54 (1951) pp. 347-353.

22. _____, "A Property of 4-Chromatic Graphs and Some Remarks on Critical Graphs," J. London Math. Soc., 27 (1952) p. 85.

23. _____, "Some Theorems on Abstract Graphs," Proc. London Math. Soc., (3) 2 (1952) p. 69.

24. _____, "Map Colour Theorems," Canadian J. Math., 4 (1952) p. 480.

25. _____, "Theorems Related to the Four-colour Conjective," J. London Math. Soc., 29 (1954) p. 143.

26. _____, "Map Colour Theorems Related to the Heawood Colour Formula," J. London Math. Soc., (1956) p. 460.

27. _____,"Percy John Heawood,"J. London Math. Soc., 38 (1963) pp. 263-277.

28. Dybkin, E. B. and Uspenskii, V. A., Multicolor Problems, (1952), D. C. Heath and Co., Boston.

29. Erdos, P. and Rado, R., "Partition Relations Connected with the Chromatic Number of Graphs," J. London Math. Soc. 43 (1959) pp. 63-72.

30. Errera, A., "Une Contribution au Probleme des Quatre Couleurs," Bul. de la Soc. Math. de France, 53 (1925) p. 42.

31. _____, "Histroique du Probleme des Quatre Couleurs," Periodico di Mathematiche, 7 (1927) p. 20.

32. Franklin, Philip, "The Four Color Problem," Amer. J. of Math., 44 (1922) p. 225.

33. _____, "A Six Color Problem," J. of Math. and Physics, 13 (1934) p. 363.

34. _____, "Note on the Four Color Problem," J. of Math. and Physics, 16 (1938) p. 172.

35. _____, "The Four Color Problem," Scripta Mathematica, (5) Yesheva Univ. (1941).

36. Frink, O., Jr., "A Proof of Petersen's Theorem," Annals of Math., 27 (1926) p. 491.

37. Gardner, Martin, "The Four Color Problem," Scientific American, (Sept. 1960).

38. Hall, D. W., Spencer, G. L., II, Elementary Topology, John Wiley and Sons, Inc., N. Y. (1955).

39. Heawood, P. J., "Map-colour Theorem," Quart. J. of Pure and Applied Math., 24 (1890) p. 332-338.

40. _____, "On the Four-colour Map Theorem," Quart. J. of Pure and Applied Math., 29 (1897) p. 285.

41. _____, "On Extended Congruences Connected with the Four-colour Map Theorem," Proc. London Math. Soc., (2) 33 (1932) pp. 253-286.

42. _____, "Failures in Congruences Connected with the Four-colour Map Theorem, Proc. London Math. Soc., (2) 40 (1936) pp. 189-2-2.

43. _____, "Note on a Correction in a Paper on Map Congruences," J. London Math. Soc., 18 (1943) pp. 160-167.

44. _____, "Note on a Correction in a Paper on Map Congruences," J. London Math. Soc., 19 (1944) pp. 18-22.

45. _____, "Map-colour Theorem," Proc. London Math. Soc., (2) 51 (1949) pp. 161-175.

46. Heffter, L., "Uber das Problem der Nachbargebiete," Math. Ann., 38 (1891) p. 477.

47. Hilbert, O. and Cohn-Vossen, S., Geometry and the Imagination, Translated by P. Ne men yi, N. Y., Chelsea Pub. Co., (1952).

48. Kagno, I. N., "A Note on the Heawood Colour Formula," J. Math. and Physics, 14 (1935) p. 228.

49. Kempe, A. B., "On the Geographical Problem of Four Colours," American J. of Math., 2 (1879) pp. 193-200.

50. _____, "How to Colour a Map with Four Colours," Nature 22 (1880) pp. 399-400.

51. Lefschetz, Solomon, Introduction to Topology, Princeton, N. J., Princeton University Press, (1949).

52. Rademacher, H. and Toeplitz, O., Enjoyment of Math., Princeton, N. J., Princeton Univ. Press (1957).

53. Reynolds, C. N., "On the Problem of Coloring Maps in Four Colors," Annals of Math., 28 (1926) p. 1.

54. _____, "Some New Methods of Approaching the Four Color Problem," Proceedings West Virginia Ac. of Science, 1 (1927) p. 91.

55. Tait, P. G., "Remarks on the Colouring of Maps," Proc. Roy. Soc. Edinburgh, 10 (1880) p. 792.

56. Tietze, Hernrich, Famous Problems in Mathematics, Graylock Press, N. Y., (1965).

57. Turan, P., On the Theory of Graphs, Colloquium Math., 3 (1954) pp. 19-30.

58. Tutte, W. T., "On Hamiltonian Circuity," J. London Math. Soc., (1946) p. 99.

59. _____, "A Family of Cubical Graphs," Proc. Camb. Phil. Soc., 43 (1947) pp. 459-474.

60. Ungar, P., "On Diagrams Representing Maps," J. London Math. Soc., 28 (1953) pp. 336-342.

61. Veblen, O., "On the Four-color Problem," Annals of Math., (2) 14 (1912) pp. 86-94.

62. _____, Analysis Situs, N. Y., (1931).

63. Wallman, Herold and Hurewicz, Witold, Dimension Theory, Princeton, N. J., Princeton Univ. Press, (1948).

64. Winn, C. E., "A Case of Coloration in the Four Color Problem," Amer. J. of Math., 49 (1937) p. 515.

65. _____, "On Certain Reductions in the Four Color Problem," J. of Math. and Physics, 16 (1938) p. 159.

66. _____, "On the Minimum Number of Polygons in an Irreducible Map," Amer. J. of Math., 62 (1940) p. 406.

APPENDIX

Jordan Curve Theorem

The unit circle is a subspace of the coordinate plane consisting of all points (x,y) that satisfy the equation $x^2 + y^2 = 1$.

A space C is said to be a simple closed curve if and only if it is homeomorphic to the unit circle. A homeomorphism, between a space S and a space T, as you recall, is a one-to-one, open, continuous mapping of S onto T. S and T are said to be homeomorphic.

The Jordan curve theorem is an important and frequently used result in topology. It states roughly, that there is an inside and an outside of a simple closed curve in a plane. More exactly, if a simple closed curve C lies in a plane and if the points of C are removed from the plane, the remainder of the plane is composed of exactly two connected pieces (components) and the curve C is the boundary of each of these pieces. Intuitively it is impossible to get from one of these pieces to the other in the plane without crossing the curve C.

Let S be a topological space. A subset X of S is said to be a component of S if and only if it satisfies the following conditions:

i) X is non-empty

ii) X is connected

iii) If Y is any connected subset of S satisfying $Y \cap X \neq \emptyset$ then $Y \subset X$.

With this definition and the following notation we can state the Jordan

Curve Theorem. We shall write E^2 for the Euclidean plane with its usual topology and $\bar{A}$ for the union of A with its set of limit points.

(The Jordan Curve Theorem) Let C be a simple closed curve in E^2. Then E - C consists of exactly two components A and B. Moreover, $C = \bar{A} - A = \bar{B} - B$.

For a rigorous development and proof of this theorem see [38]. For a more elementary discussion and a simple proof for the case of a polygon see [2].

The stereographic projection of the extended plane onto the sphere that was considered in Chapter II is a homeomorphism. The one-to-one and onto properties follow from the geometry of the transformation. That the transformation maps open sets into open sets can be seen from an analytical point of view. To prove this it is sufficient to show that circles on the plane transform into circles on the sphere and circles on the sphere, not having N as a point, map into circles on the plane.

Let x and y be coordinates on the plane such that the coordinates of T are (0,0) and let a,b, and c be coordinates in space such that the coordinates of T are (0,0,0) and of N are (0,0,1) and such that (x,y) is (x,y,0). We have the following relationship:

$$a = \frac{x}{1 + r^2} \qquad b = \frac{y}{1 + r^2} \qquad c = \frac{r^2}{1 + r^2} \qquad r^2 = x^2 + y^2$$

and conversely:

$$x = \frac{a}{1 - c} \qquad y = \frac{b}{1 - c} \qquad r^2 = \frac{c}{1 - c}$$

A circle on the sphere is the intersection of the sphere with a plane $Pa + Qb + Rc = W$ with $P^2 + Q^2 > 4W(W - R)$ to ensure actual intersection in a circle. Using the formulas for the transformation we get the equation for the corresponding points on the plane to be; $(R - W)(x^2 + y^2) + Px + Qy = W$. This is a real circle in the plane unless $R = W$, when it is a straight line. But if $R = W$ then the circle on the sphere must pass through N. In any case the sphere is cut into two disjoint open sets one mapping inside a circle on the plane and the other outside unless the circle on the sphere maps into a straight line in which case one open set will map into the set on one side of this line and the other set maps on the other side of this line. In any case open sets map onto open sets. This argument can be easily reversed to show that a circle on the plane maps into the intersection of the sphere with a plane. Thus, the mapping is inverse open and thus continuous. Consequently, the transformation is a homeomorphism.

VITA

Charles J. Searcy

Candidate for the Degree of

Doctor of Education

Thesis: HISTORY OF THE FOUR COLOR PROBLEM

Major Field: Higher Education

Minor Field: Mathematics

Biographical:

Personal Data: Born in Beaver City, Oklahoma, February 14, 1935, the son of Victor and Maxie Searcy.

Education: Attended elementary schools in Beaver City, Oklahoma; was graduated from Beaver City Senior High School, Beaver City, Oklahoma, in 1953; received the Bachelor of Science degree from Panhandle A and M College, Goodwell, Oklahoma, with a major in mathematics, in May 1957; was a participant in National Science Foundation Summer Institutes during the summers of 1961 and 1962; was a participant in the National Science Foundation Academic Year Institute at Oklahoma State University, 1963-64; received the Master of Science degree at Oklahoma State University, in August, 1964; completed requirements for the Doctor of Education degree at Oklahoma State University, Stillwater, Oklahoma, in May 1967.

Professional Experience: Taught mathematics at Tulia Public Schools, Tulia, Texas, 1958-60; Holyoke High School, Holyoke, Colorado, 1960-61; Logan County High School, Sterling, Colorado, 1961-63; Graduate Assistant, Oklahoma State University, Stillwater, Oklahoma, 1964-65; Central State College, Edmond, Oklahoma, 1965-67.