Available online at www.sciencedirect.com
SCIENCE DIRECT®
Discrete Mathematics 302 (2005) 145–172
DISCRETE MATHEMATICS
www.elsevier.com/locate/disc

Reformulating the map color theorem

Louis H. Kauffman

Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, 851 South Morgan Streets, Chicago, IL 60607-7045, USA

Received 17 December 2001; received in revised form 1 July 2003; accepted 22 July 2004
Available online 21 September 2005

Abstract

This paper discusses reformulations of the problem of coloring plane maps with four colors. We include discussion of the Eliahou–Kryuchkov conjecture, the Penrose formula, the vector cross-product formulation and the reformulations in terms of formations and factorizations due to G. Spencer-Brown.

1. Introduction

In this paper we give a concise introduction to the work of Spencer-Brown [8] on the four color theorem and some of the consequences of this work in relation to other reformulations of the four color problem. This work involves a rewriting of the coloring problem in terms of two-colored systems of Jordan curves in the plane. These systems, called *formations*, are in one-to-one correspondence with cubic plane graphs that are colored with three edge colors, so that three distinct colors are incident to each vertex of the graph. It has long been known that the four color problem can be reformulated in terms of coloring such cubic graphs.

We first concentrate on proving two key results. The first is a Parity Lemma due to Spencer-Brown [8]. This lemma is also implied by work of Tutte [9] via translation from edge colorings to formations. The second result, depending on the Parity Lemma, is a proof that a certain principle of irreducibility for formations is *equivalent* to the four color theorem. Spencer-Brown takes this principle of irreducibility (here called the *Primality Principle*) to be axiomatic and hence obtains a proof of the four color theorem that is based upon it. He

E-mail address: kauffman@uic.edu.

0012-365X/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.disc.2004.07.031

also gives proofs of the Primality Principle (see [8, Theorem 17, pp. 168–170]) that depend upon a subtle notion of inverse distinction. This work of Spencer-Brown deserves careful consideration.

The present paper is an expansion of [4]. In that paper, we also prove the Parity Lemma and discuss the primality principle. However, the discussion of factorizability of formations is imprecise in [4] and I have taken the opportunity of this paper to rectify that fault. I hope that this paper attains the desired clarity in regard to parity and primality. In the author's opinion these concepts are central to understanding the nature of the four color theorem, and it is worth a second try at explication.

There are seven sections in the present paper. In Section 2, we give the basics about cubic maps and formations. In Section 3, we prove the Parity Lemma. In Section 4, we give the equivalence of the four color theorem and the Primality Principle. In Section 5, we discuss an algorithm, the parity pass, discovered by Spencer-Brown. The parity pass is an algorithm designed to color a map that has been colored except for a five-sided region. The language of the algorithm is in terms of formations. It is an extraordinarily powerful algorithm and may in itself constitute a solution to the four color problem. It is worth conjecturing that this is so. In Section 6, we discuss an application of formations to the workings of a chromatic counting formula due to Roger Penrose. In Section 7, we apply ideas from formations to the Eliahou–Kryuchkov (EK) conjecture, showing that it can be reformulated in terms of coloring and re-coloring trees, and in terms of the vector cross-product reformulation of the four color theorem.

2. Cubic graphs and formations

A *graph* consists of a vertex set V and an edge set E such that every edge has two vertices associated with it (they may be identical). If a vertex is in the set of vertices associated with an edge, we say that this vertex *belongs* to that edge. If two vertices form the vertex set for a given edge we say that edge *connects* the two vertices (again the two may be identical). A *loop* in a graph is an edge whose vertex set has cardinality one. In a *multi-graph* it is allowed that there may be a multiplicity of edges connecting a given pair of vertices. All graphs in this paper are multi-graphs, and we shall therefore not use the prefix "multi" from here on.

A *cubic graph* is a graph in which every vertex either belongs to three distinct edges, or there are two edges at the vertex with one of them a loop. A *coloring* (proper coloring) of a cubic graph G is an assignment of the labels r (red), b (blue) and p (purple) to the edges of the graph so that three distinct labels occur at every vertex of the graph. This means that there are three distinct edges belonging to each vertex and that it is possible to label the graph so that three distinct colors occur at each vertex. Note that a graph with a loop is not colorable.

The simplest uncolorable cubic graph is illustrated in Fig. 1. For obvious reasons, we refer to this graph as the *dumbell*. Note that the dumbell is planar.

An edge in a connected plane graph is said to be an *isthmus* if the deletion of that edge results in a disconnected graph. It is easy to see that a connected plane cubic graph without isthmus is loop-free.

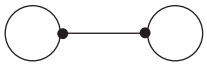


Fig. 1. The simplest uncolorable cubic graph.

Heawood reformulated the four color conjecture (which we will henceforth refer to as the *Map Theorem*) for plane maps to a corresponding statement about the colorability of plane cubic graphs. In this form the theorem reads

Map Theorem for Cubic Graphs. A connected plane cubic graph without isthmus is properly edge-colorable with three colors.

We now introduce a diagrammatic representation for the coloring of a cubic graph. Let G be a cubic graph and let $C(G)$ be a coloring of G. Using the colors r, b and p we will write purple as a formal product of red and blue:

$$p = rb.$$

One can follow single colored paths on the coloring $C(G)$ in the colors red and blue. Each red or blue path will eventually return to its starting point, creating a circuit in that color. The red circuits are disjoint from one another, and the blue circuits are disjoint from one another. Red and blue circuits may meet along edges in G that are colored purple ($p = rb$). In the case of a plane graph G, a meeting of two circuits may take the form of one circuit crossing the other in the plane, or one circuit may share an edge with another circuit, and then leave on the same side of that other circuit. We call these two planar configurations a *cross* and a *bounce*, respectively.

Definition. A *formation* [8] is a finite collection of simple closed curves, with each curve colored either red or blue such that the red curves are disjoint from one another, the blue curves are disjoint from one another and red and blue curves can meet in a finite number of segments (as described above for the circuits in a coloring of a cubic graph).

Associated with any formation F there is a well-defined cubic graph $G(F)$, obtained by identifying the shared segments in the formation as edges in the graph, and the endpoints of these segments as vertices. The remaining (unshared) segments of each simple closed curve constitute the remaining edges of $G(F)$. A formation F is said to be a formation for a cubic graph G if $G = G(F)$. We also say that F *formates* G.

A *plane formation* is a formation such that each simple closed curve in the formation is a Jordan curve in the plane. For a plane formation, each shared segment between two curves of different colors is either a bounce or a crossing (see above), that condition being determined by the embedding of the formation in the plane.

Since the notion of a formation is abstracted from the circuit decomposition of a colored cubic graph, we have the proposition:

Proposition. *Let G be a cubic graph and* $\mathrm{Col}(G)$ *be the set of colorings of G. Then* $\mathrm{Col}(G)$ *is in one-to-one correspondence with the set of formations for G.*

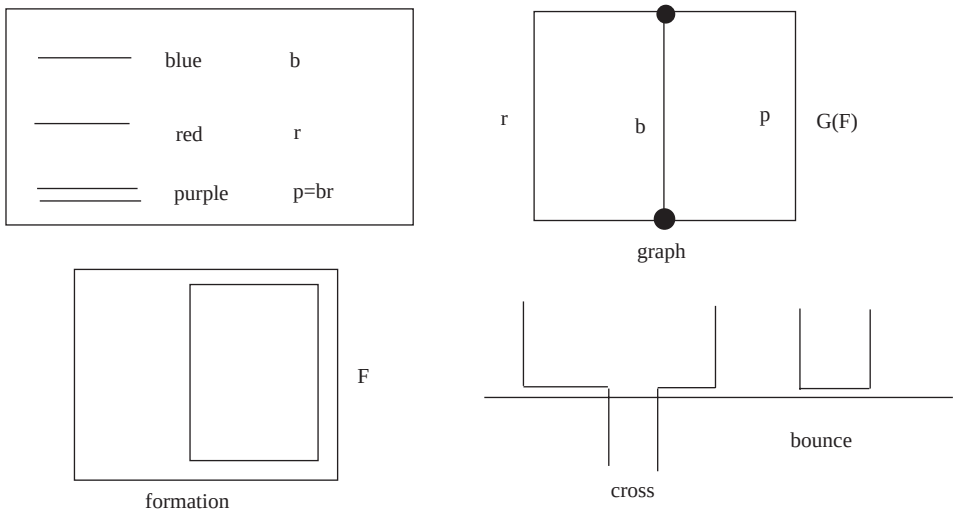


Fig. 2. Coloring and formation.

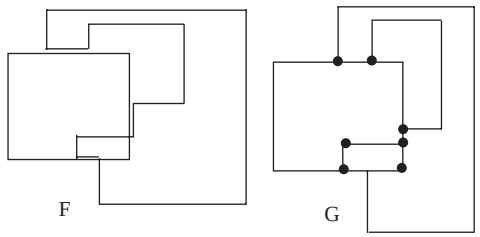


Fig. 3. Second example of coloring and formation.

In particular, the Map Theorem is equivalent to the

Formation Theorem. *Every connected plane cubic graph without isthmus has a formation.*

This equivalent version of the Map Theorem is due to Spencer-Brown [8]. The advantage of the Formation Theorem is that, just as one can enumerate graphs, one can enumerate formations. In particular, plane formations are generated by drawing systems of Jordan curves in the plane that share segments according to the rules explained above. This gives a new way to view the evidence for the Map Theorem, since one can enumerate formations and observe that all the plane cubic graphs are occurring in the course of the enumeration! See Figs. 2 and 3 for illustrations of the relationship of formation with coloring.

Remark. In the figures the reader will note that graphs are depicted with horizontal and vertical edges. This means that some edges have corners. These corners, artifacts of this form of representation, are not vertices of the graph. In depicting formations, we have

endeavored to keep the shared segments slightly separated for clarity in the diagram. These separated segments are amalgamated in the graph that corresponds to the formation.

3. Simple operations and the Parity Lemma

Recall that a *circuit* in a graph G is a subgraph that is equivalent to a circle graph (i.e. homeomorphic to a circle).

Let G be a cubic graph. Suppose that C is a coloring of G with three colors (so that three distinct colors are incident at each vertex of G). Let the colors be denoted by r (red), b (blue) and p (purple). Then, we can classify circuits in G relative to the coloring C. We shall be concerned with those circuits that contain exactly two colors. The possible two-color circuits are r–b (red–blue), r–p (red–purple) and b–p (blue–purple). Let $\Delta(G, C)$ denote the number of distinct two-color circuits in G with the coloring C.

Definition. Call the *parity* of the coloring C, denoted $\pi(G, C)$, the parity of the number of distinct two-color circuits, $\Delta(G, C)$.

Definition. If C is a coloring of G and d is a two-color circuit in G (called a *modulus* in [8]), then we can obtain a new coloring $C' \neq C$ of G by interchanging the colors on d. Call the operation of switching colors on a two-color circuit a *simple operation* on the coloring C.

In this section, we will prove a basic Parity Lemma due to Spencer-Brown [8] in the category of formations. A similar result due to Tutte [9] in the category of plane cubic graphs implies the Parity Lemma, but is proved by a different method. The lemma states that simple operations on planar graphs or planar formations preserve parity. Note that by the results of Section 1, colorings of cubic graphs and formations for cubic graphs are in one-to-one correspondence. The proof of the parity lemma given here is due to the author of this paper.

Note that for a formation F composed of red and blue curves, the two-color circuits are counted by $\Delta(F) = R + B + \mathrm{Alt}$ where R denotes the number of red curves, B denotes the number of blue curves, and Alt denotes the number of red–blue alternating circuits in the corresponding coloring. These red–blue circuits are characterized in the formation as those two-colored circuits that avoid the places where there is a superposition of red and blue (these places correspond to purple edges in the coloring). The red curves in the formation correspond to red–purple circuits in the coloring, and the blue curves in the formation correspond to blue–purple circuits in the coloring.

Each formation corresponds to a specific graph coloring. Simple operations on the coloring induce new formations over the underlying graph. Simple operations can be performed directly on a formation via a graphical calculus. This calculus is based on the principle of *idemposition* saying that: *superposition of segments of the same color results in the cancellation of those segments*. The result of an idemposition of curves of the same color is a mod-2 addition of the curves. Two curves of the same color that share a segment are joined at the junctions of the segment, and the segment disappears. In order to perform a simple

operation on a blue loop, superimpose a red loop upon it and perform the corresponding idemposition with the other red curves that impinge on this red loop along the blue loop. Similarly, in order to perform a simple operation on a red loop, superimpose a blue loop on it and idempose this blue loop with the blue curves that impinge on the red loop. Finally, in order to perform a simple operation on a red–blue alternating circuit in a formation, superimpose a red and a blue loop on this circuit and perform the corresponding idempositions. These instructions for performing simple operations are illustrated in Fig. 4. In this figure some of the edges that are intended to be superimposed are drawn at a short distance from one another in order to enhance the reader's ability to trace the curves.

Parity Lemma. *If C' and C are colorings of a planar cubic graph G with C' obtained from C by a simple operation, then the parity of C' is equal to the parity of C, $\pi(C') = \pi(C)$. Equivalently, parity is preserved under simple operations on planar formations.*

In order to prove the Parity Lemma, we need to consider elementary properties of idempositions of curves in the plane.

First, consider the idemposition of two curves of the same color, as illustrated in Fig. 5. We can distinguish three types of interaction denoted by L (left), R (right) and B (bounce). A bounce (B) is when the second curve shares a segment with the first curve, but does not cross the first curve. Crossing interactions are classified as left and right accordingly, as the person walking along the first curve, first encounters the second curve on his right (R) or on his left (L). After an encounter, there ensues a shared segment that the walker leaves in the direction of the opposite hand. Let $|L|$ denote the number of left crossings between the first and second curves, $|R|$ the number of right crossings, and $|B|$ the number of bounces. (Note the $|L|$ and $|R|$ depend upon the choice of direction for the walk along the first curve.) Let $P(A, A')$ denote the parity of $(|L| - |R|)/2 + |B|$ for an interaction of curves A and A'.

Idemposition Lemma. *Let A and A' be two simple closed curves in the plane of the same color. The parity of the number of simple closed curves resulting from the idemposition of A and A' is equal to $P(A, A') = (|L| - |R|)/2 + |B| \pmod 2$ where the terms in this formula are as defined above.*

Proof. The proof is by induction on the number of crossing interactions between the two curves. It is easy to see that the removal of a bounce changes the parity of the idemposition (see Fig. 6). Fig. 7 illustrates a collection of unavoidable crossing interactions between two curves. That is, if there are crossing interactions, then one of the situations in Fig. 7 must occur. (To see this note that if you follow curve A' and cross curve A, then there is a first place where A' crosses A again. The unavoidable configurations are a list of the patterns of crossing and crossing again.) It is then clear from Fig. 7, by counting parity after the indicated idempositions, that the result follows by induction. □

Proof of the Parity Lemma. Consider a formation F consisting of one red loop A that is touched by a set of n disjoint blue curves. It is clear by construction that the number of alternating (red/blue) circuits in F is equal to the number of curves in the idemposition obtained after letting all the blue curves become red (so that they cancel with the original red

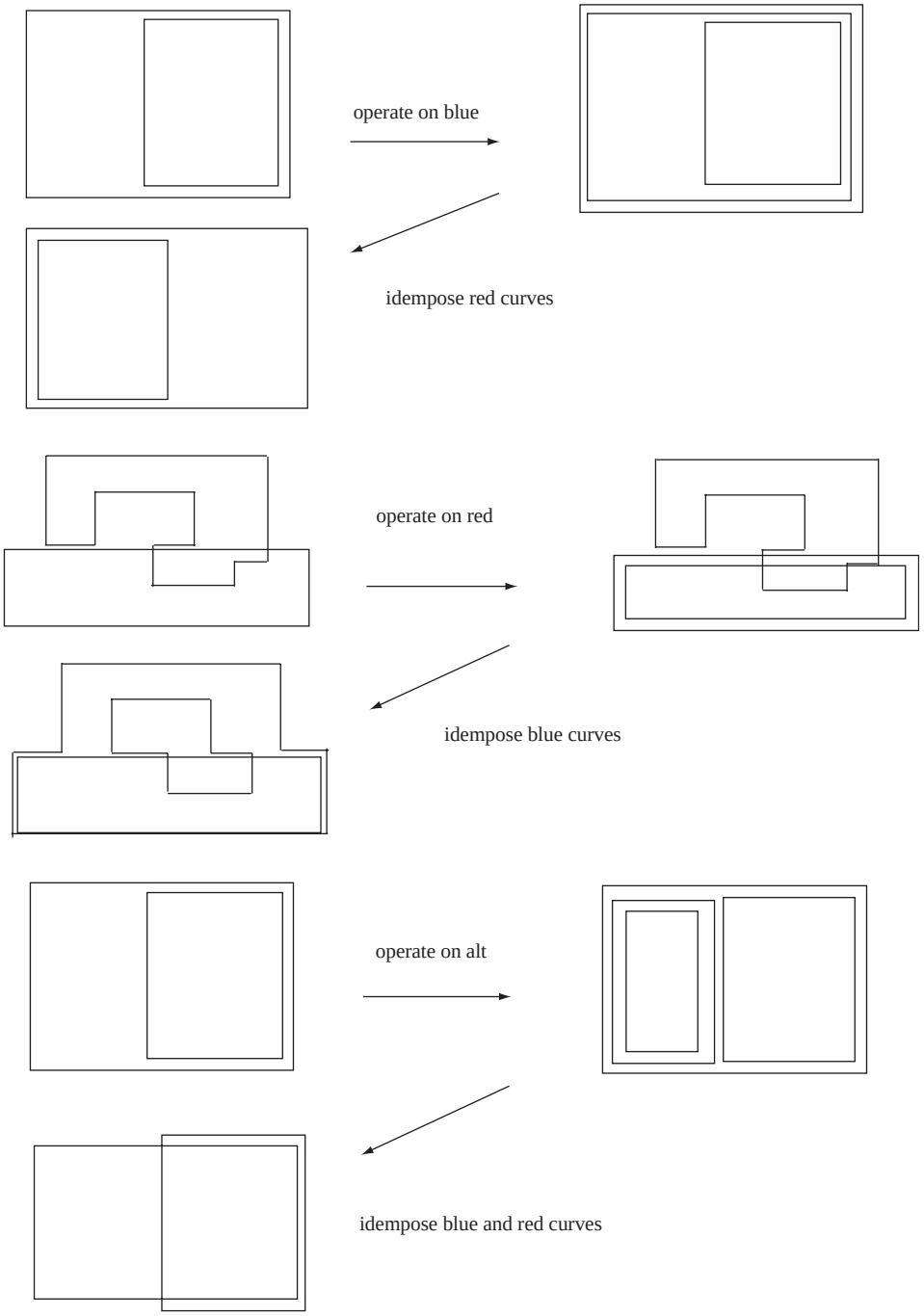
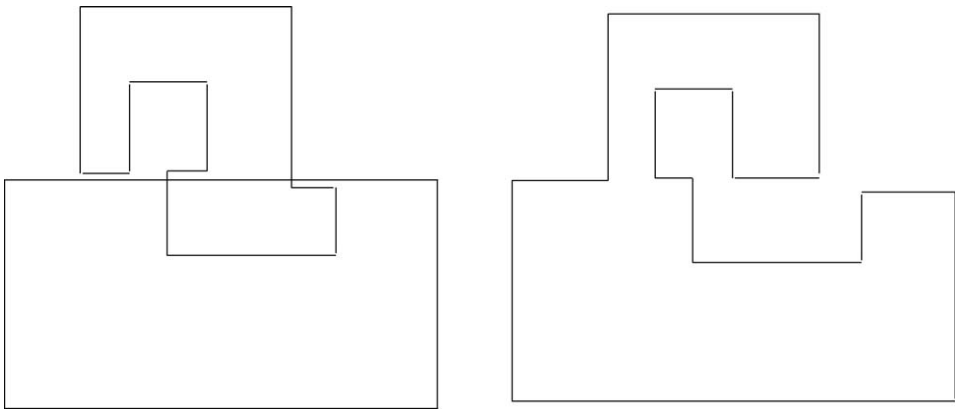


Fig. 4. Simple operations.



(|L|-|R|)/2 + |B| = (1-1)/2 + 1 = 1

Fig. 5. Idemposition.

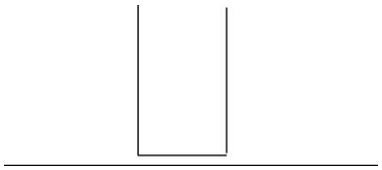


Fig. 6. Bounce.

loop where blue meets red). As a result, we can apply the Idemposition Lemma to conclude that

$$\Delta(F) = 1 + n + (|L| - |R|)/2 + |B| \pmod 2$$

where $|L|$,$|R|$ and $|B|$ denote the total number of left, right and bounce interactions between the red curve and the blue curves and n is the number of blue curves. (Apply the lemma to each blue curve one at a time.) The main point is that the parity of F is determined by a count of local interactions along the red curve A. In this case, when we perform a simple operation on A, the curve count does not change. We simply interchange the roles of blue circuits (i.e. blue/purple circuits) and alternating circuits (i.e. red/blue circuits). Thus, in this case we have that $\Delta(F) = \Delta(F')$, where F' is obtained by a simple operation on the curve A in F. Hence, parity is certainly preserved.

In the general case we have a red curve A that interacts with a collection of blue curves, and these blue curves interact with the rest of the formation. Call the whole formation F, and let G denote the subformation consisting of the curve A and all the blue curves that interact with A. If F' is the result of operating on A in F, then F' will contain G', the result of operating on A in G. G' will consist of the curve A plus all blue curves in F' that touch the curve A in F'. In counting the change of Δ from $\Delta(F)$ to $\Delta(F')$, we actually count the change in the count of blue curves and the change in the count of alternating circuits. Each

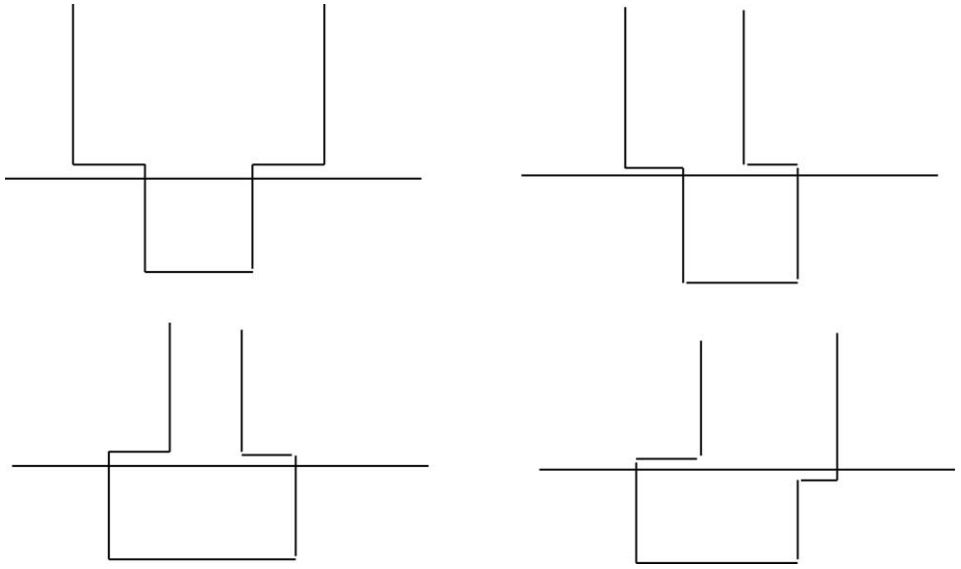


Fig. 7. Innermost cross and recross.

of these changes can be regarded as the result of a single color idemposition originating at A. The change in Δ from F to F' is the sum of the change in the number of blue curves and the change in the number of alternating circuits. Each of these changes is determined by local interactions along the curve A. The parity of the change again depends only on these local interactions. Since the transformation from G to G' has identical local interactions, and since G and G' have the same Δ and hence the same parity, it follows that F and F' have the same parity. This completes the proof of the Parity Lemma. □

Remark. In performing a simple operation, the curve count may change without changing the parity of this count. Figs. 8 and 9 illustrate an example of this phenomenon.

Remark. The Parity Lemma fails for a non-planar formation. For example, consider the formation in Fig. 10. This is a formation for the Petersen graph with one edge removed. As the figure indicates, parity is not preserved by a simple operation on this graph. The curve count in the first formation is five and the curve count in the second formation (after the simple operation) is four. This shows that the underlying graph of these two formations is non-planar.

4. A principle of irreducibility

The main result of this section is the equivalence of the four color theorem with a property of formations that I call the *Primality Principle*. In order to state this property we need to explain the concept of a *trail* in a formation, and *how a trail can facilitate or block an attempt to extend a coloring.*

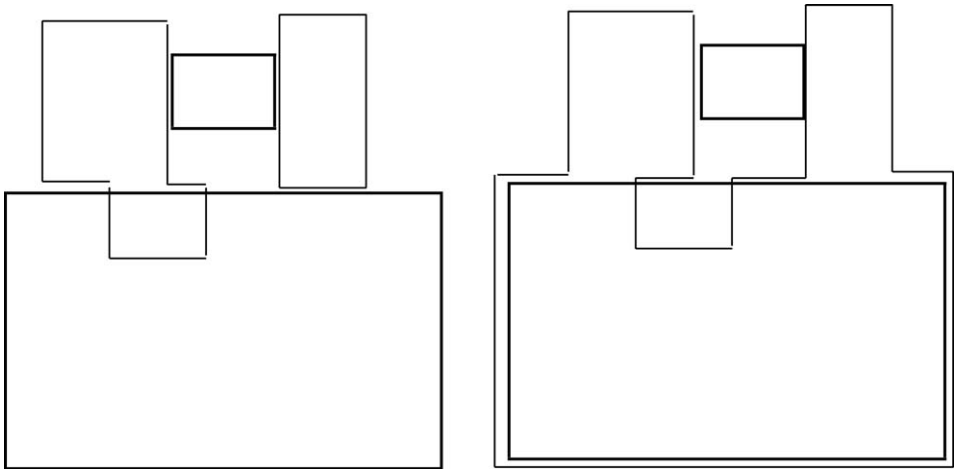


Fig. 8. Changing curve count under simple operation.

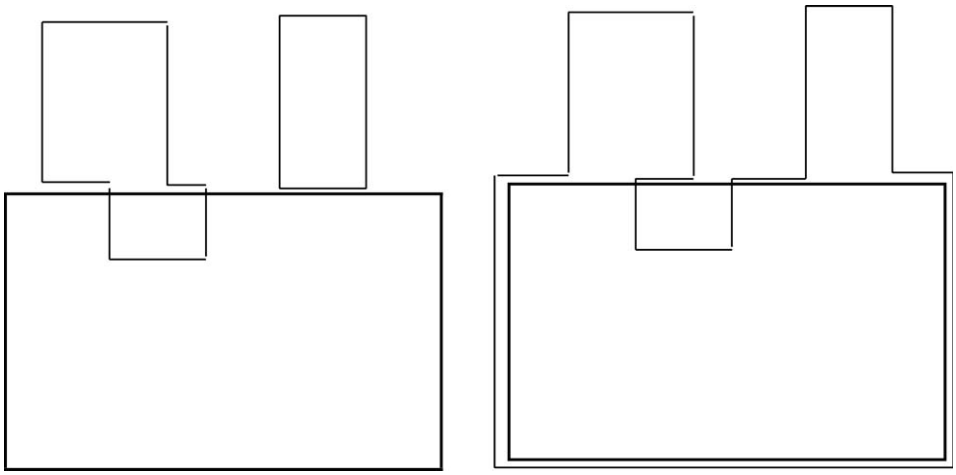


Fig. 9. Unchanging curve count under simple operation.

Consider a formation with two blue curves and a single red curve that interacts with the two blues. See Fig. 11 for an illustration of this condition. I shall call the red curve a *trail* between the two blues. Call the blue curves the *containers* or *contextual curves* for the trail. Call the *graph of the trail T* the cubic graph $G(T)$ corresponding to the formation consisting in the two blues and the red curve between them as in Fig. 12. In Fig. 11 we have also indicated a double arrow pointing between the two blue curves and disjoint from the trail. The double arrow is meant to indicate an edge that we would like to color, extending the given formation to a new formation that includes this edge. We shall refer to this double

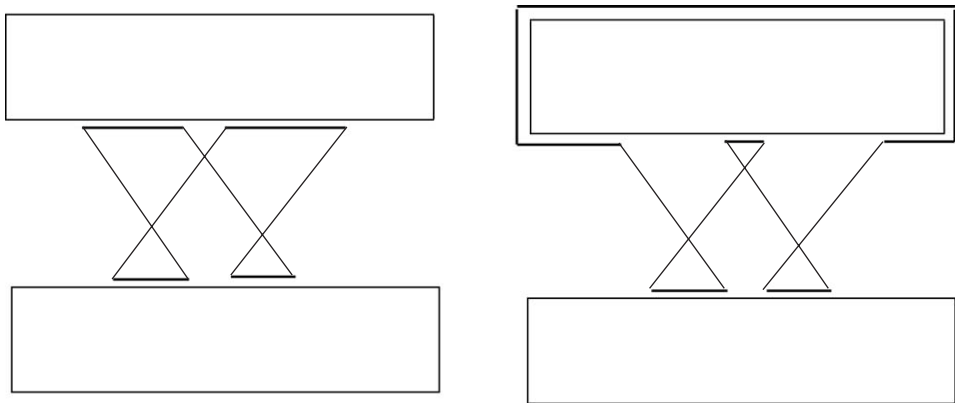


Fig. 10. Parity reversed in the one-deleted Petersen.

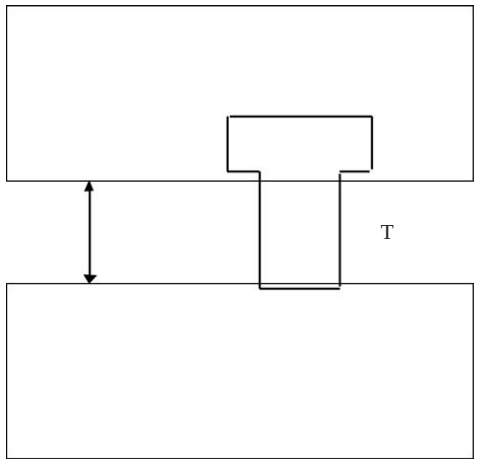


Fig. 11. A trail between two blue curves.

arrow as the *empty edge*. In the example shown in Fig. 13, we obtain this extension by drawing a purple (blue plus red) curve that goes through the empty edge. The part of the purple curve that is not on the arrow forms a pathway in the given formation from one arrow-tip to the other that uses only two colors (red and blue). After idemposition, this purple curve effects a two-color switch along this pathway and the formation is extended as desired. Under these circumstances we say that the formation is *completable over the empty edge.* If simple operations on a given formation with an empty edge can transform it so that the formation is completable over the empty edge, we say that the formation is *completable by simple operations.* Since the final action of completing the formation changes the empty edge to a colored edge, this last operation (described above) will be called a *complex operation.*

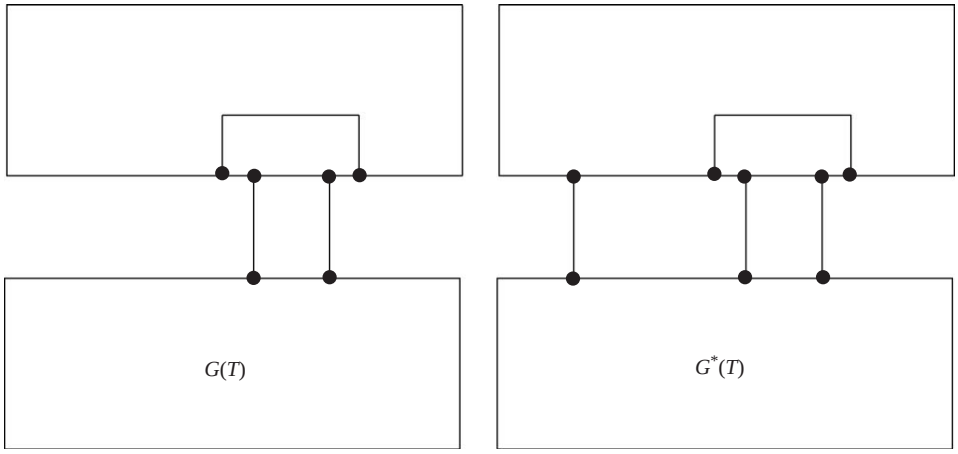


Fig. 12. The graphs $G(T)$ and $G^*(T)$.

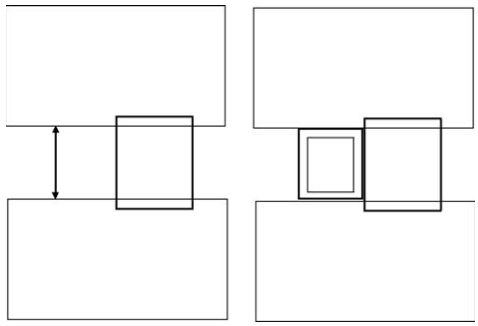


Fig. 13. A colorable trail.

Another example of a trail is shown in Fig. 14. Here, no extension is possible since the extended graph is the Petersen graph, a graph that does not admit a coloration.

In a trail the endpoints of the empty edge are specified, since one would like to complete the formation over the empty edge. The simplest example of uncolorability is just two curves and an empty edge. Then no matter how the curves are colored there is no way to extend the formation over the empty edge.

There are two cases in the coloring structure of a trail: the two contextual curves have the same color or they have different colors. We shall distinguish these two cases by *defining* those colors of the contextual curves to be the colors incident at the endpoints of the empty edge. Note that when we refer to a curve in a formation we mean either a blue curve, a red curve or a cycle that alternates in red and blue when we are performing a parity count. On the other hand, one can also consider purple curves, but these will appear in the formation as alternations of purple with blue or red at those sites where the purple is idemposed with red or blue, respectively. When counting curves, we shall only count blue, red and alternating (red and blue).

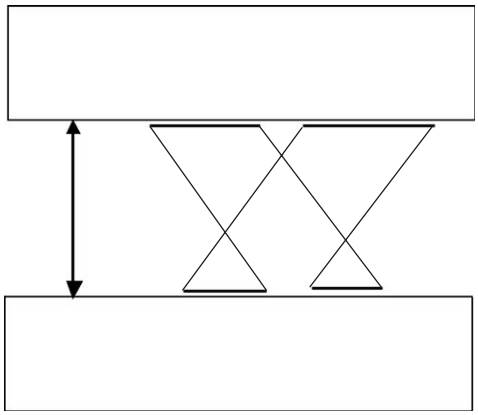


Fig. 14. The Petersen trail.

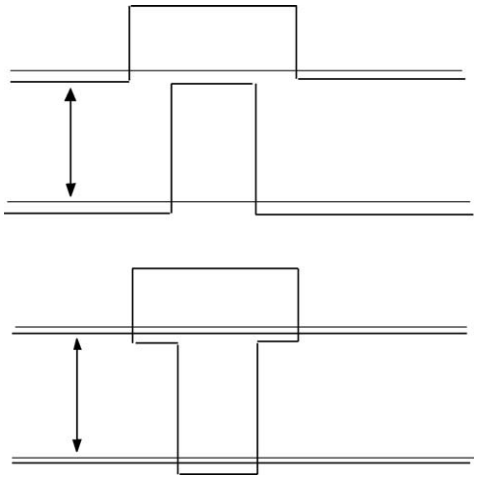


Fig. 15. A trail between two purples.

We first consider contextual curves of the same color. Suppose that both contextual curves are purple. Then any trail between them must be drawn either in blue or in red. It will be called a *non-purple trail*. Thus, one could insert a Petersen trail drawn as a red curve and then idemposed between the purple curves, or a Petersen trail drawn as a blue curve and then idemposed between the two purples. We will say that a trail between two curves of the same color (red, blue or purple) is *factored* if after removing the two contextual curves (by idemposing them with curves of the same color) the remaining trail structure has multiple components.

See Fig. 15 for an illustration of this removal process. The "trail" that we uncover by the removal process is *not* the color of the two curves and it does not touch the endpoints of

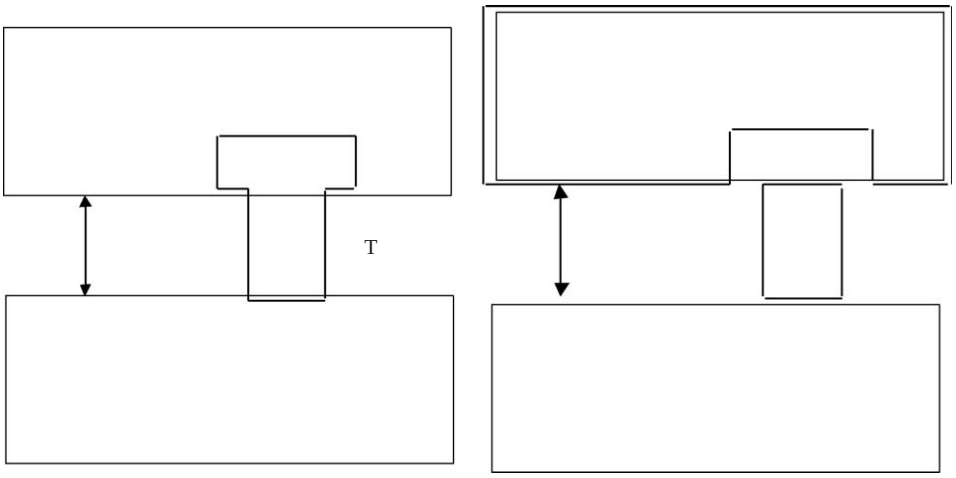


Fig. 16. A factorizable trail.

the empty edge. In Fig. 15, we illustrate a trail between two purples. That is, each endpoint of the empty edge touches the color purple. In the second part of the figure we reveal the purples so that an idemposition of this figure gives the first part and a removal of the two purple curves gives the single trail component. Since there is only one component, this trail is not factored.

For the case where both contextual curves have the same color there is no loss of generality in assuming that the two contextual curves are both red or both blue. Then any extra curves produced in a factorization can be seen directly, in their appearance as alternating, blue or red.

Secondly, suppose that the two contextual curves have different colors. And suppose that the formation has a non-empty trail structure between the two curves. We say that this formation is *factored* with respect to the empty edge if there is an extra curve in the formation that does not pass through either endpoint of the empty edge. For example, perform a simple operation on the Petersen trail as in Fig. 14, making the top curve purple (at an endpoint of the empty edge). Note that in this example every curve in the formation passes through one of the endpoints of the empty edge, so it is not factored. Second example: operate on the top curve in Fig. 11. You will find that this produces a red curve that is isolated from the empty edge, giving a factorization.

We shall say that a formation is *unfactored* if it is not factored.

We shall say that a trail T *factorizes* if there is a formation for the graph $G(T)$ (see definition above) of this trail that is factored. Note that we do not require that the original version of the trail be factored. A new version can be obtained by simple operations on the original formation, or by more complicated re-colorings. A trail is said to be *prime* if it does not admit any factorization.

Sometimes a trail can factorize by simple operations as in the example in Fig. 16. In the example in Fig. 16 we perform a simple operation on the upper blue curve. Note that in the resulting factorization the arrow is now between a lower blue curve and a part of the upper

blue curve that has a superimposed red segment from one of the factors. The Petersen trail of Fig. 14 is a significant example of a prime trail. Recolorings of the graph of the formation of this trail just return the Petersen trail in slightly disguised form.

A trail is said to be *uncolorable* if the graph $G^*(T)$ obtained from $G(T)$ by adding the edge corresponding to the double arrow is an uncolorable graph. Thus, the Petersen trail is uncolorable since $G^*(T)$ is the Petersen graph. A trail is said to be a *minimal uncolorable* trail if the graph $G^*(T)$ is a smallest uncolorable graph. Now, the Petersen graph is the smallest possible uncolorable graph other than the dumbell shown in Fig. 1. In particular, the Petersen is the smallest non-planar uncolorable. This does not, in itself, rule out the possibility of planar uncolorables other than the dumbell (that is, the essence of the four color theorem). Hence, we can entertain the *possibility* of minimal planar uncolorable trails.

Now, we can state the

Primality Principle. A minimal planar (non-empty) uncolorable trail is prime.

In other words, this principle states that there is no possibility of making a minimal planar uncolorable trail that is factored into smaller planar trails. The principle lends itself to independent investigation since one can try combining trails to make a possibly uncolorable formation (i.e. that the graph $G^*(T_1, \ldots, T_n)$ is uncolorable where this graph is obtained from the formation consisting in the trails $T_1, \ldots, T_n$ placed disjointly between two blue curves.) The combinatorics behind this principle are the subject of much of the research of Spencer-Brown. Spencer-Brown regards the Primality Principle as *axiomatic* (see [8, p. 169]). It is one purpose of this paper to point out the equivalence of the four color theorem and the Primality Principle.

Theorem. *The Primality Principle is* equivalent *to the four color theorem*.

Proof. First, suppose the Primality Principle—that minimal uncolorable trails are prime. Let T be a minimal uncolorable non-empty planar trail. Without loss of generality, T is defined by a formation consisting in a single red curve (the trail) drawn between two disjoint blue curves. We call this formation $F(T)$, the formation induced by the trail T. The formation can be depicted so that the two blue curves appear as parallel lines (to be completed to circuits—above for the top line and below for the bottom line) and the trail T is interacting between the two parallel blue lines. In this depiction, we can set a double arrow indicator between the two parallel lines, with this indicator entirely to the left of T. This double arrow indicator represents an edge that we would like to complete to form a larger formation/coloring. Uncolorability of the trail means that there is no coloring of the graph obtained by adding to the underlying graph of $F(T)$ an edge corresponding to the double arrow.

Note that an uncolorable trail is necessarily incompletable (across the empty edge) by simple operations. This implies that there is no two-color pathway in the given formation of the trail from one endpoint of the empty edge to the other endpoint. We can use these facts to count the number of curves in a minimal uncolorable trail.

First, consider a prime uncolorable trail with two blue contextual curves, and an existing trail between them in red. This trail must consist in a single red curve. There can be no other

red curves in the formation. There is an alternating curve incident to each endpoint of the empty edge. Thus, there are at most two alternating curves, one for each endpoint of the empty edge. (Other alternating curves would become red components after the removal of the contextual curves.) If there is one alternating curve, then there is a two-color pathway between the endpoints of the empty edge, and the formation is completable over this edge. Hence, there are two alternating curves. Thus, we see that the curve count (one red, two blue, two alternating) for a prime uncolorable formation with two blue contextual curves is *five.*

Second, consider a prime, uncolorable trail with one purple contextual curve and one blue contextual curve. The trail structure will then consist in red curves woven between the two contextual curves. Once these red curves are idemposed with the purple, the formation can be regarded as two blue curves with the trail structure passing through (say) the upper endpoint of the empty edge, so that this upper endpoint rests on purple. Such a formation is unfactored if and only if all curves pass through the endpoints of the empty edge. Thus, we have a single blue curve and a single alternating curve passing through the lower endpoint, and one red curve and one blue curve passing through the upper endpoint. This makes a total of two blues, one red and one alternator, hence a curve count of *four* for a prime uncolorable formation with contextual curves of different colors.

Now, consider a planar formation $F(T)$ that is minimal, prime and uncolorable. Suppose that it has contextual curves of the same color. Then it has curve count five by the above reasoning. By performing a simple operation on one of the contextual curves, we obtain a formation F' with contextual curves of different colors. The curve count of F' cannot be four, since four and five have different parity. Therefore, the curve count of F' must be five or greater and we conclude that F' is factorized. Similarly, if we begin with a formation that is unfactored and incompletable between two curves of different color, then by operating on one of them we obtain a formation between curves of the same color. The original curve count is four and the new curve count, being of the same parity, is either less than five (and hence solvable) or greater than five (and hence factored). This shows that there does not exist a minimal prime uncolorable (incompletable over the empty edge) planar trail $F(T)$. If there are uncolorables then there are minimal uncolorables. Therefore, no minimal uncolorable planar trail is prime. (The trail factors cannot themselves be uncolorable, since this would contradict minimality.) But this is a direct contradiction of the Primality Principle. Hence, the Primality Principle implies that there are no uncolorable non-empty planar trails.

Now, consider a minimal uncolorable cubic graph. Such a graph entails the possible construction of a minimal uncolorable non-empty trail. Drop an edge from the graph and color the deleted graph. The missing edge cannot have its endpoints on a single curve (red, blue or alternating) in the corresponding formation since that will allow the filling in of the missing edge and a coloration of an uncolorable. Therefore, we may take the missing edge to be between two blues. If there is more than one trail factor between these two blues then we would have a factored minimal uncolorable trail. Primality implies that there is only one factor. Therefore, the Primality Principle in conjunction with the Parity Lemma implies the non-existence of a minimal uncolorable cubic graph with a non-empty trail in the coloration of the deletion (by one edge) of the graph. The only remaining possibility is that after deleting one edge, the graph is identical to two curves. The dumbell (see Fig. 1) is the only such graph. Therefore, the Primality Principle implies the four color theorem.

Conversely, assume the four color theorem. Then indeed there does not exist a minimal uncolorable non-empty planar prime trail (since that by definition implies an uncolorable plane cubic graph with no isthmus). Hence, the statement of the Primality Principle is true. This completes the proof of the Theorem. □

Remark. This Theorem constitutes a reformulation of the four color theorem, in terms of the Primality Principle. This reformulation takes the coloring problem into a new domain. In the work of Spencer-Brown this reformulation has been investigated in great depth. The capstone of this work is an algorithm called the *parity pass* [8, pp. 182–183], that is, intended to extend formations across an uncompleted five region whenever the given formation does not already solve by simple operations. Spencer-Brown has stated repeatedly that this approach gives a proof of the four color theorem. It is not the purpose of this paper to give full review of that work. We recommend that the reader consult Spencer-Brown [8].

5. The parity pass

We shall say that a formation is 1-*deficient* relative to a graph G if it formates all but one edge of G. We shall say that a formation is *planar uncolorable* 1-*deficient* if the underlying graph is uncolorable if one adds a single designated edge to it. In discussing the Primality Principle in the last section, we have considered situations where a graph is formatted all except for a single edge. In that section we showed that the four color theorem was equivalent to the Primality Principle which states that one cannot build planar uncolorable 1-deficient formations by combining colorable trail factors. Experience in working with the calculus of formations tends to bolster one's belief in this principle. There is another approach to coloring, also due to Spencer-Brown that sheds light on this issue. It is well known since Kempe [5] that if one could give an algorithm that would color a cubic map in the plane when a coloring was given at all but one five-sided region, then any map could be colored with four colors. In this section, we describe an algorithm (the parity pass) that is designed to handle the five region in the context of formations. Spencer-Brown asserts that *given a planar formation that is* 1-*deficient at a five-region, it is either completable by simple operations, or at some stage in the parity pass algorithm the resulting formation is completable by simple operations*. We refer the reader to [8] for more details about the context and possible proof of this algorithm. The purpose of this section is to give a condensed description of the parity pass, and to urge the reader to try it out on "hard" examples.

View Fig. 17. This figure contains a complete diagrammatic summary of the parity pass. There is an initial diagram and five successive transformations (A, B, C, D, E) to related diagrams. The last diagram is locally identical to the first diagram. Each transformation consists in a single idemposition of a closed curve on the given formation. In some cases, this curve goes through one of the empty edges, coloring it, while transforming one of the edges at the five region into an empty edge. This is a *complex operation*. In other cases, the transformation is a simple operation on the given formation. In fact A, C and D are complex operations, while B and E are simple operations. Each operation can be performed if the given formation is not completable by simple operations at the five region. Conversely, if one of the operations in the parity pass cannot be performed, then that starting configuration

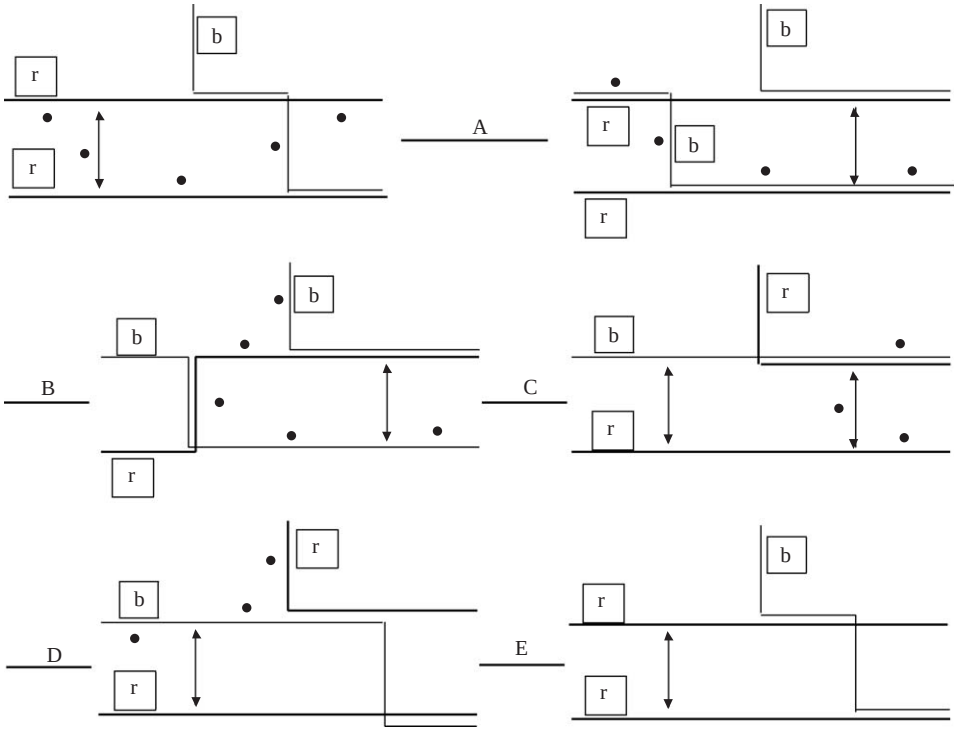


Fig. 17. The parity pass.

can be solved by simple operations. We will not prove these statements here, but we will give a worked example after some further discussion.

At each stage of Fig. 17 we have indicated with small dark circles the edges along which the idemposition is to be performed to get to the next stage. Specifically, performing step *A* requires an idemposition in blue along an alternating curve plus the drawing of this curve across the missing edge and the cancellation of a blue edge by the operating curve. The existence of this idemposition is required to perform step *A*. Step *B* entails idemposition in red along a purple/blue alternator. This is the same as following the indicated blue curve with a red idemposition. It is required that the indicated blue curve is distinct from the other blue curve indicated in the local diagram. Step *C* entails idemposition in purple along a blue/red alternator. Step *D* entails idemposition in blue along a red curve and demands that the two local red segments are part of one curve. Step *E* entails purple idemposition along a blue/red alternator that must be distinct from the other alternator locally indicated. If all steps of the parity pass can be performed, then one returns to a local configuration at the five region that is the same as the starting position.

In a given example, the reader can deduce from each transformed diagram the locus of the putative operation that produces it. This locus, and the type of operation can be deduced by comparing the changes between the diagram and its transform. We also leave to the reader

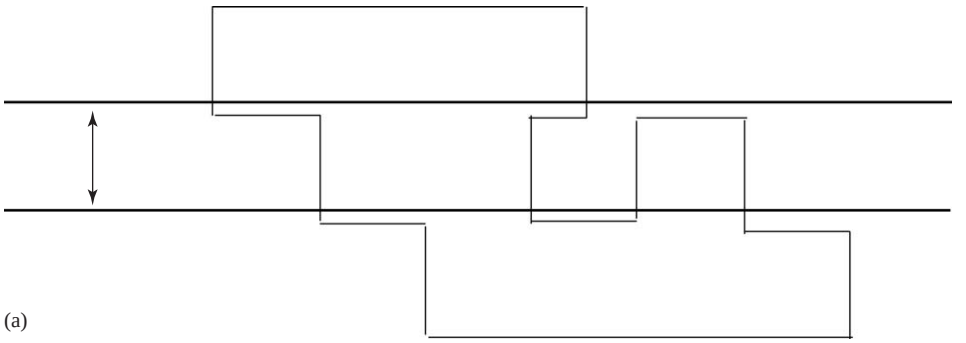


Fig. 18. Culprit number one: (1) Stage A of parity pass applied to culprit number one; (2) Stage B of parity pass applied to culprit number one; (3) Stage C of parity pass applied to culprit number one; (4) Stage D of parity pass applied to culprit number one. Idemposition cannot be performed and formation is completable at the five region.

the verification that *when a transform cannot be accomplished, then the domain formation can be completed by simple operations originating at the five region*.

It is a fascinating exercise to perform this algorithm on examples. Success consists in being *unable* to apply one of the four operations of the parity pass, since this inability implicates a solvable formation. Figs. 18 and 19 give two examples for the reader to examine. These exercises involve quite a bit of diagrammatic work, but it is worth the effort. In Figs. 18(1)–(4) we show the work involved in solving the example of Fig. 18 by the parity pass algorithm. In this case, one can apply steps *A*, *B* and *C* of the algorithm. Step *D* cannot be applied and one can see (Fig. 18(4)) that the formation at this stage can be solved by simple operations. To see this, examine Fig. 18(4) and note that after a simple operation on the curve *T*, one can place a purple curve around the five region that fills in the missing edges and does not cancel any remaining edges.

In these examples, culprit number one (Fig. 18) solves via parity pass after application of *A*, *B* and *C*. Culprit number two (Fig. 19) will solve via parity pass after application of *A* and *B*. Culprit number two is an example of a "good try" at making a factorized minimal uncolorable in the plane. As we mentioned, Spencer-Brown asserts that either the parity pass solves any five region extension problem, or the problem could have been solved by simple operations at the outset.

6. The Penrose formula

Roger Penrose [7] gives a formula for computing the number of proper edge 3-colorings of a plane cubic graph G. In this formula each vertex is associated with the "epsilon" tensor

$$P_{ijk} = \sqrt{-1}\varepsilon_{ijk}$$

as shown in Fig. 20.

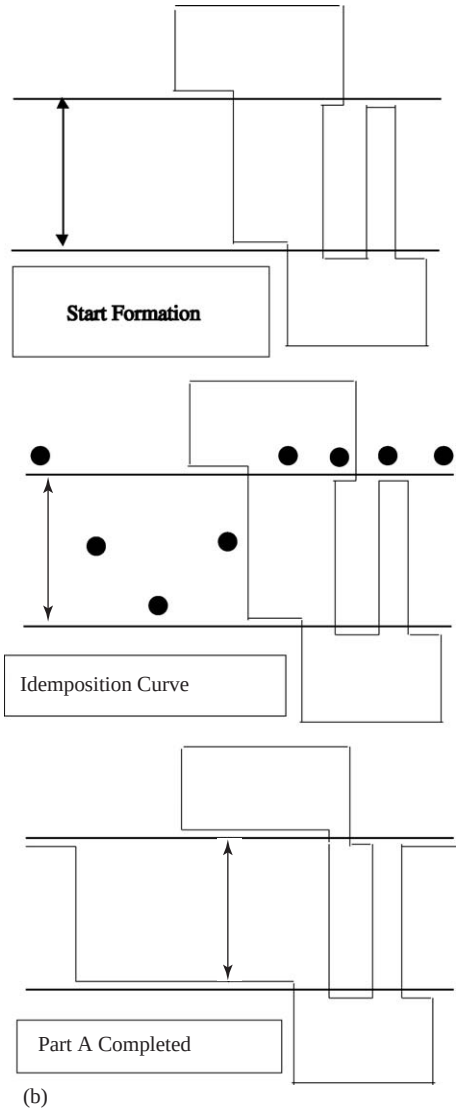


Fig. 18. (*continued*).

One takes the colors from the set {1, 2, 3} and the tensor ε_{ijk} takes value 1 for $ijk =$ 123, 231, 312 and -1 for $ijk = 132, 321, 213$. The tensor is 0 when *ijk* is not a permutation of 123. One then evaluates the graph *G* by taking the sum over all possible color assignments to its edges of the products of the P_{ijk} associated with its nodes. Call this evaluation $[G]$.

Theorem (*Penrose*). *If G is a planar cubic graph, then* $[G]$*, as defined above, is equal to the number of distinct proper colorings of the edges of G with three colors* (*so that every vertex sees three colors at its edges*).

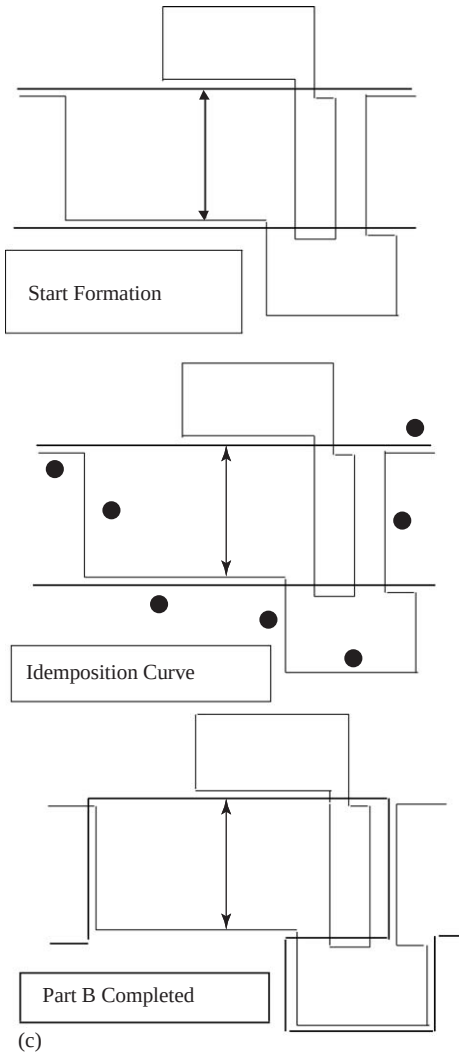


Fig. 18. (*continued*).

Proof. It follows from the above description that only proper colorings of G contribute to the summation $[G]$, and that each such coloring contributes a product of $\pm\sqrt{-1}$ from the tensor evaluations at the nodes of the graph. In order to see that $[G]$ is equal to the number of colorings for a plane graph, one must see that each such contribution is equal to $+1$. The proof of this assertion is given in Fig. 21, where we see that in a formation for a coloring each bounce contributes $+1 = -\sqrt{-1}\sqrt{-1}$, while each crossing contributes -1. Since there are an even number of crossings among the curves in the formation, it follows that the total product is equal to $+1$. This completes the proof of the Penrose Theorem.

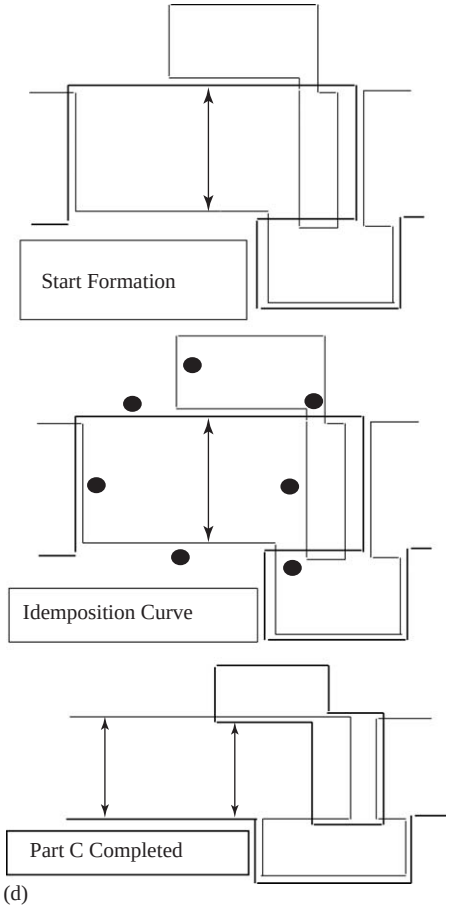


Fig. 18. (*continued*).

It is easy to see from the properties of the epsilon tensor that $[G]$ satisfies the recursive identity shown in Fig. 22. Here, we have that $[O] = 3$, where O denotes an isolated curve, and the recursion formula includes graphs with extra crossings as shown in the figure. This use of formations gives a vivid access to the theory of the Penrose formula.

7. The Eliahou–Kryuchkov conjecture

The EK conjecture [6,1] is about "reassociating" signed trees. The term *reassociation* comes from the algebraic transform of a product $(ab)c$ to a product $a(bc)$. In a non-associative algebra these two terms can represent distinct algebraic elements. In a tree, a trivalent vertex can be regarded as a representative for an algebraic product in the sense that two edge labels are multiplied to give the third edge label at that vertex. See Fig. 23

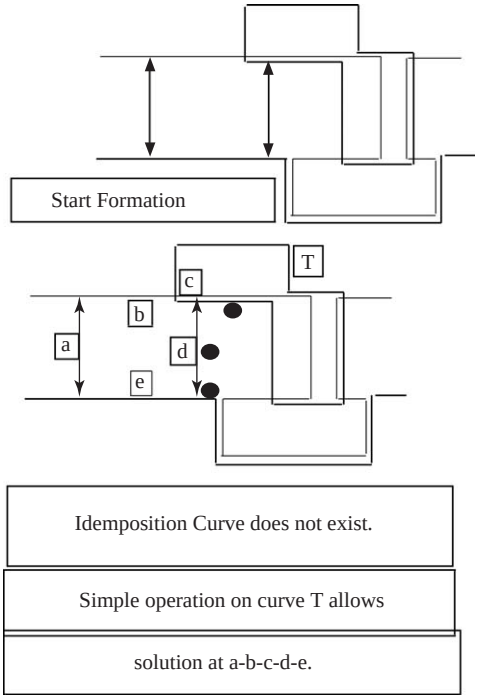


Fig. 18. (*continued*).

for an illustration of this pattern. In this figure we show how two distinct trees correspond to the two associated products $(ab)c$ and $a(bc)$.

The basic reassociation pattern in binary tree form (each vertex is incident to three edges) is shown in Figs. 23 and 24. Note that in the lower half of Fig. 24 the trees are labeled with the colors *p*, *r* and *b* with the product of any two of these colors equal to the third color. In this case, we see that the tree diagram has illustrated the identity $(rb)r = pr = b = rp = r(br)$. Thus, in this case the multiplication is associative. If we decide that $rr = pp = bb = 0$ with $0r = r0 = 0b = b0 = 0p = p0 = 0$, then the system $\{r, b, p, 0\}$ is not associative and two-colored trees changed by a reassociation move as indicated in the figure may have different coloring properties. Note that if we have a colored tree (three distinct colors at a vertex) we can take the two cyclic color orders (*rbp* clockwise or *rpb* clockwise) as denoting two possible signs (plus and minus, respectively) that can be assigned to the vertices of the tree. In the context of the conjecture we are about to discuss one considers arbitrary assignments of signs to the vertices of a tree. The relation with coloring is left out of the game momentarily.

Here is a remarkable game! We shall give signs to the vertices of a tree. We allow the reassociation move inside a larger tree only when the two adjacent vertices in the reassociation are assigned the same sign, and then both vertices receive the opposite of this sign after the reassociation. Such moves are called *signed reassociation moves.*

Eliahou–Kryuchkov Conjecture. Given any two connected trees (with cubic vertices and, at the ends, vertices incident to single edges) and the same number of twigs (a *twig* is an edge

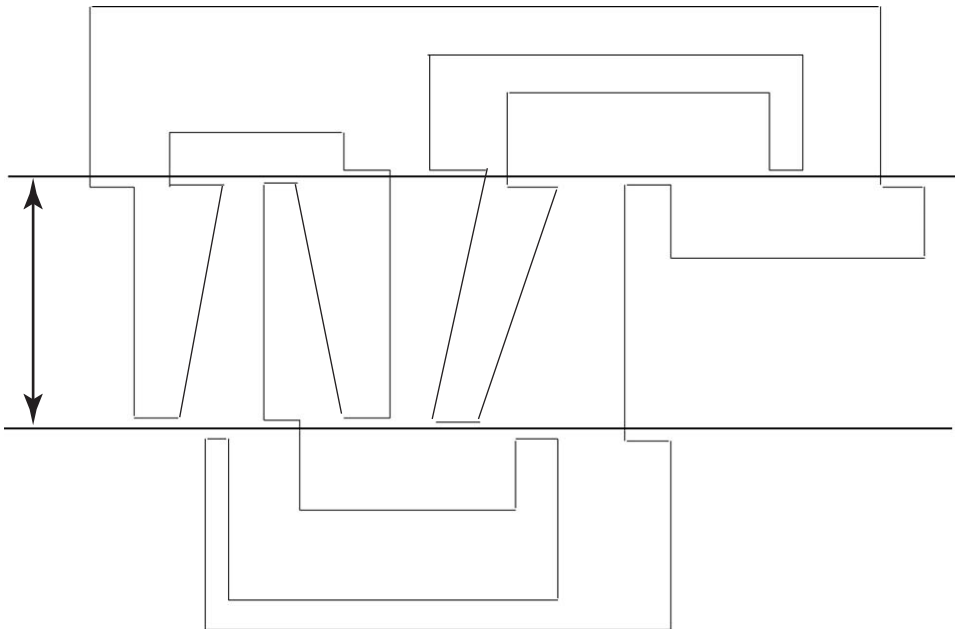


Fig. 19. Culprit number two.

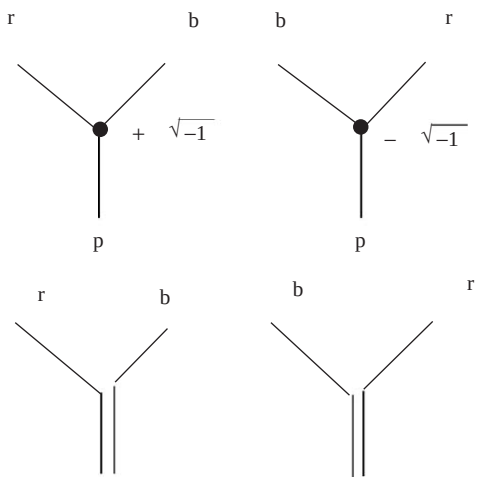


Fig. 20. Epsilon tensor.

incident to an end vertex in the tree), *then there exist assignments of signs to the vertices of the two trees so that one signed tree can be transformed to the other signed tree by a series of signed reassociation moves.* We shall refer to this conjecture as the EK conjecture.

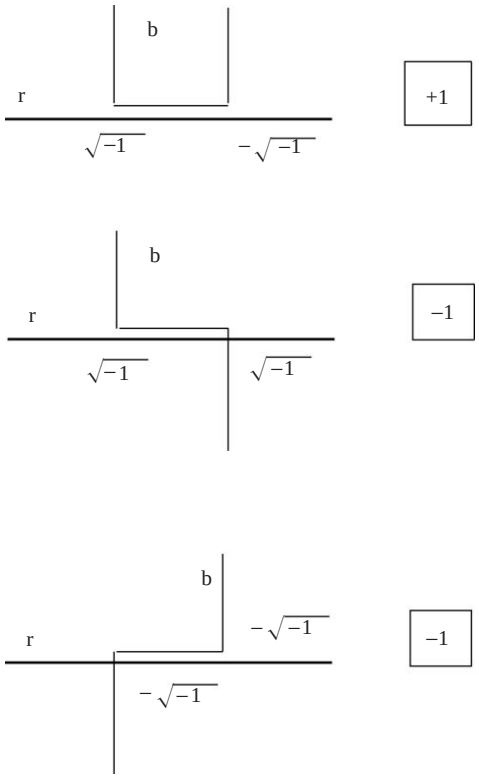


Fig. 21. Cross and bounce.

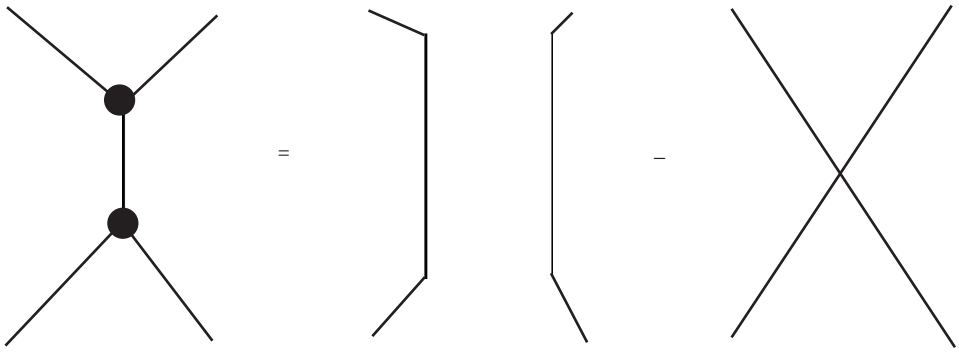


Fig. 22. Penrose formula.

It was known to the authors of this conjecture that the four color theorem follows from it. In [2] it has been shown that in fact the EK conjecture is equivalent to the four color theorem. We mention the EK conjecture here to point out that it implies that *one can edge color the*

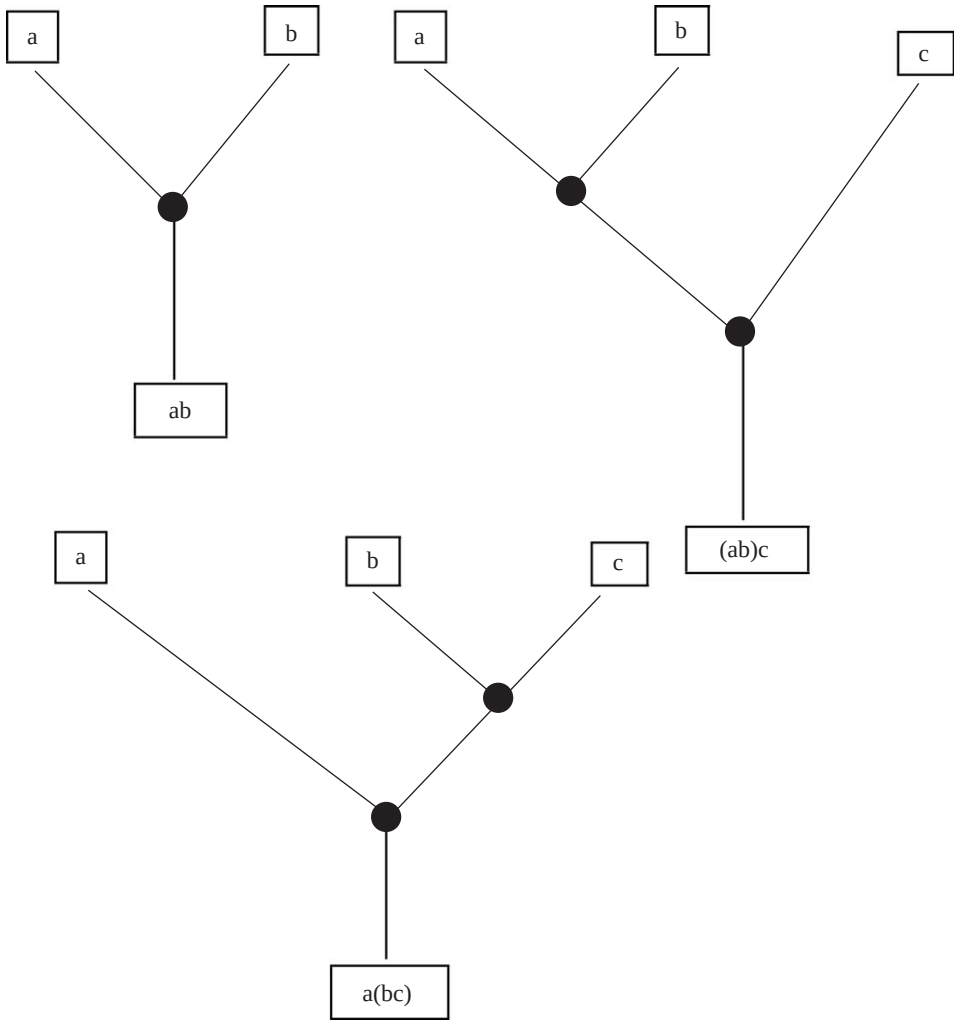


Fig. 23. Multiplication, trees and associated products.

two trees so that one tree can be obtained from the other by reassociation moves on the colorings as shown in Fig. 24 *using formations.* These reassociation moves on the colorings are particularly nice in that they do not involve changing the colors only reconfiguring the graph. The proof of this statement follows directly from the local coloring depicted in Fig. 24. There is a particularly nice pathway of colorings leading from one colored tree to the other.

In this way, the formations make the nature of the reassociation move clear and show how coloring is related to the EK conjecture. It is remarkable that the four color theorem is equivalent to this very specific statement about coloring trees.

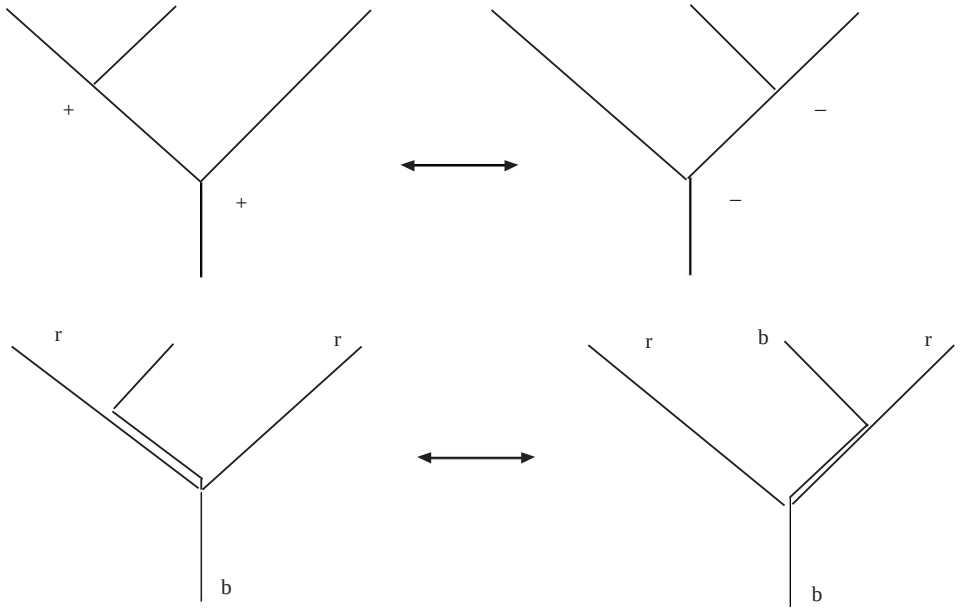


Fig. 24. Signed reassociation.

We can also see just how the EK conjecture is related to the vector cross-product reformulation of the four color theorem [3]. In the vector cross-product reformulation of the four color theorem, we are given two associated products of the same ordered sets of variables. The *vector product conjecture* then states that there exist assignments to the variables from the set of generators $\{i, j, k\}$ of the vector cross-product algebra in three-dimensional space, such that each of the two given products is non-zero in this algebra. This is sufficient to make the two products equal, since if they are non-zero then all partial products are non-zero and hence each product may be viewed in the quaternions. Since the quaternions are associative, it follows that the two products are equal. In this sense, the vector product conjecture is actually a conjecture about the structure of the quaternions.

The relationship of the vector product conjecture with graph coloring is obtained by forming a plane graph consisting of the two trees, tied at their single roots and tied at their branches by non-intersecting arcs, so that the left-most branch of the left tree has the same variable as the right-most branch of the right tree and the product order in the left tree is left-to-right, while the product order in the right tree is right-to-left. *Solving the equality of the two products is equivalent to coloring the graph consisting in two tied trees.*

Let the two associations of the product of n variables be denoted L and R.

Proposition. *The EK conjecture implies that there exists a solution to the equation $L = R$ in the vector cross-product algebra plus a series of algebraic reassociations taking L to R such that all of the intermediate terms in the sequence of reassociations are non-zero when evaluated as vector cross products.*

Proof. The proof of this assertion is easy to see using the formalism of formations by translating signs to colors as we have illustrated in Fig. 24, and using the interpretation of products via trees as shown in Figs. 23 and 24. The signs at the vertices are derived from the fact that in the cross-product algebra we have $ij = +k$ and $ji = -k$. One replaces r, b, p by i, j, k. Local signs in the partial products in the trees can then be used to decorate the vertices of the tree. This completes the sketch of the proof. □

This extra texture in the vector cross-product formulation, and its relationship with the quaternions may provide new algebraic insight into the nature of the four color theorem.

Acknowledgements

It gives the author pleasure to thank James Flagg and Karanbir Sarkaria for helpful conversations in the course of constructing this paper.

References

[1] S. Eliahou, Signed diagonal flips and the four color theorem, European J. Combin. 20 (1999) 641–646.
[2] S. Gravier, C. Payan, Flips signés et triangulations d'un polygone, European J. Combin. 23 (7) (2002) 817–821.
[3] L.H. Kauffman, Map coloring and the vector cross product, J. Combin. Theory B 48 (2) (1990) 145–154.
[4] L.H. Kauffman, On the map theorem, Discrete Math. 229 (2001) 171–184.
[5] A.B. Kempe, On the geographical problem of the four colors, Amer. J. Math. 2 (1879) 193–201.
[6] S.I. Kryuchkov, The four color theorem and trees, I.V. Kruchatov, Institute of Atomic Energy, Moscow, 1992, IAE-5537/1.
[7] R. Penrose, Applications of negative dimensional tensors, in: D.J.A. Welsh (Ed.), Combinatorial Mathematics and Its Applications, Academic Press, New York, 1971.
[8] G. Spencer-Brown, Laws of Form, Gesetze der Form, Bohmeier Verlag, 1997.
[9] W.T. Tutte, On the four-color conjecture, Proc. London Math. Soc. Ser. 2 50 (1948) 137–149.