Why No Planar Map Needs More Than Four Colours

Bhupinder Singh Anand

Update of April 4, 2015.

Abstract

Although the Four Colour Theorem is passé, we give a simple model-theoretic proof that explains why four colours suffice to chromatically differentiate any set of contiguous, simply connected, and bounded planar spaces.

1 Introduction: A historical perspective

It would probably be a fair assessment that the mathematical significance of any new proof of the Four Colour Theorem would today be perceived as lying not in any ensuing theoretical or practical utility of the Theorem per se, but in whether the proof can address the philosophically unsatisfying, and despairing ([Ty79], [Sw80], [Go08], [Cl01]) lack of, mathematical ‘insight’, ‘simplicity’ and ‘elegance’ in currently known proofs of the Theorem (eg. [AH77], [AHK77], [RSST], [Go08])—an insight and simplicity that this investigation hopes to address and reveal.

To place the investigation in a wider historical perspective, we note this extract from Robertson, Sanders, Seymour, and Thomas’s 1995-dated (apparently pre-publication) web-survey of their 1997 proof ([RSST]) of The Four Colour Theorem:

“The Four Color Problem dates back to 1852 when Francis Guthrie, while trying to color the map of counties of England noticed that four colors sufficed. He asked his brother Frederick if it was true that any map can be colored using four colors in such a way that adjacent regions (i.e. those sharing a common boundary segment, not just a point) receive different colors. Frederick Guthrie then communicated the conjecture to DeMorgan. The first printed reference is due to Cayley in 1878 ([Ca79]).

A year later the first ‘proof’ by Kempe appeared; its incorrectness was pointed out by Heawood 11 years later. Another failed proof

1 Accessible at http://people.math.gatech.edu/ thomas/FC/fourcolor.html; see also [Th88], [Cl01], and the survey of The Four Colour Theorem by Leo Rogers on the University of Cambridge’s Millenium Mathematics Project weblog ‘NRICH’ at http://nrich.maths.org/6291.
is due to Tait in 1880; a gap in the argument was pointed out by Petersen in 1891. Both failed proofs did have some value, though. Kempe discovered what became known as Kempe chains, and Tait found an equivalent formulation of the Four Color Theorem in terms of 3-edge-coloring.

The next major contribution came from Birkhoff whose work allowed Franklin in 1922 to prove that the four color conjecture is true for maps with at most 25 regions. It was also used by other mathematicians to make various forms of progress on the four color problem. We should specifically mention Heesch who developed the two main ingredients needed for the ultimate proof - reducibility and discharging. While the concept of reducibility was studied by other researchers as well, it appears that the idea of discharging, crucial for the unavoidability part of the proof, is due to Heesch, and that it was he who conjectured that a suitable development of this method would solve the Four Color Problem.

This was confirmed by Appel and Haken in 1976, when they published their proof of the Four Color Theorem [1.2] (sic).

Why a new proof?

There are two reasons why the Appel-Haken proof is not completely satisfactory.

- Part of the Appel-Haken proof uses a computer, and cannot be verified by hand, and
- even the part that is supposedly hand-checkable is extraordinarily complicated and tedious, and as far as we know, no one has verified it in its entirety.”

The motivation for the present investigation, however, lies in the author’s 1982 handwritten noting—“Proof?”—in the margin on p.169 of his 1981 edition of Ian Stewart’s fascinating survey, ‘Concepts of Modern Mathematics’ ([St81]), where Stewart notes in connection with “…the famous (infamous?) four-colour problem” that (referring to the 4-colourability of a map):

“…it is possible to show that no map can have 5 faces, each touching the other 4 (along an edge). However, this does not prove that 4 will always be enough.”

The current investigation is a seriously belated attempt to understand and prove the former remark from first principles; which apparently disproves the latter!

1.1 Neighbourly planar 4-spaces: DeMorgan’s Theorem

Definition 1 Simply connected planar space: A bounded planar space is simply connected if, and only if, every path between any two points of the space can be continuously transformed, staying within the boundary of the space, into any other such path while preserving the two endpoints in question.

2We note that topologically the figure eight, i.e. 8, is not simply connected.
Definition 2 Neighbourly: A set of simply connected planar spaces is neighbourly if, and only if, any two members share a common boundary that is not a point-set (i.e., the boundary is not of zero length).

Theorem 1 DeMorgan’s Theorem: Any set of five or more contiguous, simply connected, and bounded spaces is not neighbourly.

Proof: Let \{A, B, C, D\} be a neighbourly set of four contiguous, simply connected, bounded and chromatically differentiated planar spaces. Without any loss of generality for the central argument of this investigation, any geometric model of such a set can be treated as topologically equivalent to Fig.1.

We begin by assuming that there is a simply connected, bounded, space \(E\) which can be constructed, without crossing the boundary of \(A \cup B \cup C \cup D\), by the transformations \(A \rightarrow A^*, B \rightarrow B^*, C \rightarrow C^*\) and \(D \rightarrow D^*\) (see Figs. 2-4), such that \(\{A^*, B^*, C^*, D^*, E\}\) is also a neighbourly set of contiguous, simply connected, bounded and chromatically differentiated spaces.

We note that \(E\) must then have at least one point in common with each of \(A, B, C\) and \(D\), say \(a, b, c, d\) respectively (as indicated in Figs. 2-4: Proposed).

The boundary of any such simply connected space \(E\) (assumed non-empty) must therefore pass through each of some such \(a, b, c, d\) (each of which—except one, say \(c\)—may also lie on the boundary of the parent space, and therefore need not all be distinct).

We note that there are only the \(3C_2 = 3\) possible, topologically distinct, configurations of the boundary of \(E\) with respect to \(A^*, B^*, C^*\) and \(D^*\), as shown in Figs. 2-4 (Final) respectively:

(i) \(abeda \leftrightarrow bcdab \leftrightarrow cdabc \leftrightarrow dabc \leftrightarrow abeda\)
(ii) \(abeda \leftrightarrow bdca \leftrightarrow dcabd \leftrightarrow cabdc \leftrightarrow abedca\)
(iii) \(adcbca \leftrightarrow adcbad \leftrightarrow bcdab \leftrightarrow cadbc \leftrightarrow adcba\)

However we now note that:

(i) The \(E\)-boundary configuration \(abeda\): In the configuration \(abeda\), the proposed boundary segment \(cda\) of \(E\) (see Fig.2: Proposed) must have a common boundary segment (which may be a point-set, unless otherwise specified) with \(D^*\) (see Fig.2: Final).
Hence at least one of the two boundary segments of $E$, either $abc$ or $bcd$ (see Fig.2: Proposed), can have no point in common with $D^*$, and is not a point-set, if $E$ is simply connected (see Fig.2: Final).

Fig.2: The boundary configuration of $E$: $a \rightarrow b \rightarrow c \rightarrow d \rightarrow a$.

Whence we conclude that at least one pair of spaces $(D^*, B^*)$, $(D^*, C^*)$, or $(E, A^*)$ cannot have any common, non-zero length, boundary; such a pair may thus be similarly coloured in any map (see Fig2: Final).

(ii) The $E$-boundary configuration $abdca$: In the configuration $abdca$, the proposed boundary segment $bdc$ of $E$ (see Fig.3: Proposed) must have a common boundary segment (which may be a point-set, unless otherwise specified) with $D^*$ (see Fig.3: Final).

Hence at least one of the two boundary segments of $E$, either $cab$ or $abd$ (see Fig.3: Proposed), can have no point in common with $D^*$, and is not a point-set, if $E$ is simply connected (see Fig.3: Final).

Fig.3: The boundary configuration of $E$: $a \rightarrow b \rightarrow d \rightarrow c \rightarrow a$.

Whence we conclude that at least one pair of spaces $(D^*, A^*)$, $(D^*, B^*)$, or $(E, C^*)$ cannot have any common, non-zero length, boundary; such a pair may thus be similarly coloured in any map (see Fig3: Final).
(iii) The E-boundary configuration adbca: In the configuration adbca, the proposed boundary segment adb of E (see Fig.4: Proposed) must have a common boundary segment (which may be a point-set, unless otherwise specified) with D* (see Fig.4: Final).

Hence at least one of the two boundary segments of E, either bca or cad (see Fig.4: Proposed), can have no point in common with D*, and is not a point-set, if E is simply connected (see Fig.4: Final).

![Proposed vs Final](image)

Fig.4: The boundary configuration of E: a → d → b → c → a.

Whence we conclude that at least one pair of spaces (D*, C*), (D*, A*), or (E, B*) cannot have any common, non-zero length, boundary; such a pair may thus be similarly coloured in any map (see Fig4: Final).

Since the above cases are exhaustive, we conclude that there can be no neighbourly set of five contiguous, simply connected, bounded and chromatically differentiated planar spaces such as \{A*, B*, C*, D*, E\}.

Moreover, since we have nowhere assumed that a, b, c, d are distinct, nor that—except c—they are not boundary points of their corresponding spaces, the above argument would hold (with obvious modifications) for the general case where the construction of E results in the transformations A → \{A', A''\}, B → \{B', B''\}, and D → \{D', D''\}, where we note that any or all of \{A'', B'', D''\} may be empty spaces.

Reason: Any such pair of spaces \{A', A''\}, for instance, would be spatially separated and retain their original colour even if A' ∩ A'' is not empty (i.e., even if A' ∩ A'' is a point set at the most).

The theorem follows.

1.2 Extending the argument to non-neighbourly planar n-spaces: The Four Colour Theorem Overview

We now extend the above argument and illustrate, in the following Figs.5 to 8, why we cannot merge proper sub-spaces of four or more chromatically differentiated spaces of any, necessarily 4-coloured, n-space map so as to yield a map of n + 1 contiguous, simply connected, and bounded planar spaces which necessarily requires introduction of a 5th colour.
I: Topological model of any 4-coloured map

Model of any set of $n$ non-neighbourly spaces in any 4-coloured map

5: Original map: $n$ spaces; 4 colours

$\begin{tabular}{|c|c|c|c|c|}
\hline
Space code & F & A & B & C & D \\
\hline
Colour code & m & 1 & 2 & 3 & 4 \\
\hline
\end{tabular}$

$m$: 4-colour scheme of sub-map $F$; $n-4$ spaces; code: $m$ (initial), $m^*$ (after merger)

Fig.5: A 4-coloured map of $n$ spaces, where $\{A, B, C, D\}$ is any non-neighbourly set of 4 contiguous, simply connected, bounded, and chromatically differentiated spaces; and $F$ is a sub-map of $n-4$ contiguous, simply connected, bounded, and chromatically differentiated spaces. We note that the following representations remain valid even if $A$ and $B$ share the same colour. We also note in what follows that $C$ is always selected so that $c$ is a boundary point of $E$, and an interior point of $C$.

II: Remaps of E-boundary configuration $abcda – adcba$

6.1a: E-boundary $abcda – adcba$

$\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
Space & F & A & B & C & D & E \\
\hline
Colour & m & 1 & 2 & 3 & 4 \\
\hline
\end{tabular}$

$\begin{tabular}{|c|c|c|c|c|}
\hline
Space & F* & A* & B* & C* \\
\hline
Colour & m & 1 & 2 & 3 \\
\hline
\end{tabular}$

$m$: 4-colour scheme of sub-map $F$; $n-4$ spaces; code: $m$ (initial), $m^*$ (after merger)

6.1b: E-boundary $abcda – adcba$

$\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
Space & F* & A* & B* & C* & D* & E \\
\hline
Colour & m & 1 & 2 & 3 & 4 & 5 \\
\hline
\end{tabular}$

$m$: 4-colour scheme of sub-map $F$; $n-4$ spaces; code: $m$ (initial), $m^*$ (after merger)
Fig.6-1: Remap $A_1 \rightarrow A^*1; B_2 \rightarrow B^*2; C_3 \rightarrow C^*3; D_4 \rightarrow D^*4; E_5 \rightarrow E^5; Fm \rightarrow F^*m$. Note that only proper sub-spaces of $F$ which abut $A$, $B$ and $D$ are merged into $E$, whence $F$ and $F^*$ remain equivalent both topologically and chromatically. Also that, since $E$ must pass through some interior point, $C$ is selected so that $c$ is always such an interior point.

6.2a: E-boundary $abcda \rightarrow adcba$

First remap: $n+1$ spaces; 5 colours

6.2b: E-boundary $abcda \rightarrow adcba$

Second remap: $n$ spaces; 4 colours

6.3a: E-boundary $abcda \rightarrow adcba$

Second remap: $n$ spaces; 4 colours

6.3b: E-boundary $abcda \rightarrow adcba$

Final remap: $n+1$ spaces; 4 colours

$m$: 4-colour scheme of sub-map $F$; $n-4$ spaces; code: $m$ (initial), $m^*$ (after merger)

Fig.6-2: Remap $A^*1 \rightarrow A^*1; B^*2 \rightarrow B^*2; C^*3 \rightarrow C^*3; D^*4 \rightarrow D^*4; E5 \rightarrow E4; F^*m \rightarrow F^*m$.

Fig.6-3: Remap $A^*1 \rightarrow A^*1; B^*2 \rightarrow B^*2; C^*3 \rightarrow C^*3; D^*4 \rightarrow D^*2; E4 \rightarrow E4; F^*m \rightarrow F^*m$. 

7
III: Remaps of E-boundary configuration *abdca* – *acdba*

7.1a: E-boundary *abdca* – *acdba*

*Proposed remap: n + 1 spaces; 4 colours*

![Diagram of E-boundary *abdca* – *acdba*](image1)

- **Space**: F, A, B, C, D
- **Colour**: Code
- **Code**: m, 1, 2, 3, 4

---

7.1b: E-boundary *abdca* – *acdba*

*First remap: n + 1 spaces; 5 colours*

![Diagram of E-boundary *abdca* – *acdba*](image2)

- **Space**: F*, A*, B*, C*, D*, E
- **Colour**: Code
- **Code**: m, 1, 2, 3, 4, 5

---

7.2: E-boundary *abdca* – *acdba*

*First remap: n + 1 spaces; 5 colours*

![Diagram of E-boundary *abdca* – *acdba*](image3)

- **Space**: F*, A*, B*, C*, D*, E
- **Colour**: Code
- **Code**: m, 1, 2, 3, 4, 5

---

7.2b: E-boundary *abdca* – *acdba*

*Second remap: n spaces; 4 colours*

![Diagram of E-boundary *abdca* – *acdba*](image4)

- **Space**: F*, A*, B*, C*, D*, E
- **Colour**: Code
- **Code**: m, 1, 2, 3, 4

---

`m`: 4-colour scheme of sub-map F; n – 4 spaces; code: m (initial), m* (after merger)

**Fig.7-1**: Remap A1 → A*1; B2 → B*2; C3 → C*3; D4 → D*4; E → E5; Fm → F*m.

**Fig.7-2**: Remap A*1 → A*1; B*2 → B*2; C*3 → C*3; D*4 → D*4; E5 → E4; F*m → F*m.
7.3a: E-boundary abdca – acdba
Second remap: n spaces; 4 colours

---

7.3b: E-boundary abdca – acdba
Final remap: n + 1 spaces; 4 colours

---

8.1a: E-boundary abdca – acdba
Proposed remap: n + 1 spaces; 4 colours

---

8.1b: E-boundary abdca – acdba
First remap: n + 1 spaces; 5 colours

---

m: 4-colour scheme of sub-map F; n − 4 spaces; code: m (initial), m* (after merger)

---

IV: Remaps of E-boundary configuration abdca – acdba

---

9
8.2a: E-boundary adbca – acbda

First remap: \( n + 1 \) spaces; 5 colours

8.2b: E-boundary adbca – acbda

Second remap: \( n - 1 \) spaces; 4 colours

8.3a: E-boundary adbca – acbda

Second remap: \( n - 1 \) spaces; 4 colours

8.3b: E-boundary adbca – acbda

Final remap: \( n + 1 \) spaces; 4 colours

Explanatory Remarks

1. The basis of the argument here is that, in any map, a fifth colour is only needed at some precise 4-coloured configuration when the
creation of a new region $E$, by merging some proper sub-regions of existing regions, results in a boundary of $E$ that must pass through the four points $a, b, c, d$ within four differently coloured regions.

2. Without loss of generality, the four differently coloured regions can be represented topologically by Fig.5.

3. We note that the following induction argument, illustrated by Figs.1-8, remains valid even if the spaces $A$ and $B$ share the same colour in Figs.6-8 and the $4^{th}$ colour is one of the areas in $F$ through which $E$ passes. Further, by the stipulation that $F$ and $F^*$ remain topologically equivalent, the argument does not depend on the specific chromatic differentiation of $F$ and $F^*$.

4. Note that $C$ is simply the topological equivalent of an all-encompassing ocean. Any region of the map can be selected as $C$, so long as it is one of the regions through which the boundary of $E$ will pass through an interior point $c$.

5. There are now only 3 different topological configurations for the boundary of $E$ that passes through these four points, Figs.6.1, 7.1 and 8.1.

6. Note that the boundary of $E$ crosses a bunch of regions in $F$, so that $F$ transforms into $F^*$ (in Figs.6.1, 7.1 and 8.1). DeMorgan’s Theorem takes care of the case where $F$ is empty.

7. However, the boundary of $E$ can only pass through some internal points, or along the boundaries, of these regions in $F$ since, by stipulation, creation of $E$ cannot reduce the number of regions in $F$.

8. After the merging of regions in Figs.6.2, 7.2 and 8.2, the boundary of $E$ that abuts $F^*$, and passes through the regions that it invaded, separates the same colours in $E$ and $F^*$ as the original boundary that abutted these regions separated in $D$ and $F$ (since $F$ and $F^*$ are stipulated as topologically and chromatically equivalent).

9. However, the number of regions is now reduced by 1 or 2.

10. Hence the reduced map can be 4-coloured by the Induction Hypothesis.

11. In the reduced map (newly re-coloured to yield $A^*1$ from $A1$, ..., $F^*m^*$ from $F^*m$ if needed as in 8.3), there is now a region abutting $E$ that no longer requires introduction of a fifth colour (as shown in Figs.6.3, 7.3 and 8.3).

12. Each configuration now contains two of the original four regions that are spatially separated by the creation of the boundary
of $E$, and can have the same colour after applying the Induction Hypothesis.

13. Note that the above argument involves recognising the physical separation of areas under a sub-division, whereas in the conventional graph-theoretical perspective of the four colour problem areas are represented as nodes, boundaries disappear, and colours are represented as edges joining nodes.

2 Why no planar map needs more than four colours: The Four Colour Theorem

We conclude from the argumentation of Theorem 1, when applied to the exhaustive set of configurations in Figs.5-8, that:

Corollary 1 The Four Colour Theorem: No more than four colours are required to colour any set of contiguous, simply connected, and bounded planar spaces such that no two spaces with a common non-zero boundary have the same colour.

Proof: It follows from Theorem 1 that any map $M_5$ of 5 contiguous, simply connected, bounded and chromatically differentiated, planar spaces can be 4-coloured, by which we mean that no more than four colours are required to colour $M_5$ such that no two spaces with a common non-zero boundary share the same colour (or that, equivalently, the two spaces are chromatically differentiated).

Assume that for all $n \geq i \geq 5$, any map $M_i$ of $i$ contiguous, simply connected, bounded and chromatically differentiated, planar spaces can be 4-coloured.

It also then follows from the argumentation of Theorem 1 that we cannot construct (see the exhaustive configurations in Figs.5-8) an additional contiguous, simply connected, bounded and chromatically differentiated, planar space $E$ in $M_n$ that:

(a) has a common non-zero boundary with four points in four contiguous, simply connected, bounded and chromatically differentiated, planar spaces; and,

(b) requires introduction of a $5^{th}$ colour because $E$ must necessarily have a common non-zero boundary with four contiguous, simply connected, bounded and differently coloured, planar spaces of $M_n$ as in Figs.6-1, 7-1 and 8-1.
Reason: If we assume that we can construct such an additional contiguous, simply connected, bounded and chromatically differentiated, planar space $E$ in $M_n$ which cannot be 4-coloured (thus yielding a map $M_{n+1}$ of $n+1$ contiguous, simply connected, bounded and chromatically differentiated, planar spaces that necessarily requires 5 colours) then $E$ would be the only contiguous, simply connected, bounded and chromatically differentiated, planar space in $M_{n+1}$ that requires a 5th colour as in Figs.6-1, 7-1 and 8-1.

It would follow that:

(i) Either some subset $m_5$ of $M_{n+1}$ that contains $E$ and consists of 5 contiguous, simply connected, bounded and chromatically differentiated, planar spaces cannot be 4-coloured since $m_5$ is neighbourly; which would contradict Theorem 1;

(ii) Or every subset $m_5$ of $M_{n+1}$ that contains $E$ and consists of 5 contiguous, simply connected, bounded and chromatically differentiated, planar spaces is not neighbourly; and can:

- first be displayed topographically as in Fig.5; and
- second shown to contain two simply connected, and bounded planar spaces that are spatially separated and share the same colour as in Figs.6-1, 7-1 and 8-1.

Hence either or both of these spaces can be merged into an adjoining space without affecting the essential 5-colourability of $M_{n+1}$.

The latter now consists of either $n$ or $n-1$ contiguous, simply connected, bounded and chromatically differentiated, planar spaces as in Figs.6-2, 7-2 and 8-2, which can be 4-coloured by our induction hypothesis.

The merged spaces can then be restored, and chromatically differentiated, to yield a 4-coloured map of $M_{n+1}$ as in Figs.6-3, 7-3 and 8-3, which would contradict our assumption that $M_{n+1}$ cannot be 4-coloured.

Since any map of less than 5 contiguous, simply connected, bounded and chromatically differentiated, planar spaces can be 4-coloured, we conclude by finite induction that every map of contiguous, simply
connected, bounded and chromatically differentiated, planar spaces can be 4-coloured, as shown in Figs.6-3, 7-3 and 8-3, where we note that the induction argument illustrated in Figs.1-8 is valid even if the spaces $A$ and $B$ share the same colour in Figs.6-8. The corollary follows.

References


MSC: 05C15

This is a work in progress, and a copy of the most recent update is accessible here. The author invites, and appreciates, critical comments or observations.

Author’s postal address: #1003 B Wing, Lady Ratan Tower, Dainik Shivner Marg, Worli, Mumbai - 400 018, Maharashtra, India. Mbl: +91 93 22 591328. Email: bhup.anand@gmail.com. Weblog: Foundations of Mathematics, Logic & Computability.