Human-readable four color map theorem proof?

Klaus Kassner 3 September 2019

When I was a student, my university was visited by Wolfgang Haken who gave a talk on the then recently proved four color theorem. I was deeply impressed by the talk, because it seemed to me to start a new era of approaches to mathematical proofs. Haken and Appel needed a computer to analyse a finite but huge number of case differentiations. The proof could not be reproduced by a human by hand (at least not by someone who was not willing to invest years of diligent work with enough dedication to get rid of all errors in the course of time), so you could call it a proof only, if you trusted in the correctness of the program doing the work. Of course, it is possible to analyse and reanalyse this program, so conviction may be finally achieved. This final process is not so different from analysing a complicated hand-made proof.

In any case, it seemed obvious to me that the borders of provability had been pushed back by the introduction of computer assistance. And even if it is desirable and possible in many cases to finally find a simpler proof that is indeed verifiable by a human without using a computer, there will always be more refined theorems at the border of knowledge, for which only computer-assisted proofs are available, simply because there is no limit to the number of cases that may be necessary to check on a finite domain.

Nevertheless, it would clearly be interesting to have a simple proof for a theorem that is so easy to formulate as the four color theorem: Every planar map is four colorable. Or alternatively: Every (loopless) planar graph is four colorable. Of course one has to explain that four-colorability means that all countries/regions or nodes may be colored using no more than four different colors, with no two neighbouring regions on the map (neighbouring meaning they share a non-zero length piece of boundary) or no two connected nodes on the graph getting the same color. There are some side conditions in the map version – the countries must be connected[1] and must not have fractal boundaries, otherwise more than four of them can have a common (and in the fractal case infinite-length) boundary [1].

Last year, Robert Shuler published an alleged proof of the four color theorem that is human readable [2]. I suspected that this proof as so many before would be wrong but read the paper nonetheless, because *if* there was indeed a simple proof, I was definitely interested in it. To be sure, the "proof" is incorrect, as I will discuss in more detail below. However, it is published in a mathematical journal and up to now I was under the impression that in mathematical journals reviewing a paper involves checking any proof on correctness.[2] This suggested to me that the review process in that journal may be flawed, another indication being that it is an open access journal, and they make money from an accepted paper, whether it is correct or incorrect, whereas they will not make any money from a rejected one.[3] However, in order to make sure that the journal does not meet reasonable standards, I checked whether I could easily spot another wrong paper.

[1] But not necessarily simply connected.

[2] This is why it often takes a year or more to get a mathematical paper through the reviewing process. It is also why I have consistently refused to be a reviewer for a mathematical journal – I never have the time and rarely the competence to verify a cutting-edge research proof. I serve as a referee for theoretical physics papers, where fully rigorous mathematical proofs rarely appear. And I don't have to verify the correctness of computer programs which normally are not sent along with the papers.

[3] Moreover, if I decided to write a comment on Shuler's paper and send it to the journal, I would probably have to pay for its publication, an interesting obstacle.

This turned out to be no problem at all. Reference [3] in the same journal claims to prove the Collatz conjecture.[4] I have a colleague in my group who is quite knowledgable about algorithms and graph theory. He told me he would know if the Collatz conjecture had been proven – it had not. Moreover, I read that paper and while it is much more rigorous mathematically than the Shuler paper, most of its sub-propositions and lemmata are correct and some very interesting, it does not prove what it claims to prove, ending up in a probability argument. Essentially it says that moves that reduce the number in the sequence are more frequent than moves that increase it and so each sequence will finally decrease towards one. But that does not prove the conjecture for *all* positive integers. There could be a small fraction (that is encountered with correspondingly small probability) for which the sequence keeps increasing forever. Or else there might be a different cycle than {4,2,1}, involving only huge numbers.

Given the celebrity of both the four color theorem and the Collatz conjecture and that they both are notorious for their difficulty, any referee should be wary, if someone claims to have something new to say about them, and ought to really examine the submission in depth. This has not happened for the papers [2] and [3], so my recommendation is to not trust papers appearing in the Pure and Applied Mathematics Journal without extensive scrutiny. It seems likely to me that the journal belongs into the list of predatory journals (`https://en.wikipedia.org/wiki/Predatory_publishing`, `https://predatoryjournals.com/journals/`).

Critique of Shuler's approach

Now let us return to Shuler's "proof" of the four color theorem.

The basic idea is the following. A variable is introduced on each node, called flexibility and taking values between 0 and 4. In principle, this variable is supposed to indicate how many colors are available for the node considered. Clearly, flexibilities are ill-defined as long as the sequence of coloring is unspecified, because we can always assign flexibility 4 to the node, where coloring is started and all the other nodes then must have flexibility smaller than 4, because each node added will have at least one neighbor that takes away a possible color from it. Moreover, as I will argue below, flexibilities may depend on the actual choice of colors for previously considered nodes, so it is not even possible for each planar graph to assign flexibilities uniquely by indicating the order in which nodes are to be colored. Rather, it must also be specified whether some previous nodes have the same color or different ones. In short, the flexibilities of different nodes are not independent quantities.

Once we have assigned a consistent set of flexibilities to a map, the number of possible colorings corresponds at least to the product of the flexibilities, so if none of them is zero, we know the map to be four colorable.

Then Shuler sets up a proof by induction. Starting from a graph of three fully connected nodes[5] which all have a flexibility greater than or equal to two, he claims that if the boundary nodes of a graph with n nodes all have a flexibility ≥ 2 an $(n+1)$st node can always be added in a way

[4] Take a positive integer. If it is odd, multiply it by 3 and add one. If it is even, divide it by 2. Repeat to obtain a sequence. The Collatz conjecture is that the sequence will always end in the cycle {4,2,1}.

[5] Not all three-region maps are representable by this graph. For example, the graph corresponding to three regions in the form of concentric rings will consist of three nodes that are not all mutually connected, as the outermost region does not have a common border with the innermost one. But all that is required is that the graphs corresponding to maps can be obtained from fully triangulated graphs by elimination of (zero or more) edges. If the triangulated graph is four colorable, so is any subgraph obtained by elimination of edges.

that the new boundary nodes still all have flexibility ≥ 2. In the process, the flexibility of some nodes that become interior nodes may have to be reduced by at most one, so no flexibility will ever go down to zero and the new graph of size $n+1$ is four colorable. Moreover, it still has the property that its boundary nodes all have flexibility ≥ 2, so the induction can be continued.

It is easy to show *that* this proof must be wrong. A more detailed consideration is necessary to see *why* it is wrong.

To see that the proof must be wrong, it is sufficient to note that by the very claim that the newly added node has a flexibility of at least 2, two colors must be available for this new node. Suppose the graph of size n is *three colorable*. Then we can choose for the newly added node one of the three colors already used, because if two out of four colors are available, one of them must be in the subset of three colors that already have been employed. So the set of $n+1$ nodes is three colorable, if the set of n nodes is. But then we may conclude by induction that any planar graph is three colorable, because the starter graph contains only three nodes and hence must be three colorable. Therefore, if Shuler's claim about flexibilities of added nodes is true, any planar graph is not only four colorable, but even three colorable. This is however incorrect, as the counterexample Fig. 1 shows. The 4-node graph corresponding to the map is fully connected, each node has links to the three others, so four colors are indeed needed to color the graph (or the map).

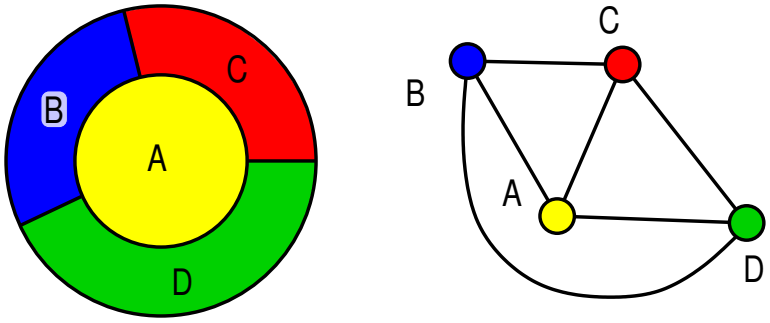

Fig. 1: Map that needs four colors and corresponding graph.

Next, let us try to understand more thoroughly, *why* the proof offered by Shuler is incorrect. First, it is useful to get a more precise idea how flexibilities work. Consider the map of Fig. 2, initially without the outer ring denoted by the letter I, i.e., the map has a central region A and regions B through H, forming a closed ring about it and each one touching its predecessor as well as its successor and A. An important aspect for the following discussion is that the number of these sectorial regions is odd. This implies that the ring consisting of B through H requires three colors. That it cannot be colored with two colors, is proved easily by contradiction. Assume that two colors are sufficient, then they must alternate around the ring. If B holds color 1, C holds color 2, D color 1 again, etc. Since the number of regions in the ring is odd, the last one (H here) would hold the same color as the first, B. But that is forbidden, because they share a common border. Therefore, in order to color the ring, we need three colors and A must then necessarily hold the fourth one. After all, it has a common border with all the ring regions.

Now let us assign flexibilities to regions A through H. First, we start coloring at A, hence we may assign $f_A = 4$. Then B has $f_B = 3$; C touches both A and B, so we can set $f_C = 2$ (at most). This holds for all the following regions except H, so we may assign flexibility 2 to D, E, until G. H has borders with three countries, A, B and G. If B and G happen to have the same

color (B and A can't, nor can G and A), then we may set the flexibility of H to two, otherwise we must choose $f_H = 1$.[6] This already demonstrates that the flexibility of H cannot be fixed uniquely by indicating the sequence of coloring. In some of the possible colorings of the partial map A through G, we may set $f_H = 2$, in others, we have only $f_H = 1$. To be on the safe side, we set $f_H = 1$. This then suggests that the minimum number of possible colorings for the map A through H is $4 \times 3 \times 2^5 \times 1 = 384$.[7]

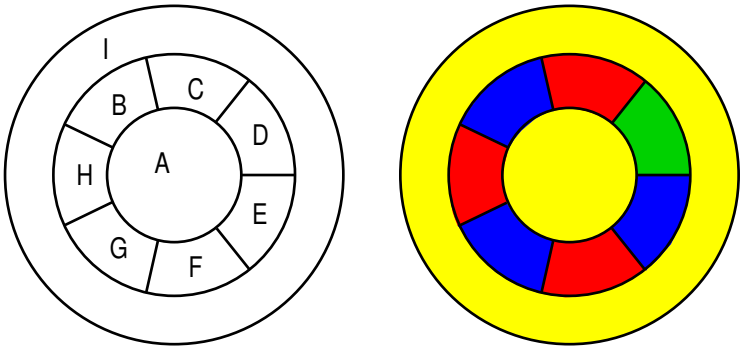

Fig. 2: Map with regions/countries on alternating rings. An odd-numbered ring that is not a single country cannot be colored with fewer than three colors.

In this example, a branching with different possibilities for the flexibility occurred only at the last country, H. However, let us now check what happens if we start coloring at B and first go round the ring. Then we have $f_B = 4$, $f_C = 3$, and $f_D = 3$. Suppose we want next to assign f_A. Obviously, that value would depend on whether so far we have used three colors or only two (i.e., set the color of D equal to that of B). In the first case, we can only set $f_A = 1$, in the second, we may set $f_A = 2$, but since this determines D to have the same color as B, we must reset $f_D = 1$. If instead of setting f_A after coloring D, we continue assigning flexibilities to E through H, we can set each of them except f_H equal to three. H touches two already colored regions, so for safety we would set $f_H = 2$ only. But some of the so-generated colorings of the ring B through H will now have four colors, since we did not restrict the possibilities after the coloring of B, C, and D. This would leave no color for A, i.e., $f_A = 0$. To avoid this, we should, after the coloring of B, C, and D, reserve (one of) the so far unused color(s) for A, which means that each of the following flexibilities can be no larger than two, because both the preceding color and the reserved color cannot be chosen. This gives $f_E = f_F = f_G = 2$ and for f_H we sometimes will be able to choose two and sometimes only one (if B and G have different colors). If we take the safe option $f_H = 1$, we have $4 \times 3^2 \times 2^3 \times 1^2 = 288$ guaranteed possible colorings, which is fewer than in the first assignment of flexibilities which is therefore "more efficient".

Next, let us consider the basic argument of Shuler's proof. Since he works – reasonably – in the realm of graphs rather than maps, I first give in Fig. 3 the graph corresponding to the map of Fig. 2. The node *Inf* is introduced to represent the complete area outside a finite map.[8] Normally, a graph (at least an undirected one) represents only topological information, not

[6] If B and G are forced to have the same colors to allow $f_H = 2$, one of f_B and f_G will have to be set equal to one, because choosing the color of B then implies that of G and vice versa.

[7] In fact, since setting $f_H = 2$ requires to set either $f_B = 1$ or $f_G = 1$ (or else fixing another flexibility on the ring to be equal to one), one of the flexibilities on the boundary will always be one, if coloring is started at A (and there are more than three regions on the ring).

[8] In fact, the four color theorem also holds for infinite maps, whereas Shuler considers only finite ones. If his proof were right, it could however be easily extended to the infinite case.

geometric one.[9] So boundary countries of the map should be characterized by being connected to the node *Inf* rather than by whether they are located at the boundary of the graph or not (in fact, the node *Inf* could be drawn inside one of the triangles involving the node I without changing the topological properties of the graph).

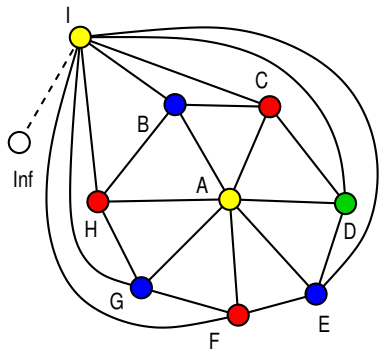

Fig. 3: Graph corresponding to the map of Fig. 2 (with colors on the nodes corresponding to the map on the right-hand side).

This is not what Shuler does. Rather, he defines boundary nodes as nodes on the border of the graph. Why not, if it is useful. I still suspect that if he had used the approach I suggest, which would mean that a newly added node corresponding to a region added at the boundary of a map is connected to the node *Inf* and the connections with *Inf* of those nodes that no longer represent boundary countries are removed and replaced with connections to the newly added node, this would have been less error-prone than his method.

Anyway, let us (for clarity) consider the subgraph consisting of nodes A through H a representative of a map with size n and add node I to get the graph of a map of size $n+1$. First, Shuler argues, there will always be two corner nodes. This would be the nodes F and E in the example. Since no "physical" property in the map (see Fig. 2) corresponds to the corner property (all the countries are inside I, none of them is connected to "infinity" (or the outside of the map) anymore, their choice is somewhat arbitrary, but they must be neighbours (i.e., share a link). For example, we can flip the link IE in Fig. 3 to the left, then node F will be inside the graph and the corner nodes will be E and D.

Next, Shuler says that the newly added node (which is I) will have its flexibility reduced from 4 (which it had as long as it was unconnected with the graph) to 2, because its color must be different from that of the two corner nodes, and since these will remain boundary nodes, we do not want to reduce their flexibility below 2. All the other nodes (G, H, B, C, D in the example) will become inner nodes and in order to keep the flexibility of the added node at 2, the links between them and the added node will be accommodated by reducing their flexibility by 1 each. By the induction assumption, this is possible without producing zero flexibilities, because before addition of I the boundary nodes (all except A) had at least a flexibility of 2 each. Then the new boundary nodes (F, E, and I) will all have flexibility of 2 at least and the construction may continue by addition of node $n+2$ (which we would call J in the example).

The assumption that is supposed to make this work is a certain independence of flexibilities, expressed by the idea that when establishing a link between two nodes with predefined flexibilities, we can accommodate the condition of non-equality of the two colors by reducing the flexibility

[9] It may be argued that planar graphs which are relevant here, have a geometric meaning as well. However, there are theorems (Kuratowski's and Wagner's) allowing to characterize a planar graph in purely topological terms.

of *either* of the two nodes by one. Unfortunately, this freedom of choice whose flexibility is to be reduced is no longer true (in general) when we are dealing with addition of a new node that is to be linked with a *whole group of nodes*. The reason is that by reducing the flexibilities of a group of nodes, connected to the new one, we independently take off some unspecified color from each of them to make it available to the new node. But what is needed is to take away the *same* color from all of the group members, otherwise no additional color will truly be made available. This is however not ensured by a mere reduction of flexibilities.

Then, here is why the Shuler procedure does not work for a structure as the one given in Fig. 3. Suppose we have done the first step of setting the flexibility of I to 2 to accommodate the links with F and E. That is certainly possible. F and E are connected, so must have different colors, which uses up two of the four colors possible for I giving it a preliminary flexibility of 2. Suppose now we wish to keep that flexibility at 2, then this means that all the other nodes to which I is connected must have colors different from the two colors we wish to remain available for I. But only two out of the four colors have this property, and these are the colors of E and F. Hence, all the nodes "between" E and F and connected to I can have only one of the two colors given to E and F. But these nodes form an odd-numbered ring and we have shown before that such a ring cannot be colored with two colors alone. Hence, it is not possible to keep $f_I = 2$. We must make one more color available for the ring, which means $f_I = 1$, and the induction proof fails for maps having the structure of Fig. 2 or graphs having the structure of Fig. 3. And it is obvious that I cannot have a flexibility of 2, because it clearly must have the same color as A, due to the fact that all the other colors must appear in the ring.

Note that this argument is valid independent of the question whether it is indeed possible to assign flexibilities ≥ 2 to all elements of the ring before addition of the new node. Consider the simple map of Fig. 1. Here we definitely can assign flexibilities $f_B = 4$, $f_C = 3$, and $f_D = 2$, so each of the boundary nodes has a flexibility of at least 2. But then we have $f_A = 1$ and hence any closed or open ring region added to the map and touching B, C, and D must have flexibility 1, because it has to have the same color as A. This case is particularly transparent. If we manage to satisfy the induction condition, the newly added node will violate it, and if we assign a flexibility higher than 1 to A in order to have more possibilities for the added region, we will violate the induction assumption by one of the nodes B, C, or D. Presumably, this argument can be extended to more than three outer nodes (as long as their number is odd), providing another refutation of the Shuler approach.

Clearly, the concept of flexibilities is too coarse to construct a proof along the lines suggested by Shuler, as it does not provide a way of describing how to reduce the options of a group of nodes by excluding the same color for all of them. This might be achieved by adding another set of variables describing correlations between colors in the map. For example, we could introduce a correlation matrix g_{ij} defined by $g_{ij} = 1$ for every pair of nodes i, j having the same color and $g_{ij} = 0$ otherwise. Every graph corresponding to a map would then have the constraint that $g_{ij} = 0$ for all nodes being connected by a link. Additional constraints may follow from the structure of the map.[10]

Of course, all this is not to say that the four color theorem is wrong. It has been proven in a number of ways. Up to now, all the correct proofs seem to involve heavy usage of computer power,

[10] For the graph of Fig. 3, we would have $g_{AI} = 1$. Then a reduction of flexibilities of a group of nodes to avoid reducing that of the newly added node I would be required to occur under the side condition that the constraints on g_{ij} are respected. It would then be immediately obvious that the constraint $g_{AI} = 1$ would not allow a flexibility greater than one for node I.

and all the proofs that allegedly do without, including Shuler's, seem to be wrong. Unfortunately. Some ideas what it takes to prove the theorem may be gathered from Ref. [4].

Regarding the discussion part of Shuler's paper,[11] one of his conjectures about wormholes is trivially answerable, because generalizations of the four color theorem to surfaces with higher genus are known. In fact, these generalizations were proved before the four color theorem and without computer assistance. The case of the torus, corresponding to genus 1, turned out to be much simpler than the case of the plane (genus 0). A torus is known to be seven colorable. If we assume that every wormhole can be deformed to spherically symmetric without a topology change, then we can say the following about four colorability (since the metric for a spherically symmetric wormhole is known): All maps on surfaces of constant radial coordinate will be topologically equivalent to planar or spherical maps – those are known to be four colorable. If the two-dimensional surface considered extends through the wormhole and reconnects across the outer space it may become topologically equivalent to either a cylinder, in which case it will still be four colorable, or if reconnection is made along two independent directions, it will become topologically equivalent to a torus, in which case up to seven colors will be necessary for maps drawn on it. Of course, we do not need a wormhole to obtain these assertions, because these topologies of 2D surfaces are already available in ordinary Euclidean space. Non-trivial 3D topology will not add any relevant aspects, since 3D versions of the four color theorem don't make much sense. In three dimensions there are so many ways to connect different volumes with each other that to color n different volumes filling a piece of space, no smaller number of colors than n will suffice in the general case.

As to the remainder of the discussion part of Shuler's article, I consider it to consist of mostly untenable and unscientific speculations. I call a speculation unscientific, if either established scientific knowledge is sufficient to show it to be wrong or at least there is no justification on the basis of established scientific knowledge for making that speculation. To give an example, I do not consider the possibility of time travel – for example via wormholes – an unscientific speculation, but I classify self-contradicting causal loops (like the grandfather paradox) in that category.

It is not impossible for unscientific speculations to change their status under a shift of scientific paradigm. Nevertheless, restrictions apply. For example, after Young's interference experiments, Newton's particle theory of light was dead. Henceforth, it would have been an unscientific speculation that light is made up of particles. With the advent of quantum mechanics that speculation became scientific again. However, and this is the restriction, major predictions from Newton's particle theory remain wrong. In order to explain refraction, it states that light is faster in a medium than in vacuum; moreover, in it the light speed in a gravitational field increases due to gravitational acceleration when light approaches the source of gravity. Both statements are known to be incorrect. Light is slower in a medium than in vacuum which is most easily explained within electrodynamics, a wave theory of light.[12] And general relativity tells us that light becomes slower when moving towards a gravitational center.[13]

References

[1] H. Hudson, Four Colors Do Not Suffice, *The American Mathematical Monthly* **110**, 417 – 423 (2003)

[11] Which is a weird addition to a paper on a mathematical theorem.
[12] A particle theory may explain the slowing down by continual absorption and reemission of photons.
[13] It also tells us that the answer to that question is to some extent coordinate dependent.

[2] R. L. Shuler, Entropy-Like State Counting Leads to Human Readable Four Color Map Theorem Proof, *Pure and Applied Mathematics Journal* **7**, 37 – 44 (2018)

[3] O. de Oliveira Santos, Proving the Collatz Conjecture With Binaries Numbers, *Pure and Applied Mathematics Journal* **7**, 68 – 77 (2018)

[4] F. R. Bernhart, A Digest of the Four Color Theorem, *J. Graph Theory* **1**, 207 – 225 (1977)