An Unavoidable Set of D-Reducible Configurations

John P. Steinberger

August 13, 2010

Abstract

We give a new proof of the four-color theorem by exhibiting an unavoidable set of 2822 D-reducible configurations. The existence of such a set had been conjectured by several researchers including Stromquist [18], Appel and Haken [2] and Robertson, Sanders, Seymour and Thomas [15].

1   Introduction

The four-color theorem states that every map of simply connected countries can be colored in four colors such that countries with a common boundary segment are given different colors. Equivalently, one can say that every planar graph is vertex 4-colorable. The statement was conjectured in 1852 but remained open until 1976, when Appel and Haken [2, 3] announced their proof. Subsequently Robertson, Sanders, Seymour and Thomas [15] obtained a substantially more compact proof. Both proofs are computer-aided.

In this paper we give a new proof similar to the one of Robertson et al. Our proof is technically simpler than both existing proofs because we use only the most basic type of reducible configuration, known as "D-reducible". However the proof is still computerized and, in fact, requires more computer time because it uses more configurations to make up for the loss of non-D-reducible configurations.

All existing proofs of the four-color theorem follow an argument outlined by Heesch [12]. The strategy is to exhibit an "unavoidable set" of "reducible configurations". Reducible configurations are local arrangements of countries that cannot appear in a smallest counterexample because their presence in a map implies the map can be colored from a smaller map by induction. A set of configurations is "unavoidable" if every map contains

1

at least one configuration from the set. The unavoidable set of Appel and Haken has 1476 elements and the unavoidable set of Robertson et al. has 633 elements. Our set has 2822 elements. b Reducible configurations come in two main flavors: C- and D-reducible configurations. Testing a configuration for D-reducibility is easy, but C-reducibility is more delicate. To show a configuration is C-reducible one needs to find a "reducer" that is safe to use in the induction step. There are a large number of possible reducers to consider and there is no single established safety test. Robertson et al. only use reducers that fit a specific safety criterion, whereas Appel and Haken use a wider variety of reducers which lead to a wider variety of complications. There is no safety check necessary for D-reducible configurations, as there are no reducers.

Over half the configurations in the unavoidable sets of Robertson et al. and about one third the configurations in the unavoidable set of Appel and Haken are C-reducible. However both teams conjectured the existence of an unavoidable set of D-reducible configurations. This paper settles the conjecture. Stromquist [18] made the same conjecture a little before Appel and Haken [2].

Proving that a set of configurations is unavoidable requires the use of a "discharging argument", a technique invented by Heesch. Appel and Haken's discharging argument is specified by 487 discharging rules. The bulk of their proof is devoted to showing the correctness of the discharging rules; the task is immense and, apparently, no full independent check has ever been made. Robertson et al. use a different type of discharging procedure, which enables them to have vastly fewer rules (only 32) but checking the procedure by hand is of the same order of tedium as for Appel and Haken. However, Robertson et al. supply machine-readable proof scripts that allow the discharging procedure to be checked by a computer in a few minutes. These proof scripts are probably the most significant difference between Robertson et al.'s proof and Appel and Haken's (though there are other notable differences).

Our own discharging procedure is quite similar to Robertson et al.'s. We have a few more rules (42 instead of 32) but the essential features are the same. As a consequence, the correctness of our discharging procedure can be established with the same type of machine-readable script. The script is verified by Robertson et al.'s original program with some very minor modifications (for example the constant specifying the maximum recursion depth needs to be incremented). The second program Robertson et al. left available, which is for checking configuration reducibility, can also be used to check our own set of configurations after a one-line change to accomodate

2

Ring Size	6–8	9	10	11	12	13	14	15	16
%	0.26	0.53	2.30	6.60	16.7	26.1	26.3	15.5	5.7

Table 1: Percentage breakdown of the unavoidable set of 2822 configurations according to ring-size.

larger configurations, so altogether no new programs are necessary for checking our proof. All one needs to do is swap Robertson et al.'s configurations and rules (contained in two separate files) for our own, and replace the proof scripts. The necessary files and programs are in the 'aux' folder accompanying this arxiv submission. The changes made to the programs are also documented there.

The cost of testing a configuration for reducibility increases roughly by four-fold with its "ring-size" (the number of countries surrounding the configuration). Appel and Haken used only configurations of ring-size at most 14, which they correctly estimated would give them enough configurations to find an unavoidable set. The unavoidable set of Robertson et al. also has only configurations of ring-size 14 or less. Our unavoidable set uses configurations of ring-size up to 16 (a configuration of ring-size 16 takes a few minutes on average to test, but the exact amount of time varies greatly). The ring-size percentage breakdown of our unavoidable set is shown in Table 1. We were not able to find an unavoidable set using only configurations of ring-size 14 or less, but we have no opinion as to whether such a set exists or not.

Altogether we tested over 130'000 configurations during the course of our work. Of these, 42.8% were D-reducible and about 90% were either C- or D-reducible (from a random sample, eschewing[1] safety tests on the reducers, and not counting non-minimally reducible configurations). By comparison, Robertson et al. tested around 14'000 configurations, of which 92.7% were either D- or C-reducible and 54.4% were D-reducible. Thus the "loss rate" consisting of configurations tested that turn out not to be usable climbs from approximately 10% to 50% when C-reducible configurations are disallowed.

In a recent tour de force, Gonthier [10] announced a formal language implementation of the Robertson et al. proof. In this setting, the terms "simple map" and "four-colorable" are first formalized from axioms. Then the statement "M simple map $\implies$ M four-colorable" is verified by a

---

[1]While ensuring the safety of reducers is a necessary and annoying aspect of C-reducibility, it is rare to loose a configuration because none of the reducers are safe.

proof checker which reads through 60'000 lines of formal language proof. In order to trust the theorem has been proved, however, one does not need to check the 60'000 lines; one only needs to check the file containing the definitions, the file containing the statement, and to trust the proof checker itself. This constitutes an improvement over traditional proofs because the proof-checker verifies all details of the proof, including manual combinatorial arguments that are only sketched or left entirely to the reader in the original proof. Presumably, since our proof is structurally identical to Robertson et al.'s, it should also be possible to transcribe it formally.

Our reducibility computations use boolean closures as opposed to the more powerful "block-count" closures [7, 9]. Thus all our configurations are D-reducible in the classical sense of Heesch [12], which fits the purpose of having the proof as non-technical as possible. This is also a stronger result since any boolean-D-reducible configuration is block-count-D-reducible but not vice-versa. Appel and Haken and Robertson et al. do not use any block-count reducible configurations in their unavoidable sets either (the concept was not even known when Appel and Haken published). Robertson et al. did investigate the possibility, however, and advertised an alternate unavoidable set of only 591 configurations containing some block-count reducible configurations.

Section 2 of the paper formulates the graph-theoretic problem and gives basic results on minimal counterexamples; Section 3 introduces the notion of a configuration and of D- and C-reducibility; Section 4 gives the discharging procedure and its proof of correctness. Part of the purpose of the paper is to give an account of the proof that is as self-contained and as accessible as possible.

2   Minimal Counterexamples

The graph-theoretic formulation of the problem is obtained from a map by creating a vertex for every country and joining vertices of adjacent countries by an edge, with one edge added for every common border between two countries (there may be several). The resulting graph is called the *dual* of the map, is planar, and can be vertex 4-colored if and only if the original map is four-colorable. The dual is loopless or else could not be colored at all. In map-theoretic language, this means that a country is not allowed to have a border with itself. Giving precise topological definitions of the terms "map", "country", "border" and so on, and deducing the existence of a well-defined dual graph leads to complications that would take us afield

4

[8, 10]. Instead we will take for granted that the four-color theorem reduces to the 4-colorability of loopless planar graphs (or one can simply adopt the latter formulation as the statement to be proved).

A *minimal counterexample* is a simple (no multiple edges) loopless planar graph that is not 4-colorable but such that every loopless planar graph with fewer vertices is 4-colorable. Clearly, if the four-color theorem is false then minimal counterexamples must exist, so if minimal counterexamples do not exist then the theorem is true.

One can assume that a minimal counterexample is a triangulation, as triangulating the faces of a planar graph only makes the graph harder to color. Or one can prove that, in fact, any minimal counterexample must be a triangulation: if the graph has a face of 4 or more sides then there is always a pair of non-adjacent vertices on the boundary of the face (Jordan curve theorem); identifying these vertices yields a loopless graph with fewer vertices whose coloring gives a coloring for the original graph. It is equally easy to see that a minimal couterexample cannot have any vertices of degree less than four, as one could color the graph by removing and then reinserting these vertices.

Our next, less trivial observation about minimal counterexamples is due to Birkhoff [5]. A *circuit* C is a closed, non-self-intersecting walk on the edges of a graph. We write $E(C)$ for the set of edges of a circuit.

Definition 2.1. *A* short circuit *of a planar graph G is a circuit C with $|E(C)| \leq 5$ such that the two open regions bounded by C contain at least one vertex each of G if $|E(C)| \leq 4$ and at least two vertices each if $|E(C)| = 5$.*

Lemma 2.2 (Birkhoff)**.** *A minimal counterexample has no short circuits.*

Thus, for example, a minimal counterexample G cannot contain a vertex of degree 4, since the 4-circuit of edges around a vertex of degree 4 would be a short circuit of length 4 in G (checking that the two regions defined by the 4-circuit both contain vertices of G uses the fact that G is a simple triangulation of minimum degree 4). Kempe [13] had already shown that a minimal counterexample cannot have a vertex of degree 4.

The import of Lemma 2.2 is best conveyed by an alternate characterization of the result also due to Birkhoff. The *distance* between two vertices u, v of a connected graph G is the length of the shortest path between them (zero if $u = v$) and the *second neighborhood* of a vertex v is the subgraph induced by vertices at distance at most 2 from v. In an arbitrary triangulation the second neighborhood of a vertex may look different than two neatly concentric triangular strips. For example a second neighborhood can

disconnect its complement (this could even be effected by the first neighborhood). But Birkhoff observed that the absence of short circuits implied well-structuredness of second neighborhoods in the sense of the following definition:

Definition 2.3. *The second neighborhood of a vertex v of a planar triangulation is* well-behaved *if the subgraph of G induced by vertices at distance i from v is a circuit for $i = 1, 2$.*

Lemma 2.4 (Birkhoff)**.** *The vertices of a minimal counterexample have well-behaved second neighborhoods.*

Lemma 2.2 implies Lemma 2.4. Lemma 2.4 also implies Lemma 2.2 if one assumes the knowledge that minimal counterexamples have no vertices of degree 4, which can be demonstrated independently of Lemma 2.2.

Proof of Lemma 2.4 from Lemma 2.2. Let v be a vertex of a minimal counterexample G with no short circuits. Thus G has no vertices of degree less than 5. Let d be the degree of v and let $e_1, \ldots, e_d$ be a clockwise ordering of the edges around v. Let $u_1, \ldots, u_d$ be the endpoints of e_1, ..., e_d. Because G is loopless and simple, $v, u_1, \ldots, u_d$ are distinct. Because G is a triangulation, u_i, u_{i+1} are adjacent for $1 \le i \le d$ (setting $u_{d+1} = u_1$). Moreover there are no other adjacencies between the u_i's or else G would have a short circuit of length 3. Therefore the vertices at distance 1 from v induce a circuit of length d in G.

Let $d_1, \ldots, d_d$ be the degrees of $u_1, \ldots, u_d$ in G. For $1 \le i \le d$, let $w_{i,1}, \ldots, w_{i,d_i-3}$ be the $d_i - 3$ vertices adjacent to u_i not in the set $\{v, u_1, \ldots, u_d\}$, and labeled such that $w_{i,1}, \ldots, w_{i,d_i-3}, u_{i+1}, v, u_{i-1}$ is a clockwise list of the vertices adjacent to u_i (with indices having values in the set $\{1, \ldots, d\}$ taken cyclically). Then $w_{i,d_i-3} = w_{i+1,1}$ because G is a triangulation, and $w_{i,j}$ is adjacent to $w_{i,j+1}$ for $1 \le j < d_i - 3$ for the same reason. Let S be the sequence of vertices $w_{1,1}, \ldots, w_{1,d_1-4}, w_{2,1}, \ldots, w_{2,d_2-4}, \ldots, w_{d,1}, \ldots, w_{d_d-4}$ (S is nonempty because $d_i \ge 5$ for all i). Then S contains all vertices at distance 2 from v (because every neighbor of a u_i has been accounted) and the elements of S are cyclically adjacent as just observed. One can easily check that the vertices of S are distinct from the fact that G has no short circuits of length ≤ 4, and that there are no other adjacencies in S besides the cyclic adjacencies from the fact that G has no short circuits of length ≤ 5 (details omitted). Thus the vertices at distance 2 from v also induce a circuit in G. $\square$

We sum up our knowledge of minimal counterexamples with the following definition:

Definition 2.5. *A triangulation is* internally 6-connected *if it has no short circuit and has minimum degree 5.*

Thus one can restate Lemma 2.2:

Lemma 2.6 (Birkhoff)**.** *A minimal counterexample is an internally 6-connected triangulation.*

To keep the paper self-contained we finish this section with a proof of Lemma 2.6, or equivalently of Lemma 2.2. The ideas used—Kempe chain arguments—recur in Section 3, so beginners should at least read part of the proof.

Proof of Lemma 2.6. Let G be a minimal counterexample. We wish to show that G cannot contain a short circuit of length 3, 4 or 5. The fact that G cannot contain a short circuit of length 3 is trivial, since if G has a short 3-circuit one can color the graph consisting of G minus the interior of the 3-circuit and the graph consisting of G minus the exterior of the 3-circuit separately, and recover a coloring for G by permuting one of the colorings to match the other.

To see that G cannot contain a short circuit of length 4 let a, b, c, d be the vertices in order of a short circuit of length 4 of G. Let A, B be the graphs obtained by deleting respectively the interior and exterior of the circuit $abcd$. One can color A and B by induction, but there is no guarantee the two the colorings induced on $abcd$ will be equivalent under permutations. In fact there are four equivalence classes under permutations for colorings of the 4-circuit; if the colors are $\{\mathtt{r}, \mathtt{g}, \mathtt{b}, \mathtt{y}\}$, representatives for the four classes are $\mathtt{rgby}$, $\mathtt{rgry}$, $\mathtt{rgbg}$ and $\mathtt{rgrg}$.

Let A' be the graph obtained from A by adding an edge between a and c; then A' is a loopless planar graph because a and c are distinct (definition of a circuit). Since A' has fewer vertices than G it is 4-colorable, and induces a coloring on the circuit $abcd$ where a and c have distinct colors. Hence A either admits a 4-coloring where $abcd$ are colored $\mathtt{rgby}$ or admits a coloring where $abcd$ are colored $\mathtt{rgbg}$. One could instead add an edge between b and d, which would lead to the conclusion that A either admits a 4-coloring where $abcd$ are colored $\mathtt{rgby}$ or admits a 4-coloring where $abcd$ are colored $\mathtt{rgry}$. Thus either the coloring $\mathtt{rgby}$ extends to A or else both the colorings $\mathtt{rgbg}$ and $\mathtt{rgry}$ extend to A.

As the same observation applies to B, we can assume without loss of generality that `rgby` extends to A and that `rgbg` and `rgry` extend to B, or else A and B induce some common coloring on *abcd*. Now consider a coloring of A where *abcd* are colored `rgby`. There may or may not be a path of $\{\mathtt{r}, \mathtt{b}\}$-vertices from the vertex a (which is `r`) to the vertex c (which is `b`) in A. Assume first that there isn't. Then the maximal connected component of `r`-`b` vertices attached to c does not include a. If we switch the colors `r` and `b` inside of the component we obtain another valid coloring of A in which the colors of a, b and d remain `r`, `g` and `y`, but the color of c is switched to `r`, so the coloring `rgry` extends to A, which settles the case because that coloring also extends to B.

In the other case there is a path of `r`-`b` vertices from a to c in A. In this case there cannot be a `g`-`y` path from b to d (Jordan curve). Therefore we can switch the colors `g` and `y` in, say, the maximal connected component of `g`-`y` vertices containing d without affecting b, and obtain the non-equivalent coloring `rgbg` of *abcd*. But this coloring extends to B, so we are again done. Thus a minimal counterexample cannot contain a short 4-circuit.

Finally to see that G cannot contain a short circuit of length 5 let *abcde* be a short circuit of length 5 in G, and again let A and B be the graphs obtained by deleting respectively the interior and exterior of the 5-circuit *abcde* from G. Both A and B can be 4-colored by induction, and each 4-coloring induces a 4-coloring of the 5-circuit *abcde*. There are only two different equivalence classes under color permutations of colorings of the 5-circuit up to rotation—representatives are `rgrgb` and `rgryb`—for a total of 10 different equivalence classes of colorings of the 5-circuit. We let $\mathtt{P}_1, \ldots, \mathtt{P}_5$ be the 5 equivalence classes whose representatives are the cyclic permutations of `rgrgb` and $\mathtt{Q}_1, \ldots, \mathtt{Q}_5$ the 5 equivalence classes whose representatives are the cyclic permutations of `rgryb` (with the cyclic shifts being arranged such that, say, $\mathtt{P}_2$ has representative `brgrg`, etc).

If H is a graph with boundary *abcde* we write $H|\mathtt{P}_i$ to mean that colorings from $\mathtt{P}_i$ can be extended to H, and likewise for $H|\mathtt{Q}_i$. The conclusion will quickly follow from the following two observations, valid for $H = A, B$ (indices are taken cyclically):

(#1) $H|\mathtt{P}_i \wedge \neg H|\mathtt{P}_{i+1} \implies H|\mathtt{Q}_i$

(#2) $\neg H|\mathtt{P}_{i+1} \wedge \neg H|\mathtt{Q}_{i+1} \implies H|\mathtt{P}_i$

To see (#1), assume for example that $A|\mathtt{P}_1$ but that $\neg A|\mathtt{P}_2$. In particular, the coloring `rgrgb` extends to A, but the coloring `rgbgb` does not. Fix a coloring of A that induces `rgrgb` on *abcde*. If there is no `r`-`b` path from a
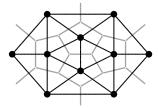
to c in A then a r-b color interchange on the r-b component containing c yields the coloring rgbgb, a contradiction, so there must be a r-b path from a to c. But then there is no g-y path from b to d, so by doing a g-y color interchange on the g-y component containing d we get the coloring rgryb from $\mathtt{Q}_1$, as desired.

To prove (#2) form a smaller graph by identifying the vertices a and c in A; the coloring induced on $abcde$ must give equal colors to a and c, so colorings from one of the equivalence classes $\mathtt{P}_1, \mathtt{Q}_1, \mathtt{P}_5$ must extend to A, which is equivalent to saying that $\neg A|\mathtt{P}_1 \wedge \neg A|\mathtt{Q}_1 \implies A|\mathtt{P}_5$, as desired. This slips a technicality: if a and c are already adjacent in A then identifying them makes a loop, and the theorem cannot be applied by induction (only loopless planar graphs can be 4-colored; if the reader feels this is unfair and that loops should simply be ignored, they should consider that the coloring of the smaller graph with a and c identified does not pull back to a valid coloring of A). The way around this case is to notice that if a and c are adjacent in A, then G already contains a short circuit of length 3 or 4.

To finish the proof that G cannot contain a short circuit of length 5, form a new graph from A by adding a vertex in the middle of the circuit $abcde$, connected to all vertices of the circuit. The new graph is planar, loopless, and is smaller than G because the interior of $abcde$ has at least two vertices in G. By induction it can be 4-colored, which shows that A must have a coloring that uses only 3 colors for the circuit $abcde$, so $A|\mathtt{P}_i$ for some i. Since likewise $B|\mathtt{P}_j$ for some j, and since we can assume that there is no equivalence class whose colorings extend both to A and B, there must be some i such that $A|\mathtt{P}_i$ and $\neg A|\mathtt{P}_{i+1}$. Without loss of generality, say that $i = 1$. Then (#1) implies $A|\mathtt{Q}_1$ and (#2) applied with $H = B$ and $i = 5$ implies $B|\mathtt{P}_5$. Since $\neg B|\mathtt{P}_1$ (because $A|\mathtt{P}_1$) one can apply (#1) again to get $B|\mathtt{Q}_5$ and then (#2) again to get $A|\mathtt{P}_4$. Keeping on one reaches the conclusion that $A|\mathtt{P}_2$, a contradiction. This finishes the proof that a minimal counterexample cannot have a short circuit. $\square$

3   Reducible Configurations

Birkhoff also investigated circuits of length 6 but concluded that similarly simple results wouldn't be forthcoming for circuits of that length. Without going into details, Birkhoff convinced himself by finding two disjoint sets of colorings of the 6-circuit such that (i) each set was self-consistent with respect to Kempe chain arguments, and (ii) each set accomodated colorings imposed on the 6-circuit by replacing the interior of the circuit with
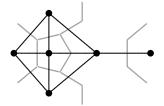
Figure 1: The Birkhoff diamond A (left) and its reducer A' (right), shown with their dual maps.

any small[2] graph, including all combinations for identifying vertices of the 6-circuit or adding chordal edges to the 6-circuit. Thus, it seemed, two arbitrary graphs A and B bounded by a common 6-circuit could *a priori* induce two disjoint sets of colorings on the circuit.

Instead of working with two unknown graphs on either side of the circuit Birkhoff then tried fixing one of the graphs. He found a pair of graphs (A, A') with 6-sided external faces where A' had fewer vertices than A and such that any graph B bounded by a 6-circuit and having at least one coloring of the 6-circuit in common with A' also had a coloring in common with A (this uses a Kempe chain argument). Thus no minimal counterexample could contain the graph A, or else one could replace A by A' and argue the existence of a coloring by induction. Birkhoff's graph A and its "reducer" A' are shown in Fig. 1. The graph A is known as the "Birkhoff diamond", and is the first "reducible configuration" discovered (it is also the smallest).

To correctly apply Birkhoff's argument one must show that the smaller graph with A replaced by A' is loopless. One can do this using short circuits. However, a different argument which does not use A' avoids the technicality entirely. Indeed, one can show that any graph B bounded by a 6-circuit must have a coloring of the 6-circuit in common with A just by a more systematic application of Kempe chain arguments (it is unknown whether Birkhoff knew this). The difference between the two types of arguments is precisely the difference between C- and D-reducibility, with the latter argument being D-reducibility. When D-reducibility fails one must resort to C-reducibility, and endure the looplessness check on the reducer.

---

[2]Birkhoff carried out his analysis to include every graph with 3 or fewer internal vertices (more precisely: any planar graph with an external 6-sided face and at most 3 internal vertices). Extending the analysis to include graphs with more than three internal vertices seemed unlikely to change the result, as the number of colorings that extend to the 6-circuit only tends to grow with the number of internal vertices.

To go further we must systematize our understanding of Kempe chain arguments. The key notion is that of a *consistent set* of colorings of a circuit. Broadly put, a Kempe chain argument shows that if a coloring $\mathtt{C}$ of a circuit extends to a graph G bounded by the circuit then there are sets of colorings $\mathcal{C}_1$, …, $\mathcal{C}_k$ of the circuit, independent of G, such that all the colorings from at least one of the $\mathcal{C}_i$'s extend to G. If we list the elements of $\mathcal{C}_i$ as $\mathtt{C}_{i,1}, \ldots, \mathtt{C}_{i,n_i}$, we can encode this observation as an implication

$$\mathtt{C} \implies (\mathtt{C}_{1,1} \wedge \cdots \wedge \mathtt{C}_{i,n_1}) \vee \cdots \vee (\mathtt{C}_{k,1} \wedge \cdots \wedge \mathtt{C}_{k,n_k}). \tag{1}$$

A set of colorings $\mathcal{C}$ of a circuit is then called "consistent" if every such implication is satisfied by setting colorings in $\mathcal{C}$ to "true" and colorings not in $\mathcal{C}$ to "false".

Let R be a circuit graph. A *signed path arrangement* for R is a partition of $V(R)$ into subsets V_1, …, V_n such that V_i is a subset of a connected component of $R \backslash V_j$ for all $i \neq j$, together with a function $s : \{V_1^1, \ldots, V_n^{m_n}\} \to \{0,1\}$ where $V_i^1, \ldots, V_i^{m_i}$ are the connected components induced by V_i in R. A *color partition* is a partition of the set $\{\mathtt{r}, \mathtt{g}, \mathtt{b}, \mathtt{y}\}$ into two sets of two elements each. If $\mathtt{C}$ is a coloring of R, $P = (V_1, \ldots, V_n, s)$ is a signed path arrangement for R, and $\theta = \{\{\mathtt{x}, \mathtt{y}\}, \{\mathtt{w}, \mathtt{z}\}\}$ is a color partition, then we say that $\mathtt{C}$ (resp. P) θ-fits P (resp. $\mathtt{C}$) if (i) the set of maximally connected subgraphs of R colored only with $\mathtt{x}$ and $\mathtt{y}$ or only with $\mathtt{w}$ and $\mathtt{z}$ is equal to the set $\{V_1^1, \ldots, V_n^{m_n}\}$ and (ii) $s(V_i^{j_1}) = s(V_i^{j_2})$ if and only if the first color of $V_i^{j_1}$ is equal to the first color $V_i^{j_2}$ as encountered when traveling clockwise around R. We simply say that "$\mathtt{C}$ fits P" if $\mathtt{C}$ θ-fits P for some θ. A set of colorings $\mathcal{C}$ of R is *consistent* if for every $\mathtt{C} \in \mathcal{C}$ and every color partition θ there is a signed path arrangement P that θ-fits $\mathtt{C}$ such that every coloring of R that fits P is in $\mathcal{C}$.

It is easy to see that the set of colorings induced by a planar graph G on a circuit R bounding a region of G is consistent. To check this, take a coloring $\mathtt{C}$ of R that extends to G and a color partition $\theta = \{\{\mathtt{x}, \mathtt{y}\}, \{\mathtt{w}, \mathtt{z}\}\}$. Fix a 4-coloring of G equal to $\mathtt{C}$ on R and declare vertices u, v of R to be in the same set V_i if and only if there is an $\mathtt{xy}$-path or a $\mathtt{wz}$-path from u to v in G (in particular, u and v must both be colored $\mathtt{x}$ or $\mathtt{y}$ or must both be colored $\mathtt{w}$ or $\mathtt{z}$). Say that n different sets $V_1, \ldots, V_n$ are obtained this way. Because G is planar, each V_i is a subset of the same component of $R \backslash V_j$ for all $j \neq i$. Define s such that $\mathtt{C}$ θ-fits $P = (V_1, \ldots, V_n, s)$, in the obvious way (there is more than one s that works; any will do). We need to check that every coloring of R that fits P extends to G. Let $H_1, \ldots, H_n$ be the $\mathtt{xy}$- or $\mathtt{wz}$-components of G containing $V_1, \ldots, V_n$ respectively. Making

a color interchange (switching $\mathtt{x}$ and $\mathtt{y}$, or $\mathtt{w}$ and $\mathtt{z}$) in any H_i results in a valid coloring of G whose restriction to the circuit R again θ-fits P, and such that the $\mathtt{xy}$- and $\mathtt{wz}$-components of G containing $V_1, \ldots, V_n$ are still $H_1, \ldots, H_n$. By repeating the process one can thus make any combination of color interchanges to the H_i's, and it is easy to see that the colorings of R generated this way and by permutations of $\{\mathtt{x}, \mathtt{y}, \mathtt{z}, \mathtt{w}\}$ are all the colorings that fit P.

It is useful to keep in mind the order of the quantifiers that define whether a set $\mathcal{C}$ is consistent: for *every* coloring $\mathtt{C} \in \mathcal{C}$ and *every* color partition θ there is *some* signed path arrangement P that θ-fits $\mathtt{C}$ such that *every* coloring that fits P is in $\mathcal{C}$. Thus for every coloring $\mathtt{C}$ there are three implications like (1): one for each color partition. The disjunction is taken over all signed path arrangements that fit $\mathtt{C}$, with each implicant being the conjunction of the colorings that fit the path arrangement.

In practice one most often wants to know whether a given set $\mathcal{C}_1$ of colorings of a k-circuit has a nonempty consistent subset or not, and, if so, what is the biggest consistent subset (because the union of two consistent sets is consistent, there is a unique maximal consistent subset). To do this one first computes the set $\mathcal{P}_1$ of signed path arrangements such that $P \in \mathcal{P}_1$ if and only if every coloring that fits P is in $\mathcal{C}_1$. Then one computes the set $\mathcal{C}_2$ of colorings consisting of all $\mathtt{C} \in \mathcal{C}_1$ that θ-fit a signed path arrangement in $\mathcal{P}_1$ for each color partition θ. Then one constructs $\mathcal{P}_2$ from $\mathcal{C}_2$ like $\mathcal{P}_1$ is constructed from $\mathcal{C}_1$, and so forth until $\mathcal{C}_i = \mathcal{C}_{i-1}$ or until $\mathcal{C}_i = \emptyset$. (This shows the idea, but does not describe the reducibility program we actually used, which is the one of Robertson et al. Indeed, Robertson et al. do all their reducibility computations using Tait's dual formulation of vertex 4-colorability in terms of edge 3-colorability. Readers should consult their paper and the note accompanying their program [16] for further details.)

For example if $\mathcal{C}$ is the set of 4-colorings that the Birkhoff diamond (the graph on the left of Fig. 1) induces on its 6-circuit and $\mathcal{C}^*$ is the set of all colorings of the 6-circuit, then the above algorithm shows that $\mathcal{C}^* \backslash \mathcal{C}$ has no nonempty consistent subset. Thus any graph obtained by gluing the Birkhoff diamond to a 6-circuit bounding the face of a planar graph G cannot be a minimal counterexample, because G will be 4-colorable by induction and the set of colorings that G induces on the 6-sided face, being consistent, must include some element of $\mathcal{C}$, leading to a coloring of the entire graph.

Not every face of a connected planar graph G is bounded by a circuit, given that G could have cut-vertices or bridges (edges with the same face on either side), but it is useful to glue graphs bounded by circuits to those faces as well. To glue a circuit R to an arbitrary face f one simply traces a

walk around the boundary of f in the natural way. The length of R must be equal to the number of edges plus the number of bridges incident with f. The walk defines a map $\phi : V(R) \to V(G)$. We say that ϕ *wraps* R *around* f. Every coloring of G induces a coloring of R by setting the color of $v \in V(R)$ to the color of $\phi(v) \in V(G)$. We call this coloring of R the *lift* of the coloring by ϕ. The set of 4-colorings of R that are obtained in this way is consistent, as can easily be seen from the same argument as above. In conclusion, we have:

Lemma 3.1. *Let G be a planar graph, f be a face of G, and let ϕ map a circuit R around f. Let $\mathcal{C}$ be the set of lifts by ϕ of 4-colorings of G. Then $\mathcal{C}$ is consistent.*

We have not yet given a precise meaning to the term "configuration". A *near-triangulation* is a connected planar graph in which all regions are triangles except possibly the infinite region. Following Robertson et al., a *configuration K* is a near-triangulation $G(K)$ and a map $\gamma_K : V(G(K)) \to \mathbb{N}$ with the following properties:

(i) for every vertex v, $G(K) \backslash v$ has at most two components, and if there are two then $\gamma_K(v) = d(v) + 2$ where $d(v)$ is the degree of v in $G(K)$,

(ii) for every vertex v, if v is not incident with the infinite region, then $\gamma_K(v) = d(v)$, and otherwise $\gamma_K(v) > d(v)$; and in either case $\gamma_K(v) \geq 5$,

(iii) K has ring-size ≥ 2, where the *ring-size* of K is defined to be $\sum_v (\gamma_K(v) - d(v) - 1)$, summed over all vertices v incident with the infinite region such that $G(K) \backslash v$ is connected.

A configuration is meant to be embedded in a bigger graph, and the meaning of the various components of the definition is fairly transparent: γ_K specifies the degree of a vertex in the larger graph; the ring-size is the arity of the face left when K is deleted from the larger graph; a configuration may have one or several cut-vertices, and these must be adjacent to exactly two vertices outside the configuration. The embedding notion is made precise below.

For the Birkhoff diamond, $G(K)$ is the graph on four vertices consisting of two adjacent triangles and γ_K is identically 5. To draw a configuration one usually uses vertex shapes (an idea of Heesch) to show the values of γ_K, which is possible because in practice $\gamma_K \leq 11$. The list of vertex shapes is shown in Fig. 2. Some configurations are shown in Figs. 3, 5, 6 and 14. The first configuration of Fig. 6 is the Birkhoff diamond[3].

---

[3] We have been referring to the graph on the top left of Fig. 1 as the "Birkhoff diamond" but this is somewhat unorthodox; the first configuration of Fig. 6 is the Birkhoff diamond, and the graph of Fig. 1 is the *free completion* of the Birkhoff diamond.

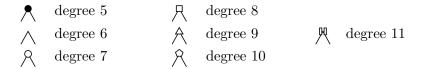|	degree 5		degree 8		
	degree 6		degree 9		degree 11
	degree 7		degree 10		

Figure 2: Shapes for showing the degree of a vertex.

The 2822 configurations of our unavoidable set require over 30 pages to draw at a reasonable size. Instead of appending them as a figure to the paper, which would be of limited use, they are available online in both machine- and human-readable forms. For the rest of the paper we let $\mathcal{U}_{2822}$ denote the set of configurations in the file U_2822.conf and drawn in U_2822.pdf. The format of U_2822.conf is explained in the 'README' file.

Let K be a configuration. A near-triangulation S is a *free completion of K with ring R* if:

(i) R is a circuit of length ≥ 2 of S and bounds the infinite region of S,

(ii) $G(K) = S \backslash V(R)$ and $G(K)$ is an induced subgraph of S and

(iii) every vertex of S not in $V(R)$ has degree $\gamma_K(v)$ in S.

One can easily check that $|V(R)|$ is the ring-size of K as defined in point (iii) of the definition of a configuration. The condition that a cut-vertex v of a configuration have $\gamma_K(v) = d(v) + 2$ ensures that the free completion of a configuration is unique (up to planar homeomorphism), so we speak of "the" free completion of a configuration. Checking the existence of a free completion from the definition of a configuration is fairly straightforward.

Definition 3.2. *Let K be a configuration and let S be the free completion of K with ring R. Then K is D-reducible if the maximal consistent subset of $\mathcal{C}^* \backslash \mathcal{C}$ is empty, where $\mathcal{C}^*$ is the set of all 4-colorings of R and $\mathcal{C} \subseteq \mathcal{C}^*$ is the set of colorings of R that extend to S.*

Let G, H be two planar graphs. We say that G and H are *spherically homeomorphic* if the point-at-infinity completions of G and H are homeomorphic (otherwise put: the stereographic projections of G and H onto a sphere are homeomorphic subsets of the sphere).

Definition 3.3. *A configuration K appears in a triangulation T if $G(K)$ is spherically homeomorphic to an induced subgraph of T such that the image of every vertex $v \in V(G(K))$ has degree $\gamma_K(v)$ in T.*

If we had embedded all our graphs on spheres from the start then Definition 3.3 would seem more natural, but in that case we could not have referred to the "infinite region" so easily, as was practical for the definition of configurations and free completions.

If a configuration K appears in a triangulation T the finite faces of K are not necessarily mapped to faces of T by the (not necessarily unique) spherical homeomorphism, but this is will be the case if T is internally 6-connected (in particular, has no short-circuits of length 3) and $G(K)$ is not a triangle.

If a configuration K appears in an internally 6-connected triangulation T then the free completion S of K is not necessarily a subgraph of T: some vertices from the ring could be identified in T. However it is easy to check that removing the image of $G(K)$ from T leaves a face f whose arity is the ring-size of K, and that there is a map ϕ wrapping the ring R of the free completion S around f such that a vertex $v \in V(G(K))$ is adjacent to a vertex w from R if and only if the image of v in T is adjacent to $\phi(w)$. Let T' be the graph obtained by removing the image of $G(K)$ from T. Then a 4-coloring of T' extends to T if and only if the coloring induced on R by the T'-coloring extends to S. If T is a minimal counterexample then T' is 4-colorable and the set of colorings induced on R by colorings of T' is consistent by Lemma 3.1. Since none of the colorings induced on R should extend to S or else T would be 4-colorable, K cannot be D-reducible. In other words:

Lemma 3.4. *A D-reducible configuration cannot appear in a minimal counterexample.*

Thus if one can exhibit a set $\mathcal{U}$ of D-reducible configurations such that some member of $\mathcal{U}$ appears in every internally 6-connected triangulation, minimal counterexamples cannot exist. This is the property of our set $\mathcal{U}_{2822}$. Of course one must not forget to show that every configuration in $\mathcal{U}_{2822}$ is actually D-reducible, but this must be done by computer. The reader interested in running the computation should consult the README file. Verifying all configurations takes about 10 hours on a modern processor.

Generally speaking, configurations that have many internal vertices compared to the ring-size are more likely to be reducible because they accept more colorings. For example one can show that configurations having vertices adjacent to more than three ring vertices are either not D-reducible or else are not minimally reducible: some subportion of the configuration is D-reducible. Thus all vertices of our configurations have $\gamma_K(v) \leq d(v) + 3$. This result was conjectured by Heesch [12] and proved by Tutte and Whitney
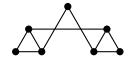
15

Figure 3: Configurations with blocks of size 2 and 3. The first and third configurations have hanging 5-5 pairs.

[19] (ideas for the proof are already implicit in Shimamoto [11]). Heesch also conjectured that minimally reducible configurations could not have "hanging 5-5 pairs", which are two adjacent vertices v, w of degree 2 in $G(K)$ with $\gamma_K(v) = \gamma_K(w) = 5$ (Fig. 3). One can prove this for D-reducible configurations and show more generally that minimal D-reducible configurations cannot have blocks of size 3 or less (Fig. 3) or cut-vertices adjacent to more than two ring vertices (such cut-vertices were excluded by our definition of a configuration for reasons of simplicity). These irreducibility results were obtained by several researchers independently in the early 1970's, see Stromquist [18] for an account. Conjecturally, these conditions are all necessary for C-reducibility as well.

It is not necessary to understand C-reducibility in order to understand our proof, but it seems fitting we give some description of it since part of the interest of our work lies in the difference between D- and C-reducibility. Let K be a configuration with free completion S and ring R. Let $\mathcal{C}$ be the set of colorings of R that extend to S, and let $\mathcal{C}^*$ be the set of all colorings of R. We say that K is C-reducible if the maximal consistent subset of $\mathcal{C}'$ of $\mathcal{C}^* \backslash \mathcal{C}$ is nonempty (meaning K is not D-reducible) and there exists a graph S' with fewer vertices than S whose external face has arity equal to the ring-size of K and such that the set of colorings $\mathcal{C}''$ of the external face of S' that extend to S' does not intersect $\mathcal{C}'$ (one makes the set of colorings of the external face of S' correspond to a set of colorings $\mathcal{C}''$ of R by wrapping R around the external face of S').

When using a C-reducible configuration K to show that a graph G where K appears is not a minimal counterexample one replaces S (embedded in G) with the graph S'. If the resulting graph is loopless one can color it by induction, which shows some coloring not in $\mathcal{C}'$ extends to the graph A obtained by removing $G(S)$ from G. But the set of colorings which extends to A is consistent, and being nonempty and not a subset of $\mathcal{C}'$ must intersect $\mathcal{C}$, so a coloring of G can be recovered. Loops can appear in fairly pernicious ways, given that vertices on the boundary of S may have non-cyclic

16

adjacencies in G, or may even be identified in G (and in S'). The difficulty of checking that the smaller graph is loopless regardless of G is proportional to how different S' is from S. Robertson et al. keep their looplessness check fairly simple by contracting at most four edges of S to make S'. Appel and Haken used arbitrary constructions for S', and hence had more to worry about.

4 Discharging

As we have stated, configurations tend to be reducible when the ring-size is small compared to the number of vertices of the configuration. This is favored by having vertices of small degree (quite intuitively, vertices of small degree "defend better" against coloring). In particular, configurations with many vertices of degree 5 and 6 are likely to be reducible. It was already observed by Kempe [13] that every triangulation must contain a vertex of degree 5 or less, and the number of such vertices augments rapidly if a triangulation contains many vertices of degree greater than 6. Hence the hope that any triangulation has a reducible configuration somewhere.

Concerning the appearance of vertices of degree 5, one can more exactly show that

$$\sum_{v \in V(T)} (6 - d(v)) = 12 \tag{2}$$

for any triangulation T where $d(v)$ is the degree of v. This is easily obtained from Euler's formula and from the formulas $2E = 3F$ and $\sum_v d(v) = 2E$. Every triangulation of degree ≥ 5 must thus have at least 12 vertices of degree 5, more if the triangulation contains vertices of degree ≥ 7.

One way to interpret (2) is to place a "charge" of $6 - d(v)$ on every vertex, so that the charges sum to 12. Then the presence of positive charge is somewhat correlated to the appearance of reducible configurations, but not exactly. The process of discharging consists in "recalibrating" the charge distribution via a series of rules that move the charge between the vertices so that, after the recalibration, the presence of positive charge on a vertex exactly coincides with the appearance of a reducible configuration. Since the total charge remains positive during the recalibration, some vertex must have positive charge and some reducible configuration must appear somewhere.

To find a discharging procedure one starts with an incomplete set of rules and one identifies "overcharging situations" where vertices have positive charge but have no reducible configurations in their neighborhoods. One
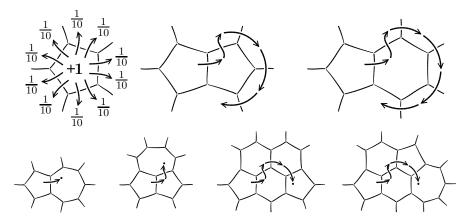
Figure 4: A sketch of Mayer's discharging procedure, using map form. Top: a pentagon divides its charge of 1 into 10 packets and sends two packets to each neighbor. A packet settles if the neighbor is a major vertex, otherwise travels clockwise or counterclockwise around the neighbor until it reaches a major vertex or until a reducible configuration appears. Bottom: sample packet trajectories. In the third drawing, the charge stops because of the appearance of a reducible configuration (the fifth configuration of Fig. 6). In the fourth drawing, the charge would stop before the heptagon if Bernhart's diamond (Fig. 5) were an allowed configuration. If no major vertex appears one of the configurations of Fig. 6 necessarily appears.

then simply adds new discharging rules to move the charge away from these vertices in the hope of obtaining a more exact calibration. The art lies in choosing rules that improve rather than worsen the situation and in choosing rules that help several overcharging situations simultaneously, if possible. The process stops when there are no overcharging situations left. There is no theoretical guarantee that the process will stop, of course, otherwise one would not need to produce a set of discharging rules in the first place.

It is important to get the discharging procedure started off on the right foot. Our discharging rules are inspired from those of Robertson et al., who

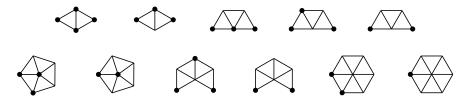Figure 5: The Bernhart diamond, a C-reducible configuration.

Figure 6: The *D*-reducible configurations used for Mayer's discharging procedure.

themselves start off with a discharging procedure of Mayer [14]. Let *minor vertices* be vertices that start with a charge ≥ 0 (vertices of degree 5, 6) and *major vertices* be vertices that start with a negative charge (vertices of degree ≥ 7). The rules of Mayer, shown in Fig. 4, remove the charge from vertices of degree 5 and place it on major vertices. After Mayer's discharging rules are applied minor vertices have zero charge unless they are in the neighborhood of a reducible configuration. The correctness of Mayer's rules depends on the configurations of Fig. 6. In the case when *C*-reducible configurations are allowed the configuration of Fig. 5 ("Bernhart's diamond") also helps; with this configuration, the charge may stop traveling before encountering a major vertex that would otherwise receive charge. Mayer's discharging rules divide the charge into fractions of 1/10, which remains the unit of exchange for the remainder of the discharging procedure.

After Mayer's discharging rules are applied the overcharged vertices are those of degree 7 and 8 (it is generally hard to overcharge vertices of degree $\geq$ 9). These overcharged vertices must then negotiate among themselves. Thus the discharging procedure is two-staged: in the first stage charge is removed from minor vertices and in the second stage charge is redistributed among major vertices (the distinction between the two stages is psychological; there is no notion of sequentiality in the application of discharging rules). The fact that a discharging rule may have both a source and a sink of degree 7 (say) allows a few simple rules to combine among themselves to achieve an overall complex effect. By contrast, Appel and Haken use a diffferent discharging paradigm in which discharging rules all have sources of degree 5 and sinks of degree ≥ 7. In this case the method for obtaining the discharging procedure is a bit different: when an overcharged vertex is identified, one of the vertices of degree 5 responsible for overcharging the major vertex looks for a different major vertex to pass some of its charge to. Rules cannot combine, which results in a much larger number of rules. Appel and Haken have 487 rules, compared to 32 for Robertson et al. (we have 42). Allaire [1] also used a

two-stage discharging procedure built on Mayer's rules in his unpublished proof, but it seems he never completed the details.

In our case we formally define a *discharging rule* L to be a tuple $(q, G(L), s, t, \gamma_L^-, \gamma_L^+)$ where $q \in \mathbb{Q}$, $G(L)$ is a near-triangulation with no cut-vertices, s and t are adjacent vertices of $G(L)$ each at distance at most two from every vertex in $G(L)$, and γ_L^-, γ_L^+ are functions from $V(G(L))$ to the set $\{5, 6, 7, \ldots\} \cup \{\infty\}$ such that $\gamma_L^- = \gamma_L^+(v) = d(v)$ if v is an internal vertex of $G(L)$ and $d(v) < \gamma_L^+(v) \geq \gamma_L^-(v) \geq d(v)$ if v is not an internal vertex of $G(L)$, where $d(v)$ is the degree of v in $G(L)$. To draw discharging rules we use the same vertex shapes as for configurations. For a vertex v with $\gamma_L^+(v) = \infty$ we put the shape for a vertex of degree $\gamma_L^-(v)$ with a '+' next to the vertex. If $\infty > \gamma_L^+(v) > \gamma_L^-(v) = 5$ we put the shape for a vertex of degree $\gamma_L^+(v)$ with a '−' next to the vertex. Otherwise $\gamma_L^-(v) = \gamma_L^+(v)$, except for rule 13 of Fig. 8 that has an exceptional vertex with $(\gamma_L^-(v), \gamma_L^+(v)) = (6, 7)$, for which we use an empty circle[4]. All our rules have $q \in \{\frac{1}{10}, \frac{1}{5}\}$. To show the placement of s, t one puts an arrow on the edge from s from t (in that direction), with a double arrow when $q = \frac{1}{5}$ and a single arrow otherwise. See Figs. 7, 8.

Let T be a triangulation and H an induced subgraph of T. We say that a rule L *appears with image* H, *source* u *and sink* w *in* T if there is a spherical homeomorphism from $G(L)$ to H taking s to u and t to w such that the image under the homeomorphism of each vertex v in $G(L)$ has degree at least $\gamma_L^-(v)$ and at most $\gamma_L^+(v)$. If u, w are adjacent in T the *total charge carried from* u *to* w *by* L is $q(L)$ times the number of distinct images H in T such that L appears with image H, source u and sink w in T (there can be more than one image H because of reflections of the rule; but if $G(L)$ has mirror symmetry and T is internally 6-connected there is at most one image). We put $r(u, w, L)$ for the total charge carried from u to w by L. We put $r_{\mathcal{L}}(u, w) = \sum_{L \in \mathcal{L}} r(u, w, L)$ for any finite set $\mathcal{L}$ of discharging rules.

For a triangulation T, a set of rules $\mathcal{L}$ and $u \in V(T)$ we define the *charge* $c_{\mathcal{L}}(u)$ *of* u *with respect to* $\mathcal{L}$ to be

$$c_{\mathcal{L}}(u) = 6 - d(v) - \sum_w r_{\mathcal{L}}(u, w) + \sum_w r_{\mathcal{L}}(w, u) \tag{3}$$

where the two sums are taken over all vertices adjacent to u. It follows

---

[4]Rules 14 and 23 of Robertson et al. could also be combined by using this representation. Thus a fair comparison of the two sets of rules would either state the number of rules of Robertson et al. as 31 or our number of rules as 43. However we did not want to waste figure space for our set, and did not want to create confusion by quoting the number of rules of Robertson et al. as 31 when it is quoted as 32 elsewhere.
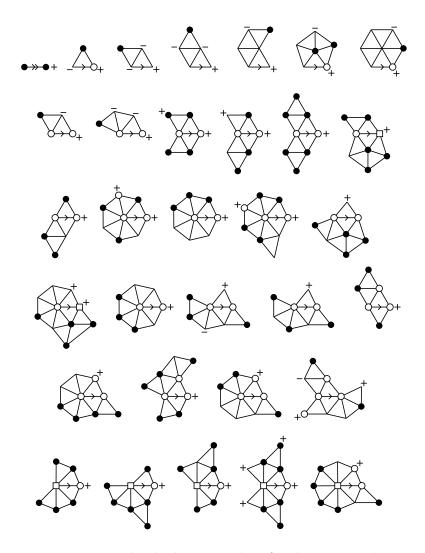
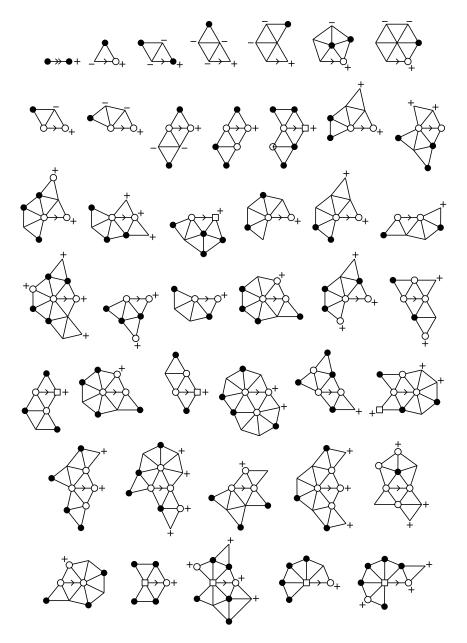Figure 7: The discharging rules of Roberston et al.

Figure 8: Our set of discharging rules, $\mathcal{L}_{42}$.

directly from (3) and (2) that

$$\sum_{u \in V(T)} c_{\mathcal{L}}(u) = 12, \tag{4}$$

since every term $r_{\mathcal{L}}(u, w)$ appears once positively and once negatively in the sum. Thus for any triangulation T and any set of rules $\mathcal{L}$ there is some $u \in V(T)$ such that $c_{\mathcal{L}}(u) > 0$.

Fig. 8 contains our rules. For comparison, Fig. 7 shows the rules of Robertson et al. The first seven rules of each set implement Mayer's discharging procedure; note there are slight differences in the first seven rules between our set and the set of Robertson et al., which are due to the loss of Bernhart's diamond (Fig. 5).

Some of our rules from the first line transfer charge between two vertices even when there is a reducible configuration present; for example the third rule if all the vertices have their smallest degree. This is harmless. One could split such rules into two or more rules of equivalent effect that would avoid such "unnecessary" charge transfers, but the resulting discharging procedure would not be any easier to check and the number of rules would be greater.

Note that $c_{\mathcal{L}}(u)$ only depends on the second neighborhood of u, because every vertex of a rule is at distance at most two from both the source and sink. Later this will greatly simplify the enumeration of cases for which $c_{\mathcal{L}}(u) > 0$, given that second neighborhoods are well-behaved in internally 6-connected triangulations (Lemma 2.4). Another consequence is that reducible configurations must be sought within second neighborhoods, because only the second neighborhood determines the charge of a vertex. Thus all the configurations of $\mathcal{U}_{2822}$ are contained in the second neighborhood of a vertex, as is the case for the unavoidable set of Robertson et al. The existence of such a set was originally conjectured by Heesch [12].

To show the correctness of our set of rules we must show that if u is a vertex of an internally 6-connected triangulation with $c_{\mathcal{L}_{42}}(u) > 0$ then the second neighborhood of u contains a configuration from $\mathcal{U}_{2822}$. The proof is divided into two main parts: the cases when $5 \leq d(u) \leq 11$ and the case when $d(u) \geq 12$. For the first case, separate machine-readable proofs are written for each degree $d(u)$; these are essentially exhaustive analyses of the possible second neighborhoods of u. In the second case one shows that $r_{\mathcal{L}_{42}}(v, u) \leq \frac{1}{2}$ for any neighbor v of u as long as no reducible configurations appear around u. Then $c_{\mathcal{L}_{42}}(u) \leq 6 - d(u) + \frac{1}{2}d(u) \leq 6 - 12 + \frac{1}{2}12 = 0$, so u is not overcharged.

While this gives the idea some precise definitions are missing. For example we have not defined what it means for a second neighborhood to

"contain" a reducible configuration, even though this may seem intuitively obvious. In fact a second neighborhood, as defined, is an induced subgraph of the triangulation and carries no information about the degrees of its outer vertices, so it would be meaningless to define configuration appearance only with respect to this subgraph. What we need is a data structure that carries degree information about all vertices in the second neighborhood, whence the upcoming definition.

A *cartwheel* W is a near-triangulation $G(W)$ with a function γ_W : $V(G(W)) \to \{5, 6, 7, \ldots\}$ such that W contains two disjoint induced circuits C_1 and C_2 and a vertex u not on C_1 or C_2, such that

   (i) $V(G(W)) = \{u\} \cup V(C_1) \cup V(C_2)$,

   (ii) u is adjacent to all vertices of C_1 and to no vertices of C_2,

   (iii) the set of vertices adjacent to the infinite region is $V(C_2)$, and

   (iv) $\gamma_W(v) = d(v)$ for every internal vertex v of $G(W)$, where $d(v)$ is the degree of v in $G(W)$.

One can check that u, C_1 and C_2 are uniquely determined for any cartwheel W. We say u is the *hub* of W.

Let W be a cartwheel and let u, w be adjacent vertices of $G(W)$, one of which is the hub of W. We say that a rule L *appears with image H, source u and sink w in W* if there is a (planar) homeomorphism from $G(L)$ to H taking s to u and t to w such that $\gamma_L^-(v) \le \gamma_W(v') \le \gamma_L^+(v)$ for every $v \in V(G(L))$, where v' is the image of v under the homeomorphism. We then define the charge $o_{\mathcal{L}}(u, w)$ carried from u to w by a set of rules $\mathcal{L}$ analogously as $r_{\mathcal{L}}(u, w)$ is defined for triangulations. The *charge $c_{\mathcal{L}}(W)$ of W with respect to $\mathcal{L}$* is

$$c_{\mathcal{L}}(W) = 6 - d(u) - \sum_w o_{\mathcal{L}}(u, w) + \sum_w o_{\mathcal{L}}(w, u)$$

where u is the hub of W and the sums are taken over all vertices adjacent to u.

If T is a triangulation and $v \in V(T)$ a cartwheel W *appears in T with hub v* if there is a spherical homeomorphism taking $G(W)$ to the second neighborhood of v, such that the hub of W is mapped to v and such that each vertex $w \in V(G(W))$ is mapped to a vertex with degree $\gamma_W(w)$ in T. It is a short step from Lemma 2.4 to see that for every vertex v of an internally 6-connected triangulation T there is a cartwheel $W = W(v)$ appearing[5] in

---

[5]This is where the proof necessitates spherical homeomorphisms; this statement would be false if cartwheel appearance were defined using planar homeomorphisms.

T with hub v. The cartwheel W is unique up to homeomorphism and we speak of "the" cartwheel appearing with hub v.

A configuration K *appears* in a cartwheel W (or that W *contains* K) if there is a planar homeomorphism from $G(K)$ to an induced subgraph of $G(W)$ such that $\gamma_L(v) = \gamma_W(v')$ for every vertex $v \in V(G(K))$, where v' is the image of v in W.

The following lemmas relate what happens in a cartwheel (charge-wise and configuration-wise) to what happens in an internally 6-connected triangulation containing the cartwheel. The omitted proofs are straightforward.

Lemma 4.1. *Let T be an internally 6-connected triangulation and $\mathcal{L}$ a set of rules. Then $c_{\mathcal{L}}(v) = c_{\mathcal{L}}(W(v))$ for every vertex v of T.*

Lemma 4.2. *If a configuration K appears in a cartwheel W and W appears in a triangulation T then K appears in T.*

By Lemma 4.1 and (4) we have

$$\sum_{v \in V(T)} c_{\mathcal{L}}(W(v)) = 12 \tag{5}$$

for any internally 6-connected triangulation T and any set of rules $\mathcal{L}$. Thus the following lemma will prove the four-color theorem:

Lemma 4.3. *An element of $\mathcal{U}_{2822}$ appears in every cartwheel W such that $c_{\mathcal{L}_{42}}(W) > 0$.*

Indeed, any internally 6-connected triangulation T must have a vertex v such that $c_{\mathcal{L}_{42}}(W(v)) > 0$ by (5) so a configuration from $\mathcal{U}_{2822}$ appears in $W(v)$ by Lemma 4.3 and in T by Lemma 4.2. Then T cannot be a minimal counterexample by Lemma 3.4. But every minimal counterexample is an internally 6-connected triangulation by Lemma 2.6 so minimal counterexamples do not exist, QED.

As explained, we will divide the proof of Lemma 4.3 according to the degree of the hub of W. We first describe the proof for the case when W has a hub of degree ≤ 11. In this case the proof is given by machine-readable scripts called "presentation files" by Robertson et al. There is one presentation file for each hub degree from 5 to 11. Here we can only describe the format of a presentation file; "doing" the proof means running the computation. For cartwheels of hub degree 5 and 6 one can also prove Lemma 4.3 by hand, as do Robertson et al., but we will not take the time to do this. (However readers should easily be able to convince themselves

from Fig. 4, given that the configurations of Fig. 6 are in $\mathcal{U}_{2822}$.) While describing the files we refer to a generic unavoidable set $\mathcal{U}$ and to a generic set of rules $\mathcal{L}$, these two sets being loaded from files at the beginning of the computation.

The upcoming material explaining presentation files is more formally covered in Robertson et al.'s original article and in the notes accompanying their program [15, 17]. Here we just explain enough to point out some aspects of their implementation that affected our proof in particular ways. For example, certain choices of Robertson et al. limited the type of configuration we could use, and also put restrictions on the set of rules. For reasons of simplicity our terminology does not always match the definitions of Robertson et al. As such our account is just an overview of ideas, not a faithful description of the discharging program.

We give names to the different parts of a cartwheel. Let W be a cartwheel with hub u. The *spokes* of W are the vertices adjacent to u. The *hats* of W are the vertices at distance 2 from u that are adjacent to exactly two spokes. The *fan vertices* of W are the vertices at distance 2 from u that are adjacent to exactly one spoke. It is easy to check that every vertex is either the hub, a spoke, a hat or a fan vertex. The fan vertices adjacent to a spoke v are called the *fan vertices over v*.

Let W be a cartwheel. We say that an induced subgraph H of $G(W)$ is a *part graph of W* if the hub as well as every spoke and hat of W is in H and if all fan vertices of a spoke v of W are in H when one fan vertex of v is in H. We say that a graph H is a *part graph* if it is a part graph of some cartwheel W. The *hub*, *spokes*, *hats* and *fan vertices* of a part graph are defined in the natural way (as for a cartwheel, these are uniquely defined by the part graph).

A *part P* consists of a part graph $G(P)$ and two functions $\gamma_P^-, \gamma_P^+ :$ $V(G(P)) \rightarrow \{5, 6, 7, \ldots\} \cup \{\infty\}$ such that (i) $\gamma_P^-(v) = \gamma_P^+(v) = d(v)$ for every internal vertex v of P where $d(v)$ is the degree of v in P, (ii) every spoke v with $\gamma_P^-(v) = \gamma_P^+(v)$ is an internal vertex of $G(P)$, and (iii) $\infty >$ $\gamma_P^-(v) \leq \gamma_P^+(v)$ for every $v \in V(G(P))$. A cartwheel W *fits a part P* if there is a homeomorphic copy of $G(P)$ in $G(W)$ such that $\gamma_P^-(v) \leq \gamma_W(v') \leq \gamma_P^+(v)$ for every $v \in V(G(P))$ where v' is the image of v in $G(W)$. Note the image of $G(P)$ may not be an induced subgraph of $G(W)$, as two hats may be adjacent in $G(W)$ but not in $G(P)$.

The *trivial part of degree d* is the part P with no fan vertices whose hub has degree d and such that $\gamma_P^-(v) = 5$, $\gamma_P^+(v) = \infty$ for every spoke and hat v. We say that a part P is *dispatched* if every cartwheel W that fits P such that $c_{\mathcal{L}}(W) > 0$ contains a configuration from $\mathcal{U}$. Our goal is thus to dispatch

the trivial part of degree d for $d = 5, 6, \ldots, 11$ since every cartwheel of hub degree d fits the trivial part of degree d. The strategy is, given a part P that we do not know how to dispatch, to refine the part into two complementary parts P' and P'' such that every cartwheel W that fits P either fits P' or P'', and to hope that P' and P'' contain enough information to be dispatched by some direct observation—otherwise one keeps on refining the parts.

Robertson et al. use three tests $\tau_{\mathtt{R}}$, $\tau_{\mathtt{H}}$ and $\tau_{\mathtt{S}}$ to directly dispatch a part P. If the test $\tau_{\mathtt{R}}(P)$ returns true then every cartwheel W that fits P contains an element of $\mathcal{U}$, so P is dispatched. If the test $\tau_{\mathtt{H}}(P)$ returns true then every cartwheel W that fits P has $c_{\mathcal{L}}(W) \leq 0$ unless W contains an element of $\mathcal{U}$, so P is also dispatched. Finally the test $\tau_{\mathtt{S}}(P)$ returns true if a part P is dispatched by reason of symmetry with previously dispatched parts. To dispatch the trivial part of degree d one starts by placing the trivial part onto an empty stack. Then the following steps are repeated until the stack is empty: pop the last part P on the stack if either $\tau_{\mathtt{R}}(P)$ or $\tau_{\mathtt{H}}(P)$ or $\tau_{\mathtt{S}}(P)$, otherwise choose parts P', P'' such that every cartwheel W that fits P fits either P' or P'', replace P with P', add P'' onto the top of the stack, and return to the first step. Clearly if the process terminates then the trivial part of degree d is dispatched. The presentation file instructs the program whether to test for $\tau_{\mathtt{R}}$, $\tau_{\mathtt{H}}$ or $\tau_{\mathtt{S}}$ or else which parts P', P'' to choose. Thus every line in a presentation file is one of four types: an order to test for either $\tau_{\mathtt{R}}$, $\tau_{\mathtt{H}}$ or $\tau_{\mathtt{S}}$, or a "branching order" specifying two complementary parts P', P'' that refine P.

We give a brief overview of the tests $\tau_{\mathtt{R}}$ and $\tau_{\mathtt{H}}$ with special attention to implementation details that place restrictions on $\mathcal{U}$ and $\mathcal{L}$. Say that a part P is *reducible* if a configuration from $\mathcal{U}$ appears in every cartwheel fitting P. Thus if $\tau_{\mathtt{R}}(P)$ returns true P should be reducible. Say that a configuration K *appears* in a part P if there is an induced subgraph H of $G(P)$ homeomorphic to $G(K)$ such that $\gamma_P^-(v') = \gamma_K(v) = \gamma_P^+(v')$ for all vertices $v \in V(K)$, where v' is the image of v in H. Note that K may appear in P without appearing in every cartwheel W fitting P. Indeed, while H is an induced subgraph of $G(P)$, $G(P)$ is not necessarily an induced subgraph of $G(W)$, so the image of H is not necessarily induced in $G(W)$ (recall that two hats may be adjacent in $G(W)$ but not in $G(P)$).

To remedy this Robertson et al. say a subgraph H of $G(P)$ is *well-positioned* if H contains a spoke v whenever it contains the two hats adjacent to v in P. It is easy to see that if a configuration K appears well-positioned in P it appears in every cartwheel fitting P. This relationship is not if-and-only-if: a configuration could appear in every cartwheel fitting a part, but not appear well-positioned in the part itself, or indeed in any part (Fig. 9).
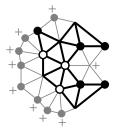
Figure 9: The configuration in bold appears in every cartwheel fitting the drawn part but is never well-positioned. Vertices with $\gamma_P^+(v) = \infty$ are lightened for clarity.

Figure 10: Preparing a spanning tree of faces.

The test τ_R only recognizes well-positioned configurations. While this may seem like a drawback configurations such as the one of Fig. 9 are quite few. An advantage of this definition of appearance, used in the program, is that modifying the values of $\gamma_P^-(v)$ or $\gamma_P^+(v)$ for a vertex v not in the image of the configuration does not affect whether the configuration appears well-positioned or not.

   To check whether a given configuration $K \in \mathcal{U}$ appears in a part P a "spanning subtree of faces" is first prepared for K by removing edges from $G(K)$ until no internal vertices are left (Fig. 10); then the program tries to "unfold" the spanning subtree onto $G(P)$ triangle by triangle, choosing an initial edge and an orientation; the unfolding fails if a vertex is missing or if vertex degrees are incompatible. If the unfolding completes successfully
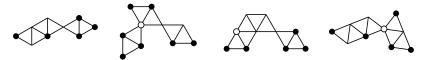
Figure 11: Some D-reducible configurations of radius greater than two that fit in a second neighborhood.

it is not actually clear that K appears in P—the mapping from $G(K)$ to $G(P)$ constructed by the unfolding may not be injective, or its image may not be induced, and so forth. Robertson et al. have a lemma showing that K appears in P if the unfolding completes successfully as long as: (i) K has radius ≤ 2, where the "radius" of a configuration is the smallest r such that there is a vertex in $G(K)$ at distance $\leq r$ from every other vertex in $G(K)$, (ii) P has hub of degree ≥ 6. The program checks (i) at the start of the program when the set $\mathcal{U}$ is loaded. If $\mathcal{U}$ contains a configuration of radius > 2 the program terminates with an error message. The latter behavior forced us to exclude certain D-reducible configurations that could have been interesting to use (Fig. 11). These are configurations which, despite having radius greater than 2, can fit inside a cartwheel by having their image exclude the hub. There are relatively few such configurations, though, and their loss does not seem to significantly impact the size of the unavoidable set.

The discharging program also contains a safety check function which verifies "rather crudely, from first principles" that the mapping produced by the unfolding process is one-to-one, induced and well-positioned. Thus the correctness of the computation does not depend on the lemma; the lemma simply shows that, when conditions (i), (ii) are met and well-positionedness is checked, the mappings produced by the unfolding routine are *theoretically* correct, and should never result in a complaint from the safety check function. The presence of the safety check function means that it is valid to use the program even if the conditions of the lemma aren't met. For example, one could remove the radius check by commenting out a line and use a set $\mathcal{U}$ containing configurations of radius > 2; a successful computation would still indicate a correct discharging procedure. (We didn't do so because configurations of radius > 2 are not all that valuable, and we preferred not to modify the program in any qualitative way.) It is also correct to use the program with cartwheels of hub degree 5, eschewing condition (ii) of the lemma. (Despite the lemma, the program only checks that the hub degree is between 5 and 11.)

The test $\tau_{\mathrm{H}}(P)$ should return true only if every cartwheel W fitting P contains a reducible configuration or else has $c_{\mathcal{L}}(W) \leq 0$. The test depends on being able to upper bound the maximum net amount of charge that can be transferred from two spokes u, v of a part P to its hub in any cartwheel W fitting P that does not contain a configuration $K \in \mathcal{U}$. We note this maximum net amount of charge as $\zeta(P, u, v)$. For every call to τ_{H} a sequence of triplets $S = ((u_1, v_1, q_1), \ldots, (u_k, v_k, q_k))$ is specified in the presentation file, where u_i, v_i are spokes of P and $q_i \in \mathbb{Q}$ such that every spoke appears

Figure 12: A rule that cannot be encoded as a part.

in exactly two triplets. The test $\tau_{\mathtt{H}}$ returns true if $\zeta(P, u_i, v_i) \le q_i$ for $i = 1, \ldots, k$ and if $6 - d + \lfloor (q_1 + \cdots + q_k)/2 \rfloor_{10} \le 0$, where d is the hub degree and $\lfloor \cdot \rfloor_{10}$ denotes rounding down to the nearest fraction[6] of $\frac{1}{10}$ (i.e. $\lfloor x \rfloor_{10} = \lfloor 10x \rfloor / 10$). The sequence S is called a *hubcap* for P. Possibly (and in fact quite often) a hubcap contains triplets where $u_i = v_i$. If $u = v$ then $\zeta(P, u, v)$ represents the maximum net charge transfer frum u to the hub in any cartwheel fitting P, and u still appears in exactly two triplets.

One upper bounds $\zeta(P, u, v)$ by enumerating all the ways incoming rules with source u or v can be placed onto the part. A combination is discarded if it triggers $\tau_{\mathtt{R}}$; if not, the net charge transfer is computed by subtracting the charge of rules induced outward with sink u or v from the charge of rules induced inward with source u or v. Doing this mainly requires a way of "gluing" rules onto a part to produce a new part, in order to check whether $\tau_{\mathtt{R}}$ is triggered. The elegant method used by Robertson et al. is to encode rules themselves as parts[7] and to have a procedure for superposing or "anding" two parts together, which is easy to do. Each rule becomes two parts, one in which the sink becomes the hub and one in which the source becomes the hub (if the degrees of the sink or source are not compatible with the hub degree under consideration, that part is not created). However not all rules can be encoded as parts, given that fan vertices do not exist in a part until the degree of the spoke beneath them has been fixed. For example the rule of Fig. 12 cannot be encoded as a part where its sink coincides with the hub of the part, because the "spoke" that would be attached to the vertex of degree 5 has undetermined degree. Thus such a rule cannot be used with the program of Robertson et al. Once again, however, we did not find this to be particularly bothersome.

We do not discuss the test $\tau_{\mathtt{S}}$, which dispatches a part by reason of symmetry with a previously dispatched part, as the reader should find such a

---

[6]Naturally, the program multiplies all charge values by 10 to make quantities integral; then one rounds down to the nearest integer.
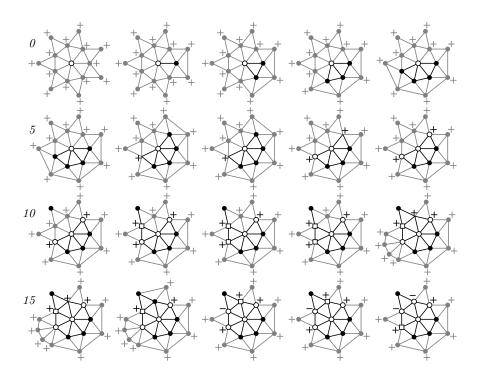
[7]In the program, rules are actually encoded as a more compact data structure called an *outlet*. Outlets are efficient for representing parts that have few vertices v with $(\gamma_P^-(v), \gamma^+(v)) \ne (5, \infty)$.

test fairly intuitive. Fig. 4 illustrates a sample presentation file. To understand the figure some syntax is needed. Each line of the file starts with an 'L' followed by the recursion depth, which is the number of parts on the stack at that point. Lines that specify a branching order, called *condition lines*, start with a 'C'. Lines that indicate a call to τ_R, τ_H or τ_S are called *disposition lines* and start with R, H or S respectively (H is for "hubcap"). A condition line specifies a vertex v of the last part P on the stack such that $\gamma_P^-(v) < \gamma_P^+(v)$ and tells the program to branch either by raising $\gamma_P^-(v)$ to a value $b \le \gamma_P^+(v)$ or by lowering $\gamma_P^+(v)$ to a value $b \ge \gamma_P^-(v)$. Each order has a *complementary order*: in the first case, to lower $\gamma_P^+(v)$ to $b-1$, in the second case, to raise $\gamma_P^-(v)$ to $b+1$. The part P' created by the original order is placed on top of the stack, whereas the part P'' created by the complementary order replaces P. (If v is a spoke, an order which sets $\gamma_P^+(v) = \gamma_P^-(v)$ entails the creation of fan vertices for v, which are set to have lower degree 5 and upper degree ∞.)

The following numbering scheme is used for vertices: numbers $1, \ldots, d$ are the spokes in clockwise order, numbers $d+1, \ldots, 2d$ are the hats in clockwise order such that hat 1 is adjacent to spoke 1 and spoke 2, and finally number $k + (l+1)d$ where $1 \le k \le d$ and $l \ge 1$ is the l-th fan vertex clockwise above the k-th spoke, if the fan vertex exists. Altogether the format of a condition line is "C m n" where m is the number of the vertex v to be affected and $n \in \{\ldots, -6, -5, 6, 7, \ldots\}$ specifies the modification to the degree function. If $n > 0$ the order is to increase $\gamma_P^-(v)$ to n, otherwise the order is to decrease $\gamma_P^+(v)$ to $-n$.

The triplets of numbers in an 'H' line list the hubcap. Identical triplets are only listed once and charge values are multiplied by 10 in order to avoid fractions. When reading a presentation file the hubcap lines are by far the most difficult to check, so difficult that it is debatable to what degree they are really "readable". One can muster some help with the 'verbose' output option of the program, which gives information as to which reducible configurations are appearing along with other helpful details. While the resulting output may be readable at a normal pace it is also quite large: over 3'000'000 lines for the Robertson et al. proof, over 13'000'000 lines for our proof. A mathematician checking these proofs at the rate of one line per second and working 9 hours a day would take over 3 months to read the Robertson et al. proof and over a year to read ours. We refer readers with further interest in presentation files to [17].

The proof of Lemma 4.3 requires a separate argument to show that vertices of degree ≥ 12 are not overcharged. For this it suffices to show that $r_{\mathcal{L}_{42}}(u,v) \le \frac{1}{2}$ for any two adjacent vertices u, v of an internally 6-connected

```
0    Degree 7
1    L0 C 1 -5
2      L1 C 2 -5
3        L2 C 3 -5
4          L3 C 4 -5
5            L4 R
6          L3 C 7 -5
7            L4 R
8          L3 C 4 7
9            L4 C 7 7
10             L5 C 12 -5
11               L6 C 5 8
12                 L7 C 4 8
13                   L8 H (2,2,5) (4,4,-2) (5,5,0) (1,3,7) (6,7,0)
14                 L7 C 6 6
15                   L8 H (1,1,4) (2,2,5) (5,5,0) (6,6,0) (7,7,-1) (3,4,2)
16                 L7 H (1,1,4) (2,2,5) (5,5,0) (6,6,2) (7,7,-3) (3,4,2)
17               L6 C 6 8
18                 L7 S 4 1 6 18
19               L6 C 4 8
```

Figure 13: Illustration of a sample presentation file. Vertices v with $(\gamma^-(v), \gamma^+(v)) = (5, \infty)$ are lightened for clarity.
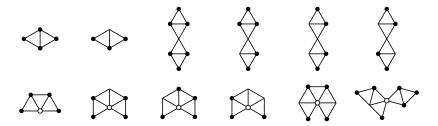
Figure 14: The D-reducible configurations used for the proof of Lemma 4.4.

triangulation T where v has degree ≥ 12 as long as T does not contain an element of $\mathcal{U}_{2822}$, since the initial negative charge of vertices of degree ≥ 12 is at least half their degree in absolute value. This is proved in Lemma 4.4 below. The configurations that are used for the proof appear in Fig. 14 (these configurations all appear on the first page of U_2822.pdf—which is ordered identically to U_2822.conf—making it easy to check they are in $\mathcal{U}_{2822}$). The proof refers to rules by the order which they appear in Fig. 8, from left to right and top to bottom. Serious readers are encouraged to pencil in some numbers.

Lemma 4.4. *Let u, v be adjacent vertices of an internally 6-connected triangulation T, where v has degree at least 8. Then $r_{\mathcal{L}_{42}}(u, v) \leq \frac{1}{2}$ if no members of $\mathcal{U}_{2822}$ appear in T.*

Proof. We say that a rule *appears* in T to mean it appears with source u and sink v. If u has degree ≥ 9 then $r_{\mathcal{L}_{42}}(u, v) = 0$ because there are no rules with source of degree ≥ 9 in $\mathcal{L}_{42}$. Thus we can assume that u has either degree 5, 6, 7 or 8, and we analyze each of these cases separately. Except for the first case rules all have value $\frac{1}{10}$ and the object is to show that no more than 5 rules (multiplicities counted) can appear simultaneously, without having a configuration in $\mathcal{U}_{2822}$ appear.

We proceed with the four cases.

Vertex u has degree 5. In this case only the first four rules of $\mathcal{L}_{42}$ can appear. Each can appear at most twice except for rule 1 which can (and will) appear only once. Say rule 4 appears twice. Then rules 2 and 3 cannot appear at all because of configurations 1 and 2 (Fig.), so $r_{\mathcal{L}_{42}}(u, v) = \frac{2}{5} < \frac{1}{2}$. Otherwise say rule 3 appears twice. Then rule 2 cannot appear because of configuration 1 and rule 4 cannot appear because of configurations 1 and 2

33

again, so $r_{\mathcal{L}_{42}}(u,v) = \frac{2}{5}$ again. Likewise if rule 2 appears twice then rule 4 cannot appear and rule 3 can appear only once, so $r_{\mathcal{L}_{42}}(u,v) \leq \frac{2}{5} + \frac{1}{10} = \frac{1}{2}$. Therefore we can assume rules 2, 3 and 4 appear at most once each, for a possible total of $\frac{1}{5} + \frac{3}{10} = \frac{1}{2}$.

Vertex u has degree 6. In this case the only rules which can appear are rules 2 through 7. Say rule 7 appears twice. Then rules 2, 3 and 6 cannot appear, and rules 4, 5 can only appear twice between the two of them because of configuration 2. Thus the charge transferred is at most $\frac{1}{10} + \frac{1}{10} + \frac{2}{10} < \frac{1}{2}$. Now say rules 2 and 7 both appear. If rules 3, 4, 5 total 4 appearances between them then rule 4 must appear twice and configuration 2 appears, so rules 3, 4, 5 can only total 3 appearances between them, which makes a possible total of 5 rule appearances. If rule 7 appears and rule 2 does not appear then rule 6 does not appear and rules 4, 5 cannot both appear twice because of configuration 2, so the total number of rule appearances is at most 5, since rule 3 can only appear once. Therefore we can assume rule 7 does not appear.

If rule 2 appears twice then rule 4 cannot appear twice because of configurations 3, 4, 6 and rules 3, 6 can only appear twice between the two of them so there are most 5 rule appearances. If rule 2 appears once then rules 3, 5, 6 can each appear only once and if they each appear once then rule 4 cannot appear twice because of configuration 7, so there are at most 5 configuration appearances. On the other hand if rule 2 does not appear then rule 6 does not appear and rules 3, 4, 5 cannot each appear twice because of configuration 2, so total at most 5 rules appearances between them.

Vertex u has degree 7. This is the lengthiest case, as there are the most rules with source of degree 7 and sink of unbounded degree. To ease the analysis we have shown the cases in Fig. 15.

If u has degree 7 then one of the rules in the left column of Fig. 15 appears with source u and sink v in T, as is easy to check (these rules are not part of $\mathcal{L}_{42}$, they are just a device for organizing the case analysis). Next to each of these rules we list the numbers of the rules in $\mathcal{L}_{42}$ that can appear in T when the rule appears; a number is listed twice if a rule can appear twice, in which case one should think of the first occurence of the number as representing the "right side up" appearance of the rule as drawn in Fig. 8, and the second occurence of the number as representing "upside down" occurence of the rule obtained by a horizontal mirror symmetry. If a rule cannot appear without creating the appearance of a configuration from Fig. 14 its number is crossed out. A set of rules that cannot appear simultaneously or that cannot appear
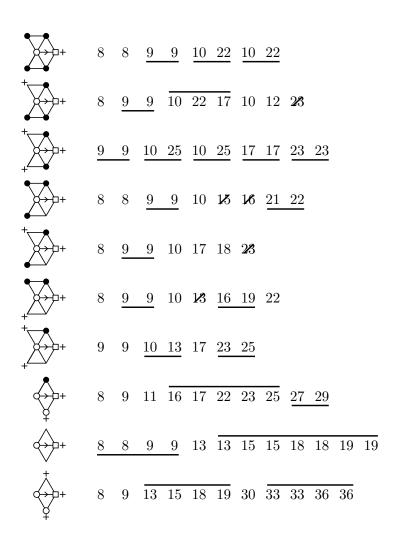
Figure 15: The case analysis showing that a vertex of degree 7 cannot transfer more than 1/2 units of charge to a vertex of degree ≥ 8.

simultaneously without creating a configuration from Fig. 14 is underlined. A set of rules that are pairwise incompatible is overlined.

In all cases it is easy to check that at most 5 rules can appear simultaneously. For example, in the next-to-last case, only three rules from the set $\{8, 8, 9, 9\}$ can appear because that group is underlined and only one rule from the overlined group can appear; this makes a maximum of $3 + 1 + 1 = 5$ rule appearances.

Vertex u has degree 8. There are only 4 rules in $\mathcal{L}_{42}$ with hub degree 8 and unbounded sink degree, rules 39 through 42. Since only rule 43 is compatible with its mirror image (rule 41 has no mirror image) only 5 rules can appear among these.

$\square$

References

[1] F. Allaire, *Another proof of the four colour theorem*-Part I, Proc. of the 7th Manitoba Conf. on Numerical Math. and Computing, (1977), Congressius Numerantum XX, Utilitas Math., Winnipeg, 1978, 3–72.

[2] K. Appel and W. Haken, Every planar map is four-colorable, Part I: Discharging, *Illinois Journal of Mathematics* **21** (1977), 429–490.

[3] K. Appel, W. Haken and J. Koch, Every planar map is four-colorable, Part II: Reducibility, *Illinois Journal of Mathematics* **21** (1977), 491–567.

[4] K. Appel and W. Haken, Every planar map is four colorable, A.M.S. Contemporary Math. **98** (1989).

[5] G.D. Birkhoff, *The reducibility of maps*, Amer. J. Math. **70** (1913), 114–128.

[6] A. Bernhart, *Another reducible edge configuration*, Amer. J. Math. **35** (1948), 144–146.

[7] D.I.A. Cohen, *Block-count reducibility and the four-color theorem*, manuscript.

[8] R. Fritsch and G. Fritsch, *The four-color theorem*, Springer-Verlag (1998).

[9] S.J. Gismondi and E.R. Swart, *A new type of four-colour reducibility*, Congr. Numer. **82** (1991), 33–48.

[10] G. Gonthier, A computer-checked proof of the four-color theorem, http://research.microsoft.com/∼gonthier/fourcolor.pdf.

[11] W. Haken, *Shimamoto's construction*, manuscrpit (1971), available at http://www.math.ubc.ca/∼jpsteinb/index.html.

[12] H. Heesch, *Untersuchungen zum Vierfarbenproblem*, Hochschulskriptum 810/a/b, Bibliographisches Institüt, Mannheim, 1969.

[13] A. B. Kempe, *On the geographical problem of the four colors*, Amer. J. Math. **2** (1879), 193–200.

[14] J. Mayer, *Une propriété des graphes minimaux dans le problème des quatres couleurs*, Problèmes Combinatoires et Théorie des Graphes, Colloques internationaux CNRS No. 260, Paris, 1978.

[15] N. Robertson, D.P. Sanders, P.D. Seymour and R. Thomas, The four-colour theorem, *J. Combin. Theorey Ser. B*, (1997) **70**, 2–42.

[16] _____, *Reducibility in the four colour theorem*, available at http://math.gatech. edu/∼thomas/OLDFTP/four.

[17] _____, *Discharging cartwheels*, available at http://math.gatech. edu/∼thomas/OLDFTP/four.

[18] W.R. Stromquist, Some aspects of the four color problem, Ph.D., Harvard University, 1975. Available at http://www.walterstromquist.com.

[19] H. Whitney and W.T. Tutte, *Kempe chains and the four colour problem*, Studies in Graph Theory (D. R. Fulkerson, eds.), Part II, Math. Assoc. of America, 1975, pp. 378–413.