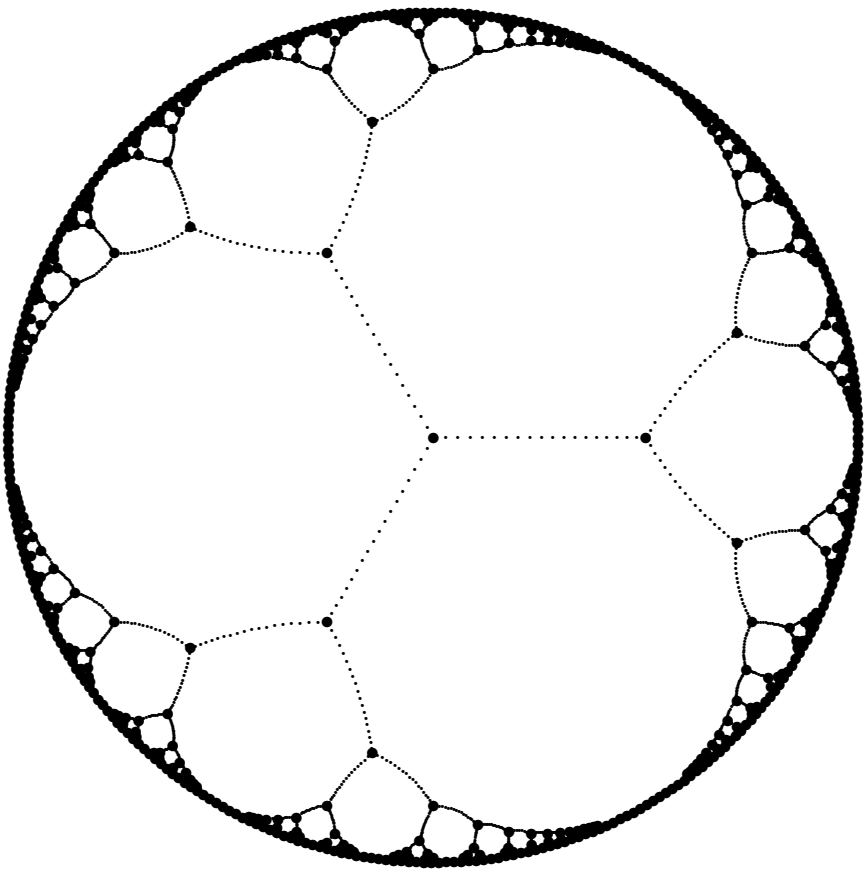
Topics in Trivalent Graphs

Marijke van Gans

31 January 2007



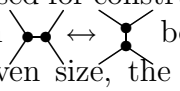
A thesis submitted to
The University of Birmingham
for the degree of
Doctor of Philosophy

Abstract

Chapter 0 details the notation and terminology used.

Chapter 1 introduces the usual linear algebra over $\mathbb{F}_2$ of edge space $\mathbf{E}$ and its orthogonal subspaces $\mathbf{Z}$ (cycle space) and $\mathbf{Z}^\perp$ (cut space). **Reduced vectors** are defined as elements of the quotient space $\mathbf{E}/\mathbf{Z}^\perp$. Reduced vectors of edges give a simple way of characterising edges that are bridges (their reduced vector is null) or 2-edge cuts (their vectors are equal), and also of spanning trees (the edges outside the tree are a basis) and form to the best of my knowledge a new approach. They are also useful in later chapters to describe Tait colorings, as well as cycle double covers. Perhaps the most important property of $\mathbf{E}/\mathbf{Z}^\perp$ is the **Unique graph theorem**: unlike in $\mathbf{E}$, a list of which reduced vectors are edges uniquely determines graph structure (if edge connectivity is high enough; that covers certain "solid" components every trivalent graph can be decomposed into).

Chapter 2 gives a brief intoduction to graph embeddings and planar graphs.

Chapter 3 deals specifically with trivalent graphs, listing some of the ways in which they are different from graphs in general. Results here include two versions of **Bipolar growth theorem** which can be used for constructive proofs, and (after defining "halftrees" and a "flipping" operation  between them) a theorem enumerating the set $\boldsymbol{C}_n$ of halftrees of a given size, the **Caterpillar theorem** showing $\boldsymbol{C}_n$ is connected by flipping, and the **Butterfly theorem** derived from it. Graphs referred to here as **solid** are shown to play an important structural rôle.

Chapter 4 deals with the 4-coloring theorem. The first half shows the older results in a unified light using edge spaces over $\mathbb{F}_4$. The second half applies methods from **coding theory** to this. The 4-color theorem is shown to be equivalent to a variety of statements about cycle-shaped words in codes over $\mathbb{F}_4$ or $\mathbb{F}_3$, many of them tantalisingly simple to state (but not, as yet, to prove).

Chapter 5 deals with what has been variously called polyhedral decompositions and (specifically for those using cycles) cycle double covers, as in the cycle double cover conjecture. The more general concept is referred to as a **map** in this paper, and identified with what is termed here **cisness structures**, which is a new approach. There is also a simpler proof of a theorem by Szekeres. Links with the subject of the previous chapter are identified, and some approaches towards proving the conjecture suggested.

Several planned **appendices** were left out of the version submitted for examination because they would make the thesis too big, and/or were not finished. Of the ones that remain, appendix **H** (on embedding infinite 4- and 3-valent trees X_∞ and Y_∞ in the hyperbolic plane) now seems disjointed from the body of the text (a planned appendix dealt with colorings of finite graphs as the images of homomorphisms from embeddings of Y_∞). Appendix **B** enumerates cycle maps (cycle double covers) on a number of small graphs while appendix **D** investigates $\dim \mathbf{Z} \cap \mathbf{Z}^\perp$.

Concrete

Nodes are in SMALL CAPS, edges (and halfedges) in lower case $\gamma\rho\varepsilon\varepsilon\kappa$.
Integers and field elements are in lower case *italic*, mostly.
Spaces (linear codes) over $\mathbb{F}_2$ are in **UPPER** and their vectors in **lower** case bold.
Spaces (linear codes) over $\mathbb{F}_4$ are in ***UPPER*** and their vectors in ***lower*** case bold italic; sets (codes) that aren't subspaces are written like $\mathcal{THIS}$.

n nodes, m edges, k components; $o := m - n + k$; if trivalent $n/2 =: h := m/3$.

We is the mathematical community at large when the subject of verbs such as "call"; it is the readers and author together in phrases like "... will come across an example of this in the next section". *I* is used when referring to me.

This paper was written in LaTeX with a sprinkling of plain TeX.
The contents of the LaTeX `picture` environments were designed with $\begin{smallmatrix}\mathrm{d}raw\\ \mathrm{p}i\not{c}t\end{smallmatrix}$.

Thanks

- to Rob T. Curtis and the gang for putting up with me all this time.
- to the examiners for their helpful corrections and suggestions.
- to Rudy J. List and Bernard B. Beard for believing in me.
- to Bill Daly for rekindling my interest in trivalent graphs.
- and everybody else at `CIS:SCIMATH` for all the great stuff, 1994–2004.
- to h for, among other things, feeding me when other funding dried up.

Contents

Appendices B, **D** and **H** are in separate files (with their own lists of references).

0 Graphs

0.0 Terminology

Graph theory nomenclature being notoriously variable, it is perhaps best to briefly define the terms as used in this paper. A **graph** (formerly **simple graph**) consists of a set $\mathcal{N}$ of items called **nodes** (*vertices, points* etc.), with a set $\mathcal{E} \subseteq$ the set of (unordered) pairs of nodes. The pairs that are in $\mathcal{E}$ are called **edges** (*links, arcs* etc.) and an edge is said to "run between" its two nodes, its "endpoints" (cf. the customary depiction of nodes as dots on the page and edges as line segments joining them). A node P and an edge ε are **incident** if P is one of the endpoints of ε. Note that, $\mathcal{E}$ being a set, there can be at most one edge between any given pair of nodes. Following several authors, let $n := |\mathcal{N}|$ and $m := |\mathcal{E}|$.

In the older usage, a graph is allowed to have more than one edge between the same two nodes; such a thing is now called a **multigraph**. Here an edge can't be defined uniquely as a pair of nodes anymore so we must use an $\mathcal{E}$ independent from $\mathcal{N}$, and the graph structure (incidence between certain edges and nodes) must be expressed more generally: we have an **incidence structure** $\Im \subseteq \mathcal{N} \times \mathcal{E}$ such that for every edge ε there are precisely two ordered pairs $(\mathrm{P}, \varepsilon)$ and $(\mathrm{Q}, \varepsilon)$ in $\Im$, the ones using ε's endpoints P and Q.

An even wider concept, a **pseudograph**, allows the "two" endpoints of an edge to be one and the same node (such an edge is called a **loop**). This breaks not only the idea of an edge as pair of nodes, but also the notion of a graph as an incidence structure. I will not have occasion to use pseudographs here (other than to mention why they are excluded from certain theorems).

Incidence structures can be written out as matrices (using rows for edges and columns for nodes, say) with entries 1 where there is incidence and 0 where there isn't. Those representing [multi]graphs are special in that each row has exactly two 1s. Generally, an incidence structure where each "edge" (the general term is *block*) is incident with the same number u of "nodes" (*points*) is called **u-uniform**. In short, a multigraph is a 2-uniform incidence structure.

The dual of an incidence structure (obtained by transposing the matrix) is again an incidence structure. The dual notion of uniform is **regular**. A v-regular multigraph is an incidence structure that is both v-regular and 2-uniform. In graph theory the number of edges incident to a node (counting loops twice) is known as the **valency** (or *degree*) of that node. Hence a v-regular graph is also known as **v-valent** graph.

A **trivalent** (3-valent) graph is often called a *cubic graph* (probably after the cube, one such graph).

An **empty** graph is a 0-regular one (it has nodes but no edges). A **null graph** has no nodes (and hence no edges either); many authors exclude this possibility from the definition of a graph, but Tutte [Tut84] and Diestel [Die05] allow it.

We saw that, formally, the view that an edge *is* a special pair of nodes is tenable for simple graphs, where the pair of endpoints uniquely identifies the edge; in a multigraph an edge can only ever *have* a pair of nodes as endpoints.

What about doing it the other way round, associating a node with a set of edges (those it is incident with)? This time, two nodes being incident with exactly the same (nonempty) set of edges automatically makes those edges unavailable for any other nodes, so the nodes form a separate component with two nodes. The only *connected* [multi]graphs for which the view that a node *is* (rather than *has*) a set of edges is untenable are thus •—•, •⟷•, •⟹• and so on; the only such [multi]graphs the ones that have these as a component, or more than one • component.

In the next chapter (on spaces) we will effectively do just that, treat nodes (and other features) as the sets of their edges.

0.1 Connectivity

Any partition of $\mathcal{N}$ into k sets $\mathcal{N}_i$ induces a partition of $\mathcal{E}$ into $\frac{k(k+1)}{2}$ (possibly empty) sets $\mathcal{E}_{ij}$, where $\mathcal{E}_{ij} = \mathcal{E}_{ji}$ denotes the set of edges linking a node in $\mathcal{N}_i$ to one in $\mathcal{N}_j$ (in particular, $\mathcal{E}_{ii}$ is the set of edges both whose endpoints are in $\mathcal{N}_i$). This notation will be used throughout this paper.

A **subgraph** G_0 of G is a graph with some $\mathcal{N}_0 \subseteq \mathcal{N}$ as node set and some $\mathcal{E}_0 \subseteq \mathcal{E}$ as edge set, with incidences as in G. As every edge needs both its endpoints, this implies $\mathcal{E}_0 \subseteq \mathcal{E}_{00}$. An **induced subgraph** is one where $\mathcal{E}_0$ is all of $\mathcal{E}_{00}$; thus any subset $\mathcal{N}_0$ of $\mathcal{N}$ uniquely defines the subgraph of G **induced by** $\mathcal{N}_0$.

A **cut** is the $\mathcal{E}_{\bullet\circ}$ belonging to a bi-partition of $\mathcal{N}$ into some $\mathcal{N}_\bullet$ and $\mathcal{N}_\circ$ (think black and white nodes). Any $\mathcal{N}_\bullet \subseteq \mathcal{N}$ determines such a bi-partition of $\mathcal{N}$ and hence a cut $\mathcal{E}_{\bullet\circ}$ (throughout the paper $\mathcal{N}_\circ$ will denote $\mathcal{N} \setminus \mathcal{N}_\bullet$ unless stated otherwise). Each of the complementary subsets $\mathcal{N}_\bullet$ and $\mathcal{N}_\circ$ determines the same cut.

A **null cut** $\mathcal{E}_{\bullet\circ}$ is one where one of $\mathcal{N}_\bullet$ and $\mathcal{N}_\circ$ is $\{\}$ so the other one is $\mathcal{N}$.

A **bipartite** graph is one that has a non-null cut $\mathcal{E}_{\bullet\circ}$ such that $\mathcal{E}_{\bullet\circ} = \mathcal{E}$.

A graph with a non-null cut that is nevertheless **empty** (as a set, i.e. without edges) is called **disconnected**. Let $\textsc{p} \approx \textsc{q}$ be the relation "there does not exist an empty cut $\mathcal{E}_{\bullet\circ}$ such that $\textsc{p} \in \mathcal{N}_\bullet$ and $\textsc{q} \in \mathcal{N}_\circ$". The relation is clearly reflexive ($\textsc{p} \approx \textsc{p}$) and symmetric ($\textsc{p} \approx \textsc{q}$ iff $\textsc{q} \approx \textsc{p}$). It is also transitive (if $\textsc{p} \approx \textsc{q}$ and $\textsc{q} \approx \textsc{r}$ then $\textsc{p} \approx \textsc{r}$), by contradiction: any bipartition that puts p in $\mathcal{N}_\bullet$ and r in $\mathcal{N}_\circ$ without there being any edges in $\mathcal{E}_{\bullet\circ}$ could only put q in the same part as one of p and r, and then there would be an empty cut between q and the other one.

So $\approx$ is an equivalence relation. The equivalence classes $\mathcal{N}_0$, $\mathcal{N}_1$, ... in which it partitions the nodes induce subgraphs known as the **components** of the graph. For this particular partition, all $\mathcal{E}_{ij}$ (for $i \neq j$) are empty by definition, and it is the finest such partition. So every edge of the graph falls in some component's $\mathcal{E}_{ii}$.

The cardinality of the smallest non-null cut is known as the **edge connectivity** κ' of the graph. The phrase k**-edge-connected** means κ' is *at least* k.

The null graph and the 1-node graph only have null cuts so their edge connectivity is not well defined this way (and customarily taken as 0). For other graphs,

- Edge connectivity 0 means the existence of one or more non-null 0-edge cuts (the graph is **disconnected**).
- Edge connectivity 1 means the existence of one or more 1-edge cuts. Such an edge is now usually known as a **bridge** (beware: the word has also been used in a different sense, e.g. [Ore67]), or sometimes *isthmus*.
- Edge connectivity 2 means the existence of one or more 2-edge cuts. Such pairs of edges are not necessarily disjoint (consider the triangle graph). It is not hard to see the relation $\alpha \equiv \beta$ meaning "α and β form a cut, or $\alpha = \beta$" is an equivalence relation; we will see an algebraic proof of this below. I will call an equivalence class of $\equiv$ (a set of edges any two of which form a cut) a **cobridge** if it contains more than one edge.

Lemma: $\kappa' \leq \delta$ where δ is the smallest valency occurring in the graph.

Proof: Let v have valency δ. Take $\mathcal{N}_\bullet = \{\textsc{v}\}$, now $|\mathcal{E}_{\bullet\circ}| = \delta$. ■

So in **trivalent graphs** $\kappa' \leq 3$. Nevertheless, it would be useful to call them

- honorary 4-edge-connected in some sense if cuts around a single node are the *only* cuts of 3 or fewer edges. In fact this is such a useful notion that we need a shorter term for it. I suggest calling such trivalent graphs **solid**[0].

[0] As with almost any adjective in the language, there is always a chance it is already in use for

- Perhaps also honorary 5-edge-connected if moreover among the other cuts those around the endpoints of an edge are the *only* cuts of 4 or fewer edges.

We saw any graph with more than 1 node (so 0-edge-connected) fall apart in one or more (1-edge-connected) components, where every node and every edge belongs to one of the components. In the same way, those 1-edge-connected components fall apart further into 2-edge-connected structural elements (by removing the bridges); 2-edge connected (sub)graphs fall apart further into 3-edge-connected structural elements (by removing the co-bridges); and so on. Each level of structural element cleanly partitions the nodes, but some edges fall between the elements.

There is also a notion of **node** or **vertex connectivity**, or just **connectivity** for short, called κ. This time the structural elements cleanly partition the edges, but some nodes are shared between them. Node connectivity appears to be deeper in some sense than edge connectivity, but for our purposes (trivalent graphs) its numeric value turns out to coincide with that of edge connectivity. This can be proven quite straightforwardly from first principles (proof, and definition of κ, omitted here for reasons of space).

It will often be useful to exclude **trivial cuts**, that is, $\mathcal{E}_{\bullet\circ}$ for which at least one of $\mathcal{N}_\bullet$ and $\mathcal{N}_\circ$ consists of a single node (this term is singularly apt in the case of trivalent graphs, as the original meaning of *trivial* is "three-way"). Thus we can rephrase the definition of a trivalent graph being **solid** as follows: that all its non-null and non-trivial cuts contain at least 4 edges.[1]

0.2 Cyclicity

A **walk** of length s (or s-walk) consists of a sequence of nodes P_i (for $0 \leq i \leq s$, not necessarily distinct), and a set of edges ε_i (for $0 \leq i < s$, not necessarily distinct) such that each ε_i has endpoints P_i and P_{i+1}. A **trek** [Cam94] is a walk where no two consecutive edges are the same, a **trail** one where all edges are distinct, and a **path** one where all nodes are distinct. Mnemonic: inclusion follows lexicographic order where *paths* $\subset$ *trails* $\subset$ *treks* $\subset$ *walks*. We usually allow $s = 0$.

another graph-theoretical property.

[1] In many ways being solid is a trivalent graph's way of being if not literally 4-edge-connected then at least something very much like it. Solid trivalent graphs even satisfy a version of Menger's theorem: take any two edges that don't share an endpoint, and contract them to two 4-valent nodes. Now there are four independent paths between those nodes. I found a proof of this after submitting the thesis for examination, which is why it does not appear here.

A **closed walk**, **trek**, or **trail** of length s is one where $\mathrm{P}_0 = \mathrm{P}_s$. A **cycle** is the closed counterpart of a path; all nodes are distinct except $\mathrm{P}_0 = \mathrm{P}_s$. Existence of P_0 means we can't allow $s = 0$; a 1-cycle is a loop; a 2-cycle is a double edge; the smallest possible cycle length in a simple graph is therefore 3.

While closed trails (all edges distinct) aren't necessarily cycles (all nodes distinct) in general, they are the same thing in the case of trivalent graphs: it is impossible to visit a node P (in via α, out via β say) and revisit it (in via γ, out via ?) without revisiting an edge as well, when there aren't at least 4 edges at the node.

Let $\mathrm{A} \approx \mathrm{Z}$ denote the existence of a walk from node A to node Z, that is an s-walk for some s, whose $\mathrm{P}_0 = \mathrm{A}$ and $\mathrm{P}_s = \mathrm{Z}$. Two nodes are called **adjacent** (with customary notation $\sim$) if they are the endpoints of some edge, so $\mathrm{A} \approx \mathrm{Z}$ just means the existence of intermediate nodes such that $\mathrm{A} \sim \mathrm{B} \sim \cdots \sim \mathrm{Y} \sim \mathrm{Z}$. Although we don't need this here, it is easy to see the existence of such a walk implies the existence of a path from A to Z (streamline the walk by, whenever it visits a node more than once, cutting out the piece of walk between those occurrences) so that would be an alternative definition of $\mathrm{A} \approx \mathrm{Z}$. Now $\approx$ is transitive (by concatenating walks) and symmetric (walk backwards), and allowing 0-walks makes it reflexive too, so it is an equivalence relation. It is the same $\approx$ we found above, whose equivalence classes induce the **components**.

- From now on, let k denote the number of components of the graph.

A graph with $k = 1$ is called **connected**; we saw one with $k > 1$ is called **disconnected**. The **null graph** with $k = 0$ is neither of these (much like 1 is neither a prime nor a composite number). Calling the null graph connected would upset some of the formulæ we'll encounter next.

If edge β between P and Q is a bridge it cannot occur in a cycle (in a cycle there would be another way to travel from P to Q so $\{\beta\}$ isn't a cut) and if β isn't a bridge it must occur in some cycle (there now must be an alternative route from P to Q, travel out by one route and back by the other, streamline the closed walk to a cycle). So an equivalent definition of **bridge** is: an edge that doesn't feature in any cycle.

A graph where no edge is a bridge is called **bridge-free**. A graph where every edge is a bridge is called a **forest**, that is, a forest is a graph without cycles. Its connected components are called trees so a **tree** is a connected forest.

- From now on, let n denote the number of nodes of a graph,

- and m the number of edges.

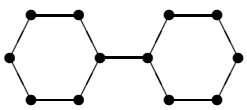
When a graph is trivalent $3n = 2m$ (because for every node there are three edges, but this counts every edge exactly twice), so

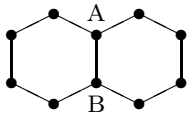
- for trivalent graphs let h denote the integer for which $n = 2h$ and $m = 3h$.

It is an elementary result [Wil02, Die05] that in a tree, $n - m = 1$ (this is one example why we must not call the null graph connected) and conversely, a connected graph with one fewer edge than nodes is a tree. In forests with k components, we have therefore $n - m = k$.

If a graph with k components has just one cycle we find $n - m = k - 1$. If it has two disjoint cycles as in the biphenyl molecule (left in the picture) we find $n - m = k - 2$. And so on, $n - m = k - o$ when the number of *disjoint* cycles is o.

- The number $o = m - n + k$ that appears here is known as the **cyclomatic number**[2], it appears to count the number of cycles.





As cycles start getting entangled things get hairier. For instance in the naphthalene molecule (right in the picture, disregarding double bonds) any two of the three paths from A to B (lengths 5, 1, 5) can be taken together as a cycle (sizes 6, 10, 6) and yet o is still 2 here, not 3. In a sense, o only counts the number of *independent* cycles. We must investigate what sense that is next.

[2]Also known as the first **Betti number** $p_1(G)$; the zeroth Betti number $p_0(G)$ is k.

1 Spaces

1.0 Edge space E

Let $\mathbb{F}$ be any field. An **edge space** is a vector space $\mathbb{F}^m$ one of whose bases consists of vectors $\mathbf{e}_\iota$ that stand in one-to-one correspondence with the edges ι of the graph. Later on we'll come across an example that uses another field, but for the moment let the field be $\mathbb{F}_2 = \{0, 1\}$ with the usual rules of addition and multiplication modulo 2. Let $\mathbf{E}$ stand for the edge space $\mathbb{F}_2^m$.

Vectors in a space over $\mathbb{F}_2$ can also be interpreted as **sets**, in our case subsets of $\mathcal{E}$. A vector $\mathbf{v}$ in $\mathbf{E}$ is $\sum v_\iota \mathbf{e}_\iota$ where v_ι is 0 or 1 (as field element) and we can identify $\mathbf{v}$ with a set of edges $\{\iota \mid v_\iota = 1\}$. Vector addition $\mathbf{u} + \mathbf{v}$ translates to symmetric difference $\mathbf{u} \triangle \mathbf{v} = \mathbf{u} \cup \mathbf{v} \setminus \mathbf{u} \cap \mathbf{v}$ when applied to sets.

In the general theory [SK77, Die05] there is also a node space $\mathbb{F}_2^n$ defined the analogous way. Vectors of edge space are sometimes called *1-chains* and those of node space *0-chains*, with two linear operators between them: the boundary operator ∂ mapping each edge to the sum of its endpoints in node space, and the coboundary operator δ mapping each node to the sum of its incident edges [SK77 Appendix A].

I will use a slightly different approach here, working in edge space throughout. Just as "an edge" is now a single basis vector of our chosen basis (a singleton set), let "a node" be just the vector sum (or set) of its incident edges (δ of its *0-chain*).

1.1 Cut space $\mathbf{Z}^\perp$

Any set of edges is a vector of $\mathbf{E}$, for instance a cut is now just the sum of the edges in the cut. Note this identifies a node with the cut around that node. More generally, for any partition of $\mathcal{N}$ into $\mathcal{N}_\bullet$ and $\mathcal{N}_\circ$ the cut $\mathcal{E}_{\bullet\circ}$ is simply the sum of all the nodes in $\mathcal{N}_\bullet$ because edges of $\mathcal{E}_{\bullet\bullet}$ occur twice in that sum of nodes but $2 = 0$, edges of $\mathcal{E}_{\bullet\circ}$ occur once and survive, while edges of $\mathcal{E}_{\circ\circ}$ don't occur at all in the sum.

Let **cut space** $\mathbf{Z}^\perp$ be the subset of $\mathbf{E}$ consisting of all cuts. Because the cuts are the sums of nodes, the set is the subspace spanned by the nodes. Its dimension is therefore no more than n. It is in fact less because e.g. the nodes in $\mathcal{N}_\bullet$ sum to the same cut as those in $\mathcal{N}_\circ$, so the nodes form a superset of a basis. To determine how much less than n the dimension is, we must find linear dependencies between the

nodes, that is, sets of nodes that sum to 0. But we know the answer already: the only empty cuts are those that partition $\mathcal{N}$ along whole components. So there are k independent linear constraints (nodes in any component sum to 0) and no other (if nodes sum to zero they can have no edge to the outside world). The dimension of $\mathbf{Z}^\perp$ is $n - k$.

1.2 Cycle space Z

We can also interpret a cycle the set of its edges, and so as a vector. Now define

- $\mathbf{X}$ as the set of all $\mathbf{x} = \sum v_\iota \mathbf{e}_\iota$ such that for the edges α, β, $\gamma \ldots$ at any node, an even number are in $\mathbf{x}$ ($\mathbf{x}$ is an *Euler set*): $v_\alpha + v_\beta + v_\gamma \cdots = 0 \pmod 2$.
- $\mathbf{Y}$ as the subset of $\mathbf{E}$ consisting of all sums of edge-disjoint cycles.
- $\mathbf{Z}$ as the subspace of $\mathbf{E}$ spanned by the cycles, that is, we take the vector sum of any collection of cycles whether edge-disjoint or not, obliterating an edge in the sum if it occurs an even number of times.

Lemma: $\mathbf{X} = \mathbf{Y} = \mathbf{Z}$ in finite graphs.

Proof (trivalent case): By definition, $\mathbf{Y} \subseteq \mathbf{Z}$ (sums of disjoint cycles are sums of cycles) and we also have $\mathbf{Z} \subseteq \mathbf{X}$ because every single cycle, if it visits a node, must leave it again by a second edge, and parity of the number of edges used at that node is preserved under addition over $\mathbb{F}_2$.

Lastly, in a *finite* graph we must have $\mathbf{X} \subseteq \mathbf{Y}$ too. This is where trivalence comes in (and we'll only need the result for trivalent graphs). Choose any vector $\mathbf{x}$ in $\mathbf{X}$, any of its edges, and a direction to walk, following edges of $\mathbf{x}$. The walk never ends ($v_\alpha + v_\beta + v_\gamma$ is nonzero as we're on an edge of $\mathbf{x}$ but even, so at least 2) and we never have to choose how to walk ($v_\alpha + v_\beta + v_\gamma$ is even and at most 3 so exactly 2, the edge we came in from and the next edge). The walk not ending means, in a finite graph, that sooner or later we revisit a node we had before. We cannot rejoin the walk mid-walk in a ρ shape ($v_\alpha + v_\beta + v_\gamma$ is never 3) so we rejoin where we started. Remove this cycle from $\mathbf{x}$, repeat, in a finite graph all edges present in $\mathbf{x}$ will eventually be allocated to a cycle. ■ Note how we needed finiteness twice.

It is also true in graphs in general, but now cycles being edge-disjoint no longer implies node-disjoint. **Sketch of proof** in this case: you *do* get ρ shapes when walking, and the trick is to mark the stalk of the ρ as not yet walked and only harvest the cycle portion. This removes 2 from $v_\alpha + v_\beta + v_\gamma$ at the nodes visited and the remaining vector is again in $\mathbf{X}$, but with fewer edges. ■

Now we know $\mathbf{Z} = \mathbf{Y}$, in other words a sum of cycles is always the edge-disjoint sum of *some* (other) cycles. Also $\mathbf{X} = \mathbf{Z}$, i.e. Eulerian trails form a sub*space*.

There's some skewness in terminology. All elements of cut space $\mathbf{Z}^\perp$ are **cuts** as defined here, following [Die05]. While there is some precedent [SK77] of calling all elements of cycle space $\mathbf{Z}$ *1-cycles* (when elements of $\mathbf{E}$ are referred to as *1-chains*) the usual convention is to reserve the word **cycle** for single, well, cycles. If we call any member of $\mathbf{Z}$ a "**cyc**" for the moment then a cycle is a **minimal** "cyc" in the sense that no proper subset of its edges already forms a "cyc". In the same way a **minimal** cut ("cutle" anyone?) is a cut such that no proper subset of its edges forms a cut. In the interest of readability this paper uses *cut* and *cycle* in their conventional, albeit asymmetric, meanings and doesn't use *cyc* and *cutle*.

1.3 Orthogonality

Let the dot product (inner product) $\mathbf{u} \cdot \mathbf{v}$ be defined as usual, $\sum u_\iota v_\iota$. Like the u_ι and v_ι it is an element of the field, for instance in our $\mathbf{E}$ it is 0 (when $\mathbf{u}$ and $\mathbf{v}$ as sets share an even number of edges) or 1 (when they share an odd number). Two vectors are **orthogonal** (**perpendicular**) $\mathbf{u} \perp \mathbf{v}$ if their dot product is zero.

Caveat: that means every set with an even number of edges is, as a vector, perpendicular to itself (non-$\mathbf{0}$ vectors being perpendicular to themselves is quite common in spaces over a finite field). Let $\mathbf{A}$ denote the subset (indeed sub*space*) of $\mathbf{E}$ consisting of such vectors. Borrowing the terminology **weight** (number of nonzero coefficients of a vector) from coding theory, they are the ones with even weight.

The **orthogonal complement**[3] $U^\perp$ of a subspace U of V is the set $\{\mathbf{v} \in V \mid \forall \mathbf{u} \in U, \mathbf{v} \perp \mathbf{u}\}$. It is easy to prove it *is* a subspace too. Moreover, linear algebra teaches us $U^{\perp\perp} = U$, and $\dim U + \dim U^\perp = \dim V$. Beware though: for spaces over finite fields this doesn't usually mean $U \oplus U^\perp = V$, because in general $U \cap U^\perp$ is not just $\{\mathbf{0}\}$ but a subspace containing several more vectors.

Our $\mathbf{Z}$ was, in its guise $\mathbf{X}$, already defined as what amounts to the orthogonal complement of $\mathbf{Z}^\perp$. Cut space and cycle space being complements is a fundamental result. The dimension of $\mathbf{Z}$ is now $m - (n - k)$ which is the cyclomatic number o. It is in this sense that o counts the number of independent cycles of the graph.

[3] This terminology is less suited to spaces over a finite field because we'll see in a moment that U and $U^\perp$ do not necessarily span the whole space. In coding theory $U^\perp$ is the **dual code** of U [MS77] so *dual subspace* would seem the obvious choice. However, **dual** applied to spaces has another meaning. I need that meaning in section **1.6**.

By the way, here too $\mathbf{Z}$ and $\mathbf{Z}^\perp$ do not in general span $\mathbf{E}$. A small survey (appendix **D**) reveals there isn't any obvious relation between $\dim \mathbf{Z} \cap \mathbf{Z}^\perp$ and any of the usual concepts of graph theory.

What vectors are these that are both in $\mathbf{Z}$ and in $\mathbf{Z}^\perp$? Clearly they must be $\perp$ themselves but that is not a sufficient condition; every vector with even weight is $\perp$ itself in a space over $\mathbb{F}_2$. Being in $\mathbf{Z}$ they are disjoint sums of cycles, but being in $\mathbf{Z}^\perp$ they only contain edges of some $\mathcal{E}_{\bullet\circ}$. Now if every edge runs between a "black" and a "white" node each of the disjoint cycles must have an even number of edges:

- $\mathbf{Z} \cap \mathbf{Z}^\perp \subseteq \mathcal{Y}$

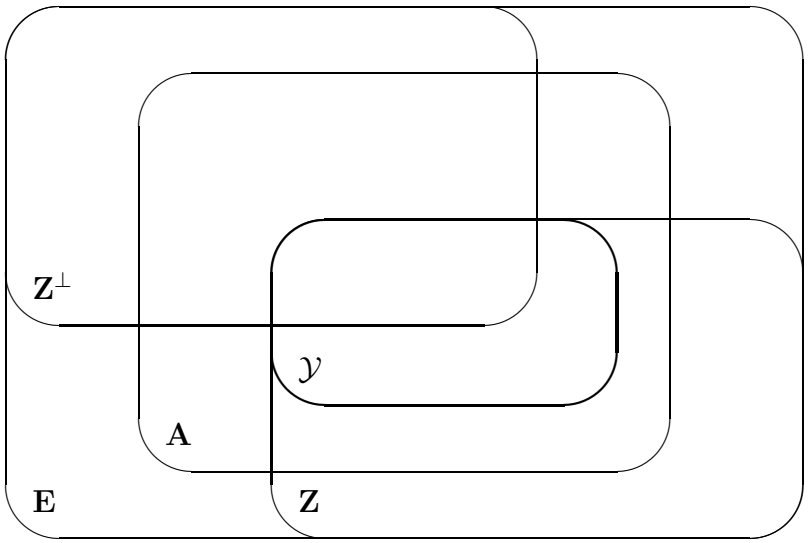
where $\mathcal{Y}$ is the set of those vectors that are **disjoint sums of even cycles**. It should not be confused with $\mathbf{A} \cap \mathbf{Z}$, the set of vectors that are **even disjoint sums of cycles** (that is, only the overall number of edges need be even). Obviously,

- $\mathcal{Y} \subseteq \mathbf{A} \cap \mathbf{Z}$

but this inclusion is in general strict (in the ordinary prism, the sum of the two triangle faces is in $\mathbf{A} \cap \mathbf{Z}$ but not in $\mathcal{Y}$). The other inclusion is in general strict too (in the same prism, a quadrangle face is in $\mathcal{Y}$ but not in $\mathbf{Z}^\perp$ so not in $\mathbf{Z} \cap \mathbf{Z}^\perp$).

Surprisingly, while $\mathbf{Z} \cap \mathbf{Z}^\perp$ and $\mathbf{A} \cap \mathbf{Z}$ are sub*spaces* (of $\mathbf{E}$ and of $\mathbf{Z}$), $\mathcal{Y}$ is only a sub*set* of $\mathbf{Z}$ and not a space at all (in the prism again, the three quad faces which were in $\mathcal{Y}$ sum to the disjoint sum of the triangles which wasn't).

A Venn diagram:



Of course this $\mathcal{Y}$ is not the only meaningful subset of $\mathbf{E}$ that's not a subspace. The most fundamental one is $\mathcal{E}$ itself! In a sense, all that can be done by linear algebra

is easy and the hard part of graph theory starts where it gets non-linear. The thing to do is to postpone that until it really becomes unavoidable.

1.4 Halfgraphs

An extension to the graph concept, useful in chapters 3 and 4, is the notion of a **halfgraph**. It consists of a number of nodes as usual, incident to edges and/or **halfedges** where a halfedge is incident to only one node. Any cut $\mathcal{E}_{\bullet\circ}$ cuts a graph into two halfgraphs. One of these (call it $H_\bullet$) has the nodes of $\mathcal{N}_\bullet$, the edges of $\mathcal{E}_{\bullet\bullet}$ as proper edges, and those of $\mathcal{E}_{\bullet\circ}$ as halfedges. The other ($H_\circ$) has the same elements with indices $_\bullet$ and $_\circ$ swapped. Let $n_\bullet$ be the number of nodes in $H_\bullet$ and let $e_\bullet$ and f be the number of proper edges and halfedges respectively.

One way to define a halfgraph $H_\bullet$ would be as an ordinary graph $U_\bullet$ using f extra univalent nodes U_j capping the halfedges. I want to extend the definition of cut space and cycle space to halfgraphs in a different way however. While such a $U_\bullet$ would have a cut space spanned by nodes including the U_j, those nodes are not going to feature in the cut space $\mathbf{Z}_\bullet^\perp$ of $H_\bullet$. Let $\mathbf{E}$ have one dimension for every edge (whether proper or half) and $\mathbf{Z}_\bullet^\perp$ be spanned by only the actual nodes of $H_\bullet$.

Let's look at each component in isolation (in other words only at connected halfgraphs, $k = 1$). Now if $f = 0$ we have an ordinary graph on our hands and we know the sum of all nodes is zero (it covers any edge $2 = 0$ times). In a **proper halfgraph** ($f \neq 0$) this is no longer true: the sum of all nodes now covers halfedges only once; it is the sum of all f of them. Define a **cut** as any sum of nodes; in a halfgraph it no longer equals the sum of the complementary set of nodes. In a connected proper halfgraph $\dim \mathbf{Z}_\bullet^\perp = n$.

This is $f+1$ less than the dimension of $U_\bullet$'s cut space ($n+f-1$ for its $n+f$ nodes) so consequently $\mathbf{Z}_\bullet$, defined again as orthogonal complement of $\mathbf{Z}_\bullet^\perp$, has dimension $f+1$ more than $U_\bullet$'s cycle space. The extra "cycles" needed to span $\mathbf{Z}_\bullet$ are the paths that start at a halfedge and end at another one.

1.5 Reduced edge space $\mathbf{E}^\circ$

Let **reduced edge space** $\mathbf{E}^\circ$ be defined as the quotient group $\mathbf{E}/\mathbf{Z}^\perp$ (with vector addition as group operation). The kernel of the homomorphism $\rho : \mathbf{E} \to \mathbf{E}^\circ$ is $\mathbf{Z}^\perp$, that is, ρ maps $\mathbf{u}$ and $\mathbf{v}$ to the same reduced vector iff $\mathbf{u} - \mathbf{v}$ (or what's the same thing, $\mathbf{u} + \mathbf{v}$) is in $\mathbf{Z}^\perp$. Let $\mathbf{u} \equiv \mathbf{v}$ denote that this is the case.

Note this definition of $\equiv$ refers to any $\mathbf{u}$ and $\mathbf{v}$ in $\mathbf{E}$. If $\mathbf{u}$ and $\mathbf{v}$ are singleton sets $\{\alpha\}$ and $\{\beta\}$ then this definition agrees with the relation $\equiv$ defined for edges on page 8 (i.e. that they feature in the same 2-cut), by the following lemma:

Lemma: $\{\alpha\} \equiv \{\beta\}$ iff $\alpha \equiv \beta$.

Proof: if $\{\alpha\} \equiv \{\beta\}$ (as defined in this section) then by definition $\{\alpha\} + \{\beta\}$ is in $\mathbf{Z}^\perp$, so α and β form a 2-cut. Conversely, if α and β form a 2-cut then $\{\alpha\} + \{\beta\} \in \mathbf{Z}^\perp$ so by the current definition $\{\alpha\} \equiv \{\beta\}$. ■

Incidentally, this is the promised algebraic proof showing the transitivity of the $\equiv$ defined on edges. Equivalence classes of that $\equiv$ are such that *any* two edges in an equivalence class form a 2-cut; equivalence classes of 2 or more edges are the cobridges while classes of 1 edge contain the edges that do not partake in 2-cuts.

Earlier we identified every set of edges (including every singleton edge, cycle, or other graph feature expressible as a set of edges) with a vector $\mathbf{v}$ of $\mathbf{E}$. We can now also assign to any such set its **reduced vector**

$$\mathbf{v}^\circ := \rho(\mathbf{v})$$

By definition it is $\mathbf{0}^\circ$ for nodes and other denizens of cut space.

While $\mathbf{E}^\circ$ has the same dimension ($m - \dim \mathbf{Z}^\perp$) that $\mathbf{Z}$ has, there is no unique one-to-one correspondence between them. There is one for every **spanning tree**, and it preserves addition. Without proof here (we won't need the result until later): every connected graph has a spanning tree (a tree containing all the nodes). Call the edges not occurring in the spanning tree **lianes**. Trees have the property that there is a unique path along the tree from any node to any other node. Any single liane, combined with the unique tree path between its endpoints, forms a cycle. Now the correspondence (for this spanning tree) is as follows: these cycles span $\mathbf{Z}$, and the reduced vectors of the lianes span $\mathbf{E}^\circ$.

As long as a graph is 3-edge-connected, each edge has a distinct reduced vector (as $\mathbf{u} \equiv \mathbf{v}$ would indicate the existence of a 2-cut). The dimension of $\mathbf{E}^\circ$ is far less than the number of edges; many edges are now the sum of other edges, for instance of a set of lianes as basis. We have an even stronger result:

Unique graph theorem: if 3-edge-connected G with largest valency Δ is such that all cuts other than those around a single node contain *more than* Δ edges, the list of reduced vectors of single edges (strictly speaking, of singleton sets of edges) uniquely determines the graph structure.

Examples of this situation are the solid trivalent graphs introduced on page 8 (the theorem is particularly well suited to regular Δ-valent graphs).

The **proof** is immediate: we have a list of vectors that represent single edges, by 3-edge-connectedness specific edges, and by assumption the only at-most-Δ-tuples (Δ-tuples, in the regular case) of them that sum to zero are the nodes. So now we know the cuts that are nodes, and which edges they are incident with. Finally the only case where a cut doesn't uniquely determine a node is in a (multi)graph consisting of 2 nodes joined by Δ edges (which then is the whole graph as the conditions of the theorem imply connectedness). This situation is easily recognised so here too we can reconstruct all nodes. ■

In chapter **3** we will see every trivalent graph decomposes in a straightforward way into solid components such that for certain proofs about trivalent graphs it suffices to consider those components; the unique graph theorem covers all of them and in that sense is applicable to all trivalent graphs.

While the theorem is almost tautological its impact is potentially far-reaching. Ordinary edge space $\mathbf{E}$ says very little about graph structure. Even supplying a list of edges (equivalently, specifying which is the usual basis whose basis vectors are the edges) tells us nothing, on its own. We need at least $\mathbf{Z}^\perp$ (or equivalently $\mathbf{Z}$) to find the nodes — *as well as* the list of edges that is, because without this the only specific feature, apart from $n-k$ and $m-n+k$, we can tell from $\mathbf{Z}^\perp$ and $\mathbf{Z}$ is the dimension of $\mathbf{Z} \cap \mathbf{Z}^\perp$ which is the same for large numbers of graphs. Actually the situation is murkier than that: we need a definition of dot product to derive $\mathbf{Z}$ and $\mathbf{Z}^\perp$ from each other and this goes some way towards pinpointing an implied basis (one where $\mathbf{u} \bullet \mathbf{v}$ is $\sum u_\iota v_\iota$) but not all the way. In summary, edge space only starts to determine graph structure if we are given a bunch of additional data. In sharp contrast, with reduced edge space $\mathbf{E}^\circ$ we *only* need a list of which vectors are the single edges (and not with respect to any particular basis either, just the coördinate-free linear dependencies between those edge vectors will do). Provided the graph is connected enough, as detailed in the theorem. In other words: earlier we saw the hard part of graph theory starts where it is no longer linear. The theorem identifies a very small packet of information on top of the linear algebra of edge spaces that is **sufficient to specify the whole graph**.

Of course not every list of reduced vectors will determine a valid graph (after we construct the nodes there better be two at each edge; the same is true for a graph specified by its $\mathcal{E} \subset \mathbf{E}$ and $\mathbf{Z}^\perp$). In later sections we will see which lists do.

1.6 $\mathbf{E}^\circ$ is dual to $\mathbf{Z}$

We saw $\mathbf{E}^\circ$, like $\mathbf{Z}$, has dimension $o = m - n + k$. Let f be a **linear function** $f : \mathbf{E}^\circ \to \mathbb{F}_2$, that is, $f(\mathbf{u}^\circ + \mathbf{v}^\circ) = f(\mathbf{u}^o) + f(\mathbf{v}^\circ)$ for all $\mathbf{u}^\circ$ and $\mathbf{v}^\circ \in \mathbf{E}^\circ$. Such a function is uniquely determined by an o-tuple of values f_i from $\mathbb{F}_2$: writing the components of $\mathbf{v}^\circ$ with raised indices v^i for a moment, $f(v) := \sum f_i v^i$. The set of these linear functions forms a vector space known as the **dual space** of (in our case) $\mathbf{E}^\circ$.

This concept of duality is quite distinct from the *dual codes* (orthogonal complements) introduced earlier. There the dimensions added to a constant (the bigger U was, the smaller $U^\perp$) and the vectors all lived in the same enclosing space. Here, dual spaces V and V^* will have exactly the same size (and don't have to be subspaces of anything), and their vectors are quite different animals. We do have $V^{**} = V$ (when the number of dimensions is finite).

Let V be any vector space; its vectors are called "contravariant vectors" in tensor calculus parlance; it is then customary to use raised indices as above, and a vector is denoted by its generic component such as v^i. A notation with subscript index such as f_i then denotes (components of) "covariant vectors", the denizens of the dual space V^*. One way to define covariant vectors in terms of contravariant ones is as above, that they are the linear functions that map V to a scalar[4] $\sum f_i v^i$.

If a fixed basis for V (and hence a particular way of making dot products) is used implicitly, the vectors of V^* will appear to stand in a fixed one-to-one correspondence with those of V, to such an extent that the distinction becomes immaterial. I did not introduce a dual space to $\mathbf{E}$ for this reason (the natural basis consisting of edges is the obvious one to use). If one were to make the distinction, dot products are only well-defined between vectors of $\mathbf{E}$ and those of $\mathbf{E}^*$ (in particular, one of $\mathbf{Z}$ and $\mathbf{Z}^\perp$ would have to be defined as subspace of $\mathbf{E}$ and one as subspace of $\mathbf{E}^*$).

Dual spaces become quite useful when there is no obvious one "natural" basis.

[4]Real tensor theorists leave out the Σ as well — repeated indices $f_i v^i$ imply "contraction" such as here while not repeating an index does not contract, so a tensor product $f_i v^j$ is a square array of m^2 components if the vectors have m components. With tensors, order of factors is immaterial because the indices carry the information what goes with what. For instance, matrix products AB and BA would be $A^i{}_j B^j{}_k$ and $A^j{}_k B^i{}_j$ respectively in tensor notation. The distinction between co- and contra-variant becomes very useful when the square length of a vector (dot product with self) is not given by the usual $\sum v_i^2$ but a more general quadratic; the square length will still be $v_i v^i$ where $v_i = g_{ij} v^j$ with g_{ij} the "metric". Dot products in general are also of this form $a_i b^j = a^i g_{ij} b^j$. Curved spaces can be accommodated painlessly just by making the metric vary from place to place.

Theorem: the dual space to $\mathbf{E}^\circ$ is, effectively, $\mathbf{Z}$.

Proof: Consider $\mathbf{Z}$ as a set of linear functions $\mathbf{c} : \mathbf{E} \to \mathbb{F}_2$ by $\mathbf{c}(\mathbf{v}) := \mathbf{c} \bullet \mathbf{v}$. As we didn't made a distinction between $\mathbf{E}$ and $\mathbf{E}^*$ we can keep all indices lower and write the dot product $\mathbf{c} \bullet \mathbf{v}$ as $\sum c_\iota v_\iota$ (summing over all edges $\iota \in \mathcal{E}$). These linear functions are not the complete set acting on $\mathbf{E}$ because any $\mathbf{v} \in \mathbf{Z}^\perp$ is mapped to $0 \in \mathbb{F}_2$ and likewise vectors from the same coset of $\mathbf{Z}^\perp$ in $\mathcal{E}$ are mapped to the same scalar. That is the *only* restriction on where $\mathbf{Z}$ maps things, as its dimension is all of $\dim \mathbf{E} - \dim \mathbf{Z}^\perp$. So $\mathbf{Z}$ is a complete set of linear functions on (that is, the dual space of) the space of cosets of $\mathbf{Z}^\perp$ in $\mathbf{E}$ (that is, $\mathbf{E}^\circ$). ■

Note this identification of $\mathbf{Z}$ with the dual space of $\mathbf{E}^\circ$ uses the dot product *in the surrounding* $\mathbf{E}$ to define the action of an element of $\mathbf{Z}$ as a function.

A 1-dimensional subspace of a space V is always the set of all scalar multiples of a nonzero vector $\mathbf{u}$ in V. In spaces over $\mathbb{F}_2$ that is just the set $\{\mathbf{0}, \mathbf{u}\}$ so the subspace uniquely determines the nonzero $\mathbf{u}$ that spans it. In the same way, a "hyperplane" (subspace of dimension one less than the full V) is determined by a vector $\mathbf{n}$ from the dual space, and vice versa in spaces over $\mathbb{F}_2$. The hyperplane is simply the set of $\mathbf{v}$ in V for which $\mathbf{v} \bullet \mathbf{n} = 0$, that is, $\{\mathbf{v} \in V \mid \mathbf{v} \perp \mathbf{n}\}$.

Now for $\mathbf{v} \in \mathbf{E}$ a single edge, and $\mathbf{c} \in \mathbf{Z}$, we have that $\mathbf{c}(\mathbf{v}) = 1$ iff the edge is in $\mathbf{c}$. Likewise $\mathbf{c}(\mathbf{v}^\circ) = 1$ iff the edge of which $\mathbf{v}^\circ$ is the reduced vector is in $\mathbf{c}$. Any hyperplane $\mathbf{C}^\circ$ of $\mathbf{E}^\circ$ is the set of $\mathbf{v}^\circ$ in $\mathbf{E}^\circ$ for which $\mathbf{c}(\mathbf{v}^\circ) = 0$, for some $\mathbf{c} \in \mathbf{Z}$. That's the same as saying $\mathbf{E}^\circ \setminus \mathbf{C}^\circ$ is the set of $\mathbf{v}^\circ$ in $\mathbf{E}^\circ$ for which $\mathbf{c}(\mathbf{v}^\circ) = 1$, for that $\mathbf{c}$. Intersecting $\mathbf{E}^\circ \setminus \mathbf{C}^\circ$ with $\mathcal{E}^\circ$, the set of reduced vectors that are single edges, now gives precisely those edges that were in that $\mathbf{c}$. This is how (sums of) cycles appear in $\mathbf{E}^\circ$.

In particular, if $\mathcal{B}$ is a basis of $\mathbf{E}^\circ$, and $\mathbf{b}_i^\circ$ the i-th basis vector, then the set of reduced vectors that have a coefficient 1 in i-th position with respect to basis $\mathcal{B}$ is such a complement of a hyperplane; intersecting with $\mathcal{E}^\circ$ gives an element $\mathbf{c}_i$ of $\mathbf{Z}$. By linearity, bases of $\mathbf{E}^\circ$ correspond to bases of $\mathbf{Z}$ in this way, and vice versa.

Finally, all the things said here for the spaces of a graph hold equally well for those of a **halfgraph**, for all the same reasons.

2 Planarity

2.0 Embedding in surfaces

Some graphs can be drawn on the page with none of the edges crossing where they shouldn't (i.e. only meeting at common endpoints). Other graphs cannot, but they can if you allow the page to be a surface Σ of high enough topological genus. An **embedding** of a graph is a mapping of its nodes to points on Σ, and of its edges to curve segments, where a curve segment is a continuous function $\gamma : [0, 1] \to \Sigma$. The points $\gamma(0)$ and $\gamma(1)$ should then coincide with the nodes, and no point $\gamma(x)$ for $0 < x < 1$ should coincide with any other point of this edge, or with any point of another edge. The set Σ^* of points that do not lie on any of the (curve segments representing the) edges will in general be disconnected. A piece of Σ that is simply connected (topologically equivalent to an open disc) will have (curve segments representing the edges of) a closed walk as its boundary, and is known as a **face**.

We know we can embed every graph in some surface (for example, embed a surface in 3 dimensions that has a skinny cyclinder for each edge stuck to a little sphere for each node and draw the graph on that) and even such that every piece of Σ^* is a face (if two faces end up joined by a "handle" of the surface we can remove that, and if a region forms a "crosscap" replace it by a simply connected region). There is a detailed topological treatment in [FF94] and [Die05] also has a section on it.

It is conjectured that (for bridge-free graphs) the embedding can always be done such that face boundaries are cycles. In that case each edge of the graph will occur in exactly two of the cycles. Such a set of cycles is known as a simple polyhedral decomposition, or a cycle double cover, and we'll look at them in more detail in chapter **5**. Note that by Euler's formula the genus g of the surface and the number f of faces are then related by $n - m + f = 2 - 2g$ (for connected G), in other words $o - f = 2g - 1$ (where o was the cyclomatic number, $m - n + k$ in general, so $m - n + 1$ for a connected graph). Keep in mind n and m, and hence o, are fixed by the graph which leaves us with a straightforward relation between f (number of faces) and g (genus of the embedding).

An equivalent but purely combinatorial expression of this topological conjecture is the "cycle double cover conjecture": for every bridge-free trivalent graph there exists a set of cycles such that every edge of the graph occurs in exactly two of them. It also makes the genus g of this cycle double cover a purely combinatorial

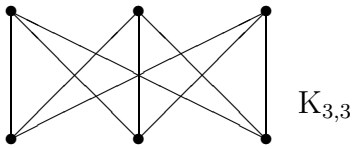
notion. It will be shown to be nonnegative (just as in topology) so the value of $f = o + 1$ for genus 0 is the maximum number of faces for any genus.

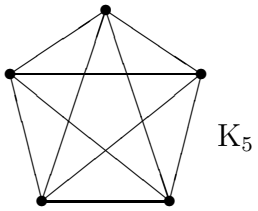
For trivalent graphs, that maximum number of cycles used in the double cover is $h + 2$. And a cycle passes through at most all $2h$ nodes, giving $2h$ edges, so the absolute minimum number of cycles to cover the $3h$ edges twice is 3 here, making the maximum genus $(h - 1)/2$.

Cycle double covers form the subject of the last chapter. Suffice to mention here they do not just reproduce genus but also that other topological property of surfaces, whether they are orientable (i.e. the distinction between mirror image features, such as clockwise or anticlockwise order of edge colors at a node, is globally meaningful) or not (like Möbius strip and Klein bottle). Genus 0 implies orientable, other integer genera come in both flavors, whereas half-odd genus implies non-orientable.

2.1 Planar graphs

A **planar** graph is one that can be embedded in a genus 0 surface (i.e. on the sphere, or equivalently on the plane). Again there are equivalent purely combinatorial criteria for a graph to be planar. **Kuratowski's theorem** [Wil02, Die05 etc.] says a graph is non-planar iff it contains a $K_{3,3}$ or K_5 *minor*[5].





MacLane's criterion [SK77, Die05] says a graph is planar iff it has a basis $\mathcal{C}$ for $\mathbf{Z}$ where every edge of the graph occurs in at most two of the basis vectors $\mathbf{c}_i$ of $\mathcal{C}$. If G is connected and bridge-free each edge must occur in at least one of the $\mathbf{c}_i$ (so some edges occur once, some twice). Now $\mathbf{c}_+ := \sum \mathbf{c}_i$ contains exactly the

[5]**Contracting** a graph H to another graph K can be pictured as (repeatedly) shrinking one of H's edges to zero length, merging its endpoints to a single node in K. Of course the edge is discarded, and if other nodes had edges to both the two old nodes, the resulting double edge is replaced by one. One says K is a **minor** of G if there is a subgraph $H \subseteq G$ that can contract to (a graph isomorphic to) K.

There is a different concept of **topological minor** but for trivalent graphs the two coincide. Note G has a K_5 minor iff its $\mathcal{N}$ can be partitioned into $\mathcal{N}_0$ (which may be empty), and $\mathcal{N}_1$ through $\mathcal{N}_5$ that each induce a connected subgraph, such that all ten $\mathcal{E}_{ij}$ $(i \neq j)$ are nonempty.

edges covered only once in $\mathcal{C}$. But then $\mathcal{C}^+ := \mathcal{C} \cup \{\mathbf{c}_+\}$ covers every edge exactly twice so the disjoint cycles in these vectors form a cycle double cover. We know the maximum number of cycles in a cover is $o+1$ but that's the number of vectors used (basis plus one). So every vector used here, including $\mathbf{c}_0$, is a single cycle. This cover is of course the obvious genus 0 one for planar graphs, that using the single faces of the embedding. Any face cycle is the sum of all the other ones, which form a basis. Note in passing the use of $\cup$ assumed $\mathbf{c}_0$ didn't equal one of the $\mathbf{c}_i$; it *does* so if the graph is a single cycle.

A "**plane graph**" is not a graph, but a planar graph together with a choice of embedding. An embedding on the sphere can be punctured in any of its f faces and then spread out as an embedding in the plane. Intuitively it is clear that 1- and 2-edge-connected graphs can be embedded on the sphere in more than one way (make a wireframe model in 3-D, rotate the 2-edge-connected portions independently from each other before flattening them onto the sphere surface) and hard to see how this could be done with 3-edge-connected ones; it is actually a theorem [Die05] that the latter only embed on the sphere in essentially one way.

The embedding Γ of planar graph G is **dual** to the embedding Δ of D if each has one of its nodes inside every face of the other, with one edge crossing every edge of the other. Now G and D are **Whitney duals** i.e. there is a bijection between their $\mathbf{E}$'s that maps edges to edges, and $\mathbf{Z}$ and $\mathbf{Z}^\perp$ of one to $\mathbf{Z}^\perp$ and $\mathbf{Z}$ of the other. **Whitney's criterion** says G is planar iff it has a Whitney dual [SK77].

3 Trivalence

3.0 All the graphs you'll ever need?

One soon realises that pretty much anything that happens in a disconnected graph happens inside one of its components, that the other components might just as well not have been there. In order to describe what goes on in graphs it is sufficient to only deal with the individual components, that is, only with connected graphs.

Connected 0-valent graphs are frankly boring. There's only one, the single isolated edgeless node. Connected 1-valent graphs aren't much better. Again there's only one, a pair of nodes linked by an edge. Connected 2-valent graphs are a bit more promising; at least there's more than one. Each is a single cycle and you get to choose how many repeating units (one node, one edge) there are, but that's all there's to it. Mixing valencies 1 and 2 gives us linear paths as well. Everything changes with valency 3. Suddenly there's this bewildering variety of shapes.

In some sense here is all the graph connectivity you'll ever need. Every graph can be obtained as contraction of some trivalent graph, because nodes of any valency can be emulated by connected bunches of 3-valent nodes. The upshot is that some kinds of things (but by no means all) will be true for all graphs if they can be proven just for trivalent graphs. We'll see one example of that with coloring plane maps in the next chapter.

Of course, you miss out on a lot of beauty and surprise if you just deconstruct every graph into a trivalent graph that will contract to it, but in this one sense the germs of all the ways graphs can hang together are already there in trivalent graphs.

With this status as simplest "real" graphs comes some exceptional behaviour too:

- While edge connectivity κ' and [node] connectivity κ are different concepts in general, they coincide for trivalent graphs. The simplest cases of features that make $\kappa' > \kappa$, such as "part of the graph $\succ\!\!\prec$ rest of the graph" all require at least 4 edges at a single node.
- While closed trails (all edges distinct) aren't necessarily closed paths aka cycles (all nodes distinct) in general, they are the same thing in the case of trivalent graphs (page 10).
- While edge-disjoint cycles aren't necessarily node-disjoint in general, they are in the case of trivalent graphs. It is impossible for there to be two edge-

disjoint cycles at a node, when there aren't at least 4 edges there. In trivalent graphs, Eulerian edge sets have a unique decomposition into cycles.

- While topological minors are minors (see footnote in previous section), the converse isn't true in general. For graphs that are no more than 3-valent however the concepts do coincide [Die05].

- The existence of a loop automatically causes there to be a bridge between the rest of the graph and  if the the relevant node is trivalent. Conversely, demanding 2-edge-connectivity automatically excludes pseudographs now.

- The existence of a double edge  automatically makes a cobridge if the nodes involved are trivalent. Conversely, demanding 3-edge-connectivity automatically excludes multigraphs now, except for the one triple-bonded .

Various induction techniques can be used in graph theory. In order to prove things about finite sets without internal structure, it is enough to show that a property holds for the empty set and that, if it holds for a set with a certain number of objects, it'll also hold for one with one more object. Graphs do have internal structure though so we must be more careful. It's not enough to only add edges (between existing nodes), and not enough to only add new nodes (initially without edges, for example). Proving a property survives *both* these operations would do, but it wouldn't work if the property only holds for certain kinds of graphs.

Trivalent graphs, for example, only exist with $n = 2h$ and $m = 3h$ for integer h, so here the single induction quantum would have to be a net increase by 2 nodes and 3 edges (it will also increase o by 1). Incidentally, one nice way to do it is to pick two distinct[6] edges  and re-route them as  to make a graph of the next bigger size, and start with K_4 (tetrahedron). It's easy to prove the result is always 3-edge-connected. Can all 3-edge-connected trivalent graphs be made this way?

3.1 Growing trivalent graphs

The following theorem takes a slightly different approach to growing a graph. It is stated in terms of a graph that is already given; we just pick our way around it, marking the nodes as "done" one by one, subject to some constraint.

[6]We can allow the edges to be same one (taking care with the order of operations). Now the result need not be more than 2-edge-connected and can even be a multigraph. Multigraphs are a necessary precursor step for some 2-edge-connected graphs, notably those with  elements.

Weak Bipolar Growth Theorem: If G is a 2-edge-connected trivalent graph there exists a sequence $\{\,\} = \mathcal{N}_\bullet^0 \subset \mathcal{N}_\bullet^1 \subset \mathcal{N}_\bullet^2 \subset \cdots \subset \mathcal{N}_\bullet^{n-2} \subset \mathcal{N}_\bullet^{n-1} \subset \mathcal{N}_\bullet^n = \mathcal{N}$ (this implies each $\mathcal{N}_\bullet^i$ contains i nodes) such that *both*
– the subgraph $G_\bullet^i$ induced by $\mathcal{N}_\bullet^i$ is connected (for $i > 0$), *and*
– the subgraph $G_\circ^i$ induced by $\mathcal{N}_\circ^i := \mathcal{N} \setminus \mathcal{N}_\bullet^i$ is connected (for $i < n$).

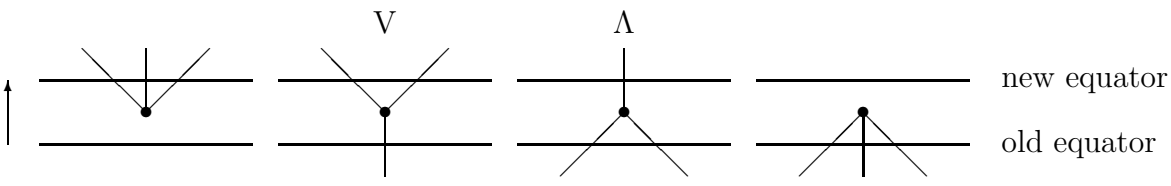
Call $G_\bullet^i$ (with edge set $\mathcal{E}_{\bullet\bullet}^i$) the Southern graph, $G_\circ^i$ (with edge set $\mathcal{E}_{\circ\circ}^i$) the Northern graph, and the cut $\mathcal{E}_{\bullet\circ}^i$ the equator.

To **prove** the first half of the theorem, that the growing Southern $G_\bullet$ can be made to remain connected, it is enough to just co-opt a node adjacent to one or more of the the existing $\mathcal{N}_\bullet$ at each step (just like when building a spanning tree). If that were ever impossible we would have an empty cut $\mathcal{E}_{\bullet\circ}$ so G wasn't connected.

Thus far, we don't really need the fact that the final trivalent graph is already known. We could still end up with any G that contains the present $G_\bullet$. This is just like building by induction, except that the partially built graphs aren't trivalent. To prove the second half of the theorem, that the shrinking Northern graph can be kept connected, we'll have to look ahead across the fence and avoid certain choices.

Lemma: We can start. **Proof:** Take any node O to form $\mathcal{N}_\bullet^1 = \{\text{O}\}$ with OA, OB, OC the equatorial edges (removed from North). Suppose this already disconnects North. Now A, B, C cannot end up in the same component Y of North (let Z be a different component, it does not contain O so no edges between Y and Z were cut by the equator, so G must have started off already disconnected). So A, B, C split over at least two components but then at least one component received only one of them, WLOG A. But then OA was a bridge in G contrary to the givens. ■

There are four kinds of Northern nodes, $\mathcal{N}_\circ^i = \mathcal{N}_0^i \cup \mathcal{N}_1^i \cup \mathcal{N}_2^i \cup \mathcal{N}_3^i$ where $\mathcal{N}_v^i$ is the set of nodes with v edges to other Northern nodes (and hence $3 - v$ edges to Southern nodes, that is, edges in the equatorial cut). At the last stage (when $i = n - 1$) there is only one Northern node left which perforce is in $\mathcal{N}_0^i$. At any stage before that, there better be no nodes at all in $\mathcal{N}_0^i$ because any such would be a separate component of $G_\circ^i$.



The four pictures show the effect of co-opting a node from $\mathcal{N}_3^i$, $\mathcal{N}_2^i$, $\mathcal{N}_1^i$ or $\mathcal{N}_0^i$

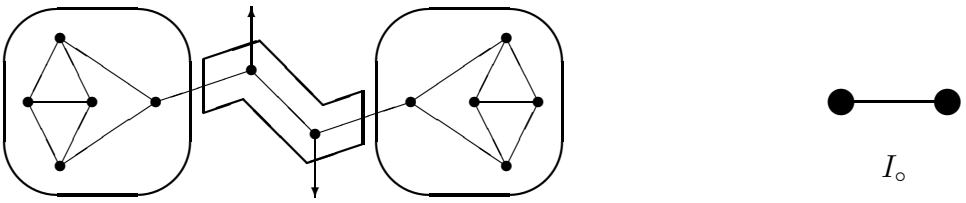
respectively to the growing South. The first kind happens on going from $\mathcal{N}_\bullet^0$ to $\mathcal{N}_\bullet^1$ and will not happen again as we keep $G_\bullet$ connected. The last kind happens on going from $\mathcal{N}_\bullet^{n-1}$ to $\mathcal{N}_\bullet^n$ and must not happen any other time because a node co-opted that way would be the last remnant of a piece that got disconnected from the rest of North. I will call the two proper ways to proceed (in the $n-2$ steps between $\mathcal{N}_\bullet^1$ and $\mathcal{N}_\bullet^{n-1}$ stages) V-steps and Λ-steps respectively.

Lemma: A Λ-step never disconnects North. **Proof:** obvious, the only Northern edge cut by it is between a single node (that will stop being part of North anyway) and the rest of North. ■

This suggest a "greedy" strategy whereby we always take Λ-steps as long as we can, that is, as long as $\mathcal{N}_1^i$ isn't empty. Now we only have to worry about the situation where it *is* empty. In that case, all Northern nodes are in $\mathcal{N}_2^i$ or $\mathcal{N}_3^i$ (no $\mathcal{N}_0^i$ either because we're doing an induction argument so may assume North isn't disconnected yet). The nodes can't all be $\mathcal{N}_3^i$ because then $\mathcal{E}_{\bullet\circ}^i$ would be empty and G disconnected. So there *are* some $\mathcal{N}_2^i$.

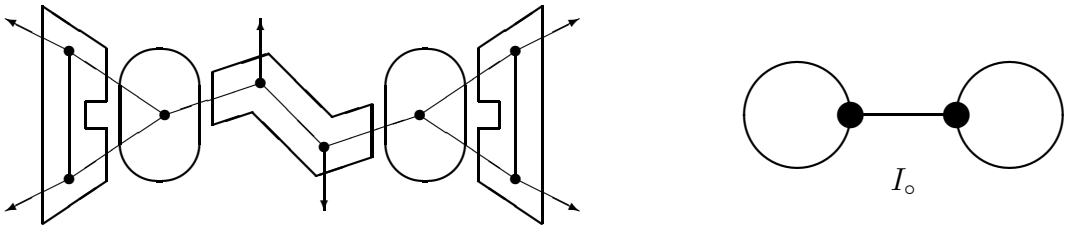
Lemma: If we can only do V-steps, we can do so without disconnecting North. **Proof:** The only nodes we can co-opt to South are $\mathcal{N}_2^i$ ones; the ones to avoid are those with a bridge on one (and hence also the other) side. To prove it is impossible for all the $\mathcal{N}_2^i$ nodes to be like that.

Let G_2 be the subgraph of $G_\circ^i$ induced by the $\mathcal{N}_2^i$ nodes, and G_3 that induced by the $\mathcal{N}_3^i$ nodes. There are three kinds of edges: those internal to G_2, those internal to G_3, and those between $\mathcal{N}_2^i$ and $\mathcal{N}_3^i$ nodes; call the latter kind **clamps**. Firstly, suppose there are no $\mathcal{N}_3^i$ nodes. In that case $G_\circ^i = G_2$ is regular 2-valent, so (being connected) is a cycle, and has no bridges. Otherwise each component of G_2 is a path; let a **lead** in $G_\circ^i$ be such a component together with the clamp at each end.



Now construct a new graph $I_\circ$ from $G_\circ^i$ by contracting each component of G_3 in it to a single new node, and replacing each lead by a new edge. Clearly, if an edge in a lead (and that includes the clamps) was a bridge of $G_\circ^i$ then the lead is a bridge of $I_\circ$, and if a lead is a bridge of $I_\circ$ then all its edges were bridges of $G_\circ^i$.

If *all* the leads in $I_\circ$ are bridges then $I_\circ$ is a *tree*. The crucial feature of trees now is that they have 1-valent nodes. And if a node of $I_\circ$ is 1-valent then the clamp leading up to it in $G^i_\circ$ is a bridge not just there, but also in the original G (the nodes making up this G_3 component are all 3-valent in $G^i_\circ$ so have no more edges in the original G than the ones we see here in $G^i_\circ$). Absence of bridges in G was given; it excludes this possibility.



So $I_\circ$ can't be a tree, and has edges that aren't bridges there. The edges of such a lead aren't bridges in $G^i_\circ$ either and we can safely co-opt any of its $\mathcal{N}_2$ nodes. ■

And with these lemmata the theorem follows. ■

While some 1-edge-connected graphs can also be constructed this way, not all of them can. The proof the construction exists relied on the absence of bridges.

Corollary: when building a trivalent graph this way we will only need V-steps and Λ-steps (in stages $i = 1$ through $n - 1$).

This is implicit in the proof as given. ■ It is the most useful form of the theorem for the purpose of actually constructing an arbitrary bridge-free trivalent graph.

The BGT can also be stated in terms of trivalent halfgraphs $H^i_\bullet$ and $H^i_\circ$ which are $G^i_\bullet$ and $G^i_\circ$ each with $\mathcal{E}^i_{\bullet\circ}$ added as halfedges. It will be used that way later.

If a **planar** G is embedded on a sphere, we can at each stage draw a representation of the equator as a closed curve separating North and South such that it intersects all the edges of $\mathcal{E}_{\bullet\circ}$ and no others. This imposes a **cyclic ordering** on these edges in the planar case.

Note in passing that in the proof we could take any node at all as our first node, the **South Pole**. Of course the construction is reversible (taking the old $\mathcal{N}^{n-j}_\circ$ as our new $\mathcal{N}^j_\bullet$ again satisfies connectedness both North and South) so every node can also feature in the rôle of **North Pole**, the last node taken. This result can be strengthened to the

Strong Bipolar Growth Theorem: as the Weak BGT above, but we can choose both poles independently (distinct, of course).

Proof: Start as in the weak version above with the desired South Pole, and mark the desired North Pole as node to keep for last.

There will be at least two $\mathcal{N}_2^i$ nodes at any stage $< n-1$ (otherwise $\mathcal{E}_{\bullet\circ}^i$ would be a bridge of G), but we must prove there are at least two in leads that aren't bridges of $I_\circ$ (then we can always avoid taking the marked node). If $I_\circ$ has only one edge we already know it's not a bridge and has all the $\mathcal{N}_2^i$ nodes so we're done. If it has more edges let's prove at least two of them aren't bridges.

Suppose there is only one non-bridge edge XY (no need to demand $\text{X} \neq \text{Y}$). Now removing it makes $I_\circ^* := I_\circ \setminus \{\text{XY}\}$ a tree (it's connected because XY wasn't a bridge, and all the remaining edges are bridges because by assumption they were already bridges in $I_\circ$). We saw $I_\circ$ could not have any 1-valent nodes so $I_\circ^*$ can have at most two, X and Y. A tree with no 1-valent nodes is a single node without edges, which contradicts $I_\circ$ having other edges besides XY. A tree with one 1-valent node doesn't exist. And a tree with two 1-valent nodes is a path, then $I_\circ$ was a cycle and its edges not bridges after all, contradicting the assumption. ■

3.2 Spanning trees

In the BGT construction we allocated, each time we had a $\mathcal{N}_\bullet$, all the edges of $\mathcal{E}_{\bullet\bullet}$ to $G_\bullet$ ($G_\bullet$ is the **induced** subgraph on that set of nodes). Equivalently, we allocated all the edges of $\mathcal{E}_{\bullet\bullet}$ as proper edges to the halfgraph $H_\bullet$, and all those of $\mathcal{E}_{\bullet\circ}$ as halfedges (which we could likewise take as a definition of **induced** for halfgraphs).

Relaxing this definition of induced somewhat (*weakly* induced perhaps?) we could merely demand that, for every $\text{P} \in \mathcal{N}_\bullet$, all edges incident to it in G are found back as proper edges or halfedges in some halfgraph $T_\bullet^i$. The difference is that an edge PQ in $\mathcal{E}_{\bullet\bullet}$ now need not be a proper edge. Rather, it could feature in such a $T_\bullet^i$ in the form of two halfedges from P and Q respectively. The two halfedges would still have to be tied together (not with any Λ-step but just as two halves of an edge) to re-create the way P and Q were connected in G. Such a situation is specifically excluded in BGT by taking proper edges from $\mathcal{E}_{\bullet\bullet}$ and halfedges from $\mathcal{E}_{\bullet\circ}$.

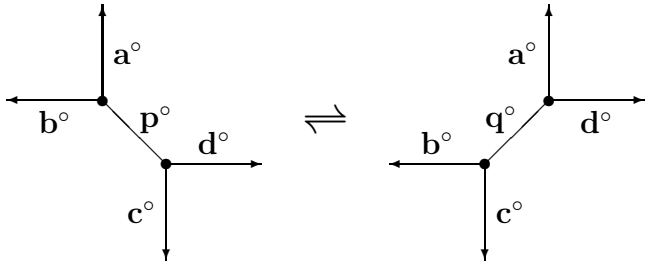
The weakest (as in least connected) such construction that still leaves such a "Southern" $T_\bullet^i$ connected would each time add only one proper edge when a new node is co-opted to the growing South. The resulting final $T_\bullet^n$ is of course (when shorn of its $2h+2$ halfedges) a **spanning tree**. Label the halfedges and let $\mathcal{S}$ be the set of all trivalent graphs obtained by pairing up the halfedges. Clearly there

are $(2h+1)!! := (2h+1)(2h-1)(2h-3)\cdots 3\cdot 1$ ways to do the pairing (treating the two halfedges on 1-valent nodes of $T^b_\bullet$ as distinct). These are all the possible trivalent graphs that admit a given tree (with valency nowhere greater than 3) as spanning tree.

In the **planar** case (keeping $T^n_\bullet$ embedded in the plane in the same way as in a given embedding Γ of the graph it is a part of), halfedges must be joined up without crossing. It is well known such non-crossing pairings of $2(h+1)$ cyclically arranged items can be done in a **Catalan number** $C_{h+1} := \frac{1}{h+2}\binom{2h+2}{h+1}$ of ways. This constitutes all the connected trivalent planar graph embeddings Γ that admit a given tree (with valency nowhere greater than 3), with given planar embedding, as spanning tree. This again treats two halfedges on one node as distinct but here it does not lead to overcounting; a pairing obtained by swapping two halfedges on the same node would not have been counted among the C_{h+1} non-crossing pairings.

3.3 Flips of trivalent graphs

Consider the following local modification to a trivalent graph or halfgraph:



where edges are labelled by their reduced vectors.

I will call this a **flip** (as opposed to a concept of *flop* introduced later). Two nodes are replaced by two other nodes; an edge by another edge (the rest of the graph beyond the arrows stays the same). In terms of $\mathbf{E}$ etc. this is *almost* the same graph: one edge gets projected out of the basis and another one put in, cycle and cut spaces get tweaked, etc. Alternatively one *could* say $\mathbf{q}^\circ$ is the "same" edge as $\mathbf{p}^\circ$ that got connected differently, but that doesn't work too well (re-creating the same graph shape after flipping some more edges around doesn't necessarily put $\mathbf{p}^\circ$ back in the rôle of $\mathbf{p}^\circ$).

In terms of reduced vectors however we can use *exactly* the same space. In the left image we have $\mathbf{a}^\circ + \mathbf{b}^\circ = \mathbf{p}^\circ = \mathbf{c} + \mathbf{d}^\circ$ which implies $\mathbf{a}^\circ + \mathbf{b}^\circ + \mathbf{c} + \mathbf{d}^\circ = \mathbf{0}^\circ$ (i.e. any of these four could be expressed as the sum of the other three, they form a

cut). In the right image we have $\mathbf{a}^\circ + \mathbf{d}^\circ = \mathbf{q}^\circ = \mathbf{b} + \mathbf{c}^\circ$ which implies exactly the same relation $\mathbf{a}^\circ + \mathbf{b}^\circ + \mathbf{c} + \mathbf{d}^\circ = \mathbf{0}^\circ$. Here it makes sense to say the underlying space stayed the same. Both $\mathbf{p}^\circ$ and $\mathbf{q}^\circ$ exist as vectors in the space and continue to exist, the "only" difference between the images is that $\mathbf{p}^\circ$ gets dropped from the list of reduced vectors that represent single edges, and $\mathbf{q}^\circ$ gets chosen instead — and that determines which sets of edges form a node, etc.

This is an interesting alternative way to **walk the space of trivalent graphs of a given size**. We do not keep labelled nodes and then span edges between them, nor keep labelled edges and then tie some together as nodes (indeed, any one node or edge may get dropped from the graph at some stage). Instead, we keep the underlying $\mathbf{E}^\circ$ fixed (its basis vectors are what is "labelled" so to say) and change which vectors are edges. Each step of the walk is as in the picture: you pick on any edge $\mathbf{p}^\circ$ and do this flip. If planarity is not important there is even a third possible edge there, equal to $\mathbf{a}^\circ + \mathbf{c}^\circ$ or what's the same thing $\mathbf{b}^\circ + \mathbf{d}^\circ$.

The procedure when carried out mindlessly (by a computer on a random walk, say) may produce less connected graphs. For example, if (in the left picture) $\mathbf{a}^\circ\mathbf{b}^\circ$ and $\mathbf{d}^\circ\mathbf{c}^\circ$ are two of the four links between two portions of a solid graph then in the right picture those two paths have been replaced by a single one via $\mathbf{q}^\circ$ making the graph 3-edge-connected. Another such flip could make the graph 2-edge-connected and then a reduced vector for some edge, e.g. $\mathbf{q}^\circ$, will be the same as the vector for the other edge of the relevant 2-cut. Yet another flip can bring 1-edge-connectedness (the vector of the bridge is now $\mathbf{0}^\circ$) and if we reach a stage where there is a loop, attempts to flip that loop around have no effect. All these changes are reversible *if* you know what to reverse i.e. if you have an independent record what's connected to what. If you don't then plain 3-edge-connectedness is already a problem: nodes cannot be distinguished from other 3-cuts now, and you need to know what the nodes are to be able to make a flip.

All this is bound to happen (as detailed in section **1.5**) since $\mathbf{E}^\circ$ together with a list of the reduced vectors that are edges only uniquely names edges in a trivalent graph if it is 3-edge-connected, and only distinguishes nodes from other 3-cuts if it is solid. So if the list of vectors of $\mathbf{E}^\circ$ that are single edges is the *only* way the graph is represented (by a proof, or by a probabilistic computer survey, etc.) then you need to restrict the allowed flips to those that keep it solid.

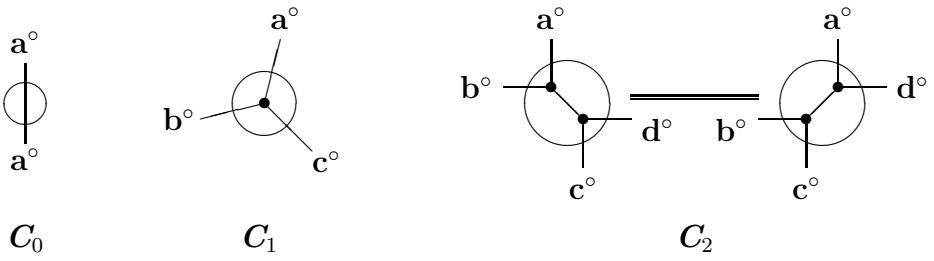
For a proof that just means you demand that the graph stays connected that way (and then you need to worry whether there are any flips still allowed). Time constraints prevent me from investigating whether the set of all solid trivalent

graphs of a given size — each represented by a node in some huge graph, with edges between those that are one flip away — is connected, in general. We *will* see an answer to this question in the planar case though.

In the case of a computer walk (random or systematic) you need to check the new edge ($\mathbf{q}^\circ$ in the picture) does not make any more 3-cuts than the two that represent its endpoints. Probably the most efficient way to do this is by maintaining a list of all the $\binom{m}{2}$ sums of edges. Checking which of these equal the new edge is an $O(m^2)$ process and updating the list an $O(m)$ one. The list could also be used to find the adjacent edges (needed to carry out the flip) in the first place, another $O(m^2)$ operation, if that information isn't already kept in another form.

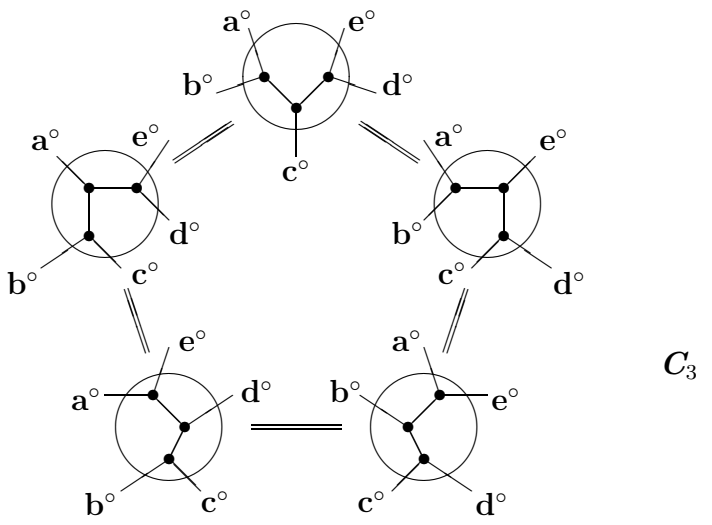
3.4 All planar trivalent halftrees

In this section we will (at first at least) not worry about representing a graph uniquely by its list of edges in $\mathbf{E}^\circ$ but simply use flipping as a means to hop from one halfgraph to the next. Also, we will only look at **plane** embeddings. Now we need to interpret the image on page 30 as representing a portion of a graph embedded (as in the picture) in the plane. Let's say the image is that of a **halftree** (a halfgraph without cycles), the arrows are its halfedges, and we maintain their cyclic order. Under those circumstances the two ways depicted are the only planar ways to tie up the inside trivalently. Here they are again (right in the picture):

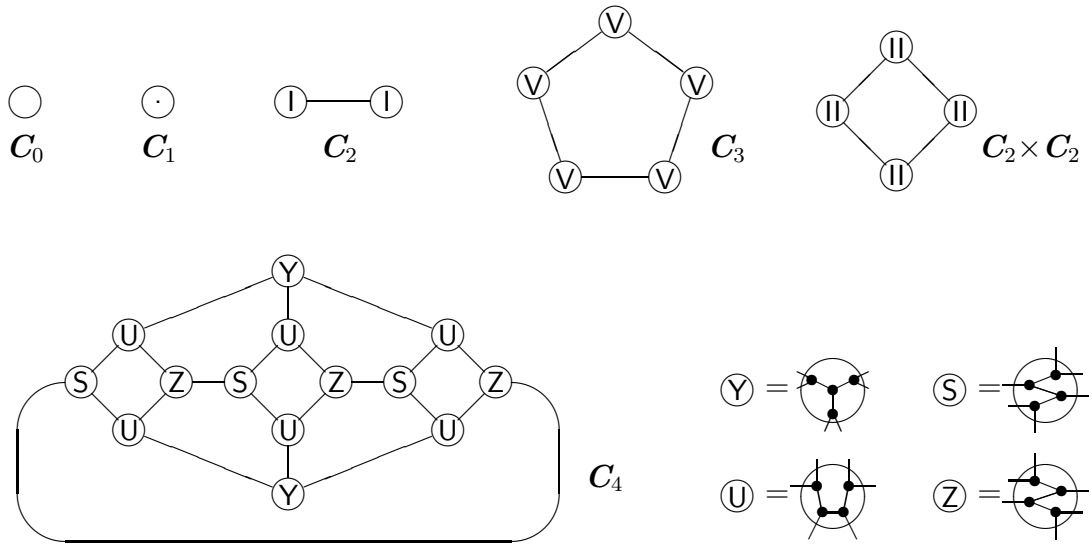


Making a planar trivalent patch like that with three halfedges crossing the "equator" is even easier, now there is just one node inside and no proper edges. Only one way to do that (second picture). With two halfedges there are no nodes inside; the halfedges have to be the same edge going straight through (left picture).

If on the other hand we make the number of halfedges larger (still without there being cycles inside) things rapidly get more interesting. With five halfedges crossing the "equator" (in fixed cyclic order) there are five planar ways to tie things up inside without cycles (needing three nodes and two proper edges) and the five arrangements can be reached from each other by means of flips:



The set $\boldsymbol{C}_n$ of such halftrees (using $n+2$ halfedges, no cycles inside so n nodes and $n-1$ proper edges) forms itself a meta-graph if we represent each halftree by a meta-node, and being linked via a single flip by a meta-edge (the prefix *meta* here has no other function than to avoid confusion with the nodes and edges of the individual halfgraphs). Here they are again on a smaller scale (with halftree shapes indicated schematically), including $\boldsymbol{C}_4$ we hadn't seen yet:



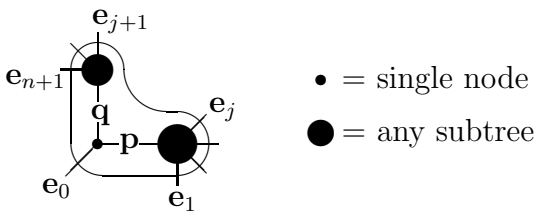
In $\boldsymbol{C}_4$ we see for the first time some internal structure. The 5-gons ($\boldsymbol{C}_3$ subgraphs) occur where two adjacent halftree edges are being flipped; 4-gons indicate two non-adjacent halftree edges being flipped (they are the product of two $\boldsymbol{C}_2$ that don't interfere with each other). All larger $\boldsymbol{C}_n$ have these 5- and 4-cycles, for the same

reasons (relatively fewer 5-cycles as n gets bigger), and each meta-graph $\boldsymbol{C}_n$ is regular $(n-1)$-valent because every proper edge in a halftree can be flipped.

Thus far the numbers of meta-nodes are 1, 1, 2, 5, 14, ... We have a surprise appearance of the **Catalan numbers** again.

Theorem: The number of halftrees on $n+2$ halfedges is $C_n := \binom{2n}{n} \Big/ (n+1)$, the n-th Catalan number.

Proof: We must count the number of ways, N_n say, to tie up the $n+2$ given halfedges $\mathbf{e}_i$ into planar trivalent halftrees with n nodes, $n-1$ proper edges and maximum valency 3, counting rotated versions separately only in as far as their orientation can be distinguished relative to the $\mathbf{e}_i$. To prove $N_n = C_n$, the n-th Catalan number.



We saw the induction starts for the first few n. Now assume it has been shown for numbers of nodes up to and including $n-1$, to prove for n.

Of the $n+2$ halfedges, follow one ($\mathbf{e}_0$ for instance) inward until it forks. The right branch $\mathbf{p}$ leads to some $j > 0$ halfedges. By planarity they must be contiguously $\mathbf{e}_1$ through $\mathbf{e}_j$ and by the induction assumption there are N_{j-1} ways to tie these $j+1$ halfedges ($\mathbf{p}$ and the $\mathbf{e}_i$). The left branch $\mathbf{q}$ leads to the rest, $n+1-j > 0$ halfedges, independently N_{n-j} ways to tie the $n+2-j$ halfedges including $\mathbf{q}$. So

$$N_0 \;=\; 1 \quad \text{(for } n+2=2 \text{ there are no } \mathbf{p} \text{ and } \mathbf{q}\text{, so the above doesn't apply)}$$
$$N_n \;=\; \sum_{j=1}^{n} N_{j-1}N_{n-j} \;=\; \sum_{i=0}^{n-1} N_i N_{n-1-i} \quad \text{(for } n+2 \geq 3 \text{ halfedges i.e. } n \geq 1\text{)}$$

But these together are a well known recurrence relation for $N_n = C_n$. ■

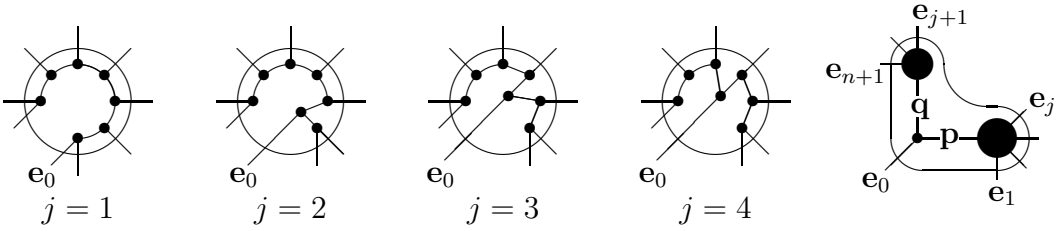
The first few Catalan numbers, after $C_0 \;=\; 1$, are

$$C_1 \;=\; C_0C_0 \;=\; 1\cdot 1 \;=\; 1$$
$$C_2 \;=\; C_0C_1 + C_1C_0 \;=\; 1\cdot 1 + 1\cdot 1 \;=\; 2$$
$$C_3 \;=\; C_0C_2 + C_1C_1 + C_2C_0 \;=\; 1\cdot 2 + 1\cdot 1 + 2\cdot 1 \;=\; 5$$
$$C_4 \;=\; C_0C_3 + C_1C_2 + C_2C_1 + C_3C_0 \;=\; 1\cdot 5 + 1\cdot 2 + 2\cdot 1 + 5\cdot 1 \;=\; 14$$
$$C_5 \;=\; C_0C_4 + C_1C_3 + C_2C_2 + C_3C_1 + C_4C_0 \;=\; 1\cdot 14 + 1\cdot 5 + 2\cdot 2 + 5\cdot 1 + 14\cdot 1 \;=\; 42$$

Caterpillar Theorem: $\boldsymbol{C}_n$ is connected.

Proof by induction can start as we saw the first few $\boldsymbol{C}_i$ are connected. Now assume the lemma is true for all $\boldsymbol{C}_i$ with $i < n$, to prove for $\boldsymbol{C}_n$.

We must show every planar halfgraph tree embedding with $n+2$ halfedges is linked to all the others (by some sequence of successive flips). It will suffice to show they are all linked to one such, for instance T_n whose $n-1$ proper edges form a path from halfedge $\mathbf{e}_0$ via proper edges $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}\ldots$ to halfedge $\mathbf{e}_{n+1}$ with the other halfedges branching out each time to the right. By successive flips on $\mathbf{a}$ $\mathbf{b}$, $\mathbf{c}\ldots$ from T_n (whose right branch from $\mathbf{e}_0$ has 1 halfedge, its left branch n of them) we can reach an instance of j halfedges hanging off the right branch from $\mathbf{e}_0$ ($\mathbf{p}$ say) and $n+1-j$ from the left branch ($\mathbf{q}$ say), for any $1 \le j \le n$.



From there we can reach all C_{j-1} possible sub-halftrees on $\mathbf{p}$ as halfedge and j more halfedges $\mathbf{e}_1$ through $\mathbf{e}_j$ (because $\boldsymbol{C}_{j-1}$ is connected by the induction assumption), combined with all C_{n-j} possible sub-halftrees on $\mathbf{q}$ as halfedge and $n-j+1$ more halfedges $\mathbf{e}_{j+1}$ through $\mathbf{e}_{n+1}$ (because $\boldsymbol{C}_{n-j}$ is connected by the induction assumption), that is, all $C_{j-1} \times C_{n-j}$ combinations of possibilities for that j. ■

We saw earlier $\mathbf{Z}$ for halfgraphs has dimension $o + f - 1$ where o is the cyclomatic number of the underlying graph (the halfgraph shorn of its halfedges) and f the number of halfedges; for halftrees $o = 0$ so the dimension is $f - 1$. We also saw $\mathbf{E}^\circ$ has the same dimension as $\mathbf{Z}$. One possible basis for the $\mathbf{E}^\circ$ of a halftree is formed by the halfedges leaving any one of them out, as they are clearly linearly independent: all halfedges form a cut and sum to zero; no subset of them do.

Incidentally the cut formed by the halfedges is the sum of all nodes: this includes proper edges twice so zero times, but halfedges only once. Recall in a halfgraph, unlike in a graph, that's not the same thing as the sum of no nodes — cuts are sums of specific sets of nodes here, not of either part of a partition of the nodes.

Despite appearances, the proper edges in our halftrees aren't bridges at all as the portion on either side is connected to the rest of the world not just by that edge but also by all the halfedges on its side of the divide. The reduced vector of the

edge isn't $\mathbf{0}^\circ$ but the sum of all the halfedges on one side, or equivalently the sum of all the halfedges on the other side of the divide.

In a trivalent halftree, every cut that's the sum of k nodes cuts through at least $k+2$ edges (proper edges and/or halfedges), so this is the ideal ground for representing graph structure by nothing more than $\mathbf{E}^\circ$ and a list of its vectors that are edges.

3.5 All planar trivalent graphs

As before let, for a trivalent graph, h be the integer such that $n = 2h$ and $m = 3h$; now $o = h + 1$. Given a space $\mathbf{E}^\circ$ over $\mathbb{F}_2$ of the right dimension o, when is a set $\mathcal{E}^\circ \subset \mathbf{E}^\circ$ such that there is a graph whose reduced edge space is $\mathbf{E}^\circ$ with $\mathcal{E}^\circ$ the set of reduced vectors representing the single edges? We're now in a position to answer that question. Let $\boldsymbol{\Gamma}_n$ be the set of all such $\mathcal{E}^\circ$ that do, and $\boldsymbol{\Pi}_n \subset \boldsymbol{\Gamma}_n$ the planar ones among them.

Firstly, every connected graph G has a spanning tree T (in general in many different ways). It has $n - 1$ edges ($2h - 1$ in the trivalent case). That leaves a set $\pounds$ of $m - (n + 1) = o$ (in our case $h + 1$) **lianes**, edges of G not in T.

Lemma: The reduced vectors of the lianes form a basis for the graph's $\mathbf{E}^\circ$.

Proof: The number of them is right (as $\dim \mathbf{E}^\circ = o$). Remains to prove they are linearly independent. In spaces over $\mathbb{F}_2$ the only nonzero coefficient in a linear expression could be 1, so we just need to prove no subset of the lianes sums to zero. In $\mathbf{E}^\circ$ that means no subset of them must form a cut. But that is obvious, otherwise they would leave T disconnected. ■

It is instructive to see how an edge $\mathbf{e}^\circ$ that is not a liane (i.e. one in T) appears as sum of the lianes as basis vectors. Every edge of T is a bridge in T so cutting it partitions $\mathcal{N}$ into some $\mathcal{N}_\bullet$ and $\mathcal{N}_\circ$. Now $\pounds \cap \mathcal{E}_{\bullet\circ}$ (those lianes that link the two parts of the tree) form, together with $\mathbf{e}^\circ$ itself, a cut (sum to zero as reduced vectors) so $\mathbf{e}^\circ$ is the sum of the lianes in $\pounds \cap \mathcal{E}_{\bullet\circ}$.

All this is true for graphs in general. From now on, let's restrict ourselves to the trivalent case and inch our way towards $\boldsymbol{\Gamma}_n$ and $\boldsymbol{\Pi}_n$. It will be easier to first assess the extent of the larger sets $\boldsymbol{G}_n$ where we count each pair (G, T), a graph with choice of spanning tree, as separate, and $\boldsymbol{P}_n$ where we do the same for planar graphs (also counting different embeddings separately of those, the less-than-2-edge-connected G, that have more than one).

Starting from the other end, how many bases does $\mathbf{E}^\circ$ have? Counting ordered

bases, there are $2^{h+1}-1$ choices for the first basis vector $\mathbf{e}_0^\circ$ (any nonzero vector), $2^{h+1}-2$ ways to choose $\mathbf{e}_1^\circ$ (any one other than the first) and in general $2^{h+1}-2^i$ ways to choose $\mathbf{e}_i^\circ$ (any vector not in the space spanned by the i vectors chosen already). That's

$$B_{h+1} := \prod_{i=0}^{h} (2^{h+1} - 2^i)$$

bases $\pounds^\circ$, or $B_{h+1}/(h+1)!$ unordered ones.

Next, let's see how many different (G,T) have one such $\pounds^\circ$ as the set of reduced vectors of their lianes. Taking the planar case first, we saw in the previous section there are a Catalan number $C_n = C_{2h}$ different planar halftrees given a fixed cyclic order of their $2h+2$ halfedges. And in section **3.2** we saw there are another Catalan number C_{h+1} ways of pairing up those halfedges in a planar way. That gives

$$C_{2h}C_{h+1} \;=\; \frac{(4h)!}{(2h)!\,(2h+1)!} \times \frac{(2h+2)!}{(h+1)!\,(h+2)!} \;=\; \frac{2\,(4h)!}{(2h)!\,h!\,(h+2)!}$$

different planar (G,T) pairs for each basis, so the whole $\boldsymbol{P}_{2h}$ has B_{h+1} times that many. Of course we're horribly overcounting graphs, as each G has many different spanning trees; $\boldsymbol{\Pi}_{2h}$ is the set of equivalence classes in $\boldsymbol{P}_{2h}$ of (G,T) with the same G. In terms of the vectors, each denizen (G,T) of $\boldsymbol{P}_{2h}$ is a set of $3h$ vectors that are single edges with $h+1$ linearly independent ones marked as lianes; each denizen G of $\boldsymbol{\Pi}_{2h}$ is just a set of $3h$ vectors that are single edges; the equivalence classes are less than $\binom{3h}{h+1}$ or equivalently $\binom{3h}{2h-1}$ in size as not every set of $h+1$ edges in G is linearly independent or equivalently not every set of $2h-1$ edges a spanning tree.

More interesting than the exact number of them is whether $\boldsymbol{\Pi}_{2h}$ is connected, when viewed as a meta-graph with meta-edges between two G's that can be reached from each other by a single flip. We already know the set of graphs that share a planar halfedge pairing pattern (with all possible spanning tree shapes inside) are connected as a meta-graph; it is isomorphic to $\boldsymbol{C}_{2h}$ which is connected by the Caterpillar Theorem (page 35). And any G is a member of many such sets. To prove $\boldsymbol{\Pi}_{2h}$ is connected following on from that we would only need to show *one* graph from the set with any one halfedge pairing pattern is connected to *one* graph each from the other such sets.

Butterfly Theorem: the set of trivalent plane graphs of a given size is connected by flipping.

Proof: One way to represent each of the C_x non-crossing pairings of $2x$ objects arranged along a cycle is as strings of matching brackets: break the cycle at some arbitrary fixed point, choose a fixed direction (anticlockwise say) and then write "(" when (what is now) the first item of any pair is encountered and ")" when the (now) second item comes along. This ploy is well known, see e.g. [CG96] so the salient facts can be stated here without proof: the string consisting of $2x$ brackets (x of each flavor) carries all the information needed to know which one pairs up with which (as well known to any reader). This pairing is governed by what we could call **nesting level**: it is 0 before the string, goes up by one each time we read a "(", and down by one on each ")". If a "(" increases nesting level from λ to $\lambda + 1$, the first subsequent ")" that decreases the level from $\lambda + 1$ to λ is its mate. Valid strings are those for which, firstly, nesting level is 0 again after the string (this is guaranteed by there being equal numbers of both flavor, and there are $\binom{2x}{x}$ such strings) and secondly, nesting level never dips below zero in between (this rules out things like "())(()" and there are several lovely proofs why this condition reduces the number of valid strings by a factor $x + 1$).

To prove connectedness of the set of C_{h+1} pairings of our halfedges it suffices to show they are all connected to one of them. Let's choose "(((···)))" as target; to reach that it suffices to show any string that contains ")(" is connected to the string with that ")(" replaced by "()". We can do even better: given any string that isn't yet equal to the target there will always be a *first* occurrence of ")(" and we can replace that one first. Because valid strings start with "(", that first occurrence can only be preceded by one or more "(". That means it will be part of "()(" and we replace that with "(()". It is easy to see the pairs here: if before the switch some later ")" as in "()(···)" matches the third bracket printed here then, by following the nesting level, after the switch the very same ")" will match the first bracket in "(()···)".

Recall we can choose the spanning tree in any way as all choices are already connected by flips. In the picture below curved segments represent lianes; the slanting and vertical lines are tree edges while horizontal dashed edges can be either.[7]

[7] If there are only two nodes there are only two possibilities, neither strictly speaking graphs (one a triple-edged multigraph and the other a pseudograph with two loops and a bridge). The argument given won't work here but it is easy to see a flip will go from the former to an incarnation of the latter with any of the three edges replaced by the bridge.

The graph on the right is exactly the same as on the left; we've just chopped $\mathbf{b}^\circ$ off the tree making it a liane, replacing $\mathbf{a}^\circ$ as liane which got stuck onto the tree. The pairing pattern on the left is "`()(···)`" and that on the right "`(()···)`" with the same last bracket. ■

Note the proof does not say that in every trivalent plane graph we can choose which edges are lianes, then choose again, and again, until we find a set of lianes that is draped in "`(((···)))`" fashion. What it does say is that for every pairing pattern we can find a tree inside that lets us change the pairing pattern by one step, then find a completely different tree inside that pairing pattern that lets us change it by one more step, and so on, until we reach "`(((···)))`". And because we can get there from any pairing pattern we can get from there *to* any other pairing pattern.

Every trivalent plane graph of a given size can now be reached from any other by flipping. This opens the way for **proving things for all trivalent plane graphs** (if the thing is true for one of them and is preserved by flipping).

The proof does not yet say that $\boldsymbol{\Pi}_{2h}$ is connected, that is, that from every trivalent plane graph of a given size *in any one of its representations as edge list in* $\mathbf{E}^\circ$ we can reach any other such. I'm not ruling out the possibility that $\boldsymbol{\Pi}_{2h}$ isn't connected.

A related question: consider K_4 with $\mathbf{x}^\circ$, $\mathbf{y}^\circ$, $\mathbf{z}^\circ$, $\mathbf{x}^\circ + \mathbf{y}^\circ + \mathbf{z}^\circ$ one of its 4-cycles. Embedded on the sphere the remaining edges $\mathbf{x}^\circ + \mathbf{y}^\circ$ and $\mathbf{y}^\circ + \mathbf{z}^\circ$ are on opposite sides of this cycle; they could not be on the same side given planarity. This opens an intriguing prospect. Can we generalise this — given a cycle in any plane graph, can we tell somehow from the reduced vector immediately which side of the cycle, if the vector is an edge, that edge would be? How is the idea of one or the other side expressed by the vectors? Is it related to (counter)clockwice orientation of edges at a node? How does it break down in 2- and 1-edge-connected graphs?

3.6 All trivalent graphs

The number of (G, T) pairs in $\boldsymbol{G}_n$ or G's in $\boldsymbol{\Gamma}_n$ (that is, no planarity assumed) can be calculated by similar methods. We start with the same B_{h+1} bases; in stead of

C_{h+1} planar pairings we get all $(2h+1)!!$ possible pairings; on the other hand we get fewer spanning trees because their embedding in the plane no longer matters.

I haven't yet proved the connectedness by flipping of the non-planar equivalent of $\boldsymbol{C}_n$ here, let alone that of all of $\boldsymbol{\Gamma}_n$. It seems highly likely they are connected though, and that shouldn't be too hard to prove given a little time.

Recall in the non-planar case there are each time two more alternatives for an edge. Given a graph and a cycle double cover, at every edge one of the two alternatives gives the new graph automatically a cycle double cover again by re-routing the cycles, and the other one doesn't (it would cover the new edge four times). So given only a small set of graphs with known cover (at least one such pair for each possible genus) together with a stronger kind of connectivity (using only the flips that preserve the cover) might lead to a proof of the cycle double cover conjecture.

4 Tait colorings

4.0 Colorings

A **coloring** of certain items, such as nodes or edges of a graph, or faces of a graph embedding, is a mapping from those items to some set X, where the elements of X are traditionally given color names. An x-coloring is one where $|\mathrm{X}| = x$. Usually, there is an additional restriction on colorings. Unless stated otherwise, the restriction is that **adjacent** items must have different colors (with the relevant definition of *adjacent*).

For edges, *adjacent* is interpreted as being incident to the same node; for nodes, as being incident to the same edge. As a loop makes a node adjacent to itself, and is an edge adjacent to itself[8] pseudographs can have neither edge- nor node-colorings by this definition.

When colorings do exist, the phrase "G **needs** x colors" means that no coloring (of nodes, edges etc. as clear from the context) of G exists with fewer than x colors, but there does exist one or more x-coloring.

For edge colorings there is a celebrated theorem proven by G. I. Vizing in 1964, while still a grad student (in Novosibirsk). If Δ is the largest valency occurring in G, there can obviously not be a coloring using fewer colors[9], so any graph needs at least Δ different colors. The theorem states that G never needs more than $\Delta + 1$ colors. For the proof see e.g. [Ore67], [FW77], [Wil02], or the relevant encyclopædia entry by my hand in [PM∞].

It is a special case of Vizing's theorem for multigraphs (1965) which states no multigraph needs more than $\Delta + \mu$ colors where μ is the largest number of edges between the same two nodes.

For graphs, the theorem has the effect of dividing them into two kinds: those of "class I" needing Δ and those of "class II" needing $\Delta + 1$ colors. The classification problem (which graphs are in which class) is unsolved in general. Many particular kinds of graphs can be proven to be of "class I", needing only Δ colors. Notably, bipartite graphs by König's theorem [FW77, Wil02].

[8] We'll see a justification for this point of view in the next footnote.

[9] Let N be a node which attains the maximum valency. Any two edges adjacent to each other by being both incident to N are required to have different colors so we already need Δ colors. This argument would break down if two of the valencies of N are accounted for by a loop, unless we also call a loop adjacent to itself.

4.1 Planar face-4-coloring

Francis Guthrie was born in England in 1832, obtained firsts in both mathematics and law, taught mathematics in South Africa until his death in 1899 and also found the time to get involved with railroad construction and to make his name in botany. One day he noticed something when coloring a map of the counties of England — the original four-color conjecture. And on 23 October 1852, in London, his brother Frederick asked his math professor Augustus de Morgan about that observation ("With my brother's permission"). De Morgan consulted Hamilton ("A student of mine asked me to day to give him a reason for a fact which I did not know was a fact — and do not yet."). The latter didn't think much of it and quipped he didn't have time for that "quaternion of colours". The problem didn't go away though, Cayley dug it out again, Kempe gave his flawed proof that stood for 11 years (and the still useful concept of Kempe chains), and the rest is history [FF94].

The four-color conjecture (now four-color theorem) has loomed large over graph theory and a lot of effort in the field was partly motivated by it. When Kenneth Appel, Wolfgang Haken and John Koch [AHK77] finally found a proof it was instantly recognised as one of the all-time landmarks of mathematics[10]. As with all large proofs there was some discussion afterwards and minor details were cleaned up. One thing that exercised many was the use of a computer (then a novelty) to both identify and deal with about 2000 separate cases. Nowadays the view has gained ground that very large proofs are always hard to check by the community, no less so if they were written by a human in longhand. If anything, the use of computers can make it *easier* to spot errors.

Perhaps it has been fortunate the theorem resisted proof so long, because it has been a motivating force for many exciting developments in graph theory. This includes important results in graph coloring such as Vizing's theorem, but also work that transcends the bounds of graph theory, such as that by Whitney, MacLane, Tutte et al. on what is now called matroids. Let's finish this section by looking at the theorem, and why it was enough to prove it for **trivalent** graphs.

Four-Color Theorem (Appel, Haken, Koch 1976): Every trivalent bridge-free planar graph can be face-4-colored.

[10]The proof was found in 1976 and announced in [AH76]. For a while, the University of Illinois stamped outgoing letters with FOUR COLORS SUFFICE. The details followed in a lengthy two-part paper [AHK77]. I briefly describe the method in the next footnote. Very readable accounts of the proof can be found in [SK77, FF94, Wil02].

Corollory: Every bridge-free planar graph can be face-4-colored.

To see why the corollary is equivalent we must first note that *adjacent* for faces doesn't mean merely sharing a node; only sharing one of more edges counts. If we allowed faces to be "adjacent" (and demand them to be colored differently) if they only shared a node the problem how many colors were needed would become uninteresting; a pie chart of enough segments would need any number of colors.

Now that we only demand faces to have a different color if they are adjacent along an edge, consider a 4-valent node $\times$ in a planar embedding of a planar graph, and a small boundary adjustment $\rangle\!\vdash\langle$ near that node (we split the 4-valent node into two 3-valent ones and run an edge between the two). Before the adjustment, each of the North and South faces were adjacent to each of the East and West ones and that was all. After the adjustment North becomes adjacent to South as well. We keep all the existing restrictions (which face cannot have the same color as which) and add one more such restriction, so the problem only becomes harder: if the new map can still be face-4-colored the old one definitely can (with the same colors assigned to the same faces even). In the same way, any v-valent node ($v > 3$) can be split into $v - 2$ trivalent ones only making the coloring harder.

We can omit 0-valent nodes (merely funny features in the landscape) and 2-valent border stones B (merging two edges AB and BC to one AC) without affecting the surrounding face coloring. And 1-valent nodes give rise to bridges which we must demand do not occur, because they have the same face at both sides so make a face adjacent to itself, and you cannot color a face differently from itself.

The only valency left now is 3, and we saw that (for any x) if the trivalent graph we get after adjustment can still be face-x-colored, the original graph certainly can. This shows why it is enough to prove the theorem for trivalent graphs.

Another thing that comes to mind when considering the origins of the problem is that Guthrie attempted to color regions of a **bounded** portion of the surface. Suppose you color the countries of a continent (or indeed counties of an island such as Britain). Countries bordering the coast can have any of the 4 available colors. If you now consider the sea to be a country as well, call it Oceania, and give it one of the four colors (WLOG blue) the coastal countries are restricted to the three remaining colors (just like any other set of countries with a shared neighbour). So this too only has the effect of making the coloring harder. It suffices to prove we can color a map where every area, including the outer one, counts as a face to be colored (then we can certainly color maps on bounded portions of surface).

Finally, coloring a graph embedded on a **sphere** is equivalent to coloring the same graph embedded on a **plane**, now that we color Oceania too. To go from sphere to plane puncture a hole in the interior of any face and call it the outer face of the plane map. Sphere embeddings are essentially simpler because each time one sphere map with f faces punctures to f seemingly different plane maps. And of course, the problem was already put in terms of spherical embeddings right from the start: Guthrie's map of England represented a portion of the Earth's surface.

The nature of the Appel-Haken-Koch proof[11] is such that nobody is left with a feeling they can see **why** it had to be that way. What is it about planarity that forces a coloring, while even some quite simple non-planar graphs (such as the Petersen one) have none? A short "aha!" proof would still be welcome.

The rest of this chapter divides in two parts. Sections **4.2**, **4.3** and **4.4** describe three alternative but equivalent formulations of face-4-colorability for planar trivalent graphs. Each time different features of the graph are colored and each time the number of colors drops by one. These formulations were known in the 19th century but are presented here in a unified framework as a discrete analogue of vector differentiation and integration, using elements of $\mathbb{F}_4$ as color names. For the first two steps, the equivalence holds for graphs embedded in surfaces of any genus when going down (the differentiation step); it is going back up (the integration step) that requires planarity. The last step is a bit different.

Section **4.5** describes these four equivalent colorings in terms of vectors in edge spaces over $\mathbb{F}_4$ and shows how they relate to the spaces over $\mathbb{F}_2$ we've been working with. Admissable colorings form a non-space subset to which methods from **coding theory** can be applied. The various subsections of **4.6** show some of the forms such an approach could take. The 4-color theorem is shown to be equivalent to a bizarre variety of statements about cycle-shaped words in codes over $\mathbb{F}_4$ or $\mathbb{F}_3$, many of them tantalisingly simple to state (but not, as yet, to prove).

---

[11] About 2000 features (clusters of adjacent faces with specific numbers of sides) are identified that are both **unavoidable** (every trivalent bridgefree graph has at least one of them) and **reducible** (to an alternative feature with fewer faces such that the original graph can be colored if the alternative can — enabling induction on the number of faces). The computer programs kept adjusting sets of unavoidable and reducible features until they coincided.

Reducibility of each feature is straightforward. Unavoidability is a property of the whole set. By Euler's formula the numbers of sides of f faces must sum to $6f - 12$ (for example twelve 5-gons and the rest 6-gons). The missing 12 can be thought of as a quantity of **charge** (the discrete analogue of the sphere's curvature) that can be **discharged** (moved around by changing the graph) but is conserved globally, so if you don't have *this* you must have *that*, or *that*...

The number of different features required has been whittled down since. Robertson, Sanders, Seymour and Thomas (1997) found a version that needs about 600.

4.2 Tait's edge-3-coloring

For trivalent graphs $\Delta = 3$, so Vizing's theorem splits them into those that need 3 edge colors (class I) and those that need 4 (class II). The ubiquitous Petersen graph is an example of the latter kind. Trivalent graphs of class II (excluding certain trivial cases) were dubbed **snarks** by Martin Gardner because they are so hard to find (the word comes from *The Hunting of the Snark* by Lewis Carroll).

There is a nice result by Peter G. Tait (1831–1901, Scottish physicist and mathematician) that trivalent graphs with a **Hamiltonian cycle**[12] only need 3 colors. Paint the edges of the cycle alternatingly red and green; now the third edge at each node can become blue. This works because the number of nodes in a trivalent graph is even (to satisfy $3n = 2m$). Edge-3-colorings of trivalent graphs are called **Tait colorings** in his honor. Another theorem by Tait, linking edge-3-colorings to face-4-colorings, appears further down in this section.

Decomposing trivalent graphs by their edge connectivity we have that

- disconnected trivalent graphs (for instance a 2-component graph $H\ \ K$) can be Tait colored iff all the components (here H and K) can. This is true for any edge- or node-coloring: any restrictions on colorings due to adjacency happen within a component.
- trivalent graphs $H—K$ containing a bridge can never be Tait colored. We'll see an algebraic proof shortly.
- trivalent graphs $H{=}{=}K$ with edge connectivity 2 can be split into their constituents, each having its two half-edges joined to a single proper edge. We will see below that if a Tait coloring of $H{=}{=}K$ exists it must give the edges in the 2-cut the same color. Therefore $H{=}{=}K$ can be Tait colored iff both $H\supset$ and $\subset K$ can (permute the colors in one of the graphs as necessary to make the match).
- trivalent graphs $H{\equiv}K$ with edge connectivity 3 can be split into their constituents, each given an extra node to tie the three halfedges together. We will see the edges of the 3-cut must have all three colors once, so $H{\equiv}K$ can be Tait colored iff both $H\ni$ and $\in K$ can (permute to match).

Note that if the graph is solid (as defined on page 8), i.e. if the only 3-cuts occur around nodes, that last splitting operation doesn't actually make the graph smaller

[12] A cycle visiting all the nodes of the graph.

so would be useless for induction arguments. So the question which trivalent graphs have Tait colorings reduces to the question which solid ones do. Then any more weakly connected ones can be split into their solid cores to find the answer.

The classification into classes I and II is as unsolved for trivalent graphs as it is for graphs in general. For **planar** graphs however the answer is known. Surprisingly, it is the same problem as that in the preceding section.

Theorem (Tait, 1880): A trivalent plane graph can be face-4-colored iff it can be edge-3-colored. [Tai80, Tai80']

I will prove this using elements of the finite field $\mathbb{F}_4$ as color names, for reasons that will become clear. The face coloring will use all four elements and the edge coloring only the three nonzero ones.

$\mathbb{F}_4$ has four elements customarily called 0, 1, ω and $\overline{\omega}$. As far as addition is concerned $\mathbb{F}_4$ behaves like a 2-dimensional vector space over $\mathbb{F}_2$ where any two nonzero elements can be taken as the basis, their sum is the third. In particular $\mathbb{F}_4$, like $\mathbb{F}_2$, has "characteristic" 2 which means anything added to itself becomes zero, and subtraction equals addition. The multiplication of the three nonzero elements is cyclic, with each of ω and $\overline{\omega}$ being the inverse of the other, and the square of the other. Multiplication by 0 of course gives 0.

To prove a face-4-coloring implies an edge-3-coloring, we employ a process of **differentiation**. Each edge gets a color that's the difference (equivalently, the sum) of the colors of the faces it separates. Edge color is clearly an element of $\mathbb{F}_4$ this way, and because adjacent faces have different colors it becomes a *nonzero* element of $\mathbb{F}_4$, hence only three edge colors get used.

Remains to prove the edge coloring is valid, that is, that it gives different colors to adjacent edges. In other words, that the edge colors a, b and c at a node are all different. Travel round the node. Let the first face color be x. The next (after crossing the edge with color a) is $x+a$. The next is $x+a+b$. The next $x+a+b+c$. But wait! That's the one we started with, it had color x. So the sum of all the differences one time round, $a+b+c$, must be zero. And by inspection it is easy to verify the requirement that $a+b+c$ is 0 while none of a, b, c themselves are is only satisfied by them being one each of the three nonzero elements. ■

Note the proof that way round is valid for graphs embedded on any surface. The converse implication requires planarity.

To prove an edge-3-coloring implies a face-4-coloring we would employ a process

of **integration** — and indeed, "plus a constant": we can start by giving the first face an arbitrary color. From there on, travel around and accumulate the color of each edge crossed into a sum. The total reached at each stage is the color of the face we're in.

Differentiating a "potential", a function of spatial position, gives a "gradient", but conversely not everything is a gradient that can be integrated to a well-defined potential. The same holds for our discrete toy potential and gradient. We must still prove that the face color accumulated this way is unique, independent of journey taken. Or what's the same thing, that closed journeys always sum to zero. Colors of journeys here being the sum of colors of edges crossed. Note [closed] journeys are [closed] paths in the dual graph; here color is the sum of those of edges travelled.

One way of proving it is by noting that (and here planarity comes in) any closed journey encloses a bunch of nodes (or its complement, on a sphere) and the closed journey can be decomposed as the sum of little closed journeys each around a single node (tile up the area enclosed so each tile has one node). When tiling the area enclosed by closed journeys the colors do add: let areas A and B be enclosed by closed journeys $P \cup Q$ and $Q \cup R$ respectively, Q being the whole shared portion. The boundary of the combined area is $P \cup R$, and Q also drops out of the sum of colors because it is added twice. Once we're down to single node journeys, we already know the journey around each node amounts to color 0. ■

After proving the above theorem, and noting Hamiltonian paths ensure edge-3-coloring (p 45), Tait tried hard to show every trivalent planar graph has a Hamiltonian path. Unfortunately, some don't, as shown much later by Tutte. But a Hamiltonian path, single-handedly visiting all the nodes, isn't necessary. If we find a set of disjoint *even*-length cycles that together visit all the nodes, we can still paint these with alternating red-green edges, and the remaining edges blue. Conversely, if there is a Tait coloring then the red and green edges together will form such a set of one or more cycles, and so do the blue and red edges, and again the green and blue edges. Incidentally, all three sets together then also form a cycle double cover.

4.3 Heawood's node-2-coloring

Percy Heawood was a contemporary of Tait and did much work in graph colorings, including finding the numbers of colors needed to face-color graphs embedded in surfaces of any genus except 0 (curiously, this problem was solved long before the

planar case). Here we will look at a feature of planar graph colorings discovered more than once independently, but probably first by Heawood.

Definition: a **Heawood coloring** is a node-2-coloring (black and white say) of a plane graph where, for once, the condition for a coloring to be valid is *not* that adjacent ones have different colors, but that the number of black and white nodes in every *face* cycle are equal (mod 3).

Theorem: A planar trivalent bridge-free graph can be edge-3-colored iff it can be Heawood node-2-colored.

For embeddings of trivalent graphs in the plane (or any orientable surface) Heawood node coloring derived from Tait edge coloring can be described very simply: color those nodes black where edge colors red, green, blue occur anticlockwise, and those white where it happens clockwise. There is another way to say this using the cyclic multiplication of edge colors 1, ω and $\overline{\omega}$. There are two kinds of nodes:

- Those where, if you travel through it by turning left, you see edge color getting multiplied by ω (and therefore by $\omega^{-1} = \overline{\omega}$ if you turn right).
- Those where, if you travel through it by turning left, you see edge color getting multiplied by $\overline{\omega}$ (and therefore by $\overline{\omega}^{-1} = \omega$ if you turn right).

To prove edge-3-coloring implies Heawood node coloring we can again use **differentiation** (of the discrete log of field elements, an integer mod 3). By inspection it is clear that given three different nonzero $\mathbb{F}_4$ colors at the three edges of a node, it doesn't matter which edge you come from: the quotient of edge colors encountered on turning left will be the same, for any one node. Here too the number of colors drops by one on differentiation: because successive edges never have the same color the quotient is never 1. We're left with colors ω and $\overline{\omega}$ for nodes.

To prove the node coloring derived this way has the Heawood property, consider that the product of node colors around a face cycle (a single face of the embedding, a face we can perambulate by turning left each time) must be 1, to get the same edge color back after the round trip. As that accumulated product of all the quotients is ω^{b-w} where b and w are the numbers of black and white nodes in the cycle, $b - w \equiv 0 \pmod 3$ is indeed satisfied. ■

Note the proof that way round is valid for graphs embedded on any orientable surface. The converse implication requires planarity.

To prove Heawood coloring implies Tait coloring we must **integrate** again, "plus a constant". Assign any color to the first edge and start walking. To prove a single

face cycle does sum to zero (so the edge gets the same color after a round trip) given the Heawood property is immediate: $\omega^{b-w} = 1$ when $b - w = 0 \pmod 3$. To prove any cycle sums to zero consider that the area it encloses exists in the planar case and can be decomposed into single cycles. The summing process is interesting: if the contribution of a node is ω say (on turning left) then enclosing two of the faces there causes us to go right, but ω^{-1} is indeed the same as ω^2. And enclosing all three faces inside a larger cycle contributes $\omega^3 = 1$ (i.e. no contribution to the product) as it should. ■

Note that if every face of a trivalent graph has a number of sides divisible by 3 (call this a **threefold** graph), it has at least two valid Heawood colorings: just make all nodes white or all black (it is not certain these are the *only* valid colorings, for instance a hexagon face allows three black and three white nodes). Conversely if a coloring is valid and uses only one color, the graph *must* be threefold.

Now consider the following method of **tweaking** an arbitrary trivalent graph that is already validly Heawood colored. We set out to make all nodes black. Replace every white node by a little triangle consisting of 3 black nodes (alternatively, if we happen to come across an all-white triangle at any stage we could contract it to a black node). It is easy to see this preserves the congruences in all neighbouring faces. So when we've finished (and all the nodes are black) the coloring is still valid. But we just saw that means the tweaked graph is threefold. Of course, we could also have set out to make all nodes white, with the same result.

Finally consider a graph of which we don't know a coloring. We don't know which nodes to replace with triangles. But *if* there exists any choice of nodes to tweak that turns the graph into a threefold one, *then* we can Heawood color that in monochrome and trace the tweaks backward to find a coloring of the original graph.

Now we see yet another equivalent formulation of the four-color theorem: a trivalent graph can be colored thus iff it is possible to tweak it into a threefold graph by judiciously adding (or deleting) triangle faces.

4.4 The elusive 1-coloring

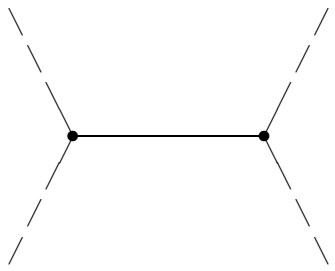
After applying discrete differentiation twice, getting the number of colors down from 4 (for faces) via 3 (for edges) to 2 (for nodes) it looks tempting to find yet another differentiation argument making the four-color theorem equivalent to the existence of 1-colorings of some other graph elements. After that it shouldn't be hard, surely, two prove 1 color suffices for those elements, opening the way to

possibly a straightforward alternative proof of the theorem. Unfortunately, we seem to have run out of kinds of elements to paint.

The reason the number of face colors dropped by one going to edge colors is that the latter were differences of the former and never 0 because *adjacent* faces have distinct colors. Again, the number of colors dropped by another one going to Heawood node colors because they were ratios and never 1 (equivalently, their discrete logs never 0) because *adjacent* edges have distinct colors. Those reasons would break down at the next step at any rate, because adjacent nodes (by the usual definition of adjacent) can shamelessly have the same Heawood color.

Taking a step back for a more general view, it appears that the link between face-4-colorings and edge-3-colorings hinged on the *additive* properties of $\mathbb{F}_4$ (and the color that dropped out is the additive identity, 0) and the link between edge-3-colorings and Heawood node-2-colorings hinged on the *multiplicative* properties of $\mathbb{F}_4$ (and the color that dropped out is the multiplicative identity, 1). There is a third feature of finite fields, the Frobenius automorphisms. In $\mathbb{F}_4$ addition is a group of all four face colors, multiplication is a group for the three edge colors, and the automorphism group has two elements (the identity, and *conjugation* that fixes 0 and 1 but swaps the two Heawood node colors ω and $\overline{\omega}$). If there is a final step the identity that drops out would be that of the automorphism group, and the single "color" left that of conjugation.

And this final step exists. A graph with $4!T$ face colorings (for some T) has only T essentially different colorings (under permutation of color names). The $4!T$ distinct face colorings show up as only $3!T$ distinct edge colorings and $2!T$ distinct Heawood colorings. At that stage we have a bipartition of $\mathcal{N}$ into some $\mathcal{N}_\bullet$ with color ω (red/green/blue edges counterclockwise) and $\mathcal{N}_\circ$ with $\overline{\omega}$ (clockwise). That is, a cut $\mathcal{E}_{\bullet\circ}$. By leaving it immaterial which are the "black" and which the "white" nodes, the cuts follow the spirit of the other two differentiations leaving only T distinct cuts. An example. A piece of plane graph



has 4! face colorings that use every color once, and another 4! that leave one color

unused, and give the same color to the two non-adjacent faces (so T is 2 here). When we move to edge colorings the numbers get divided by 4: there are two sets of 3! colorings now. That divides further by 3 when we move to Heawood colorings where there are twice 2! colorings (at each node independently the edge colors could go clockwise or counterclockwise). It gets divided finally by a factor 2 to only two essentially different ways to do the coloring (the nodes turn the same way so the central edge is not in the cut, or the opposite way so the central edge is in the cut).

So now our colorings are certain elements of $\mathbf{Z}^\perp$. There doesn't seem to be any obvious algebraic criterion which cuts, nor an induction argument for them. We'll come across them again as an instance of *cisness structures* in the next chapter.

4.5 Spaces and codes for Tait colorings

We saw that (for bridgeless trivalent planar graphs) face-4-colorings, Tait edge-3-colorings, Heawood node-2-colorings, and now what we could call Heawood cuts all imply one another. Let's refer to the whole syndrome of interrelated colorings as a Tait coloring of the graph.

Let $\boldsymbol{E}$ be the vector space over $\mathbb{F}_4$ analogous to $\mathbf{E}$ over $\mathbb{F}_2$. That is, it has a basis that stands in 1–1 correspondence with the set $\mathcal{E}$ of edges but we allow scalars ω and $\overline{\omega}$ as well as 0 and 1 now.

Taking the sum of edges of a cycle (each with the same coefficient 1) again as the vector representing a cycle, and the same for a cut, we get cycle space $\boldsymbol{Z}$ and cut space $\boldsymbol{Z}^\perp$. These have the same dimensions as the corresponding $\mathbf{Z}$ and $\mathbf{Z}^\perp$, because it is easy to see there are no new linear dependencies between the relevant basis vectors: for any $\boldsymbol{v}$ in $\boldsymbol{E}$,

$$\boldsymbol{v} = \sum_\iota v_\iota \boldsymbol{e}_\iota \; = \sum_\iota (x_\iota + \omega y_\iota)\boldsymbol{e}_\iota \; = \sum_\iota x_\iota \boldsymbol{e}_\iota + \omega \sum_\iota y_\iota \boldsymbol{e}_\iota + \omega \; := \; \boldsymbol{x} + \omega \boldsymbol{y}$$

where a_ι and $b_\iota \in \{0,1\} \subset \mathbb{F}_4$. Now any scalar multiple of a basis vector $\boldsymbol{z}_j$ of $\boldsymbol{Z}$ decomposes with its $\boldsymbol{x}$ either $\boldsymbol{z}_j$ or $\mathbf{0}$ and its $\boldsymbol{y}$ also $\boldsymbol{z}_j$ or $\mathbf{0}$, so any $\mathbb{F}_4$-linear relation between the $\boldsymbol{z}_j$ would imply two independent relations on their $\boldsymbol{x}$'s and $\boldsymbol{y}$'s separately, acting as $\mathbb{F}_2$-linear relations between the basis vectors $\mathbf{z}_j$ of $\mathbf{Z}$, which we know doesn't happen. Curiously, the multiplication of $\mathbb{F}_4$ doesn't really come into this. Of course not every subspace of $\boldsymbol{E}$ is an extension of a subspace of $\mathbf{E}$ in this way. Our $\boldsymbol{Z}$ and $\boldsymbol{Z}^\perp$ are, due to the special form of their basis vectors.

Using 1, ω and $\overline{\omega}$ again for the three Tait colors, an edge-3-coloring is of the form

$\sum v_\iota \boldsymbol{e}_\iota$ where edge ι gets color $v_\iota \in \mathbb{F}_4$, in other words the whole coloring is a vector in $\boldsymbol{E}$. The requirement for the coloring to be valid, i.e. that for edges α, β, γ occurring at one node v_α, v_β, v_γ must be 1, ω and $\overline{\omega}$ in some order, can be teased apart into two conditions:

- The three edge colors occurring around a node must sum to 0. This condition on its own reduces the $4^3 = 64$ possibilities of assigning colors (from all of $\mathbb{F}_4$ including 0) to $4^2 = 16$: apart from 1, ω and $\overline{\omega}$ occurring once each (in six permutations) we could also have any of them occurring twice with 0 as third color (nine ways of doing that), or thrice color 0 (in one way).
- Edge color 0 must not occur anywhere. This condition now reduces the 16 further to just those 6 arrangements we want.

Usually one would take those conditions the other way round (firstly that we only use three colors, and secondly that there's one each at a node). Doing it as listed here has the advantage that the first condition is pure linear algebra. The set of vectors satisfying it is a subspace, and it is one we already know: it is none other than $\boldsymbol{Z}$, because the condition amounts to a vector being $\perp$ the basis vectors of $\boldsymbol{Z}^\perp$, and hence to all of $\boldsymbol{Z}^\perp$. The condition also tells us straightaway that edge colors of any cut must sum to zero. This is the algebraic reason for the assertions about 3- and 2-cuts in the preceding section; it also shows why a graph with a bridge cannot have a Tait coloring (the bridge being a cut would get color 0).

The second condition has a coding theory flavor. Recall m-dimensional vectors can be described as **code words** of m **letters**, notation $abc\ldots$ rather than $(a, b, c\ldots)$, and that a **code** is a subset of the space. The concept of a code is wider than that of a vector space as the letters need not be elements of a field. If they are, and the code is a subspace, we call it **linear**. Important concepts are **weight** of a word (the number of nonzero letters in it) and **distance** between words (the number of places where they differ). For a linear code the difference $\mathbf{u}-\mathbf{v}$ exists, the distance between $\mathbf{u}$ and $\mathbf{v}$ is its weight (and conversely the weight of any $\mathbf{w}$ is the distance between it and $\mathbf{0}$) so distances and weights take on the same set of values. The second condition now says that the coloring is a code word of $\boldsymbol{Z}$ with weight m. Let the **Tait code** $\mathcal{Z} \subset \boldsymbol{Z}$ of a trivalent graph be the set of those code words of weight m. It too is a code (just not a linear one); any set of code words is.

Let an **extended Tait coloring** be an assignment of the colors 0, 1, ω, $\overline{\omega}$ to the edges that only satisfies the first condition (i.e. nothing other than a word of the

linear code $\boldsymbol{Z}$, the cycle space over $\mathbb{F}_4$) and a **Tait coloring**, as before, one that also satisfies the second condition (a word of $\boldsymbol{Z}$ of weight m, a word of $\mathcal{Z} \subset \boldsymbol{Z}$).

The question whether some G has a Tait coloring now resolves to the question whether its $\boldsymbol{Z}$ has a word of weight m (equivalently, whether its $\mathcal{Z}$ has any words at all).

Note in passing that $\boldsymbol{Z}$ over $\mathbb{F}_4$ brings us back full-circle to face-4-colorings: using a basis for cycle space consisting of all faces $\boldsymbol{z}_i$ of the embedding except one face $\boldsymbol{z}_+$ (which is then the sum of the basis cycles) the face coloring that assigns c_i to each $\boldsymbol{z}_i$ corresponds to the edge coloring represented by a vector $\boldsymbol{v} = \sum_i c_i \boldsymbol{z}_i$ (adding a constant to each c_i gives a different face coloring with the same edge coloring).

4.6 Variations on a theme

4.6.0 Taking stock

The big question that comes to mind when considering Tait coloring trivalent graphs is of course which graphs can be colored that way — and why. The most basic incarnation of a Tait coloring must be as its edge-3-coloring, because in this form the question is applicable to every trivalent graph regardless of any embedding. If the graph comes with an embedding in a surface (so "faces" are well-defined) we saw face-4-colorings imply edge-3-colorings; if the surface is orientable we saw an edge-3-coloring implies a Heawood node-2-coloring (and a corresponding Heawood cut "1-coloring"); finally if the surface is planar the reverse implications are true as well. Let "a Tait coloring" refer to this whole syndrome of related colorings.

The problem is solved for bipartite trivalent graphs (edge-3-colorings exist) and for trivalent graphs with bridges (here they don't exist) and in both these cases we can easily see why. It is also solved for planar trivalent graphs by the 1976 Appel-Haken-Koch proof, but the nature of that proof left some appetite for a simpler proof. What kind of approaches could one try? We should ask

- what it means to be planar. We looked at this in Chapter **2**, in terms of embeddings in surfaces. The fundamental topological property of surfaces of genus 0 is that any closed curve on them can be contracted continuously to a single point. The discrete version of this property featured in the proofs that a face-4-coloring is implied by an edge-3-coloring, and an edge-3-coloring by a Heawood node-2-coloring.

When building a graph with the bipolar growth theorem, planarity in a similar way maintains a cyclic order of the edges that cross the "equator" at any stage of the build. This is one property of planar graphs we could exploit.

Chapter **2** briefly mentioned other hallmarks of planarity, such as the absence of K_5 and $K_{3,3}$ minors (as $K_{3,3}$ is edge-3-colorable it is likely only the absence of a K_5 minor matters) and having a cycle double cover of $h + 2$ faces.

- what it means (for a planar graph) to have a Tait coloring. We should look at various ways to express the existence of such a coloring. We already saw
  - the face-4-coloring version in **4.1**,
  - the edge-3-coloring version in **4.2**,
  - a node-2-coloring version in **4.3**,

  and how using elements of **the field** $\mathbb{F}_4$ as color names brings out the relation between these colorings. In **4.3** we also saw the existence of a coloring is
  - equivalent to the graph being able to be transformed (by changing nodes into triangles) into one where all the faces have 3, 6, 9, 12... edges,
  - and finally in **4.4** how the coloring is equivalent to a certain cut.

  The rest of this chapter will be devoted to describing yet other variations.

  Note one can of course also look at the dual plane graph, a triangulation if the original is trivalent, where the problem is couched in terms of node-4-colorings. The 1976 proof is usually described in this form [Wil02].

More globally, what overall strategy could the proof use?

- We could use induction over the (infinite) set of (finite) trivalent bridge-free graphs by size (the 1976 proof did that).
- We could also build an arbitrary trivalent graph (for instance via the bipolar growth theorem) and attempt to prove that if colorings exist at one stage they still exist at the next stage. Some of the suggestions below follow this route.
- Yet another approach would be by contradiction: proving that if a trivalent bridge-free graph cannot be Tait colored it isn't planar (e.g. has a K_5 minor),
- or that if a trivalent planar graph cannot be Tait colored it has a bridge.

4.6.1 The constructive view

We already know a trivalent graph with bridges can't be Tait colored, so assume bridge-freeness. We may also assume connectedness (otherwise just deal with each component separately). Now our graph is 2-edge-connected, so we can build it along the lines of the bipolar growth theorem (section **3.1**). At each stage we will need to keep track of valid colorings for the Southern graph $G^i_\bullet$ thus far, or of the Southern halfgraph $H^i_\bullet$. The latter will turn out to be easier (it is already trivalent at each node).

A Tait-coloring of a trivalent halfgraph can be defined the same way as for a trivalent graph. From now on, let "edge" be shorthand for halfedges and proper edges alike. At each node the three edges must get distinct colors. Equivalently, by the twin criteria listed earlier, colors (as $\mathbb{F}_4$ elements) sum to 0 at each node, and no edge individually has color 0. We're not going to cap the halfedges with 1-valent nodes (restricting edge color to sum to 0 at those nodes too would give all halfedges color 0). So the appropriate cycle space $\boldsymbol{Z}^i_\bullet$ over $\mathbb{F}_4$ to take colorings from is that as defined for halfgraphs (section **1.4**).

There we saw that (for a connected halfgraph with $f > 0$ halfedges) the dimension of the cycle space $\boldsymbol{Z}^i_\bullet$ is $f - 1$ more than the dimension o of the cycle space of the underlying graph (obtained by capping the halfedges with 1-valent nodes). So a subspace of $\boldsymbol{Z}^i_\bullet$ of dimension o is spanned by the actual cycles, and the remaining $f - 1$ dimensions are provided by paths starting and ending at a halfedge. Our $G^i_\bullet$ has i nodes; with e proper edges and f halfedges we have $3i = 2e + f$, or equivalently for the underlying capped graph with i trivalent and f univalent nodes, and $e + f$ edges, $3i + f = 2(e + f)$. The cyclomatic number o of the latter is now $(e + f) - (i + f) + 1 = e - i + 1$ so the dimension of the halfgraph's cycle space is $e + f - i$. That agrees with $e + f$ for the entire edge space and i for the cut space (recall nodes aren't linearly dependent because there are halfedges).

As we build, we never lose edges or adjacency between them (some halfedges become proper edges and we get new halfedges) so each next $G^{i+1}_\bullet$ can only have colorings that assign to those edges already present in $G^i_\bullet$ colorings that were valid there. Some of those may survive in $G^{i+1}_\bullet$ (perhaps even extended in more than one way) and some may not. Coloring is an incremental process this way.

By the same token the whole graph can only have a valid Tait coloring if all the $G^i_\bullet$ we encountered had one; they were just colorings of part of the graph. And we will see below that if we get as far as $G^{n-1}_\bullet$ we can also color the whole graph.

Here too, let an **extended Tait coloring** be any element of $\boldsymbol{Z}_\bullet^i$ (we only demand that colors add to 0 at each node) and the **Tait code** $\mathcal{Z}_\bullet^i$ be the set of valid Tait colorings, the code words of full weight (here $e+f$). Proving the theorem amounts to showing that for each i there are still such codewords left.

4.6.2 Equatorial codes

Let the **equatorial code** $\boldsymbol{O}_\bullet^i$ consist of the code words of $\boldsymbol{Z}_\bullet^i$ with all letter positions deleted except for the ones that refer to halfedges. In terms of vectors, we're projecting the $(e+f)$-dimensional edge space $\boldsymbol{E}_\bullet^i$ to f of its dimensions, and see what's left of the subspace $\boldsymbol{Z}_\bullet^i$ of $\boldsymbol{E}_\bullet^i$ under this projection. Some code words may become the same under the projection. The vectors of $\boldsymbol{O}_\bullet^i$ can also be regarded as equivalence classes of vectors of $\boldsymbol{Z}_\bullet^i$.

Lemma: the colors of the halfedges in any extended Tait coloring (the letters in any word of $\boldsymbol{O}_\bullet^i$) sum to zero.

Sum the colors (of the three edges) at every node of the halfgraph. This is 0 for each node so 0 for all nodes together. But that takes in every proper edge twice (which is 0 times) and halfedges once, so it is the sum of halfedge colors. ■

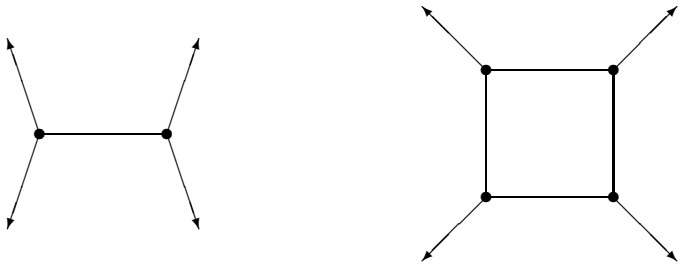
Lemma: if we get as far as stage $n-1$ we can color the whole graph. The three halfedges left have nonzero colors in a valid coloring of $H_\bullet^{n-1}$, and by the previous lemma they now sum to 0. That means we can tie them up into the final node. ■

Lemma: the dimension of $\boldsymbol{O}_\bullet^i$ is $f-1$.

We saw earlier that the dimension of $\boldsymbol{Z}_\bullet^i$ was $o+f-1$ where o is that of the subspace of proper cycles. The latter are the kernel of the projection from $\boldsymbol{Z}_\bullet^i$ to $\boldsymbol{O}_\bullet^i$. ■

This means $\boldsymbol{O}_\bullet^i$ consists of *all* the 4^{f-1} code words whose letters sum to zero. We won't be needing this fact; we will focus on a subset of $\boldsymbol{O}_\bullet^i$ next.

Let the **equatorial Tait code** $\mathcal{O}_\bullet^i$ be the image of $\mathcal{Z}_\bullet^i$ under the projection from $\boldsymbol{Z}_\bullet^i$ to $\boldsymbol{O}_\bullet^i$. Note that $\mathcal{O}_\bullet^i$ is *not* simply the collection of words in $\boldsymbol{O}_\bullet^i$ of full weight (like $\mathcal{Z}_\bullet^i$ was to $\boldsymbol{Z}_\bullet^i$), only a subset of them (there are many arrangements of nonzero colors on just the halfedges that would not extend back to a coloring of proper edges with nonzero colors). We cannot reconstruct $\mathcal{O}_\bullet^i$ from $\boldsymbol{O}_\bullet^i$, it depends on what is inside the graph already "done". An example.



The halfgraph on the left admits halfedge colorings $\begin{smallmatrix}*&*\\\circ&\circ\end{smallmatrix}$ and $\begin{smallmatrix}*&\circ\\\circ&*\end{smallmatrix}$ (where $*$ and $\circ$ are any two distinct colors) while the one on the right admits $\begin{smallmatrix}*&*\\\circ&\circ\end{smallmatrix}$ and $\begin{smallmatrix}*&\circ\\*&\circ\end{smallmatrix}$ and $\begin{smallmatrix}\circ&\circ\\\circ&\circ\end{smallmatrix}$. Both are viable codes but they are different, bearing the imprint of what happened before.

The reason we only need equatorial codes is that the question whether we can color $G^{i+1}_\bullet$, given a set of valid colorings of $G^i_\bullet$, only depends on the colors at the halfedges. Those are going to be the ones we split with V-steps, or tie together with Λ-steps. Using $\mathcal{O}^i_\bullet$ rather than $\mathcal{Z}^i_\bullet$ just means travelling lightly. We only carry what is still relevant. Only the patterns of color still viable on the halfedges will determine what we can do at the next step. Any (by now) proper edges have already asserted their influence by thinning out what colorings are left in $\mathcal{O}^i_\bullet$. Another advantage is that valid coloring of planar graphs becomes a (cyclic) 1-dimensional rather than 2-dimensional problem. Of course the second dimension is still there, in the form of the succession of $\mathcal{O}^i_\bullet$ for subsequent i. So what the bipolar growth really does in the planar case is impose a **polar coördinate frame**: the radial coördinate i refers to successive stages (linked by V- and Λ-steps) and the tangential coördinate labels the halfedges in cyclic order, the letters of the equatorial code at that stage.

Of course, each V-step causes there to be one more letter in the next equatorial code $\mathcal{O}^{i+1}_\bullet$ than in $\mathcal{O}^i_\bullet$, and each Λ-step causes there to be one fewer. The crucial feature of the cyclic order of halfedges for plane graphs is that a Λ-step only ever ties together two halfedges that were successive in the cyclic order i.e. two letters adjacent in the code word written cyclically, the combined edge taking their place, and likewise a V-step produces two successive halfedges (letters) that replace the one that they split.

4.6.3 Additive description

The following are immediate consequences of the colors at a node summing to 0 (and of subtraction being addition in a field of characteristic 2):

- Tying two halfedges on a Λ-step produces a halfedge with the sum of the colors of the old halfedges. This is only a valid coloring if the old halfedges had different colors (else it gives 0).
- Likewise, on a V-step we replace a single color by two colors that sum to the original color. This is always possible, in two valid ways even (e.g. blue becomes red/green or green/red).

So a V-step doubles the number of valid colorings, but a Λ-step decreases the number in a more unpredictable way: it all depends whether we had code words where the relevant two halfedges (letter positions) had different colors (letters). Of course it doesn't matter if individual code words at the $\mathcal{O}^i_\bullet$ stage go extinct (leave no descendants at the $\mathcal{O}^{i+1}_\bullet$ stage) by giving the same color to halfedges about to be joined on a Λ-step, what matters is whether the code as a whole survives (by always having some codewords that do have descendants). We'll work with codes, not individual code words.

The strategy, for a putative alternative proof of the four-color theorem along these lines, is gradually becoming clear. We must show the code is versatile enough that it always has descendants. If it is to hold for any trivalent graph then we must be able to cope with pretty much any two successive halfedges being tied by a Λ-step, with any halfedge being split with a V-step, and the two flavors of step occurring in any order. When building our graph we won't restrict the order in which nodes are picked any further than already demanded by the bipolar growth theorem. No further peeks over the fence. Tait coloring a planar trivalent graph becomes a kind of game for two players. One player is the graph; it decides what structure to throw at us, which halfedges to tie together next or which to fork next. We're the other player and we don't get to decide anything; we just keep the score and see if there are going to be any colorings left. There are only a few things we are entitled to expect:

- As $\mathcal{O}^1_\bullet$ has three halfedges and $\mathcal{O}^{n-1}_\bullet$ has too, the numbers of V-steps (which increase the number of halfedges by 1) and Λ-steps (which decrease it by 1) must be equal in the end.
- The number of halfedges never dips below two (one would mean a bridge, zero a disconnected graph). In fact, if it is convenient to demand the number never dips below *three* halfedges we may do so (we saw the existence of Tait colorings only needs to be proven for 3-edge-connected graphs); that would mean that at any stage we can have had at most as many Λ-steps as V-steps.

We could even demand it never dips below *four* (at stages 2 through $n-2$) if that helps, because it is sufficient to prove it just for solid graphs (by the decomposition argument on page 45); that would mean we've always had more V-steps than Λ-steps (at those stages).

Let a **string** be a sequence of halfedges (letter positions) that is contiguous in the cyclic order, and the **color of the string** in any one coloring (code word) the sum of the colors of the halfedges in that string. A long as there is no Λ-step tying the first or last halfedge of the string to halfedges outside it, the string can be thought of as persisting from one stage to the next, even if there is plenty Λ- and V-activity within.

The observations at the beginning of this section now amount to *the color of a string remaining constant* (in code words and their descendants) as long as the string persists. The rules for colors after Λ- and V-steps simply mean that any knitting done doesn't change overall color in any string big enough to contain the knitwork.

We can see there are certain things a code should never do. Let a **yoke** be any string whose color is zero in *all* the code words. An equatorial code must not have any yoke (other than a null string or the entire equator, whose color is always 0). The reason is that the graph could always decide to tie all the halfedges of the string together (with enough Λ-steps) to a single halfedge, and this would then get color 0. In this way strings are much like single halfedges.

Incidentally, this is why it is essential to construct the successive cuts via the BGT which keeps North in one piece. If we allow North to get disconnected we *would* have two separate pieces of equator that each by necessity sum to zero. Of course, the same graph would be Tait-colorable whichever way we built it, but things would just be so much harder to prove.

While absence of yokes is necessary for a code, it is not sufficient. Let a **polyyoke** be a set of one or more non-overlapping strings such that in every word of the code there is at least one of those strings that gets color 0. A code must also not have a polyyoke (this is a stronger condition than not having yokes). The reason this time is that the graph could decide to tie each of those strings up to a single halfedge. Then each codeword would fail on at least one of them.

The phrase **non-overlapping** here doesn't just mean disjoint. We must allow strings $A \subset B$ to count as non-overlapping (a graph could first tie up A to a single halfedge and later do the same with all of B). Only if both $A \setminus B$ and $B \setminus A$ are

nonempty should A and B count as overlapping. We don't need to demand the absence of overlapping strings at least one of which always has color 0, because the graph cannot tie all of them to single halfedges *simultaneously*. If strings A and B overlap, a code could give A (but not B) color 0 in some codewords and give B (but not A) color 0 in some other codewords (and even color 0 to both in yet other codewords). Then if the graph decides to do one thing the code could survive via some of its codewords' descendants, and if the graph does something else it survives via other descendants.

To find an alternative proof of the existence of Tait colorings for all bridge-free planar trivalent graphs, we need to find a set of criteria for "**liveness**" of an equatorial code. They should be (a) such that a live code has valid colorings, and (b) such that a code at stage i being live implies that the code at stage $i+1$ is again live whatever V-step or Λ-step the graph throws at us. Finding criteria with these properties (and showing the simple $\mathcal{O}^1_\bullet$ code satisfies them) amounts to a proof.

The biggest hurdle is the way Λ-steps can deplete a code. Suppose our code has some nice properties that allow it to weather certain kinds of storms. If we tie edges α and β together we have to throw away all code words where they had the same color. How do we know how much bio-diversity is going to be left in the code's gene pool after that? The concept of a polyyoke is tailor-made to cope with this problem.

Lemma A: the code $\mathcal{O}^1_\bullet$ has no polyyoke. By inspection. ■

Lemma Λ: absence of polyyokes is preserved under a Λ-step.

Let $\mathcal{O}^{i+1}_\bullet$ be obtained from $\mathcal{O}^i_\bullet$ by tying two successive halfedges, say π and σ, together in a Λ-step to the new halfedge ψ. We must prove that if $\mathcal{O}^i_\bullet$ has no polyyokes then $\mathcal{O}^{i+1}_\bullet$ has none either. By contradiction: suppose there is now a polyyoke $\wp_{i+1} = \{A, B, C \ldots\}$. The old code $\mathcal{O}^i_\bullet$ had two kinds of code words: those that gave $\pi\sigma$ color 0 (they didn't survive to $\mathcal{O}^{i+1}_\bullet$) and the rest (which are subject to $\wp_{i+1}$). Now construct $\wp_i = \{\pi\sigma, A, B, C \ldots\}$ where each of A, B etc. are interpreted as containing both π and σ if they contain ψ in $\mathcal{O}^{i+1}_\bullet$. Every codeword that doesn't make $\pi\sigma$ zero-colored does so with one of the other strings so $\mathcal{O}^i_\bullet$ already had $\wp_i$ as a polyyoke (it is non-overlapping because any of A, B etc. contain $\pi\sigma$ whole or not at all). ■

If it there was a corresponding lemma preserving the absence of polyyokes under a V-step this could be the whole liveness criterion. Unfortunately, while V-steps appear to go easy on colorings, they don't co-operate in this particular way.

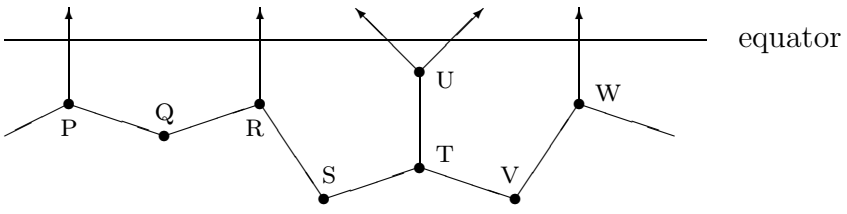
Suppose we have three halfedges red-green-red in succession. If we do a V-step on the green halfedge, it must either split red/blue or blue/red. We can keep both possibilities in the code: red-red-blue-red and red-blue-red-red. But what if the graph is really devious and ties the first two together *and* the last two? Then we are sunk whichever way round we split it. So red-green-red can't be live (and by the same reason red-blue-red can't). Again, we must treat this as a condition on whole strings (they can always become single halfedges).

What about red-red-red? Clearly we can't have a code forcing the colors of three successive strings to be red-red-red (let alone red-0-red) for different reasons. Now if we're going to have to outlaw these, and we saw red-green-red and red-blue-red aren't alright either, couldn't we simply say red-any-red kills liveness? Well no, because it doesn't! A code that allows some red-green-red code words *and* some red-red-red codewords, on the same three strings, could be perfectly alright. The graph cannot do everything. If the second string (as single halfedge) gets split by a V-step and then its halves tied up with its neighbours, those strings are not available anymore to be tied up a different way. If the graph does one thing, some code words survive and if it does the other thing, others survive.

4.6.4 Multiplicative description

Because none of the halfedges have color 0, successive halfedge colors have well-defined quotients. We could just as well write down our codewords by means of these quotients. It would cut down on the number of code words by a factor three, just like using Heawood node colors rather than edge colors — because that's exactly what it is! Physically, two successive halfedges form parts of a **halfface**, an unfinished face of the graph embedding, and the quotient is just the product of the Heawood colors (ω or $\overline{\omega}$) of the nodes along that unfinished cycle.

Example: in the picture below the first halfface quotient is the product of Heawood colors of nodes P, Q, R; the next that of R, S, T, U; the next that of U only (such a quotient cannot be 1); the next that of U, T, V, W; and so on.



Just as we had strings of successive halfedges, summing their colors, we can have

stretches of successive halffaces, multiplying the quotients. The meaning of the product of quotients in such a stretch is simply the quotient of the colors of the bounding halfedges. As is evident from the picture, individual nodes' Heawood colors may feature two or three times in such a product.

Now let the quotients of individual halffaces occurring along the equator be a, b, c, ... If the edge just before quotient a has color λ then the next ones have colors λa, λab, λabc and so on. When we come across the same edge again, one time round the equator, it has color $\lambda abc\cdots xyz$ but also the same color λ as before, so

$$abc\cdots xyz \;=\; 1$$

This condition is also handy to express the fact that none of the quotients can be zero. Talking about zero, we still have that the sum of all edge colors $\lambda + \lambda a + \lambda ab + \cdots + \lambda abc\cdots xy \;=\; 0$ (there's no term with z at the end, unless we remove the first term). But this is a product with a factor λ which, being an edge color, can't be 0 itself. So for the other factor we must have

$$1 + a + ab + \cdots + abc\cdots xy \;=\; 0$$

Consider two successive halfedges. The condition that a code should not force them to have the same color in all its code words was expressed additively by the sum of the colors not being 0 in all the words. Multiplicatively, the condition is that the quotient shouldn't be 1 in all the words. Let the colors be λ and λa, then $\lambda + \lambda a = \lambda(1 + a) \neq 0$. As we already know $\lambda \neq 0$ this means $1 + a \neq 0$, that is $a \neq 1$. Surprisingly though, multiplicative considerations do not always (at first sight at least) give the same result as additive ones, as we will see.

First let us investigate what V- and Λ-steps do to quotients. Splitting a halfedge with color μ into two halfedges gives the other two nonzero colors, that is $\mu\omega$ and $\mu\overline{\omega}$ in either order. Say μy and $\mu\overline{y}$ where y is ω or $\overline{\omega}$. Now let the quotients μ/λ and ν/μ with previous and next edges be a and b respectively. That means

$$\begin{array}{ccccccc} \lambda & & \lambda a & & \lambda ab \\ | & a & | & b & | \end{array} \quad \text{turns under a V-step into} \quad \begin{array}{ccccccc} \lambda & & \lambda ay & & \lambda a\overline{y} & & \lambda ab \\ | & ay & | & y & | & yb & | \end{array}$$

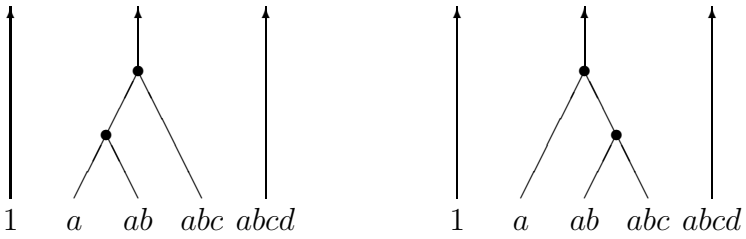
(recall $y^{-2} = y$). In a Λ-step the reverse happens. Now successive μ and μy add to $\mu(1 + y)$ which is $\mu\overline{y} = \mu y^{-1} = \mu y^2$ when y, as required here, is either ω or $\overline{\omega}$. So

$$\begin{array}{ccccccccc} \lambda & & \lambda x & & \lambda xy & & \lambda xyz \\ | & x & | & y & | & z & | \end{array} \quad \text{turns under a Λ-step into} \quad \begin{array}{ccccccc} \lambda & & \lambda x\overline{y} & & \lambda xyz \\ | & x\overline{y} & | & \overline{y}z & | \end{array}$$

Clearly when things happens only between some two bounding halfedges α and ζ the overall quotient ζ/α afterwards will be the same it was before. The product of a stretch of quotients stays the same under any knitting that stays inside the stretch because it is just the quotient of edge colors of ζ and α.

Note how with a V-step the new factor y appeared in three places (the new halfface and its two neighbours), leaving the stretch quotient unchanged as $1 = y^3$, and how with a Λ-step the factor y of the disappearing face that gets closed appears as $\overline{y}$ in two places (its erstwhile neighbours), here $y = y^4 = (y^2)^2$.

Now compare the following two situations.



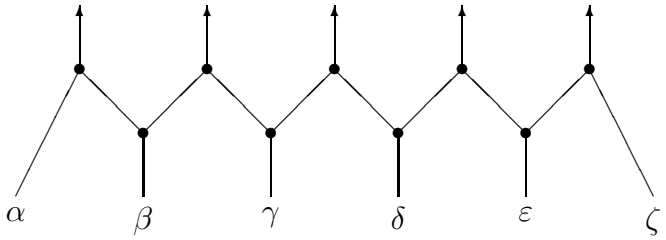
In the left picture we start with quotients $\mid a \mid b \mid c \mid d \mid$. By the rules for Λ-steps just discussed, this changes into $\mid a\overline{b} \mid \overline{b}c \mid d \mid$ at the first step and then into $\mid a\overline{b}\overline{(\overline{b}c)} \mid \overline{(\overline{b}c)}d \mid$ which is $\mid a\overline{c} \mid b\overline{c}d \mid$ at the second step. In the right picture on the other hand the first step gives us $\mid a \mid b\overline{c} \mid \overline{c}d \mid$ and the second $\mid a\overline{(b\overline{c})} \mid \overline{(b\overline{c})}\overline{c}d \mid$ which is $\mid a\overline{b}c \mid \overline{b}d \mid$. But how can this be? In both cases the middle edge (of three) at the top is the *sum* of the middle three edges (out of five) at the bottom, so they should be the same. But they are here calculated as $a\overline{c}$ and $a\overline{b}c$ respectively. The solution is that (in both pictures) $\overline{b}c$ cannot be 1 (or equivalently $b\overline{c}$ cannot). If the pictures are comparable at all then moreover neither b nor c can be 1. That means one of them must now be ω and the other $\overline{\omega}$ and under those restrictions indeed $a\overline{c} = a\overline{b}c$.

So while addition and multiplication strictly speaking give results that imply each other, the number of choices how to represent the same thing starts to proliferate soon. In this way each of the additive and multiplicatibe approaches has a unique slant on things; looking at either complements the other.

The restriction we saw in the previous section on codes forcing e.g. red-green-red or red-blue-red (but not red-red-red) at a certain place in all their codewords can be expressed more easily multiplicatively (it also frees us from using WLOG "red" for any edge color). It now takes the form that two successive halfface quotients should not be ω and $\overline{\omega}$ in either order.

Let a **neutral stretch** be a stretch whose quotient is 1 (in certain codewords) and a **minimal neutral stretch** (**m.n.s.** for short) be one such that none of its proper substretches is already neutral (in those codewords). A two-halfface stretch whose quotients are always $\omega\overline{\omega}$ or $\overline{\omega}\omega$ is a m.n.s. but allowing 1 1 as well (i.e. "red-red-red") would destroy minimality. So the restriction on "red-green-red" can be put succinctly in the form that there should not be a two-halfface m.n.s. that holds for all words in the code.

In fact we can dispense with the *two*-halfface qualifier: there is a similar restriction on m.n.s. of any size, and for exactly the same reason. Suppose a m.n.s. applies (to all codewords) on the stretch between edges α and ζ say. We can say WLOG that halfedge α is red; the fact that we have a neutral sequence means ζ is red too and the fact that it is minimal means there's no red halfedge before we get to ζ. Now suppose a graph splits all the intervening halfedges β, γ, δ and ε with V-steps, and then uses Λ-steps to tie up neighbours as follows:



Here β is some non-red color, so we split it red-something or something-red. The first is out of the question because the left branch will get joined with the red α. So β's right branch must be the red one. But then γ's left branch can't be red so its right branch is, and so on, until we clash with the red ζ at the far end. So indeed a code should not force such an m.n.s., of any number of halffaces, on all its codewords. It is clear why lack of minimality would destroy the argument: any red edge in between would split in two branches that are both non-red.

While a one-halfface m.n.s. is the same thing as a two-halfedge yoke (two successive halfedges forced in the same color by the code) larger m.n.s. are not related to larger yokes in any direct way. Demanding absence of yokes and of m.n.s. are two quite different constraints; a code must satisfy both.

Just as we can replace the restriction on yokes by the stronger demand that polyyokes are absent we can do the same for m.n.s. However, any such "poly-m.n.s." (a set of stretches such that each codeword makes at least one of them an m.n.s.) would have to consist of stretches that are disjoint and not even contiguous to each other (because if they share bounding halfedges the graph could not carry

out its threat to tie *all* of them up as in the diagram above).

A code that stays alive whatever the graph throws at it must avoid these poly-m.n.s. as well as polyyokes. Unfortunately the absence of the former isn't inherited under Λ-steps and absence of the latter not under V-steps. We must find a more comprehensive liveness criterion that includes these as special cases and *does* get inherited. It seems likely it will be some interplay of additive and multiplicative properties of $\mathbb{F}_4$.

4.6.5 log Heawood color space

As long as all the edge colors and their quotients are in the set $\{\overline{\omega}, 1, \omega\}$ we might as well treat their multiplication and division as the addition and subtraction of their "discrete logarithms" $\{-1, 0, +1\}$ which are residues (mod 3). The fact that we now compose additively, and that $\mathbb{F}_3$ is a field, means we can use some vector space formalism. Let the **equatorial Heawood code**, at stage i, consist of the discrete logs of the halfface quotients (quotients between colors of successive halfedges). Again, we have a succession of codes whose codewords mimic a (cyclic) one-dimensional arrangement, not of halfedges but (just as in the preceding section) of the gaps between them.

Until now we've just used single letters rather arbitrarily for those single halfface quotients occurring at some stage i. Quotients in later stages then got expressed as products (now sums) of these single letters. We could take more care exactly which (compound) quotients merit a single letter. This section will make one choice; the next section another.

It is easy to see that in the code $\mathcal{O}^1_\bullet$ the three halfface "quotients" (now differences) are all the same, the Heawood color of the first node (ω or $\overline{\omega}$). Call this a say. At each of the $h+1$ V-steps a new halfface is born, label these with single letters too. This means single letters will always have the value ± 1 (ω or $\overline{\omega}$ before taking the log), never 0 (formerly 1). Each time such a letter, say b, is born it gets added to (formerly multiplied into) the two neighbouring halffaces.

When it comes to closing a halfface (to a complete face), on a Λ-step, the accumulated "quotient" will in general consist of several letters. These sums (formerly products) must be non-zero (formerly non-1) as well, just like single letters. We saw the inverse of the disappearing "quotient" got added (formerly multiplied) into its neighbours; this may or may not make any particular letter disappear from the expressions that are left.

All this mimics exactly the way earlier equatorial codes doubled their number of code words on a V-step and got randomly depleted on a Λ-step (of course, as quotients and now their logs are straightforward transformed versions of those codes). The usefulness, if any, of the present incarnation lies in the new way it represents the constraints on the code vectorially.

Let the individual letters (however many we have had already at the i-th stage) form the basis vectors of a space H_i over $\mathbb{F}_3$. Its vectors are the formal sums of these letters, with coefficients from $\mathbb{F}_3$. We can identify (single or multiletter) "quotients" with vectors of H_i. It is the *graph structure* that tells us which formal sums of letters are going to appear in the halffaces. Any particular code word (coloring) takes the form of assigning actual values from $\mathbb{F}_3$ to each of the letters (and hence to all their linear combinations). That is, a linear function $H_i \to \mathbb{F}_3$.

The space of all codewords, even invalid ones (an **extended equatorial Heawood code** if you will) arises thus as the dual space H_i^* (being a vector space it is a linear code). The codewords we really want (forming a subset of H_i^*, a non-linear code) are those that assign nonzero values to all the single letters born at the very first step or at V-steps, and also nonzero values to the linear combinations of letters killed off at Λ-steps, or at the very last step.

4.6.6 Quark confinement

With the formalism as introduced in the previous section it is relatively easy to explain what we're going to do here. In principle, all that happens is a different choice of basis for H_i. Again, we will introduce a new letter (basis vector) on each V-step. But the new letter need not be simply the new halfface "quotient". Suppose the new basis vector is the k-th one. Then we *will* choose it such that the new halfface "quotient", and its two neighbours, each receive 1 as their k-th component (and everything else along the equator gets 0 in that position). But we reserve the right to make the other components of the new halfface "quotient" something other than all-0. The neighbours of course still get the new halfface "quotient", whatever its vectorial expression, added bodily.

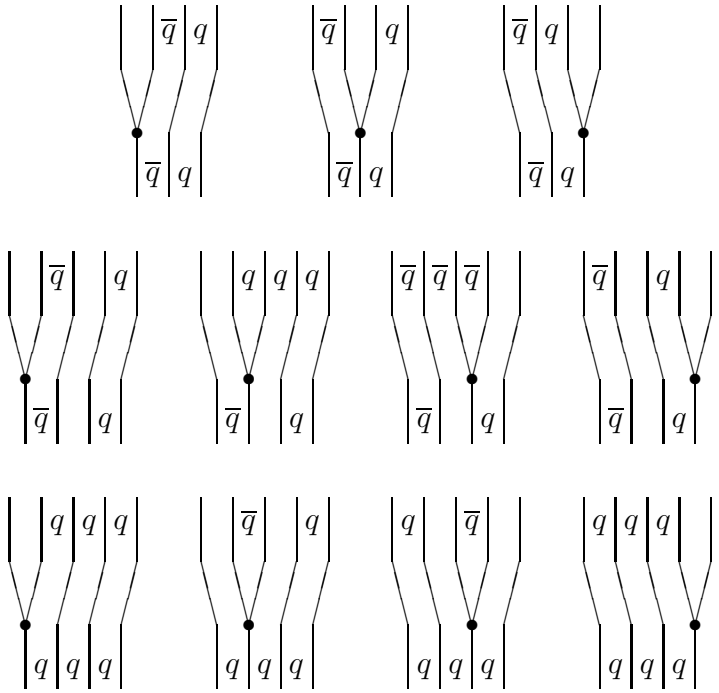
If the new letter to allocate is x then the new halfface may get plain x or it may get $x+$some combination of old letters; the basis vector x would be defined accordingly. An example. If the neighbours of the new halfface are $a+$ this and $a+$ that, the new halfface gets $x-a$ which makes the neighbours $x+$ this and $x+$ that, relative to that particular definition of what the new x is. In summary, the new letter is

always given to the new halfface and its two neighbours, but the new distribution of the existing letters will depend on their current distribution. Single letters under this allocation scheme (spelled out in full below) will be termed **quarks**.

The reason for this quirky scheme is that we can rig it so that the following holds:

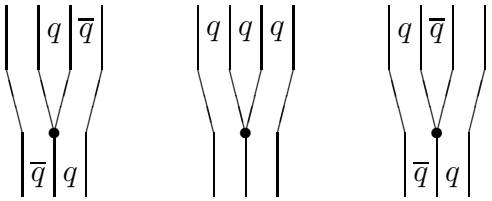
Quark confinement theorem: Any quark (q say) will, at any stage i, only occur distributed in one of the following ways:
– in three successive halffaces: $|\ldots+q\,|\ldots+q\,|\ldots+q\,|$
– same, but using its inverse: $|\ldots-q\,|\ldots-q\,|\ldots-q\,|$
– once, adjacent to its inverse once: $|\ldots+q\,|\ldots-q\,|$ or $|\ldots-q\,|\ldots+q\,|$
– id. with one halfface in between: $|\ldots+q\,||\ldots-q\,|$ or $|\ldots-q\,||\ldots+q\,|$
– or absent altogether at this stage.

To show this is true we must spell out what to do in each possible situation. With a V-step on any edge of $|\,\overline{q}\,|\,q\,|$ or $|\,\overline{q}\,|\quad|\,q\,|$ or $|\,q\,|\,q\,|\,q\,|$ respectively:



For a V-step on any edge of $|\,q\,|\,\overline{q}\,|$ or $|\,q\,|\quad|\,\overline{q}\,|$ or $|\,\overline{q}\,|\,\overline{q}\,|\,\overline{q}\,|$ use the same, with q and $\overline{q}$ interchanged.

If a Λ-step occurs we have no choice, but it works out alright: we get all the same ones read upside down. That covers all cases, except the ones obtained by reading the following upside down:



When read the right way up, the outer two here are alternative ways to deal with a V-step and the inner one represents the V-step that first created the letter. ■

Quark confinement[13] is of course a direct consequence of planarity; otherwise we would have had none of these *local* goings-on along the equator.

As before, the vectors for the "quotients" of halffaces that disappear at Λ-steps, or appear at V-steps, are the ones a coloring must give a nonzero value (single letter quarks are not necessarily among them). Showing a coloring exists amounts, as it would in the previous section, to solving a system of simultaneous inequalities (this $\neq 0$, that $\neq 0$, ...) that are linear in all the single-letter unknowns.

The way quark confinement helps (if indeed it does) is by keeping letters local, not just with respect to each other, but also with respect to the grid of edges as boundaries of the halffaces. Let the **central axis** of a quark be either a halfedge or the halfface between two successive ones, and defined as follows: it starts off as the halfface newly created with the letter, and while still a halfface keeps its identity on any V-step. When the halfface disappears on a Λ-step it becomes the halfedge between its old neighbours. While a halfedge, it keeps its identity even when merging due to a Λ-step, and becomes a halfface again when the halfedge undergoes a V-step. Inspection of the proof above now also shows ■ the following:

Lemma: the 2 or 3 occurrences of a quark remain centered on its central axis.

This puts limits on which (and how many?) of the unknowns can occur in which of the simultaneous inequalities (and how far apart?). I have not yet tabulated all the possibilities; it is possible that this could be turned into a proof that a solution (coloring) always exists.

---

[13] I've borrowed these terms from physics because they seemed so apt. There, **quarks** are the building blocks of all *hadrons* (particles subject to the "strong force") where they occur either paired with an anti-quark ($q\bar{q}$ is a *meson*) or as triple (qqq is a *baryon* such as a proton or neutron, $\bar{q}\bar{q}\bar{q}$ an *anti-baryon*). The word was taken by Murray Gell-Mann from a line in James Joyce's *Finnegans Wake*: "Three quarks for muster mark."

Quark **confinement** refers to the impossibility of creating other combinations: when you try to split qqq into $q \cdots qq$ so much energy is needed that a quark/antiquark pair is formed at the place of the physical separation so you get $q\bar{q} \cdots qqq$ instead, separating into two new hadrons.

4.6.7 Weights and the dual code

Recall the weight, or Hamming weight, of a codeword is the number of nonzero letters it contains. If a code has N_i codewords of weight i, for each i, this list of values is usually presented as a polynomial $\sum N_i x^i$ in an arbitrary formal variable x, or equivalently as a homogenous polynomial $\sum N_i w^{m-i} x^i$ (where m is the number of letters in a codeword). Either polynomial is known as the (**Hamming**) **weight enumerator**. One reason for giving the list in the form of a polynomial is that it combines like one. For any flavor of weight enumerator we'll encounter

- If a code $\mathcal{C}$ is composed as $\mathcal{A} \mid \mathcal{B}$ (that is, $\mathcal{A}$ forms the first so-and-somany digits positions of $\mathcal{C}$, and $\mathcal{B}$ the rest, and all combinations occur) the weight enumerators of $\mathcal{A}$ and $\mathcal{B}$ only need to be multiplied (as polynomials) to give that of $\mathcal{C}$.
- If two codes $\mathcal{A} \subset \mathcal{V}$ and $\mathcal{B} \subset \mathcal{V}$ (i.e. they both live on the same digit positions) with $\mathcal{A} \cap \mathcal{B} = \{\}$ (so at most one of them can be linear) are merged to a code $\mathcal{A} \cup \mathcal{B}$ their weight enumerators get summed.

As we have access to the dual code (orthogonal complement), the MacWilliams theorems [MS77] are of interest. They relate the weight enumerator of a linear code to that of the dual code. For $\mathbf{C}$ and $\mathbf{C}^\perp$ over any $\mathbb{F}_q$, if $\mathbf{C}$ has Hamming weight enumerator $W(w,\, x)$, that of $\mathbf{C}^\perp$ is obtained by substituting $W(w+(q-1)x,\, w-x)$ and dividing by the number of codewords in $\mathbf{C}$ (the next section has more on weight enumerators and MacWilliams theorems).

For our purposes, if $\boldsymbol{Z}^\perp$ has Hamming weight enumerator $N(w,x) = \sum_i N_i w^{m-i} x^i$ and $\boldsymbol{Z}$ has $Z(w,x) = \sum_i Z_i w^{m-i} x^i$ then $Z(w,\, x) = N(w+3x,\, w-x) \,/\, |\boldsymbol{Z}^\perp|$ so the coefficient Z_m of the latter, the number of words of full weight in $\boldsymbol{Z}$, is found to be

$$Z_m \;=\; \frac{1}{4^{n-k}} \sum_{i=0}^{m} 3^{m-i} (-1)^i N_i \;=\; \frac{(-1)^m}{4^{n-k}} \sum_{i=0}^{m} (-3)^{m-i} N_i$$

by multiplying out and gathering terms, and this is the number of valid Tait colorings; all we need to prove is that it is nonzero! That is, we must show the weighted sum of terms where $m-i$ is even exceeds (the absolute value of) the odd terms.

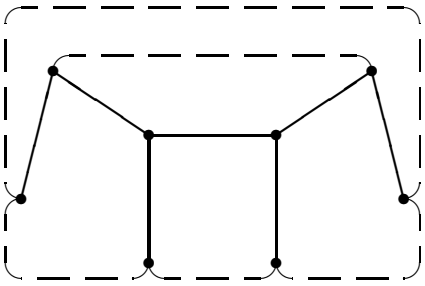
A basis for $\boldsymbol{Z}^\perp$ is given by any $n-1$ nodes $\boldsymbol{u}_i$ leaving one out, so $\boldsymbol{Z}^\perp$ has all 4^{n-1} vectors $\sum c_i \boldsymbol{u}_i$; the simplest way to do the enumeration is via all 4^n vectors $\sum c_i \boldsymbol{u}_i$ summing over all nodes and dividing the numbers by 4 (the last node is the sum of all others so each vector occurs 4 times). The letters in the code words correspond

to edges, that for the edge between i-th and j-th node is now $c_i + c_j$, so for each of these 4^n what effectively are 4-colorings of the nodes count the number of edges between nodes of different color, that's the weight. In K_4 for example

partition of the c_i	weight i	N_i	$3^{m-i}(-1)^i N_i$
4	0	4/4	+729
1,3	3	48/4	−324
2,2	4	36/4	+81
1,1,2	5	144/4	−108
1,1,1,1	6	24/4	+6
		256/4	+384

and $384/4^3 = 6$, which is indeed the number of Tait colorings of K_4.

The space $\boldsymbol{Z}^\perp$ is rather more manageable than $\boldsymbol{Z}$. For instance (taking $k = 1$ for simplicity) over the $n-1$ edges of a **spanning tree** we get all possible 4^{n-1} combinations exactly once, so N_i-thus-far is simply $\binom{n-1}{i}3^i$ giving each term a factor 3^m in all; now the sum of binomial coefficients of alternating sign yields 0.



Next we must put in the **lianes** (non-tree-edges, dashed in the example above for the cube) one by one. Any one is easy; $\boldsymbol{Z}^\perp$ structure mandates the edges of the cycle we close must have colors summing to 0. This skews the distribution in a predictable way that indeed makes Z_m-thus-far positive. There are nice relations that allow us to take all even terms in Z_m together (with the right number of factors 3), and the odd ones together, without worrying about individual N_i.

The hard part is combining the contributions from all lianes. If we build by BGT rather than start with a spanning tree the problem is counting separately the portions of the N_i that refer to edges about to be combined on a Λ-steps having the same or different $\boldsymbol{Z}^\perp$-color.

There cannot be a simple arithmetic inequality that keeps Z_m positive because we already know that it is zero for some (non-planar) graphs. So a proof must make

use of the particular non-crossing constraint on lianes that planarity imposes, or equivalently the cyclic order of halfedges in BGT.

If we use an induction argument and so rely on Z_m of smaller graphs it might help to be really devious and use a fact that comes for free: Z_m can't possibly be *negative* for any graph.

4.6.8 Complete weights

The Hamming weight enumerator did not distinguish between different nonzero letters. The **complete weight enumerator** is a polynomial with (in the case of $\mathbb{F}_4$) four formal variables such that the coefficient N_{ijk} in the homogenous polynomial $\sum N_{ijk} w^{m-i-j-k} x^i y^j z^k$ counts the number of codewords with i 1's, j ω's, k $\overline{\omega}$'s and the rest 0's, distinguishing between all flavors of letter[14].

The MacWilliams theorem for complete weights is a bit involved, not least because we must introduce a deliberate asymmetry that will disappear again in use.

Any field $\mathbb{F}_q$ ($q = p^d$, p prime) is, where its addition is concerned, an abelian group F of form $C_p \times C_p \times \cdots C_p$; in other words, a vector space $\mathbb{F}_p^d$ under addition. Let the vector $(a_0, a_1, \ldots a_{d-1})$ denote the element $a \in \mathbb{F}_q$ w.r.t. any basis[15] where the a_i are in $\mathbb{F}_p$ i.e. residues (mod p). For any d-tuple c_0, $c_1 \ldots$ c_{d-1} of such residues, the mapping $\chi : (a_0, a_1, \ldots a_{d-1}) \mapsto e^{\frac{2\pi i}{r} \sum c_i a_i}$ is one of the characters of F. For any χ with not all the c_i zero (MacWilliams and Sloane use $\chi_1 : (a_0, a_1, \ldots a_{d-1}) \mapsto e^{\frac{2\pi i}{r} a_0}$) MacWilliams' Theorem takes the following form:

Let the complete weight enumerator of m-digit linear code $\mathcal{C}$ over $\mathbb{F}_q$ be $W(\ldots x_a, \ldots)$ where the variables x_a are indexed by field element a. Then the complete weight

[14]For codes over $\mathbb{F}_2$ the two kinds of enumerator coincide. Intermediate versions exist such as the **Lee weight enumerator** for codes over fields of odd prime order that only brackets residues $\pm r$ together; one could also e.g. just distinguish quadratic residues from the others, and from 0.

The most elaborate form is the **exact weight enumerator** that (for codes of m letters out of an alphabet of q) uses mq distinct formal variables referring to both flavor and position of letters occurring. It gives a complete specification of the code. Note each time a less explicit enumerator can be obtained from a more specific one by evaluating the latter polynomial with some of its variables substituted by the same formal quantity. This is why proofs about weight enumerators tend to apply to all versions simultaneously.

[15]One possible basis is 1, x, x^2 ... x^{d-1} where x is any primitive element. There are nicer *normal* bases y, y^p, y^{p^2}, ... $y^{p^{d-1}}$ where the Frobenius automorphism $a \mapsto a^p$ just cycles the coördinates, including ones where y is primitive too, by Davenport's theorem [MS77].

enumerator of $\mathcal{C}^\perp$ is obtained by substituting

$$\text{any } x_a \text{ by } \sum_b \chi(ab)x_b$$

and dividing the polynomial by $|\mathcal{C}|$. In binary codes for example, x_0 is replaced by $x_0 + x_1$ and x_1 by $x_0 - x_1$ as $\chi(0 \cdot 0) = \chi(0 \cdot 1) = \chi(1 \cdot 0) = +1$ but $\chi(1 \cdot 1) = -1$. More generally in $\mathbb{F}_p$, $\chi(ab) = e^{2\pi iab/p}$. In $\mathbb{F}_4$ (writing again w, x, y, z for x_0, x_1, x_ω, $x_{\bar{\omega}}$) if $\boldsymbol{Z}^\perp$ has weight enumerator $N(w,x,y,z) = \sum N_{ijk} w^{m-i-j-k} x^i y^j z^k$ then the weight enumerator $Z(w,x,y,z) = \sum Z_{ijk} w^{m-i-j-k} x^i y^j z^k$ of $\boldsymbol{Z}$ is given by the substitution

$$\begin{aligned} Z(w,x,y,z) \;=\; & N(w+x+y+z, \\ & \quad w-x+y-z, \\ & \quad w+x-y-z, \\ & \quad w-x-y+z) \end{aligned}$$

which looks like it couldn't possibly be right as it treats z differently from x and y. Linear codes only depend on the *additive* properties of elements so all nonzero elements should appear on the same footing. The solution is that the expressions above are only one possible choice to express the MacWilliams theorem for $\mathbb{F}_4$. A particular χ must be chosen to express the theorem but the effect of the choice goes away again in the result on the level where a Z_{ijk} is expressed as linear combination of various N_{ijk} because in a linear code $N_{ijk} = N_{jik} = N_{jki}$ etc.

For the particular $\boldsymbol{Z}$ that occurs as cycle space of trivalent graph, the only code words of Hamming weight $m = 3h$ that occur are actually ones with exactly h 1's, h ω's and h $\overline{\omega}$'s. So we have a choice how to express the number of Tait colorings using complete weight enumerators. We can say it is Z_{hhh}, or we can say it is the sum of all Z_{pqr} with $p + q + r = m$, or any possibility in between. They all give different expressions in terms of the N_{ijk}. The fact that they are all numerically equal is a property of the particular $\boldsymbol{Z}^\perp$ we're using, derived from trivalent graphs.

For any choice the expressions become quite unwieldy. And just as in the previous section, arithmetic alone that disregards position cannot prove Z_{hhh} or etc. is nonzero because the number of colorings *is* zero for some (nonplanar) graphs. We need to make use of planarity again in some form or another.

4.6.9 Tait coloring reduced vectors

For connected trivalent graphs ($n = 2h$ and $m = 3h$) the dimension of $\mathbf{Z}^\perp$ is $2h-1$; that of $\mathbf{Z}$ and $\mathbf{E}^\circ$ is $h+1$ (reduced space $\mathbf{E}^\circ$ manages to condense the graph

structure in about one third of the number of bits $\mathbf{E}$ does, losing some redundant information).

At every node, each of the three edges there is (as reduced vector) the sum of the other two. This can be applied to Tait colorings as follows: the color of each edge is the sum of colors of two other edges there. Adding nodes this extends to all cuts: the Tait colors of edges in a cut sum to zero (we used this already to show that bridges cannot get nonzero color, and that if $\mathbf{a} \equiv \mathbf{b}$ then $\mathbf{a}$ and $\mathbf{b}$ must have the same color). So whenever an edge itself is, as reduced vector, the sum of certain other edges, its color is the sum of their colors too. But that means color (in any Tait coloring) is a property of vectors of $\mathbf{E}$ *only through it being a property of their reduced vectors*.

Yet another way of saying the same thing: let $\mathbf{V}$ be the additive group of $\mathbb{F}_2^2$, the Klein 4-group (isomorphic to the additive group of $\mathbb{F}_4$). Any given Tait coloring is a mapping from $\mathcal{E}$ to $\mathbf{V}$ which can be extended to a homomorphism φ from $\mathbf{E}$ (as additive group) to $\mathbf{V}$. Now we see that $\mathbf{E}^\circ$ sits between them: φ is the product of $\rho : \mathbf{E} \to \mathbf{E}^\circ$ and another homomorphism $\psi : \mathbf{E}^\circ \to \mathbf{V}$ because $\ker \rho \subset \ker \varphi$ (all what sums to zero as reduced vectors has color summing to zero).

Note all this defines Tait colorings in terms of vectors over $\mathbb{F}_2$, not $\mathbb{F}_4$. As $\mathbf{V}$ has dimension 2, $\ker \varphi$ is a subspace of $\mathbf{E}$ of dimension $m - 2$ (the vectors that get color 0), and again $\ker \mathbf{E}^\circ$ is a subspace of $\mathbf{E}^\circ$ of dimension $o - 2$ (the reduced vectors that get color 0). So proving a particular ψ is a valid Tait coloring amounts to proving *all the vectors representing single edges are outside* $\ker \psi$. Showing a valid Tait coloring exists amounts to finding a large enough kernel (subspace of dimension 2 less than everything) that fits in the embedding space without hitting any of the single edges scattered across the space like raisins in a raisin loaf.

The problem can be visualised in terms of sets of possible kernels. A Λ-step doesn't increment the dimension of $\mathbf{E}^\circ$ (it adds an edge but that's just the sum of existing ones). So the kernel to fit doesn't get bigger either. It does get harder to find kernels that fit, because one more vector is now marked as "a single edge" that must be avoided. A V-step increments the dimension by 1 (not 2, the other new edge is the sum of this edge and an existing one). Kernels must also grow one dimension, and for each existing kernel this can happen in precisely two ways. Why is that? We must choose one of the new vectors to help span the new kernel, but as it remains of dimension 2 less than the embedding space it'll then colonise one quarter of the new vectors. Choosing any of those to help span it would give the same kernel so there are really only four choices. Of them two are not allowed: the new edges

must remain outside the new kernel (and they would not have amounted to the same choice because their difference, the edge being split by the V-step, isn't in the old kernel). Specifically if $\mathbf{c}^\circ$ is the edge that got split, into $\mathbf{d}^\circ$ and $\mathbf{e}^\circ$ (so $\mathbf{c}^\circ + \mathbf{d}^\circ + \mathbf{e}^\circ = \mathbf{0}$), and if $\mathbf{a}^\circ$ and $\mathbf{b}^\circ$ are the edges adjacent to $\mathbf{c}^\circ$ at the other end (so $\mathbf{a}^\circ + \mathbf{b}^\circ + \mathbf{c}^\circ = \mathbf{0}$) then we can choose $\mathbf{d}^\circ + \mathbf{a}^\circ = \mathbf{e}^\circ + \mathbf{b}^\circ$ or we can choose $\mathbf{d}^\circ + \mathbf{b}^\circ = \mathbf{e}^\circ + \mathbf{a}^\circ$.

Yet another tantalising approach that makes it look so easy. I better stop here. Hoping someone cleverer than me will run with the ball and manages to do what has thus far eluded me — craft a straightforward proof, perhaps using some of the ideas suggested here, of *why* four colors suffice.

5 Cycle double covers

5.0 Maps

Recall a **cycle** was defined in chapter **0** as the closed counterpart of a *path* (all nodes distinct except first and last) and that in a trivalent graph closed *trails* (all edges distinct) are necessarily cycles. There we also borrowed Cameron's term *trek* (where edges may be re-used, as long as you don't backtrack on the previous edge). Now let a **circuit** be a closed trek. For trivalent graphs this means that after each edge you have a choice of precisely two edges to follow next. While the term *circuit* is indeed used by some authors in this sense, there are half a dozen other usages.

Let a **circuit map** be a collection of circuits of G such that every edge is visited twice (formal definitions below). Because a circuit can visit an edge more than once (in the same or opposite direction) the two instances where a circuit map visits an edge can be part of the same circuit. Let a **cycle map** be a circuit map where all circuits are cycles. Now the two visits to an edge are by two different cycles of the map. Faces of a polyhedron are an example of a cycle map.

G. Szekeres, who first studied them, called circuit maps *polyhedral decompositions* and cycle maps *simple polyhedral decompositions*. In recent literature cycle maps are called *cycle double covers* instead but there does not seem to be a corresponding modern term for circuit maps, which aren't discussed anymore, except here. Mixing one older term and one unrelated new term for two so closely related concepts would be unfortunate. Any way of resolving the dilemma has disadvantages. My choice of using (cycle and circuit) **map** has at least the virtue of being brief.

Following Szekeres [Sze72, Sze73], a cycle map will be **proper** if no two cycles share more than a single edge, and **even** if all the cycles are of even length. The following tables may serve to facilitate comparisons with the older literature:

here (defined in chapter **0**)	Tutte, Szekeres
walk	*path*
trek	*semisimple path*
trail	(Sz.) *simple path*
path	(T.) *simple path*
closed	*re-entrant*
closed walk	*circuit*
closed trek = **circuit**	*semisimple circuit*
"closed path" = **cycle**	*simple circuit*

but note the authors are careful to define the various flavors of their *circuits* as equivalence classes of *re-entrant paths* starting at any point of the *circuit*. Also, their *paths* and *circuits* are all directed.

here (defined in this section)	Szekeres
circuit map	*polyhedral decomposition*
cycle map	*simple polyhedral decomposition*
proper cycle map	*proper polyhedral decomposition* where *proper* is defined to imply *simple* too
even cycle map	*even polyhedral decomposition* where *even* is defined to imply *simple* too
oriented circuit map (defined in next section)	*coherent polyhedral decomposition* note *coherent* doesn't imply *simple*
oriented cycle map	*coherent simple polyhedral decomposition*

Embedding planar graphs on the sphere (as polyhedra), and more generally embedding graphs in surfaces of higher genus, formed the inspiration of the concept. On the one hand, there are far more maps than the obvious ones: the dodecahedron for example admits 4 294 967 296 circuit maps, 30 843 of which are cycle maps! Appendix **B** gives a flavor of this diversity by enumerating cycle maps for a few dozen cubic graphs. On the other hand, it is still open whether every (bridgefree) graph actually has a cycle map.

One vector space formulation of $\mathcal{M}$ being a cycle map would be

$$\mathcal{M} \subset \mathcal{C} \quad \text{and} \quad \forall \mathbf{e} \in \mathcal{E} \;\; \#\{\mathbf{c} \in \mathcal{M} \mid \mathbf{e} \bullet \mathbf{c} = 1\} = 2$$

where $\mathcal{C}$ is the set of single cycles (not a *space*) and # denotes cardinality. Note that the similar looking condition

$$\mathcal{M} \subset \mathbf{Z} \quad \text{and} \quad \forall \mathbf{e} \in \mathcal{E} \;\; \#\{\mathbf{c} \in \mathcal{M} \mid \mathbf{e} \bullet \mathbf{c} = 1\} = 2$$

is slightly too wide: if $\mathcal{M} = \{\mathbf{c}_0, \mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \ldots\}$ is a cycle map, implying all the $\mathbf{c}_i$ are single cycles, and (say) $\mathbf{c}_0$ and $\mathbf{c}_1$ are disjoint, then replacing $\mathcal{M}$ by $\{\mathbf{c}_0 + \mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \ldots\}$ would still satisfy the wider condition. Of course, proving the existence of such a collection of elements of $\mathbf{Z}$ would still prove the existence of a cycle map, because any element of $\mathbf{Z}$ can be decomposed into disjoint cycles.

There is no vector space formulation of the concept of circuit maps. A circuit is more than its edge set; the order of edges matters (as does the direction an edge is visited, at least relative to that of the other edges). We will see a formal definition

in the next section when we turn to directed circuit maps, which are also needed to define the concept of oriented maps.

For all pseudographs (including multigraphs (including graphs)) we have

Lemma: a pseudograph with a bridge has no cycle map.

Lemma: a pseudograph with a cobridge has no proper cycle map.

Proofs: we saw earlier that if β is a bridge there is no cycle through β $\blacksquare$ and that if α and β are in a cobridge their reduced vectors are equal, so every cycle through α passes through β; the two cycles of the map that cover α intersect in α and β. $\blacksquare$

Conjecture: every bridgefree pseudograph has a cycle map.

Conjecture: "every" bridgefree cobridgefree pseudograph has a proper cycle map.

When trying to find cycle maps, it is evident that

- nodes of valency 0 can be disregarded as we are trying to cover edges.
- nodes of valency 1 give rise to a bridge.
- nodes of valency 2 give rise to cobridges; if we don't mind cycle maps being improper (on the original graph) such nodes can be disregarded and the two edges there amalgamated.
- nodes $\times$ of valency 4 can be teased apart into two trivalent ones $\rangle\!\!-\!\!\langle$ and more generally nodes of valency $\rho > 3$ into $\rho - 2$ trivalent ones, in a choice of ways. This only makes the problem "harder" in the sense that if a [proper] cycle map exists on the resulting trivalent graph it is certainly still a [proper] cycle map after contracting the nodes back to the original ρ-valent node.

The only nodes left are trivalent ones, so the conjectures will be true if they are true for trivalent graphs:

Conjecture: every bridgefree trivalent multigraph has a cycle map.

Conjecture: "every" bridgefree cobridgefree trivalent graph has a proper cycle map.

Note the term is only *graph* in the second conjecture because we saw trivalent multigraphs have cobridges (except for the one connected one with a triple edge which has an obvious proper cycle map). It is only *multigraph* in the first conjecture because a trivalent pseudograph has a bridge. There is some confusion in the literature about the earliest enunciation of these conjectures. Both are mentioned already by Szekeres [Sze73]. The first conjecture is the famous **cycle double**

cover conjecture. The second must be read as excluding at least $K_{3,3}$ which is clearly a counterexample. No counterexamples are known to the cycle double cover conjecture, not even to the stronger

Conjecture: every bridgefree trivalent multigraph has, for each of its cycles, a cycle map containing that cycle.

Conjecture: every bridgefree trivalent multigraph has an oriented cycle map.

5.1 Oriented maps

A **directed graph** is like a graph except that it has **directed edges**. I will adopt Szekeres' term **arc** for a directed edge as it is considerably shorter. In stead of there either being an edge *between* two nodes P and Q or not, there can now either be an arc *from* P *to* Q or not, and independently from that either an arc *from* Q *to* P or not. Having both is much like having an old-fashioned two-way edge. Conversely, every ordinary graph G can be considered as a directed graph $\vec{G}$ that happens to have, for each edge of G, arcs in both directions on that edge. Thus edges become pairs of arcs.

Let a **directed** circuit or cycle be a circuit or cycle with a choice which way to traverse it. To every circuit map (of c circuits) correspond 2^c **directed circuit maps** formed by choosing a direction of traversal for each of the c circuits. Every time one of the directed circuits passes through an edge of G it visits only one of the two arcs on that edge. An **oriented circuit map** is a directed circuit map where the direction of the circuits is chosen in such a way that, of each edge, both arcs are visited. In other words it is a single cover of the arcs of $\vec{G}$.

A (**non-**)**orientable** circuit map is a circuit map that can(not) be turned into an oriented map by a suitable choice of direction of its circuits.

We will see below when we discuss embeddings of graphs in topological surfaces that the existence of an orientable map implies the surface is orientable in the topological sense. The definition in terms of maps is the purely combinatorial analogue.

Theorem: Every trivalent (pseudo- or multi)graph has an oriented circuit map.

Szekeres notes this is a special case of a more general theorem by Petersen and König, but prefers to give an independent proof which centers on the existence of

a mapping L from arcs to arcs, with the two properties

$$\left.\begin{array}{l}\bullet\ \text{if arc } \mu \text{ runs to node P then } L(\mu) \text{ runs from node P} \\ \bullet\ \text{if } L(\mu)=\nu \text{ then } L(\overline{\nu})\neq\overline{\mu}\end{array}\right\} \qquad (***)$$

(where the overbar indicates reversal: $\overline{\mu}$ is the other arc on the same edge as μ). Clearly any such L is a recipe for walking: after passing through μ to reach P, choose $L(\mu)$ next. And given an oriented circuit map, the recipe for walking its circuits in their given direction satisfies $(***)$.

I will follow the tradition and give a new proof again (a bit shorter too). First we must show that, conversely, an L satisfying $(***)$ determines an oriented circuit map. Define the orbit $[\mu]$ of an arc μ as $\{\mu,\ L(\mu),\ L(L(\mu)),\ L(L(L(\mu))),\ \ldots\}$.

Let α, β, γ be the arcs from a node P. The second condition says, among other things, that $L(\overline{\alpha})$ is never α, $L(\overline{\beta})$ never β and $L(\overline{\gamma})$ never γ. But we have more: if $L(\overline{\alpha})=\beta$ then also $L(\overline{\beta})=\gamma$ and $L(\overline{\gamma})=\alpha$; or else if $L(\overline{\alpha})=\gamma$ then also $L(\overline{\gamma})=\beta$ and $L(\overline{\beta})=\alpha$. In pictorial terms: embed the node and its edges locally in a plane in a Y shape, now either L tells us always to turn left at this node, whatever direction we come from, or it always tells us to turn right.

So there's always a unique μ for which $L(\mu)=\nu$, call it $L^{-1}(\nu)$, which makes $\langle L\rangle$ a group acting on $\mathcal{A}$. In finite G it acts cyclically on each orbit making that a closed directed walk and, as $L(\overline{\mu})$ is never μ, a directed circuit. Finally, every μ (and hence also $\overline{\mu}$ for any μ) occurs in exactly one orbit $[\mu]$ so the orbits of such an L form an oriented circuit map.

Now we just need to show the existence of such an L. Label the $2m=6h$ arcs with distinct ordinals. At any node P, the three arcs α, β, γ from there can now uniquely be identified as $\min(\alpha,\ \beta,\ \gamma)$ and $\max(\alpha,\ \beta,\ \gamma)$, the ones with the smallest and largest label respectively, and $\mathrm{med}(\alpha,\ \beta,\ \gamma)$, the remaining one. Defining

$$\begin{aligned} L\Big(\overline{\min(\alpha,\ \beta,\ \gamma)}\Big) &= \mathrm{med}(\alpha,\ \beta,\ \gamma) \\ L\Big(\overline{\mathrm{med}(\alpha,\ \beta,\ \gamma)}\Big) &= \max(\alpha,\ \beta,\ \gamma) \\ L\Big(\overline{\max(\alpha,\ \beta,\ \gamma)}\Big) &= \min(\alpha,\ \beta,\ \gamma) \end{aligned}$$

gives $L(\overline{\mu})$ for arcs $\overline{\mu}$ that go *to* P (to find the value of $L(\mu)$ for arcs μ going *from* P apply the definition at the node where the arc goes *to*) and this L suffices. ■

If we have a (multi)graph a slight simplification is possible in that we only need $3h$ distinct ordinals for the edges (the directionality of walking still matters for

which edge is visited before which). For pseudographs we do need arc numbers, to distinguish the two ways into a loop.

We now have an L with the desired properties that generates a circuit map. It is by no means the only one. There's no reason to derive the epithets min, med and max from some global labeling, all we need them to do is uniquely identify arcs locally. This suggests a **local labeling** of arcs: at each node, arbitrarily assign to the arcs sprouting from that node three local labels, for instance the nonzero elements $\overline{\omega}$, 1, ω of the field $\mathbb{F}_4$. Let $\lambda(\mu)$ be the local label of arc μ and let $y(\mathrm{P}, f)$ denote the arc from P that has local label f there. Just as before we can set

$$\begin{aligned} L\Big(\overline{y(\mathrm{P},\overline{\omega})}\Big) &= y(\mathrm{P},1) \\ L\Big(\overline{y(\mathrm{P},1)}\Big) &= y(\mathrm{P},\omega) \\ L\Big(\overline{y(\mathrm{P},\omega)}\Big) &= y(\mathrm{P},\overline{\omega}) \end{aligned}$$

Let us apply the term **turning left** to the direction at each node where such an L sends us, relative to the arc we came from. If the reverse of the arc we came in from had local number $\overline{\omega}$, 1 or ω we now take 1, ω or $\overline{\omega} = \omega^2$ respectively, that is, we multiply the local label of the reverse of the arc just traversed by ω to turn left. Note we do not assume any relation between $\lambda(\alpha)$ and $\lambda(\overline{\alpha})$.

Clearly there are at each node six ways to assign local numbers; three of them are completely equivalent in terms of which way is "left"; the other three make the opposite choice. Each choice of what constitutes "left" at a node can be made independently from choices at other nodes, and each combination of choices gives a valid oriented circuit map. Conversely, for every such map there is a local labeling that makes the direction it sends us look "left" at each node. So there are 2^{2h} different oriented circuit maps on a G with $n = 2h$ nodes.

Now let $\mathcal{O}_1$ and $\mathcal{O}_2$ be two oriented circuit maps with corresponding local labelings λ_1 and λ_2. Clearly, in terms of their own local labelings each map sends us left everywhere. But describing the situation in terms of λ_1 say, $\mathcal{O}_2$ may multiply λ_1 by ω (send us "left$_1$") at some nodes, and multiply λ_1 by $\overline{\omega}$ (send us "right$_1$") at other nodes,

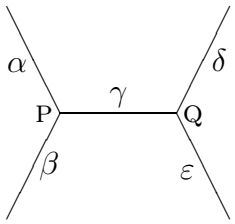
The simplest way to describe the situation is by fixing one canonical local labeling λ_0 (which could be the one based on ordering of global labels described above, or any other one) and *then* look at any of these 2^n oriented maps, $\mathcal{O}_i$ say. Now $\mathcal{O}_i$ has us going always left$_0$ at some nodes and always right$_0$ at some others. There is a function τ_i corresponding to $\mathcal{O}_i$ that assigns to each node the kind of turn taken,

$\tau_i(\text{P}) = \omega$ or $\overline{\omega}$ if we turn left$_0$ or right$_0$ at node P (signifying λ_0 gets multiplied by ω or $\overline{\omega}$ there). A choice of map $\mathcal{O}_i$ is equivalent to a choice of where-next function $L_i : \mathcal{A} \to \mathcal{A}$ (where $\mathcal{A}$ is the set of arcs), and expressed algebraically as a turning function $\tau_i : \mathcal{N} \to \{\omega, \overline{\omega}\}$.

5.2 Cisness

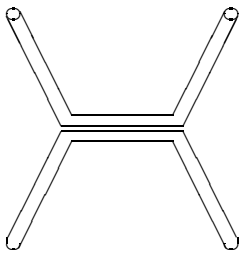
A map must have two circuits (cycles) that pass through edge γ in the picture below. Even leaving orientation (orientability and direction of travel) aside, one of these must pass through α and the other through β (with a third circuit visiting α and β in succession bypassing γ, on this visit of the node at least). This is a consequence of circuitness: if both circuits took the $\alpha\gamma$ route then β would still need to be visited by another circuit, but with α and γ doubly covered we would have to visit β and backtrack along β (not a circuit).

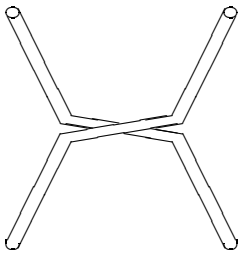
Applying this to both nodes in the picture we see that if one circuit takes the $\alpha\gamma\delta$ route the other must follow $\beta\gamma\varepsilon$ and vice versa; or else if one takes the $\alpha\gamma\varepsilon$ route the other takes $\beta\gamma\delta$ and vice versa.



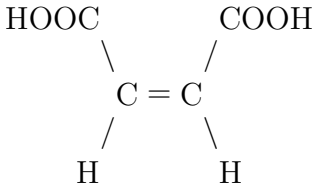
Thus there are precisely two possibilities with respect to edge γ between P and Q: either a map contains (undirected) ...$\alpha\gamma\delta$... and ...$\beta\gamma\varepsilon$... or it contains ...$\alpha\gamma\varepsilon$... and ...$\beta\gamma\delta$... We can regard this as a property of the edge γ, under a given map. I call it the **cisness** of the edge, in this map (the derivation of *cis* follows shortly).

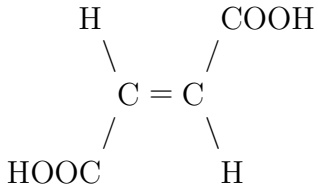
Firstly, consider an oriented circuit map $\mathcal{O}_i$. As in the previous section, it multiplies canonical local arc numbers λ_0 by $\tau_i(\text{P})$ at each node P. In other words, at some nodes it always sends us left$_0$ and at other nodes always right$_0$. Now we can express the cisness of the edge as $\pi_i(c) = \tau_i(\text{P}) - \tau_i(\text{Q})$. This π_i is, like the τ_i, an element of $\mathbb{F}_4$ but because any τ_i is ω or $\overline{\omega}$ any π_i is either 0 (if the τ_i are equal) or 1 (the difference between ω and $\overline{\omega}$ either way). So the differences are in $\mathbb{F}_4$ but effectively just numbers modulo 2, as only values 0 and 1 occur. Let's refer to those two possibilities as **cis**$_0$ for $\pi_i(c) = 0$ and **trans**$_0$ for $\pi_0(c) = 1$:





These terms were borrowed from systematic nomenclature in organic chemistry[16], where e.g. maleic acid (below left) and fumaric acid (below right) are called *cis* and *trans* isomers of 1,2-dicarboxyethene (or but-2-ene-1,4-dioic acid) respectively[17].





The nice thing of this description of a map is that the cis_0-or-trans_0 property of edges (unlike the left_0-or-right_0 property of nodes) is invariant under reversing a directed cycle. If $\mathcal{O}_i$ has c directed circuits, all the 2^c directed circuit maps obtained by reversing circuits (in general no longer oriented maps) have the same cisnesses. But then cisness isn't just a property of an edge in (both directed circuits through that edge in) any one of these directed circuit maps, but also a property of the edge in (both circuits in) the circuit map formed by forgetting direction. We can retire the notions of left_0 and right_0 turns which were only needed as scaffolding.

We can define cisness in an even more general way. Consider that in a graph embedded in a plane (or any orientable surface) one *could* choose left_0 and right_0 in a particular consistent way, but that we didn't actually bother to do so. Rather, left_0 and right_0 were assigned locally in an arbitrary way, independently at each node. We can do the same thing again: assign cis_0 and trans_0 locally in an arbitrary way, independently at each edge. For instance, arbitrarily set $\pi_0(\gamma) = 0$ (call it cis_0) if the routes taken were $...\alpha\gamma\delta...$ and $...\beta\gamma\varepsilon...$, and $\pi_0(\gamma) = 1$ (trans_0) if they were $...\alpha\gamma\varepsilon...$ and $...\beta\gamma\delta....$ Likewise define $\pi_0(\iota)$ arbitrarily for every edge ι. We can still refer to such a $\pi_0(\gamma)$ as the **cisness** of γ and describe any map in terms of this π_0.

[16] Which borrowed them from the Latin for "on this side" and "on the other side" respectively.

[17] The rate at which transition between such isomers occurs by rotation about the sp^2 "double" carbon-carbon bond is, unlike for an sp^3 "single" bond, low enough at room temperature for them to persist as separately identifiable substances.

Let's say a circuit **agrees** with a given cisness function π_0 if, on the edges it visits, it has that cisness. The importance of the concept of cisness lies in the following

Cisness theorem: cisnesses are circuit maps, circuit maps are cisnesses.

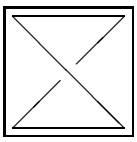
- Every circuit map determines a cisness function π_0, because the two times the edge is visited it must happen with the same cisness (as shown in the beginning of this section) and every edge has a cisness determined for it by the map. ■
- Conversely in a G with $m = 3h$ edges, each of the 2^{3h} different π_0 (choosing cisness independently at each edge) determines a circuit map by the following argument. For each edge to start with there are two arcs. For each arc (edge and direction to travel) there are two choices for the second edge. For each of those choices, follow the π_0 cisness of the second edge to (given the first edge) find the third edge, then follow the π_0 cisness of the third edge to (given the second edge) find the fourth edge, and so on. This never makes us backtrack on the same edge (although we may revisit the edge later) so is exactly a trek, and in a finite graph eventually a circuit (the next section details the ways in which we can meet up with ourselves). The procedure only finds circuits that agree with π_0 (the same s-circuit is found $2s$ times, once from each of its arcs) and there are only two possible undirected ways to travel along any edge that agree with its cisness, both extend to a unique circuit by following cisness on successive edges, so the procedure construct a set of circuits that covers every edge twice. ■

Moreover, the cisness formalism incorporates all the *relative* orientation at adjacent nodes we need (a little patch of planarity if you will) without any global orientation. We don't need orientable surfaces for this kind of relative orientation. And we don't need directed circuits now that absolute direction of turn at one node is removed as the scaffolding it was for relative direction of turn at adjacent nodes.

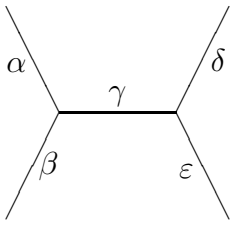
The relation with the 2^{2h} oriented maps of the previous section is less than straightforward. Let $2^{3h} = m_1 + m_2 + m_3 + \cdots$ where m_c denotes the number of maps with c (undirected) circuits. For each such map there are 2^c ways to assign directions to the circuits, giving $2m_1 + 4m_2 + 8m_3 + \cdots$ directed circuit maps. Many of these will not be oriented though (we visit some edges in the same direction both times); we saw above there must be exactly 2^{2h} left that are. Example: K_4 (the tetrahedrom) has $m_1 + m_2 + m_3 + m_4 = 28 + 28 + 7 + 1 = 2^6$ circuit maps, only one of which (the one with 4 cycles) orientable, giving 2^4 oriented maps.

5.3 Cycle maps

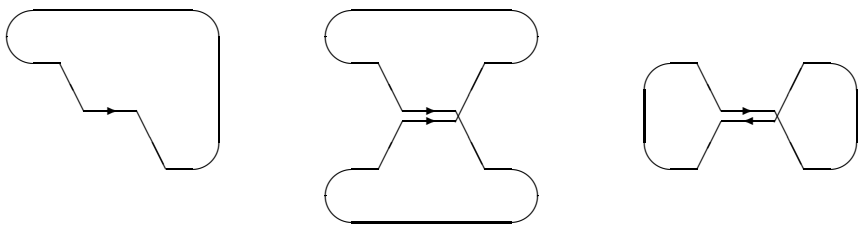
Continuing with the K_4 example, that one orientable circuit map consisting of the four 3-cycles is also a cycle map. It is not the only one; the map consisting of the three 4-cycles is the other one. It is neither proper (any two 4-cycles share two non-adjacent edges) nor orientable (any two 4-cycles, considered directed, traverse one shared edge in the same direction and one in opposite directions). Any two of the 4-cycles are related as



Let G be trivalent and bridgefree, in particular it has no loops. We saw any (undirected) circuit map is given by a choice of cisness π_0 on G; now let's investigate under which circumstances the circuit map given by π_0 is a cycle map.



Let Γ be one of the circuits of the map; choose a direction for it as directed circuit $\vec{\Gamma}$, and let γ be the i-th edge visited by $\vec{\Gamma}$, and WLOG let it pass through edges $\alpha\gamma\varepsilon$ in that order at that occasion. If we keep following the circuit it must visit γ again; let the next time it crops up be as j-th edge. It can do so in three ways:



- Again via $\alpha\gamma\varepsilon$ (left picture): we just close the circuit, it repeats after $j - i$ steps. If the circuit is this well behaved at every edge it's a cycle.

  The other time γ is visited (via $\beta\gamma\delta$ or $\delta\gamma\beta$) happens in another circuit.

- Via $\beta\gamma\delta$ (middle picture): while this isn't a cycle, if nothing worse happens it can easily be repaired by choosing the other cisness at γ. This disconnects the 8 shape and reconnects it into two separate cycles.
- Via $\delta\gamma\beta$ (right picture): again the two visits of γ happen in the same circuit, but this case isn't so easy to turn into part of a cycle map.

We see the problematic case is that where, after visiting an edge PQ in that order of vertices, we come across it as QP the next time.

Note that a fourth case, after γ being the i-th edge via $\alpha\gamma\varepsilon$ it being the j-th edge

- via $\varepsilon\gamma\alpha$

cannot happen. By the given cisness, the part of the circuit leading up to the j-th edge would have to retrace the part after the i-th edge. Now if $j-i$ is even the $\frac{1}{2}(i+j)$-th edge would have to run from and to the same node (contradicting the absence of loops); if $j-i$ is odd the $\frac{1}{2}(i+j+1)$-th edge would have to retrace the $\frac{1}{2}(i+j-1)$-th edge (contradicting circuitness).

The following lemma might be useful as last step for any proof of the existence of cycle maps:

Lemma Ω: if π is a cisness function on G such that for every edge ε of G there is a cycle $\mathbf{c}(\varepsilon)$ through ε that agrees with π everywhere, then there exists a cycle map on G. Proof: let

$$\mathcal{U} = \bigcup_{\varepsilon \in \mathcal{E}} \{\mathbf{c}(\varepsilon)\}$$

(that is, a set of cycles such that for every edge ε it contains one cycle $\mathbf{c}(\varepsilon)$ through it that had to exist, but being a set any duplicate $\mathbf{c}(\varepsilon') = \mathbf{c}(\varepsilon'')$ now occurs only once).

Any set of cycles will cover any edge ε a certain number k_e times (that is, the set contains k_e cycles that pass through ε), and all we would normally know is that $k_e \geq 0$ and an integer. Of our $\mathcal{U}$ however we know two more things:

- For every ε, $k_e \geq 1$ (it has at least one cycle through ε), and
- for every ε, $k_e \leq 2$ (we saw earlier that a complete π allows precisely two circuits through ε that follow π everywhere, any cycle in $\mathcal{U}$ is one of those two circuits).

So we now have a set of cycles that covers any edge either once or twice. Let $\mathbf{u}$ be the sum of all cycles in $\mathcal{U}$, it contains exactly those edges covered only once by $\mathcal{U}$. Being in $\mathbf{Z}$, it is a sum of zero or more disjoint cycles $\mathbf{c}_i$.

It is easy to see none of the $\mathbf{c}_i$ can be in $\mathcal{U}$. Let γ be an edge of $\mathbf{c}_i$, P one of its endpoints, α and β the other edges at P. Now $\mathbf{c}_i$ passes through two edges at P, WLOG β and γ. Because $\mathbf{c}_i$ is a component of $\mathbf{u}$, $\mathcal{U}$ covered β and γ only once. $\mathcal{U}$ can't cover α zero times (it covered every edge at least once) nor one time (sum of covers of α, β, γ is even) so it covered α twice. But if $\mathbf{c}_i \in \mathcal{U}$ then $\mathcal{U} \setminus \mathbf{c}_i$ covers α twice, and β and γ not at all, which is impossible to realise by a set of cycles.

Now that none of the $\mathbf{c}_i$ is in $\mathcal{U}$, we can construct $\mathcal{M} = \mathcal{U} \cup \{\mathbf{c}_0, \mathbf{c}_1, \mathbf{c}_2, \ldots\}$ as disjoint union. Every edge that was covered only once by $\mathcal{U}$ is now covered a second time by one of the $\{\mathbf{c}_i\}$ so $\mathcal{M}$ is a cycle map. ■

The cycle map is of course the circuit map determined by π. As every edge already contains at least one cycle in $\mathcal{U}$ that follows the cisness, a second cycle there will also have to follow it, by $\mathcal{M}$ being a map.

We can also describe the concept of a cycle double cover in terms of reduced vectors rather than the complete $\mathbf{E}$. In a 3-edge-connected G find a basis $\mathcal{B}^\circ$ for $\mathbf{E}^\circ$ and define $f_i(\mathbf{v}^\circ)$ as the i-th component (0 or 1) of $\mathbf{v}$ when written with respect to $\mathcal{B}^\circ$. Let $\mathbf{C}_i^\circ$ be the set of reduced vectors for which the component is 0, an $o-1$-dimensional subspace. Intersect its complement with $\mathcal{E}^\circ$, the set of reduced vectors that represent single edges. $(\mathbf{E}^\circ \setminus \mathbf{C}_i^\circ) \cap \mathcal{E}^\circ$ represents the edges of some element $\mathbf{c}_i$ of $\mathbf{Z}$. Now restrict the basis to a set $\mathcal{B}^\ddagger \subseteq \mathcal{B}^\circ$ and let the restricted weight of $\mathbf{v}^\circ$ be the sum of $f_i(\mathbf{v}^\circ)$ for those i where the i-th basis vector is in $\mathcal{B}^\ddagger$. Choosing $\mathcal{B}^\ddagger$ such that the restricted weight of all vectors in $\mathcal{E}^\circ$ is exactly 2 makes the set $\mathbf{c}_i$ (for those i where the i-th basis vector is in $\mathcal{B}^\ddagger$), together with their sum (in $\mathbf{Z}$), a cycle double cover (or rather, their disjoint cycles are). And vice versa.

5.4 Combinatorial genus

In chapter **2**, one concept was borrowed from topology: genus. It was *defined* combatorially (for cycle maps, rather than surfaces) using

$$n - m + f \;=\; 2 - 2g$$

so depends, given a G (and hence n and m) only on the number of cycle faces f used (we can adopt the same definition for *circuit* maps).

If we avoid topological proofs and the connexion between topological and combinatorial genus, then we will need independent proofs of a few facts of life of genus, for trivalent G.

Theorem: combinatorial genus g is a nonnegative integer or half-odd[18].

Proof: That g is integer or halfodd is clear from the definition; remains to prove that $g \geq 0$. First consider the case that $\mathcal{M}$ is a cycle map with f cycles $\mathbf{c}_i$ on a connected trivalent G. Clearly $\sum \mathbf{c}_i = \mathbf{0}$ as every edge occurs twice in the sum. More generally, let $\mathcal{M}$ be any circuit map with f circuits γ_i, again on a connected trivalent G, and let $\mathbf{c}_i$ represent the sum of edges occurring in γ_i. Each $\mathbf{c}_i$ is still in Z but may now be the sum of disjoint cycles (an edge covered twice by γ_i cancels in $\mathbf{c}_i$). Again, $\sum \mathbf{c}_i = \mathbf{0}$ as every edge is covered either twice (once each by different $\mathbf{c}_i$) or not at all (twice by a single γ_i and hence zero times by every $\mathbf{c}_i$).

The question arises whether the $\mathbf{c}_i$ corresponding to the γ_i of any subset $\mathcal{M}'$ of $\mathcal{M}$ can already sum to $\mathbf{0}$. In that case the same would be true for $\mathcal{M}'' := \mathcal{M} \setminus \mathcal{M}'$ as well. Every edge is covered exactly twice by the circuits of $\mathcal{M}$, and must now be covered an even number of times by $\mathcal{M}'$ and $\mathcal{M}''$ separately, that is, covered twice by the one and then not by the other, or vice versa. Keeping in mind that G is connected there must now be nodes where those two populations of edges meet. At some node P there would be WLOG one edge covered by $\mathcal{M}'$ and the other two by $\mathcal{M}''$. But there is no circuit or combination of circuits that can give $\mathcal{M}'$ that pattern of coverage at P as circuits don't backtrack.

So there can be no linear combination involving only some of the $\mathbf{c}_i$ that already sums to $\mathbf{0}$. This puts an upper bound on f. Let o again be the dimension of $\mathbf{Z}$. There are no linearly independent sets of vectors in $\mathbf{Z}$ containing more than o vectors. The set of f vectors $\mathbf{c}_i$ sums to $\mathbf{0}$ but we just saw any of its subsets is linearly independent. So

$$f - 1 \;\leq\; o \;=\; \dim Z \;=\; m - n + 1 \;=\; h + 1$$

where we use the connectedness of G in setting k to 1, and its trivalency in setting $m - n = 3h - 2h$. In terms of the genus g for which, in a connected trivalent G,

$$2 - 2g = n - m + f \;=\; f - h$$

we now find $2 - 2g \leq 2$. ■

[18]Half of an odd integer (multiple of $\frac{1}{2}$ that's not integer). I adopted it long ago from P. A. M. Dirac in *Principles of Quantum Mechanics*, but did find it is rare (the OED doesn't have it).

5.5 Combining maps

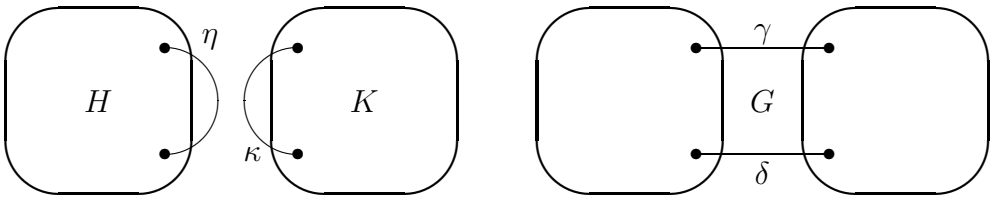
Let $\mu(G)$ denote the number of cycle maps of G, and $\mu_\circ$ the number of them that are orientable. Now let G be split into smaller graphs H and K in some fashion, then we can investigate how μ or $\mu_\circ$ of G depends on those of H and K. Summarising the results first:

$$\begin{aligned}
\mu(H\ K) &= \mu(H)\,\mu(K) \\
\mu(H-K) &= 0 \\
\mu(H=K) &= 2\,\mu(H)\,\mu(K) \\
\mu(H\equiv K) &= \mu(H)\,\mu(K)
\end{aligned}$$

and the same relations for $\mu_\circ$.

The simplest case is $\mu(H\ K)$ where (if H and K are both connected) G has them as its two components, or (more generally) G consists of all the connected components of H together with all the connected components of K. Clearly every one of the maps on H can be combined with any map on K, making the total number $\mu(G)$ the product of $\mu(H)$ and $\mu(K)$.

In the $\mu(H=K)$ case, we construct G by taking any edge η in H and any edge κ in K, breaking the edges and reconnecting each loose end in H with one in K, forming new edges γ and δ:

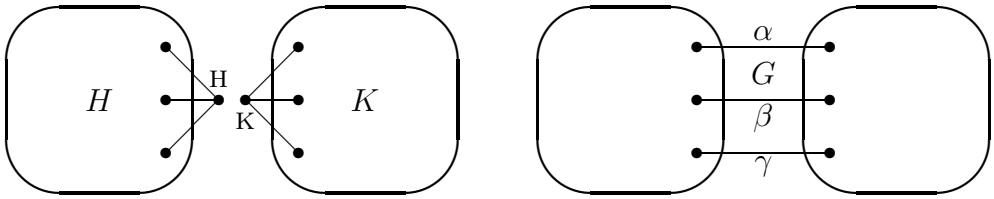


Of course we could have crossed over the edges, and moreover we could have picked any other edge in H, and in K. So "$H=K$" does not uniquely define a graph G. However, the expression $\mu(H=K)$ is still well-defined, it denotes the map count for any *one* G constructed this way.

Let $\mathcal{H}$ be any map on H with cycles A and B through η, and $\mathcal{K}$ any map on K with cycles Γ and Δ through κ. In a given G, we can combine these maps by hooking up either A with Γ and B with Δ, or A with Δ and B with Γ, creating $2\mu(H)\mu(K)$ maps in all. Conversely, every map on G induces maps on H and K along these

lines, so these $2\mu(H)\mu(K)$ maps are all of them. And it proves $\mu(H=K)$ does not depend on the way we splice the graphs together.

For $\mu(H \equiv K)$ we go a step further. This time we will lose a node H from H and one K from K, and the total number of edges will go down by three. In H, let α', β' and γ' be the edges to H; in K, let α'', β'' and γ'' be the edges to K; now in G we connect α' to α'', β' to β'', γ' to γ'', forming new α, β, γ:



We saw any map $\mathcal{H}$ on H has at H one cycle A' that goes via γ' and β', one B' via α' and γ', and one Γ' via β' and α'. Likewise, a map $\mathcal{K}$ on K has at K one cycle A'' via γ'' and β'', one B'' via α'' and γ'', and one Γ'' via β'' and α''. We can only combine A' with A'', B' with B'', and Γ' with Γ'', which always gives a valid cycle map on G, $\mu(H)\mu(K)$ of them. Conversely it is easily seen every cycle map on G has one cycle A via γ and β, one B via α and γ, and one Γ via β and α, which split into valid maps on H and K.

Note also that anything said thus far for μ also holds for $\mu_\circ$, because in each of these constructions it is both a necessary and sufficient condition for a map of G to be orientable that the two constituent maps are.

We haven't yet mentioned $\mu(H-K)$. While there are ways to tweak H and K into having one loose end (for instance, put an extra node along an existing edge) it doesn't matter: the main reason for inclusion of $\mu(H-K)$ in the list is to remind us that any scheme that produces a G with a bridge produces a G with no maps at all.

We already knew finding $\mu(G)$ for disconnected G reduces to a product $\mu(H)\mu(K)$, and that $\mu(G) = 0$ for 1-edge-connected G. Thanks to the analysis above we can now also decompose all 2-edge-connected G, and "all" 3-edge-connected ones.

The latter "all" is between scare quotes because decomposition is not always useful. If G is solid (the only 3-edge cuts are those around a node) then the only way to decompose it as $H \equiv K$ is with either H or K equal to the original G (and the other a triply bonded 2-edge multigraph). So the decompositions carried out here are possible for graphs that are less than 3-edge-connected and those that are "really"

only 3-edge-connected (have a 3-edge cut that's not a node) and stops at solid graphs.

It would seem the next step is combining graphs by 4 edges (removing two adjacent nodes from each of H and K), 5 edges, and so on. However, with such compositions $\mu(G)$ can no longer be predicted merely from $\mu(H)$ and $\mu(K)$. For instance, suppose G has four edges α, β, γ, δ to be split, in H we pair up α' and β' to join at a new node P, γ' and δ' at a new node Q, with a fifth edge ε' between P and Q, and we do the same things in K with a new edge ε''. Now maps $\mathcal{H}$ on H and $\mathcal{K}$ on K do not necessarily combine. If the maps are both cis or both trans at ε' and ε'' they do, but if they are one of each then two cycles in each will loop round as one single big circuit in G. If $\mathcal{H}$ and $\mathcal{K}$ are cycle maps we can only say we have a circuit map on G, we need $\mathcal{H}$ and $\mathcal{K}$ to be proper cycle maps to make the combined map just a cycle map, we need them to be proper cycle maps with an additional property (that neighbouring cycles have no more than two shared neighbours) to make the combined map just a proper cycle map, and so on. There is no flavor of map of which the numbers simply propagate.

So solid graphs are irreducible as far as decomposition goes. Any graph more flimsy than that can be decomposed into solid components, with $\mu(G)$ and $\mu_\circ(G)$ able to be derived from those of its components, but there it ends.

Note the decompositions of trivalent graphs here are exactly those of chapter **3**. There we were interested in the existence of Tait colorings. The connexion is not accidental: the existence of a Tait coloring is equivalent to the existence of a cycle map using the red-green cycles, the blue-red cycles, and the green-blue cycles. This is necessarily an *even* map and in chapter **3** we were careful to match colors of edges, that is, preserve evenness of the cycles. Here the construction doesn't care about cycles being of odd or even length but the argument goes a step further in being quantitative, number of maps rather than existence.

5.6 Existence lemmata

Let G be bridge-free trivalent, and let $\mathbf{Z}$ be again the subspace of $\mathbf{E}$ spanned by the cycles.

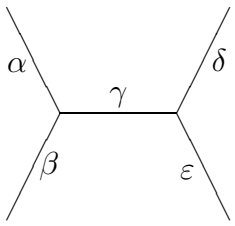
Lemma I: for any edge ε, there exists a cycle through ε.

Proof: by definition since ε is not a bridge, as we saw earlier. ■

Lemma V: for any two adjacent edges α and β (meeting at node P say), there exists a cycle through α and β.

Proof: Let γ be the third edge at P. There are cycles $\mathbf{u}$ through α and $\mathbf{v}$ through β by the previous lemma. If any one one of them goes via $\alpha\beta$ we're done. Otherwise, $\mathbf{u}$ and $\mathbf{v}$ go via $\alpha\gamma$ and $\beta\gamma$ respectively and we consider $\mathbf{u}+\mathbf{v}$, which is like $\mathbf{u}$ and $\mathbf{v}$ an element of $\mathbf{Z}$. Recall (from the remark on page 14 following the proof of the **XYZ** lemma) that such an element is a disjoint union of cycles, say $\mathbf{u}+\mathbf{v} = \mathbf{c}_0+\mathbf{c}_1+\cdots$. Edge γ cancels in the sum $\mathbf{u}+\mathbf{v}$ but α and β don't so $\mathbf{u}+\mathbf{v}$ contains them. Whichever one of the cycles $\mathbf{c}_i$ contains α must also go via β, because it doesn't contain γ. ■

For the next lemma, let α, β and γ be again the edges at one node, and γ, δ and ε the edges at the other endpoint of γ.



Lemma H: if a cycle exists that takes the $\alpha\gamma\delta$ route (let's call this way round cis), then one exists that takes the the other cis route $\beta\gamma\varepsilon$. Of course this must then be true vice versa, and likewise for the two trans routes.

One proof: let $\mathbf{u}$ be the cycle via $\alpha\gamma\delta$. By the previous lemma there is a cycle $\mathbf{u}'$ via $\beta\gamma$; if it goes via $\beta\gamma\varepsilon$ we're done so assume it goes via $\beta\gamma\delta$. There is also cycle $\mathbf{u}''$ via $\gamma\varepsilon$; if it goes via $\beta\gamma\varepsilon$ we're done so assume it go via $\alpha\gamma\varepsilon$. Now $\mathbf{u}+\mathbf{u}'+\mathbf{u}''$ contains β, γ and ε (counting the parities of numbers of times it covers the edges) so one of its disjoint cycles goes via $\beta\gamma\varepsilon$. ■

If we write vectors of $\mathbf{E}$ in binary with the coördinates for edges α through ε as the first five digits, then the addition is `10110...` + `01110...` + `10101...` = `01101...`

Another proof: a cycle $\mathbf{v}'$ through $\alpha\beta$ stays away from γ, so it takes both δ and ε or neither. Likewise, a cycle $\mathbf{v}''$ through $\delta\varepsilon$ stays away from γ, so takes both α and β or neither. By lemma **V** both exist. Now $\mathbf{v}'$ may or may not pass through $\delta\varepsilon$; if does so, call it $\mathbf{w}$. Else check if $\mathbf{v}''$ passes through $\alpha\beta$; if it does, call it $\mathbf{w}$. If neither did, call their sum $\mathbf{w}$. This way we always have an element $\mathbf{w}$ of $\mathbf{Z}$ passing through both $\alpha\beta$ and $\delta\varepsilon$ (which can be added to a cis path to give the other cis

path, or to a trans path to give the other trans path). ■ Note we do not necessarily have something that can be added to a cis path to turn it into trans and vice versa.

The reason the first proof worked straightaway is because not just the digits `1`, but also the digits `0` were guaranteed in the various cycles (for instance, $\alpha\gamma\varepsilon$ stays away from β and δ).

We already know there are *some* cycles through γ, and any cycle must take a cis path or a trans path. So both cis paths though γ are taken by cycles, or both trans paths are, or all four are.

The second, longer, proof of the last lemma also told us that all four paths are taken whenever cycles $\mathbf{v}'$ exist that visit $\alpha\beta$ without going via $\delta\varepsilon$, or cycles $\mathbf{v}''$ exist that visit $\delta\varepsilon$ without going via $\alpha\beta$ (if either exists the other does too, by adding $\mathbf{w}$ that goes via $\alpha\beta$ and $\delta\varepsilon$). The converse is also true: if cis *and* trans paths exist we can add a cis path to either of the trans paths to give such $\mathbf{v}'$ and $\mathbf{v}''$. So there are two possibilities as regards to any edge γ:

- All four paths (both cis and both trans) are taken by cycles, and there are cycles that take one of $\alpha\beta$ and $\delta\varepsilon$ but not the other, and vice versa.

- Only two of the four paths (both cis or both trans) are taken by cycles, and the only cycles via $\alpha\beta$ go via $\delta\varepsilon$, and vice versa. Such a cycle has γ as a **chord** (Szekeres calls a chord a *canal*).

Lemma X: there exist cycles via $\alpha\gamma\delta$ and via $\beta\gamma\varepsilon$ (both the cis routes say), and/or there exist cycles via $\alpha\gamma\varepsilon$ and via $\beta\gamma\delta$ (both the trans routes).

Of course, this follows immediately from lemma H as we just saw. However, it is instructive to see how it is also possible to prove this weaker result straight from lemma V without relying on addition in $\mathbf{Z}$ (we will need such arguments in the next chapter).

By lemma V there is a cycle through $\alpha\gamma$ which must then go via $\alpha\gamma\delta$ or $\alpha\gamma\varepsilon$, WLOG let it go via $\alpha\gamma\delta$. Now there is also a cycle through $\beta\gamma$, if it goes via $\beta\gamma\varepsilon$ we're done ($\alpha\gamma\delta$ and $\beta\gamma\varepsilon$) so assume it goes via $\beta\gamma\delta$. There is also a cycle through $\gamma\varepsilon$, if it goes via $\beta\gamma\varepsilon$ we're done ($\alpha\gamma\delta$ and $\beta\gamma\varepsilon$) so assume it goes via $\alpha\gamma\varepsilon$. But now we're done the other way ($\beta\gamma\delta$ and $\alpha\gamma\varepsilon$). ■

5.7 Path counts

Here the same lemmata are revisited. This time, existence is sharpened with quantitative arguments. We can still assume G to be bridge-free, so every edge has cycles passing through it.

If $\alpha...\omega$ is an open or closed path (i.e. a path or cycle), let $\zeta(\alpha...\omega)$ denote the number of elements of $\mathbf{Z}$ that contain every edge of $\alpha...\omega$ (in other words, that are the disjoint sum of one or more cycles one of which passes through $\alpha...\omega$).

Recall that $\dim Z = o := m - n + k$, so for the empty path

Lemma O#: $\zeta(\,) = 2^o$. ■

Some of these will pass through a given edge ε, and some not.

Lemma I#: for every edge α, $\zeta(\alpha) = 2^{o-1}$. There is always some element $\mathbf{v}_0$ of $\mathbf{Z}$ that passes through α (because G is bridge-free) and some that doesn't, $\mathbf{u}_0 := \mathbf{0}$ for instance. Let there be y distinct $\mathbf{v}_j$ that do and x distinct $\mathbf{u}_i$ that don't, so $x + y = 2^o$. Every $\mathbf{u}_i + \mathbf{v}_0$ is a distinct $\mathbf{v}_j$ so $y \geq x$, and every $\mathbf{v}_j + \mathbf{v}_0$ is a distinct $\mathbf{u}_i$ so $x \geq y$, so $x = y = 2^o/2$. ■

Lemma V#: for any adjacent edges α and β, $\zeta(\alpha\beta) = 2^{o-2}$. Let the third edge at this node again be γ, and let $\mathbf{u}_0$ be an element of $\mathbf{Z}$ passing through $\alpha\gamma$, $\mathbf{v}_0$ one through $\beta\gamma$, and $\mathbf{w}_0$ one through $\alpha\beta$ (they all exist by lemma V). Every $\mathbf{u}_i$ through $\alpha\gamma$ is $\mathbf{v}_j + \mathbf{w}_0$ for some $\mathbf{v}_j$ through $\alpha\gamma$, etc. etc. So the numbers of each are equal. Finally, every $\mathbf{t}_k$ through none of the three edges is $\mathbf{u}_i + \mathbf{u}_0$ for some $\mathbf{u}_i$ and vice versa. There are $2^o/4$ of each kind.

For the next two lemmata, let $\alpha...\omega$ run from node A to node Z, and let ω, ρ and σ be the three edges at Z (if $\alpha...\omega$ is closed then Z = A and one of ρ and σ is α).

Lemma Y#: $\zeta(\alpha...\omega) = \zeta(\alpha...\omega\rho) + \zeta(\alpha...\omega\sigma)$. Obvious. ■

If $\zeta(\alpha...\omega\sigma) = 0$ then $\zeta(\alpha...\omega\rho) = \zeta(\alpha...\omega)$. We will say $\alpha...\omega$ **forces** $\alpha...\omega\rho$ in such a situation, and that it **forces** it **non-trivially** if $\zeta(\alpha...\omega)$ (and hence $\zeta(\alpha...\omega\rho)$) wasn't already zero.

Lemma Z#: either $\zeta(\alpha...\omega\rho)$ and $\zeta(\alpha...\omega\sigma)$ are equal, or one of them is zero.

Suppose neither is zero. Now there is an element $\mathbf{u}_0$ of $\mathbf{Z}$ through $\alpha...\omega\rho$, and an element $\mathbf{v}_0$ through $\alpha...\omega\sigma$. Their sum $\mathbf{t} = \mathbf{u}_0 + \mathbf{v}_0$ passes through $\rho\sigma$ and not through any edge of $\alpha...\omega$. Let there be x distinct $\mathbf{u}_i$ through $\alpha...\omega\rho$ and y distinct $\mathbf{v}_j$ through $\alpha...\omega\sigma$. Every $\mathbf{u}_i + \mathbf{t}$ is a distinct $\mathbf{v}_j$ so $y \geq x$, and every $\mathbf{v}_j + \mathbf{t}$ is a distinct $\mathbf{u}_i$ so $x \geq y$. ■

Note that by the last two lemmata, if $\zeta(\alpha...\omega) = z$ then each of $\zeta(\alpha...\omega\rho)$ and $\zeta(\alpha...\omega\sigma)$ is z or $z/2$ or 0.

Lemma B#: every $\zeta(\alpha...\omega)$ is a power of two, or zero. We saw $\zeta(\,)$ and $\zeta(\alpha)$ are, and each time we lengthen the path by another edge the number stays the same, gets exactly halved, or becomes zero. ■

Alternative proof: $\mathbf{Z}$ is a subspace of $\mathbf{E}$, the set $\mathbf{P}$ of vectors of $\mathbf{E}$ whose α-th, ... ω-th bits are all 0 is also a subspace, and so $\mathbf{Z}\cap\mathbf{P}$ is too, and hence has 2^d members for some d. Now let $\mathcal{Q}$ be the set of vectors of $\mathbf{E}$ whose α-th, ... ω-th bits are all 1. Either $\mathbf{Z}\cap\mathcal{Q}$ is empty, so $\zeta(\alpha...\omega) = 0$, or it isn't and contains some $\mathbf{v}_0$, now every $\mathbf{v}_i$ it contains is $\mathbf{u}_i + \mathbf{v}_0$ for some $\mathbf{u}_i$ in $\mathbf{Z}\cap\mathbf{P}$ and vice versa and so $\zeta(\alpha...\omega) = 2^d$. ■

Let $\alpha...\omega$ again run from node A to node Z with ω, ρ and σ the three edges at Z, and let α, λ and μ be the three edges at A.

Lemma W#: if $\alpha...\omega$ doesn't force $\alpha...\omega\rho$ but $\lambda\alpha...\omega$ does force $\lambda\alpha...\omega\rho$ non-trivially, then $\alpha...\omega$ doesn't force $\lambda\alpha...\omega$ but $\alpha...\omega\rho$ does force $\lambda\alpha...\omega\rho$ non-trivially.

Crucially, $\lambda\alpha...\omega\rho$ has nonzero ζ because it's forced nontrivially, so all of its ancestors have nonzero ζ. Set $\zeta(\alpha...\omega) = z$. Each of the paths down to $\lambda\alpha...\omega\rho$ uses two steps, so its ζ could only be z (forced at both steps), $z/2$ (forced once), or $z/4$ (forced at neither step). We are given that one path contains a force step, and that one path contains a non-force step. So only $z/2$ is possible. Now we can deduce the forced or non-forced status of the two remaining steps. ■

Lemmata I and V were sharpened into I# and V# by quantifying the number of elements of $\mathbf{Z}$ involved, and we can do something like it for lemma H too.

Lemma H#: $\zeta(\alpha\gamma\delta) = \zeta(\beta\gamma\varepsilon)$. We saw there was an element $\mathbf{w}$ of $\mathbf{Z}$ passing through both $\alpha\beta$ and $\delta\varepsilon$, but not γ. Adding this to all the distinct $\mathbf{u}_i$ through $\alpha\gamma\delta$ (and hence not through β nor ε) gives distinct $\mathbf{v}_j$ through $\beta\gamma\varepsilon$ (and hence not through α nor δ), and vice versa. ■

Finally, when does forcing happen, and why? Recall $\mathbf{Z}$ is a binary code with m-

letter words, but only 2^o of the 2^m possible words. Because it is linear there is a quite specific scheme which words exist, relative to each other. There are several sets of $o = h + 1$ letter positions that are maximally independent, that is, all 2^o combinations of values of those letters occur, and all the other $m - o = 2h - 1$ letters are then determined by them (as a linear combination of values of letters in the independent set). In the graph, such an independent set consists of the lianes outside some spanning tree. Suppose the d'th letter is the sum of the a'th, b'th and c'th. That means all four sum to zero (so any of the four is the sum of the others), and the corresponding set of edges is of course none other than a cut.

Being a subspace, the code never forces a set of letter positions to sum to 1, only to 0. That means that

- If a path contains all but one of the edges of a cut with an **even** number of edges, the cut forces the path to take the remaining edge of the cut too (if the path is to have any cycles passing through it). Wherever it is in the graph, the path will have to go there eventually.
- If a path contains all but one of the edges of a cut with an **odd** number of edges, the cut forces the path to *not* take the remaining edge of the cut (if the path is to have any cycles passing through it). If the path reaches that edge (which it doesn't have to) it is forced to take the alternative instead.

This is also obvious from what a cut $\mathcal{E}_{\bullet\circ}$ is. If we are in $\mathcal{N}_\bullet$ we want a cycle to bring us back again. That means an even number of hops across the cut.

Keep in mind that when a path is forced to do something it may be due to only some subset of its edges being in a cut. The coding theoretic considerations effectively extend the notion of **forcing** from contiguous paths to various sets of edges.

5.8 Mapping the graph

Let's construct a path Ω as follows. To start, take any edge (e.g. α in the diagram on page 91) and any edge adjacent to it (e.g. γ in that diagram). Thus far our Ω consists of $\alpha\gamma$. By lemma V there are cycles through this elbow joint.

To extend the path, consider such cycles must either pass through δ next or through ε. There may be some of both kinds, or only one, but cycles of at least one kind exist. So WLOG say there are some that take the $\alpha\gamma\delta$ route (call this the cis route). Then by lemma H there also exist cycles that take the other cis route $\beta\gamma\varepsilon$. We

extend our path with (in this case) δ, such that there are still cycles through this path $\alpha\gamma\delta$.

We can extend the path again in the same way (with γ and δ taking the rôle of α and γ in the previous step) and by the same argument the new path, $\alpha\gamma\delta\rho$ say, will be such that there is still a cycle through it.

Note we construct the path such that there are elements of $\mathbf{Z}$ passing through all the edges which is possible by the vector space arguments in the preceding sections. An element of $\mathbf{Z}$ consists of edge-disjoint cycles, but in a trivalent graph they are also node-disjoint, so the construction does ensure there is at least one cycle that on its own passes through the whole path.

We can keep extending the path in the same way. Sooner or later (in a finite graph) we run out of edges so must come across ourselves again. Say the path is $\alpha\gamma\cdots\psi\omega$ and one (or both) of the edges adjacent to ω in the forward direction (i.e. not sharing the same node with it that ψ does) is an edge already in the path. Let "the path **meets itself**" refer to this situation.

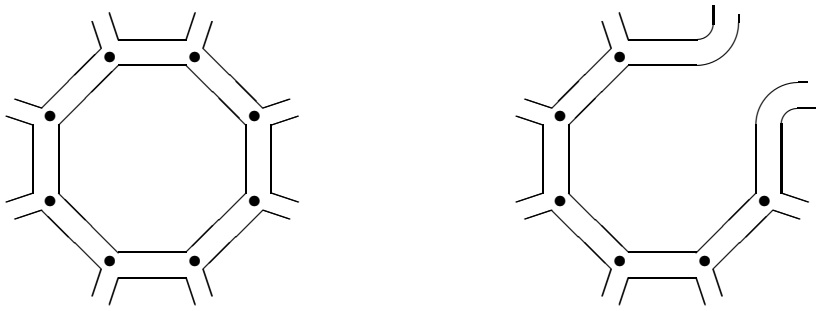
Non-rho lemma: the first time the path meets itself it does so by the last edge ω being adjacent to the first edge α (at the node it doesn't share with the second edge γ), and not adjacent to to any other edge of the path.

Any other way to meet ourselves would make it impossible for a cycle to contain all the edges of our path thus far, as it would use all three edges at some node. ■

So our path has *become* the one cycle left that still passes through it. The elements of $\mathbf{Z}$ still containing all the edges of our path are those that contain this cycle (and by lemmata in the previous section there are a power-of-two number that do).

The construction of the path seems obvious, if all it did was find a cycle. But it also tells us a less obvious thing: at every edge of the way we picked up a set of cycles that shared the same cisness as our cycle but took the other route; let's refer to these as its **neighbours**.

One way to extend the argument would be to start walking from one of the loose ends. However, trying to prove something like the non-rho lemma for neighbouring cycles simultaneously soon becomes daunting.



Another approach would be to file the cycle away as "done" and remove it from the graph somehow, then focus on the smaller graph that's left. The image shows one way to remove one cycle. If O is the cycle we just mapped, and $\psi = \text{PS}$ one of its edges, we remove ψ as well as P and S, making the other two edges through P, OP and PQ say, one edge OQ, and the other two edges through S, RS and ST say, one edge RT. The former O is merged with any potential neighbour across ψ to a new partial cycle Ω. This makes n go down by 2, m by 3, and o by 1, as it should.

Let's call one such step an **iteration**. Its effect on n, m and o can be summarised by saying it decrements h by one.

If we **preserve the cisness** on the edges of O we keep, we can later put it back. So the strategy to map the whole graph would be to map a cycle, retire it which gives us a new partially mapped Ω, extend that to a cycle and repeat. Ending up with the tetrahedron K_4 which can be mapped. Then put everything back in reverse order. By induction, every graph would be able to be mapped, proving the cycle double conjecture. Moreover we can pick any cycle of a given G as the first cycle (the one that doesn't inherit a partial Ω) proving the conjecture in the stronger form.

One problem with this scheme is that edge-connectivity can go down by 1 at any iteration, because one edge is removed altogether. The construction is the reverse of the | | to ⟩⟨ construction briefly discussed on page 25 but our problems are different here. We haven't got a complete free hand where the path leads us (it may be forced) so may not be able to avoid the graph becoming first 2-edge-connected and then 1-edge-connected which would scupper mapping.

This is actually one one form of a bigger problem: when we merge O with one of its neighbours (A say, across edge ψ) how do we know that neighbour isn't one of O's other neighbours as well (across edge ϕ say)? If so we're sunk, because the new Ω will border itself across edge ϕ. This problem is bound to occur when edge connectivity is 2 (now there must be two faces bordering each other in every

cycle map) and goes down to 1 on removing one edge of the co-bridge (now every circuit map has a circuit bordering itself at the bridge and there are no cycle maps). However, it can occur locally at any time even if edge connectivity is high.

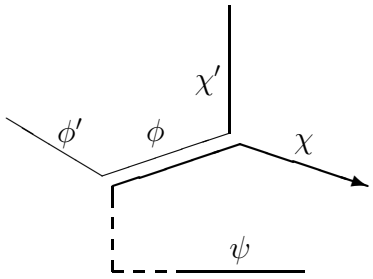
Demanding all maps are proper (no cycles neighbouring each other more than once) so that on the next iteration it is still a map is not, on its own, a solution as it only staves off the problem by one iteration. There is an endless regression of stronger conditions there. Besides, some graphs ($K_{3,3}$) only have improper maps.

Good neighbour lemma: we can arrange for A not to clash with the other neighbours.

Of course when we talk about neighbours we haven't yet chosen them. In stead of "a neighbour A" we have (for each of the edges of O) just three edges, determining cisness on only the middle one of them (the one shared with O). That is, a whole set of cycles that pass through that U-shape, and all we know is that the set is nonempty. If the graph is connected enough, can we rig it so that there are neighbours in the set that don't give us grief?

Suppose we pick which edge ψ of O to delete, this iteration. Now we only need to make sure the neighbour there, A say, is not the same as any of the other neighbour across all the other edges. Say U is the neighbour across the edge ϕ. What we need to know (for one ψ and all other ϕ) is whether there are cycles through ϕ (with the chosen cisness there) that don't insist on passing through ψ. We can pick ψ right from the start, the first edge of O for instance, and vet any proposed ϕ as we choose it.

Choosing ϕ and its neighbours in U is a two-step process. At one stage we choose ϕ rather than its alternative ϕ' in the growing O. Only at the next choice of edge of O (when we choose χ rather than χ', say) is the cisness of ϕ (in O and hence U) determined; then the neighbours of ϕ in U are ϕ' and χ'. So ϕ and ϕ' being adjacent in U is already determined even before we choose either of them for O.



In a 3-edge-connected graph, choosing two edges (such as ϕ and ϕ') for cycles to go through never forces them to go through a third (such as ψ), it can only force the

cycles *not* to go through ψ. Three edges can force a cycle to go through a fourth though (if the four are a cut). Now it is not impossible for $\{\phi, \phi', \psi, \chi\}$ to be a cut and $\{\phi, \phi', \psi, \chi'\}$ to be a cut too, but only if $\{\chi, \chi'\}$ is a cut, and we can avoid that in a 3-edge-connected graph.

Note we only need to choose between χ and χ', there's always one we can choose that's alright, so the choices can be made incrementally. The upshot is yes, we can construct O such that at every edge there are neighbours that don't pass through ψ and so don't coincide with any A we might choose there. ■

This isn't yet a proof of the cycle double cover conjecture because there is still the issue that the graph may become 2-edge-connected after an iteration. It would seem that 2-edge-connected graphs are easy, just treat them component-wise, but what about the cisness we already imposed on edges? Loose ends need to be tied up. Unfortunately, the time for writing up this thesis has come to an end. Though not my interest in and attention to the matter.

Appendix **B** has some lists of cycle maps on a few small trivalent graphs. There always appear to be some orientable ones too. The section on h-gonal prisms has the observation that the numbers of maps for h bears no relation to those for $h-1$ but does closely follow that for $h-2$, and why maps for $h-2$ (for these particular graphs) can be extended to h but not to $h-1$. The fact that we *can* go from $h-2$ to h may be peculiar to these prisms but the fact that we *can't* go from $h-1$ to h, even only for these graphs, already means it isn't true we *can* for all graphs. Should we look for induction arguments that jump by 4 nodes and 6 edges (and 2 cycles) at a time? On the other hand, the fact that induction on increasing h by 1 at a time cannot work for *enumeration* of maps might not apply to *existence* arguments.

Index

References

[Tai80] P. G. Tait, "Remarks on the Colouring of Maps", *Proc. Roy. Soc. Edinburgh* **10** (1880), p 729

[Tai80'] P. G. Tait, "Note on a Theorem in Geometry of Position", *Trans. Roy. Soc. Edinburgh* **29** (1880), pp 657–60

[Hea98] P. J. Heawood, "On the Four-Colour Map Theorem", *Quart. J. Math* **29** (1898), pp 270–85

[Tut66] W. T. Tutte, *Connectivity in Graphs*, Mathematical Expositions **15**, Univ. of Toronto Pr. 1966

[Ore67] Oystein Ore, *The Four-Color Problem*, Acad. Pr. 1967 ISBN 0 12 528150 1

[HP73] Frank Harary and Edgar M. Palmer, *Graphical Enumeration*, Acad. Pr. 1973, ISBN 0 123 24245 2

[Sze72] G. Szekeres, "Oriented Tait graphs", *J. Austral. Math. Soc.* **16** (1973) pp 328–31

[Sze73] G. Szekeres, "Polyhedral decompositions of cubic graphs", *Bull. Austral. Math. Soc.* **8** (1973) pp 367–87

[Sze74] G. Szekeres, "Polyhedral decomposition of trivalent graphs", pp 125–7 in *Combinatorial Mathematics* (ed. D. A. Holton), Lecture Notes in Math. **403**, Springer 1974, ISBN 3 540 0 387 06903 8

[Sze75] G. Szekeres, "Non-colourable trivalent graphs", pp 227–33 in *Combinatorial Mathematics III* (ed. Anne Penfold Street, W. D. Wallis), Lecture Notes in Math. **452**, Springer 1975, ISBN 3 540 0 387 07154 7

[AH76] K. I. Appel & W. Haken, "Every Planar Map is Four Colorable", *Bull. Amer. Math. Soc.* **82** (1976) pp 711–2

[AHK77] K. I. Appel, W. Haken, J. Koch "Every Planar Map is Four Colorable", *Illinois J. of Math.* **21** (1976), Part I: "Discharging" pp 429–90, Part II: "Reducibility" pp 491–567

[FW77] S. Fiorini & R. J. Wilson, *Edge-colourings of graphs*, Pitman 1977, ISBN 0 273 01129 4

[MS77] F. J. MacWilliams & N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland 1977, repr. 1983, ISBN 0 444 850 09 0 10 4

[SK77] Thomas L. Saaty & Paul C. Kainen, *The Four-Color Problem: assaults and conquest*, McGraw-Hill 1977; repr. Dover 1986, ISBN 0 486 65092 8

[RW78] Ronald C. Read & Robin J. Wilson, *An Atlas of Graphs*, Oxford Univ. Pr. 1978, ISBN 0 19 853289 X

[BM79] J. A. Bondy & U. S. R. Murty, eds, *Graph Theory and Related Topics* (proceedings of the 1977 conference in honor of Tutte's 60th birthday), Acad. Pr. 1979, ISBN 0 12 114350 3

[Sey79] P. D. Seymour, "Sums of Circuits", pp 341–55 in the above

[HW79] G. H. Hardy & W. M. Wright, *An Introduction to the Theory of Numbers*, Oxford Univ. Press 1938, 5th ed. 1979, ISBN 0 19 85317 $^{0\,2}_{1\,0}$

[CFP81] H. S. M. Coxeter, Roberto Frucht & David L. Powers, *Zero-Symmetric Graphs, Trivalent Graphical Regular Representations of Groups*, Acad. Pr. 1981, ISBN 0 12 194580 4

[RW83] R. W. Robinson & N. C. Wormald, "Numbers of Cubic Graphs", *J. Graph Theory* **7** (1983) pp 463–7

[Tut84] W. T. Tutte, *Graph Theory*, Encyclopedia of Mathematics **21**, Addison-Wesley 1984, ISBN 0 201 13520 5

[Gib85] Alan Gibbons, *Algorithmic Graph Theory*, Camb. Univ. Pr. 1985, ISBN 0 521 $^{24659\,8}_{28881\,9}$

[AK93] E. F. Assmus & J. D. Key, *Designs and their Codes* (pbk. ed. w. corr.), Camb. Univ. Pr. 1993, ISBN 0 521 45839 0

[Cod93] P. D. Coddington, "Analysis of Random Number Generators using Monte Carlo Simulation" (Sep 1993) http://www.arXiv.org/abs/cond-mat/9309017

[Cam94] Peter J. Cameron, *Combinatorics: topics, techniques, algorithms* Camb. Univ. Pr. 1994, ISBN 0 521 45761 0, http://www.maths.qmul.ac.uk/~pjc/comb/ (solutions, errata &c.)

[FF94] Rudolf Fritsch & Gerda Fritsch, *Der Vierfarbensatz*, Brockhaus 1994; *The Four-Color Theorem*, tra. Julie Peschke, Springer 1998, ISBN 0-387-98497-6

[MCl94] John McCleary, *Geometry from a differentiable viewpoint*, Camb. Univ. Pr. 1994, ISBN 0 521 42480 1

[CG96] John H. Conway & Richard K. Guy, *The Book of Numbers*, Springer 1996, ISBN 0 387 97993 X

[Knu98] Donald E. Knuth, *The Art of Computer Programming* Vol 2 *Seminumerical Algorithms*, 3rd ed. Addison Wesley 1998, ISBN 0 201 89684 2

[Wil02] Robert A. Wilson, *Graphs, Colourings and the Four-colour Theorem*, Oxford Univ. Pr. 2002, ISBN 0 19 851062 4 (pbk), http://www.maths.qmul.ac.uk/~raw/graph.html

[Die05] Reinhard Diestel, *Graph Theory*, Graduate Texts in Math. **173**, Springer 1997, 2000, 3rd ed. 2005, ISBN $\begin{smallmatrix}3\,540\\0\,387\end{smallmatrix}$ 26182 6 http://www.math.uni-hamburg.de/home/diestel/books/graph.theory/download.html (full text as pdf, errata &c.)

[LB05] Rodrigo S. C. Leão & Valmir C. Barbosa, "6-Cycle Double Covers of Cubic Graphs" (preprint Jun 2005) http://www.arXiv.org/abs/cs.DM/0505088

[Bae∞] John Baez, *This Week's Finds* (ongoing publication), http://math.ucr.edu/home/baez

[PM∞] Aaron Krowne & al., *PlanetMath* online math encyclopædia (ongoing publication), http://www.planetmath.org/

[Slo∞] N. J. A. Sloane, *On-Line Encyclopedia of Integer Sequences* (ongoing publication), http://www.research.att.com/projects/OEIS

[Wei∞] Eric Weisstein's *World of Mathematics* (ongoing publication), http://mathworld.wolfram.com/

Appendix B: cycle maps of some trivalent graphs

As before, h denotes the integer that is half the number of nodes, and one third of the number of edges, of a trivalent (pseudo- or multi)graph.

There is one trivalent simple graph of $h = 2$ (tetrahedron, p 28); two of $h = 3$ (the prism on p 6, utilities on p 2); six of $h = 4$ of which five connected (p 3) and one disconnected (two tetrahedra); 21 of $h = 5$ of which 19 connected; they proliferate more than exponentially [RW83], see integer sequence `A002851` [Slo∞] for the number of connected trivalent simple graphs.

The cycle maps of all connected simple graphs of $h \leq 4$, and various larger graphs, are listed here. Some tables spill over several pages. Throughout,

- the `g` column holds the genus; suffix $_\circ$ indicates orientable.
- the $|M|$ column holds the number of cycles in a map; $2\mathtt{g} + |M| = h + 2$.
- cycle sizes gives the sizes (in increasing order) of cycles that occur in a map.
- the $\times$ column lists the number of times a cycle map with that cycle pattern occurs. In many cases this will be due to isomorphic maps (many graphs here have large automorphism groups); no attempt was made to identify non-isomorphic maps sharing the same pattern of cycle sizes.

Rows are ordered by genus; within each genus by increasing size of largest cycle, then next largest cycle, and so on.

- the bottom row gives the total number of cycle maps for the graph, and how many of them are orientable.

These tables were produced as summary log files by my map finding program `map`, see `http://web.mat.bham.ac.uk/marijke/g3/map.html` for C source code and a downloadable version for DOS. Maximum capacity is 20 nodes, 30 edges. The latest version, used here, can identify which cycle maps are orientable.

A newer program `mapp` is being developed from scratch at the time of writing. It uses a custom `vec2` class enabling graphs of near unlimited size, and the running time may go up with graph size as $2^{m/2}$ ($\approx$ the number of maps) as opposed to 2^m for `map` because it uses some of the recent ideas discussed in this thesis.

The **purpose of writing such programs** is twofold. Firstly, there is no better test whether ideas about e.g. graph structure are sound than actually coding it up, writing a program that can run on its own on arbitrary valid data and therefore needs to know what is the right thing to do in all possible situations. Any combination of possibilities you hadn't thought of is mercilessly revealed at this stage. Secondly, it produces output like that on the following pages. It can never do "all" graphs, but the human perusing the listings may spot some pattern or trend. It's always a good idea to have a few examples of the thing you're studying.

B.0 Null graph, $h = 0$

The number of valid maps of the null graph is one: the empty map.

Null ($h = 0$, *not a connected graph*)

g	$\|M\|$	cycle sizes	×
$0_\circ$	0		1
maps			$1_\circ = 1$

B.1 Trivalent multigraph of $h = 1$

The "nitrogen molecule" multigraph N_2 consists of 2 nodes joined by a triple edge. It is the only trivalent multigraph of $h = 1$; there are no trivalent simple graphs of this size. There is one pseudograph of $h = 1$ listed among the prisms (p 6).

Nitrogen N_2 ($h = 1$, *bipartite, multigraph*)

g	$\|M\|$	cycle sizes	×
$0_\circ$	3	(2) (2) (2)	1
maps			$1_\circ = 1$

B.2 Trivalent graphs of $h = 2$

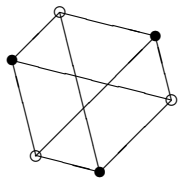
The tetrahedron (p 28) is the only trivalent simple graph of $h = 2$. There is a trivalent multigraph of $h = 2$ listed among the prisms (p 6).

B.3 Trivalent graphs of $h = 3$

The utilities graph has 6 nodes and 9 edges traditionally representing three houses A, B, and C, each connected by an edge to each of E[lectricity], G[as], and W[ater]. The only other simple graph of $h = 3$ is the (trigonal) prism (p 6).

Utilities graph, Thomsen's graph, $K_{3,3}$ ($h = 3$, *bipartite*)

g	$\|M\|$	cycle sizes	×
$\frac{1}{2}$	4	(4) (4) (4) (6)	6
$1_\circ$	3	(6) (6) (6)	2
maps			$2_\circ \subset 8$

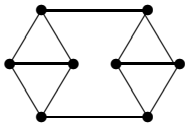


B.4 Trivalent graphs of $h = 4$

The cube is listed already with the prisms (p 6) and the regular solids (p 28).

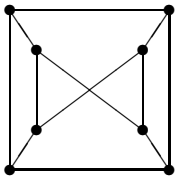
Di-diamond ($h = 4$, *2-edge-connected*)

g	$\lvert M\rvert$	cycle sizes	×
$0_\circ$	6	(3) (3) (3) (3) (6) (6)	2
$\frac{1}{2}$	5	(3) (3) (4) (7) (7)	4
1	4	(4) (4) (8) (8)	2
maps			$2_\circ \subset 8$



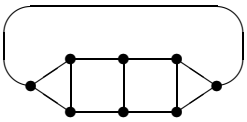
Twisted cube ($h = 4$)

g	$\lvert M\rvert$	cycle sizes	×
$\frac{1}{2}$	5	(4) (4) (5) (5) (6)	4
		(4) (4) (4) (4) (8)	1
1	4	(6) (6) (6) (6)	1
		(5) (5) (7) (7)	8
		(4) (4) (8) (8)	2
$1_\circ$	4	(5) (5) (7) (7)	4
		(4) (6) (6) (8)	4
maps			$8_\circ \subset 24$



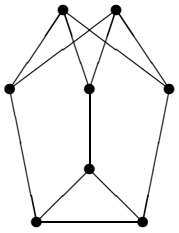
Bracelet ($h = 4$)

g	$\lvert M\rvert$	cycle sizes	×
$0_\circ$	6	(3) (3) (4) (4) (5) (5)	1
$\frac{1}{2}$	5	(3) (4) (5) (6) (6)	2
		(3) (3) (6) (6) (6)	1
1	4	(5) (5) (7) (7)	1
		(3) (7) (7) (7)	2
$1\frac{1}{2}$	3	(8) (8) (8)	1
maps			$1_\circ \subset 8$



Mitre ($h = 4$)

g	$\lvert M\rvert$	cycle sizes	×
$\frac{1}{2}$	5	(3) (4) (5) (5) (7)	6
1	4	(4) (6) (6) (8)	6
$1_\circ$	4	(3) (7) (7) (7)	2
$1\frac{1}{2}$	3	(8) (8) (8)	2
maps			$2_\circ \subset 16$

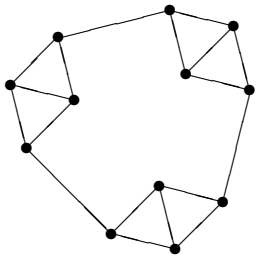


B.5 Polydiamonds

The di-diamond (p 3), tri-diamond, tetra-diamond... consist of two, three, four... $-<|>-$ "diamonds" arranged cyclically, giving them edge-connectivity 2. The mono-diamond is the 3-edge-connected tetrahedron (p 28). Maps for these graphs are especially easy to enumerate: $\frac{1}{2}2^h$ cycle maps of which $\frac{1}{2}2^{h/2}$ orientable.

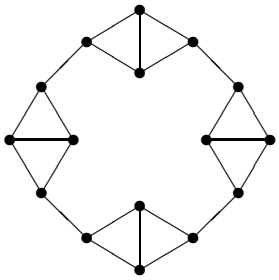
Tri-diamond ($h = 6$, *2-edge-connected*)

g	$\|M\|$	cycle sizes	×
$0_\circ$	8	(3) (3) (3) (3) (3) (3) (9) (9)	4
$\frac{1}{2}$	7	(3) (3) (3) (3) (4) (10) (10)	12
1	6	(3) (3) (4) (4) (11) (11)	12
$1\frac{1}{2}$	5	(4) (4) (4) (12) (12)	4
maps			$4_\circ \subset 32$



Tetra-diamond ($h = 8$, *2-edge-connected*)

g	$\|M\|$	cycle sizes	×
$0_\circ$	10	(3) (3) (3) (3) (3) (3) (3) (3) (12) (12)	8
$\frac{1}{2}$	9	(3) (3) (3) (3) (3) (3) (4) (13) (13)	32
1	8	(3) (3) (3) (3) (4) (4) (14) (14)	48
$1\frac{1}{2}$	7	(3) (3) (4) (4) (4) (15) (15)	32
2	6	(4) (4) (4) (4) (16) (16)	8
maps			$8_\circ \subset 128$

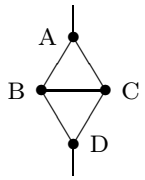


Penta-diamond ($h = 10$, *2-edge-connected*)

g	$\|M\|$	cycle sizes	×
$0_\circ$	12	(3) (3) (3) (3) (3) (3) (3) (3) (3) (3) (15) (15)	16
$\frac{1}{2}$	11	(3) (3) (3) (3) (3) (3) (3) (3) (4) (16) (16)	80
1	10	(3) (3) (3) (3) (3) (3) (4) (4) (17) (17)	160
$1\frac{1}{2}$	9	(3) (3) (3) (3) (4) (4) (4) (18) (18)	160
2	8	(3) (3) (4) (4) (4) (4) (19) (19)	80
$2\frac{1}{2}$	7	(4) (4) (4) (4) (4) (20) (20)	16
maps			$16_\circ \subset 512$

• • •

Let G^0 be a graph, e one of its edges, $\mathcal{M}$ any cycle map on G^0, with X and Y its two cycles through e. Now replace e by a diamond ABCD to form G^1 for which h is two higher than before.



To extend $\mathcal{M}$ to a map on G^1 we can route either of X and Y via either of ABD and ACD, and add the two triangles to form two maps with two more cycles than $\mathcal{M}$, and so the same genus as $\mathcal{M}$. Or we can route either of X and Y via either of ABCD and ACBD, and add the quadangle to form two maps with only one more cycle than $\mathcal{M}$, and hence a genus $\frac{1}{2}$ higher than that of $\mathcal{M}$.

Conversely every map on G^1 derives uniquely from a map on G^0 in this way, so if G^0 had μ cycle maps G^1 has 4μ such maps, half of which with the same genus as the corresponding map of G^0 and half of them with genus one half higher. Moreover, the twist introduced in the second two choices makes it impossible to orient the quadrangle the same way as X, so if G^0 had $\mu_\circ$ orientable cycle maps G^1 has $2\mu_\circ$, each with the same genus as the corresponding map of G^0.

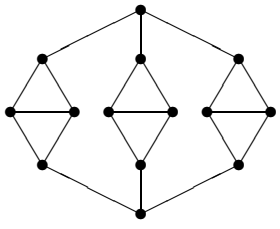
Doing the same thing with d edges to form G^d gives likewise $4^d\mu$ cycle maps, $2^d\mu_\circ$ of them orientable. Starting with the tetrahedron as our G^0, with $\mu = 2$ (one each of genus 0 and $\frac{1}{2}$) and $\mu_\circ = 1$, we get $4^d \cdot 2$ and $2^d \cdot 1$ for the $(d+1)$-diamond as G^d.

Even better, let the tetrahedron be the G^1 member of the series, the d-diamond the G^d member. Purely formally set $\mu = \mu_\circ = \frac{1}{2}$ by hand for the now nonexistent G^0, and the genus of that "half of a map" to 0, just to make the number of maps and their genera 0 and $\frac{1}{2}$ come out right for G^1. There is even a kind-of-justification in terms of building the tetrahedron by inserting a diamond in a nodeless cycle (the factor 2 is lost because we don't start off with distinguishible X and Y). The numbers of maps and the **binomial distribution** of genera for d-diamonds follow.

König's graph ($h = 7$)

König's counterexample (1936) to a conjecture by Tait (it's trivalent and planar but has no Hamiltonian cycle). Not a polydiamond, but similar considerations make it easy to predict the number of maps here too.

g	$\|M\|$	cycle sizes	×
$0_\circ$	9	(3) (3) (3) (3) (3) (3) (8) (8) (8)	8
$\frac{1}{2}$	8	(3) (3) (3) (3) (4) (8) (9) (9)	24
1	7	(3) (3) (4) (4) (9) (9) (10)	24
$1\frac{1}{2}$	6	(4) (4) (4) (10) (10) (10)	8
maps			$8_\circ \subset 64$

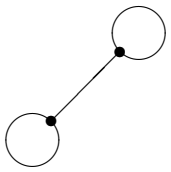


B.6 Prisms

The h-gonal prism consists of two h-gon cycles ("top" and "bottom" face) with edges joining successive nodes in the one face to successive nodes in the other face. The ordinary prism is the $h = 3$ case, the cube the $h = 4$ case.

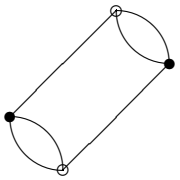
Note how for even h there are only maps with integer genus, from 0 to $(n-2)/2$; when h is odd no halfodd genera occur except $(n-2)/2$.

Monogonal Prism ($h = 1$, *pseudograph, 1-edge-connected*)



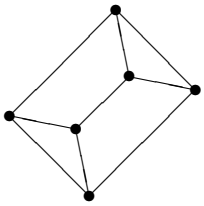
g	$\|M\|$	cycle sizes	×
maps			0

Digonal Prism ($h = 2$, *bipartite, multi-, 2-edge-connected*)



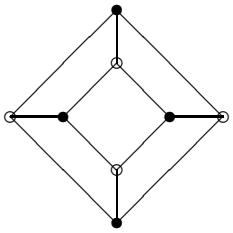
g	$\|M\|$	cycle sizes	×
$0_\circ$	4	(2) (2) (4) (4)	2
maps			$2_\circ = 2$

[Trigonal] Prism ($h = 3$)



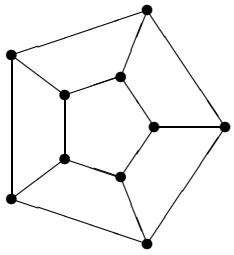
g	$\|M\|$	cycle sizes	×
$0_\circ$	5	(3) (3) (4) (4) (4)	1
$\frac{1}{2}$	4	(3) (5) (5) (5)	2
1	3	(6) (6) (6)	1
maps			$1_\circ \subset 4$

Cube *aka* **Square Prism** ($h = 4$, *bipartite*)



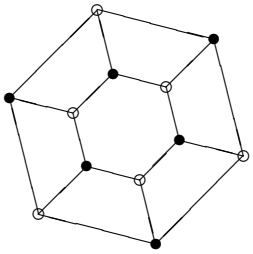
g	$\|M\|$	cycle sizes	×
$0_\circ$	6	(4) (4) (4) (4) (4) (4)	1
1	4	(6) (6) (6) (6)	6
		(4) (6) (6) (8)	12
$1_\circ$	4	(6) (6) (6) (6)	4
		(4) (4) (8) (8)	3
maps			$8_\circ \subset 26$

Pentagonal Prism ($h = 5$)



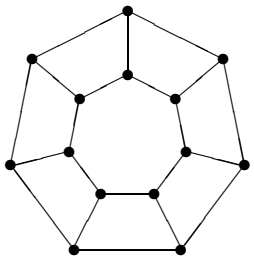
g	$\lvert M\rvert$	cycle sizes	×
$0_\circ$	7	(4) (4) (4) (4) (4) (5) (5)	1
1	5	(6) (6) (6) (6) (6)	1
		(4) (6) (6) (7) (7)	5
		(4) (4) (7) (7) (8)	5
		(4) (4) (6) (8) (8)	5
		(4) (4) (6) (6) (10)	5
$1_\circ$	5	(4) (6) (6) (6) (8)	5
		(4) (4) (4) (8) (10)	5
$1\frac{1}{2}$	4	(7) (7) (7) (9)	30
		(5) (7) (9) (9)	10
maps			$11_\circ \subset 72$

Hexagonal Prism ($h = 6$, *bipartite*)



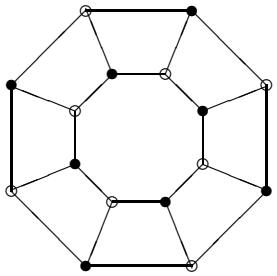
g	$\lvert M\rvert$	cycle sizes	×
$0_\circ$	8	(4) (4) (4) (4) (4) (4) (6) (6)	1
1	6	(4) (6) (6) (6) (6) (8)	6
		(4) (4) (6) (6) (8) (8)	9
		(4) (4) (4) (8) (8) (8)	8
		(4) (4) (4) (6) (8) (10)	12
		(4) (4) (4) (6) (6) (12)	6
$1_\circ$	6	(6) (6) (6) (6) (6) (6)	1
		(4) (4) (6) (6) (8) (8)	9
		(4) (4) (6) (6) (6) (10)	6
		(4) (4) (4) (4) (10) (10)	3
		(4) (4) (4) (4) (8) (12)	6
2	4	(8) (8) (10) (10)	135
		(6) (10) (10) (10)	14
		(8) (8) (8) (12)	8
		(6) (8) (10) (12)	12
		(4) (10) (10) (12)	6
$2_\circ$	4	(8) (8) (10) (10)	15
		(6) (6) (12) (12)	1
maps			$42_\circ \subset 258$

Heptagonal Prism ($h = 7$)



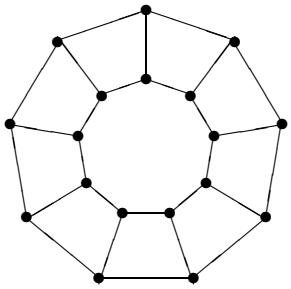
g	$\lvert M\rvert$	cycle sizes	$\times$
$0_\circ$	9	(4) (4) (4) (4) (4) (4) (4) (7) (7)	1
1	7	(6) (6) (6) (6) (6) (6) (6)	1
		(4) (4) (6) (6) (6) (8) (8)	14
		(4) (4) (4) (6) (6) (9) (9)	14
		(4) (4) (4) (4) (8) (9) (9)	7
		(4) (4) (6) (6) (6) (6) (10)	7
		(4) (4) (4) (4) (8) (8) (10)	7
		(4) (4) (4) (4) (6) (10) (10)	7
		(4) (4) (4) (4) (6) (8) (12)	14
		(4) (4) (4) (4) (6) (6) (14)	7
$1_\circ$	7	(4) (6) (6) (6) (6) (6) (8)	7
		(4) (4) (4) (6) (8) (8) (8)	7
		(4) (4) (4) (6) (6) (8) (10)	21
		(4) (4) (4) (6) (6) (6) (12)	7
		(4) (4) (4) (4) (4) (10) (12)	7
		(4) (4) (4) (4) (4) (8) (14)	7
2	5	(6) (6) (8) (11) (11)	7
		(4) (8) (8) (11) (11)	7
		(4) (6) (10) (11) (11)	14
		(4) (4) (11) (11) (12)	7
$2\frac{1}{2}$	4	(9) (11) (11) (11)	392
		(9) (9) (11) (13)	238
		(7) (11) (11) (13)	28
		(7) (9) (13) (13)	14
maps			$57_\circ \subset 842$

Octagonal Prism ($h = 8$, *bipartite*)



g	$\lvert M\rvert$	cycle sizes	×
$0_\circ$	10	(4) (4) (4) (4) (4) (4) (4) (4) (8) (8)	1
1	8	(4) (6) (6) (6) (6) (6) (6) (8)	8
		(4) (4) (4) (6) (6) (8) (8) (8)	16
		(4) (4) (4) (6) (6) (6) (8) (10)	32
		(4) (4) (4) (4) (6) (6) (10) (10)	20
		(4) (4) (4) (4) (4) (8) (10) (10)	16
		(4) (4) (4) (6) (6) (6) (6) (12)	8
		(4) (4) (4) (4) (4) (8) (8) (12)	8
		(4) (4) (4) (4) (4) (6) (10) (12)	16
		(4) (4) (4) (4) (4) (6) (8) (14)	16
		(4) (4) (4) (4) (4) (6) (6) (16)	8
$1_\circ$	8	(6) (6) (6) (6) (6) (6) (6) (6)	1
		(4) (4) (6) (6) (6) (6) (8) (8)	20
		(4) (4) (4) (4) (8) (8) (8) (8)	2
		(4) (4) (6) (6) (6) (6) (6) (10)	8
		(4) (4) (4) (4) (6) (8) (8) (10)	24
		(4) (4) (4) (4) (6) (6) (10) (10)	12
		(4) (4) (4) (4) (6) (6) (8) (12)	24
		(4) (4) (4) (4) (4) (4) (12) (12)	4
		(4) (4) (4) (4) (6) (6) (6) (14)	8
		(4) (4) (4) (4) (4) (4) (10) (14)	8
		(4) (4) (4) (4) (4) (4) (8) (16)	8
2	6	(6) (6) (6) (6) (12) (12)	2
		(4) (6) (6) (8) (12) (12)	24
		(4) (4) (8) (8) (12) (12)	12
		(4) (4) (6) (10) (12) (12)	24
		(4) (4) (4) (12) (12) (12)	8
3	4	(12) (12) (12) (12)	604
		(10) (12) (12) (14)	1 664
		(10) (10) (14) (14)	284
		(8) (12) (14) (14)	64
		(6) (14) (14) (14)	8
		(10) (10) (12) (16)	56
		(8) (12) (12) (16)	12
		(8) (10) (14) (16)	16
		(4) (14) (14) (16)	8
$3_\circ$	4	(12) (12) (12) (12)	35
		(10) (10) (14) (14)	28
		(8) (8) (16) (16)	1
maps			$184_\circ \subset 3\,118$

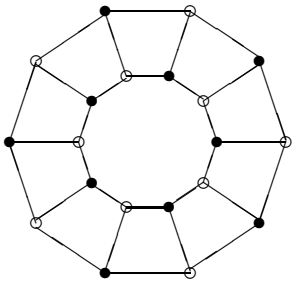
Enneagonal Prism ($h = 9$)



g	$\lvert M\rvert$	cycle sizes	×
$0_\circ$	11	(4) (4) (4) (4) (4) (4) (4) (4) (4) (9) (9)	1
1	9	(6) (6) (6) (6) (6) (6) (6) (6) (6)	1
		(4) (4) (6) (6) (6) (6) (6) (8) (8)	27
		(4) (4) (4) (4) (6) (8) (8) (8) (8)	9
		(4) (4) (6) (6) (6) (6) (6) (6) (10)	9
		(4) (4) (4) (4) (6) (6) (8) (8) (10)	54
		(4) (4) (4) (4) (6) (6) (6) (10) (10)	18
		(4) (4) (4) (4) (4) (4) (10) (10) (10)	3
		(4) (4) (4) (4) (4) (6) (6) (11) (11)	27
		(4) (4) (4) (4) (4) (4) (8) (11) (11)	9
		(4) (4) (4) (4) (6) (6) (6) (8) (12)	36
		(4) (4) (4) (4) (4) (4) (8) (10) (12)	18
		(4) (4) (4) (4) (4) (4) (6) (12) (12)	9
		(4) (4) (4) (4) (6) (6) (6) (6) (14)	9
		(4) (4) (4) (4) (4) (4) (8) (8) (14)	9
		(4) (4) (4) (4) (4) (4) (6) (10) (14)	18
		(4) (4) (4) (4) (4) (4) (6) (8) (16)	18
		(4) (4) (4) (4) (4) (4) (6) (6) (18)	9
$1_\circ$	9	(4) (6) (6) (6) (6) (6) (6) (6) (8)	9
		(4) (4) (4) (6) (6) (6) (8) (8) (8)	30
		(4) (4) (4) (6) (6) (6) (6) (8) (10)	45
		(4) (4) (4) (4) (4) (8) (8) (8) (10)	9
		(4) (4) (4) (4) (4) (6) (8) (10) (10)	27
		(4) (4) (4) (6) (6) (6) (6) (6) (12)	9
		(4) (4) (4) (4) (4) (6) (8) (8) (12)	27
		(4) (4) (4) (4) (4) (6) (6) (10) (12)	27
		(4) (4) (4) (4) (4) (6) (6) (8) (14)	27
		(4) (4) (4) (4) (4) (4) (4) (12) (14)	9
		(4) (4) (4) (4) (4) (6) (6) (6) (16)	9
		(4) (4) (4) (4) (4) (4) (4) (10) (16)	9
		(4) (4) (4) (4) (4) (4) (4) (8) (18)	9
2	7	(4) (6) (6) (6) (6) (13) (13)	9
		(4) (4) (6) (6) (8) (13) (13)	54
		(4) (4) (4) (8) (8) (13) (13)	18
		(4) (4) (4) (6) (10) (13) (13)	36
		(4) (4) (4) (4) (12) (13) (13)	9

3	5	(8) (8) (8) (15) (15)	3
		(6) (8) (10) (15) (15)	18
		(4) (10) (10) (15) (15)	9
		(6) (6) (12) (15) (15)	9
		(4) (8) (12) (15) (15)	18
		(4) (6) (14) (15) (15)	18
		(4) (4) (15) (15) (16)	9
$3\frac{1}{2}$	4	(13) (13) (13) (15)	5 058
		(11) (13) (15) (15)	4 266
		(9) (15) (15) (15)	44
		(11) (13) (13) (17)	936
		(11) (11) (15) (17)	450
		(9) (13) (15) (17)	108
		(9) (11) (17) (17)	18
maps			$247_\circ \subset 11\,620$

Decagonal Prism ($h = 10$, *bipartite*)



g	$\lvert M\rvert$	cycle sizes	×
$0_\circ$	12	(4) (4) (4) (4) (4) (4) (4) (4) (4) (4) (10) (10)	1
1	10	(4) (6) (6) (6) (6) (6) (6) (6) (6) (8)	10
		(4) (4) (4) (6) (6) (6) (6) (8) (8) (8)	50
		(4) (4) (4) (4) (4) (8) (8) (8) (8) (8)	2
		(4) (4) (4) (6) (6) (6) (6) (6) (8) (10)	60
		(4) (4) (4) (4) (4) (6) (8) (8) (8) (10)	40
		(4) (4) (4) (4) (4) (6) (6) (8) (10) (10)	60
		(4) (4) (4) (6) (6) (6) (6) (6) (6) (12)	10
		(4) (4) (4) (4) (4) (6) (6) (8) (8) (12)	60
		(4) (4) (4) (4) (4) (6) (6) (6) (10) (12)	40
		(4) (4) (4) (4) (4) (4) (4) (10) (10) (12)	10
		(4) (4) (4) (4) (4) (4) (6) (6) (12) (12)	35
		(4) (4) (4) (4) (4) (4) (4) (8) (12) (12)	20
		(4) (4) (4) (4) (4) (6) (6) (6) (8) (14)	40
		(4) (4) (4) (4) (4) (4) (4) (8) (10) (14)	20
		(4) (4) (4) (4) (4) (4) (4) (6) (12) (14)	20
		(4) (4) (4) (4) (4) (6) (6) (6) (6) (16)	10
		(4) (4) (4) (4) (4) (4) (4) (8) (8) (16)	10
		(4) (4) (4) (4) (4) (4) (4) (6) (10) (16)	20
		(4) (4) (4) (4) (4) (4) (4) (6) (8) (18)	20
		(4) (4) (4) (4) (4) (4) (4) (6) (6) (20)	10
$1_\circ$	10	(6) (6) (6) (6) (6) (6) (6) (6) (6) (6)	1
		(4) (4) (6) (6) (6) (6) (6) (6) (8) (8)	35
		(4) (4) (4) (4) (6) (6) (8) (8) (8) (8)	25
		(4) (4) (6) (6) (6) (6) (6) (6) (6) (10)	10
		(4) (4) (4) (4) (6) (6) (6) (8) (8) (10)	100
		(4) (4) (4) (4) (6) (6) (6) (6) (10) (10)	25
		(4) (4) (4) (4) (4) (4) (8) (8) (10) (10)	15
		(4) (4) (4) (4) (4) (4) (6) (10) (10) (10)	10
		(4) (4) (4) (4) (6) (6) (6) (6) (8) (12)	50
		(4) (4) (4) (4) (4) (4) (8) (8) (8) (12)	10
		(4) (4) (4) (4) (4) (4) (6) (8) (10) (12)	60
		(4) (4) (4) (4) (4) (4) (6) (6) (12) (12)	15
		(4) (4) (4) (4) (6) (6) (6) (6) (6) (14)	10
		(4) (4) (4) (4) (4) (4) (6) (8) (8) (14)	30
		(4) (4) (4) (4) (4) (4) (6) (6) (10) (14)	30
		(4) (4) (4) (4) (4) (4) (4) (4) (14) (14)	5
		(4) (4) (4) (4) (4) (4) (6) (6) (8) (16)	30
		(4) (4) (4) (4) (4) (4) (4) (4) (12) (16)	10

$1_\circ$	10	(4) (4) (4) (4) (4) (4) (6) (6) (6) (18)	10
		(4) (4) (4) (4) (4) (4) (4) (4) (10) (18)	10
		(4) (4) (4) (4) (4) (4) (4) (4) (8) (20)	10
2	8	(4) (4) (6) (6) (6) (6) (14) (14)	25
		(4) (4) (4) (6) (6) (8) (14) (14)	100
		(4) (4) (4) (4) (8) (8) (14) (14)	25
		(4) (4) (4) (4) (6) (10) (14) (14)	50
		(4) (4) (4) (4) (4) (12) (14) (14)	10
3	6	(6) (6) (8) (8) (16) (16)	15
		(4) (8) (8) (8) (16) (16)	10
		(6) (6) (6) (10) (16) (16)	10
		(4) (6) (8) (10) (16) (16)	60
		(4) (4) (10) (10) (16) (16)	15
		(4) (6) (6) (12) (16) (16)	30
		(4) (4) (8) (12) (16) (16)	30
		(4) (4) (6) (14) (16) (16)	30
		(4) (4) (4) (16) (16) (16)	10
4	4	(14) (14) (16) (16)	24 100
		(12) (16) (16) (16)	4 380
		(14) (14) (14) (18)	4 380
		(12) (14) (16) (18)	9 560
		(10) (16) (16) (18)	180
		(12) (12) (18) (18)	485
		(10) (14) (18) (18)	110
		(8) (16) (18) (18)	10
		(6) (18) (18) (18)	10
		(12) (14) (14) (20)	180
		(12) (12) (16) (20)	100
		(10) (14) (16) (20)	40
		(10) (12) (18) (20)	20
		(4) (18) (18) (20)	10
$4_\circ$	4	(14) (14) (16) (16)	210
		(12) (12) (18) (18)	45
		(10) (10) (20) (20)	1
maps			$758_\circ \subset 45\,290$

Orientable maps on prisms

The orientable maps on prisms form (like the maps on polydiamonds earlier) a collection simple and regular enough to enumerate completely. If $M_\circ(h)$ represents the total number of orientable maps on the h-gonal prism, we find

$$M_\circ(h+2) \;=\; 4M_\circ(h) + (3h-2)$$

both for even and odd $h > 2$ (for $h = 2$ the multiplier is 2 rather than 4). This suggests that, in general, by inserting an extra  element each orientable map of the h-gonal prism extends to four orientable maps on the $(h+2)$-gonal prism. Any such induction cannot proceed by merely inserting  (increasing h by one) at a time, because the numbers of maps are quite different for even and odd h.

Splitting out the numbers of orientable maps over genus

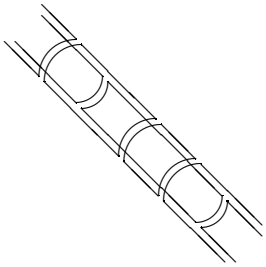
h	$M_\circ(h)$	$\mathsf{g} = 0_\circ$	$1_\circ$
1			
3	1	1	
5	11	1	10
7	57	1	56
9	247	1	246

h	$M_\circ(h)$	$\mathsf{g} = 0_\circ$	$1_\circ$	$2_\circ$	$3_\circ$	$4_\circ$
2	2	$1+1$				
4	8	1	$3+4$			
6	42	1	25	16		
8	184	1	119		64	
10	758	1	501			256

sheds some more light. We see there is each time a planar (genus 0) map, and that the behaviour of approximate quadrupling is entirely due to the occurrence of $2^{h-1} - (h+1)$ maps of orientable genus 1. When h is even, there are additionally 2^{h-2} maps of orientable genus $\frac{1}{2}(h-2)$, simply quadrupling at each step.

The **genus 0 map** is the obvious polyhedron one with top, bottom and h side faces as its cycles. When $h = 2$ the top and bottom digon can be swiveled independently giving two maps on an equal footing.

The **oriented genus 1 maps** are toroidal and best understood as such, with the two h-gons surrounding the hole rather than having surface area spanned across them. The circular ladder formed by the edges must now be embedded on a circular tube of surface, and we can visualise some of the rungs of the ladder as being draped along the "front" of the tube facing us (making a $2i+2$-cycle if the next "front" rung is i rungs on) and the other rungs around the "back" of the surface (making cycles there).



(extend tube to cover h rungs and glue the ends together)

There being h rungs means 2^h choices which are two-by-two the same as only the distinction between "front" and "back" matters, not which is which; moreover there must be at least 2 out of h rungs running along the front and at least 2 along the back to form cycles (1 would mean a circuit sharing an edge with itself, and 0 two more separate cycles giving the planar map). So the actual number isn't $\frac{1}{2} \cdot 2^h$ but

$$\tfrac{1}{2}\big[\tbinom{h}{2} + \tbinom{h}{3} + \cdots + \tbinom{h}{h-3} + \tbinom{h}{h-2}\big] \;=\; 2^{h-1} - (h+1)$$

which is what we saw.

The extra **oriented genus $\frac{1}{2}(h-2)$ maps** occurring for **even** h are best understood by noting they all consist of four cycles. Split out over cycle size pattern there are

$\binom{h}{0} = 1$ map of pattern (h) (h) $(2h)$ $(2h)$

$\binom{h}{2}$ ones of pattern $(h+2)$ $(h+2)$ $(2h-2)$ $(2h-2)$

$\binom{h}{4}$ ones of pattern $(h+4)$ $(h+4)$ $(2h-4)$ $(2h-4)$ ... and so on.

The sequence ends with $\binom{h}{h/2-1}$ when $h \equiv 2 \pmod 4$ but with $\frac{1}{2}\binom{h}{h/2}$ when $h \equiv 0 \pmod 4$ to prevent double counting ($h + h/2$ and $2h - h/2$ are indistinguishable cycle sizes). Note the total is indeed 2^{h-2} as each $\binom{h}{2i}$ is the sum of two items in the row above it in the Pascal triangle, so we have a contiguous half of that row.

The first map has the top and bottom faces as (h) cycles; with the other two cycles complete $(2h)$ cogwheels, cogs of each overlapping those of the other wheel so as to cover all edges twice. In the type that occurs $\binom{h}{2}$ times, the cycles following top and bottom faces are replaced by two that at at some rung swap over, and at some other rung swap back; the cogwheels skip an "in" or "out" at the corresponding place. Likewise the next type has a choice of four places where to swap over, and so on.

Now we know what the maps look like we see how they extend by **induction**. Each orientable genus-1 map for some h generates four such maps for $h+2$ by adding two rungs and making all possible choices. In fact here we *don't* have to go from h to $h+2$, we can simply go to $h+1$ adding one rung (with two choices, "front" or "back") at a time. We still don't get *all* orientable maps that way (e.g. when we add the i-th rung there is a map with only the i-th and any j-th rung "front" and the rest "back"; it is not an extension of a cycle map on a graph without the i-th rung, only of a circuit map); that would not be a problem for an *existence* proof.

The extra maps for even h show that in some cases induction really does need two pairs of nodes (although they might not need to be be contiguous). This isn't just because we restricted ourselves to orientable maps (to keep the numbers down); attempting to grow the cogwheels to $h+1$ along the same lines gives not merely a non-orientable map but a circuit map — and for purely global reasons (two cycles are merged that already shared an edge). **Induction from h to $h+1$ is doomed; induction to $h+2$ might just work** for an existence proof of cycle maps, as part of some proof of the cycle double cover conjecture.

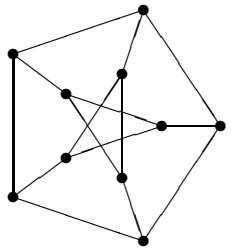
B.7 Petersen-like graphs

The Petersen graph is the $\mathrm{P}(5,2)$ member of a more general pattern $\mathrm{P}(h,d)$ consisting of two h-gons where each time the i-th node of the one h-gon is linked to the di-th node (mod h) of the other h-gon. Of course h and d must be coprime. $\mathrm{P}(h,d)$, $\mathrm{P}(h,h-d)$, $\mathrm{P}(h,p)$, $\mathrm{P}(h,h-p)$ where $dp \equiv 1 \pmod h$ are all the same, and $\mathrm{P}(h,1)$ are the h-gonal prisms of the previous section.

Petersen graph P(5,2), aka (3,5)-cage ($h = 5$, *class II*)

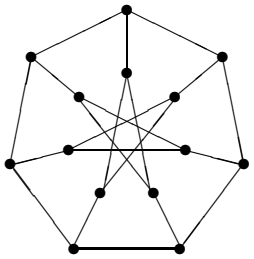
One of many realisations of the Petersen graph: let V be a set of 5 items, now the nodes are pairs $\subset V$, and an edge exists between pairs that are disjoint. See also the double cover notes under Desargues graph (p 25) and dodecahedron (p 28).

g	$\lvert M\rvert$	cycle sizes	×
$\frac{1}{2}$	6	(5) (5) (5) (5) (5) (5)	2
1	5	(5) (5) (6) (6) (8)	30
$1_\circ$	5	(5) (5) (5) (6) (9)	20
maps			$20_\circ \subset 52$



Generalised Petersen P(7,2) ($h = 7$)

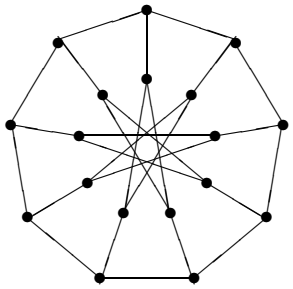
g	$\lvert M\rvert$	cycle sizes	×
$\frac{1}{2}$	8	(5) (5) (5) (5) (5) (5) (5) (7)	1
1	7	(5) (5) (5) (5) (7) (7) (8)	7
$1_\circ$	7	(5) (5) (5) (5) (5) (7) (10)	7
$1\frac{1}{2}$	6	(6) (6) (7) (7) (8) (8)	14
		(5) (7) (7) (7) (8) (8)	14
		(6) (6) (6) (8) (8) (8)	7
		(6) (6) (7) (7) (7) (9)	7
		(6) (6) (6) (7) (8) (9)	21
		(5) (6) (7) (7) (8) (9)	14
		(5) (5) (7) (8) (8) (9)	7
		(5) (6) (7) (7) (7) (10)	7
		(5) (6) (6) (7) (8) (10)	28
		(5) (5) (7) (7) (8) (10)	21
		(5) (5) (6) (8) (8) (10)	7
		(5) (5) (5) (7) (10) (10)	14
		(5) (5) (7) (7) (7) (11)	7
		(5) (5) (6) (7) (8) (11)	28
		(5) (5) (6) (6) (9) (11)	7
		(5) (5) (5) (6) (8) (13)	7



2	5	(7) (8) (9) (9) (9)	7
		(7) (8) (8) (9) (10)	28
		(7) (7) (9) (9) (10)	21
		(6) (8) (9) (9) (10)	14
		(7) (7) (8) (10) (10)	7
		(6) (8) (8) (10) (10)	14
		(6) (7) (9) (10) (10)	7
		(5) (8) (9) (10) (10)	14
		(7) (8) (8) (8) (11)	7
		(6) (8) (8) (9) (11)	14
		(6) (7) (8) (10) (11)	28
		(5) (8) (8) (10) (11)	7
		(5) (7) (9) (10) (11)	14
		(6) (6) (8) (11) (11)	7
		(5) (7) (8) (11) (11)	21
		(5) (6) (9) (11) (11)	28
		(5) (5) (10) (11) (11)	7
		(7) (7) (8) (8) (12)	21
		(6) (6) (9) (9) (12)	7
		(5) (7) (8) (10) (12)	42
		(5) (5) (10) (10) (12)	14
		(6) (6) (7) (11) (12)	14
		(5) (6) (7) (12) (12)	7
		(7) (7) (7) (8) (13)	7
		(6) (7) (7) (9) (13)	7
		(5) (6) (9) (9) (13)	7
		(5) (7) (7) (10) (13)	7
		(5) (5) (9) (10) (13)	7
		(5) (6) (7) (11) (13)	14
		(5) (5) (6) (13) (13)	7
		(6) (6) (8) (8) (14)	7
$2_\circ$	5	(6) (8) (9) (9) (10)	7
		(6) (8) (8) (10) (10)	7
		(7) (8) (8) (8) (11)	7
		(5) (7) (8) (10) (12)	14
$2\frac{1}{2}$	4	(10) (10) (10) (12)	7
		(8) (10) (12) (12)	28
		(6) (12) (12) (12)	14
		(9) (9) (11) (13)	28
		(7) (11) (11) (13)	56
		(7) (9) (13) (13)	7
		(5) (11) (13) (13)	21
		(8) (10) (10) (14)	14
		(6) (10) (12) (14)	14
maps			$42_\circ \subset 862$

Generalised Petersen P(9,2) ($h = 9$)

P$(9, 2)$ is one of those graphs that can be edge-3-colored in essentially only one way [FW77 p 120]. Only some of the 241 map types (rows of the table) are shown here, including all the orientable ones.



g	$\|M\|$	cycle sizes	×
$\frac{1}{2}$	10	(5) (5) (5) (5) (5) (5) (5) (5) (5) (9)	1
1	9	(5) (5) (5) (5) (5) (5) (8) (8) (8)	9
$1_\circ$	9	(5) (5) (5) (5) (5) (5) (5) (8) (11)	9
$1\frac{1}{2}$	8	(5) (5) (5) (5) (7) (8) (8) (11)	36
		(5) (5) (5) (5) (7) (7) (9) (11)	9
		(5) (5) (5) (5) (5) (7) (8) (14)	9
2	7	(not listed)	
$2_\circ$	7	(5) (8) (8) (8) (8) (8) (9)	9
		(5) (5) (7) (7) (9) (10) (11)	18
		(5) (5) (5) (8) (8) (11) (12)	18
		(5) (5) (7) (7) (8) (9) (13)	9
		(5) (5) (5) (5) (9) (11) (14)	9
$2\frac{1}{2}$	6	(not listed)	
3	5	(not listed)	
$3_\circ$	5	(9) (11) (11) (11) (12)	3
		(9) (9) (12) (12) (12)	1
		(10) (10) (10) (11) (13)	9
		(8) (10) (10) (12) (14)	9
		(7) (10) (10) (11) (16)	18
		(8) (8) (8) (12) (18)	3
$3\frac{1}{2}$	4	(9) (15) (15) (15)	10
		(10) (14) (14) (16)	9
		(8) (14) (16) (16)	18
		(11) (13) (13) (17)	9
		(11) (11) (15) (17)	9
		(9) (11) (17) (17)	9
		(5) (15) (17) (17)	9
4	3	(18) (18) (18)	1
maps		$115_\circ \subset 9\,685$	

Generalised Petersen P(10,3) ($h = 10$)

P$(10, 3)$ is the Desargues graph listed on p 25.

B.8 Snarks

Recall that by Vizings theorem, graphs are Class I (can be ρ-edge-colored) or Class II (cannot, but can be $\rho+1$-edge-colored) where ρ is the largest valency occurring. For trivalent graphs $\rho = 3$; an edge-3-coloring of a Class I graph here is known as a **Tait coloring**. A Tait coloring implies a cycle map with even cycles (the connected components of the red-green, blue-red, and green-blue 2-valent subgraphs).

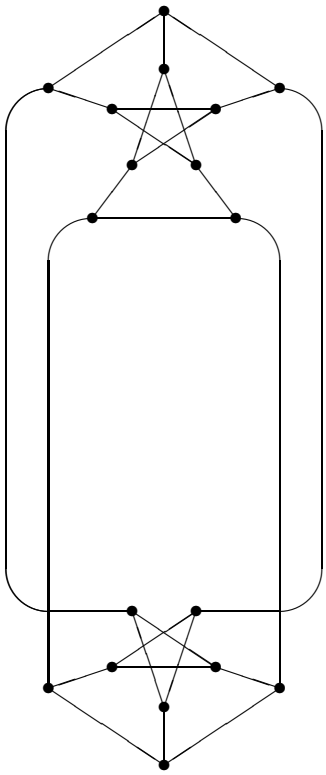
Trivalent graphs are necessarily Class II if they have a bridge, because such a graph has no cycle maps. Bridge-free ones are Class I if they are planar (by the 4-color theorem, and the Tait coloring derived from that face coloring). Most non-planar bridge-free ones are Class I as well. Bridge-free trivalent Class II graphs are rare, and those of girth at least 5 are so hard to find that they were dubbed **snarks** by Martin Gardner. Nowadays the label *snark* is further restricted to cases where removal of three edges only disconnect the graph if those edges join up at a node.

The Petersen graph (p 16) is the smallest snark.

One way (due to Isaacs) to describe the Blanuša snark is stitching it together from two incomplete Petersen graphs. Due to its scant symmetry, its 6389 maps are of no fewer than 398 different cycle size patterns, most of which are omitted here.

Blanuša snark #2 ($h = 9$, *class II*)

g	$\|M\|$	cycle sizes	$\times$
1	9	(5) (5) (6) (6) (6) (6) (6) (7) (7)	1
		(5) (5) (5) (5) (6) (6) (7) (7) (8)	1
		(5) (5) (5) (5) (5) (6) (6) (8) (9)	2
		(5) (5) (5) (5) (5) (5) (6) (9) (9)	1
		(5) (5) (5) (5) (5) (5) (7) (7) (10)	1
$1_\circ$	9	(5) (5) (5) (5) (6) (6) (6) (8) (8)	1
		(5) (5) (5) (5) (5) (6) (7) (7) (9)	2
		(5) (5) (5) (5) (5) (5) (6) (8) (10)	1
$1\frac{1}{2}$	8	(not listed)	

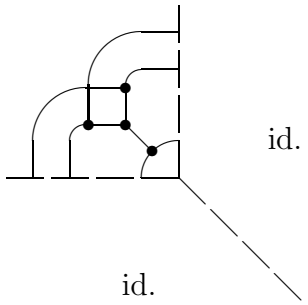


2	7	(not listed)	
$2_\circ$	7	(6) (6) (8) (8) (8) (9) (9)	1
		(5) (5) (8) (9) (9) (9) (9)	1
		(5) (5) (7) (9) (9) (9) (10)	4
		(6) (6) (6) (8) (8) (10) (10)	2
		(5) (6) (6) (8) (9) (10) (10)	4
		(5) (5) (6) (9) (9) (10) (10)	6
		(5) (5) (5) (9) (10) (10) (10)	4
		(6) (6) (7) (8) (8) (8) (11)	4
		(5) (5) (6) (9) (9) (9) (11)	4
		(5) (6) (6) (8) (8) (10) (11)	4
		(5) (5) (5) (9) (9) (10) (11)	12
		(5) (5) (5) (8) (10) (10) (11)	4
		(5) (5) (6) (8) (8) (11) (11)	4
		(5) (5) (5) (8) (9) (11) (11)	4
		(5) (5) (6) (8) (9) (9) (12)	2
		(5) (5) (5) (8) (9) (10) (12)	4
		(5) (5) (6) (6) (10) (10) (12)	2
		(5) (5) (5) (7) (10) (10) (12)	4
		(5) (5) (5) (7) (9) (11) (12)	4
		(6) (6) (6) (7) (8) (8) (13)	2
		(5) (6) (6) (7) (8) (9) (13)	4
		(5) (5) (6) (7) (9) (9) (13)	4
		(5) (5) (6) (7) (8) (10) (13)	4
		(5) (5) (6) (6) (9) (10) (13)	8
		(5) (5) (5) (6) (10) (10) (13)	4
		(5) (5) (6) (6) (7) (12) (13)	2
		(5) (5) (6) (6) (6) (13) (13)	2
		(5) (5) (6) (7) (8) (9) (14)	4
		(5) (5) (6) (6) (9) (9) (14)	5
		(5) (5) (5) (6) (9) (10) (14)	8
		(5) (5) (5) (5) (10) (10) (14)	2
		(5) (5) (5) (6) (8) (11) (14)	4
		(5) (5) (5) (5) (9) (11) (14)	2
		(5) (5) (6) (6) (6) (9) (17)	2
$2\frac{1}{2}$	6	(not listed)	
3	5	(not listed)	
$3_\circ$	5	(6) (10) (12) (12) (14)	8
		(5) (11) (12) (12) (14)	8
		(5) (10) (12) (13) (14)	16
		(6) (10) (10) (12) (16)	8
		(5) (10) (11) (12) (16)	16
		(5) (10) (10) (13) (16)	8
maps			$199_\circ \subset 6\,389$

Isaacs' flower snarks form an infinite family. The first member of the series is formed by replacing one node of the Petersen graph by a triangle. While it is bridgefree and Class II, the girth 3 means it is not really a snark. It consists of 3 repeated four-node units, one of which is drawn below. The other graphs of the series are obtained by extending it to 5, 7, ... four-node units; these are all snarks.

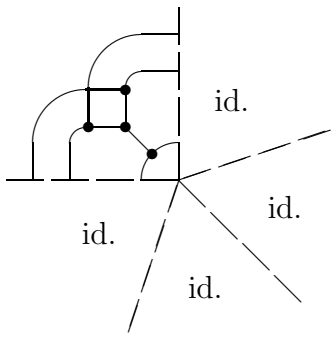
Isaacs 3-flower ($h = 6$, *class II*)

g	$\|M\|$	cycle sizes	×
$\frac{1}{2}$	7	(3) (5) (5) (5) (6) (6) (6)	2
1	6	(5) (5) (5) (7) (7) (7)	2
		(3) (5) (6) (7) (7) (8)	6
		(3) (6) (6) (6) (6) (9)	6
		(3) (5) (6) (6) (7) (9)	12
		(3) (5) (5) (7) (7) (9)	6
$1_\circ$	6	(3) (6) (6) (6) (6) (9)	2
		(3) (5) (6) (6) (6) (10)	6
		(3) (5) (5) (6) (7) (10)	12
$1\frac{1}{2}$	5	(5) (7) (8) (8) (8)	6
		(6) (7) (7) (7) (9)	2
		(6) (6) (7) (7) (10)	6
		(5) (6) (7) (8) (10)	12
		(5) (5) (8) (8) (10)	6
		(5) (6) (7) (7) (11)	6
		(5) (5) (7) (8) (11)	12
maps			$20_\circ \subset 104$



Isaacs 5-flower ($h = 10$, *class II*)

g	$\|M\|$	cycle sizes	×
$1\frac{1}{2}$	9	(5) (6) (6) (6) (6) (6) (7) (9) (9)	10
		(5) (6) (6) (6) (6) (6) (7) (7) (11)	30
2	8	(not listed)	
$2_\circ$	8	(6) (6) (6) (7) (7) (9) (9) (10)	20
		(5) (6) (6) (6) (7) (9) (10) (11)	40
		(5) (6) (6) (6) (7) (7) (10) (13)	20
		(6) (6) (6) (6) (6) (7) (9) (14)	10
		(5) (6) (6) (6) (6) (6) (11) (14)	10
		(5) (6) (6) (6) (6) (6) (10) (15)	2
		(5) (6) (6) (6) (6) (6) (7) (18)	10
$2\frac{1}{2}$	7	(not listed)	



3	6	(not listed)	
$3_\circ$	6	(8) (8) (9) (11) (12) (12)	20
		(6) (9) (10) (11) (12) (12)	20
		(6) (10) (10) (10) (11) (13)	10
		(6) (9) (10) (11) (11) (13)	20
		(7) (8) (8) (12) (12) (13)	20
		(6) (7) (10) (12) (12) (13)	10
		(8) (8) (9) (10) (11) (14)	40
		(7) (9) (9) (10) (11) (14)	60
		(6) (9) (10) (10) (11) (14)	50
		(6) (7) (10) (10) (13) (14)	20
		(5) (7) (10) (11) (13) (14)	20
		(6) (6) (9) (11) (14) (14)	50
		(6) (6) (7) (13) (14) (14)	20
		(5) (6) (10) (12) (12) (15)	10
		(6) (7) (9) (10) (13) (15)	40
		(6) (6) (8) (12) (13) (15)	20
		(6) (6) (9) (10) (14) (15)	10
		(6) (8) (8) (10) (11) (17)	10
		(6) (6) (6) (10) (15) (17)	10
		(6) (7) (9) (9) (11) (18)	60
		(6) (6) (6) (11) (13) (18)	20
		(5) (6) (7) (11) (13) (18)	20
		(6) (6) (6) (7) (17) (18)	10
$3\frac{1}{2}$	5	(not listed)	
maps			$682_\circ \subset 45\,930$

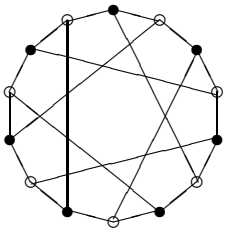
B.9 Graphs from projective geometry

The Fano, Pappus, and Desargues graphs are highly symmetric distance-regular graphs formed from self-dual arrangements of points and lines occurring in the classical projective planes. Both the points and the lines are represented by nodes; edges represent incidence between them, making these graphs bipartite.

Fano incidence graph, aka Heawood (3,6)-cage ($h = 7$, *bipartite*)

The Fano graph represents a whole projective plane, the classical one of order two (7 points & 7 lines), the Fano plane. One cyclic representation is shown here.

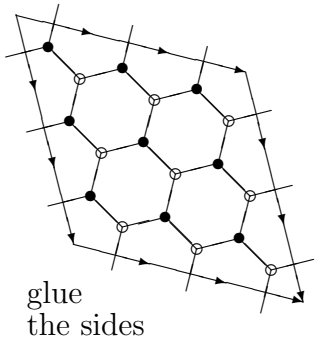
g	$\lvert M\rvert$	cycle sizes	×
$1_\circ$	7	(6) (6) (6) (6) (6) (6) (6)	8
$1\frac{1}{2}$	6	(6) (6) (6) (8) (8) (8)	56
		(6) (6) (6) (6) (8) (10)	168
2	5	(6) (6) (10) (10) (10)	168
		(6) (6) (8) (10) (12)	168
		(6) (6) (6) (10) (14)	168
$2_\circ$	5	(6) (8) (8) (8) (12)	56
$2\frac{1}{2}$	4	(10) (10) (10) (12)	112
		(6) (12) (12) (12)	56
$3_\circ$	3	(14) (14) (14)	8
maps			$72_\circ \subset 968$



Pappus incidence graph ($h = 9$, *bipartite*)

The Pappus arrangement is the affine plane of order 3 without vertical lines: point (x, y) lies on line $[\,m, b\,]$ if it satisfies $y = mx + b$; all variables (mod 3).

g	$\lvert M\rvert$	cycle sizes	×
$1_\circ$	9	(6) (6) (6) (6) (6) (6) (6) (6) (6)	2
2	7	(6) (8) (8) (8) (8) (8) (8)	36
		(6) (6) (8) (8) (8) (8) (10)	432
		(6) (6) (6) (8) (8) (10) (10)	162
		(6) (6) (6) (6) (10) (10) (10)	72
		(6) (6) (6) (8) (8) (8) (12)	324
		(6) (6) (6) (6) (8) (10) (12)	216
		(6) (6) (6) (6) (6) (12) (12)	54
		(6) (6) (6) (6) (8) (8) (14)	108
$2_\circ$	7	(6) (6) (6) (8) (8) (10) (10)	54
		(6) (6) (6) (8) (8) (8) (12)	36



glue
the sides
into a torus

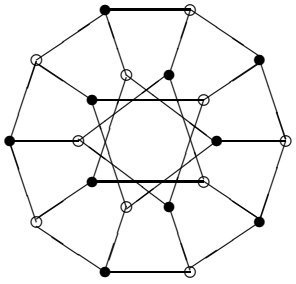
$2\frac{1}{2}$	6	(8) (8) (8) (8) (10) (12)	432
		(6) (8) (8) (10) (10) (12)	324
		(6) (8) (8) (8) (12) (12)	1 296
		(6) (6) (8) (10) (12) (12)	1 080
		(6) (6) (6) (12) (12) (12)	132
		(6) (8) (8) (8) (10) (14)	648
		(6) (6) (8) (8) (12) (14)	972
		(6) (6) (6) (8) (12) (16)	216
		(6) (6) (8) (8) (8) (18)	72
3	5	(10) (10) (10) (12) (12)	54
		(8) (10) (12) (12) (12)	432
		(6) (12) (12) (12) (12)	216
		(8) (10) (10) (12) (14)	216
		(8) (8) (12) (12) (14)	540
		(6) (10) (12) (12) (14)	648
		(8) (8) (10) (14) (14)	270
		(6) (8) (12) (14) (14)	432
		(6) (6) (14) (14) (14)	54
		(8) (8) (10) (12) (16)	540
		(6) (10) (10) (12) (16)	216
		(6) (8) (12) (12) (16)	864
		(6) (6) (12) (14) (16)	540
		(6) (8) (8) (16) (16)	378
		(6) (6) (10) (16) (16)	162
		(6) (8) (10) (12) (18)	216
		(6) (8) (8) (14) (18)	108
		(6) (6) (10) (14) (18)	108
$3_\circ$	5	(10) (10) (10) (12) (12)	18
		(8) (8) (10) (14) (14)	54
		(6) (6) (14) (14) (14)	18
		(6) (6) (6) (18) (18)	18
$3\frac{1}{2}$	4	(12) (12) (14) (16)	648
		(10) (12) (16) (16)	216
		(8) (14) (16) (16)	216
		(12) (12) (12) (18)	144
		(8) (12) (16) (18)	216
maps			$200_\circ \subset 14\,210$

Desargues incidence graph ($h = 10$, *bipartite*)

One representation of the Desargues arrangement: let V be a set of 5 items, points pairs $\subset V$, and lines triples $\subset V$; now incidence is by pair $\subset$ triple. This is a double cover of the Petersen graph: identify a pair and its complement triple (a node and its unique opposite at maximal distance 5) with a single node of the latter.

The picture here shows it as the generalised Petersen graph P(10, 3), cf. p 16.

g	$\lvert M\rvert$	cycle sizes	×
2	8	(6) (6) (6) (8) (8) (8) (8) (10)	120
		(6) (6) (6) (6) (8) (8) (10) (10)	510
		(6) (6) (6) (6) (6) (10) (10) (10)	240
$2_\circ$	8	(6) (6) (8) (8) (8) (8) (8) (8)	20
		(6) (6) (6) (6) (6) (10) (10) (10)	40
$2\frac{1}{2}$	7	(8) (8) (8) (8) (8) (10) (10)	12
		(6) (8) (8) (8) (10) (10) (10)	320
		(6) (6) (8) (10) (10) (10) (10)	1 080
		(6) (6) (8) (8) (10) (10) (12)	660
		(6) (6) (6) (10) (10) (10) (12)	360
		(6) (6) (8) (8) (8) (12) (12)	240
		(6) (6) (6) (8) (10) (12) (12)	960
		(6) (6) (6) (6) (12) (12) (12)	40
		(6) (6) (8) (8) (8) (10) (14)	360
		(6) (6) (6) (8) (10) (10) (14)	960
		(6) (6) (6) (8) (8) (12) (14)	240
		(6) (6) (6) (6) (10) (12) (14)	720
		(6) (6) (6) (6) (8) (14) (14)	60
		(6) (6) (8) (8) (8) (8) (16)	120
		(6) (6) (6) (8) (8) (10) (16)	360
		(6) (6) (6) (6) (10) (10) (16)	240
		(6) (6) (6) (6) (6) (10) (20)	24



3	6	(10) (10) (10) (10) (10) (10)	602
		(8) (10) (10) (10) (10) (12)	240
		(8) (8) (10) (10) (12) (12)	1 020
		(6) (10) (10) (10) (12) (12)	840
		(6) (8) (10) (12) (12) (12)	960
		(6) (6) (12) (12) (12) (12)	240
		(8) (8) (10) (10) (10) (14)	960
		(6) (10) (10) (10) (10) (14)	2 040
		(8) (8) (8) (10) (12) (14)	240
		(6) (8) (10) (10) (12) (14)	2 400
		(6) (6) (10) (12) (12) (14)	2 400
		(8) (8) (8) (8) (14) (14)	240
		(6) (8) (8) (10) (14) (14)	1 200
		(6) (6) (10) (10) (14) (14)	2 100
		(6) (6) (8) (12) (14) (14)	960
		(6) (6) (6) (14) (14) (14)	160
		(8) (8) (8) (10) (10) (16)	360
		(6) (8) (10) (10) (10) (16)	720
		(6) (8) (8) (10) (12) (16)	600
		(6) (6) (10) (10) (12) (16)	960
		(6) (6) (8) (12) (12) (16)	360
		(6) (8) (8) (8) (14) (16)	240
		(6) (6) (8) (10) (14) (16)	1 200
		(6) (6) (6) (12) (14) (16)	720
		(6) (6) (8) (8) (16) (16)	420
		(6) (6) (6) (10) (16) (16)	120
		(6) (6) (10) (10) (10) (18)	240
		(6) (6) (8) (10) (12) (18)	480
		(6) (6) (6) (12) (12) (18)	120
		(6) (6) (8) (8) (14) (18)	480
		(6) (6) (6) (10) (14) (18)	120
		(6) (6) (6) (8) (16) (18)	240
		(6) (6) (6) (6) (18) (18)	30
		(6) (6) (6) (10) (12) (20)	120
$3_\circ$	6	(10) (10) (10) (10) (10) (10)	40
		(8) (8) (10) (10) (12) (12)	60
		(6) (10) (10) (10) (10) (14)	120
		(6) (6) (10) (12) (12) (14)	120
		(6) (6) (10) (10) (14) (14)	120
		(8) (8) (8) (10) (10) (16)	120
		(6) (6) (10) (10) (10) (18)	20

$3\frac{1}{2}$	5	(10) (12) (12) (12) (14)	120
		(10) (10) (12) (14) (14)	2 160
		(8) (12) (12) (14) (14)	360
		(8) (10) (14) (14) (14)	840
		(6) (12) (14) (14) (14)	240
		(10) (10) (12) (12) (16)	360
		(8) (12) (12) (12) (16)	240
		(10) (10) (10) (14) (16)	960
		(8) (10) (12) (14) (16)	720
		(6) (12) (12) (14) (16)	840
		(8) (8) (14) (14) (16)	360
		(6) (10) (14) (14) (16)	480
		(8) (10) (10) (16) (16)	360
		(8) (8) (12) (16) (16)	240
		(6) (10) (12) (16) (16)	480
		(10) (10) (10) (12) (18)	360
		(8) (10) (12) (12) (18)	240
		(6) (12) (12) (12) (18)	40
		(8) (10) (10) (14) (18)	240
		(8) (8) (12) (14) (18)	360
		(6) (10) (12) (14) (18)	1 800
		(6) (8) (14) (14) (18)	720
		(6) (10) (10) (16) (18)	600
		(6) (8) (12) (16) (18)	480
		(6) (6) (14) (16) (18)	480
		(6) (6) (12) (18) (18)	120
		(10) (10) (10) (10) (20)	120
		(6) (10) (10) (14) (20)	240
		(6) (6) (10) (18) (20)	120
4	4	(14) (14) (16) (16)	840
		(12) (16) (16) (16)	240
		(14) (14) (14) (18)	520
		(12) (14) (16) (18)	360
		(10) (16) (16) (18)	540
		(12) (12) (18) (18)	420
		(10) (14) (18) (18)	720
		(8) (16) (18) (18)	360
		(6) (18) (18) (18)	160
		(10) (12) (18) (20)	120
$4_\circ$	4	(10) (10) (20) (20)	12
maps			$672_\circ \subset 51\,390$

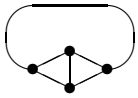
B.10 Platonic solids

The cube was listed before among the prisms (p 6) and here again for reference. Octahedron and icosahedron do not arise as their graphs are not trivalent. The graph diagrams on this page are not centered on a node (vertex) or a cycle (face) but on an edge, for a change.

Note the dodecahedron is a double cover of the Petersen graph, and the cube of the tetrahedron, each time by identifying opposite points on the former with single nodes of the latter. Only some of the maps survive these homomorphisms.

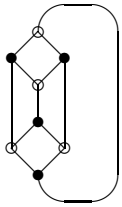
Tetrahedron $\mathbf{K_4}$ ($h = 2$)

g	$\lvert M\rvert$	cycle sizes	×
$0_\circ$	4	(3) (3) (3) (3)	1
$\frac{1}{2}$	3	(4) (4) (4)	1
maps			$1_\circ \subset 2$



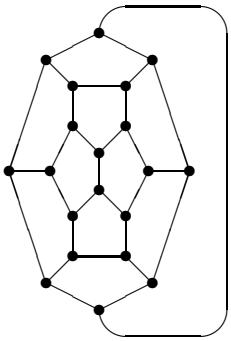
Cube ($h = 4$, *bipartite*)

g	$\lvert M\rvert$	cycle sizes	×
$0_\circ$	6	(4) (4) (4) (4) (4) (4)	1
1	4	(6) (6) (6) (6)	6
		(4) (6) (6) (8)	12
$1_\circ$	4	(6) (6) (6) (6)	4
		(4) (4) (8) (8)	3
maps			$8_\circ \subset 26$



Dodecahedron ($h = 10$)

g	$\lvert M\rvert$	cycle sizes	×
$0_\circ$	12	(5) (5) (5) (5) (5) (5) (5) (5) (5) (5) (5) (5)	1
$1\frac{1}{2}$	9	(5) (5) (5) (5) (5) (5) (8) (11) (11)	60
2	8	(5) (5) (5) (8) (8) (8) (10) (11)	60
		(5) (5) (5) (5) (8) (8) (12) (12)	150
		(5) (5) (5) (5) (8) (8) (11) (13)	60
		(5) (5) (5) (5) (8) (8) (10) (14)	120
		(5) (5) (5) (5) (8) (8) (9) (15)	60
		(5) (5) (5) (5) (5) (8) (12) (15)	180
$2_\circ$	8	(5) (5) (5) (8) (8) (8) (9) (12)	20
		(5) (5) (5) (5) (5) (5) (15) (15)	30
		(5) (5) (5) (5) (5) (5) (12) (18)	10



$2\frac{1}{2}$	7	(8) (8) (8) (9) (9) (9) (9)	20
		(5) (9) (9) (9) (9) (9) (10)	12
		(5) (8) (8) (9) (10) (10) (10)	60
		(5) (8) (9) (9) (9) (9) (11)	60
		(5) (8) (8) (9) (9) (10) (11)	120
		(5) (8) (8) (8) (10) (10) (11)	180
		(5) (8) (8) (8) (9) (11) (11)	120
		(5) (5) (9) (9) (10) (11) (11)	60
		(5) (5) (8) (9) (11) (11) (11)	240
		(5) (8) (8) (9) (9) (9) (12)	120
		(5) (5) (9) (9) (9) (11) (12)	120
		(5) (5) (8) (9) (10) (11) (12)	240
		(5) (5) (8) (8) (11) (11) (12)	600
		(5) (5) (5) (11) (11) (11) (12)	40
		(5) (5) (8) (9) (9) (12) (12)	180
		(5) (5) (8) (8) (10) (12) (12)	60
		(5) (5) (5) (9) (12) (12) (12)	80
		(5) (5) (8) (9) (9) (11) (13)	60
		(5) (5) (8) (8) (9) (12) (13)	120
		(5) (5) (5) (8) (12) (12) (13)	120
		(5) (5) (8) (9) (9) (10) (14)	120
		(5) (5) (8) (8) (9) (11) (14)	240
		(5) (5) (5) (9) (11) (11) (14)	240
		(5) (5) (8) (8) (8) (12) (14)	60
		(5) (5) (5) (8) (11) (12) (14)	360
		(5) (5) (5) (9) (9) (13) (14)	60
		(5) (5) (8) (9) (9) (9) (15)	60
		(5) (5) (8) (8) (8) (11) (15)	120
		(5) (5) (5) (8) (11) (11) (15)	120
		(5) (5) (5) (9) (9) (12) (15)	60
		(5) (5) (5) (8) (8) (14) (15)	240
		(5) (5) (8) (8) (9) (9) (16)	60
		(5) (5) (8) (8) (8) (10) (16)	60
		(5) (5) (5) (9) (9) (11) (16)	60
		(5) (5) (5) (8) (9) (12) (16)	120
		(5) (5) (5) (5) (12) (12) (16)	60
		(5) (5) (5) (9) (9) (9) (18)	20
		(5) (5) (5) (8) (8) (11) (18)	120
		(5) (5) (5) (5) (8) (14) (18)	60

3	6	(10) (10) (10) (10) (10) (10)	71
		(9) (10) (10) (10) (10) (11)	360
		(9) (9) (10) (10) (11) (11)	360
		(8) (10) (10) (10) (11) (11)	450
		(9) (9) (9) (11) (11) (11)	40
		(8) (9) (10) (11) (11) (11)	600
		(8) (8) (11) (11) (11) (11)	450
		(5) (11) (11) (11) (11) (11)	24
		(9) (9) (10) (10) (10) (12)	90
		(8) (10) (10) (10) (10) (12)	120
		(8) (9) (10) (10) (11) (12)	360
		(8) (8) (10) (11) (11) (12)	600
		(5) (10) (11) (11) (11) (12)	360
		(8) (8) (10) (10) (12) (12)	180
		(5) (10) (10) (11) (12) (12)	240
		(5) (9) (11) (11) (12) (12)	120
		(8) (8) (8) (12) (12) (12)	60
		(5) (8) (11) (12) (12) (12)	120
		(8) (9) (10) (10) (10) (13)	120
		(8) (9) (9) (10) (11) (13)	240
		(8) (8) (10) (10) (11) (13)	360
		(8) (8) (9) (11) (11) (13)	240
		(5) (10) (10) (11) (11) (13)	480
		(5) (9) (11) (11) (11) (13)	120
		(8) (8) (9) (10) (12) (13)	240
		(5) (10) (10) (10) (12) (13)	120
		(8) (8) (8) (11) (12) (13)	120
		(5) (9) (10) (11) (12) (13)	360
		(5) (8) (11) (11) (12) (13)	480
		(5) (8) (10) (12) (12) (13)	480
		(8) (8) (9) (9) (13) (13)	60
		(8) (8) (8) (10) (13) (13)	60
		(5) (8) (10) (11) (13) (13)	240
		(5) (8) (9) (12) (13) (13)	180
		(5) (5) (12) (12) (13) (13)	240
		(5) (5) (11) (13) (13) (13)	60
		(8) (9) (9) (10) (10) (14)	180
		(8) (9) (9) (9) (11) (14)	120
		(8) (8) (9) (10) (11) (14)	120
		(8) (8) (8) (11) (11) (14)	120
		(5) (9) (10) (11) (11) (14)	240
		(5) (8) (11) (11) (11) (14)	240

3	6	(5) (8) (10) (11) (12) (14)	240
		(5) (5) (12) (12) (12) (14)	60
		(5) (9) (9) (10) (13) (14)	120
		(5) (8) (10) (10) (13) (14)	120
		(5) (8) (9) (11) (13) (14)	240
		(5) (8) (8) (12) (13) (14)	120
		(5) (5) (10) (13) (13) (14)	60
		(5) (8) (8) (11) (14) (14)	240
		(5) (5) (11) (11) (14) (14)	120
		(5) (5) (10) (12) (14) (14)	60
		(8) (8) (8) (10) (11) (15)	120
		(5) (9) (10) (10) (11) (15)	120
		(5) (9) (9) (11) (11) (15)	120
		(5) (8) (10) (11) (11) (15)	240
		(8) (8) (8) (9) (12) (15)	120
		(5) (8) (10) (10) (12) (15)	120
		(5) (8) (9) (11) (12) (15)	120
		(5) (8) (8) (12) (12) (15)	480
		(5) (5) (11) (12) (12) (15)	120
		(5) (8) (9) (10) (13) (15)	120
		(5) (8) (8) (11) (13) (15)	240
		(5) (5) (11) (11) (13) (15)	120
		(5) (5) (10) (12) (13) (15)	120
		(5) (5) (9) (13) (13) (15)	60
		(5) (8) (9) (9) (14) (15)	240
		(5) (8) (8) (10) (14) (15)	120
		(5) (5) (10) (11) (14) (15)	120
		(5) (5) (8) (13) (14) (15)	120
		(5) (8) (8) (9) (15) (15)	120
		(5) (5) (10) (10) (15) (15)	60
		(5) (5) (9) (11) (15) (15)	120
		(5) (5) (8) (12) (15) (15)	420
		(5) (5) (5) (15) (15) (15)	40
		(5) (8) (8) (11) (12) (16)	240
		(5) (5) (10) (12) (12) (16)	60
		(5) (8) (9) (10) (11) (17)	120
		(5) (8) (8) (11) (11) (17)	240
		(5) (5) (11) (11) (11) (17)	120
		(5) (5) (10) (11) (12) (17)	120
		(5) (5) (8) (11) (14) (17)	360
		(5) (5) (8) (8) (17) (17)	60

$3_\circ$	6	(10) (10) (10) (10) (10) (10)	5
		(8) (10) (10) (10) (11) (11)	30
		(9) (9) (10) (10) (10) (12)	10
		(8) (8) (8) (12) (12) (12)	20
		(5) (8) (10) (11) (13) (13)	60
$3\frac{1}{2}$	5	(11) (11) (12) (12) (14)	180
		(10) (11) (12) (13) (14)	240
		(10) (11) (11) (14) (14)	360
		(10) (10) (12) (14) (14)	180
		(9) (11) (12) (14) (14)	120
		(8) (12) (12) (14) (14)	60
		(8) (11) (13) (14) (14)	180
		(9) (9) (14) (14) (14)	40
		(8) (10) (14) (14) (14)	120
		(5) (13) (14) (14) (14)	60
		(11) (11) (11) (12) (15)	240
		(10) (10) (11) (14) (15)	120
		(9) (10) (12) (14) (15)	120
		(8) (11) (12) (14) (15)	240
		(8) (9) (14) (14) (15)	60
		(5) (12) (14) (14) (15)	180
		(11) (11) (11) (11) (16)	60
		(10) (11) (11) (12) (16)	180
		(10) (10) (10) (14) (16)	60
		(9) (10) (11) (14) (16)	240
		(8) (11) (11) (14) (16)	360
		(8) (10) (12) (14) (16)	180
		(5) (11) (14) (14) (16)	300
		(9) (10) (10) (15) (16)	120
		(8) (9) (12) (15) (16)	120
		(8) (9) (11) (16) (16)	120
		(8) (8) (12) (16) (16)	60
		(10) (11) (11) (11) (17)	360
		(10) (10) (11) (12) (17)	240
		(8) (11) (12) (12) (17)	120
		(9) (10) (10) (14) (17)	120
		(8) (10) (11) (14) (17)	600
		(8) (9) (12) (14) (17)	120
		(5) (10) (12) (16) (17)	120
		(5) (8) (14) (16) (17)	120

$3\frac{1}{2}$	5	(10) (10) (11) (11) (18)	180
		(9) (10) (11) (12) (18)	120
		(8) (11) (11) (12) (18)	240
		(8) (10) (12) (12) (18)	60
		(9) (9) (10) (14) (18)	60
		(8) (10) (10) (14) (18)	60
		(8) (9) (11) (14) (18)	120
		(8) (8) (12) (14) (18)	120
		(5) (11) (12) (14) (18)	360
		(5) (9) (14) (14) (18)	60
		(5) (10) (12) (15) (18)	120
		(5) (8) (14) (15) (18)	120
		(5) (10) (11) (16) (18)	120
		(5) (9) (12) (16) (18)	120
		(5) (8) (11) (18) (18)	120
		(5) (5) (14) (18) (18)	60
		(8) (10) (10) (12) (20)	60
		(8) (8) (12) (12) (20)	30
		(8) (8) (10) (14) (20)	240
4	4	(15) (15) (15) (15)	30
		(13) (13) (17) (17)	60
		(11) (15) (17) (17)	360
		(9) (17) (17) (17)	80
$4\frac{1}{2}$	3	(20) (20) (20)	10
maps			$186_\circ \subset 30\,843$

References

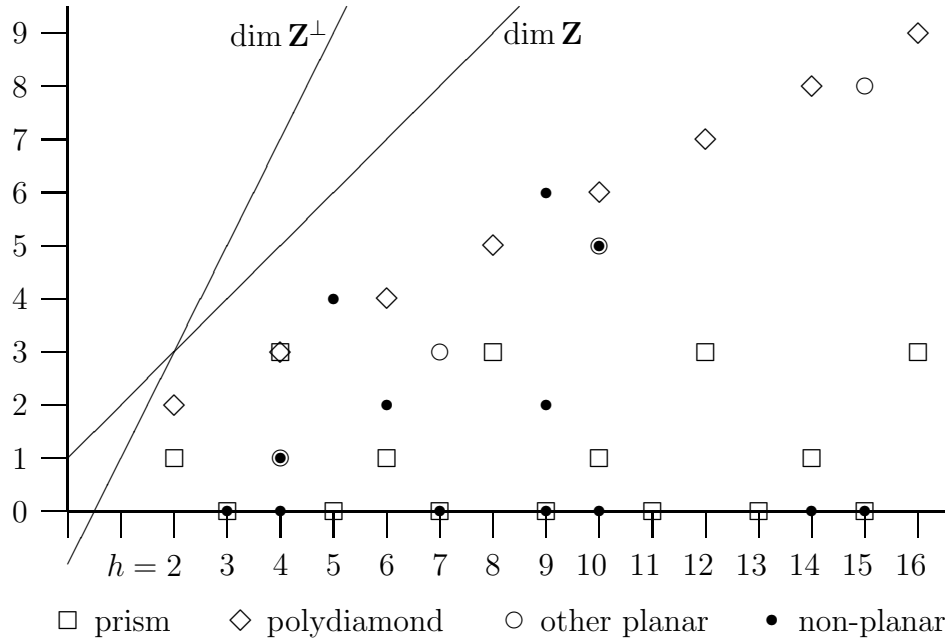
[FW77] S. Fiorini & R. J. Wilson, *Edge-colourings of graphs*, Research Notes in Math. **16**, Pitman 1977, ISBN 0 273 01129 4

[RW83] R. W. Robinson & N. C. Wormald, "Numbers of Cubic Graphs", *J. Graph Theory* **7** (1983) pp 463–7

[Slo∞] N. J. A. Sloane, *On-Line Encyclopedia of Integer Sequences* (ongoing publication), http://www.research.att.com/projects/OEIS

Appendix D: Dimension of $\mathbf{Z}^\perp \cap \mathbf{Z}$ surveyed

For trivalent graphs G, let h again be the integer such that G has $2h$ nodes and $3h$ edges; when G is connected $\dim \mathbf{Z}^\perp = 2h - 1$ and $\dim \mathbf{Z} = h + 1$. Here $\dim \mathbf{Z}^\perp \cap \mathbf{Z}$ is plotted against h for a selection of connected bridge-free trivalent graphs.



$h = 2$ (4 nodes, 6 edges): $\dim N = 3$, $\dim Z = 3$, $\dim N \cap Z =$

1 — digonal prism
2 — tetrahedron K_4 (mono-diamond)

$h = 3$ (6 nodes, 9 edges): $\dim N = 5$, $\dim Z = 4$, $\dim N \cap Z =$

0 — triangular prism — utilities graph $K_{3,3}$

$h = 4$ (8 nodes, 12 edges): $\dim N = 7$, $\dim Z = 5$, $\dim N \cap Z =$

0 — mitre (section **B.4**)
1 — bracelet — twisted cube (section **B.4**)
3 — di-diamond — cube (square prism)

$h = 5$ (10 nodes, 15 edges): $\dim N = 9$, $\dim Z = 6$, $\dim N \cap Z =$

0 — pentagonal prism
4 — Petersen $P_{5,2}$

$h = 6$ (12 nodes, 18 edges): $\dim N = 11$, $\dim Z = 7$, $\dim N \cap Z =$

1 — hexagonal prism
2 — Isaacs 3-flower
4 — tri-diamond

$h = 7$ (14 nodes, 21 edges): $\dim N = 13$, $\dim Z = 8$, $\dim N \cap Z =$

0 — heptagonal prism — generalised Petersen $P_{7,2}$ — Fano
3 — König's counterexample from section **B.5**

$h = 8$ (16 nodes, 24 edges): $\dim N = 15$, $\dim Z = 9$, $\dim N \cap Z =$

3 — octagonal prism
5 — tetra-diamond

$h = 9$ (18 nodes, 27 edges): $\dim N = 17$, $\dim Z = 10$, $\dim N \cap Z =$

0 — enneagonal prism — generalised Petersen $P_{9,2}$
2 — a Blanuša snark — smallest 0-symmetric graph [CFP81]
6 — Pappus

$h = 10$ (20 nodes, 30 edges): $\dim N = 19$, $\dim Z = 11$, $\dim N \cap Z =$

0 — Isaacs 5-flower
1 — decagonal prism
5 — dodecahedron — Desargues
6 — penta-diamond

$h = 11$ (22 nodes, 33 edges): $\dim N = 21$, $\dim Z = 12$, $\dim N \cap Z =$

0 — endecagonal prism

$h = 12$ (24 nodes, 36 edges): $\dim N = 23$, $\dim Z = 13$, $\dim N \cap Z =$

3 — dodecagonal prism
7 — hexa-diamond

$h = 13$ (26 nodes, 39 edges): $\dim N = 25$, $\dim Z = 14$, $\dim N \cap Z =$

0 — trisdecagonal prism

$h = 14$ (28 nodes, 42 edges): $\dim N = 27$, $\dim Z = 15$, $\dim N \cap Z =$

0 — Isaacs 7-flower
1 — tetradecagonal prism
8 — hepta-diamond

$h = 15$ (30 nodes, 45 edges): $\dim N = 29$, $\dim Z = 16$, $\dim N \cap Z =$

0 — pentadecagonal prism — pair-syntheme graph
8 — `nohampath` (non-Hamiltonian-path graph of Appendix ??

$h = 16$ (32 nodes, 48 edges): $\dim N = 31$, $\dim Z = 17$, $\dim N \cap Z =$

3 — hexadecagonal prism
9 — octa-diamond

Beyond the diagram of the previous page:

$h = 23$ (46 nodes, 69 edges): $\dim N = 45$, $\dim Z = 24$, $\dim N \cap Z =$

0 — Tutte's non-Hamiltonian planar graph [Wil02 p46 and cover]

Comments on the $\dim \mathbf{Z}^\perp \cap \mathbf{Z}$ survey above

Let $d = \dim \mathbf{Z}^\perp \cap \mathbf{Z}$. In this survey $d \leq (h+3)/2$ (higher d were found below).

The k-diamonds, $(\text{  })^k$ arranged cyclically, have $d = k+1 = h/2+1$. The other planar high scorers (König's, `nohampath`) also consist mainly of diamonds, but higher $d = (h+3)/2$ are attained by the Petersen and Pappus graphs.

The h-gonal prisms, $(\text{  })^h$ arranged cyclically, have $d = 3$ if h is divisible by 4, $d = 1$ for other even h, and $d = 0$ for odd h. It is not hard to exhibit the type of $\mathbf{Z}^\perp \cap \mathbf{Z}$ element that only exists for even h, and the types that only exist when $h \equiv 0 \pmod 4$.

Overall, there seems to be little correspondence with graph structure! Planar or not, bipartite or not, Class I (Tait-colorable) or not, Hamiltonian or not doesn't predispose for low or high $\dim \mathbf{Z}^\perp \cap \mathbf{Z}$. The Petersen has $d = 4$, the 3-flower derived from it $d = 2$ and the higher flower snarks $d = 0$. Tutte's graph has $d = 0$ and another non-Hamiltonian one $d = 8$.

Cyclical symmetry (as in prisms, generalised Petersen $\mathrm{P}_{h,2}$ and flower snarks) doesn't seem to do much for d (the polydiamonds, also cyclically symmetric, share their high d with other graphs with high diamond content).

More complex symmetry appears to boost the d of cube, Petersen graph and dodecahedron... but then the pair-syntheme graph has a resounding $d = 0$. Likewise, Desargues and Pappus have $d = 5$ and 6 while the Fano graph (also a point-line incidence graph from projective geometry) has $d = 0$.

Random graphs' $\dim \mathbf{Z}^\perp \cap \mathbf{Z}$

There are various approaches to generating a random sample of graphs, and they will give different distributions (certain kinds of graphs more often, others less or not at all). To generate a trivalent graph of $2h$ nodes and $3h$ edges, we can

- start with $2h$ nodes each with 3 "handles" and then $3h$ times randomly pick two of the still remaining "handles" to run an edge between. Detail: to avoid pseudographs we can demand the "handles" belong to different nodes; redo the graph if for the last edge only "handles" of the same node remain.
- start with $3h$ edges each with 2 "handles" and then $2h$ times randomly pick three of the still remaining "handles" to join into a node. Detail: to avoid pseudographs we can demand the "handles" belong to different edges; etc.
- start with the "nitrogen" multigraph $\bullet \equiv \bullet$ and then $h-1$ times randomly pick edges e_0 ($\mathrm{X}_0\mathrm{Z}_0$ say) and e_1 ($\mathrm{X}_1\mathrm{Z}_1$) and "turn II into H", that is, re-route each e_i from X_i to a new node Y_i with a new edge from Z_i to Y_i, and finally draw an edge $\mathrm{Y}_0\mathrm{Y}_1$. Allowing $e_0 = e_1$ (taking care to finish re-routing e_0 before working on it as e_1) gives multigraphs. Pseudographs never occur.

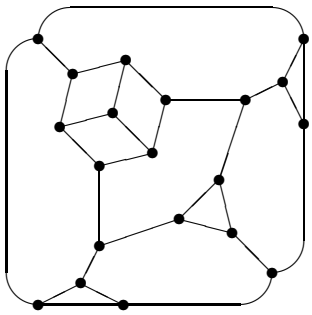
- "turn ‖ into H" as in the preceding method, but do not allow $e_0 = e_1$. At the first step the $\bullet \equiv \bullet$ multigraph turns into the tetrahedron, and no more multigraphs are formed.

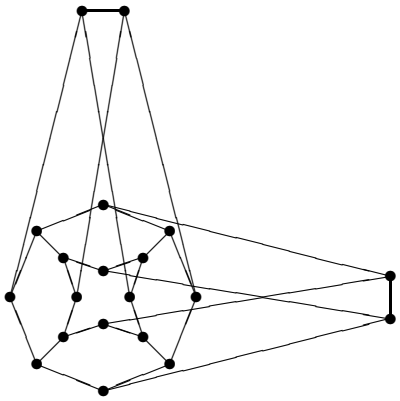
The first two methods can generate more than one component, bridges, and cobridges. The fruits of the ‖ to H method are always 3-edge-connected if multigraphs are banned, and 2-edge-connected otherwise (not only is the $-\bullet=\bullet-$ portion itself 2-edge-connected to the rest of the graph, but also e.g. diamonds are only formed from a double edge precursor). Of course, a graph consisting of $(-\bullet=\bullet-)^h$ arranged cyclically attains $d = h - 1$, larger than anything in the survey above.

Using each method several million times, for $h = 10$ without pseudographs, and analysing the graphs, the different values of $d = \dim \mathbf{Z}^\perp \cap \mathbf{Z}$ occur with the following frequencies (the $_2$ or $_3$ suffix indicates minimum edge connectivity):

random...	$d=0$	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$	$d=7$
...edges	0.370	0.418	0.171	0.036	0.005	0.000 4	0.000 03	
...nodes	0.368	0.417	0.173	0.037	0.005	0.000 4	0.000 03	0.000 003
...‖ to H_2	0.416	0.420	0.141	0.021	0.002	0.000 04	0.000 004	
...‖ to H_3	0.425	0.418	0.137	0.018	0.001	0.000 03	0.000 001	

Two 3-edge-connected graphs of $d = 6$ found:





Note that there are, up to isomorphism, only 516 344 different trivalent graphs of $h = 10$ (not counting multigraphs), 510 489 of them connected [RW83], and most of those again 3-edge-connected, so our random procedures hit on any one of them *on average* about twice per million. Of course some more often, missing other ones. An exhaustive survey of all trivalent graphs of this (or even larger) size would seem feasible. Due to time constraints no attempt was made to code up the necessary algorithms (to search all graphs of a given size in an ordered way, omitting isomorphic duplicates) to do a $\dim \mathbf{Z}^\perp \cap \mathbf{Z}$ survey under those terms.

Calculating $\dim \mathbf{Z}^\perp \cap \mathbf{Z}$ in the pedestrian way used by me is much slower for graphs with $h = 20$, but the frequencies (for smaller samples) look much the same there.

The random search is useful in highlighting the fact that graphs with higher d are progressively thinner on the ground, something obscured by the survey above of specific (often highly symmetric) graphs.

References

[CFP81] H. S. M. Coxeter, Roberto Frucht & David L. Powers, *Zero-Symmetric Graphs, Trivalent Graphical Regular Representations of Groups*, Acad. Pr. 1981, ISBN 0 12 194580 4

[RW83] R. W. Robinson & N. C. Wormald, "Numbers of Cubic Graphs", *J. Graph Theory* **7** (1983) pp 463–7

[Wil02] Robert A. Wilson, *Graphs, colourings and the four-colour theorem*, Oxford Univ. Pr. 2002, ISBN 0 19 8510624 (pbk), http://www.maths.qmul.ac.uk/~raw/graph.html (errata &c.)

Appendix H: free graphs on the hyperbolic plane

H.0 Introduction to the hyperbolic plane

When you attempt to draw a portion of a free graph, you will find the amount of room you need grows exponentially with the distance from the starting point. But there is a space with that property, and every free graph can be embedded in it as a regular polytope (i.e. with constant edge length, and same angles at each node).

The **hyperbolic plane** is a 2-dimensional manifold of constant negative curvature. It is analogous to the **sphere**, a 2-dimensional manifold of constant positive curvature. Spheres are of course all isomorphic apart from a scale factor; if the sphere is embedded in Euclidean 3-dimensional space $\mathbb{R}^3$ this scale factor appears as the radius R of the sphere. Choosing units of length such that $R = 1$, the curvature everywhere becomes $+1$. Hyperbolic planes are likewise all isomorphic apart from a scale factor, which we may as well call R again, and choosing units of length such that $R = 1$ the curvature everywhere becomes -1.

The hyperbolic plane can also be embedded in a 3-dimensional space, just not an Euclidean one but 3-dimensional Minkowski space $\mathbb{M}_3$. This makes calculations of distances and angles very easy, in close analogy to the way calculations in spherical geometry can make use of its embedding in $\mathbb{R}^3$.

Here we will recall (mostly without proof) a few salient facts of hyperbolic geometry, but no more than we will need. This is not the place for a thorough development of this well known subject, see e.g. [MCl94] for a modern treatment. In the same vein the properties of Minkowski space are not derived here. Any good relativity textbook will do so (for $\mathbb{M}_4 \supset \mathbb{M}_3$), usually in the earliest chapters before going on to general relativity; [Wey21] and [Cla79] are excellent choices. For a representation theory approach to Lorentz transformations (the non-reflecting and non-t-reflecting isometries of $\mathbb{M}_4$) see [Wey31] and [CM76].

A few words do need to be said about the way the hyperbolic plane can be embedded in Minkowski space though, as it tends not to be mentioned in texts on either subject. I have never yet seen it spelled out, in fact. Which is odd, as Minkowski spaces (anisotropic vector spaces, spaces with non-positive-definite metric) have been with us for over a century now, and are clearly the natural way to talk about the hyperbolic plane.

H.1 The hyperbolic plane in Minkowski coördinates

The special relativity theory of 1905 was reformulated in 1908 by Minkowski in what is now called **Minkowski space**, which we can define as the set of all "fourvectors" $(t; x, y, z)$, quadruples of reals (where t represents time and the other three the spatial coördinates). The usual vector addition is defined on them, and a

non-Euclidean inner product $(t;\, x,\, y,\, z) \cdot (t';\, x',\, y',\, z') = tt' - xx' - yy' - zz'$ (or sometimes with the opposite sign). Two fourvectors are considered to be orthogonal if their inner product is zero. The inner product of $(t;\, x,\, y,\, z)$ with itself, the quadratic form $t^2 - x^2 - y^2 - z^2$, is analogous to square of the length of the vector in Euclidean spaces, except that it can be positive as well as zero or negative. If t^2 is greater than, equal to or less than $x^2 + y^2 + z^2$ the fourvector is called time-like, light-like or space-like respectively.

Spaces with fewer or more dimensions, with any number of them being time-like or space-like (contributing positively or negatively to the inner product) can be defined completely analogously; all are known as anisotropic spaces. For our purposes we will need 3-dimensional Minkowski space $\mathbb{M}_3$ with vectors $(t;\, x,\, y)$ (the subspace $z = 0$ of classical Minkowski space, if you will).

We can define the sphere (centered on the origin) with $R = 1$ as the subset

$$\mathcal{S} \stackrel{\text{def}}{=} \{(x,\, y,\, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}$$

of Euclidean 3-space $\mathbb{R}^3$, and in the same way the hyperbolic plane with $R = 1$ as

$$\mathcal{H} \stackrel{\text{def}}{=} \{(t;\, x,\, y) \in \mathbb{M}_3 \mid t^2 - x^2 - y^2 = 1 \text{ and } t > 0\}$$

(without the $t > 0$ condition the set would consist of two disconnected pieces, with $t \geq +1$ and $t \leq -1$ respectively). If we define the distance between vectors $\mathbf{u}$ and $\mathbf{v}$ as $\sqrt{(\mathbf{u} - \mathbf{v}) \cdot (\mathbf{u} - \mathbf{v})}$ (which in $\mathbb{M}_3$ is only meaningful for time-like $\mathbf{u} - \mathbf{v}$) the sphere and hyperbolic plane are defined very similarly, as a sheet of points at distance 1 to $\mathbf{0}$ in their respective spaces (and everywhere $\perp$ lines from $\mathbf{0}$).

That definition embedding $\mathcal{H}$ in $\mathbb{M}_3$ uses the "positive sign convention" where the t terms are positive in the inner product, and the spatial ones negative. I will now change to the opposite sign convention for $\mathbb{M}_3$ for reasons that will become clear, and correspondingly write its vectors as $(x,\, y;\, t)$ rather than $(t;\, x,\, y)$.

H.2 Metric and curvature

The sphere is, like its embedding $\mathbb{R}^3$, a **metric** space i.e. it has a nonnegative distance function $d(\mathbf{u}, \mathbf{v})$ with the three properties (i) $d(\mathbf{u}, \mathbf{v}) = 0$ iff $\mathbf{u} = \mathbf{v}$, (ii) $d(\mathbf{u}, \mathbf{v}) = d(\mathbf{v}, \mathbf{u})$, and (iii) the triangle inequality $d(\mathbf{u}, \mathbf{v}) + d(\mathbf{v}, \mathbf{w}) \geq d(\mathbf{u}, \mathbf{w})$. In $\mathbb{R}^3$ a valid distance function is of course given by $d(\mathbf{u}, \mathbf{v}) := \sqrt{(\mathbf{u} - \mathbf{v}) \cdot (\mathbf{u} - \mathbf{v})}$. For $\mathcal{S}$, we can borrow this straight-line distance through $\mathbb{R}^3$ between points of $\mathcal{S}$, but the more usual choice of distance function is one derived from it by limits and integration.

This straight-line distance Δs satisfies $(\Delta s)^2 = (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2$, and we can approximate a path along the sphere by many small straight line steps. In the limit for small steps they become (infinitesimally small) tangent vectors, and

one integrates $\mathrm{d}s$ given by $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$ to get the length of a path. Distances *along* the sphere are then given by the length of the shortest path. With this metric, the sphere acquires its constant positive curvature.

Minkowski spaces aren't metric spaces (the "square of the distance" can be of either sign and even zero for distinct points), but surprisingly the hyperbolic plane derives a proper distance function from the Minkowski metric. The reason is that for $\mathbf{u}$ and $\mathbf{v} \in \mathcal{H}$ the difference $\mathbf{u} - \mathbf{v}$ is always of the same flavor, space-like (and so in the limit tangent vectors to $\mathcal{H}$ are all space-like too).

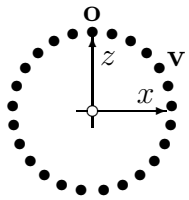
Defining $d(\mathbf{u}, \mathbf{v})$ as $\sqrt{(\mathbf{u} - \mathbf{v}) \cdot (\mathbf{u} - \mathbf{v})}$ with sign convention where spatial terms are positive i.e. $(\Delta s)^2 = (\Delta x)^2 + (\Delta y)^2 - (\Delta t)^2$ would therefore be a valid distance function when restricted to $\mathcal{H}$, but the more usual definition is again to use this only for infinitesimal distances, integrating $\mathrm{d}s$ given by $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 - \mathrm{d}t^2$ to path lengths, and defining distance *along* the hyperbolic plane as lengths of shortest paths. With this metric, the hyperbolic plane acquires its constant negative curvature.

The curvature is in general a complicated tensor, but for 2-dimensional spaces this tensor has only one independent component so it can be expressed as a scalar. In general it varies from point to point; the sphere and hyperbolic plane (and the Euclidean plane, with zero curvature) are 2-dimensional spaces where curvature is constant across the space. We will come across one interpretation of curvature (angle excess or deficit of polygons as integral of curvature over the area enclosed).

H.3 Cartesian and polar coördinates

Let's describe the lay of the land on the sphere as seen from the "North Pole" $\mathbf{o} = (0, 0, 1)$, not to be confused with $\mathbf{0} = (0, 0, 0)$ which is not on the sphere. Choosing a second point $\mathbf{v} = (x, y, z)$ at some distance r from $\mathbf{o}$ along the sphere,

$$\begin{aligned} x &= \sin r \, \cos\phi \\ y &= \sin r \, \sin\phi \\ z &= \cos r \end{aligned}$$



for some ϕ. In the picture here, $\phi = 0$, so $y = 0$. In the zx plane we plot points of $\mathcal{S}$, for various (positive and negative) r (those with negative r can also be obtained with positive r and $\phi = \pi$). Note $|r|$ is the distance from $\mathbf{o}$ *along the sphere*, points with equal increments 0.25 of r are plotted, just shy of the maximum of π. Intersecting the sphere by other planes through the z axis (other values of ϕ) gives the same picture. All this is elementary, the reason it is spelled out here is to show the analogies and differences when we move to $\mathcal{H}$. Note in passing that a circle (set of points at equal distance r to $\mathbf{o}$ for varying ϕ, not shown in the picture) has in the embedding $\mathbb{R}^3$ a

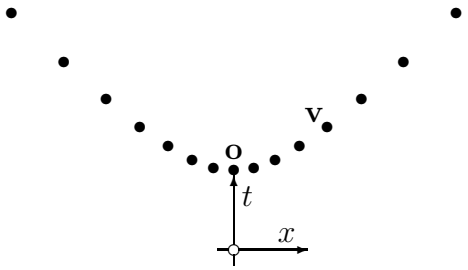
circumference 2π times its radius there, $\sin r$, and it has that same circumference

$$2\pi \sin r$$

on $\mathcal{S}$ which is not proportional to its radius r there (the growth of circles around the North Pole slows down as they reach the equator, and then they shrink again). With $2\pi \sin r : r$ varying there are no *similar* circles of different size on the sphere, and more generally no similar geometrical figures at different scale.

On $\mathcal{H}$ a very similar set of relations hold. Here too we locally effect translations on $\mathcal{H}$ by a kind of rotations (Lorentz transformations) in $\mathbb{M}_3$. Putting one point $\mathbf{o}$ again at (0, 0; 1) and a point $\mathbf{v}$ at arbitrary distance r from it at $(x, y; t)$ we have

$$\begin{aligned} x &= \sinh r \cos\phi \\ y &= \sinh r \sin\phi \\ t &= \cosh r \end{aligned}$$



for some ϕ. In the picture here, $\phi = 0$, so $y = 0$. In the tx plane we plot points of $\mathcal{H}$, for various (positive and negative) r (those with negative r can also be obtained with positive r and $\phi = \pi$). Note $|r|$ is the distance from $\mathbf{o}$ *along the hyperbolic plane*, again we plot points with equal increments 0.25 of r, up to an arbitrary cutoff. Intersecting by other planes through the t axis (other values of ϕ) gives the same picture, so the whole $\mathcal{H}$ is the surface of revolution of the slice depicted here, just as $\mathcal{S}$ was that of the circular slice on the previous page. This gives $\mathcal{H}$ a **hyperboloid bowl** shape, but keep in mind that depicting it on Euclidean paper or screen does no justice to distances in $\mathbb{M}_3$ or in $\mathcal{H}$. The dots here are at equal distance (in both), and generally distances in near 45° direction between t and a spatial axis are much shorter than they look on the page. Note in passing that a circle (set of points at equal distance r to $\mathbf{o}$ for varying ϕ, not shown in the picture) has in the embedding $\mathbb{M}_3$ a circumference 2π times its radius there, $\sinh r$, and it has that same circumference

$$2\pi \sinh r$$

on $\mathcal{H}$ which is not proportional to its radius r there (this time the growth of the circumference speeds up exponentially, as $\sinh r$ tends to $\frac{1}{2}\exp r$). Again there are no *similar* geometrical figures at different scales.

Finally, if two points on $\mathcal{S}$ have distance $2r$ there, their **straight-line distance** through $\mathbb{R}^3$ is $2\sin r$; conversely if the straight-line distance is $2s$ the distance along the sphere is $2\arcsin s$. The easiest way to see this is with a circle of which $2r$ is a diameter. It's true when the plane of the circle is perpendicular to the z-axis and remains true under 3-D rotations. In the same way, if two points on $\mathcal{H}$ have distance $2r$ there, their **straight-line distance** through $\mathbb{M}_3$ is $2\sinh r$; conversely if the straight-line distance is $2s$ the distance along the hyperbolic plane is $2\,\mathrm{arsinh}\, s$.

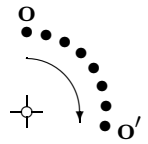
H.4 Transformations

We can choose the coördinate frame of $\mathbb{R}^3$ in such a way that any point appears as any other point on the sphere. Rotating the sphere while keeping the coördinate axes fixed has the same effect. Note how $\mathcal{S}$ as a set maps to itself while what locally act as translations on it are effected by rotations in the embedding space.

As $\mathcal{S}$ looks the same from any point on it, the lay of the land from any point is the same as the description seen from the "North Pole" $\mathbf{o}$ in the preceding section.

We will need explicit formulæ for the rotations of the embedding space that move points in $\mathcal{S}$ or $\mathcal{H}$ to other points in $\mathcal{S}$ or $\mathcal{H}$. Those on the sphere are rotations such as $R_x^r : (x,\, y,\, z) \mapsto (x',\, y',\, z')$ with

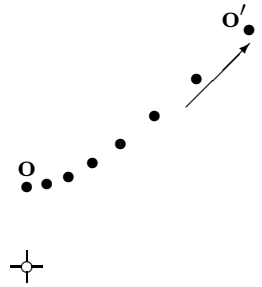
$$\begin{aligned} x' &= x\cos r + z\sin r \\ y' &= y \\ z' &= z\cos r - x\sin r \end{aligned}$$



that moves the point $\mathbf{o} = (0,\, 0,\, 1)$ to a new position $\mathbf{o}'$ by a distance r in (initially) the positive x direction, moves all points on the line $\mathbf{oo}'$ by the same amount, and other points on $\mathcal{S}$ by smaller amounts. Likewise transformations in other directions such as R_y^r that moves $\mathbf{o}$ in the y direction (not to mention rotations with other points in the rôle of our $\mathbf{o}$). The same formulæ can also be used for coördinate transformations (where the axes rotate in the opposite direction).

The corresponding ones for the hyperbolic plane are the Lorentz transformations such as $L_x^r : (x,\, y;\, t) \mapsto (x',\, y';\, t')$ with

$$\begin{aligned} x' &= x\cosh r + z\sinh r \\ y' &= y \\ t' &= z\cosh r + x\sinh r \end{aligned}$$



that moves the point $\mathbf{o} = (0,\, 0;\, 1)$ to a new position $\mathbf{o}'$ by a distance r in (initially) the positive x direction, moves all points on the line $\mathbf{oo}'$ by the same amount, and other points on $\mathcal{H}$ by bigger amounts, and e.g. its obvious counterpart L_y^r involving y (not to mention ones with other points in the rôle of our $\mathbf{o}$). Here too the same formulæ can be used for coördinate transformations with the axes rotating in the opposite direction (in $L_x^{\pm r}$ the t and x axes move both towards, or both away from, the 45° asymptote).

Of course, rotations of the embedding space can not only be used for translating things locally in $\mathcal{S}$ and $\mathcal{H}$, but also for *rotations* there. In this case, if we rotate around $\mathbf{o}$ (around the z or t axis in the embedding space, keeping z or t fixed) the transformation involves only x and y and of course uses circular rather than hyperbolic trig functions in both cases. Note what this entails: on $\mathcal{S}$, a translation here is always a rotation somewhere else on the sphere. On $\mathcal{H}$ by contrast the two are achieved by distinct sets of isometries of $\mathbb{M}_3$.

H.5 Horocycles

Intersecting $\mathcal{S}$ with a plane in $\mathbb{R}^3$ (that doesn't miss it altogether, nor grazes it at a point) gives a circle. If the plane passes through $\mathbf{0}$ in $\mathbb{R}^3$ we get a line of $\mathcal{S}$ (a geodesic, a "great circle"). Because the properties of the sphere are invariant under rotations about $\mathbf{0}$ in $\mathbb{R}^3$ every line is isomorphic to any other, e.g. to the equator on a geographical globe, and we can likewise view any ordinary circle in the orientation of one of the latitude circle on such a globe. This shows the dual properties of circles on $\mathcal{S}$. On the one hand, a circle is the set of points at **equal distance to a point** (to a geographical pole, for latitude circles). For any one circle there are two opposite points for which this is true, with distances r and $\pi - r$. On the other hand a circle is also a set of points staying at **equal distance to a line** (to the equator, in the case of a latitude circle). The whole set of points with that distance to the line consists of two circles, one on each side of the line.

Intersecting $\mathcal{H}$ with a plane in $\mathbb{M}_3$ passing through $\mathbf{0}$ also gives a line (geodesic), and $\mathcal{H}$ being invariant under Lorentz transformations that fix $\mathbf{0}$ we can put any such line in the position of e.g. the "x-axis" of $\mathcal{H}$ (the set of points with $y = 0$). This time though there are **three cases** for intersections with an arbitrary plane.

Calling the t axis vertical, if a plane is less vertical than 45° (i.e. the plane only has space-like vectors) we can apply a Lorentz transformation that makes the plane horizontal (a plane $t = C$ for some constant C). If it doesn't miss $\mathcal{H}$ altogether ($C < 1$) nor intersects it in a point ($C = 1$) the result is a circle with radius r for with $\cosh r = C$, that is, the set of points at distance r from some point (after our Lorentz transformation, that point is $\mathbf{o}$).

If the plane is steeper than 45° we can apply a Lorentz transformation that makes it purely vertical. After suitable rotation around the t axis it is now $x = S$ for some constant S. This is, for a d with $\sinh d = S$, the set of points at distance d to, and again on one side of, some line (after our transformations that line is the "y-axis"). Note that (unlike on the Euclidean plane, but just like on $\mathcal{S}$) such a set is not itself a line unless $d = 0$.

So the rôles of being equidistant to a point and being equidistant to a line are carried by two distinct classes of curves in $\mathcal{H}$. It would seem natural to call the latter class "hyperbolas". I haven't seen that term used in this way however.

Finally, a plane at exactly 45° cuts the bowl (if it doesn't miss it) in a curve that forms the limiting case of both circles and "hyperbolas". It would seem natural to call it a "parabola" but the term in use in the literature is **horocycle**. Again, we can arrange for a Lorentz transformation to make it, say, pass through $\mathbf{o}$ and a rotation about the t-axis to orient its open end towards, say, the positive x-axis. Now there are no free parameters left so all horocycles are congruent. In our chosen orientation it is the intersection of the bowl $t^2 - x^2 - y^2 = 0$ ($t > 0$) with the plane $t - 1 = x$. These two conditions are easily seen to be equivalent to $2t = y^2 + 2$ and

$2x = y^2$, so the generic point on this horocycle has parametric expression ($\frac{1}{2}y^2$, y; $\frac{1}{2}y^2+1$). This parametrisation is also the one with constant "speed" 1, that is, the length of the piece of curve between y_0 and y_1 is $|y_1 - y_0|$ as proven in section H.7 which also shows the most generally oriented horocycle has a parametrisation with x, y and t at most quadratic in the same parameter.

Note in passing that while we have now seen circles and "hyperbolas" on $\mathcal{H}$ with "parabolas" between them, and circles assuming all rôles on $\mathcal{S}$, we didn't get any ellipses this way. On an Euclidean plane ellipses arise naturally by stretching (say) the x- and y-axes by different factors. On $\mathcal{S}$ and $\mathcal{H}$ that is not such a natural thing to do, because these spaces have their own built-in length scales, so it is not surprising that our conic-section-like exercises didn't unearth any ellipses. These spaces also don't *need* ellipses the way the Euclidean plane does. In $\mathcal{H}$, keeping one point $\mathbf{p}$ of a circle fixed while making it bigger (decreasing the amount of "bend" per unit curve length) we pass through the horocycle stage where the far end recedes to infinity, and then get "hyperbolas" still curved away in the same direction from the the tangent through $\mathbf{p}$. In the Euclidean plane this doesn't work, here the circle remains finite until it degenerates to the tangent; we need to stretch the circle anisotropically into ellipses to be able to deform it further via a parabola into hyperbolas.

On the sphere, there are no **parallels**; through a point $\mathbf{p}$ outside a line ℓ there are exactly 0 lines that don't intersect ℓ. In the Euclidean plane there is famously exactly 1 such line; failure to prove this fact from the other axioms led to spherical and hyperbolic geometries the consistency of which showed the Euclidean parallel axiom indeed is independent from the others. In $\mathcal{H}$ there are infinitely many lines through $\mathbf{p}$ that do not intersect ℓ.

Wlog let ℓ be the x-axis $y = 0$ and let $\mathbf{p}$ at distance d from it be at $(0,\, s;\, c) := (0,\, \sinh d;\, \cosh d)$ with $s > 0$. The generic point on ℓ has a parametrisation $(\sinh a,\, 0;\, \cosh a)$ which is incidentally again "speed 1" (a is the distance along the line). Applying now a Lorentz transformation that moves $\mathbf{p}$ to $(0,\, 0;\, 1)$, any line through it (other than the y-axis) has equation $y = mx$ for some slope m. The generic point on ℓ moves to $(\sinh a,\, -s\cosh a;\, c\cosh a)$ under this transformation. Since $\cosh a > |\sinh a|$ the system only has a solution (a point of intersection) if $|m| > s$ (our s was positive). All the lines through $\mathbf{p}$ with $-s \le m \le +s$ do not intersect ℓ. Traditionally however only the two extreme cases $m = \pm s$ are graced with the name **parallels** to ℓ.

The connexion of parallels with horocycles is as follows. The set of lines that intersect a circle at right angles is a bundle (or "pencil") of lines passing through one point (the centre of that circle) and conversely every line in a bundle through $\mathbf{p}$ intersects any circle with $\mathbf{p}$ as centre at right angles. The set of lines that intersect a horocycle at right angles is a bundle (or "pencil") of lines all mutually parallel (in the strict sense above), and conversely a bundle of mutually parallel lines intersects every one of an infinite set of horocycles at right angles. The lines intersecting one of

our "hyperbolas" at right angles are merely mutually non-intersecting, not parallel in the strict sense.

Again, horocycles are analogous to circles, with a centre that has receded to infinity, and the transition stage between circles and "hyperbolas". We will see another example of this when we next turn our attention to polygons.

H.6 Polygons

The transformations R_x^r and L_x^r moved the point **o** in the x direction by **parallel transport**, that is without introducing any unnecessary twists. This is a local notion, not a global one. If you moved on the Earth by parallel transport from a point on the equator, facing East, first by a quarter circle along the equator, then stepping sideways to the North Pole while still facing East, and finally from the pole back to the equator walking backwards because you still don't want to turn around, you end up at the starting place, facing North. This quarter turn of orientation without ever turning is a consequence of the amount of curvature enclosed by your closed path. On $\mathcal{S}$ with its constant curvature this is proportional to the area enclosed.

Another way to see the same thing is that the three right angles of the triangular path sum to $\frac{3}{4}(2\pi)$ rather than the Euclidean $\frac{1}{2}(2\pi)$ for triangles. The extra $\frac{1}{4}(2\pi)$ is the right turn smuggled in. Thus we see that polygons have a **sum of angle excess** over the Euclidean value proportional to the area they enclose (not just proportional but equal, when $R = 1$, as the octant triangle's area is then $\frac{1}{8}4\pi$).

The same holds on the hyperbolic plane except that curvature is negative here, so there is a **sum of angle deficit**.

The lack of similarity between figures of different sizes is now obvious: very small polygons have near-Euclidean angles; larger ones have angles that deviate more.

The Platonic solids form a case in point. In the tetrahedron, octahedron and icosahedron (as painted by arcs on a sphere) there are at each vertex 3, 4 and 5 equilateral triangles coming together, as opposed to 6 as would happen on the plane. So the angles of these triangles must be $2\pi/3$, $2\pi/4$ and $2\pi/5$ as opposed to the Euclidean $2\pi/6$. And indeed the triangles of the icosahedron are most nearly Euclidean, being the smallest (relative to sphere radius) while those of the tetrahedron deviate the most and are the largest. To get 7, 8, 9... equilateral triangles meeting at a vertex we have to embed them on the hyperbolic plane and as 7 stays closest to the Euclidean value we need ever bigger triangles to get 8, 9...

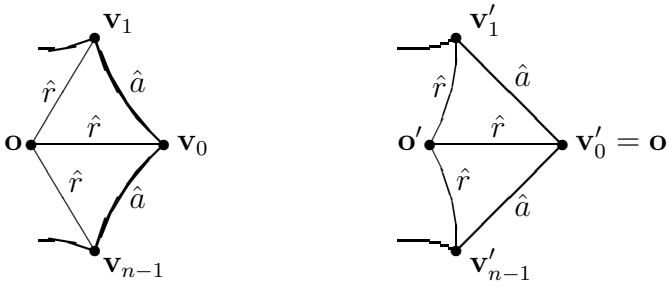
Let us now keep angle size fixed (**a right angle**, for instance) and ask ourselves what the size must be of a regular n-gon with exactly those angles, for various n. For $n = 4$ the answer will be infinitely small (on $\mathcal{S}$ or $\mathcal{H}$), because this is the Euclidean value. For $n = 3$ we saw the answer already, the octant triangle on

the sphere. For $n = 5$ and larger we need an angle deficit so we will find our right-angled polygon on the hyperbolic plane.

Centering an n-gon on $\mathbf{o}$, its vertices $\mathbf{v}_k$ are all at some distance r from $\mathbf{o}$. Using $\theta_n := 2\pi/n$ they are

$$\mathbf{v}_k \;=\; \begin{pmatrix} \sinh r \cos(k\theta_n) \\ \sinh r \sin(k\theta_n) \\ \cosh r \end{pmatrix}$$

To find the value $\hat{r}_n$ that makes the angles right for some n, consider that the lines $\mathbf{v}_0\mathbf{v}_1$ and $\mathbf{v}_0\mathbf{v}_{n-1}$ must then make $45°$ angles with the x axis $\mathbf{o}\mathbf{v}_0$. After a coördinate transformation moving the new $\mathbf{v}'_0$ to the old $\mathbf{o}$, the new $\mathbf{v}'_1$ must have x and y coördinates that sum to 0 (it must lie due "NorthWest"). The pictures here (unlike the previous ones) are inside $\mathcal{H}$, distorted as we get further from $\mathbf{o}$.



Now $\mathbf{v}'_1$ is given by

$$\mathbf{v}'_1 = \begin{pmatrix} \cosh \hat{r}_n \sinh \hat{r}_n (\cos\theta_n - 1) \\ \sinh \hat{r}_n \sin\theta_n \\ (\cosh \hat{r}_n)^2 - (\sinh \hat{r}_n)^2 \cos\theta_n \end{pmatrix}$$

and setting the sum of the x and y coördinates to zero simplifies to

$$\cosh \hat{r}_n \;=\; \frac{\sin\theta_n}{1 - \cos\theta_n}$$

which implies

$$\begin{aligned} (\cosh \hat{r}_n)^2 \;&=\; \frac{(\sin\theta_n)^2}{(1-\cos\theta_n)^2} \;=\; \frac{1-(\cos\theta_n)^2}{(1-\cos\theta_n)^2} \;=\; \frac{1+\cos\theta_n}{1-\cos\theta_n} \\ (\sinh \hat{r}_n)^2 \;&=\; (\cosh \hat{r}_n)^2 - 1 \;=\; \frac{2\cos\theta_n}{1-\cos\theta_n} \end{aligned}$$

Remains to calculate the length $\hat{a}_n$ of the sides of this n-gon. The x and y coördinates of $\mathbf{v}'_1$ were $\pm \sinh \hat{r}_n \sin\theta_n$ and must be $1/\sqrt{2}$ times $\sinh \hat{a}_n$ so

$$\begin{aligned} (\sinh \hat{a}_n)^2 \;&=\; 2\,(\sinh \hat{r}_n)^2 (\sin\theta_n)^2 \;=\; 2\,\frac{2\cos\theta_n}{1-\cos\theta_n}\big(1-(\cos\theta_n)^2\big) \\ &=\; 4\cos\theta_n(1+\cos\theta_n) \;=\; 4\cos\theta_n + 4(\cos\theta_n)^2 \\ \sinh \hat{a}_n \;&=\; 2\sqrt{\cos\theta_n + (\cos\theta_n)^2} \\ (\cosh \hat{a}_n)^2 \;&=\; 1 + (\sinh \hat{a}_n)^2 \;=\; 1 + 4\cos\theta_n + 4(\cos\theta_n)^2 \\ \cosh \hat{a}_n \;&=\; 1 + 2\cos\theta_n \end{aligned}$$

The curious fact now is that as $n \to \infty$ and $\cos\theta_n \to 1$ our $\hat{a}_n$ converge, to

$$\begin{aligned}
\cosh \hat{a}_\infty &= 3 \\
\sinh \hat{a}_\infty &= 2\sqrt{2} \\
\exp \hat{a}_\infty &= 3 + 2\sqrt{2} = (1+\sqrt{2})^2 \\
\hat{a}_\infty &= \log(3+2\sqrt{2}) \approx 1.7627471740391
\end{aligned}$$

The interpretation is that there are infinitely many discrete values $\hat{a}_n$ below $\hat{a}_\infty$ such that if we keep turning right angles (the same way round) after each time a distance of $\hat{a}_n$, we exactly close a right-angled n-gon. Choosing any other $a < \hat{a}_\infty$ will have us circling some point $\mathbf{o}$ interleaving previous steps without closing the polygon, at least the first time round. Note that $\hat{r}_n$ unlike $\hat{a}_n$ does grow without bound as n goes up.

If we choose an $a > \hat{a}_\infty$ we get a right-angled ∞-gon the vertices of which can be shown, by similar arguments as above, to lie at some equal distance d from a straight line which we could call its directrix. In other words while the vertices of a finite n-gon lie on a circle, those of such an infinite polygon lie on one of the "hyperbolas" of the previous section. Note d grows without bound as a approaches $\hat{a}_\infty$ from above.

The most interesting right-angled ∞-gon of all must be the limiting case with $a = \hat{a}_\infty$. It neither closes up, nor has a directrix; its points lie on a **horocycle**.

We can ask the same questions about **regular polygons with angle 120**$^\circ$ at each vertex. The calculation goes much like before. Let's use $\widehat{r}_n$ and $\widehat{a}_n$ for the radius and the side of such n-gons to distinguish them from the $\hat{r}_n$ and $\hat{a}_n$ for right-angled polygons we did before.

Again $\mathbf{v}_1$ is moved to $\mathbf{v}_1'$ by moving $\mathbf{v}_0$ to $\mathbf{v}_0' = \mathbf{o}$ and this time we want the tangent to be $\sqrt{3}$, that is, $(\sinh \widehat{r}_n \sin\theta_n)^2 = 3(\cosh \widehat{r}_n \sinh \widehat{r}_n (\cos\theta_n - 1))^2$ which gives

$$3\big(\cosh \widehat{r}_n\big)^2 = \left(\frac{\sin\theta_n}{\cos\theta_n - 1}\right)^2$$

We can take $\cosh \widehat{a}_n$ from the t coördinate $(\cosh \widehat{r}_n)^2(1 - \cos \widehat{r}_n) + \cos \widehat{r}_n$ of $\mathbf{v}_1'$. By substituting the above and some more algebra, similar to last time, we arrive at

$$\cosh \widehat{a}_n = \tfrac{1}{3}(1 + 4\cos\theta_n)$$

The horocyclic limiting values for 120°-angled ∞-gon sides are now

$$\begin{aligned}
\cosh \widehat{a}_\infty &= 5/3 \\
\sinh \widehat{a}_\infty &= 4/3 \\
\exp \widehat{a}_\infty &= 3 \\
\widehat{a}_\infty &= \log 3 \approx 1.0986122886681
\end{aligned}$$

H.7 Horocyclic polygons and discrete quadratics

The coördinates of the points on a circle on $\mathcal{H}$ evidently satisfy a quadratic relation: $x^2 + y^2 = s^2$ and hence $t^2 = s^2 + 1$, for some constant s (the sinh of the radius), in a coördinate frame centered on the circle's centre; Lorentz transformations, being linear, map it to another set of quadratics (this time involving t as well) in any other frame. Note in passing that for large enough t the equation for the ratio $x : y$ will have discriminant $D < 0$ as there are no points at infinity on the circle.

We saw in section H.5 that the coördinates of the points on a horocycle also satisfy a quadratic relation: $2x = y^2$ and $2t = y^2+2$ in a suitably aligned coördinate frame. Note this time, in keeping with the horocycle's "parabolic" nature, for $t \to \infty$ the equation for the ratio $x : y$ will have D approaching 0 as there is one point at infinity.

Presumably, points on the "hyperbolic" type of sections also satisfy a quadratic relation; the equation for $x : y$ given $t \to \infty$ will then have $D > 0$ because there are two points at infinity, but we will not need these sections here.

As all these curves are one-dimensional manifolds, it is possible to specify a point just by giving one parameter. For instance, on $x^2 + y^2 = s^2$ a choice of, say, x pinpoints two (or one or no) points. Much nicer of course would be a parameter that has *speed* 1 on the curve, that is, an a such that it also measures the distance along the curve. For the circle, we need to go to transcendental functions to describe $(x,\, y;\, t)$ as $(s\, \cos a,\, s\, \sin a;\, \sqrt{s^2+1})$. For the horocycle, we don't.

Recall the horocycle with $2x = y^2$, $2t = y^2 + 2$ was formed by intersecting $\mathcal{H}$ with the plane $t = x + 1$, and has a natural parametrisation with points expressed as $(x,\, y;\, t) = (\frac{1}{2}y^2,\, y;\, \frac{1}{2}y^2 + 1)$; let us call this point $\mathbf{h}(y)$. From $t = x + 1$ for all points on the curve we see that the straight-line distance, through the embedding $\mathbb{M}_3$, between points $(x,\, y;\, t)$ and $(x + \Delta x,\, y + \Delta y;\, t + \Delta t)$ both on the curve is $\sqrt{(\Delta x)^2 + (\Delta y)^2 - (\Delta t)^2} = |\Delta y|$ because $\Delta x = \Delta t$.

We have the curious situation that this straight-line distance $d(\mathbf{h}(y), \mathbf{h}(y + \Delta y))$ equals Δy, for finite Δy. In particular, $d(\mathbf{h}_0, \mathbf{h}_2)$ equals $d(\mathbf{h}_0, \mathbf{h}_1) + d(\mathbf{h}_1, \mathbf{h}_2)$ for all points on the curve, as long as $\mathbf{h}_1$ lies between $\mathbf{h}_0$ and $\mathbf{h}_2$.

We can *not* conclude the curve is a straight line though, as $\mathbb{M}_3$ is not a metric space[1]. None of this affects distances in $\mathcal{H}$ which is an honest-to-goodness metric space; for paths through $\mathbb{M}_3$ constrained to $\mathcal{H}$ we have the ordinary situation that the detour from $\mathbf{h}_0$ via $\mathbf{h}_1$ to $\mathbf{h}_2$ takes longer than the shortest path from $\mathbf{h}_0$ to $\mathbf{h}_2$. A horocycle really *is* curved in $\mathcal{H}$. As always with $\mathbb{M}_3$ and $\mathcal{H}$, straight-line

[1] A geodesic is a curve for which the variation δL of path length is zero. In metric spaces, and paths within "space-like" subspaces of $\mathbb{M}_3$, a geodesic is the *shortest* curve between two points. For paths with "time-like" tangents in $\mathbb{M}_3$ a geodesic is actually the *longest* curve between two points. The $t = x+1$ plane has geodesics where path length is neither a minimum nor a maximum but still with zero variation, just like a point of inflexion of a function can have derivative zero.

distance $2s$ through $\mathbb{M}_3$ (here $2s = \Delta y$) translates to distance $2\,\mathrm{arsinh}\, s$ in $\mathcal{H}$, and $\mathrm{arsinh}(s_0 + s_1) < \mathrm{arsinh}\, s_0 + \mathrm{arsinh}\, s_1$ for positive s'es.

What we *can* conclude is that in the limit for $\Delta y \to 0$ it is actually the distance along the curve that is measured by Δy (because any curve is nearly straight over short enough stretches); and then that Δy still measures path length over finite stretches (by integrating it).

Now let $\mathbf{v}_k$ for integer k be the successive vertices of a regular ∞-gon (with angles 90°, 120°, or any other angle) where the sides are just the right length to make the vertices fall on a horocycle. We saw all horocycles are congruent, so WLOG let the vertices lie on our canonical one with points $(\frac{1}{2}y^2,\, y;\, \frac{1}{2}y^2 + 1)$. Because the regular ∞-gon is congruent with itself shifted along by one vertex, straight-line distances S between any two successive vertices are equal. So the y coördinate of $\mathbf{v}_k$ is $\mathrm{S}k + \mathrm{K}$ for some K, and after a rotation that aligns $\mathbf{v}_0$ with the x-axis it is just $\mathrm{S}k$ and the coördinates of $\mathbf{v}_k$ are $(\frac{1}{2}\mathrm{S}^2k^2,\, \mathrm{S}k;\, \frac{1}{2}\mathrm{S}^2k^2 + 1)$. After an arbitrary Lorentz transformation, the expressions for x, y and t will all still be no worse than quadratic in k, an **integer**. Thus we have proved ■

Theorem: when $\mathbf{v}_k$ are the successive vertices of an ∞-gon, their Minkowski coordinates x, y and t are each given by a quadratic in the index.

Note that if the leading terms are $\mathrm{X}k^2$, $\mathrm{Y}k^2$ and $\mathrm{T}k^2$ we see straightaway that $\mathrm{X}^2 + \mathrm{Y}^2 = \mathrm{T}^2$ because the relative contribution of the 1 in $x^2 + y^2 = t^2 - 1$ must vanish as t increases without bound. Attempting to undo the arbitrary Lorentz transformation we could rotate around (0, 0; 1) by some Θ until $\mathrm{X} = \mathrm{T}$ and $\mathrm{Y} = 0$. Also, allowing non-integer k for a moment (to get all points on the horocycle) we can change the variable k by subtracting a constant K such that $k = 0$ occurs at the point of closest approach to (0, 0; 1) and the quadratic for t has no first degree term. Now we must merely be a shift A along the x-axis away from the canonical horocycle. Backtracking, the most general set of quadratics for an ∞-gon is

$$\begin{aligned} x\cos\Theta - y\sin\Theta &= \tfrac{1}{2}e^{\mathrm{A}}(\mathrm{S}k + \mathrm{K})^2 + \sinh\mathrm{A} \\ x\sin\Theta + y\cos\Theta &= \mathrm{S}k - \mathrm{K} \\ t &= \tfrac{1}{2}e^{\mathrm{A}}(\mathrm{S}k - \mathrm{K})^2 + \cosh\mathrm{A} \end{aligned} \tag{1}$$

where the e^{A} arises from $\cosh\mathrm{A} + \sinh\mathrm{A}$. Finally, if a **free graph** is embedded as a regular polytope (constant edge length, same angles at every vertex) the regions of surface bounded by its infinite "cycles" are all regular ∞-gons. If we choose the edge length just right they will lie on horocycles, and everything said here applies to them.

H.8 X_∞ on $\mathcal{H}$

In a previous section we found $\hat{a}_\infty$ as the side of the smallest infinite right-angled polygon, and therefore the smallest possible edge length we can use for a geometric representation of X_∞ on the hyperbolic plane, if we don't want edges to overlap. Using it rather than any $a > \hat{a}_\infty$ is not just parsimonious, it may also have nice properties.

So there is our analytical representation of X_∞: place one node at **o**, with four edges of length $\hat{a}_\infty$ sprouting from it along for instance the (projections of the) x and y axes. At the endpoints of every new edge place another node with from it three more edges of length $\hat{a}_\infty$ at $\frac{1}{4}(2\pi)$ and $\frac{1}{2}(2\pi)$ angles with the edge leading to it, and so on.

One nice property it has (albeit shared with similar graphs of larger edge length) is this: you always know which way is home, back to **o**. This is by no means obvious given the way curved spaces play fast and loose with any notions of direction: as we saw in the preceding section parallel transport doesn't preserve any *global* notion of orientation so we cannot add up the turns we took to reach some node **v**, conclude **o** lies "that way" and take of the four edges the one that points most nearly that way. However, it can be shown that for any node $\mathbf{v} \neq \mathbf{o}$ (representing some element $g \neq 1$ of the free group) there is always one of the four neighbours (representing gx, gy, gx^{-1}, gy^{-1}) whose distance to **o** is smaller, and moreover that this is the same neighbour that lies on the unique path from **v** back to **o** (in group terms, the unique one of gx, gy, gx^{-1}, gy^{-1} whose word is shorter than that of g). It is a consequence of the arguments used to prove the lemma later in this section.

The partial order $\mathbf{v} \preceq \mathbf{w}$ defined by **v** lying on a path from **o** to **w** (in terms of reduced words, the former being a front end of the latter) is *very* partial. For most pairs of nodes neither $\mathbf{v} \preceq \mathbf{w}$ nor $\mathbf{v} \succeq \mathbf{w}$. By contrast all nodes can be compared by distance to **o** (the worst that can happen is that some nodes are at the same distance) and the significance of the result above is that whenever $\mathbf{v} \preceq \mathbf{w}$ it agrees with $d(\mathbf{o}, \mathbf{v}) \leq d(\mathbf{o}, \mathbf{w})$. Just as induction arguments on word length have proved fruitful, a geometric notion of distance of group elements might be used for instance to classify the kinds of tilings that can occur, as described in appendix **G2**.

Another nice property (of the graph with edge lengths $\hat{a}_\infty$, and possibly some isolated other values) is number-theoretical. But to make it work it is easiest to rotate the graph by one-eight turn, so the initial edges are in the $x = \pm y$ directions. Without this turn, the neighbours of **o** are at $(\pm 2\sqrt{2},\ 0;\ 3)$ and $(0,\ \pm 2\sqrt{2};\ 3)$, and (as is easily checked) all nodes at $(p\sqrt{2},\ q\sqrt{2};\ t)$ for integer p, q and t. With the turn, using coördinates $(u,\ v;\ t)$ where

$$\begin{pmatrix} u \\ v \\ t \end{pmatrix} = \begin{pmatrix} (x+y)/\sqrt{2} \\ (x-y)/\sqrt{2} \\ t \end{pmatrix} \qquad \begin{pmatrix} x \\ y \\ t \end{pmatrix} = \begin{pmatrix} (u+v)/\sqrt{2} \\ (u-v)/\sqrt{2} \\ t \end{pmatrix}$$

(to be pedantic that's not a turn but a reflexion, and therefore easy to remember as it's the same both ways round) the step operators take the form

$$L_x^{\pm\hat{a}_\infty} = \begin{pmatrix} 2 & 1 & \pm 2 \\ 1 & 2 & \pm 2 \\ \pm 2 & \pm 2 & 3 \end{pmatrix} \qquad L_y^{\pm\hat{a}_\infty} = \begin{pmatrix} 2 & -1 & \pm 2 \\ -1 & 2 & \mp 2 \\ \pm 2 & \mp 2 & 3 \end{pmatrix}$$

and as we start off with $\mathbf{o}$ having integer coördinates (0, 0; 1) all our nodes have integer coördinates $(u, v; t)$. Interestingly, there is a converse.

Theorem: all triples of integers $(u, v; t)$ for which $t^2 - u^2 - v^2 = 1$ and $t > 0$ occur as a node somewhere in X_∞ (with edge length $\hat{a}_\infty$, when expressed in these coördinates). We will need the following

Lemma: if $\mathbf{t} = (u, v; t)$ with integer coördinates, $t^2 - u^2 - v^2$, $t > 0$, isn't already at $\mathbf{o}$, one of the four steps $L_x^{\pm\hat{a}_\infty}$ and $L_y^{\pm\hat{a}_\infty}$ is guaranteed to bring it closer to $\mathbf{o}$.

Of course this couldn't possibly be true for all points in $\mathcal{H}$, but we only need to prove it is true for those with *integer* coördinates (in other words, the small sliver of area hugging the x and y axes from which points would overshoot under application of a discrete step in the right direction will turn out to have no integer points other than $\mathbf{o}$). Because t goes down monotonously as the distance to $\mathbf{o}$ goes down we can use it as a proxy for distance in our proofs; the lemma says t goes down.

Proof of theorem, given the lemma: if at each stage we can find one of the four steps that makes t go down, it'll still be a positive integer, and we cannot have an infinite descent so we must after a finite number of steps arrive at $\mathbf{o}$ with $t = 1$. This incidentally also proves the earlier claim about being able to get home. Now if from any $\mathbf{t}$ a number of steps gets us to $\mathbf{o}$ then the reverse steps in reverse order get us from $\mathbf{o}$ to $\mathbf{t}$ proving the latter was in the graph. ■

Let $\mathbf{o} = \mathbf{t}_0, \mathbf{t}_1, \mathbf{t}_2, \ldots \mathbf{t}_k = \mathbf{t}$ be the nodes visited along the unique k-edge path from $\mathbf{o}$ to $\mathbf{t}$, and let L_i be the step matrix associated with the step from $\mathbf{t}_{i-1}$ to $\mathbf{t}_i$ ($L_x^{\pm\hat{a}_\infty}$ if it was in the $\pm x$ direction, $L_y^{\pm\hat{a}_\infty}$ if $\pm y$, in the coördinate frame we took with us parallel transported along all preceding steps). Now we cannot say that each $\mathbf{t}_i$ with $i > 0$ is obtained from $\mathbf{t}_{i-1}$ by multiplying it with L_i, because they only move $\mathbf{o}$ to its neighbours (attempting to build a graph by applying successive step matrices to $\mathbf{o}$ would make edges of unequal length, crossing eachother). Rather we must sneak up on it this way:

$$\begin{aligned} L_1^{-1}\mathbf{o} &= \mathbf{o}' \\ L_2^{-1}L_1^{-1}\mathbf{o} &= \mathbf{o}'' \\ &\cdots \\ L_k^{-1}\ldots L_2^{-1}L_1^{-1}\mathbf{o} &= \mathbf{o}^{[k]} \end{aligned} \tag{2}$$

where $\mathbf{o}^{[i]}$ are the coördinates of $\mathbf{o}$ in a coördinate frame centered on $\mathbf{t}_i$, with directions of axes parallel transported along the i intervening steps. Note that

the t-coördinate of $\mathbf{t}_i$ equals that of $\mathbf{o}^{[i]}$ because the distance between $\mathbf{o}$ and $\mathbf{t}_i$ is the same both ways and the relation between t and distance is the same in all coördinate frames (it is the direction, the ratio $x : y$ or $u : v$ that is unpredictably different). So we can couch a proof in terms of the t-coördinate of $\mathbf{o}^{[i]}$.

Proof of the lemma: let $\mathbf{o}^{[i]} = (u_i,\, v_i;\, t_i)$. I assert that if u_i and v_i are nonnegative, then the step $\mathbf{o}^{[i-1]} = L_x^{-\hat{a}_\infty}$ brings us closer to $\mathbf{o}$ (makes t go down). Consider first that if u_i or v_i is 0, t_i^2 must be a square that is only 1 higher than another integer square, so we are definitely at $\mathbf{o}$ with $t = 1$. Remains the case of u_i and v_i both positive, so at least 1 (this is where them being integer is used), and hence

$$\begin{aligned} 1 &< 2u_iv_i \\ u_i^2 + v_i^2 + 1 &< u_i^2 + 2u_iv_i + v_i^2 \\ t_i^2 &< (u_i + v_i)^2 \\ t_i &< u_i + v_i \\ t_i - u_i - v_i &< 0 \\ 3t_i - 2u_i - 2v_i &< t_i \\ t_{i-1} &< t_i \end{aligned} \tag{3}$$

If $(u_i,\, v_i)$ is in any of the other quadrants, the same argument with sign changes shows that then each time one of the other step matrices brings us closer to home. ■

The proof implies that the quadrant $\mathbf{o}$ appears to be in from here can be trusted to tell us (at least the first step of) the way home, the most recent step to undo.

Now imagine we are at some node $\mathbf{t}$ having come from $\mathbf{o}$ via some path V. We see $\mathbf{o}$ at $\mathbf{o}^{[k]}$ in (say) the Q-th quadrant and can reach it via the reverse path $\bar{V}$. Now given Q, for all possible positions in that quadrant where $\mathbf{o}$ could be it is true that the last step of V is of the flavor determined by Q. But that means that for all nodes in that quadrant the *first* step of the path $\bar{V}$ *to* that node corresponds to Q. Forgetting about $\mathbf{t}$ now we have the corollary that the quadrant we see a node in corresponds to the first step of the unique path to that node. The first step of the path to a node fixes what quadrant the nodes are in that we can reach via that step. In other words the four clover leaves of the graph attached to the four neighbours of $\mathbf{o}$ never overlap, which was already suggested (but not quite proven) by the fact that right-angled polygons with this edge length don't close.

We can now see from (2) how nodes in a path are built from the L_i:

$$\begin{aligned} L_1\mathbf{o} &= \mathbf{t}_1 \\ L_1L_2\mathbf{o} &= \mathbf{t}_2 \\ &\dots \\ L_1L_2\dots L_k\mathbf{o} &= \mathbf{t}_k \end{aligned}$$

For computational purposes one would keep not just the 3-component vector $\mathbf{t}_i$ at each stage, but the actual 9-component matrix product $L_1L_2\dots L_i$, and apply

successive L_i to the right. The matrix carries all the information to recover $\mathbf{t}_i$ (which is simply its t column).

We can also see how every node of the graph is the result of *some* sequence of L_i applied to $\mathbf{o}$, but piling successive factors L_i on the left isn't the way to walk a *path* because each application of an L_i doesn't in general map nodes to *adjacent* nodes. It does each time map the graph to itself. With this insight in hand we can also interpret (3) simply as referring to coördinates of a sequence of points $\mathbf{t}_i$ (each a node of X_∞ iff $\mathbf{t}_{i-1}$ is, i.e. all of them iff we end up at $\mathbf{o}$) each time closer to $\mathbf{o}$ as i goes down, without needing to stick to any path.

What values of t occur for nodes of the graph? First note that t is odd, u and v even: it is true in $\mathbf{o}$ and the step matrices perpetuate that; it is also a direct consequence of $t^2-u^2-v^2 \equiv 1 \pmod 4$. Let a **square-sum** be a sum of two squares (which may be 0^2); the prime decomposition of such a number, e.g. our u^2+v^2, parallels that of the Gaussian integer $u+vi$ and hence u^2+v^2 can contain any prime factor $q \equiv 3 \pmod 4$ only as even power q^{2k}. But $u^2+v^2 = t^2-1 = (t-1)(t+1)$. The latter two factors are even, so $\frac{1}{2}(t-1)$ and $\frac{1}{2}(t+1)$, differing by 1, are coprime. Now any q^{2k} in u^2+v^2 must go wholly into $\frac{1}{2}(t-1)$ or into $\frac{1}{2}(t+1)$. This makes $\frac{1}{2}(t-1)$ and $\frac{1}{2}(t+1)$ separately square-sums again, by well-known number theory results [HW79]. As this must hold for all t coördinates occurring in the graph we can easily prepare a list of candidate t: find the instances where two successive integers are *both* square-sums, only such numbers can be our $\frac{1}{2}(t-1)$ and $\frac{1}{2}(t+1)$:

$$\begin{array}{ccccccccccccccccccccccccccccc}
0 & \overline{1} & \overline{2} & \cdot & 4 & \overline{5} & \cdots & \overline{8} & \overline{9} & \overline{10} & \cdots & 13 & \cdots & \overline{16} & \overline{17} & \overline{18} & \cdot & 20 & \cdots\cdot & \overline{25} & \overline{26} & \cdots & 29 & \cdots & 32 & \cdot & 34 & \cdot & \overline{36} & \overline{37}... \\
 & 1 & 3 & & & 9 & & 17 & 19 & & & & & 33 & 35 & & & & & 51 & & & & & & & & 73 &
\end{array}$$

then their sum would be t. Conversely, if $\frac{1}{2}(t-1)$ and $\frac{1}{2}(t+1)$ are square-sums then t^2-1, being 4 times their product, is too, so u and v exist for which $u^2+v^2 = t^2-1$, i.e. all these candidates actually occur as t in the graph.

As to **how many u, v occur for each** t, let $t^2-1 = u^2+v^2$ (if $\neq 0$) decompose into real primes as

$$2^a \cdot \prod_r p_r^{b_r} \cdot \prod_s q_s^{2c_s}$$

where the primes $p_r \equiv 1 \pmod 4$ and the primes $q_s \equiv 3 \pmod 4$. Now there are

$$4\prod_r (b_r+1)$$

pairs (u, v) with the same u^2+v^2 because there are that many Gaussian integers $u+vi$. The b_r+1 arise because each p_r corresponds to two Gaussian primes $f+gi$ and $g+fi = i(f-gi)$; we can choose to use 0 through b_r of each flavor in $u+vi$. The factor 4 is just the choice in number of factors i we can give $u+vi$. Neither a nor c_s appear in the result because factors 2 and q_s^2 in u^2+v^2 match that many factors $1+i$ and q_s+0i in $u+vi$ in essentially only one way, as we have already accounted for an arbitrary number of factors i.

H.9 Y_∞ on $\mathcal{H}$

The regular ∞-gons appropriate for Y_∞ are those with angles 120° at each vertex, and we found the sides of the smallest such are $\widehat{a}_\infty = \log 3$. Again, we can take this as the smallest edge length that prevents the embedding of Y_∞ overlapping itself.

The (left-multiplying) matrix for a translation by a whole edge length in the $+x$ direction is easily constructed from cosh and sinh of $\widehat{a}_\infty$, and those for directions $\pm 120°$ from there by conjugating it with a rotation:

$$L_x^{+\widehat{a}_\infty} = \begin{pmatrix} +\frac{5}{3} & 0 & +\frac{4}{3} \\ 0 & +1 & 0 \\ +\frac{4}{3} & 0 & +\frac{5}{3} \end{pmatrix} \qquad \left.\begin{matrix} L_u^{+\widehat{a}_\infty} \\ L_v^{+\widehat{a}_\infty} \end{matrix}\right\} = \begin{pmatrix} +\frac{7}{6} & \mp\frac{1}{6}\sqrt{3} & -\frac{2}{3} \\ \mp\frac{1}{6}\sqrt{3} & +\frac{3}{2} & \pm\frac{2}{3}\sqrt{3} \\ -\frac{2}{3} & \pm\frac{2}{3}\sqrt{3} & +\frac{5}{3} \end{pmatrix} \qquad (4^+)$$

If the neighbours of a node lie in these three directions then, taking our coördinate system with us by parallel transport, the next node along will have neighbours in the $-x$ direction and directions $\pm 120°$ from there, for which the matrices are

$$L_x^{-\widehat{a}_\infty} = \begin{pmatrix} +\frac{5}{3} & 0 & -\frac{4}{3} \\ 0 & +1 & 0 \\ -\frac{4}{3} & 0 & +\frac{5}{3} \end{pmatrix} \qquad \left.\begin{matrix} L_u^{-\widehat{a}_\infty} \\ L_v^{-\widehat{a}_\infty} \end{matrix}\right\} = \begin{pmatrix} +\frac{7}{6} & \mp\frac{1}{6}\sqrt{3} & +\frac{2}{3} \\ \mp\frac{1}{6}\sqrt{3} & +\frac{3}{2} & \mp\frac{2}{3}\sqrt{3} \\ +\frac{2}{3} & \mp\frac{2}{3}\sqrt{3} & +\frac{5}{3} \end{pmatrix} \qquad (4^-)$$

Let's call the former kind **black** nodes and the latter kind **white** nodes. They alternate, so neighbours of black are white and vice versa.

Just as in the X_∞ section, these matrices do (in any one coördinate system) not transform each node's coördinates to those of its neighbours (rather, the neighbours of $AB...KL\mathrm{O}$ are $AB...KLM\mathrm{O}$, where M is one of the above). However, with some care they can also be used to map Y_∞ as a whole to itself, as we will see below (that will again map each node to *some* node, not in general a neighbour).

In the **node-centered** view of Y_∞ we again put one node O at (0, 0; 1). For definiteness let's make it black, so one neighbour is in the $+x$ direction at $(\frac{4}{3},\, 0;\, \frac{5}{3})$.

To get number theoretical results in terms of integer coördinates this time we should not look at Gaussian integers $a + bi$ with integer a and b but at what are sometimes known as Eisenstein integers (having been studied by Eisenstein and Jacobi), $a + b\,\omega$ with rational integer a and b, where $\omega = -\frac{1}{2} + \frac{i}{2}\sqrt{3} = e^{2\pi i/3}$. Note that $\omega^2 = \overline{\omega} = \omega - 1$ so the set is closed under addition and multiplication. To bring out the symmetry of the situation, note that $\omega^3 = 1$ so the three unit roots 1, ω and $\omega^2 = \overline{\omega} = -\frac{1}{2} - \frac{i}{2}\sqrt{3}$ are arranged 120° apart on the unit circle, which in turn implies $1 + \omega + \overline{\omega} = 0$ so any expression $a + b\,\omega + c\,\overline{\omega}$ has only two independent parameters in that it equals any other such expression obtained by adding the same value to each of a, b and c.

We can use other coördinates than a, b, c for an Eisenstein integer z, ones of the form $(x,\, u,\, v)$ where x is the Cartesian coördinate with that name, the real part

of $z = x + yi$, while u and v are the real parts of z/ω and $z/\overline{\omega}$ respectively. Here too there are only two independent parameters, but for a different reason: each z has a unique representation as $(x,\, u,\, v)$ but always $x + u + v = 0$. The nice thing about these coördinates is that rotation by 120° doesn't involve any factors $\frac{1}{2}\sqrt{3}$ but merely cyclic permutation of x, u and v.

Note that x, u and v are either integer or *halfodd* (half of an odd integer), for instance 1 appears as $(1, -\frac{1}{2}, -\frac{1}{2})$, ω as $(-\frac{1}{2}, 1, -\frac{1}{2})$ and $\overline{\omega}$ as $(-\frac{1}{2}, -\frac{1}{2}, 1)$. To convert between Cartesian and "Eisenstein" coördinates (keeping $x = x$):

$$y \;=\; \frac{u - v}{\sqrt{3}} \qquad\qquad \begin{aligned} u \;&=\; -\frac{x}{2} + \frac{y}{2}\sqrt{3} \\ v \;&=\; -\frac{x}{2} - \frac{y}{2}\sqrt{3} \end{aligned}$$

from which $x^2 + u^2 + v^2 = \frac{3}{2}(x^2 + y^2)$, so the modulus is given by

$$|(x,\, u,\, v)|^2 \;=\; \tfrac{2}{3}(x^2 + u^2 + v^2)$$

Because of the redundancy in coördinates whereby $x = -u - v$ it can also be expressed in any two of them,

$$|(x,\, u,\, v)|^2 \;=\; \tfrac{4}{3}(u^2 + uv + v^2)$$

and its permutations. Writing $(x,\, y;\, t)$ now as $(x,\, u,\, v;\, t)$ the fact that it lies on $\mathcal{H}$ is expressed by

$$3t^2 - 2(x^2 + u^2 + v^2) \;=\; 3$$

We already saw that in Y_∞ the neighbour of $\mathbf{o}$ in the x direction lies at $(x,\, y;\, t) = (\frac{4}{3},\, 0;\, \frac{5}{3})$ which in our new coördinates is $(\frac{4}{3},\, -\frac{2}{3},\, -\frac{2}{3};\, \frac{5}{3})$. The other two neighbours lie at 120° (cyclic permutations of x, u and v). The values that occur for the coordinates of nodes of Y_∞ will turn out never to need a 2 in the denominator, in fact the numerators of the spatial (non-t) coördinates have factors 2 to spare.

We do get factors 3 in the denominator. If we were to start wandering off along the plane in any of the six ± 1, $\pm\omega$, $\pm\overline{\omega}$ directions, using our new $\widehat{a}_\infty$ step size, those factors would soon proliferate. Even sticking to the x-axis we pick up an extra factor $\frac{1}{3}$ at each step. But Y_∞ isn't built like that. From $\mathbf{o}$ we can go in the 1, ω or $\overline{\omega}$ direction, but at the next node we must go in the -1, $-\omega$ or $-\overline{\omega}$ direction. And these two sets of directions keep alternating along any possible path. And it turns out that walking that way it never gets worse than one factor 3 in the denominator.

Theorem Y1: $3x$, $3u$, $3v$ and $3t$ remain integers for all nodes of Y_∞. This can be proven from first principles by congrences (omitted here for reasons of space), alternatively as a consequence of results we will turn to next.

Theorem Y2: if t, x, u, v are integers for which $3t^2 - 2x^2 - 2u^2 - 2v^2 = 27$ and $t > 0$ and $x + u + v = 0$ then $(x/3,\, u/3,\, v/3;\, t/3)$ occurs somewhere in Y_∞. Proof further down.

Consider Y_∞ in **edge-centered** orientation, starting with one edge **np** along the x-axis, with its centre at $\mathbf{o} = (0,\, 0;\, 1)$. Now $\cosh \widehat{a}_\infty = 5/3$ and $\sinh \widehat{a}_\infty = 4/3$ imply that $\cosh \frac{1}{2}\widehat{a}_\infty = 2/\sqrt{3}$ and $\sinh \frac{1}{2}\widehat{a}_\infty = 1/\sqrt{3}$, so we can put **n** at $(-1,\, 0;\, 2)/\sqrt{3}$ and **p** at $(+1,\, 0;\, 2)/\sqrt{3}$, where the former is black and the latter white.

Theorem H1: In this embedding, node coördinates are integers divided by $\sqrt{3}$.

All coördinates: x, y and t. At first this sounds impossible, seeing position of $\sqrt{3}$'s in matrices (4^+) and (4^-). But consider the form of $H \cdot \mathbf{o} = \mathbf{p}$ (where $H = X^{\widehat{a}_\infty/2}$)

$$\begin{pmatrix} 2/\sqrt{3} & 0 & 1/\sqrt{3} \\ 0 & 1 & 0 \\ 1/\sqrt{3} & 0 & 2/\sqrt{3} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{3} \\ 0 \\ 2/\sqrt{3} \end{pmatrix}$$

which (by $k/\sqrt{3} = \frac{k}{3}\sqrt{3}$) is of the general pattern

$$\begin{pmatrix} \#\sqrt{3} & \# & \#\sqrt{3} \\ \#\sqrt{3} & \# & \#\sqrt{3} \\ \#\sqrt{3} & \# & \#\sqrt{3} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \#\sqrt{3} \\ \#\sqrt{3} \\ \#\sqrt{3} \end{pmatrix}$$

where # denotes arbitrary rationals. Successive neighbouring nodes $\mathbf{p}_i$ on any path from **p**, $HA \cdot \mathbf{o} = \mathbf{p}_1$, $HAB \cdot \mathbf{o} = \mathbf{p}_2$, $HABC \cdot \mathbf{o} = \mathbf{p}_2 \ldots$ will be formed by matrix multiplications of the same pattern provided HA, HAB, $HABC \ldots$ are of the same shape as H. And indeed A, B, $C \ldots$ being those in (4^+) and (4^-) guarantees that H's shape is perpetuated in $H \cdot A = HA$, then in $HA \cdot B = HAB$, and so on:

$$\begin{pmatrix} \#\sqrt{3} & \# & \#\sqrt{3} \\ \#\sqrt{3} & \# & \#\sqrt{3} \\ \#\sqrt{3} & \# & \#\sqrt{3} \end{pmatrix} \begin{pmatrix} \# & \#\sqrt{3} & \# \\ \#\sqrt{3} & \# & \#\sqrt{3} \\ \# & \#\sqrt{3} & \# \end{pmatrix} = \begin{pmatrix} \#\sqrt{3} & \# & \#\sqrt{3} \\ \#\sqrt{3} & \# & \#\sqrt{3} \\ \#\sqrt{3} & \# & \#\sqrt{3} \end{pmatrix}$$

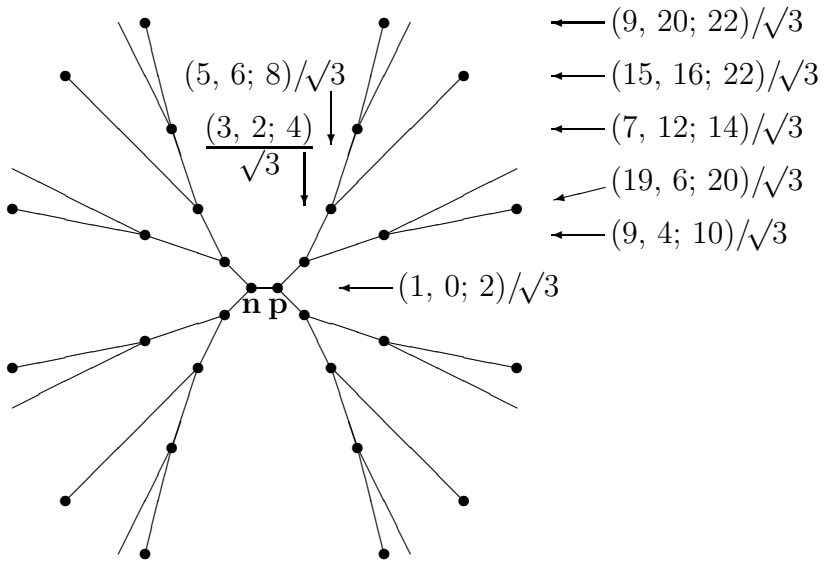
To prove we actually keep getting *integers*$/\sqrt{3}$, we need to be more specific:

Lemma H1$^+$: In this embedding of Y_∞, nodes have coördinates $(x,\, y;\, t)/\sqrt{3}$ where
(a) x, y and t are integer,
(b) x odd, y and t even,
(c) $t + y \equiv 2 \pmod 4$,
(d) for black nodes $x \equiv -1 \pmod 4$ and for white $x \equiv +1 \pmod 4$.

Rather than proving these constraints perpetuate themselves from $H\mathbf{o}$ under insertion of matrix factors A, B, C, ... it is much easier to apply matrix factors to the vectors themselves. This should, just like above with X_∞, map each node to *some* node (not in general a neighbouring one). We must be careful with the way the black and white node populations have edges in different directions though.

Let $N = HRH$ with H as above (moving **o** to the right by half an edge length) and R a counterclockwise rotation by 60° around $(0,\, 0;\, 1)$. Now N moves **n** to **p** and all the other nodes in the ∞-gon on the "North" side of **np** along by one step as well. Other nodes are moved by larger amounts. Of course, $N^{-1} = H^{-1}R^{-1}H^{-1}$ moves

the other way. Similarly, $S = HR^{-1}H$ moves **n** to **p** and all the other nodes in the ∞-gon on the "South" side of **np** along by one step as well, and $S^{-1} = H^{-1}RH^{-1}$ moves the other way.



The first few nodes of Y_∞ *in edge-centered view. Each node at* $(x, y; t)$ *is displayed at* (x, y)*. Edges are approximated by straight lines on the page.*

These halfstep-turn-halfstep matrices are

$$N = \begin{pmatrix} 1 & -1 & +1 \\ +1 & \frac{1}{2} & +\frac{1}{2} \\ +1 & -\frac{1}{2} & \frac{3}{2} \end{pmatrix} \quad N^{-1} = \begin{pmatrix} 1 & +1 & -1 \\ -1 & \frac{1}{2} & +\frac{1}{2} \\ -1 & -\frac{1}{2} & \frac{3}{2} \end{pmatrix}$$

$$S = \begin{pmatrix} 1 & +1 & +1 \\ -1 & \frac{1}{2} & -\frac{1}{2} \\ +1 & +\frac{1}{2} & \frac{3}{2} \end{pmatrix} \quad S^{-1} = \begin{pmatrix} 1 & -1 & -1 \\ +1 & \frac{1}{2} & -\frac{1}{2} \\ -1 & +\frac{1}{2} & \frac{3}{2} \end{pmatrix}$$

Now, keeping in mind these matrices are multiplied directly into node vectors (on their left) and map black to white and vice versa, the lemma follows: **n** and **p** satisfy its specifications, and these matrices perpetuate them to every node we can map to **n** or **p**. To see that we can crawl up to any node, consider that rolling along the North ∞-gon we don't just reach all its nodes, but also put all neighbouring ∞-gons in turn at the place of the South ∞-gon. Then we can do the same with neighbour ∞-gons of neighbour ∞-gons, and so on. ■ And so theorem H1. ■

We are now in a position to prove theorem Y1 from H1: applying H to $(x, y; t)/\sqrt{3}$ gives $(\frac{2}{3}x + \frac{1}{3}t, y/\sqrt{3}; \frac{2}{3}t + \frac{1}{3}x)$ where the x- and t-coördinates are already in the required form (one-third of an integer). The u- and v-coördinates are easily found to be $\frac{1}{3}x \mp \frac{1}{2}y + \frac{1}{6}t$ which (by y and t being even) are also of that form. ■

Theorem H2: if integers x (odd), y and t (even) satisfy $t^2 - x^2 - y^2 = 3$, the node $(x, y; t)/\sqrt{3}$ occurs somewhere in Y_∞ embedded edge-centered.

Theorems Y2 and H2 are related like Y1 and H1, by translating coördinates. This time though Y2 is easier to prove than H2, with **o** being one of the nodes there.

The proof goes much the same as for the corresponding theorem about X_∞, with 120° pie slices replacing quadrants. Again, we apply the step matrices, now (4^+) or (4^-) depending on node color, to node vectors and prove there is one that decreases the t-coördinate. This time the interpretation cannot be "if there is a node at $\mathbf{v}$ there must be *some* node at $A\mathbf{v}$ (because of Y_∞'s two node populations we need N and S rather than the step matrices for that) but we can still use the path interpretation where $\mathbf{o}^i$ is the position of the original **o** in a coördinate frame parallel transported along the path to $\mathbf{v}_i$ (each time $\mathbf{o}^i$ has the same t-coördinate in that frame as $\mathbf{v}_i$ has in the original frame, because distance from **o** to $\mathbf{v}_i$ equals that from $\mathbf{v}_i$ to **o** and distance to the origin is a function of t only), and use A for $\mathbf{o}^i = A\mathbf{o}^{i-1}$ so stepping back $\mathbf{o}^{i-1} = A^{-1}\mathbf{o}^i$.

The crucial part of the proof doesn't need Eisenstein integer $3(x,\, u,\, v;\, t)$ but can be stated in ordinary $(x,\, y;\, t)$ (still with integer $3x$ and $3t$, with $3y$ containing a $\sqrt{3}$) because we can WLOG consider the case where $\mathbf{o}^i$ lies in the 120° pie slice centered on the $+x$ axis (in the case of a black node, with obvious sign changes for white nodes). This means $\mathbf{o}^i = (x_i,\, y_i;\, t_i)$ with

$$x < 0 \qquad \text{and} \qquad y_i^2 \le 3x_i^2 \tag{5}$$

Using (4^-) as inverse of (4^+), $t_{i-1} = \frac{5}{3}t_i - \frac{4}{3}x$.

Note theorem Y1 implies $X = 3x_i$ and $Y = y_i\sqrt{3}$ are integers. The latter also follows from lemma H1$^+$ (the node- and edge-centered y-coördinates are the same), and that additionally tells us Y is even, and (applying H) that X is odd.

Now (5) says $Y^2 \le X^2$. We can only have $Y^2 = X^2$ (so $t_i^2 = 4x_i^2{+}1$) for $y_i = x_i = 0$, the centre we're trying to prove we will reach; $Y^2 = 3X^2 - 1$ is impossible (mod 3) where squares are never -1, $Y^2 = 3X^2 - 2$ is impossible (mod 4) for odd X, and $Y^2 = 3X^2 - 3$, a square $3(x+1)(x-1)$, implies $X = 1$ and $Y = 3$, with no solution for t_i. Now we have $Y^2 < X^2 - 3$ i.e.

$$\begin{aligned} y_i^2 + 1 &< 3x_i^2 \\ x_i^2 + y_i^2 + 1 &< 4x_i^2 \\ t_i^2 &< 4x_i^2 \\ t_i &< 2x_i \\ \tfrac{2}{3}t_i - \tfrac{4}{3}x_i &< 0 \\ \tfrac{5}{3}t_i - \tfrac{4}{3}x_i &< t_i \\ t_{i-1} &< t_i \end{aligned}$$

which finishes the proof of Y2. ■ And with it H2: note Y1 and H1 are of the form "node ⇒ integer" and Y2 and H2 of the form "integer ⇒ node" so the argument "edge-centered integer ⇒ node-centered integer" used above to show H1 ⇒ Y1 can be used unchanged to show Y2 ⇒ H2. ■

In case you wondered what happens with the Diophantine equation $t^2-x^2-y^2=3$ without parity restrictions: if all three are odd $t^2-x^2-y^2 \equiv 7 \pmod 8$ so there is no solution. Else two must be even; if they are t and y we get edge-centered Y_∞ as above; if t and x the same rotated by 90°; if x and y then $t^2-x^2-y^2 \equiv 1 \pmod 4$ so again no solution.

Corollary: from any node **v** (in node-centered view), other than **o**, *that* one of **v**'s three neighbours that lies on the path from **v** to **o** is the one that lies closest to **o** by ordinary distance along the plane (ignoring the graph). Proof: implicit in the construction used. As with X_∞, this depends critically on us having used a long enough edge length where no parts of the graph overlap.

Corollary: from any node **v** (in node-centered view), other than **o**, there is exactly one neighbour of **v** closer to **o** than **v** is. We just saw such a neighbour, and there can't be more than one because then the same argument would exhibit two paths to **o**, and Y_∞ has no cycles.

Let **p**, **q** and **r** be the neighbours of **n**. Treating **n** in the rôle of node **o** above, we see the "half of all nodes" in edge-centered view from which paths reach **p** before **n** are the "one-third of nodes" in node-centered view from which paths to **n** go via **p** rather than **q** or **r**.

Theorem H3: the nodes **v** (in edge-centered view) from which a path to edge **np** reaches **p** first are precisely the nodes with positive x-coordinate. This is obvious from the picture above, but still worth proving. No **v** has $x = 0$ in edge-centered view (as $x\sqrt{3}$ is odd), so all lie in the half-planes N and P separated by the y axis. Change to a frame centered on node **n** (or **p**) and rename **v** to $\mathbf{o}^i$, we saw **p** (or **n**) lies on the path to $\mathbf{o}^i$ iff $\mathbf{o}^i$ lies in the 120° pie slice to the right (left). It is easy to check P (or N) lies in that slice (projecting $(x, y; t)$ to plane (x, y) the y-axis of edge-centered view becomes one branch of a hyperbola when centered on **n** or **p**, and the pie slice bounds are its asymptotes). ■

H.10 Projections

If we wish to **project** geometric figures from $\mathcal{H}$ onto a flat page we can do so simply by suppressing t and plotting points at (x, y); no information is lost as t is constrained to be $\sqrt{(x^2+y^2+1)}$ on $\mathcal{H}$. This is in the same spirit as projecting (a hemisphere of) $\mathcal{S}$ by suppressing z (a kind of photographic mapping with perspective as from infinite distance). The result stil depends of course on which point we're looking straight down on, which point appears at the coördinates we called **o**. Radial distances from it are progressively squashed the further we get from **o** in the case of $\mathcal{S}$ but progressively stretched further from **o** in the case of $\mathcal{H}$. Tangential distances (to circles around **o**) are reproduced faithfully in both cases.

The **Minkowski coördinates** x, y and t used here are the most convenient for calculations, but historically **Poincaré coördinates** (ξ, η) have more commonly

been used for the hyperbolic plane. They too are a projection of 3-dimensional Minkowski coördinates, the view with perspective from the point $(0,\, 0;\, -1)$ projected to a plane $\perp$ the t axis; the inside of the unit disk on that plane is the image of all of $\mathcal{H}$. Poincaré coördinates are *conformal* i.e. they preserve all angles. Lines on the hyperbolic planes are projected to circle arcs perpendicular to the rim (the unit circle) at both their endpoints, circles to interior circles, horocycles to circles tangent to the rim, and "hyperbolas" to other circle arcs. If we project from $(0,\, 0;\, 0)$ instead we get **Klein–Beltrami coördinates** $(\mathrm{X},\, \mathrm{Y})$, their unique property is that lines are projected to straight lines.

To convert between Minkowski and Poincaré coördinates:

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \Big/ (1+t) \qquad \begin{pmatrix} t \\ x \\ y \end{pmatrix} = \begin{pmatrix} 1+\xi^2+\eta^2 \\ 2\xi \\ 2\eta \end{pmatrix} \Big/ (1-\xi^2-\eta^2)$$

and between Minkowski and Klein–Beltrami ones:

$$\begin{pmatrix} \mathrm{X} \\ \mathrm{Y} \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \Big/ t \qquad \begin{pmatrix} t \\ x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ \mathrm{X} \\ \mathrm{Y} \end{pmatrix} \Big/ \sqrt{1-\mathrm{X}^2-\mathrm{Y}^2}$$

Clearly $x : y = \xi : \eta = \mathrm{X} : \mathrm{Y}$ so direction from $\mathbf{o}$ is represented the same way in these three coördinate systems. They only differ in the way they treat radial distance r from $\mathbf{o}$, which they represent as follows:

- $\sqrt{x^2+y^2} = \sinh r$ for Minkowski,
- $\sqrt{\xi^2+\eta^2} = \tanh(r/2)$ for Poincaré,
- $\sqrt{\mathrm{X}^2+\mathrm{Y}^2} = \tanh r$ for Klein–Beltrami;

which shows how the latter two project the entire hyperbolic plane inside the unit circle (as in M. C. Escher's *Circle Limit* series of etchings, which use Poincaré coordinates). This is also obvious geometrically, as $\mathcal{H}$ is a bowl that hugs the **light cone** $t^2-x^2-y^2=0$ asymptotically and we're projecting from a point below the bowl. By contrast $(x,\, y)$, which doesn't project from a point but by lines parallel to the t-axis, paints the image onto the entire plane.

To show that $\mathcal{H}$ as defined at the start of this appendix really *is* the hyperbolic plane we know and love, consider there is up to isomorphism only one connected surface of constant negative curvature where lines don't intersect themselves. Now $\mathcal{H}$ has negative curvature somewhere (as a triangle with angle deficit can be exhibited), and has the same curvature everywhere because dot products (and hence lengths of vectors and angles between them) are invariant under Lorentz transformations. So it has constant negative curvature and, being connected and its lines being unbounded and non-self-intersecting, is isomorphic to the hyperbolic plane.

References

[Wey21] Hermann Weyl, *Raum Zeit Materie* 4th ed. 1921; translated as *Space Time Matter* by Henry L. Brose, Methuen 1950; Dover 1952 with ISBN 0 486 60267 2

[Wey31] Hermann Weyl, *Gruppentheorie und Quantenmechanik*, Leipzig 1928, 2nd ed. 1931; translated as *The Theory of Groups and Quantum Mechanics* by H. P. Robertson, Methuen 1931; Dover 1950 with ISBN 0 486 60269 9

[CM76] M. Carmeli & S. Malin, *Representations of the Rotation and Lorentz Groups: an Introduction*, Lecture Notes in Pure & Appl. Math. **16**, Marcel Dekker 1976, ISBN 0 8247 6449 8

[Cla79] C. Clarke, *Elementary General Relativity*, Edward Arnold 1979, ISBN 0 7131 2763 5

[HW79] G. H. Hardy & W. M. Wright, *An Introduction to the Theory of Numbers*, Oxford Univ. Press 1938, 5th ed. 1979, ISBN 0 19 85317 $^{0\,2}_{1\,0}$

[MCl94] John McCleary, *Geometry from a differentiable viewpoint*, Camb. Univ. Pr. 1994, ISBN 0 521 42480 1