THE FOUR-COLOUR THEOREM

Neil Robertson[1]
Department of Mathematics
The Ohio State University
231 West 18th Avenue
Columbus, Ohio 43210

Daniel Sanders[2]
School of Mathematics
Georgia Institute of Technology
Atlanta, Georgia 30332

Paul Seymour
Bellcore
445 South Street
Morristown, New Jersey 07960

and

Robin Thomas[3]
School of Mathematics
Georgia Institute of Technology
Atlanta, Georgia 30332

February 1994; revised 16 January 1997

Abstract. The four-colour theorem, that every loopless planar graph admits a vertex-colouring with at most four different colours, was proved in 1976 by Appel and Haken, using a computer. Here we give another proof, still using a computer, but simpler than Appel and Haken's in several respects.

[1] Research partially performed under a consulting agreement with Bellcore, partially supported by DIMACS Center, Rutgers University, New Brunswick, New Jersey 08903, USA, and partially supported by NSF grant no. DMS-8903132 and ONR grant no. N00014-92-J-1965.
[2] Research supported by DIMACS and ONR grant no. N00014-93-1-0325.
[3] Research partially performed under a consulting agreement with Bellcore, partially supported by DIMACS, and partially supported by NSF grant no. DMS-9303761 and ONR grant no. N00014-93-1-0325.

1. Introduction

The four-colour theorem (briefly, the 4CT) asserts that every loopless planar graph admits a vertex 4-colouring. This was conjectured by F. Guthrie in 1852, and remained open until a proof was found by Appel and Haken [3, 4, 5] in 1976.

Unfortunately, the proof by Appel and Haken (briefly, A&H) has not been fully accepted. There has remained a certain amount of doubt about its validity, basically for two reasons:

(i) part of the A&H proof uses a computer, and cannot be verified by hand, and

(ii) even the part of the proof that is supposed to be checked by hand is extraordinarily complicated and tedious, and as far as we know, no one has made a complete independent check of it.

Reason (i) may be a necessary evil, but reason (ii) is more disturbing, particularly since the 4CT has a history of incorrect "proofs". So in 1993, mainly for our own peace of mind, we resolved to convince ourselves somehow that the 4CT really was true. We began by trying to read the A&H proof, but very soon gave this up. To check that the members of their "unavoidable set" were all reducible would require a considerable amount of programming, and *also* would require us to input by hand into the computer descriptions of some 1400 graphs; and this was not even the part of their proof that was most controversial. We decided it would be easier, and more fun, to make up our own proof, using the same general approach as A&H. So we did; it was a year's work, but we were able to convince ourselves that the 4CT is true and provable by this approach. In addition, our proof turned out to be simpler than that of A&H in several respects.

The basic idea of the proof is the same as that of A&H. We exhibit a set of "configurations"; in our case there are 633 of them. We prove that none of these configurations can appear in a minimal counterexample to the 4CT, because if one appeared, it could be replaced by something smaller, to make a smaller counterexample to the 4CT (this is called proving "reducibility"; here we are doing exactly what A&H and several other authors did - for instance, [2, 9]). But every minimal counterexample is an "internally 6-connected triangulation" (defined later), and in the second part of the proof we prove that at least one of the 633 configurations appears in *every* internally 6-connected triangulation. (This is called proving "unavoidability", and uses the method of "discharging" vertices, first suggested by Heesch [11]). Consequently, there is no minimal counterexample, and so the 4CT is true. Where our method differs from A&H is in how we prove unavoidability.

Some aspects of this difference are: we confirm Heesch's conjecture that one can prove unavoidability of some reducible set without looking beyond the second neighbourhoods of "overcharged" vertices; consequently, we avoid the problems with configurations that "wrap around and meet themselves", that were a major source of complications for A&H; our unavoidable set has size about half that of the A&H set; our "discharging procedure" for proving unavoidability (derived from an elegant method of Mayer [13]) only involves 32 discharging rules, instead of the 300+ of A&H; and we obtain a quadratic time algorithm to find a 4-colouring of a

planar graph, instead of the quartic algorithm of A&H.

Our proof is also somewhat easier to check, because we replace the mammoth hand-checking of unavoidability that A&H required, by another mammoth hand-checkable proof, but this time written formally so that, if desired, it can be read and checked by a computer in a few minutes. We are making the necessary programs and data available to the public for checking. (For details see the next section.)

The paper is organized as follows. Section 2 contains some preliminary definitions and observations. In the Appendix we list the 633 configurations, using conventions described in section 2. In section 3 we prove their reducibility, and in section 4 we prove unavoidability; thus, the proof of the 4CT is complete at the end of section 4. Section 5 is a variety of comments on why we did things the way we did, and section 6 contains a quadratic time algorithm to 4-colour a planar graph. In section 7 we give more details of the "machine-checkable proof".

2. The Set of Configurations

Throughout the paper, let Σ be a fixed 2-sphere. If $X \subseteq \Sigma$, its topological closure is denoted by $\overline{X}$. A *line* is a subset of Σ homeomorphic to the closed unit interval, and its *ends* are defined in the natural way. An *open disc* is a subset of Σ homeomorphic to the real plane $\mathbb{R}^2$ and a *closed disc* is a subset of Σ homeomorphic to $\{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$.

A *drawing* G is a pair $(U(G), V(G))$, where $U(G) \subseteq \Sigma$ is closed and $V(G) \subseteq U(G)$ is finite, such that

(i) $U(G) - V(G)$ has only finitely many arc-wise connected components, called *edges*

(ii) for each edge e, $\overline{e}$ is a line and $\overline{e} - e$ consists of the two ends of $\overline{e}$.

Thus, we do not permit loops. The members of $V(G)$ are the *vertices* of G, and the set of edges is denoted by $E(G)$. An edge e is incident with a vertex v (and vice versa) if v is an end of $\overline{e}$. This defines a graph, and we use standard graph-theoretic terminology for drawings without further explanation. The *degree* of a vertex v is the number of edges incident with it, and is denoted by $d(v)$ or $d_G(v)$. A drawing G is a *subdrawing* of a drawing H if $V(G) \subseteq V(H)$ and $E(G) \subseteq E(H)$. A subdrawing G' of G is *induced* if every edge of G with both ends in $V(G')$ belongs to G'.

A *region* of G is an arc-wise connected component of $\Sigma - U(G)$. A region r is *incident* with a vertex v (and vice versa) if $v \in \overline{r}$; and a region r is *incident* with an edge e (and vice versa) if $e \subseteq \overline{r}$. A region is a *triangle* if it is an open disc and is incident with precisely three distinct edges that form a circuit of G. (By definition, a circuit has no repeated vertices or edges.) A drawing is a *triangulation* if every region is a triangle. Thus, a triangulation can have parallel edges, but no circuit of length two bounds a region.

A *minimal counterexample* means a drawing T that is not 4-colourable, such that every drawing T' with $|V(T')| + |E(T')| < |V(T)| + |E(T)|$ is 4-colourable. To prove the 4CT we shall show that there is no minimal counterexample. It is easy

to show that every minimal counterexample is a triangulation, and is 5-connected, and is 6-connected except for vertices of degree 5. More precisely, let us say a *short circuit* of a triangulation is a circuit C with $|E(C)| \leq 5$, so that for both the open discs Δ bounded by $U(C)$, $\Delta \cap V(T) \neq \emptyset$, and $|\Delta \cap V(T)| \geq 2$ if $|E(C)| = 5$. Let us say T is *internally* 6-*connected* if it has no short circuit. Then (see for example [7]) we have

(2.1) *Every minimal counterexample is an internally 6-connected triangulation.*

At the end of (6.5) we give an algorithm which in effect constructs a 4-colouring of T from a short circuit of T and 4-colourings of all triangulations smaller than T, and (2.1) is a corollary of the existence of such an algorithm. It is easy to convert (6.5) to a proof of (2.1).

A drawing G is *planar* if one region of G is designated as *infinite*, and all the others *finite*. (More exactly, a planar drawing is a pair (G, r) where G is a drawing and r is a region of G; but this seems pedantic.) A *near-triangulation* is a non-null connected planar drawing G such that every finite region is a triangle. In figures representing planar drawings, the outside will always represent the infinite region.

A *configuration* K consists of a near-triangulation $G(K)$ and a map $\gamma_K : V(G(K)) \to \mathbb{Z}_+$ ($\mathbb{Z}_+$ denotes the non-negative integers) with the following properties:

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

(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.

Suppose we wish to describe a configuration K by a figure. One way is to "draw" the drawing, and write the number $\gamma_K(v)$ next to the point representing the vertex v, but this is inconvenient. A better way, due to Heesch [11], is to use a choice of "vertex shapes" to represent the values of $\gamma_K(v)$. The shapes we use are shown in figure 1.

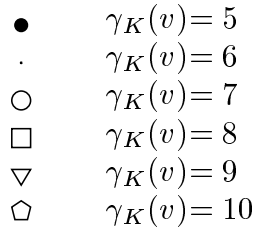

Figure 1: The shapes of vertices.

We did not make any special shape for vertices v with $\gamma_K(v) \geq 11$. Please note, therefore, that in the very last configuration of the Appendix, the vertex v of degree eight is supposed to satisfy $\gamma_K(v) = 11$. Apart from this, we shall not need to describe configurations having vertices v with $\gamma_K(v) \geq 11$.

In the Appendix we show 633 configurations, using the notation explained in figure 1. They are in lexicographic order of degree sequence. To refer to an individual configuration in this set, we use $\mathrm{conf}(x, y, z)$ to mean the configuration on row y and column z of page x of the Appendix. Some of the configurations are drawn with extra "half-edges"; for instance, $\mathrm{conf}(1, 1, 4)$ is the first. All these half-edges should be ignored for the moment. Also, certain of the edges are drawn thicker than usual – again, this should be ignored for the moment, and thickened edges regarded as normal edges. (For the reader who wishes to confirm our results, this list together with all the necessary programs can be accessed at the URL http://www.math.gatech.edu/~thomas/FC/ftpinfo.html in electronic form. The same material is also available via anonymous ftp from ftp.math.gatech.edu, in the directory pub/users/thomas/four.)

Two configurations K and L are *isomorphic* if there is a homeomorphism of Σ mapping $G(K)$ to $G(L)$ and γ_K to γ_L. Any configuration isomorphic to one in the Appendix is called a *good* configuration.

Let T be a triangulation. A configuration K *appears* in T if $G(K)$ is an induced subgraph of T, every finite region of $G(K)$ is a region of T, and $\gamma_K(v) = d_T(v)$ for every vertex $v \in V(G(K))$. We shall prove the following two statements.

(2.2) *If T is a minimal counterexample, then no good configuration appears in T.*

(2.3) *For every internally 6-connected triangulation T, some good configuration appears in T.*

From (2.1), (2.2) and (2.3) it follows that no minimal counterexample exists, and so the 4CT is true. We shall prove (2.2) in section 3, and (2.3) in section 4. The first proof needs a computer. The second can be checked by hand in a few months, or, using a computer, it can be verified in a few minutes.

Let G be a triangulation or a near-triangulation, and let $\kappa: E(G) \to \{-1, 0, 1\}$ be some function. A triangle r of G is *tri-coloured* (by κ) if $\{\kappa(e), \kappa(f), \kappa(g)\} = \{-1, 0, 1\}$, where e, f, g are the three edges incident with r. We say κ is a *tri-colouring* of G if every region is tri-coloured (if G is a triangulation), or every finite region is tri-coloured (if G is a near-triangulation). The reason for using -1, 0, 1 as the possible values instead of the more natural 1, 2, 3 (say) is just that the former was what we actually did in the computer program, and we might as well stay as close to the program as possible.

It was observed by Tait [14] that

(2.4) *A triangulation T is 4-colourable if and only if it admits a tri-colouring.*

We found tri-colourings more convenient than vertex 4-colourings to manipulate

in the computer program, although it is easy to convert one to the other, and so in what follows we shall use tri-colourings, again in an effort to stay as close to the program as possible.

In the proof of (2.2), we shall choose a small non-null set of edges of T, and contract them, producing a new drawing T'; then from the minimality of T it follows that T' has a vertex 4-colouring, and we convert this to a 4-colouring of T. There are some delicate issues involved in whether T' really exists (in particular, is loopless), but also, contracting edges in a drawing automatically produces a mass of notational difficulties (e.g. is an edge of the drawing after contraction *the same edge* as the corresponding edge in the drawing before contraction?) To avoid obscuring the serious issues with these notational ones, it seems best to break the contraction process into two stages, as follows.

Let G be a triangulation or a near-triangulation. A subset $X \subseteq E(G)$ is *sparse* if every region is incident with at most one edge in X, and the infinite region (in the case of a near-triangulation) is incident with no edge in X. If $X \subseteq E(G)$ is sparse, a *tri-colouring of G modulo X* is a map $\kappa: E(G) - X \to \{-1, 0, 1\}$ such that for every region (except the infinite region, in the case of a near-triangulation) incident with edges e, f, g,

(i) if $e, f, g \notin X$, then $\{\kappa(e), \kappa(f), \kappa(g)\} = \{-1, 0, 1\}$

(ii) if $g \in X$, then $\kappa(e) = \kappa(f)$.

Thus, a tri-colouring is a tri-colouring modulo $\emptyset$.

The following will allow us to use the fact that the drawing obtained by contracting all the edges in X is 4-colourable, without having to mention that drawing and its attendant notational complexities.

(2.5) *Let T be a minimal counterexample, and let $X \subseteq E(T)$ be sparse. Suppose that $X \neq \emptyset$, and there is no circuit C of T such that $|E(C) - X| = 1$. Then T admits a tri-colouring modulo X.*

Proof. Let F be the drawing with vertex set $V(T)$ and edge set X, and let $Z_1, \ldots, Z_k$ be the vertex sets of the components of F. Let S be the graph with vertex set $\{Z_1, \ldots, Z_k\}$ and edge set $E(T) - X$, in which $e \in E(T) - X$ is incident with Z_i if $\overline{e} \cap Z_i \neq \emptyset$. Since there is no circuit C of T with $|E(C) - X| = 1$, it follows that S is loopless; since S is obtained from T by contracting the edges in X, it is planar; since $X \neq \emptyset$, $|V(S)| + |E(S)| < |V(T)| + |E(T)|$; and since T is a minimal counterexample, S admits a vertex 4-colouring. Consequently there is a map $\phi: V(T) \to \{1, 2, 3, 4\}$ such that

(i) for $1 \leq i \leq k$, $\phi(v)$ is constant for $v \in Z_i$, and

(ii) for every edge e of T with $e \notin X$, $\phi(u) \neq \phi(v)$ where e has ends u, v.

For each edge $e \in E(T) - X$ with ends u, v, define

$$\kappa(e) = \begin{cases} -1 & \text{if } \{\phi(u), \phi(v)\} = \{1,2\} \text{ or } \{3,4\} \\ 0 & \text{if } \{\phi(u), \phi(v)\} = \{1,3\} \text{ or } \{2,4\} \\ 1 & \text{if } \{\phi(u), \phi(v)\} = \{1,4\} \text{ or } \{2,3\}. \end{cases}$$

Then κ is a tri-colouring of T modulo X, as we see as follows. Let r be a region of T, incident with edges e, f, g and vertices u, v, w, where e, f, g have ends uv, vw, uw respectively. If $e, f, g \notin X$, then $\phi(u)$, $\phi(v)$, $\phi(w)$ are all distinct, and so $\{\kappa(e), \kappa(f), \kappa(g)\} = \{-1, 0, 1\}$. On the other hand, if $g \in X$ (say), then $\phi(u) = \phi(w)$ and so $\kappa(e) = \kappa(f)$. □

Actually, when we apply (2.5), X will always be the edge-set of a forest of T.

3. Reducibility

Let R be a circuit. An *edge-colouring* of R is a map $\kappa: E(R) \to \{-1, 0, 1\}$. We wish to define what we mean by a "consistent set" of edge-colourings of R, and for that we need several definitions. A *match* m in R is an unordered pair $\{e, f\}$ of distinct edges of R, and a *signed match* in R is a pair (m, μ) where m is a match and $\mu = \pm 1$. A *signed matching* in R is a set M of signed matches, so that if $(\{e, f\}, \mu), (\{e', f'\}, \mu') \in M$ are distinct then

(i) $\{e, f\} \cap \{e', f'\} = \emptyset$, and

(ii) e, f belong to the same component of the graph obtained from R by deleting e' and f'.

If M is a signed matching, $E(M)$ denotes

$$\{e \in E(R): e \in m \text{ for some } (m, \mu) \in M\}.$$

For $\theta \in \{-1, 0, 1\}$, an edge-colouring κ of R is said to *θ-fit* a matching M in R if

(i) $E(M) = \{e \in E(R): \kappa(e) \neq \theta\}$, and

(ii) for each $(\{e, f\}, \mu) \in M$, $\kappa(e) = \kappa(f)$ if and only if $\mu = 1$.

A set $\mathcal{C}$ of edge-colourings of R is *consistent* if for every $\kappa \in \mathcal{C}$ and every $\theta \in \{-1, 0, 1\}$ there is a signed matching M such that κ θ-fits M, and $\mathcal{C}$ contains every edge-colouring that θ-fits M.

The significance of consistency stems from (3.1) below, which essentially dates back to Kempe [12] and Birkhoff [7]. Let H be a near-triangulation; then there is a closed walk

$$v_0, f_1, v_1, \ldots, f_k, v_k = v_0$$

tracing the boundary of the infinite region, in the natural sense. (In fact, there are several such walks, depending on the orientation and the choice of initial vertex.) Since H may not be 2-connected, this walk may have repeated vertices or edges. Let R be a circuit graph (not necessarily a circuit of H) of length k, with edges $e_1, \ldots, e_k$ in order; and for $1 \leq i \leq k$, define $\phi(e_i) = f_i$. We say that ϕ *wraps* R *around* H. If κ is a tri-colouring of H, let $\lambda(e) = \kappa(\phi(e))$ $(e \in E(R))$; then λ is an edge-colouring of R, called a *lift* of κ (by ϕ).

(3.1) *Let H be a near-triangulation, and let ϕ wrap a circuit R around H. Let $\mathcal{C}$ be the set of all lifts by ϕ of tri-colourings of H. Then $\mathcal{C}$ is consistent.*

Proof. Let $e_1, \ldots, e_k$, $f_1, \ldots, f_k$ be as in the definition of "wraps". Let $\lambda \in \mathcal{C}$, and let $\theta \in \{-1, 0, 1\}$. We must show that there is a signed matching M such that λ θ-fits M and $\mathcal{C}$ contains every edge-colouring of R that θ-fits M. By permuting -1, 0, 1 we may assume that $\theta = 0$.

Since $\lambda \in \mathcal{C}$, λ is the lift of some tri-colouring κ of H. A *rib* is a sequence

$$g_0, r_1, g_1, r_2, \ldots, r_t, g_t$$

such that

(i) $g_0, g_1, \ldots, g_t$ are distinct edges of H,

(ii) $r_1, r_2, \ldots, r_t$ are distinct finite regions of H,

(iii) if $t > 0$ then g_0, g_t are both incident with the infinite region of H, and if $t = 0$ then g_0 is incident with no finite region of H,

(iv) for $1 \le i \le t$, r_i is incident with g_{i-1} and with g_i, and

(v) for $0 \le i \le t$, $\kappa(g_i) \ne 0$.

In any rib the values of $\kappa(g_0), \kappa(g_1), \kappa(g_2), \ldots$ are ± 1 alternately, and $\kappa(e) = 0$ for every edge e not in the rib that is incident with a region in the rib. Thus, if we reverse the signs of $\kappa(g_0), \ldots, \kappa(g_t)$ we obtain a new tri-colouring of H.

Moreover, any two ribs are disjoint (they share neither edges nor regions); and for $1 \le i \le k$, f_i belongs to a unique rib if $\kappa(f_i) = \pm 1$, and to no rib if $\kappa(f_i) = 0$.

With each rib $g_0, r_1, \ldots, r_t, g_t$, we associate the signed match $(\{e_i, e_j\}, \mu)$, where $g_0 = f_i$, $g_t = f_j$, and $\mu = +1$ or -1 depending whether t is even or odd, respectively (or equivalently, whether $\kappa(g_0) = \kappa(g_t)$ or not, respectively). The set of all these signed matches is a signed matching M, and λ θ-fits M. Now let λ' be any edge-colouring of R that θ-fits M, and define $\kappa''(f_i) = \lambda'(e_i)$ $(1 \le i \le k)$. (This is well-defined, for if $f_i = f_j$ then $\lambda(e_i) = \lambda(e_j)$ and hence $\lambda'(e_i) = \lambda'(e_j)$.) By reversing the signs of κ in some of the ribs we can construct a tri-colouring κ' of H whose restriction to $\{f_1, \ldots, f_k\}$ is κ''. Then λ' is the lift of κ', and so $\lambda' \in \mathcal{C}$ as required. □

Since the null set is consistent, and the union of any two consistent sets is consistent, it follows that any set of edge-colourings $\mathcal{S}$ has a unique maximal consistent subset $\mathcal{S}'$. Moreover, for $|E(R)|$ sufficiently small, a computer can compute $\mathcal{S}'$ from a knowledge of $\mathcal{S}$ reasonably quickly. For instance, with $|E(R)| = 14$, which is the maximum we need, it normally takes less than a minute on a Sparc 20 workstation.

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

(i) R is an induced circuit of S, and bounds the infinite region of S

(ii) $G(K)$ is an induced subdrawing of S, $G(K) = S \backslash V(R)$, every finite region of $G(K)$ is a finite region of S, and the infinite region of $G(K)$ includes $U(R)$ and the infinite region of S, and

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

For instance, the drawing in figure 2 is a free completion of $\mathrm{conf}(1, 3, 2)$.

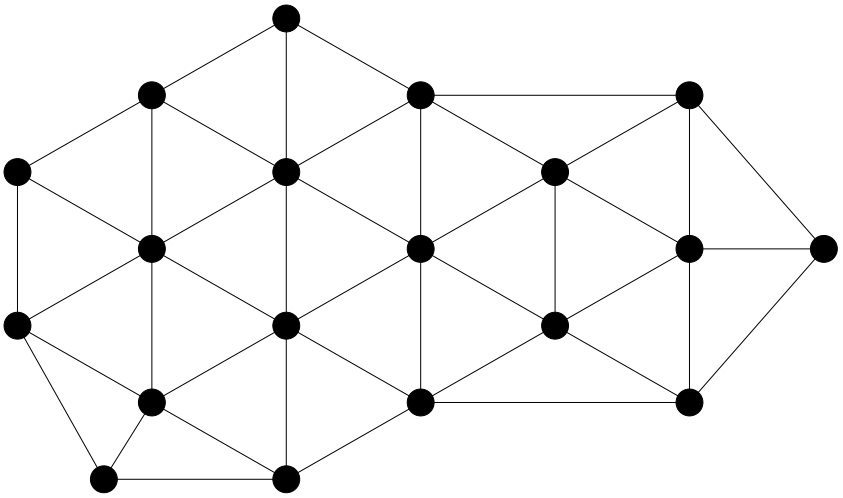

Figure 2. A free completion.

It is easy to check that every configuration has a free completion. (This is where we use the condition that ring-size ≥ 2 in the definition of a configuration – the ring-size is actually the length of the ring in the free completion, as the reader may verify.) Moreover, if S_1, S_2 are two free completions of K, there is a homeomorphism of Σ fixing $G(K)$ pointwise and mapping S_1 to S_2. (This is where condition (i) in the definition of a configuration is used.) Thus, there is essentially only one free completion, and so we may speak of "the" free completion without serious ambiguity.

Let S be the free completion of a configuration K with ring R. Let $\mathcal{C}^*$ be the set of all edge-colourings of R, and let $\mathcal{C} \subseteq \mathcal{C}^*$ be the set of all restrictions to $E(R)$ of tri-colourings of S. Let $\mathcal{C}'$ be the maximal consistent subset of $\mathcal{C}^* - \mathcal{C}$. The configuration K is said to be *D-reducible* if $\mathcal{C}' = \emptyset$. We shall see that no D-reducible configuration appears in a minimal counterexample.

There are several other ways in the literature to prove that a configuration does not appear in a minimal counterexample, but we shall not need the more difficult ones (general C-reducibility, block-count reducibility). The only other technique we require is a special case of C-reduction, the following.

With notation as before, let $X \subseteq E(S) - E(R)$. We say that X is a *contract* for K if $X \neq \emptyset$, X is sparse in S, and no edge-colouring in $\mathcal{C}'$ is the restriction to $E(R)$ of a tri-colouring of S modulo X.

Subject to one other condition (we have to check that if K occurs in an internally 6-connected triangulation T, then there is no circuit C of T with $|E(C) - X'| = 1$, where X' is the set of edges of T corresponding to X), no configuration K with a contract can appear in a minimal counterexample, as we shall see.

As we observed before, certain of the configurations drawn in the Appendix have extra "half-edges" – these are meant to indicate edges of the free completion. (Whenever we drew any half-edge incident with a vertex v, we also drew all the other edges of the free completion incident with v, in their natural cyclic order, to make clear which edge was which.) Also, in the Appendix, sometimes certain edges or half-edges are drawn thicker than usual.

By using a computer, we showed that

(3.2) *For each of the* 633 *configurations K drawn in the Appendix, let X be the set of edges of the free completion of K drawn thickened in the figure. If $X = \emptyset$, then K is D-reducible. Otherwise, $1 \le |X| \le 4$, and X is a contract for K.*

In the remainder of this section, we derive (2.2) from (3.2).

Suppose that K is a configuration that appears in a triangulation T. There need not be any subdrawing of T that is a free completion of K, and this presents certain problems when we try to deduce results about T from results about the free completion. We overcome them by means of (3.3) below.

If T is a triangulation or near-triangulation, its set of regions (excluding the infinite region, for a near-triangulation) is denoted by $F(T)$. Let T be a triangulation, and S a near-triangulation. A *projection of S into T* is a map ϕ with domain $V(S) \cup E(S) \cup F(S)$, such that

(i) ϕ maps $V(S)$ into $V(T)$, $E(S)$ into $E(T)$, and $F(S)$ into $F(T)$

(ii) for distinct $u, v \in V(S)$, $\phi(u) = \phi(v)$ only if u, v are both incident with the infinite region of S; for distinct $e, f \in E(S)$, $\phi(e) = \phi(f)$ only if e, f are both incident with the infinite region of S; and for distinct $r, s \in F(S)$, $\phi(r) \ne \phi(s)$

(iii) for $x, y \in V(S) \cup E(S) \cup F(S)$, if x, y are incident in S, then $\phi(x)$, $\phi(y)$ are incident in T.

(3.3) *Let K be a configuration appearing in a triangulation T, and let S be the free completion of K. Then there is a projection ϕ of S into T such that $\phi(x) = x$ for all $x \in V(G(K)) \cup E(G(K)) \cup F(G(K))$.*

This is a "folklore" theorem, and we omit its proof, which is straightforward but lengthy. A function ϕ as in (3.3) is called a *corresponding projection.*

(3.4) *Let K be a configuration appearing in a triangulation T, let S be the free completion of K, with ring R, and let ϕ be a corresponding projection of S into T. Let H be the planar drawing obtained from T by deleting $V(G(K))$ and designating*

as infinite the region of H including $V(G(K))$. Then H is a near-triangulation, and the restriction of ϕ to $E(R)$ wraps R around H.

This is another straightforward result, and so again we omit the proof. If G is a planar drawing, $v \in V(G)$ and $e \in E(G)$, we say that e *faces* v if e is not incident with v and there is a finite triangle of G incident with both e and v. If S is a free completion of a configuration K, and $X \subseteq E(S)$ is sparse in S with $|X| = 4$, a vertex v of S is a *triad* for X if

(i) $v \in V(G(K))$,

(ii) there are at least three vertices of S that are adjacent to v and incident with a member of X, and

(iii) if $\gamma_K(v) = 5$, then not every member of X faces v.

(3.5) *Let K be a configuration appearing in a triangulation T, let S be the free completion of K, and let ϕ be a corresponding projection of S into T. Let $X \subseteq E(S)$ be sparse in S with $|X| \leq 4$, such that if $|X| = 4$ there is a triad for X. If there is a circuit C of T with $|E(C) - \phi(X)| \leq 1$, then there is a short circuit in T.*

Proof. Let $X' = \phi(X) \cap E(C)$. Since X is sparse in S, no edge of X is incident with the infinite region of S, and consequently every edge in X is incident with two distinct finite regions of S. By (3.3), it follows that every region of T incident with an edge in X' is equal to $\phi(r)$ for some finite region r of S; and hence X' is sparse in T. Now $|E(C)| \leq |X| + 1 \leq 5$, and we may assume that C is not a short circuit of T, and so there is an open disc $\Delta \subseteq \Sigma$ bounded by C, with $|\Delta \cap V(T)| \leq 1$, and with $\Delta \cap V(T) = \emptyset$ if $|E(C)| \leq 4$. But every edge of X' is incident with a triangle of T included in Δ, and all these triangles are distinct since X' is sparse in T. Consequently, Δ includes at least $|X'| \geq |E(C)| - 1$ triangles of T. If $|E(C)| \leq 4$, then this is impossible since $\Delta \cap V(T) = \emptyset$; and so $|E(C)| = 5$, $|X| = 4$, there is a unique vertex t of T in Δ, $d_T(t) = 5$, and every edge of C faces t in T.

Since $|X| = 4$, there is a triad $v \in V(S)$ for X, by hypothesis. Either $d_K(v) \geq 6$ or some edge in X does not face v in S since v is a triad, and it follows in either case that $v = \phi(v) \neq t$. Since $v = \phi(v)$ has at least three distinct neighbours in C, and every vertex of C is adjacent to t, it follows that T has a short circuit, as required. □

Next we need

(3.6) *Let K be one of the configurations shown in the Appendix, let S be its free completion, and let X be the set of edges of S thickened in the figure. If $|X| = 4$, there is a triad for X.*

To prove this, the reader should examine individually all the configurations in the Appendix. For most of them, $|X| \leq 3$, so the task is not as difficult as it might seem.

(There is a subtle problem here. When we use a computer to check the correctness of (3.2) and later (4.9), the computer reads 633 configurations from a file, not directly from the Appendix. How can we be sure that the contents of the file matches what is drawn in the Appendix? Does the reader also have to check this somehow? One way to avoid that extra burden on the reader is to redefine "good configuration" to be one of the configurations in the file, rather than one in the Appendix, and to regard the Appendix just as an illustration. Then if for some reason there is a discrepancy between the two lists, our proof is still valid. But then the proof of (3.6) is suspect, because to prove it by hand we use the Appendix. To resolve this, we include in the programs a check that the configurations in the computer file satisfy (3.6).)

Now we prove (2.2), which we restate.

(3.7) *If T is a minimal counterexample, then no good configuration appears in T.*

Proof. Suppose that K is a good configuration that appears in T. Let S be the free completion of K, with ring R, and let ϕ be a corresponding projection of S into T. Let X be the set of edges of S corresponding to those thickened in the Appendix. Let H be obtained as in (3.4), and let ψ be the restriction of ϕ to $E(R)$.

By (3.4) H is a near-triangulation and ψ wraps R around H. Let $\mathcal{C}^*$ be the set of all edge-colourings of R, and let $\mathcal{C}_1 \subseteq \mathcal{C}^*$ be the set of all lifts of tri-colourings of H via ψ. By (3.1), $\mathcal{C}_1$ is consistent. Let $\mathcal{C}_2 \subseteq \mathcal{C}^*$ be the set of all restrictions to $E(R)$ of tri-colourings of S. Since T admits no tri-colouring by (2.4), it follows easily that $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$. Let $\mathcal{C}_3$ be the maximal consistent subset of $\mathcal{C}^* - \mathcal{C}_2$. Since $\mathcal{C}_1$ is consistent and $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$, it follows that $\mathcal{C}_1 \subseteq \mathcal{C}_3$.

It is possible to complete H to a triangulation T' by adding edges, and since T is a minimal counterexample it follows that T', and hence H, admits a tri-colouring. Consequently $\mathcal{C}_1 \neq \emptyset$, and so $\mathcal{C}_3 \neq \emptyset$, and K is not D-reducible. By (3.2), $1 \leq |X| \leq 4$ and X is a contract for K. Now X is sparse in S, and by (2.1) T has no short circuit, and so by (3.6) and (3.5) there is no circuit C of T with $|E(C) - \phi(X)| = 1$. Hence by (2.5) T admits a tri-colouring modulo $\phi(X)$, κ say. The restriction of κ to $E(H)$ is a tri-colouring of H, since $\phi(X) \cap E(H) = \emptyset$; and so its lift, λ say, via ψ belongs to $\mathcal{C}_1$ and hence to $\mathcal{C}_3$. But for $e \in E(S)$, let $\kappa'(e) = \kappa(\phi(e))$; then κ' is a tri-colouring of S modulo X, and λ is its restriction to R. This contradicts that X is a contract for S, and the result follows. □

4. Unavoidability

In this section we prove (2.3). A *cartwheel* is a configuration W such that there is a vertex w and two circuits C_1, C_2 of $G(W)$ with the following properties:

(i) $\{w\}$, $V(C_1)$, $V(C_2)$ are pairwise disjoint and have union $V(G(W))$

(ii) C_1 and C_2 are both induced subgraphs of $G(W)$, and $U(C_2)$ bounds the infinite region of $G(W)$

(iii) w is adjacent to all vertices of C_1 and to no vertices of C_2.

It follows that the edges of $G(W)$ are of four kinds; edges of C_1, edges of C_2, edges between w and $V(C_1)$, and edges between $V(C_1)$ and $V(C_2)$. We call w the *hub* of the cartwheel. See figure 3.

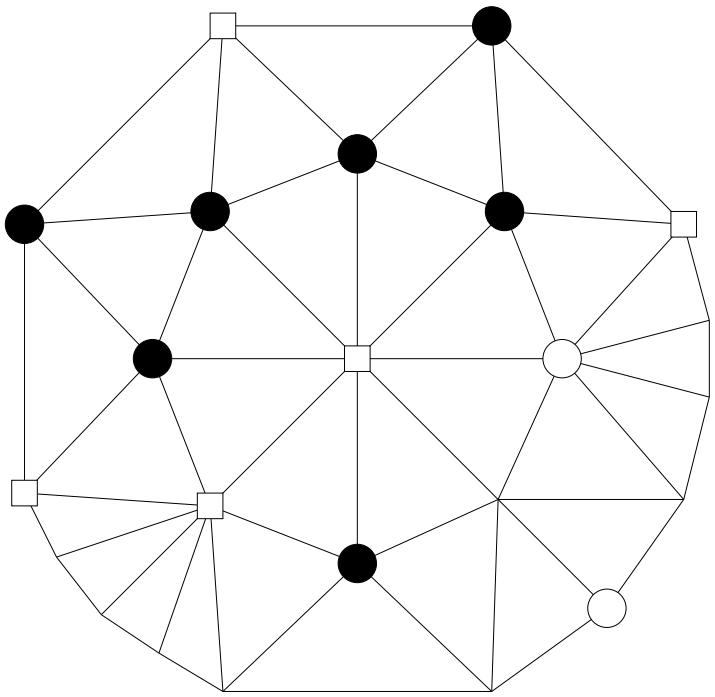

Figure 3: A cartwheel

To avoid confusion, let us stress that figure 3 is a picture of W, not of the free completion of W; the free completion would have *three* concentric circuits around w.

We need the following result of Birkhoff [7].

(4.1) *Let v be a vertex of an internally 6-connected triangulation T. There is a unique cartwheel appearing in T with hub v.*

Let W be a cartwheel. A configuration K *appears* in W if $G(K)$ is an induced subdrawing of $G(W)$, every finite region of K is a finite region of W (and hence the infinite region of K includes the infinite region of W), and $\gamma_K(v) = \gamma_W(v)$ for every $v \in V(G(K))$.

A *path* Q in a drawing G is a non-null connected subdrawing with no circuits in which every vertex has degree ≤ 2. Its length is $|E(Q)|$. (Thus, we permit paths of length 0, but we insist that paths have no repeated vertices or edges). It is a *u, v-path* if $u, v \in V(Q)$ and u, v are the vertices of Q of degree < 2.

A *pass* P is a quadruple (K, r, s, t) where

(i) K is a configuration,

(ii) r is a positive integer,

(iii) s and t are distinct adjacent vertices of $G(K)$, and

(iv) for each $v \in V(G(K))$ there is an s,v-path and a t,v-path in $G(K)$, both of length ≤ 2.

We write $r(P) = r$, $s(P) = s$, $t(P) = t$, and $K(P) = K$. We call r the *value* of the pass, s its *source*, and t its *sink*.

A pass P *appears* in a triangulation T if $K(P)$ appears in T. A pass P *appears* in a cartwheel W if $K(P)$ appears in W. Isomorphism for passes is defined in the natural way.

Let $\mathcal{P}$ be a set of passes. We write $P \sim \mathcal{P}$ to denote that P is a pass isomorphic to a member of $\mathcal{P}$. If W is a cartwheel, we define $N_{\mathcal{P}}(W)$ to be

$$\begin{aligned} 10(6 - \gamma_W(w)) + \sum &(r(P) \colon P \sim \mathcal{P}\,,\ P \text{ appears in } W\,,\ t(P) = w) \\ &- \sum (r(P) \colon P \sim \mathcal{P}\,,\ P \text{ appears in } W\,,\ s(P) = w)\,, \end{aligned}$$

where w is the hub of W.

(4.2) *Let T be an internally 6-connected triangulation, and let $\mathcal{P}$ be a set of passes. Then the sum of $N_{\mathcal{P}}(W)$, over all cartwheels W appearing in T, equals* 120.

Proof. For each $v \in V(T)$, there is a unique cartwheel W_v appearing in T with hub v, by (4.1). Thus, $\{W_v \colon v \in V(T)\}$ is the set of all cartwheels appearing in T.

Let P be a pass appearing in T, with source s. We claim that P appears in W_s. To see this, let $H = G(K(P))$, and let $G = G(W_s)$. Certainly $V(H) \subseteq V(G)$, by condition (iv) in the definition of a pass, since $V(G)$ contains every vertex v of T such that there is an s,v-path in T of length ≤ 2. Also, $E(H) \subseteq E(G)$, because $E(H) \subseteq E(T)$ and G is an induced subdrawing of T. Finally, let r be a finite region of H. Then r is a region of T since P appears in T; we must show that r is a finite region of G. Suppose not; then r is a subset of the infinite region of G. But every edge of T incident with r is an edge of G, and so r is the infinite region of G. Hence every region of G is a region of T, and so $G = T$, which is impossible since W_s has ring-size ≥ 2. This proves that r is a finite region of G, and so P appears in W_s, as claimed.

It follows that

$$\begin{aligned} &\sum (r(P) \colon P \sim \mathcal{P}\,,\ P \text{ appears in } T) \\ &\quad = \sum_{v \in V(T)} \sum (r(P) \colon P \sim \mathcal{P}\,,\ P \text{ appears in } W_v\,,\ s(P) = v)\,, \end{aligned}$$

because certainly every pass appearing in some W_v also appears in T. The same equation holds with $s(P)$ replaced by $t(P)$, and consequently

$$\sum_{v \in V(T)} N_{\mathcal{P}}(W_v) = \sum_{v \in V(T)} 10(6 - \gamma_W(v))\,.$$

Let $|V(T)| = n$. For each vertex v, $\gamma_W(v) = d_T(v)$, and

$$\sum_{v \in V(T)} d_T(v) = 2|E(T)| = 6n - 12\,,$$

by the well-known application of Euler's formula. Hence

$$\sum_{v \in V(T)} N_{\mathcal{P}}(W_v) = 60n - 10(6n - 12) = 120\,,$$

as required. □

It follows from (4.2) that

(4.3) *Let T be an internally 6-connected triangulation, and let $\mathcal{P}$ be a set of passes. Then there is a cartwheel W appearing in T with $N_{\mathcal{P}}(W) > 0$.*

We shall describe a set $\mathcal{P}$ of passes with the property that

(4.4) *For every cartwheel W with $N_{\mathcal{P}}(W) > 0$, some good configuration appears in W.*

Our second main result (2.3) follows immediately from (4.3) and (4.4); and so the objective of the remainder of this section is to describe $\mathcal{P}$ and show that (4.4) holds. Now our set $\mathcal{P}$ contains infinitely many non-isomorphic passes, but they can be divided conveniently into 32 classes, each described by what we call a "rule".

Formally, a *rule* is a 6-tuple $(G, \beta, \delta, r, s, t)$, where

(i) G is a near-triangulation, and $G\backslash v$ is connected for every vertex v,

(ii) β is a map from $V(G)$ to $\mathbb{Z}_+$; and δ is a map from $V(G)$ to $\mathbb{Z}_+ \cup \{\infty\}$ satisfying $\beta(v) \le \delta(v)$ for every vertex v,

(iii) $r > 0$ is an integer, and

(iv) s and t are distinct, adjacent vertices of G, and for every $v \in V(G)$ there is a v, s-path and a v, t-path of length ≤ 2, such that $\delta(w) \le 8$ for the internal vertex w of the path, if there is one.

A pass P *obeys* a rule $(G, \beta, \delta, r, s, t)$ if P is isomorphic to some (K, r, s, t) where $G(K) = G$ and $\beta(v) \le \gamma_K(v) \le \delta(v)$ for every vertex $v \in V(G)$.

Let us extend the conventions of figure 1 to describe rules. Conveniently, we shall not need many possibilities for the pairs $(\beta(v), \delta(v))$. In all cases, either

(a) $5 \le \beta(v) = \delta(v) \le 8$, or

(b) $\beta(v) = 5$ and $6 \le \delta(v) \le 8$, or

(c) $5 \le \beta(v) \le 8$ and $\delta(v) = \infty$.

To describe case (a), we naturally use the conventions of figure 1. For (b), we use the figure 1 convention that indicates a vertex v with $\gamma(v) = \delta(v)$, and add to the

figure a minus sign ($-$) close to the vertex. Similarly, for (c) we use the figure 1 convention for $\gamma(v) = \beta(v)$, and add a plus sign ($+$).

In addition, for each rule we indicate r, s and t by marking the edge joining s and t with an arrow (if $r = 1$) or double arrow (if $r = 2$), directed from s to t. (For all the rules we need, $r = 1$ or 2.) Thus figure 4 describes 32 rules (the reader should verify that in each case, conditions (i)-(iv) in the definition of "rule" are satisfied). Henceforth in this paper, $\mathcal{P}$ denotes the set of passes that obey one of these rules. (No pass obeys two distinct rules from figure 4, but there are a few instances of a pass P obeying a rule from the figure where the associated isomorphism from P to (K, r, s, t) is not unique – for instance, with rules 10 and 31. Let us stress that even in such a case, the pass P is counted only once in the set $\mathcal{P}$; the latter is a set, not a multiset.)

Passes obeying the first rule have value 2, and all other members of $\mathcal{P}$ have value 1. The first seven rules are different from the others. In any pass P that obeys one of them, the source s satisfies $\gamma_P(s) = 5$ or 6, while in a pass that obeys one of the other rules, $\gamma_P(s) = 7$ or 8 and the sink t satisfies $\gamma_P(t) \geq 7$. There is some system in the first seven rules, as we shall see in (4.5) below, but the other rules were chosen by trial and error, and have no particular plan or pattern.

At first sight the reader may wonder why in rule 10, the vertex v with $\beta(v) = 5$ and $\delta(v) = \infty$ cannot simply be deleted. The reason is, suppose a pass appears obeying rule 10 with v of degree 5. Then *two* passes would appear obeying the modified rule (mirror images of one another) and the net effect of the set of rules would be different.

We need to prove that our choice of $\mathcal{P}$ satisfies (4.4). To do so, we break (4.4) into three cases, depending on the degree of the hub of the cartwheel: degree at most 6, degree $7, \ldots, 11$, and degree at least 12. The first and last cases can be done by hand, as we shall see. For the first we need the following lemma.

(4.5) *Let W be a cartwheel, with hub w of degree 5 or 6. For $k = 1, \ldots, 32$ let p_k (respectively, q_k) be the sum of $r(P)$ over all passes P obeying rule k and appearing in W with sink (respectively, source) w. Suppose that no good configuration appears in W. Then:*

(i) $p_1 = q_2 + q_3$

(ii) $p_3 = q_4$

(iii) $p_4 = q_5 + q_6$, *and*

(iv) $p_5 = q_7$.

Proof. We write γ for γ_W. Let X be the set of all triples (x, y, z) of neighbours of w in W such that x,y,z are all distinct, y is adjacent to both x and z, and $\gamma(x) = 5$. Thus $p_1 = |X|$. Now q_2 is the number of $(x, y, z) \in X$ with $\gamma(y) \geq 7$; and q_3 is the number of $(x, y, z) \in X$ with $\gamma(y) \leq 6$ and $\gamma(z) \geq 6$. Since there is no $(x, y, z) \in X$

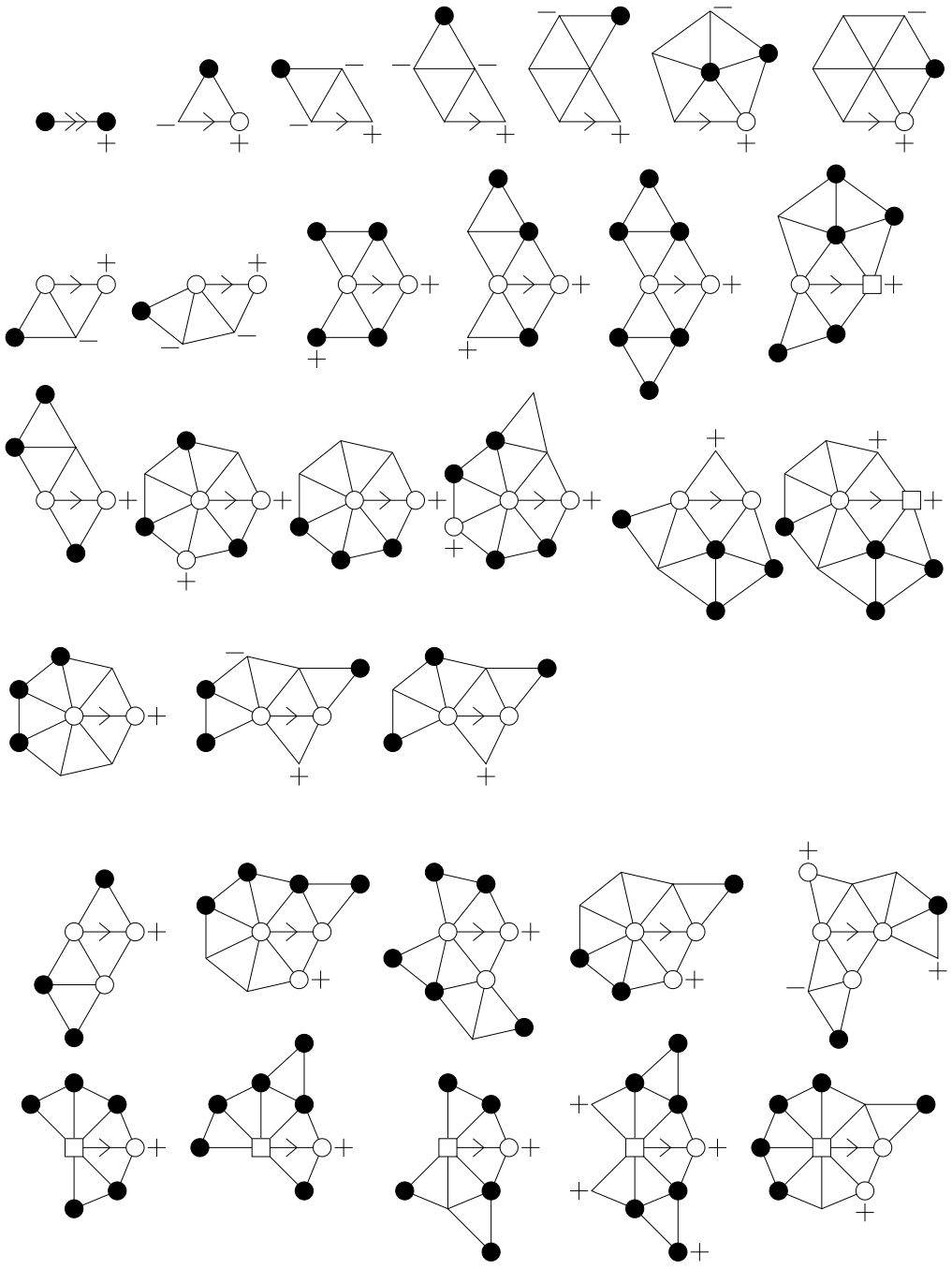

Figure 4: The rules

with $\gamma(y) \leq 6$ and $\gamma(z) = 5$ (because $\mathrm{conf}(1,1,1)$, $\mathrm{conf}(1,1,2)$ and $\mathrm{conf}(1,1,4)$ do not appear) it follows that $q_2 + q_3 = |X| = p_1$. This proves (i).

If $\gamma(w) = 5$ then p_3, p_4, p_5, q_4, q_5, q_6, q_7 are all zero and so (ii),(iii) and (iv) are true. We assume then that $\gamma(w) = 6$.

Now let X be the set of all triples (x, y, z) of neighbours of w such that x,y,z

are all distinct, y is adjacent to both x and z, $\gamma(x) \leq 6$, $\gamma(y) \leq 6$, and $\gamma(u) = 5$, where u is the vertex different from w adjacent to both x and y. Thus $p_3 = |X|$. But for each $(x, y, z) \in X$, $\gamma(z) \geq 6$ since $\mathrm{conf}(1,1,2), \mathrm{conf}(1,1,4), \mathrm{conf}(1,1,5)$ and $\mathrm{conf}(1,1,7)$ do not appear; and so $|X| = q_4$. This proves (ii).

The proofs of (iii) and (iv) are similar and we omit them. □

From (4.5) we deduce:

(4.6) *Let W be a cartwheel with $N_{\mathcal{P}}(W) > 0$, and with hub of degree 5 or 6. Then a good configuration appears in W.*

Proof. Let w be the hub, and define p_k and q_k for $k = 1, \ldots, 32$ as in (4.5). Suppose that no good configuration appears; we shall show that $N_{\mathcal{P}}(W) = 0$, a contradiction. First, let $\gamma_W(w) = 5$. Then $p_k = 0$ for $k = 2, \ldots, 32$ and $q_k = 0$ for $k = 4, \ldots, 32$, and so

$$N_{\mathcal{P}}(W) = 10 + p_1 - q_1 - q_2 - q_3 = 0$$

by (4.5) (since $q_1 = 10$). Now let $\gamma(w) = 6$. Then, again by (4.5),

$$N_{\mathcal{P}}(W) = p_1 + p_3 + p_4 + p_5 - q_2 - q_3 - q_4 - q_5 - q_6 - q_7 = 0\,.$$

In either case we have a contradiction. □

Now we prove (4.4) for cartwheels with hub of degree at least 12. For that we need a different lemma, the following.

(4.7) *Let W be a cartwheel with hub w, and let v be a neighbour of w. If no good configuration appears in W, then the sum of $r(P)$ over all passes $P \in \mathcal{P}$ appearing in W with source v and sink w is at most 5.*

Proof. For $k = 1, \ldots, 32$ let R_k be the sum of $r(P)$ over all passes P obeying rule k, appearing in W with source v and sink w. Let $R = R_1 + \cdots + R_{32}$. We must show that $R \leq 5$. We observe that for each k, $R_k \leq 2$.

First, let $\gamma_W(v) = 5$. Then $R_1 = 2$, and $R = 2 + R_2 + R_3$. Since $\mathrm{conf}(1,1,1)$ does not appear, it follows that not both R_2, $R_3 = 2$, and so $R \leq 5$ as required.

Now let $\gamma_W(v) = 6$. Then

$$R = R_2 + R_3 + R_4 + R_5 + R_6 + R_7\,.$$

Since $R_3 + R_5 + R_6 \leq 2$, $R_2 + R_7 \leq 2$, and $R_4 \leq 2$, we may assume that equality holds throughout. Since $R_4 = 2$ it follows that $R_7 = 0$, and $R_2 \leq 1$ (since $\mathrm{conf}(1,1,3)$, $\mathrm{conf}(1,1,6), \mathrm{conf}(1,2,2)$ do not appear), contradicting that $R_2 + R_7 = 2$. Thus $R \leq 5$ as required.

If $\gamma_W(v) \geq 9$ then $R = 0$, because $\gamma_P(s) \leq 8$ for the sink s of every pass $P \in \mathcal{P}$. There therefore remain the cases $\gamma_W(v) = 8$ and 7.

Let $\gamma_W(v) = 8$. Then

$$R = R_{28} + R_{29} + R_{30} + R_{31} + R_{32}\,.$$

Now R_{30},R_{31},$R_{32} \leq 1$, and if any one of them is nonzero then the other two are zero and so are R_{28} and R_{29} (since $\mathrm{conf}\,(5,7,2)$ does not appear). Hence we may assume that $R = R_{28} + R_{29} \leq 3$, as required.

Finally, let $\gamma_W(v) = 7$. This we need to break into several subcases. Let u_1 and u_2 be the two vertices adjacent to both v and w, and let $\gamma_W(u_1) = c_1$ and $\gamma_W(u_2) = c_2$. From the symmetry we may assume that $c_1 \leq c_2$.

First, let $c_1 = c_2 = 5$. Then

$$R = R_8 + R_9 + R_{10} + R_{11} + R_{12} + R_{13}\,.$$

If $R_{10} \geq 1$, then $R_{10} = 1$, $R_8 \leq 2$, $R_9 \leq 1$ (since $\mathrm{conf}\,(1,4,3)$ does not appear), $R_{12} + R_{13} \leq 1$ and $R_{11} = 0$, and hence $R \leq 5$ as required. We assume then that $R_{10} = 0$, and hence $R_{12} = R_{13} = 0$ and $R_8 \leq 1$. But $R_9 \leq 2$ and $R_{11} \leq 1$, so $R \leq 4$ as required.

Next, let $c_1 = 5$ and $c_2 = 6$. Then

$$R = R_8 + R_9 + R_{14} + R_{15} + R_{16} + R_{17} + R_{18} + R_{19}\,.$$

Not both R_{16} and R_{18} are nonzero, and both are zero if $R_8 = 2$, and so $R_8 + R_{16} + R_{18} \leq 2$. Also, $R_9 \leq 1$ (since $\mathrm{conf}\,(1,4,3)$ and $\mathrm{conf}\,(1,4,5)$ do not appear), $R_{14} + R_{19} \leq 1$, and $R_{15} + R_{17} \leq 1$, so $R \leq 5$ as required.

Next, let $c_1 = c_2 = 6$. Then

$$R = R_8 + R_9 + R_{20} + R_{21} + R_{22}\,.$$

But $R_9 \leq 2$, $R_8 + R_{21} \leq 2$ (since $\mathrm{conf}\,(1,4,3)$ and $\mathrm{conf}\,(2,10,6)$ do not appear), and $R_{20} + R_{22} \leq 1$, so $R \leq 5$ as required.

Next, let $c_1 = 5$ and $c_2 \geq 7$. Then

$$R = R_8 + R_9 + R_{18} + R_{19} + R_{23} + R_{24} + R_{25}\,.$$

But $R_8 \leq 1$, $R_9 \leq 1$,$R_{18} + R_{19} \leq 1$, and $R_{23} + R_{24} + R_{25} \leq 1$, so $R \leq 4$ as required.

Finally, let $c_1 \geq 6$ and $c_2 \geq 7$. Then

$$R = R_8 + R_9 + R_{21} + R_{22} + R_{26} + R_{27}\,.$$

But $R_8 \leq 1$, $R_9 \leq 1$, $R_{21} + R_{22} \leq 1$, and $R_{26} + R_{27} \leq 1$, so $R \leq 4$ as required.

This proves the result if $\gamma_W(v) = 7$, and hence completes the proof. □

From (4.7) we deduce:

(4.8) *Let W be a cartwheel with $N_{\mathcal{P}}(W) > 0$, and with hub of degree ≥ 12. Then a good configuration appears in W.*

Proof. Suppose that no good configuration appears. Let $\gamma_W(w) = d$ and let D be the set of neighbours of w, where w is the hub of W. For each $v \in D$, let $R(v)$ be the sum of $r(P)$ over all passes $P \in \mathcal{P}$ appearing in W with source v and sink w. Then $\sum_{v \in D} R(v) \leq 5d$ by (4.7). Hence

$$N_{\mathcal{P}}(W) = 10(6-d) + \sum_{v \in D} R(v) \leq 10(6-d) + 5d = 60 - 5d \leq 0\,,$$

a contradiction. The result follows. □

In view of (4.6) and (4.8), in order to prove (4.4) and hence (2.3) it remains to prove the following.

(4.9) *Let W be a cartwheel with $N_{\mathcal{P}}(W) > 0$, and with hub of degree 7, 8, 9, 10 or 11. Then a good configuration appears in W.*

For each of the five cases, we have a proof. Unfortunately they are very long (altogether about 13,000 lines, and a large proportion of the lines take some thought to verify), and so cannot be given here. Moreover, although any line of the proofs can be checked by hand, the proofs themselves are not "really" checkable by hand because of their length. We therefore wrote the proofs so that they are machine-readable, and in fact a computer can check these proofs in a few minutes. (More details are given in section 7.) Alternatively, one can write a computer program to check (4.9) directly, for it is easily seen to be a finite problem.

This concludes our proof of the 4CT. Of course, we have not given proofs of (3.2) and (4.9), but merely asserted that we checked them by computer. For the reader to be sure of the truth of these two statements, he needs to read the computer programs and then run them on a computer, or to write and use his own programs. To facilitate this, we are making all the necessary programs and data (and, in particular, the proofs of the five cases of (4.9)) available on the World-Wide Web and via "anonymous ftp" as described earlier. Verifying (3.2) takes about 3 hours on a Sun Sparc 20 workstation, and (4.9) takes about 20 minutes altogether. The first needs about one megabyte of RAM, and the second less. We used workstations for convenience, but the programs run on personal computers (including laptops) as well.

Gašper Fijavž, a student of one of the referees of this paper (Bojan Mohar), has independently verified the truth of (3.2) and (4.9). He wrote his own programs (in Pascal - ours were written in C) and ran them on two computers with two different processors (a 486 and a pentium based PC). No discrepancies were found. Also, Christopher Carl Heckman, a student of Robin Thomas, wrote a Pascal program to independently verify (4.9). His program is available from anonymous ftp along with the other programs.

5. Modifications, Extensions, Comments

We tried to "optimize" the proof presented here as best we could, but what precisely this means is open to debate. For instance, it seems desirable to

(a) make the size of the unavoidable set of reducible configurations as small as possible;

(b) make the number of rules as small as possible;

(c) make the running time of the computer programs as short as possible;

(d) make the non-computer parts of the proof (i.e., what is in this paper) as simple as possible.

These objectives unfortunately conflict. For instance, we could replace rules 14, 15 and 17 by one rule, simpler than all of these, at the cost of increasing the size of the unavoidable set by about 20. Also, there are several trade-offs between (a) and (d), as we explain below.

What we have presented is a somewhat uneasy compromise, and we would like to explain here some of the reasons we chose it, and what could have been done differently. Concerning (b), we could see how to make the rules a little simpler, but not much. (This is not to say that no better choice of rules exists; no doubt it does.) The place where we sacrificed the most was in the choice of ways to prove configurations reducible. As far as we know, there are basically four ways to prove that configurations do not appear in minimal counterexamples: showing they are

(i) D-reducible; this is perfectly acceptable and normal;

(ii) reducible because there is a contract in our sense, of size $\leq k$ say; let us call this being k-reducible;

(iii) C-reducible, in the sense of Heesch [11] and Appel and Haken [5]; this means that the configuration cannot appear, because if it did it could be replaced by something smaller, thereby producing a planar triangulation that is not 4-colourable, smaller than the supposedly "minimal" one ;

(iv) block-count reducible, in the sense of A. Bernhart (unpublished), Cohen [8] and Gismondi and Swart [10]; this is the same as being C-reducible, except that the definition of "consistent" is changed to something significantly more restrictive.

Using them all, we found an unavoidable set of size 591. But we decided to abandon (iv). This was a wrench, because several useful configurations were block-count reducible, and we could not reduce them in any other way (shown in figure 5). (The first of these was shown block-count reducible by Gismondi and Swart [10], and perhaps also by Bernhart, but the others seem to be new.) We were originally hoping that every "reasonable" configuration (defined later) might be block-count reducible, but with experiment our faith in this has declined.

The reason we dropped this method was that block-count reducibility takes a lot

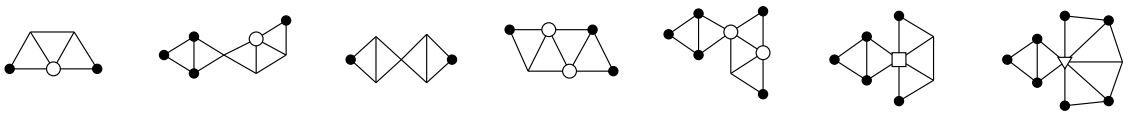

Figure 5: Block-count reducible configurations

of computing time, and also it seemed to be incompatible with getting a polynomial time algorithm to 4-colour a planar graph, which is the objective of the next section.

Second, general C-reducibility. This gives an unavoidable set of size 609, but the method is highly complicated and dangerous. It is very important to verify that if you replace a piece of a triangulation by a smaller piece, you do not introduce loops. There are two sources of complication, first that the free completion of the configuration may not appear in the triangulation, and second, that replacing the configuration by something smaller often involves identifying vertices on the ring of the free completion; and these two can conspire to create loops in most devious ways. (For instance, let a, b, c, d be vertices of the ring, in order, but not necessarily consecutive. Suppose that $\phi(a) = \phi(b)$, where ϕ is the corresponding projection, and $\phi(c)$ is adjacent to $\phi(d)$ in the triangulation, though c and d are not adjacent in the free completion. Now suppose that the "smaller piece" has a identified with d and b with c; then making this substitution creates a loop.) Of course, this can be controlled, with care, but we wanted to do this "safety check" part of the proof by hand, and it seemed to us too dangerous and complicated. So we decided not to regard as reducible everything that we could prove C-reducible, but only those for which the safety check was easy.

At the other extreme, the safety check is vacuous for D-reduction (which is a special case of C-reduction), and very easy for k-reduction, where $k \leq 3$. Now $\mathrm{conf}(1,1,4)$ is 4-reducible and not 3-reducible, but it has an easy safety check, and has been known to be reducible since 1948 [6], so it seems silly to exclude it. We decided therefore to accept this and a handful of other useful small configurations. There is an unavoidable set of configurations, all 1-reducible except for our handful, but it was rather large, about 900. (We expect that our handful could be eliminated if necessary; indeed, we expect that there is even an unavoidable set of D-reducible configurations, but have not tried to find it.) If we had stopped at 3-reducibility, the set would have size about 700. We tried allowing everything that was 4-reducible, and it turned out by accident that all members of the optimal subset had an easy safety check, given in (3.5), so that seemed very satisfactory, and we fixed on it. If we had gone to (say) 8-reducibility, the unavoidable set would have had size about 618, but the analogue of (3.5) would be much more complicated.

Incidentally, given a contract for a configuration, it is easy to verify that it really is a contract – but how did we find it in the first place? We did nothing clever here, although there are clever methods available (see [1]). We just tried all possibilities.

Another natural question – why did we insist that contracts were sparse? Because everything we needed and could reduce with a "non-sparse contract" (i.e. a set of

edges containing 3, 1 or 0 edges from each finite region) we could also reduce with a sparse one of at most the same size, so it was a free way to simplify the paper.

A third question; while it is straightforward to check that the set of the Appendix works, how did we find it? For that, we need to discuss a conjecture of Heesch [11], that guided us.

Let us say a configuration K is *reasonable* if

(i) $\gamma_K(v) \leq d(v) + 3$, for every vertex v, and

(ii) no 3-vertex block of $G(K)$ has two vertices of degree 2 in $G(K)$.

Heesch made the conjecture that every minimally reducible configuration is reasonable. This is still open, though it is proved for D-reducibility (Whitney and Tutte [15] proved (i), and we proved (ii), unpublished). But we believed Heesch's conjecture enough that we almost never tested any configuration for reducibility unless it was reasonable. (In fact, we think that condition (ii) can be replaced by the stronger condition that $G(K)$ has no block with at most three vertices.) So now, how did we find the 633 configurations? We simply examined all cartwheels W with hub of degree $7, \ldots, 11$ with $N_{\mathcal{P}}(W) > 0$, and for each such W listed all the reasonable configurations it contained, found which were minimally reducible, and discarded the others. Thus, for each cartwheel we had a set of configurations. Then we needed to choose another set meeting each cartwheel set, as small as possible. For this we used a heuristic.

Incidentally, the appealing feature of Heesch's conjecture above is in the reverse direction; most reasonable configurations seem to be reducible. We examined in total 15165 reasonable configurations, and we could reduce (by D- or C-reducibility) 14051 of them.

There is another helpful observation, due to Allaire [1]. Let K be a configuration, and let S be its free completion, with ring R. Let $\mathcal{C}^*$ be the set of all edge-colourings of R, and let $\mathcal{C}$ be all restrictions to $E(R)$ of tri-colourings of S. Let $\mathcal{C}_1$ be the maximal consistent subset of $\mathcal{C}^* - \mathcal{C}$, and let $\mathcal{C}_2$ be the maximal consistent subset of $\mathcal{C}^* - \mathcal{C}_1$. Thus, $\mathcal{C} \subseteq \mathcal{C}_2$, but equality need not hold. If $\mathcal{C}_2 = \mathcal{C}$, K is said to be *symmetrically D-irreducible* (SDIR). Allaire observed that there is a very close correspondence between being SDIR and being irreducible (by D- and C-reduction). Of our 15165 configurations, 1086 of them are SDIR, and as far as we know only one of them is reducible; and of the remainder, we can reduce all except 29. Allaire (somewhat bravely) conjectured that every non-SDIR configuration was in fact D- or C-reducible.

Finally, let us point out that the existence of a set of rules like those in figure 4, contained in the "second neighbourhood" of both the source and sink, confirms another conjecture of Heesch [11].

6. A Colouring Algorithm

In this section we convert the proof of the 4CT to an algorithm that finds a vertex 4-colouring of a drawing T, in time $O(|V(T)|^2)$. The two computer-search results,

(3.2) and (4.9), are of course still needed, but they are not part of the algorithm; they are part of the proof that every step of the algorithm can be carried out.

It is sufficient to find a vertex 4-colouring of a triangulation, since if we wish to 4-colour a drawing that is not a triangulation, it is easy to make it a triangulation by adding edges. In view of this, it suffices to find a tri-colouring rather than a vertex 4-colouring, because it is easy to convert one to the other.

A first attempt at an algorithm would be:

Step 1: Find a short circuit in T if there is one, or conclude that T is internally 6-connected.

Step 2: If there is a short circuit C, adapt the proof of (2.1) to construct a tri-colouring of T from tri-colourings of the restrictions of T to the two discs bounded by C.

Step 3: If there is no short circuit, locate a good configuration that appears in T, and the appropriate contract X; find a tri-colouring modulo X; and convert it to a tri-colouring of T by adapting the proof of (3.7).

Steps 2 and 3, which might seem the non-trivial parts, can easily be done algorithmically. The problem lies in step 1; we don't know how to do this in linear time. So we are forced to proceed more deviously. Instead, we go ahead directly with step 3, assuming there is no short circuit; then if at some stage something goes wrong, it is because there *was* a short circuit, and now we can find it and go to step 2 instead.

A word on data structure; we assume that the triangulation is input in a form comprised of a number $n \geq 0$, a vector $(d_1, \ldots, d_n)$ of non-negative integers, and for all v with $1 \leq v \leq n$, a vector $(u_v(1), \ldots u_v(d_v))$ of integers all between 1 and n. The number of vertices will be n, and for $1 \leq v \leq n$, d_v will be the number of edges incident with v, and $u_v(1), \ldots, u_v(d_v)$ will be the ends of these edges different from v, enumerated in clockwise cyclic order around v of the edges in the natural sense. This input has size $O(|V(T)|)$, because $|E(T)| \leq 3n - 6$. (The inequality $|E(T)| \leq 3n - 6$ is true even for triangulations with parallel edges.)

To input a near-triangulation, we observe that there is essentially a unique way to add a vertex v_0 in the infinite region and add edges incident with v_0, to make a triangulation. Thus, we can regard a near-triangulation on n vertices as a triangulation on $n+1$ vertices with one vertex distinguished; and this is the data structure we use.

We need algorithmic versions of (2.1), (3.1) and (4.1). Let us do them in reverse order.

(6.1) *Algorithm.*

Input: *A triangulation T, and a vertex w of T.*

Output: *Either a cartwheel appearing in T with hub w, or a short circuit of T.*

Running Time: $O(|V(T)|)$.

This is easy, and we omit the details.

(6.2) *Algorithm.*

Input: *A near-triangulation H, a circuit R, a function ϕ wrapping R around H, and a tri-colouring κ of H.*

Output: *A set of tri-colourings of H including κ, such that their lifts by ϕ are all distinct and form a consistent set.*

Running Time: *$O(|V(H)|)$, if $|E(R)|$ is at most a constant.*

Description. We start with $\mathcal{C}_1 = \{\kappa\}$ and begin the first iteration. In general, at the start of the ith iteration, we have a set $\mathcal{C}_i$ of tri-colourings of H including κ, so that their lifts by ϕ are all distinct. We test if this set of lifts is consistent, and if not, find $\kappa' \in \mathcal{C}_i$ and $j \in \{-1, 0, 1\}$ so that there is no signed matching M satisfying

(a) the lift of κ' j-fits M, and

(b) every edge-colouring that j-fits M is the lift of some member of $\mathcal{C}_i$.

This takes constant time, if $|E(R)|$ is at most a constant. If the set is consistent, we output $\mathcal{C}_i$. If not, we adapt the proof of (3.1) in the natural way to obtain from κ' (in linear time) a new tri-colouring κ_i of H whose lift is different from the lifts of all members of $\mathcal{C}_i$, set $\mathcal{C}_{i+1} = \mathcal{C}_i \cup \{\kappa_i\}$, and return for the next iteration. The number of iterations is at most a constant if $|E(R)|$ is at most a constant, since $|\mathcal{C}_i| = i$ and all the lifts of $\mathcal{C}_i$ are distinct. □

To obtain an algorithmic version of (2.1) we need two lemmas, the following. If κ, κ' are edge-colourings of a circuit R, they are *equivalent* if there is a permutation λ of $\{-1, 0, 1\}$ such that for every $e \in E(R)$, $\kappa'(e) = \lambda(\kappa(e))$. This is an equivalence relation, and it is easy to check that every consistent set is a union of equivalence classes. The following (due to Birkhoff [7]) is proved by easy case analysis, and we omit the proof.

(6.3) *Let R be a circuit of length 4, with edges e_1, e_2, e_3, e_4 in order. Let $\mathcal{C}_0$ be the set of all edge-colourings equivalent to $(0, 0, 0, 0)$ (that is, the edge-colouring κ with $\kappa(e_i) = 0$ $(i = 1, \ldots, 4)$). Similarly, let $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$ be the sets equivalent to $(0, 1, 1, 0)$, $(0, 1, 0, 1)$, $(0, 0, 1, 1)$ respectively. Every non-empty consistent set of edge-colourings of R includes one of $\mathcal{C}_0 \cup \mathcal{C}_1$, $\mathcal{C}_1 \cup \mathcal{C}_2$, $\mathcal{C}_2 \cup \mathcal{C}_3$, $\mathcal{C}_3 \cup \mathcal{C}_0$.*

Similarly, we have (also due to Birkhoff [7])

(6.4) *Let R be a circuit of length 5, with edges e_1, e_2, e_3, e_4, e_5 in order. For $1 \le i \ne j \le 5$, let $\mathcal{A}_{ij}$ be the equivalence class of edge-colourings of R equivalent to κ, where $\kappa(e_i) = 1$, $\kappa(e_j) = -1$, and $\kappa(e_k) = 0$ for $k \ne i, j$. For $1 \le i \le 5$, let $\mathcal{C}_i = \mathcal{A}_{ij} \cup \mathcal{A}_{ik} \cup \mathcal{A}_{jk}$, where $e_j, e_k \ne e_i$ are the two edges with a common end with e_i. For $1 \le i \le 5$, let the edges different from e_i be e_a, e_b, e_c, e_d in order, and let $\mathcal{D}_i = \mathcal{A}_{ac} \cup \mathcal{A}_{ad} \cup \mathcal{A}_{bc} \cup \mathcal{A}_{bd}$. Finally, let $\mathcal{E} = \mathcal{A}_{12} \cup \mathcal{A}_{23} \cup \mathcal{A}_{34} \cup \mathcal{A}_{45} \cup \mathcal{A}_{15}$. Every non-empty consistent set that meets $\mathcal{E}$ includes one of $\mathcal{C}_1, \ldots, \mathcal{C}_5$, $\mathcal{D}_1, \ldots, \mathcal{D}_5$, $\mathcal{E}$.*

Proof. Again, this is just case analysis; but since the proof is easy to give and not so easy to find, we give it. Let $\mathcal{C}$ be a consistent set.

(1) *If $\mathcal{A}_{12} \subseteq \mathcal{C}$, then $\mathcal{C}$ includes one of $\mathcal{A}_{13}$, $\mathcal{A}_{15}$ and one of $\mathcal{A}_{23}$, $\mathcal{A}_{25}$.*

For let $\kappa = (-1, 1, 0, 0, 0) \in \mathcal{C}$, with the natural notation for edge-colourings. Since $\mathcal{C}$ is consistent, there is a signed matching M such that κ (-1)-fits M, and $\mathcal{C}$ contains all edge-colourings that (-1)-fit M. Since κ (-1)-fits M it follows that M is one of

$$\{(\{e_2, e_3\}, -1), (\{e_4, e_5\}, 1)\}$$
$$\{(\{e_2, e_5\}, -1), (\{e_3, e_4\}, 1)\}.$$

If M is the first of these, then the edge-colouring $(-1, 0, 1, 0, 0)$ (-1)-fits it and so belongs to $\mathcal{C}$, and hence $\mathcal{A}_{13} \subseteq \mathcal{C}$; and otherwise, $(-1, 0, 0, 0, 1)$ belongs to $\mathcal{C}$ and hence $\mathcal{A}_{15} \subseteq \mathcal{C}$. Also, there are only two signed matchings M' such that κ 1-fits M', and the second conclusion follows similarly.

(2) *If $\mathcal{A}_{13} \subseteq \mathcal{C}$ then $\mathcal{C}$ includes one of $\mathcal{A}_{23}$, $\mathcal{A}_{35}$.*

The proof is similar.

Now, let $\mathcal{C}$ be a consistent set that meets $\mathcal{E}$. Thus, one of $\mathcal{A}_{12}$, $\mathcal{A}_{23}$, $\mathcal{A}_{34}$, $\mathcal{A}_{45}$, $\mathcal{A}_{15}$ is included in $\mathcal{C}$, but we may assume that not all of them are, for if so, then $\mathcal{E} \subseteq \mathcal{C}$ as required. From the symmetry we may assume that $\mathcal{A}_{12} \subseteq \mathcal{C}$ and $\mathcal{A}_{23} \not\subseteq \mathcal{C}$. From (1), $\mathcal{A}_{25} \subseteq \mathcal{C}$. If $\mathcal{A}_{15} \subseteq \mathcal{C}$, then $\mathcal{C}_1 \subseteq \mathcal{C}$ and the theorem holds, so we assume that $\mathcal{A}_{15} \not\subseteq \mathcal{C}$. By (1), $\mathcal{A}_{13} \subseteq \mathcal{C}$. By (2), $\mathcal{A}_{35} \subseteq \mathcal{C}$, and so $\mathcal{D}_4 \subseteq \mathcal{C}$, as required. □

Now let us put these pieces together.

(6.5) *Algorithm.*

Input: *A triangulation T.*

Output: *A tri-colouring of T.*

Running Time: $O(|V(T)|^2)$.

Description. First we test if T has two parallel edges. (This takes time $O(|V(T)|)$.) If so we go to the short circuit subroutine described later. Otherwise, T is simple and hence 3-connected (unless $|V(T)| \leq 3$, which is trivial).

We test if every vertex has degree ≥ 5. (This takes time $O(|V(T)|)$.) If not, the neighbours of the offending vertex form a short circuit (unless $|V(T)| \leq 5$ which is trivial) and we go to the short circuit subroutine. Otherwise, T has minimum degree 5.

For each vertex v, we compute $N(v)$, defined as $10(d(v) - 6) + a - b$, where a, b are the sums of $r(P)$ over all passes $P \sim \mathcal{P}$ appearing in T with sink (respectively, source) v. For each v and each rule of figure 4, this takes time $O(d(v))$, so the total running time is $O(|V(T)|)$.

Now, as in the proof of (4.2),

$$\sum (N(v) : v \in V(T)) = 120.$$

Note that this does *not* require that T be internally 6-connected; the only place that hypothesis was used in the proof of (4.2) was to apply (4.1), which we are not doing here. Consequently, we can choose a vertex w with $N(w) > 0$, in time $O(|V(T)|)$.

We apply (6.1) to T and w. This takes time $O(|V(T)|)$. If we find a short circuit, we go to the short circuit subroutine. Otherwise, we have a cartwheel W appearing in T with $N_{\mathcal{P}}(W) > 0$.

By (4.4), some good configuration appears in W. We find such a configuration K, in time $O(|V(T)|)$. Thus, K appears in T.

We construct the free completion S of K with ring R; then $|E(R)| \leq 14$. We construct the corresponding projection ϕ of S into T. If K is D-reducible, let X be any singleton subset of $E(S) - E(R)$. Otherwise, let $X \subseteq E(S) - E(R)$ be the set of edges thickened in the Appendix. Let $X' = \phi(X)$. (This all takes constant time.)

We test if there is a circuit C of G with $|E(C) - X'| \leq 1$. If so, we use the obvious algorithmic version of (3.5) to find a short circuit of T, and go to the short circuit subroutine. (This takes time $O(|V(T)|)$.)

We construct the triangulation T' obtained from T by contracting every edge in X' and deleting an edge from any two parallel edges that bound a region. We call (6.5) to find a tri-colouring of T'. We convert this to a tri-colouring κ_1 of T modulo X'.

Let H be the planar drawing obtained from T by deleting $V(G(K))$ and designating as infinite the region including $V(G(K))$. Let κ_2 be the restriction of κ_1 to $E(H)$; then κ_2 is a tri-colouring of H. Let ψ be the restriction of ϕ to $E(R)$. We apply (6.2) to H, R, ψ and κ_2, to obtain a set $\mathcal{C}$ of tri-colourings such that all their lifts by ψ are distinct and form a consistent set $\mathcal{D}$, with $\kappa_2 \in \mathcal{C}$.

Now the lift of κ_2 by ψ is the restriction to $E(R)$ of a tri-colouring of S modulo X, by the choice of κ_1. Since X is a contract for K by (3.2), $\mathcal{D}$ contains the restriction to $E(R)$ of some tri-colouring of S. We find such a tri-colouring κ_3; combine κ_2 and κ_3 to obtain a tri-colouring κ of T; and return κ.

It remains to describe the short circuit subroutine. For this, the input is T together with a short circuit C of T.

First, if $|E(C)| \geq 3$ we test if C is an induced circuit of T, and if not we find another short circuit of smaller length and put it in place of C; we repeat this until either $|E(C)| = 2$ or C is induced. (This takes time $O(|V(T)|)$.)

We find all connected components $X_1, \ldots, X_k$ of $T \backslash V(C)$. (Necessarily $k \geq 2$, since T is a triangulation; and in fact if $|E(C)| \geq 3$, then $k = 2$, as is easily seen.) For $1 \leq i \leq k$, we construct the drawing H_i consisting of X_i, C and all edges with one end in $V(C)$ and the other in $V(X_i)$.

Suppose first that $|E(C)| = 2$. Let $E(C) = \{f, g\}$, and for each i, we use (6.5) to obtain a tri-colouring κ_i of the triangulation $H_i \backslash g$, such that $\kappa_i(f) = 0$. We

define κ by: $\kappa(e) = \kappa_i(e)$ if $e \in E(H_i) - E(C)$, and otherwise $\kappa(e) = 0$. Then κ is a tri-colouring of T.

Henceforth we may assume that $|E(C)| \geq 3$, and hence $k = 2$ and C is induced, and in particular every edge of T belongs to H_1 or to H_2. If $|E(C)| = 3$ we use (6.5) to find tri-colourings of H_1 and H_2 that agree on $E(C)$, and piece them together to form a tri-colouring of T. The cases $|E(C)| = 4$, 5 are more complicated. Let the edges of C be $e_1, \ldots, e_d$ in order, where $d = |E(C)|$, and let the vertices be $v_1, \ldots, v_d$, where e_i has ends v_i, v_{i+1} $(1 \leq i \leq d)$ and v_{d+1} means v_1.

First, suppose that $|E(C)| = 4$. Add an edge to H_1 with ends v_1, v_3, forming a triangulation T_1; apply (6.5) to obtain a tri-colouring of T_1, and hence a tri-colouring κ of H_1 such that $\kappa(e_1) \neq \kappa(e_2)$. Apply (6.2) (with ϕ the identity and $R = C$) to obtain a set of tri-colourings $\mathcal{B}$ of H_1, such that $\kappa \in \mathcal{B}$, and the restrictions of the members of $\mathcal{B}$ to $E(C)$ are all different and form a consistent set $\mathcal{C}$. By (6.3) and symmetry, we may assume that either $\mathcal{C}_0 \cup \mathcal{C}_1 \subseteq \mathcal{C}$ or $\mathcal{C}_1 \cup \mathcal{C}_2 \subseteq \mathcal{C}$ (using the notation of (6.3)). If $\mathcal{C}_0 \cup \mathcal{C}_1 \subseteq \mathcal{C}$, construct T_2 from H_2 by deleting e_3 and e_4 and identifying v_2 with v_4; then apply (6.5) to T_2, to obtain a tri-colouring κ_2 of H_2 with $\kappa_2(e_1) = \kappa_2(e_4)$. Consequently, the restriction of κ_2 to $E(C)$ is in $\mathcal{C}_0 \cup \mathcal{C}_1 \subseteq \mathcal{C}$, and so there exists $\kappa_1 \in \mathcal{B}$ such that $\kappa_1(e) = \kappa_2(e)$ $(e \in E(C))$. Then piece κ_1 and κ_2 together to obtain a tri-colouring of T. On the other hand, if $\mathcal{C}_1 \cup \mathcal{C}_2 \subseteq \mathcal{C}$, add an edge to H_2 with ends v_1, v_3, and proceed similarly.

Now suppose that $|E(C)| = 5$. Let T_1 be obtained from H_1 by adding a new vertex adjacent to $v_1, \ldots, v_5$. By (6.5) and (6.2) we construct a set $\mathcal{B}$ of tri-colourings of H_1, such that their restrictions to $E(C)$ form a non-empty consistent set $\mathcal{C}$ meeting $\mathcal{E}$, using the notation of (6.4). By (6.4), one of $\mathcal{C}_1, \ldots, \mathcal{C}_5$, $\mathcal{D}_1, \ldots, \mathcal{D}_5$, $\mathcal{E}$ is included in $\mathcal{C}$. If say $\mathcal{C}_1 \subseteq \mathcal{C}$, let T_2 be obtained from H_2 by deleting e_4 and identifying v_3 with v_5. If $\mathcal{D}_1 \subseteq \mathcal{C}$, let T_2 be obtained from H_2 by adding two edges with ends v_2v_4, v_2v_5. If $\mathcal{E} \subseteq \mathcal{C}$, let T_2 be obtained from H_2 by adding a new vertex adjacent to v_1, v_2, v_3, v_4, v_5. In each case we apply (6.5) to T_2, and thereby obtain a tri-colouring κ_2 of H_2 whose restriction to $E(C)$ is in $\mathcal{C}$ and therefore equals the restriction to $E(C)$ of some member κ_1 of $\mathcal{B}$. By piecing κ_1 and κ_2 together we obtain a tri-colouring of T.

This concludes the short circuit subroutine. It is an easy exercise to check that (6.5) has running time $O(|V(T)|^2)$. □

7. The Machine-Checkable Proofs

In this section we describe in more detail the structure of our proof of (4.9). We use a branch-and-bound approach; in other words, we start with a totally general cartwheel, that we wish to show satisfies (4.9), and we repeatedly break the problem into cases until in each case the cartwheel is sufficiently restricted that we can "see" that it satisfies (4.9). Thus, as we proceed, we have partial knowledge of the cartwheel, increasing as we go deeper into subcases, and before we can explain anything else we must explain how this partial knowledge is represented.

Let W be a cartwheel, and let $\gamma = \gamma_W$. The vertices of $G(W)$ are of four kinds:

the hub, w say; the neighbours of w, called *spokes*; vertices different from w adjacent to two distinct spokes, called *hats*; and the other vertices, called *fan* vertices. Each fan vertex is adjacent to a unique spoke. For each spoke v, the *fan over* v is the set of fan vertices adjacent to v (if $\gamma(v) \geq 6$) or the edge joining the two hats adjacent to v (if $\gamma(v) = 5$).

Choose a subset X of the spokes, and delete from $G(W)$ the fan over v for each $v \in X$; let the resulting near-triangulation be K. For each vertex v of K, let $a(v) \in \mathbb{Z}_+ \cup \{\infty\}$ and $b(v) \in \mathbb{Z}_+$ such that

$$5 \leq b(v) \leq \gamma(v) \leq a(v),$$

and such that $b(v) = \gamma(v) = a(v)$ if either $v = w$ or v is a spoke not in X, and $b(v) \neq a(v)$ for $v \in X$. We call the triple (K, a, b) a *part, fitting* W. This is what we mean by "partial knowledge" of a cartwheel; we shall know a part that fits it. A given part may fit many different cartwheels. A part is *successful* if every cartwheel W that it fits in which no good configuration appears satisfies $N_{\mathcal{P}}(W) \leq 0$.

We define the hub, spokes, hats and fans of a part in the natural way. A part is *trivial* if $a(v) = \infty$ and $b(v) = 5$ for every vertex v except the hub (and consequently it has no fans and all its hats are pairwise non-adjacent). The trivial part with hub of degree k (it is unique up to isomorphism) fits (an isomorphic copy of) every cartwheel with hub of degree k, and consequently, to prove (4.9) it suffices to prove the following.

(7.1) *For $k = 7, 8, 9, 10$ and 11, the trivial part with hub of degree k is successful.*

Let us say a part (K', a', b') is a *refinement* of a part (K, a, b) if K is a subdrawing of K' with the same hub, and

$$b(v) \leq b'(v) \leq a'(v) \leq a(v)$$

for each vertex v of K. Two particular refinements are of special importance. Let (K, a, b) be a part, and let v be a vertex of K with $a(v) \neq b(v)$. Let c be an integer with $b(v) \leq c < a(v)$. Let (K_1, a_1, b_1) be the part defined as follows. If v is a spoke and $b(v) = c$, let K_1 be a near-triangulation obtained by adding to K a $(c-4)$-edge path between the two hats adjacent to v, and making every internal vertex of this path adjacent to v; and otherwise let $K_1 = K$. For $v' \in V(K_1)$, let $a_1(v') = \infty$ and $b_1(v') = 5$ if $v' \in V(K_1) - V(K)$, let $a_1(v') = a(v')$ and $b_1(v') = b(v')$ for $v' \in V(K) - \{v\}$, and let $a_1(v) = c$ and $b_1(v) = b(v)$.

Let (K_2, a_2, b_2) be defined as follows. If v is a spoke and $a(v) = c + 1$ let K_2 be a near-triangulation obtained by adding to K a $(c-3)$-edge path between the two hats adjacent to v, and making every internal vertex of this path adjacent to v; and otherwise let $K_2 = K$. For $v' \in V(K_2)$, let $a_2(v') = \infty$ and $b_2(v') = 5$ if $v' \in V(K_2) - V(K)$, let $a_2(v') = a(v')$ and $b_2(v') = b(v')$ for $v' \in V(K) - \{v\}$, and let $a_2(v) = a(v)$ and $b_2(v) = c + 1$.

It follows that (K_1, a_1, b_1) and (K_2, a_2, b_2) are both refinements of (K, a, b); we call them a *complementary pair* of refinements of (K, a, b). Moreover, for any

cartwheel W such that (K, a, b) fits W, if $\gamma_W(v) \leq c$ then (K_1, a_1, b_1) fits an isomorphic copy of W, and if $\gamma_W(v) \geq c+1$ then (K_2, a_2, b_2) fits an isomorphic copy of W; and so it follows that:

(7.2) *Let (K, a, b) be a part, and let $(K_1, a_1, b_1), (K_2, a_2, b_2)$ be a complementary pair of refinements of (K, a, b). If they are both successful then (K, a, b) is successful.*

This constitutes the "branch" mechanism of our branch-and-bound proof of (7.1). We shall have some current part, and if we cannot see directly that it is successful, we choose a complementary pair of refinements, prove individually that they are successful, and infer from (7.2) that our original part is successful.

Now we must explain what we mean by "seeing directly" that the part is successful. Here we use only three kinds of argument. In increasing order of complexity, they are:

Argument 1: symmetry. Our part is a refinement of an isomorphic copy of a part that has already been shown to be successful.

Argument 2: reducibility. For every cartwheel W that the part fits, some good configuration appears in W. For instance, if $\gamma(w) = 7$, and w has neighbours $v_1, \ldots, v_7$ in order, and $a(v_1) = a(v_5) = 5$, and $a(v_2) = a(v_3) = a(v_4) = 6$, we can infer (all the other values of a and b are irrelevant) that one of $\mathrm{conf}(1,4,3)$, $\mathrm{conf}(1,4,5)$, $\mathrm{conf}(1,6,1)$, $\mathrm{conf}(1,5,5)$, $\mathrm{conf}(1,7,5)$ appears. Verifying this involves case-checking, but generally not much, and it can easily be done by hand.

Argument 3: hubcap bounds. Here we argue about $N_{\mathcal{P}}(W)$. Let W be some hypothetical cartwheel that our part fits, with hub of degree k, and in which no good configuration appears. For each spoke v, let $R(v)$ (respectively, $S(v)$) be the sum of values of all passes $P \sim \mathcal{P}$ appearing in W with source v and sink w (respectively, source w and sink v), and let $T(v) = R(v) - S(v)$. We must show that $N_{\mathcal{P}}(W) \leq 0$, that is, that

$$\sum_v T(v) \leq 10(k-6).$$

When this argument is to be applied, we are given a *hubcap*, that is, a list

$$(x_1, y_1), \ldots, (x_k, y_k)$$

of pairs of spokes so that every spoke appears in the list $x_1, y_1, \ldots, x_k, y_k$ exactly twice. For each pair (x_i, y_i), we enumerate all combinations of passes from $\mathcal{P}$ that might appear simultaneously in W, with source either x_i or y_i and sink w. (A set of simultaneously appearing passes must agree with each other on the degree of vertices of W that appear in more than one of them, must be compatible with the given part, and must not force the appearance in W of a good configuration; so such sets are not very big, and are easy to enumerate, particularly when the part is quite highly refined, which is usually the case when this argument is applied.) Of all these combinations we see which has $T(x_i)+T(y_i)$ the largest; and this gives an upper bound on $T(x_i)+T(y_i)$. By summing this over all pairs (x_i, y_i) in the hubcap

(and dividing by two) we obtain an upper bound on $\sum_v T(v)$. So, in summary, the third argument method is: we are given a (carefully selected) hubcap; we use it to compute an upper bound on $\sum_v T(v)$; and deduce that

$$\sum_v T(v) \leq 10(k-6)$$

and hence that the part is successful.

Now we can explain the machine-checkable proof. Let $7 \leq k \leq 11$; we need to show that the trivial part with hub of degree k is successful. At each step of the proof we have a "current" part, that we are trying to show successful. The proof either specifies a complementary pair of refinements of the current part (whereupon it proceeds to tackle them both separately) or it states that the current part can be shown to be successful via one of arguments 1, 2 and 3 (and gives some helpful hints, such as the hubcap for argument 3). Of course there is some book-keeping required, to make sure that every time the process branches both branches are completed, but this is straightforward. Further, more technical details are available with the data.

Acknowledgements

We would like to express our thanks to Dan Younger for several stimulating discussions on the topic of the Appel-Haken proof, and on the desirability of finding a different proof that could more easily be checked. Thanks also to Reinhard Diestel, Tommy Jensen and Jean Mayer for carefully reading the manuscript and detecting several errors, to Tom Fowler for reading the computer programs, and to Bill Cook for his advice on computing and integer programming. Thanks to the referees for carefully checking the paper, and thanks to Gašper Fijavž, and Christopher Carl Heckman for kindly verifying the computer work.

References

1. F. Allaire, "Another proof of the four colour theorem – Part I", *Manitoba Conf. on Numerical Math. and Computing*, Proc. 7th Congressus Numerantium XX, 1977, 3-72.
2. F. Allaire and E. R. Swart, "A systematic approach to the determination of reducible configurations in the four-color conjecture", *J. Combinatorial Theory, Ser. B*, 25 (1978), 339-362.
3. K. Appel and W. Haken, *Every planar map is four colorable*, *A.M.S. Contemporary Math.* 98 (1989).
4. K. Appel and W. Haken, "Every planar map is four colorable. Part I. Discharging", *Illinois J. Math.* 21 (1977), 429-490.
5. K. Appel, W. Haken and J. Koch, "Every planar map is four colorable. Part II. Reducibility", *Illinois J. Math.* 21 (1977), 491–567.

6. A. Bernhart, "Another reducible edge configuration", *Amer. J. Math.* 70 (1948), 144-146.

7. G. D. Birkhoff, "The reducibility of maps", *Amer. J. Math.* 35 (1913), 114-128.

8. D. I. A. Cohen, "Block count consistency and the four color problem", manuscript.

9. K. Dürre, H. Heesch and F. Miehe, "Eine Figurenliste zur chromatischen Reduktion", manuscript eingereicht am 15.8.1977.

10. S.J.Gismondi and E.R.Swart, "A new type of 4-colour reducibility", *Congr. Numer.* 82 (1991) 33-48.

11. H. Heesch, *Untersuchungen zum Vierfarbenproblem*, Hochschulskriptum 810/a/b, Bibliographisches Institut, Mannheim 1969.

12. A. B. Kempe, "On the geographical problem of the four colours", *Amer. J. Math.* 2 (1879), 183-200.

13. J. Mayer, "Une propriété des graphes minimaux dans le problème des quatre couleurs", *Problèmes Combinatoires et Théorie des Graphes, Colloques internationaux C.N.R.S. No. 260*, Paris 1978.

14. P. G. Tait, "Note on a theorem in geometry of position", *Trans. Roy. Soc. Edinburgh* 29 (1880), 657-660.

15. H. Whitney and W. T. Tutte, "Kempe chains and the four colour problem", in *Studies in Graph Theory*, Part II (ed. D. R. Fulkerson), Math. Assoc. of America, 1975, 378-413.

Appendix: The unavoidable set of reducible configurations.