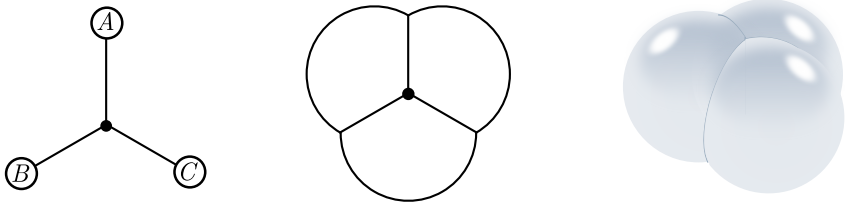
Why is a soap bubble like a railway?

David Wakeham[1]

Department of Physics and Astronomy
University of British Columbia
Vancouver, BC V6T 1Z1, Canada

Abstract

At a certain infamous tea party, the Mad Hatter posed the riddle: why is a raven like a writing-desk? We do not answer this question. Instead, we solve a related nonsense query: why is a soap bubble like a railway? The answer is that both *minimize over combinatorial structures*. We give a self-contained introduction to this genre of minimization problem, starting with minimal networks on the Euclidean plane and ending with close-packed structures for three-dimensional foams. Along the way, we touch on algorithms and complexity, graph theory, curvature, physics, space-filling polyhedra, and bees from other dimensions. The only prerequisites are high school geometry and a spirit of adventure.



[1] david.a.wakeham@gmail.com

Contents

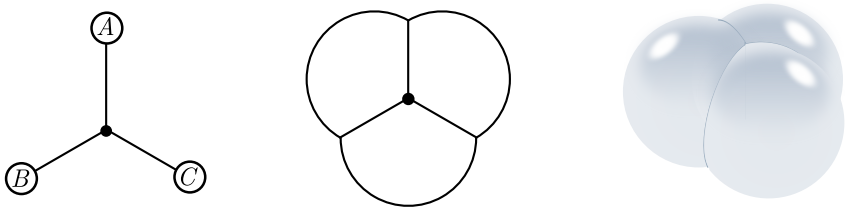
Acknowledgments

These notes were inspired by conversations with Rafael Haenel, Pedro Lopes and Haris Amiri, and the hands-on Steiner tree and soap bubble activity organized by Rafael and Pedro for Diversifying Talent in Quantum Computing. I would particularly like to thank Rafael and Pedro for encouragement and feedback at various stages of the draft, and the students of the UBC Physics Circle, where some of the material on Steiner trees was first presented. As always, their enthusiasm spurred me on. I am supported by an International Doctoral Fellowship from the University of British Columbia.

1 Introduction

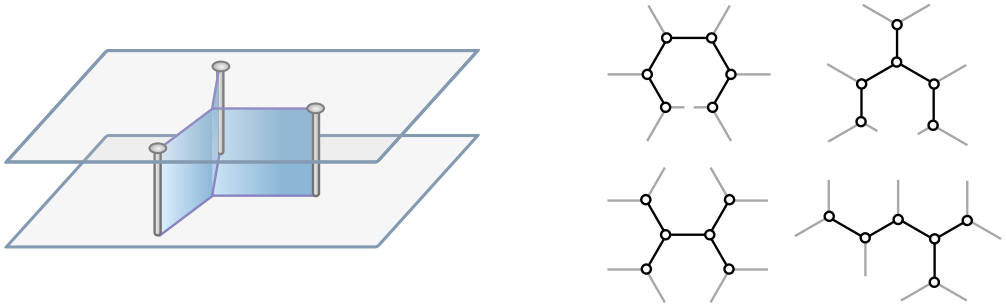
> The Hatter opened his eyes very wide on hearing this; but all he said was, "Why is a raven like a writing-desk?" "Come, we shall have some fun now!" thought Alice. "I'm glad they've begun asking riddles.—I believe I can guess that," she added aloud. "Do you mean that you think you can find out the answer to it?" said the March Hare. "Exactly so," said Alice.
>
> *Lewis Carroll*

Why is a soap bubble like a railway? I believe we can guess that. Suppose we are designing a rail network which joins three cities. If stations are cheap, our biggest expense will be rail itself, and to minimize cost we should make the network as short as possible. For three cities A, B and C, the cheapest network typically looks like the example below left. In addition to stations at each city, we add a *hub* station in the middle to minimize length. For a general triangle of cities, hub placement follows a simple rule: outgoing rail lines are equally spaced, fanning out at angles of 120°.



A two-dimensional bubble, with walls made of soapy water, solves the same problem. The molecules in the water are attracted to each other, creating surface tension. Tension pulls the surface taut, and length is once again minimized, due to the budgetary constraints of Nature itself. Like rail lines, bubble walls converge at junctions three at a time, separated by 120°. The rule even works for the soapy walls of a *three-dimensional* bubble.

Of course, rail networks in the real world connect many cities, and the problem is more complicated. But it remains true that for the cheapest network, any time we introduce a hub it must have three rail lines emerge at angles of 120°, with the same going for multiple bubbles. This makes the connection between soap bubbles and railways *useful*: by drilling screws through plexiglass, we can make a soap bubble computer, and solve network planning problems with soapy water!

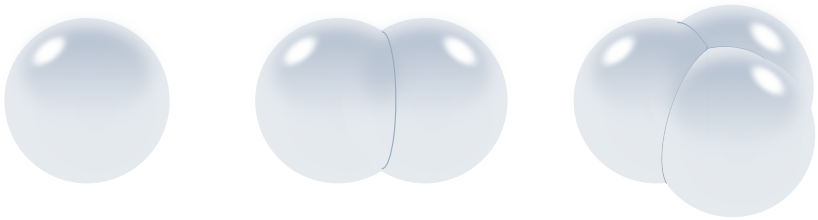


While soap bubbles can find small railways almost instantaneously, there is a deep but subtle reason they aren't useful for finding the best way to connect every city in North America. In principle, we just place a screw at the position of every city, dip into soapy water, and withdraw.

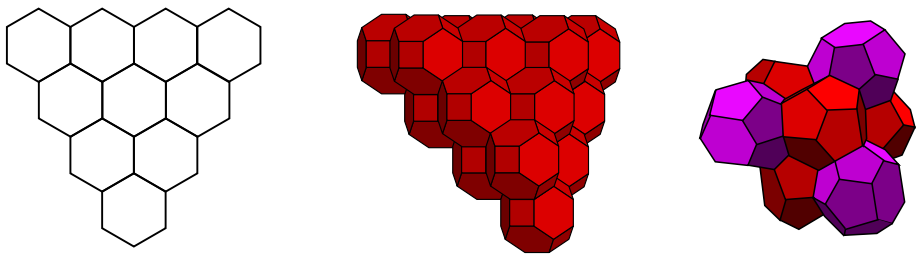
In practice, it will take longer than the age of the universe for the bubbles to settle down into something useful! The problem is just too hard. Although we know what hub stations look like *locally*—a trident of three rail lines—there are many different ways to arrange a given number of hubs. We show a few examples for six hubs above right. As the number of hubs gets large, there are so many that *no physical mechanism*, soap bubbles or quantum computers or positronic brains, can quickly search them all to find the shortest candidate. This is called the *NP Hardness Assumption*. It is ultimately a physical principle, because it makes predictions about the behaviour of physical objects which compute, such as soap bubbles.

If it takes arbitrarily large amounts of computing power to find the cheapest network, it is no longer cheap. *Approximate* answers are preferable if they can be found quickly, and we will give two methods for rapid (but suboptimal) rail planning below. But bubbles still hold surprises. Once we remove the plexiglass and screws, genuine bubbles are free to form, each cell enclosing some fixed volume. The laws of physics will now try to minimize the total area of the cell surfaces, so in a sense, the forms flowing out of the bubble blower are *conjectures* made by Nature about the best (i.e. smallest-area) way to enclose some air pockets.

For example, the humble spherical bubble harbours the following conjecture: of all surfaces of fixed volume V, the sphere has the smallest area. This is called the *isoperimetric inequality*. But surprisingly little is known about more bubbles. While the symmetric *double bubble* shown below is the most economic way to enclose two equal volumes, no one knows if the symmetric triple bubble is optimal for three equal volumes.



Our comparative ignorance of bubbles will not stop us launching, undaunted, into the problem of partitioning not two, not three, but an *infinite number* of equal volumes. As a warm-up, we can consider the problem for two-dimensional bubbles. We will show that in a large foam of bubbles, the 120° rule means that most cells are hexagonal. This helps explain why bees prefer a hexagonal lattice for building their hives. They are trying not to waste wax! In fact, the hexagonal tessellation, where each hexagon is identical, provably requires the least amount of wax per equal volume cell.



In three dimensions, more things can happen. In addition to the 120° rule, we need a few other rules for bubbles which together make up *Plateau's laws*. Unlike two dimensions, these laws don't

tell us precisely how many faces a bubble, but they do give some constraints. We can use these constraints to eliminate all but one space-filling pattern, the *Kelvin structure* (above middle), made from pruned octahedra. Surprisingly, this is *not* the best way to separate an infinite number of cells of equal volume in three dimensions. There is a mutant tessellation made from weaving together two different equivoluminous shapes called the *Weaire-Phelan structure*, shown above right. No one knows if there is a way to beat it. Although there are no four-dimensional bees to store their honey in Weaire and Phelan's cells, Nature uses this structure to make superconductors.

These notes represent a maximalist take on minimization. Below, you will encounter graph theory, algorithms and complexity, space-filling patterns, curvature, non-Euclidean bees, physics, chemistry and a plethora of unsolved mathematical problems. Come, we shall have some fun now!

1.1 Overview

In §2, we start our study of minimization with the problem of minimal networks on the triangle. In §2.1, we analyze the equilateral triangle using symmetry, and argue that a hub should be placed in the center. We deform this solution in §2.2, and give some loose arguments that the hub collides with a vertex when an internal angle opens to 120°. This is generalized in §2.3 to give the 120° rule for general minimal networks. Finally, in §2.4, we give a brief history of minimal networks.

In §3, we use tools from graph theory to take the 120° rule, which is a local constraint, and turn it into a global constraint on the structure of the network. Trees and their basic properties are introduced in §3.1, and exploited in §3.2 to put a bound on the maximum number of hubs. This allows us to solve some small but nontrivial networks. In §3.3, the bound is turned into a rough argument for the computational hardness of finding minimal networks, while §3.4 provides some easily computable alternatives, namely the minimal spanning tree and Steiner insertion heuristic.

With §4, we move laterally into the realm of soap bubbles. We build soap bubble computers in §4.1 to solve our minimal network problems, where our computational hardness results resurface as predictions about physics. In §4.2, we introduce Euler's formula and apply it to bubble networks, while in §4.3, we make a simple scaling argument that most bubbles in a large foam are hexagonal. This is related to the fact that bees build hexagonal hives, and the *honeycomb theorem* that bees know the best way to partition the plane into cells of equal size. The *planar bubble configuration problem* make its appearance in §4.4, along with a heuristic proof of the isoperimetric inequality.

The last section, §5, considers three-dimensional bubbles. After defining mean curvature in §5.1, we state Plateau's laws in §5.2, motivating them by analogy with bubble networks. With §5.3, we describe the three-dimensional bubble configuration problem and Plateau's problem for wireframes and bubble blowers. Finally, in §5.4 we generalize Euler's formula to study network constraints on three-dimensional foams, and conclude with a whirlwind tour of regular tessellations of space, the Kelvin problem, the Weiare-Phelan surprise, and the related tetrahedrally closed-packed structures.

Prerequisites. The only prerequisites are high school algebra, geometry and a little physics. These notes should therefore be suitable for high school and undergraduate mathematics enrichment. We will often resort to heuristics, pictures, and physical intuition as a result, which may not appeal to some readers. But the price of admission is lower, and we hope the rides are no less fun.

Exercises. There are around 40 problems of varying length and difficulty. Many of these are essential and used subsequently in the text. I hope this is not a weakness, but rather than an incentive to solve them! Difficult exercises are labelled with a mountain (⛰), or an icy mountain (🏔) in the case of greater abstraction or required background. Longer labels (🗻 and 🗾) inflect for length. For solutions, please contact me by email.

2 Trains and triangles

Suppose we want to join up three towns A, B and C by rail. Building railways is expensive, since we not only need to design and build the rail itself, but acquire the land beneath it. In contrast, stations can be reasonably cheap: we just slap together some sidings, a platform, and a bench or two. To minimize cost we should make the total length of the rail network as short as possible.

If the railway lets us travel from one town to any other, we say that the rail network is *connected*. A *hub* is a station built solely to connect rails. A connected rail network of minimal length is called a *minimal network* or *Steiner tree*.[2] Two possible networks for the three towns are shown in Fig. 1. The "triangle" network is built from two sides of the triangle formed by the three towns, while the "trident" network adds a hub (also called a *Steiner point*) in the middle.

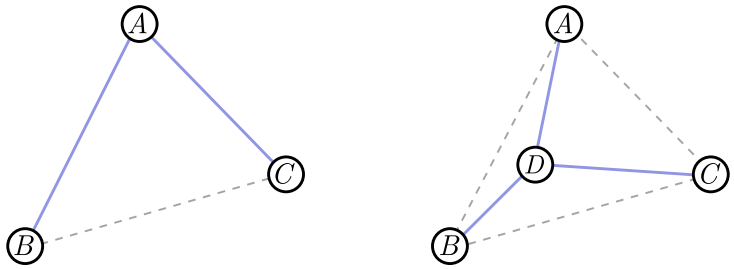


Figure 1: Rail networks (triangle and trident) connecting three towns.

Exercise 2.1. *Choosing sides.*

Suppose A, B and C are separated by distances AB, AC and BC. A triangular network consists of two sides of the triangle. Which ones should we choose?

Exercise 2.2. *Triangle or trident?*

In Fig. 1, we have two networks connecting the same towns: two sides of the triangle, and the trident-shaped network with a hub D in the middle. Check the trident is shorter. *Hint.* Measure lengths with a ruler. Simple but it works!

Already, there is a surprise. Although the simplest network consists of two sides of the triangle, this is not minimal, since (to spoil Exercise 2.2) the trident in Fig. 1 is shorter. We can go further and optimize the placement of the hub D. The case for a general triangle is tricky, but we can build most of the intuition we need by focusing on the special case of an *equilateral* triangle.

2.1 Equilateral triangles

Suppose A, B and C sit on the corners of an equilateral triangle of side length d, as in Fig. 2. The triangular network has total length $L_\Lambda = 2d$. For the trident network on the right, we place the

[2] We will see where the term "tree" comes from in §3.

hub D directly in the middle. Let's trade our engineering for math hats, and find the length of the trident network using trigonometry.

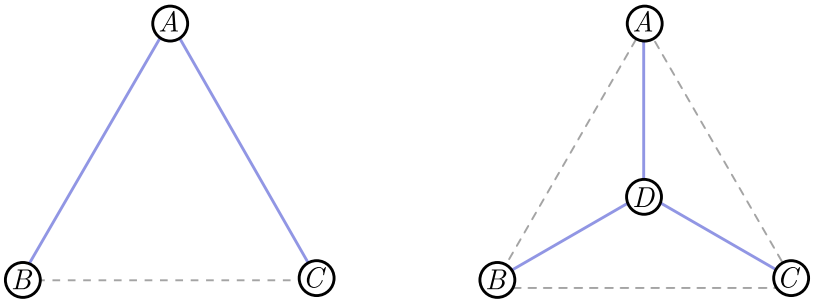


Figure 2: Rail networks on an equilateral triangle.

Exercise 2.3. *Equilateral trident.*

Show that the length of the trident network is

$$L_{\mathrm{Y}} = \sqrt{3}d.$$

Since $\sqrt{3} \approx 1.7 < 2$, the trident is shorter than the triangle.

Although this beats the triangle network, it's possible that placing D somewhere other than the center could make the network even shorter. But as it turns out, the center is optimal, and we can argue this from *symmetry*. We draw one of the triangle's axes of symmetry[3] in red in Fig. 3. We can wiggle the hub D left and right along the dark blue line in Fig. 3.

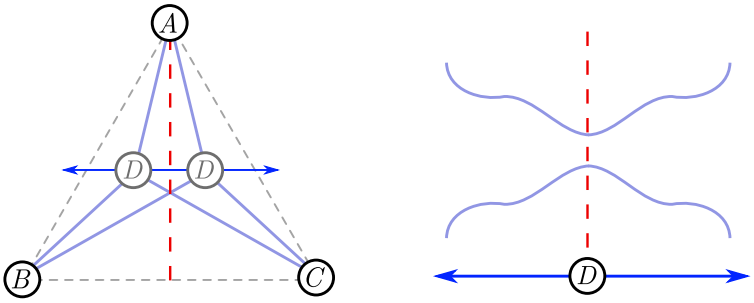


Figure 3: *Left.* Wiggling the hub. *Right.* Length is an even function of wiggle.

Because of symmetry, the total length of the network (light blue lines) is an *even function* of how far we have moved D along the dark blue line. On the right in Fig. 3, we depict two possibilities for an even function. Length could either be a *minimum* on the red line, like the curve on top, or a *maximum*, like the curve on the bottom. Of course, if we move the hub along the blue line outside

[3]This cuts the triangle into two mirror-image halves.

the triangle, the network becomes very long. This suggests it is a minimum[4] on the red line, and for a minimal network it should lie on that line, as in Fig. 4 (left). But there are two other axes of symmetry, associated with B and C. All three intersect at the center of the triangle, as shown in Fig. 4 (right). Since D should lie on each of these lines, it must lie at the center!

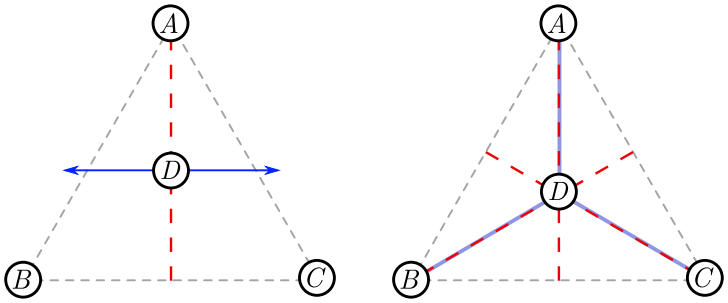


Figure 4: *Left.* Length is minimized on the red line. *Right.* Total length is minimized at the intersection of the red lines.

2.2 Deforming the triangle

We are now going to take our solution to the equilateral triangle and slowly *deform* it, sliding the corners so that the triangle so it is no longer equilateral. What will happen to the optimal position of the hub D? Since everything is sliding continuously, the optimal hub should slide continuously as well. In Fig. 5, we give an example, with the paths of the corners are depicted in purple, and the corresponding continuous change of hub in green.

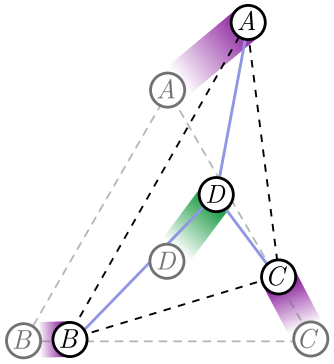


Figure 5: Optimal hub position slides as we slide the corners of the triangle.

Since the hub position changes continuously, it should stay inside the triangle for small deformations of the corners. But for triangles which are far from equilateral, the sliding hub might *collide* with a corner! In this case, the trident network collapses into a simpler triangular network, formed from two sides of the triangle. In Fig. 6, B remains fixed in position, but A and C lower symmetrically and open out the angle of the triangle, with the optimal hub D moving vertically down as they do so. At some critical angle θ_{crit}, it will coincide with B.

[4]This does not *prove* it is a minimum, since there may be other minima we have missed. See Exercise 2.8 for a rigorous proof.

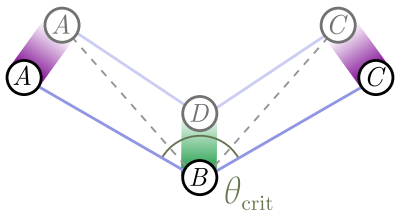


Figure 6: At some critical angle θ_{crit}, D collides with B.

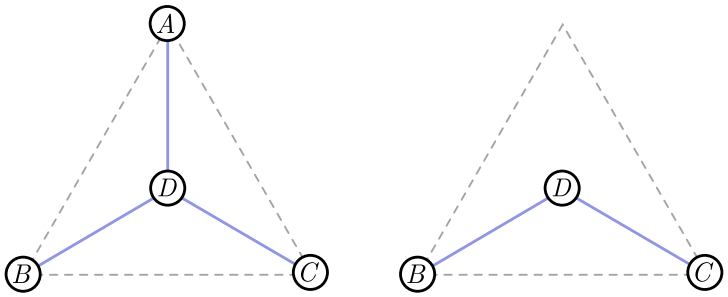


Figure 7: Removing a corner city removes a leg from the equilateral trident.

It turns out this critical angle is $\theta_{\text{crit}} = 120°$. Although we won't provide a watertight proof just yet, we can give a plausibility argument. Let's return to the equilateral triangle. Instead of adding a hub in the middle, suppose that D is in fact a *fourth city* fixed in place. Clearly, the solution in Fig. 7 (left) is still optimal, since if we could add more hubs to reduce the total length, we could add more hubs to improve the network for the equilateral triangle. If we now remove a corner city, such as A, the optimal network removes the corresponding leg of the trident, as in Fig. 7 (right).

Exercise 2.4. *Cutting corners.*

Suppose that in Fig. 7 (right), we can add a new hub E which reduces the total length of the network. Explain how adding E could reduce the length of the network in Fig. 7 (left), and thereby improve our solution for the equilateral triangle.

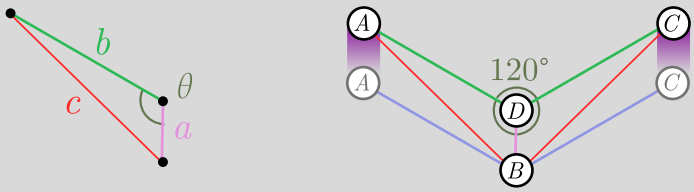
Exercise 2.4 is an example of a *proof by contradiction*, a favourite proof method among mathematicians. To show something is false, we assume it is true and use it to derive a contradiction with known facts. We then reason backwards to conclude that it cannot be true! The next exercise gives a slightly stronger indication that the critical angle is 120°. This is the best we can do without more involved math (Exercises 2.9 and 2.8).

Exercise 2.5. *Critical isosceles.* ▲

The argument above really only establishes that $\theta_{\text{crit}} \leq 120°$. In principle, the triangular network might become optimal at some angle $\theta_{\text{crit}} < 120°$. In this exercise, we will show for an isosceles triangle that this is not the case. We will need the *law of cosines*,

$$c^2 = a^2 + b^2 - 2ab\cos\theta,$$

for the triangle depicted below left:



Above right, we have a triangular network (blue lines) ABC, forming an angle of 120°. We now raise the two nodes A and B symmetrically so that the angle ABC is less than 120°. You can prove that the green and purple lines are shorter than the red lines, so that an interior hub D, making an angle 120° with green and purple lines, yields a shorter network.

(a) Show using the law of cosines(or otherwise) that

$$c^2 = a^2 + b^2 + ab.$$

(b) From part (a), argue that

$$a + 2b < 2c.$$

(c) Conclude that for an isosceles triangle ABC, the critical angle is $\theta_{\text{crit}} = 120°$.

2.3 The 120° rule

Let's state the general, n-city version of the problem we've been studying:

Box 2.1. *Minimal networks.*

Suppose we have n cities on the plane. The *minimal network* or *Steiner tree* is a configuration of edges connecting these cities which has minimal total length. We can introduce additional hubs in order to minimize this total length.

Our work with triangles pays off with a remarkable conclusion about minimal networks connecting *any* number of cities called *the* 120° *rule*. Readers who are not interested in the proof may simply internalize the the contents of the following blue box and move on.

Box 2.2. *The* 120° *rule.*

In a minimal network, every hubs has three edges separated by angles of 120°.

The argument is ingenious. Our first step is to show that it is impossible for a hub to have edges separated by less than 120°. Suppose we have cities or *fixed nodes* $A_1, A_2, \ldots, A_n$ connected by a minimal network, and a hub station H with incoming rail lines separated by less than $\theta_{\text{crit}} = 120°$, as on the left in Fig. 8. There may be other incoming lines, but these will play no role in our proof

and can be ignored.

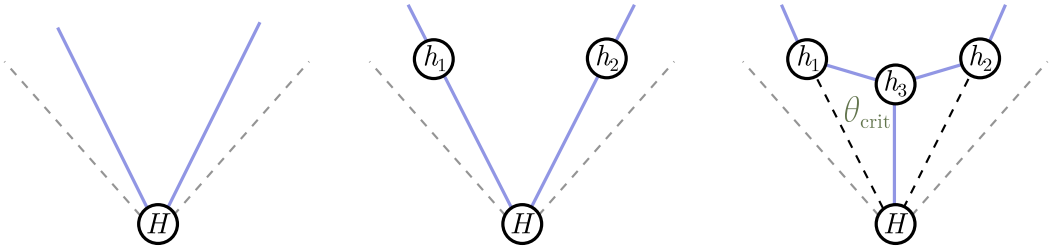


Figure 8: *Left.* A hub with incoming angle less than θ_{crit}. *Middle.* Adding two extra stations. *Right.* A shorter network.

We can build two new stations on these outgoing legs, h_1 and h_2, without changing the length of track. For simplicity, we take these new stations to be the same distance from H, as in Fig. 8 (middle). But from our work in the previous section, we know that the minimal network connecting h_1, h_2 and H is not the triangle network we have drawn! Instead, it is a trident with another hub h_3 in the middle, Fig. 8 (right). This strictly decreases the length of the network, so our original network could not be truly minimal.

This means that any hub must have spokes separated by *at least* 120°. How do we know that there are three, separated by exactly 120°? Well, suppose two lines enter H, separated by *more* than 120°. Then there can only be two incoming edges, joining H to some cities A and B, since any additional lines would have to be closer than 120° to one of these lines. We have the situation depicted on the left of Fig. 9.

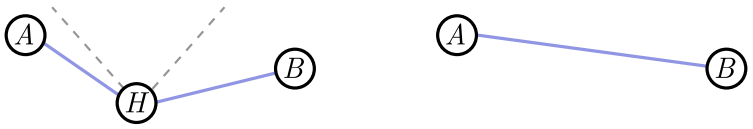


Figure 9: *Left.* A hub with incoming angle greater than θ_{crit}. *Right.* A shorter network.

Hopefully you can see what goes wrong: if there is a "kink" in the blue line, then we can obtain a shorter network be deleting H and directly connecting A and B. (Remember that H is a hub, introduced only to shorten the network, and not a city that needs to be connected.) Once again, we have a contradiction! Strictly speaking, we can have hubs with only two incoming edges, separated by 180°. But such a hub is always unnecessary, since all it does is sit on a straight line. If we delete these useless hubs, we have the general result advertised above, namely that any hub in a minimal network has three equally spaced spokes.

Exercise 2.6. *Outer rim.*

Our proof applies to hubs only, but similar arguments apply to the cities $A_1, A_2, \ldots, A_n$. Prove the following:

(a) No incoming edges can be separated by less than 120°.

(b) The number of incoming edges is between one and three.

2.4 A minimal history

French mathematician Pierre de Fermat (1607–1665) was the first to ask about minimal networks on the triangle, though he framed it as a geometric problem:

Box 2.3. *Fermat's problem.*

Given three points A, B, C in the plane, find the point D such that the sum of lengths $|DA| + |DB| + |DC|$ is minimal.

He figured out the answer himself, but according to the mathematical custom of the day, sent a letter to Galileo's student Evangelista Torricelli (1608–1647), challenging him to solve it. Torricelli found the same answer, but using a different method, so the position of the hub is called the *Fermat-Torricelli point* in joint honour of its discoverers. Jakob Steiner (1796–1863) generalized Fermat's question to n points on the plane:

Box 2.4. *Steiner's problem.*

Given n points $A_1, \cdots, A_n$ in the plane, find the point D such that the sum of lengths $|DA_1| + \cdots + |DA_n|$ is minimal.

Although minimal networks are also called *Steiner* trees, Steiner's problem is very different from the n-city problem we've been considering. Steiner wanted a *single* point such that the sum of lengths to that point is minimal, rather than a connected network of minimal length. Put differently, it is the minimal network when you are allowed to add at most one hub.

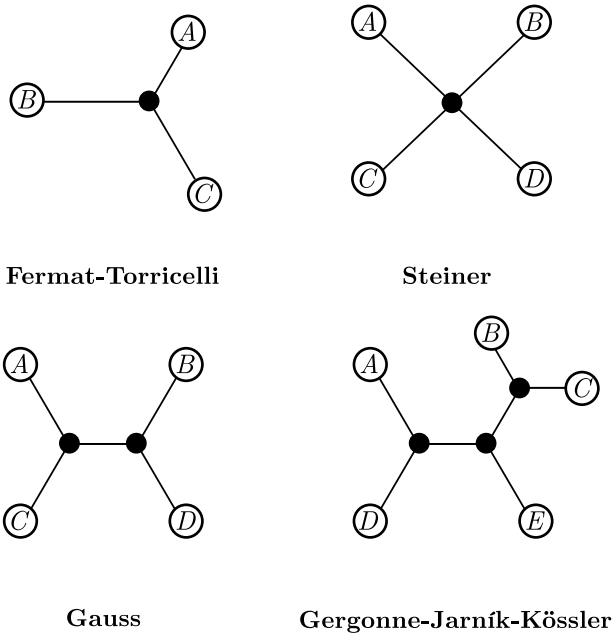


Figure 10: A visual history of minimal networks.

In 1836, 200 years later, the great German mathematician Carl Friedrich Gauss (1777–1855) mulled on the design of a minimal rail network between four German cities (Exercise 4.1). Around the same time, the French mathematician Joseph Diez Gergonne (1771–1859) considered the general n city problem (connecting them via canals rather than railways) and discovered the 120° rule. The world evidently paid no attention until 1934, when Czech mathematicians Vojtěch Jarník (1897–1970) and Miloš Kössler (1884–1961) independently rediscovered Gergonne's results [21]. The Gergonne-Jarník-Kössler version was popularized under the name *Steiner trees* by Richard Courant and Herbert Robbins in their classic 1941 text, *What is Mathematics?* [5].

We finish this section by finding the Steiner point and Steiner tree for regular polygons, a trigonometric construction of the Fermat-Torricelli point for the optimal hub placement (Exercise 2.9), and a proof that the 120° rule does indeed minimize total distance (Exercise 2.8).

Exercise 2.7. *Easy polygons.*

Consider n cities on the corners of a regular n-sided polygon.

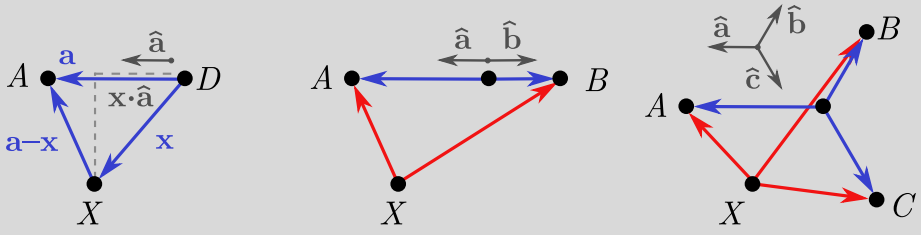
(a) Show that for $n \geq 6$, the network formed by removing a single edge from the perimeter satisfies the 120° rule and requirement (a) from Exercise 2.6. It's harder to prove, but this is in fact the minimal network![a]

(b) Use the reasoning in §2.1 to argue that the center of the polygon solves Steiner's problem in Box 2.4.

[a] You might wonder why this doesn't follow immediately. As will explore in §3, and particularly Exercise 3.4, it turns out that satisfying these rules is not enough to guarantee that a network is minimal.

Exercise 2.8. *From straight lines to Steiner's problem.* 🏂

Here, we give a rigorous proof of the 120° rule, and immediately extend it find the analogous rule for Steiner's problem. The proof makes use of vectors and the dot product, hence the higher difficulty rating. Recall that $|\mathbf{v}|$ is the length of the vector $\mathbf{v}$, and $\hat{\mathbf{v}} = \mathbf{v}/|\mathbf{v}|$ is the unit vector pointing in the same direction.



Choose a point D on the plane, which will act as the "origin". Consider another point, A, making a vector $\mathbf{a} = DA$, with unit vector $\hat{\mathbf{a}}$.

(a) Prove (visually or however you like) that for any other point X, with $\mathbf{x} = DX$,

$$|\mathbf{a}| \leq |\mathbf{a} - \mathbf{x}| + \mathbf{x} \cdot \hat{\mathbf{a}},$$

where as in the image above, $\mathbf{x} \cdot \hat{\mathbf{a}}$ is the length of $\mathbf{x}$ projected onto $\mathbf{a}$.

(b) Consider two points A and B on the plane. Using the previous exercise, show that for any point X,

$$|DA| + |DB| \leq |XA| + |XB| + \mathbf{x} \cdot (\hat{\mathbf{a}} + \hat{\mathbf{b}}).$$

(c) Conclude that if we choose D so that $\hat{\mathbf{a}} + \hat{\mathbf{b}} = \mathbf{0}$, the sum $|DA| + |DB|$ will be minimized. Geometrically, what does correspond to? Does this makes sense?

(d) Let's now introduce *three* points A, B, C on the plane, with origin D. Generalize (c) to establish that $|DA| + |DB| + |DC|$ is minimized when

$$\hat{\mathbf{a}} + \hat{\mathbf{b}} + \hat{\mathbf{c}} = \mathbf{0}.$$

Show that this is precisely the 120° rule.

(e) Finally, consider points $A_1, \ldots, A_n$ and corresponding vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$. Generalize (d) to conclude that if a point D exists such that

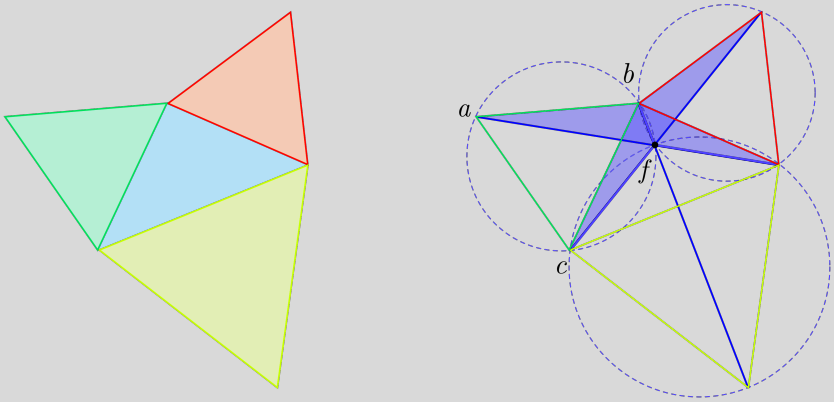
$$\hat{\mathbf{a}}_1 + \cdots + \hat{\mathbf{a}}_n = \mathbf{0},$$

then it solves Steiner's problem (Box 2.4).

(f) Exploit (e) to solve Steiner's problem for an arbitrary quadrilateral.

Exercise 2.9. *Searching for Fermat-Torricelli.*

Here, we give a geometric construction of the Fermat-Torricelli point for any triangle. Proceed if you like geometry! So, we're going to find the interior point satisfying the 120° rule for the blue triangle (below left). Start by attaching equilateral triangles (green, red, yellow) on each side, and drawing lines (dark blue) from the outer corners of the equilateral triangles to the opposite corner of our original triangle, as shown below right.



We claim these lines intersect at the point f, and moreover, are separated by angles of 120°. To prove this, draw the dotted circles circumscribing each equilateral

triangle. The exercises guide you through a demonstration that the circles intersect at 120° angles at f, using the inscribed angle theorem.

(a) Show that the shaded triangles are congruent. Argue that, in consequence, the three blue lines do interesect at a single point.

(b) From part (a), argue that $\angle baf = \angle bcf$.

(c) From (b) and the inscribed angle theorem, argue that a, b, c, f lie on a circle.

(d) Since the triangle is equilateral, $\angle cab = 60^\circ$. Using the inscribed angle theorem once more, show that $\angle cfb = 120^\circ$. Repeating this argument for the remaining two triangles gives our result!

This construction works provided all angles in the blue triangle are $< 120^\circ$.

(e) What goes wrong if an angle is $\geq 120^\circ$?

3 Graphs

In a sense, the 120° rule solves the problem of minimal networks, giving us a mathematical condition that hubs must obey. But if I hand you a list of cities and tell you to start designing, you will quickly see that the 120° rule is not enough! In this section, we will think more about the layout of networks, including how many hubs we need to consider, the number of network arrangements, the general difficulty of finding these networks and methods for approximating them.

Studying network layouts is the domain of *graph theory*. A graph is a bunch of dots connected by lines, drawn on a page. The technical term for dots is *vertices* or *nodes*, and *edges* for the lines. If an edge joins two nodes, we say they are *neighbours*. Edges must start and end at different vertices, and are allowed to overlap. Vertices can be attached to any number of edges, including zero. The rules are illustrated in Fig. 11. We let E denote the number of edges and N the number of nodes.

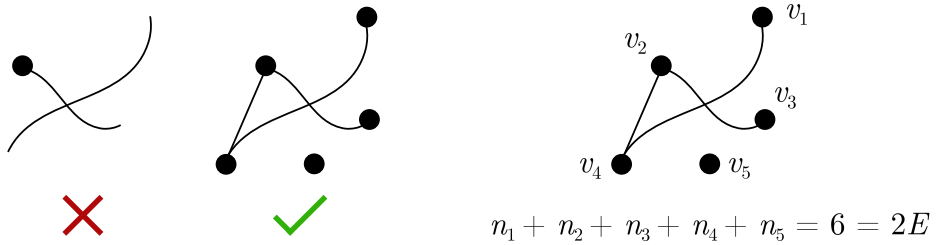


Figure 11: *Left.* "Illegal" and "legal" graphs. *Right.* The handshake lemma in action.

We will need a simple, general result called the *handshake lemma*. There are two ways to count edges. The first is simply to count the edges directly, yielding a number E. But our rules tell us that edges attach to a vertex at each end. So instead, we can go through the vertices and count the number of edges which attach to them. This will hit each edge *twice*, once for the vertex at either end, so this way of counting gives $2E$. That's the handshake lemma! More precisely, suppose there are N vertices $v_1, v_2, \ldots, v_N$. If these have $n_1, n_2, \ldots, n_N$ edges attached, the handshake lemma states that

$$n_1 + n_2 + \cdots + n_N = 2E. \tag{1}$$

The name, incidentally, comes from the fact that if vertices $v_1, \ldots, v_N$ are people, and edges are handshakes, we add the number of handshakes each person performs to get twice the total number of handshakes.

3.1 Trees and leaves

Some cities are not joined by rail, say Minsk and Darwin. But in a *connected* rail network, there is at least one route between each pair of cities. Anyone who has had the pleasure of exploring Tokyo by train will know the dizzying extent to which more than one route from A to B is possible. But in a genuinely minimal train network, A and B will be joined by a *unique* route.

The basic idea is to get rid of routes until one is left. If there is more than one way to get from A to B, the network has unnecessary edges and can be "pruned" to get something shorter. You might worry that pruning these unnecessary edges could accidentally disconnect other cities, but this is never the case! Fig. 12 shows why. Suppose A and B are connected by two paths, labelled p_1 and p_2, and potentially consisting of more than one edge. The blob to the left is all the vertices whose paths to B go through A first, and similarly, vertices on the right connect to A through B.

If two nodes are in the same blob, such as C and E, then pruning path p_2 has no effect on whether they are connected. If two nodes are in different blobs, like C and D, they can still reach each other using path p_1. We can prune the redundant paths willy nilly!

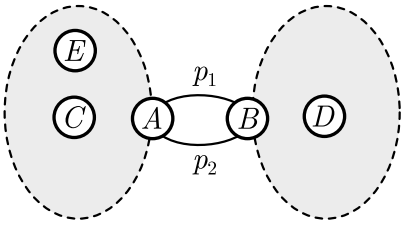


Figure 12: Pruning unnecessary paths.

Exercise 3.1. *Pruning along the path.*

Generalize the argument above to account for nodes that lie between A and B. (These are nodes which can connect to either A or B without passing through the other, and schematically lie on paths p_1 or p_2.)

Once we have completely pruned the network, there is only a single path connecting any two nodes A and B. Such a network is called a *tree* because it can be drawn so that edges look like branches. This finally explains why minimal networks are also called Steiner trees! An example of a tree is shown in Fig. 13. Every tree has a special node called a *leaf*. As the name suggests, this is at the "end" of the tree's branches. More formally, a leaf is a node with a single edge, like J, D, G, and I in Fig. 13. It may seem intuitive, but as an exercise in reasoning about trees, let's *prove* they must have leaves.

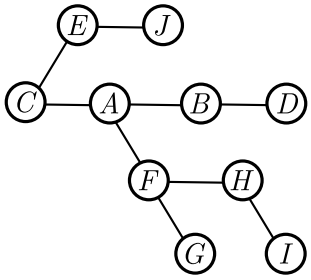
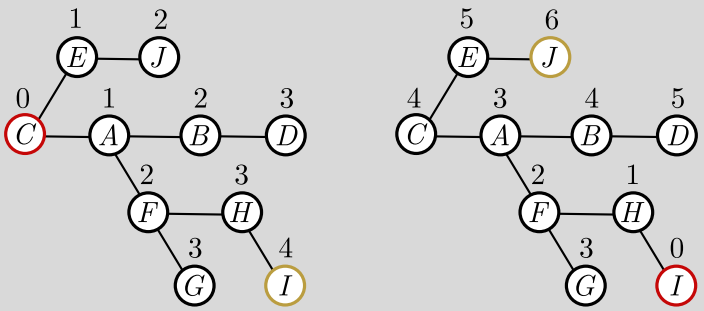


Figure 13: A tree network, with a unique path between each node.

Exercise 3.2. *Finding leaves.*

To begin our proof that each tree has a leaf, we choose a node at random (red, below left) and count the number of steps to each other node.



(a) Explain why the number of steps from the red node to any other node is well-defined in a tree.

(b) Consider the node or nodes furthest from the red node (orange, above left). Argue that these must be leaves. *Hint.* If they are not, what is the distance from red node to their neighbours?

(c) In fact, we can prove something stronger. The previous question tells us how to find a leaf. Repeat the same procedure, but start with the leaf and find the furthest node. Conclude that every tree has *two* leaves.

(d) Show, using an example, that a tree need not have more than two leaves.

3.2 Hub caps

In Fig. 13, you may have noticed the number of edges $E = 9$ is one less than the number of nodes, $N = 10$. This is not a coincidence. For any tree, it turns out that $E = N - 1$. We can prove this fact using the existence of leaves. The idea is simple: keep removing the leaf, and the single edge joining it to the rest of the tree, until you have a single node left. This requires the removal of $N-1$ nodes, and hence $N - 1$ edges. Since there are *no* edges now, and we removed one each time, we must have started with $N - 1$ edges. Hence,

$$E = N - 1 \tag{2}$$

for trees in general. Equation (2), along with the handshake lemma (1), will allow us to place a cap on the maximum number of hubs that can occur in the network.

Suppose we are trying to connect n cities, and introduce h hubs in order to do so. The total number of nodes is then $N = n + h$. The 120° rule tells that each hub attaches to exactly three edges. Each of the n cities attaches to at least one edge to ensure it is connected to the rest of the network. Thus, (1) gives

$$2E = n_1 + n_2 + \cdots + n_N \geq n + 3h. \tag{3}$$

From (2), we know that $E = N - 1 = n + h - 1$. Combining this with (3), we find

$$2(n + h - 1) = 2n + 2h - 2 \geq n + 3h \quad \Longrightarrow \quad n - 2 \geq h. \tag{4}$$

In other words, the number of hubs h is at most $n - 2$.

Exercise 3.3. *Hubs and nubs.*

While hubs always have three attached edges, Exercise 2.6 tells us that cities (fixed nodes) have between one and three edges.

(a) Show it is always possible to arrange n cities so that $h = 0$.

(b) At the other end of the spectrum, argue that the maximum number of hubs, $h = n - 2$, occurs when the fixed nodes are exactly the leaves of the network.

The 120° rule and hub cap together give us a simple tool for building minimal networks. For n fixed nodes, pick $h = n-2$ hubs, with spokes emerging at angles of 120°, and connect them together to form a tree, with the fixed nodes as leaves. Although the angles are fixed, we can extend the spokes and legs, and perform overall rotations of the network. We call this extendable configuration of hubs and spokes a *tinkertoy*, after the modular children's toy it vaguely resembles (Fig. 14).[5]

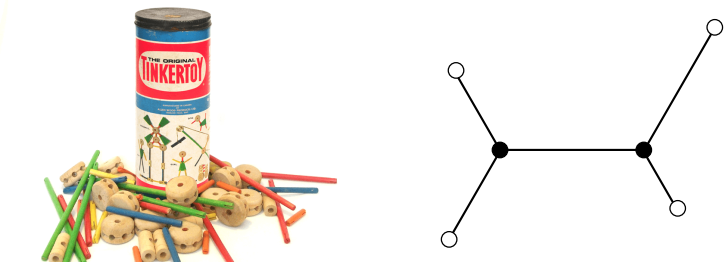
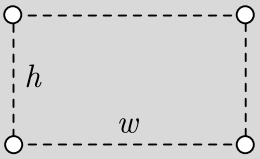


Figure 14: *Left.* A real Tinkertoy™. *Right.* A network tinkertoy.

We can play with our network tinkertoys, or program a computer to play with them, until they do what we want. The strategy of playing with tinkertoys to find the minimal network was first proposed by Zdzislaw Alexander Melzak[6] [22]. We give some examples in the following exercises.

Exercise 3.4. *Minimal rectangular network.*

Consider four cities on a rectangle of height h and width $w \geq h$:



(a) Draw the single tinkertoy for $n = 4$, and argue from Exercise 2.6 that this should describe the minimal network.

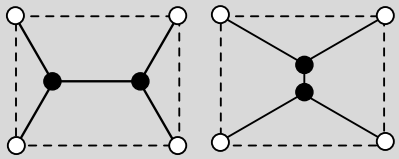
(b) Fit the tinkertoy to the city, and deduce that the minimal network has length

$$L = w + \sqrt{3}h.$$

[5] In the mathematics literature, a tinkertoy graph is called a *topology*, and the different ways of stretching the edges the *degeneracies* of that topology.

[6] Apparently, Melzak was a professor at UBC, but remarkably little information exists about him now.

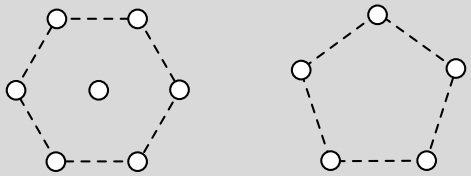
(c) Show that the tinkertoy can be oriented in *two* ways when $h < w < \sqrt{3}h$. Explain why the horizontal orientation is always minimal.



Part (c) tells us something very important. Even if a tinkertoy fits, the configuration isn't necessarily the true minimum! Put different, the 120° rule is not sufficient to guarantee that a network is minimal.

Exercise 3.5. *Harder polygons.*

When $h = n - 2$, you can use a single tinkertoy, but if $h < n - 2$, you will need multiple tinkertoys.



(a) Find the minimal network connecting a regular hexagon with a node in the center. You will need multiple tinkertoys!

(b) Draw the single tinkertoy for $n = 5$, and schematically indicate what the minimal network on a regular pentagon looks like. (No need to find the exact hub positions.)

3.3 Avoiding explosions

For a small number of hubs, tinkertoys are useful. But are they useful for many hubs? Suppose that fiddling with tinkertoys is a quick operation, and once a tinkertoy is selected, a human or a computer can quickly check whether the tinkertoy can be extruded to hit our fixed points. If there are many tinkertoys, finding one that fits could still take a while. In Fig. 15, we show a few tinkertoys for $h = 6$, suggesting that with more hubs, enumerating them all may turn out to be hard. In fact, as h gets larger, the total number of tinkertoys T_h suffers what is called a *combinatorial explosion*, growing exponentially as a function of h. A brute force approach, which simply fiddles with each tinkertoy to see if it can be made to fit the fixed point, will take an exponential amount of time. This is beginning to seem like a hard problem in general!

Counting the total number of tinkertoys is difficult.[7] To demonstrate this exponential growth, we are instead going to focus on a subset of tinkertoys we can conveniently enumerate. Trees in

[7] See Exercise 3.6 for a physicist's shortcut.

general have a complicated structure, so to simplify, we consider only *linear* tinkertoys. These are tinkertoys where the hubs lie on a "line", so that no hub has more than two neighbours, for instance Fig. 16 (left). The next problem is that even these linear tinkertoys can be rotated by 180°. To avoid counting the same tinkertoy twice, we need some way of knowing which end is which. A simple method is to start and end with a ∨-shaped segment, as in Fig. 16 (right). If we rotate 180°, the tinkertoy is bookended by ∧-shaped segments, which is clearly distinct. Not every linear tinkertoy has this form, so we call these special tinkertoys *oriented*.

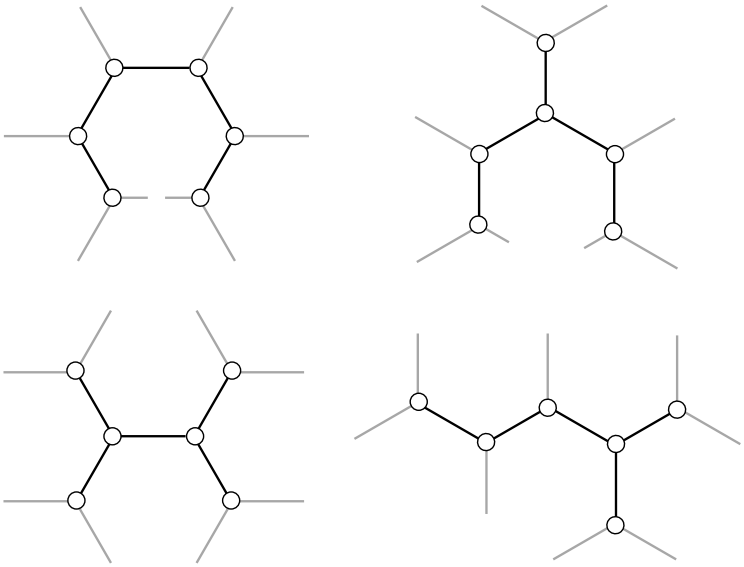


Figure 15: A selection of tinkertoys for $h = 6$.

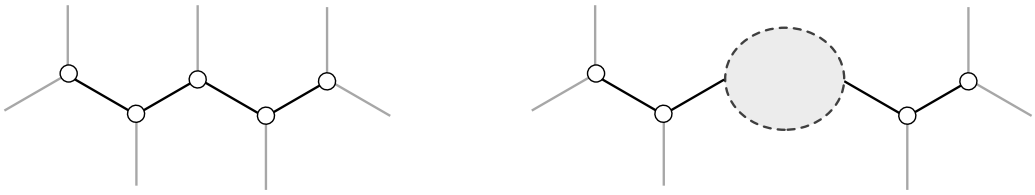
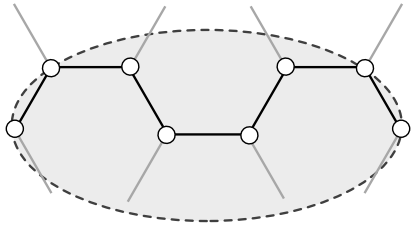


Figure 16: *Left.* A linear tinkertoy. *Right.* An oriented tinkertoy.

With the notion of oriented tinkertoys, we can immediately find an exponentially growing set! There are $h-1$ edges altogether since the hubs form a tree. We fix four (two at each end) to ensure the tinkertoy is oriented. That leaves $h-5$ edges within the grey circle of Fig. 16 (right). As we move along from the leftmost ∨, these edges constitute $h-5$ turns left or right by 60° before we exit again to hit the final ∨. At each point, either a left or a right turn is allowed, so there are 2^{h-5} possible choices altogether. To make this more transparent, we could label left and right turns with 1s and 0s respectively, so that a tinkertoy is just a sequence of binary digits, as in Fig. 17. Thus, there are an exponential number of oriented tinkertoys.[8] If you like drawing graphs, you can have a grow at findiing the *total* number T_h in the next exercise.

[8]You might worry that if we turn too many times, the tinkertoy will collide with itself and no longer be valid. For



1001100

Figure 17: The binary sequence for an oriented tinkertoy.

Exercise 3.6. *Physicist's induction.* ⛰

Calculate the number of tinkertoys T_h from $h = 0$ to $h = 6$. You should be able to find the general sequence T_h by searching for these numbers in the Online Encyclopedia of Integer Sequences. At large h, the OEIS informs us that this sequence grows exponentially, with

$$T_h \approx \frac{2^{2h-4}}{\sqrt{\pi} h^{5/2}}.$$

Of course, this doesn't prove that the sequences match up forever. But it is enough evidence to satisfy a physicist.

By now, we should be confident that there are many tinkertoys. If we have to consider even a fraction of them at large h, any computer is doomed to failure. For instance, using the counting in Exercise 3.6, suppose a computer can check a billion tinkertoys per second, and wants to design a railway network to connect the ~ 800 largest cities in North America. If it has to check every tinkertoy, it will take an unimaginably long

$$\frac{2^{2\cdot 800-4}}{\sqrt{\pi} 800^{5/2} \cdot 10^9} \text{ s} \approx 10^{456} \text{ years.}$$

Would a faster computer help? Not likely. If you do more operations per second than there are atoms in the universe, it still takes $\sim 10^{388}$ years! No realistic improvements in processing speed will make this problem solvable in the lifetime of the universe.

Notice that there are two slightly distinct problems here. The first is searching for tinkertoys that fit; and the second is singling out the truly minimal network from the shortlist of fitting tinkertoys. The two are not the same because, as we saw in Exercise 3.4, just because a tinkertoy fits doesn't mean it is minimal. The first problem is easier because if somebody hands you a tinkertoy and claims it fits, you can easily check. In fact, you yourself could make a lucky guess and find a tinkertoy which fits immediately. There is an area of computer science called *complexity theory* which classifies problems according to how hard they are. In the language of complexity theory, finding tinkertoys that fit is called NP, for "Non-deterministing Polynomial time". This is a fancy way of saying you can make a lucky guess and confirm it immediately.

instance, after six right turns, edges of equal length will form a closed hexagon! But we can always adjust the length of edges to prevent this from happening, so the count remains valid.

In fact, fitting tinkertoys is as hard as any problem in the set NP. "As hard as" is a technical term in complexity theory, meaning that you can transform any algorithm for finding good tinkertoys into an algorithm for solving *any other problem* in NP! It is a key that unlocks the rest of the set. We call such a task NP-complete, since it gives us "complete" access to every NP problem.

If someone hands you a tinkertoy configuration and claims that it's the minimal network, you must first check that it fits. So finding a minimal network is at least as hard as fitting a tinkertoy. But you can't stop there! You have to keep searching to find *all* the tinkertoys that fit, checking the lengths, and verifying that the first configuration really is the shortest. The second problem appears strictly harder, though surprisingly, no one has been able to prove it.[9] The best we can say is that it is at least as hard as fitting tinkertoys (though it may be *equally* hard), placing it in a class called NP-hard [14], which means "as hard as any problem in NP".

Exercise 3.7. *Tiny tinkertoys.*

We've been talking about fitting a *single* tinkertoy, but as we saw in Exercise 3.5, the minimal network is sometimes obtained by cobbling together multiple "tiny" tinkertoys. Let's see if this changes our story.

(a) Argue that including tiny tinkertoys makes the problem of finding the true minimal network even harder.

(b) Explain why fitting a collection of tiny tinkertoys is potentially much quicker.

At the risk of spoiling the exercise, it turns out that finding tinkertoys *above* some given size, to fit a subset of fixed nodes, is NP-complete.

We can summarize these results as follows:

Box 3.1. *Complexity I.*

Fitting tinkertoys is NP-complete. Finding minimal networks is NP-hard.

3.4 Minimum spanning trees

All these heavy-sounding results about complexity theory make life sound impossible for network planners. But while finding the exact minimal network is difficult, approximating is easy! Life, and near-optimal rail travel, go on. We'll discuss two simple approximation schemes, starting with a generalization of the very first Exercise 2.1. Recall that, for three cities, the triangle network consists of the two shortest sides of the triangle. Put differently, we draw an edge between each city, and select the two shortest ones, which happen to form a tree which connects everything.

For n cities, we do the same thing. Draw an edge between each city, forming what is called the *complete graph* on n nodes. From these edges, we select a subset which form a tree, connecting each city and of minimum total length. This is called a *minimum spanning tree (MST)*, since it "spans" the cities. We illustrate the construction for $n = 4$ in Fig. 18.

[9] Since fitting tinkertoys is NP-complete, if finding minimal networks was in NP, then we could somehow transform an algorithm for fitting tinkertoys into finding the truly minimal network. So, to prove it is harder, we need to show it is not in NP. Proving there is no "lucky guess" procedure, however, is very hard, since we somehow need to consider every possible algorithm!

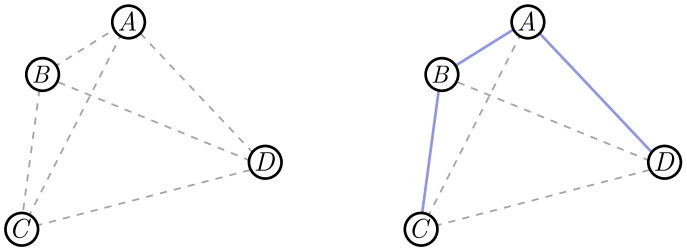


Figure 18: *Left.* The complete graph for four cities. *Right.* The minimum spanning tree.

The usefulness of MSTs depends on whether they are fast to compute and close to optimal. We start with the first question. Unlike tinkertoys, there is a procedure to construct the MST edge by edge. This procedure is very simple:

0. Pick a random vertex v_0.
1. Add the shortest edge adjacent to v_0 to form a tree T_1.
2. Add the shortest edge adjacent to T_1 to form a tree T_2.
3. Add the shortest edge adjacent to T_2 to form a tree T_3.

And so on, until we have a tree T_{n-1} which spans all the nodes. By "adjacent to", we just mean an edge which touches the tree but is not already in it. This algorithm was discovered in 1930 by Jarník [20], but subsequently rediscovered by Robert Prim in 1957 [24], so it is called the *Prim-Jarník algorithm*. We implement it for $n = 4$ in Fig. 19.

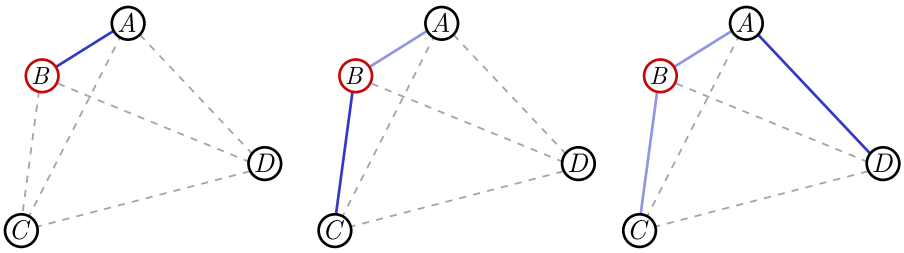


Figure 19: The Prim-Jarník algorithm for $n = 4$. Dark blue edges are added sequentially.

Exercise 3.8. *MST is easy.*

Here, we will give a very lazy bound on the number of steps required to perform the Prim-Jarník algorithm.

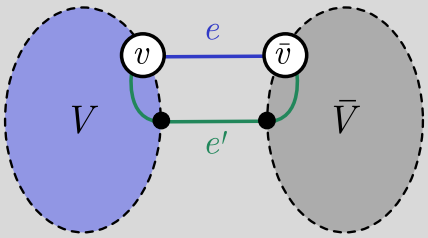
(a) Using the handshake lemma (1), show the total number of edges in the complete graph on n cities is $E_{\text{complete}} = n(n+1)/2$.

(b) The algorithm has $n-1$ steps where it adds an edge. For each step, it must consider the available edges. Call this a *substep*. Give a very lazy argument that the total number of substeps for the algorithm is $\leq n^3$.

This is a polynomial function of n, rather than an exponential function of n.

Exercise 3.9. *Correctness of Prim-Jarník.* ⛰

Suppose that the Prim-Jarník algorithm produces a tree T which is not minimal, with $T' \neq T$ the genuine MST. Then there must be a step in the construction where we first add an edge e which is not in T'. We will show that the algorithm is *correct* in the sense that this situation cannot occur! There will always be a shorter edge e' it should add instead of e. The setup is shown below.



(a) Suppose that, before the algorithm adds the "bad edge" e, it spans a set of cities V. The complementary set of cities is $\bar{V}$. Show that e connects a vertex $v \in V$ to a vertex $\bar{v} \in \bar{V}$.

(b) Argue that the MST T' has an edge e' connecting V to $\bar{V}$. *Hint.* Use the fact that there is a path from v to $\bar{v}$ in T'.

(c) Explain why removing e' from T', and replacing it with e, results in a *tree*. *Hint.* Show there is still exactly one route between any two nodes.

(d) From part (c), conclude that the Prim-Jarník algorithm is correct.

Finding MSTs is quick. But are they any good, or can they be much longer than the minimal network? Once again, our simple results on triangles provide some insight. Let's start with an equilateral triangle of side length d. In Exercise 2.3, you found that the minimal network has length $L_Y = \sqrt{3}d$. The MST for the equilateral triangle just consists of any two sides, and therefore has length $L_\Lambda = 2d$. The *ratio* of these two lengths is $\rho = L_\Lambda / L_Y = 2/\sqrt{3} \approx 1.15$, so the MST is about 15% longer than the Steiner tree. This is close enough for many practical purposes.

You might wonder, in general, how bad this ratio can get. To start with, let's see what happens when we squeeze or stretch the triangle symmetrically. If we squeeze it, like Fig. 20 (left), the MST consists of a long side of length d and the short side which shrinks to zero. Similarly, the minimal network consists of two short sides which approach zero, and a long side which approaches d. So the ratio of lengths approaches 1. Similarly, as we stretch the triangle out like Fig. 20 (right), the MST is the shorter two sides at the top, of total length $2d$, while the hub eventually hits the top vertex, so it coincides with the MST. Once again, the ratio approaches 1.

This hints that the equilateral triangle is the worst-case scenario. In fact, you can show in Exercise 3.11 that this ratio is at most $2/\sqrt{3}$ for any triangle. In Gilbert and Pollak's magisterial study [15], they conjecture that this holds for *any number* of cities! In other words, if ρ is the ratio of the length of the MST to the minimal network for any given set of cities, the *Gilbert-Pollak conjecture* states that

$$\rho \leq \frac{2}{\sqrt{3}}. \tag{5}$$

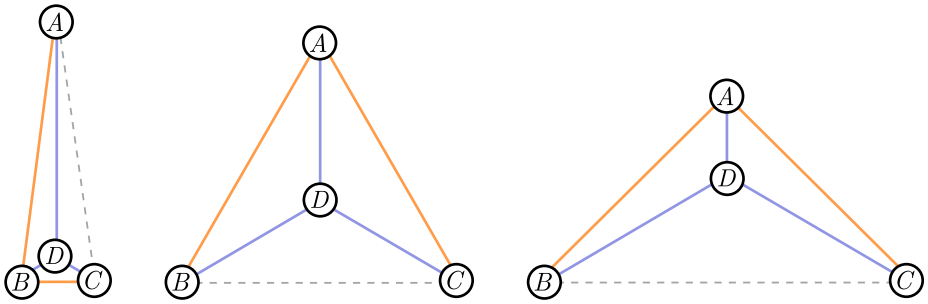


Figure 20: Stretching and squeezing the equilateral triangle.

The conjecture remains unproven. The best we can do right now is $\rho \leq 1.21$ [4].

What if we want to do better than 15%? We can tweak the MST a little to get closer to the optimal network length. One particularly simple method is the *Steiner insertion heuristic* [9], which elegantly combines the MST and our work with triangles. The basic observation is that no edges in a minimal network are separated by less than 120°, since hubs always have edges separated by *exactly* 120°, and edges at fixed nodes must be separated by at least 120° according to Exercise 2.6(a). The idea is to find edges with "bad" angles ($< 120°$) and replace them with hubs.

In more detail, the insertion heuristic works as follows. We first find the MST (using Prim-Jarník or another quick procedure), and then search for the pair of edges with the smallest angle $< 120°$. If no such angle exists, we are done! It such an angle does exist, the two edges connect a vertex, say A, to vertices B and C, as below in Fig. 21. We introduce a hub for these three vertices, which satisfies the 120° rule. And then we do the whole thing again, looking for bad angles to replace, until no more are left. That's it!

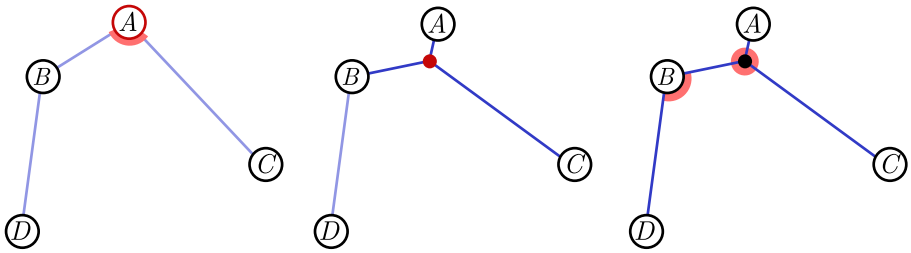


Figure 21: Applying the Steiner insertion heuristic to our MST. First, we find the smallest angle $< 120°$. Then, add a hub. No more bad angles, so we're done!

Exercise 3.10. *Steiner insertion heuristic.*

Let's explore some general properties of the insertion algorithm.

(a) Argue that the insertion of a hub can only *decrease* length.

(b) Give an example showing that the insertion heuristic need not converge to the globally minimal network. *Hint.* Exercise 3.4(c).

(c) Remember our earlier statement that fitting a tinkertoy to a set of fixed nodes is NP-complete. Explain why the Steiner heuristic can run quickly without contradicting this result. *Hint.* Exercise 3.7.

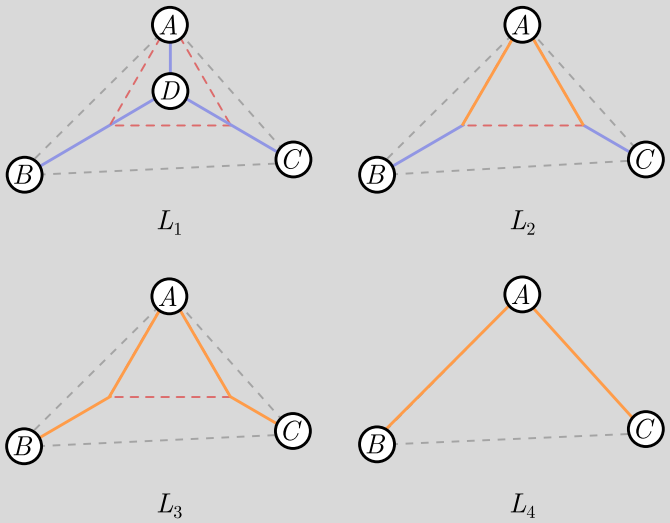
Although Steiner insertion is quick, the optimality varies. A different and less practical[10] method [3] shows that, in principle, you can approximate the minimal network on n cities as *closely as you like*, in some number of steps at most polynomial in n. For this reason, minimal networks belong to a complexity class called PTAS ("Polynomial Time Approximation Scheme"), the problems which can be easily approximated. We can update our statement about complexity:

Box 3.2. *Complexity II.*

Finding minimal networks is NP-hard but also PTAS.

Exercise 3.11. *Gilbert-Pollak for triangles.* ▲

Below, we give a visual proof of the Gilbert-Pollak conjecture for triangles. The basic idea is that, in an arbitrary triangle with angles $\leq 120°$, we can attach a small equilateral triangle to the largest angle (city A below).



The lengths L_1, L_2, L_3, L_4 are made up of lengths of coloured lines, but blue lines have weight 1, while orange lines have a weight $\sqrt{3}/2$. For instance,

$$L_1 = |DA| + |DB| + |DC|, \quad L_4 = \frac{\sqrt{3}}{2}(|AC| + |BC|).$$

In other words, L_1 is the length of the minimal network, and L_4 is $\sqrt{3}/2$ times the length of the MST.

(a) Argue that $L_1 \leq L_2 \leq L_3 \leq L_4$.

(b) Use this (along with the case where some internal angle is $\geq 120°$) to establish the Gilbert-Pollak conjecture for triangles.

[10] Impractical because each step takes a very long time.

4 Bubble networks

Humans are not the only players in the minimization game. Nature is also cheap, or rather *lazy*: it does as little as possible, formally known as the Principle of Least Action. If we play our cards right, perhaps we can hack the laws of physics to do our minimization for us. In our case, it turns out we can do network planning with *bubbles*. Bubbles are formed when a film of liquid separates two volumes of air. Surface tension tries to pull the bubble surface taut in all directions, which results in the *minimization of area*. But if there are no constraints, then the surface will shrink until nothing is left! Really, we mean that bubbles minimize the area of the wall *subject to constraints*.

For building railway networks, we want walls to be one-dimensional, and the contraints to be fixed external nodes. We'll talk about how to do this in a moment, but there is a more natural constraint associated with blowing bubbles: they enclose a pocket of air. This explains why soap bubbles are spheres! As we will show in §5.3, a sphere (Fig. 22 (left)) is the smallest surface containing a fixed volume of air. A lone bubble is direct proof of Nature's laziness.

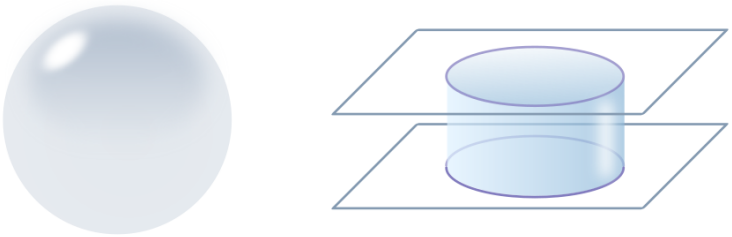


Figure 22: Soap bubbles in three and two dimensions.

If we sandwich the bubbles between two plexiglass plates, we will get a *two-dimensional* network of bubbles. The vertical walls look like a graph from above, and a single bubble will be a circle (Fig. 22 (right)). There are two questions about these networks that immediately present themselves. First, what happens at a junction of bubble walls? And second, what do walls look like *away* from a junction? The first question is easy to answer. Imagine zooming in on a junction until the walls *look straight*. Since the bubbles try to minimize wall area, or viewed from above, *wall length*, they will obey the 120° rule, since this is the local rule any length-minimizing network obeys![11]

The situation away from junctions is a little trickier, but as we will see, for both physical (Exercise 4.7) and mathematical reasons (§4.4), a bubble wall can either be straight, or it can curve along the arc of a circle. Viewing a straight line as the arc of an *infinitely* large circle, we can just say that walls are arcs of circles.

4.1 Computing with bubbles

Plexiglass gives us two-dimensional bubbles, and length rather than surface area will be minimized. But the constraint will generally be to enclose a fixed *area* of air per cell. Can we hack this setup to make a *soap bubble computer* for finding minimal networks? Yes! The key is to give the bubble walls something to hold onto. If we drill some screws between the plexiglass plates, these will act like the cities, and a network of walls can form between them.[12] Fig. 23 shows an example with

[11] Zooming in enough means that edges can be reconfigured without having any practical effect on air enclosed.
[12] For a programmable soap bubble computer, you can use suction cups with rods between them.

four screws, and the junctions that can form between bubble walls.

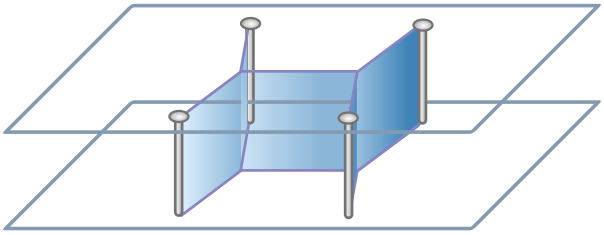


Figure 23: A soap bubble computer for finding minimal networks.

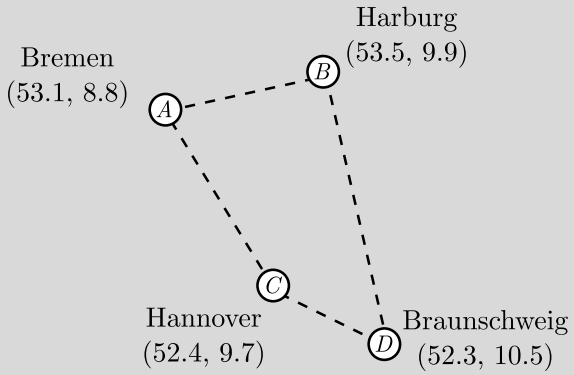
You can use a soap bubble computer to solve the original railway planning problem.

Exercise 4.1. *Trains and soap bubbles.*

As advertised in §2.4, the mathematician Gauss wanted to connect four cities with a minimal rail network. In an 1836 letter to his friend, the astronomer Heinrich Schuhmacher (1780–1850), Gauss asked:

> *How does a railway network of minimal length connect the four German cities of Bremen, Harburg, Hannover, and Braunschweig?*

The cities are drawn, along with their GPS coordinates, below:



(a) Find the minimum spanning tree using the Prim-Jarník algorithm.

(b) Assuming Gilbert-Pollak, lower bound the length of the minimal network.

(c) Improve the MST using the Steiner insertion heuristic.

(d) Build a soap bubble computer and solve the Gauss' railway problem. How does this compare to the results of the Steiner insertion heuristic?

For small networks, the soap bubble works almost instantaneously, making it easy to believe it will quickly give the right answer for large networks as well. Sadly, this cannot be true! To see why, recall that in §3.3, we argued that finding a tinkertoy is NP-complete, and finding the genuine

minimal network is NP-hard. Both problems are at least as hard as everything in NP, the class of problems where lucky guesses can be checked quickly. But just because a lucky guess can be checked quickly does not mean your chances of making a lucky guess are good. In fact, computer scientists are almost certain that most problems in NP *cannot* be solved quickly on a regular digital computer. The set of problems which can be solved quickly is called P, for "Polynomial time". To summarize, computer scientists believe that P≠NP through proving it is the most important open problem in computer science.[13]

But, you might object, a soap bubble is not a regular digital computer; it is built out of the laws of physics rather than 1s and 0s. Could it do things quickly that would take a digital computer longer than the age of the universe? The answer is almost certainly no. Computer scientist Scott Aaronson has persuasively argued [1] that the problems in NP-complete and NP-hard cannot be solved quickly by *any* computer, digital or analogue. This is called the *NP Hardness Assumption*.

One piece of evidence is that every time we think we have a loophole for quickly solving NP-complete problems, the loophole disappears on closer examination. The subtleties inevitably beat us. But there is broader philosophical reason for believing NP Hardness: roughly, *NP is OP*.[14] Many of the hardest problems we know are NP-complete, and if we could solve them, then [1]

> *...we would be almost like gods. The NP Hardness Assumption is the belief that such power will be forever beyond our reach.*

This means we cannot quickly find the minimal rail network for 800 cities using soap bubbles, a black hole, human DNA, a quantum computer, or any other conceivable mechanism. No one will ever know what the network looks like.

That raises the question: what do soap bubbles actually do? They cannot quickly find minimal networks, since this problem is potentially even harder than NP. But there are several ways for this to fail. First, they can take a long time to settle down, which Aaronson saw happening in his own soap bubble experiments, even for a few screws [1]. Secondly, they could relax into *local minima* rather than the true minima. Since fitting tinkertoys is NP-complete, even this can take a long time, unless (like the Steiner insertion heuristic) the tinkertoys are small.[15] A final possibility is that we simply solve the wrong problem, and e.g. by introducing small bubbles which change the network configuration. Based on my own experiments with soap bubble computers, it appears that all of these failure modes can be realized!

I am not trying to deflate soap bubbles, and the rest of these notes is really just a love letter to bubbles and their physico-mathematical properties. Rather, the moral is that physics and computation *interact in interesting ways*, with results about computation lead to physical predictions (e.g. Exercise 4.2). Going in the other direction, physics can lead to new insights into computer science, the most spectacular example being the advent of *quantum computers*, based on the laws of quantum mechanics rather than the classical logic of 1s and 0s. Although in their infancy, thinking about quantum computers has already taught us some remarkable things about computer science, complexity classes, and the power dwelling in Nature's laziness. Sadly, we must leave that story for another time!

[13] So important that there is a $1 million bounty on its head!
[14] Gamer speak for "overpowered".
[15] See Exercise 3.7 for more on this subtlety.

Exercise 4.2. *NP Hardness and the laws of physics.*

Here are a few fun ways to solve NP-complete problems:

(a) Create a time machine, and by sending a computer through it again and again, perform an arbitrary number of computations in finite time.

(b) Build a "Zeno hypercomputer", performing one step in 1/2 s, the second step in 1/4 s, the third step in 1/8 s, etc., so an infinite steps take 1 second.

(c) Store information in infinite precision real numbers, e.g. points on a line, and manipulate them using basic arithmetic [26].

If the NP Hardness Assumption is correct, none of these methods works! In each case, what do you think this is telling us about the nature of the universe?

4.2 The many faces of networks

While we can use soap bubbles to learn about minimal networks, we can arguably obtain more insight by going in the other direction. What do minimal networks teach us about soap bubbles? In this section, we consider the two-dimensional bubble networks with *no screws*. We'll just let the bubbles do their own thing! Fig. 24 shows a real two-dimensional soap foam.[16] We've counted the number of sides per cell, and surprisingly, most seem to be hexagonal. Is this is a coincidence, or is something deep going on?

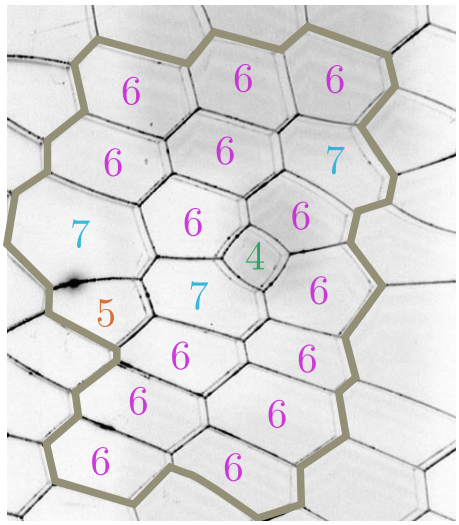


Figure 24: Most cells in a bubble network are hexagonal.

The answer is something deep. We can actually *prove* most bubbles are hexagonal using the 120° rule, some more graph theory, and a little physics. The main result we will need from graph theory is *Euler's formula*, discovered by the prolific Swiss mathematician Leonhard Euler (1707–1783)

[16] Based on a photograph by Klaus-Dieter Keller, Wikimedia Commons.

in 1735. It states a relationship between the number of nodes N, edges E, and faces F in a graph, proved below in Exercise 4.3:

$$N - E + F = 2. \tag{6}$$

Importantly, this only holds for connected graphs which can be drawn without any edges crossing, also called *planar graphs* (Fig. 25). A face is defined as any region enclosed by a loop of edges, including (counterintuitively at first) the *exterior* of the graph.

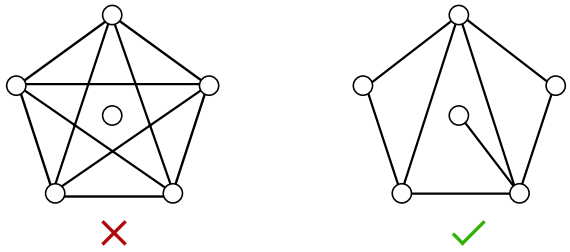


Figure 25: *Left.* A disconnected graph which cannot be drawn without edge crossings. *Right.* A planar graph. Euler's formula holds if we count the region outside the graph as a face.

One way to obtain a planar graph is to take a three-dimensional polyhedron, remove a single face, and flatten what remains onto the plane. This flattening process is shown for the cube in Fig. 26. The removed face becomes the exterior region of the graph, which is why we count it as a face.

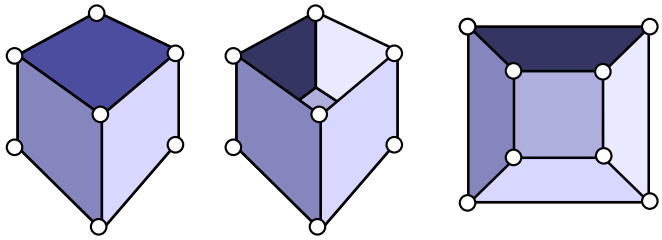


Figure 26: Remove the top of the cube and flatten.

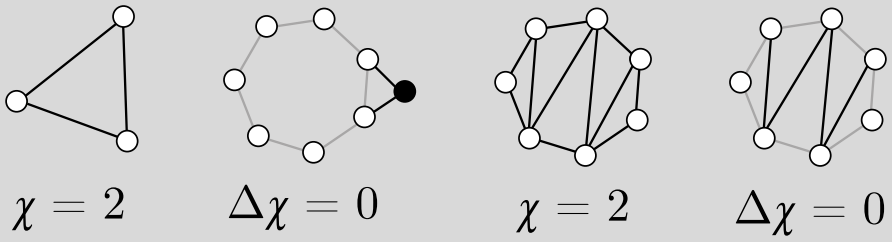
Let's check that Euler's formula works. For the cube, we have $F = 6$ faces, $E = 12$ edges, and $N = 8$ corners, so $F - E + N = 2$ just as Euler predicts. We can count either using the cube itself, or the flattened graph, provided we count the exterior as a face.

Exercise 4.3. *Euler's formula.*

Define the *Euler characteristic*

$$\chi = N - E + F.$$

Our goal will be to show $\chi = 2$ for a graph without crossings. First, we will establish Euler's formula for networks made out of triangles. We can then extend this to any graph without crossings. Below we depict stages (a), (b), (c) and (e).



(a) Show that a lone triangle in the plane obeys Euler's formula.

(b) Suppose a network obeys Euler's formula. Add a triangle (two edges and a node) to an external edge, and explain why the Euler characteristic doesn't change, $\Delta\chi = 0$. Conclude that the new network obeys Euler's formula.

(c) Explain why a network composed of triangles obeys Euler's formula.

Now we can generalize to any network without crossings.

(d) Consider a face, i.e. loop of edges, in such a network. Describe a procedure to add edges so that the face is split into triangles.

(e) Show that, after your procedure in part (d), $\Delta\chi = 0$.

(f) Conclude that any network without crossings obeys $\chi = 2$.

When there are no screws, every node in the bubble network is a hub, and therefore obeys the 120° rule, with three bubble walls meeting. By the handshake lemma (1), we have $2E = 3N$. Putting this into Euler's formula, we can eliminate N and find a relation between the number of faces and number of edges:

$$N - E + F = \tfrac{2}{3}E - E + F = 2 \quad \Longrightarrow \quad 3F - E = 6. \tag{7}$$

It will be useful to treat the external face a little differently. Let F' be the number of *internal* faces, so that $F = F' + 1$. Then (7) becomes $3F' - E = 3$.

Before proceeding, we need two additional properties of our bubble networks. First of all, an edge cannot dangle into the middle of a face. If it did, the vertex at the end of the dangling edge would not have three attached edges, only one, which is impossible by the 120° rule. It follows that every edge straddles two distinct faces.[17]

Let F_s denote the number of internal faces with s sides, and let E_b stand for the number of edges of the outer face of the collection of bubbles. The total number of internal faces is

$$F' = F_1 + F_2 + F_3 + \cdots . \tag{8}$$

But since each edge is associated with *two* faces, we can also express edges as

$$2E = E_b + 1 \cdot F_1 + 2 \cdot F_2 + \cdots + s \cdot F_s + \cdots . \tag{9}$$

If we plug (8) and (9) into $3F' - E = 3$, we finally get

$$6 + E_b = 6F' - 2E + E_b = (6-1) \cdot F_1 + (6-2) \cdot F_2 + \cdots + (6-s) \cdot F_s + \cdots .$$

[17] Counterexamples like an edge cutting across two concentric circles are also ruled out by the 120° rule.

We will call the RHS the *hexagonal difference* D_{hex}, since it counts the number of edges which do not belong to a hexagonal face, with a sign depending on whether the face is smaller (+) or larger (−) than a hexagon. So, more simply, we have

$$D_{\text{hex}} = 6 + E_b. \tag{10}$$

The hexagonal difference is 6 more than the number of boundary edges. We give a few simple examples in Fig. 27, with the contribution to D_{hex} indicated in each cell.

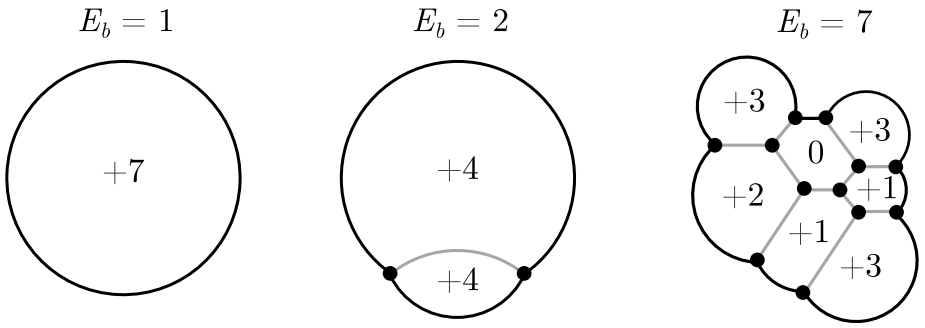


Figure 27: D_{hex}, the sum of numbers in cells, is always 6 more than E_b.

Exercise 4.4. *Large and small faces.*

Equation (10) already tells us some interesting things about bubble networks.

(a) Explain why $D_{\text{hex}} \geq 6$.

(b) Deduce that

$$6 + 5 \cdot F_1 + 4 \cdot F_2 + \cdots + 1 \cdot F_5 \geq 1 \cdot F_7 + 2 \cdot F_8 + \cdots .$$

(c) Suppose a bubble network has two bubbles with four sides and no other small faces. What is the maximum number of 10-sided bubbles?

In general, once we count the "small" faces $F_1, \ldots, F_5$, we can constrain the number of "large" faces $F_7, F_8, \ldots$.

4.3 Hexagons and honeycomb

It's still not clear why most bubbles are hexagonal. At this point, we need to introduce some basic physical intuition. Suppose the foam has overall size $\sim L$. Assuming bubbles have a typical size independent of L, the number of external edges $E_b \sim L$. The total area of the bubble network should scale as $A \sim L^2$. For instance, consider a roughly circular foam of radius L (Fig. 28). If bubbles have average edge length ℓ, independent of L, then $E_b \approx (\pi/\ell)L$, while $A = \pi L^2$.

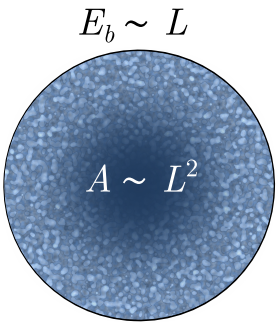


Figure 28: As the foam gets large, the number of outer edges scales as L, and the area as L^2.

It follows that, for large L, the hexagonal difference $D_{\text{hex}} = 6 + E_b \sim L$. The "density" of non-hexagonal edges d_{hex} is just the total hexagonal difference divided by the area of the foam. Since area scales as L^2, the density of non-hexagonal edges scales as

$$d_{\text{hex}} \sim \frac{D_{\text{hex}}}{L^2} \sim \frac{1}{L}. \tag{11}$$

As L becomes larger, edges belonging to non-hexagons become increasingly rare. This explains why a typical cell in a bubble network has six sides, just like Fig. 24.

Exercise 4.5. *Bubble blobs.*

A *bubble blob* is a set of contiguous bubbles in a bubble network. Let E_o denote the number of edges extending outward from the boundary, and E_i the number extending inward.

(a) Explain why the difference from hexagonality is now given by

$$D_{\text{hex}} = 6 + E_i - E_o. \tag{12}$$

(b) Verify that the blob of cells in Fig. 24 satisfies (12).

(c) Repeat the scaling argument above, and conclude that in a large blob, departures from hexagonality become rare.

You may have wondered if the hexagonality of bubbles is related to the fact that bees build honeycombs in a hexagonal lattice. It is! Bees have a clear evolutionary reason to minimize the amount of wax used. Charles Darwin (1809–1882) discusses the hive-making instinct and its relation to fitness in his *Origin of Species* [6]:

> *That motive power of the process of natural selection having been economy of wax; that individual swarm that wasted least honey in the secretion of wax, having succeeded best.*

For honeycomb walls to be minimal, they must obey the 120° rule. If honeycomb cells are equal in size (which bees might prefer for simplicity of construction), then a natural guess at the optimal arrangement is the *hexagonal lattice*. This is the only regular tessellation satisfying the 120° rule.

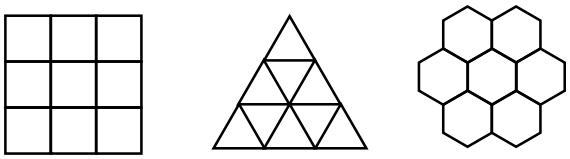
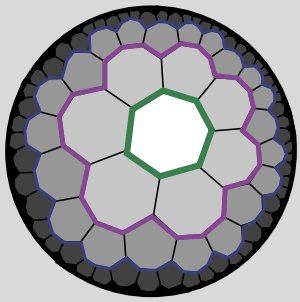


Figure 29: The three regular tessellations of the plane: square, triangle, hexagon.

The *honeycomb conjecture* states that the hexagonal lattice is the *globally* minimal solution among tessellations of the plane with equal cell size. It is hard to verify this guess, since you need to check all possible *irregular* tilings as well as the regular ones. But in 1999, it turned from conjecture into theorem after Thomas Hales gave a formal proof [17]. In Exercise 4.6, we explore the analogous problem for the saddle-shaped *hyperbolic* plane.

Exercise 4.6. *Hyperbolic honeycomb.*

Our scaling argument assumed we were on a regular, Euclidean plane. But we can see what happens if, instead of working on the Euclidean plane, we work on the strangely curved *hyperbolic plane*.



Above, we have tiled the hyperbolic plane with heptagons. Each heptagon has the same area, and sides of equal length, but the curvatures means they must be drawn with different lengths on our flat page!

(a) Find A (in heptagon units) and E_b for the regions enclosed in (i) green, (ii) purple, (iii) blue. Does the ratio E_b/A appear to be decreasing?

(b) Argue that, in general, for $n > 1$ "rings" of heptagons,

$$E_b = 4 \cdot 7^n, \quad A = \frac{1}{6}(7^{n+1} - 1) \approx \frac{7}{6} \cdot 7^n.$$

Hint. Use a geometric sum for A.

This shows that on the hyperbolic plane, the scaling $E_b \sim L$, $A \sim L^2$ no longer holds. Instead, the boundary and area *scale the same way.*

(c) Why does the 120° rule still hold for minimal networks on the hyperbolic plane? *Hint.* What happens when you zoom in on a node?

(d) Show that for ring n, $E_o - E_i = 2 \cdot 7^n$. Using part (c) and similar reasoning to the plane, conclude that a large number of heptagonal rings,

$$D_{\text{hex}} = 6 - 2 \cdot 7^n \approx -2 \cdot 7^n.$$

(e) Finally, show that our heptagonal tiling has

$$d_{\text{hex}} = \frac{D_{\text{hex}}}{A} \approx -\frac{12}{7}.$$

(f) Given Exercise 4.4(a), how can D_{hex} be negative?

The weird properties of hyperbolic space mean that the optimal tessellation depends on the size of the cells. The heptagonal tiling is optimal for the cell size pictured above, at least among *regular* hyperbolic tilings [7]. The "hyperbolic honeycomb conjecture"—that this is optimal among *all* hyperbolic tessellations with this cell size, including the irregular ones—remains open.[18] Perhaps we should breed some hyperbolic bees, and inspect their honeycomb in a few million years!

4.4 The isoperimetric inequality and bubbletoys

The preceding two sections studied two-dimensional bubble foams, assuming there were no fixed nodes. The total length is being minimized, but subject to what constraints? The answer is suggested by our earlier discussion of air pockets, and and by the honeycomb conjecture. The bees have no fixed nodes, since they are not trying to connect anything. Instead, they are trying to build cells to store honey. To simplify the problem, we have considered an infinite number of cells of the same size, but what if the bees only want six? Or want to vary their serving sizes? In general, we can ask for the *minimal length bubble network* enclosing cells of size $A_1, A_2, \ldots, A_n$.

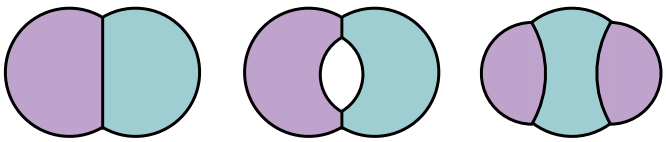


Figure 30: *Left.* The standard double bubble. *Middle.* An empty pocket. *Right.* A split bubble.

In the same way that we are allowed to add nodes to minimal networks to decrease length, we will allow *empty pockets* and *bubble splitting* (Fig. 30) if it helps us reduces length. We will always ask that the bubble network is connected for physical reasons.[19] This leads to...

Box 4.1. *The Planar Bubble Configuration Problem.*

Find the connected bubble network of smallest perimeter enclosing cells of area $A_1, A_2, \ldots, A_n$, allowing empty pockets and split bubbles.

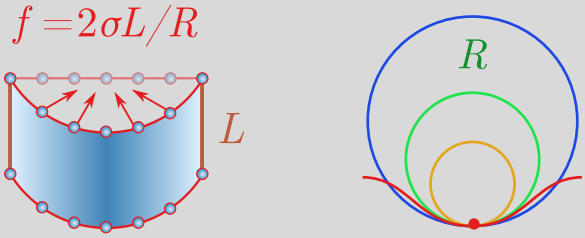
[18] As far as I know, the problem is open for all cell sizes.

[19] If the network is not connected, the disconnected parts can "float" relative to each and will soon collide, forming a connected network.

While we derived the 120° rule directly from minimizing length, the other salient property of bubble networks is that walls are straight or arcs of circles. You can see where this comes from using the physics of surface tension.

Exercise 4.7. *Young-Laplace I.*

The molecules in a bubble wall are attracted to each other. If you try to *bend* the surface, it strains the molecular bonds, which attempt to restore the unstretched state. The amount of bending at a point can be quantified by finding a circle which fits snugly onto the curve (the green circle, below right).



If this snug circle has radius R, we say the bend has *radius of curvature* R. For a bubble wall of height L, the restoring force per unit length of curve is $f = 2\sigma L/R$, where σ is the *surface tension*.

(a) Show that if there are no other forces acting on the wall, it must be straight.

(b) Now consider the effects of *pressure* at a point on the wall. If the pressure on one side is P_{out}, and inside is P_{in}, argue the bend will have radius of curvature

$$R = \frac{2\sigma}{\Delta P}, \tag{13}$$

where $\Delta P = P_{\text{out}} - P_{\text{in}}$. This is called the *Young-Laplace law*. Check it is consistent with (a).

(c) Within a cell, pressure difference equalize very quickly, so it is reasonable to assume pressure is constant on a face of the network. Deduce that bubble walls are either flat or arcs of circles.

Although Exercise 4.7 does involve surface tension, it says nothing about minimizing surface area or solving the bubble configuration problem. It seems plausible that real bubbles do solve this problem, but for the moment, let us view the bubble configurations as *physical conjectures* about minimal-length solutions. In other words, they are guesses made by Nature, awaiting the rubber stamp of mathematical proof.

The simplest physical conjecture is for a single bubble of fixed area A. The only smooth way to draw a single cell is a circle (Fig. 32 (left)), and when Nature is left to its own devices, bubbles tend to assume this form. You can also check (Exercise 4.8) that there is no way to split the single bubble, or introduce air pockets, while maintaining a connected bubble network.

Exercise 4.8. *One bubble to rule them all.*

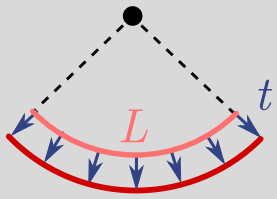
(a) Show that, if we split a bubble into parts which contain a total area A, they cannot share any edges. Explain why the same goes for an empty air pocket and the region outside the bubble configuration.

(b) Argue that splitting a single bubble, or adding empty pockets, violates (a).

Mathematicians as far back as ARCHIMEDES (287–212 BC) have suggested that the circle is the shape of smallest perimeter for a fixed area A, a guess called the *isoperimetric inequality*.[20] This guess wasn't verified until the 19th century, but the modern proof is simple enough to present in outline. The basic idea is to wobble a line and see how the length and enclosed area change. To start with, we consider wobbling the radius of a single circular arc.

Exercise 4.9. *Stretched arcs.*

Suppose an arc of length L and radius R is part of a curve enclosing some area on the plane. Consider extending the radius by a small amount t, where "small" means much smaller than L.



(a) Show that the area enclosed changes by

$$\Delta A \approx Lt. \tag{14}$$

(b) Assuming that the angle subtended by the arc is the same, explain why the length of the arc changes by

$$\Delta L \approx \frac{Lt}{R}. \tag{15}$$

(c) Check that (b) still makes sense for a straight line.

For both ΔA and ΔL, there are some additional corrections, but these will appear as higher powers of t, starting at t^2.

In general, we can take a curve on the plane and chop it up into k small pieces of length $L_1, L_2, \ldots, L_k$ and constant radius of curvature $R_1, R_2, \ldots, R_k$, setting $R_i = \infty$ for any straight lines.[21] Imagine we wobble the curve by *independently* changing the radii for each segment, adding

[20] Saying the circle has the *least perimeter* of all figures of area $A = \pi r^2$ is the same as saying it has *most area* of all figures with the same perimeter $2\pi r$. "Isoperimetric" means "same perimeter".

[21] If we wanted to be rigorous, we would actually chop the line up into an *infinite* number of pieces using calculus.

$t_1, t_2, \ldots, t_k$. Using (14), the total change in area is

$$\Delta A = \Delta A_1 + \Delta A_2 + \cdots + \Delta A_k = L_1 t_1 + L_2 t_2 + \cdots + L_k t_k. \tag{16}$$

From (15), the total change in length is

$$\Delta L = \Delta L_1 + \Delta L_2 + \cdots + \Delta L_k = \frac{L_1 t_1}{R_1} + \frac{L_2 t_2}{R_2} + \cdots + \frac{L_k t_k}{R_k}. \tag{17}$$

It's clear that if we deform the curve so as to preserve area, then $\Delta A = 0$.

But here is the clever part: if the curve is a local minimum of perimeter, then the perimeter looks like a *quadratic* function of the wobbling.[22] But we are making t small enough that we can ignore these quadratic t^2 terms, and keep only the terms proportional to t. Thus, in the approximation we have used to compute (17), a perimeter-minimizing curve has $\Delta L = 0$. You can show in the next exercise that, given the forms for ΔA and ΔL, this is only possible if

$$R_1 = R_2 = \cdots = R_k = R,$$

i.e. the radius of curvature is constant. Thus, the perimeter-minimizing curve has constant R.

Exercise 4.10. *Constant radius of curvature.*

If we vary a perimeter-minimizing curve, then $\Delta L = 0$. If the wobbles also preserve area, then $\Delta A = 0$. We will show that this implies all the radii $R_1, R_2, \ldots, R_k$ are the same.

(a) Suppose that only t_1 and t_2 are nonzero in (16). Show that $\Delta A = 0$ implies

$$t_1 = -\frac{L_2 t_2}{L_1}.$$

(b) Now substitute this into (16), and from $\Delta L = 0$, argue $R_1 = R_2$.

(c) Extend this argument to show that $R_1 = R_2 = R_3 = \cdots = R_k$.

Does constant radius of curvature mean we have a circle? Not necessarily. You could join arcs of the same circle with a "kink". But we can always approximate a kink as closely as we like by a smooth edge which encloses the same area (Fig. 31). This edge will have a *different* radius of curvature, which contradicts our argument! The only smooth, closed curve we can draw, which has the same radius of curvature R at every point, is the circle of radius R itself. This more or less proves the isoperimetric theorem.[23]

So much for a single bubble. The next simplest problems involve two and three bubbles of equal area A. The standard *double bubble* (Fig. 32 (middle)) and *tripple bubble* (Fig. 32 (right)) configurations are drawn below. Since we can now introduce air pockets and splitting, as in Fig.

[22] This is similar in spirit to the argument for equilateral triangles that the network length was an even function of wobble.

[23] Technically, we have only shown that *if* there is a perimeter-minimizing shape of fixed area, it is a circle. But our approximation strategy can also be turned into a proof that the circle does minimize.

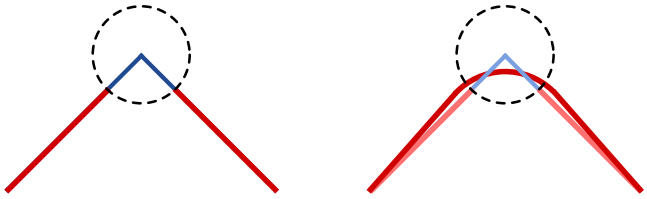


Figure 31: Replacing a kink by a smooth edge which encloses the same area.

30 for two bubbles, it is much harder to show these simple arrangements are minimal. The double bubble was only shown to be minimal in 1993 [11], and the triple bubble in 2002 [31]. As far as I know, no other finite planar bubble configurations are known to be minimal.

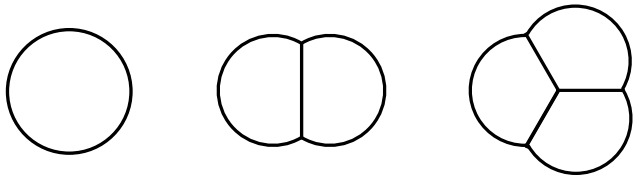


Figure 32: *Left.* A circle. *Middle.* The standard double bubble. *Right.* The standard triple bubble.

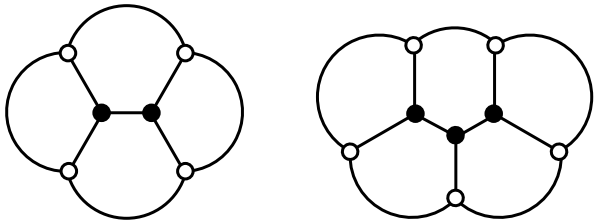


Figure 33: Tinkertoys giving rise to bubbletoys.

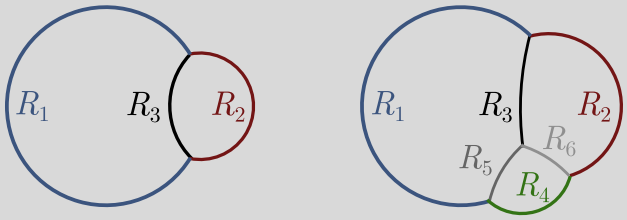
Of course, we might say: forget mathematics, and let Nature be our guide. By carrying out simple experiments, we should be able to see which configurations are predicted by physics. Right? Unfortunately, the same combinatorial explosion that plagued soap bubble computers in §3.3 afflicts planar bubble configurations. One argument is that every tinkertoy gives rise to a bubble configuration, simply by adding arcs to the outside as illustrated in Fig. 33.[24] I call these *bubbletoys*. Incidentally, the two pictured bubbletoys are conjectured to be the minimal planar configurations for four and five equal-area bubbles. This suggests that solving the planar bubble configuration problem is NP-hard, and even finding a bubble configuration which encloses the volumes $A_1, A_2, \ldots, A_n$ is NP-complete. [25] Nature will take increasingly long times to converge on her "conjectures", solve the wrong problem, or both. Either way, we cannot get physics to magically solve our NP-complete problems for us!

[24] More generally, we will have to bend the inner walls to make the sure the pressure difference is balanced by tension. See Exercise 4.11 for more details.

[25] Unlike tinkertoys, where the problem could be easier when we stiched together small tinkertoys, here, there are no external nodes so we only have *large* tinkertoys. And finding the minimal configuration is *much, much harder*, since we not only have an exponential number of tinkertoys, but an infinite slew of configurations that arise from splitting and empty pockets. No wonder we know almost nothing about bubbles!

Exercise 4.11. *Bubble radii and pressure cocycles.* ⛰

We show the standard double and triple bubble for bubbles of different radii below.



This exercise uses physics to simply relate the bubble radii! (There are also derivations from the 120° rule, but they are much messier [16].)

(a) Let's start with the double bubble. By considering pressure differences as across interfaces, explain why

$$\frac{1}{R_2} = \frac{1}{R_1} + \frac{1}{R_3}. \tag{18}$$

Hint. Use Exercise 4.7(b).

(b) We can make this observation more general. Consider moving around a loop on a bubble network. Across each interface, there are pressure differences $\Delta P_1, \Delta P_2, \ldots, \Delta P_n$. Show that, along the loop,

$$\Delta P_1 + \Delta P_2 + \cdots \Delta P_n = 0.$$

This is called the *pressure cocycle condition* in the mathematics literature.

(c) Using the pressure cocycle condition for the triple bubble, calculate that in addition to (18), we have

$$\frac{1}{R_4} = \frac{1}{R_1} + \frac{1}{R_5} = \frac{1}{R_2} + \frac{1}{R_6}.$$

Check that the results of executing a loop around the inner junction are consistent with these relations.

5 Bubbles in three dimensions

So far, we've only considered two-dimensional networks, while the real world has three dimensions. Thankfully, removing the plexiglass changes less than you might expect! Let's start by summarizing what we know about bubble networks. The key result from §2 was the 120° rule. In §4, we learned that bubble walls are straight or arcs of circles, so that they have constant radius of curvature (Exercise 4.7). We also discovered from our treatment of the isoperimetric problem that perimeter-minimizing wall do not have "kinks". We can encode these insights as "laws" for bubble networks:

Box 5.1. *Bubble network laws I.*

1. *No kinks.* Edges are smooth, i.e. no vertices attached to one or two edges.
2. *Constant curvature.* Edges have constant radius of curvature.
3. *The* 120° *rule.* Three edges meet at a junction, separated by 120°.

There is another way to motivate the 120° rule that will prove very useful in three dimensions. The law forbidding kinks means that the fewest edges that can meet at a junction is three. Moreover, meeting at angles of 120° is the most *symmetric* way for incoming edges to be separated. Way back in §2.1, we saw that symmetry had an important role to play in minimizing the length of the network on an equilateral triangle, so perhaps it's unsurprising that the two are connected here. We call this the *minsym principle*. It lets us reformulate our network laws in a slightly different way:

Box 5.2. *Bubble network laws II.*

1. *No kinks.* Edges are smooth, i.e. no vertices attached to one or two edges.
2. *Constant curvature.* Edges have constant radius of curvature.
$\bar{3}$. *Minsym.* At a junction, the minimal number of edges meet symmetrically.

Generalizing to three dimensions is now "easy"!

5.1 Mean curvature

Viewed through a dimensional lens, a planar bubble network is a set of *two-dimensional* cells separated by *one-dimensional* bubble walls. But when bubbles can roam around in three dimensions, the cells are *three-dimensional* volumes separated by *two-dimensional* walls. Although it seems like a whole differen kettle of fish, three-dimensional bubbles are governed by almost exactly the same laws as their planar counterparts. The three-dimensional laws are called *Plateau's laws*, after the Belgian physicist Joseph Plateau (1801–1883) who guessed them by assiduously observing bubbles [23].

"No kinks" seems straightforward: bubble walls are smooth and cannot suddenly terminate.[26] But there are subtleties for the remaining two rules. In a bubble network, edges have constant radius of curvature. What is the analogous statement for surfaces? It turns out in three and more dimensions, the notion of the curvature of a surface is not unique, and different definitions are useful

[26] Unless there is something for them to end on, e.g. a bubble blower. We'll return to this problem below.

for different applications. For our purposes, the relevant notion is *constant mean curvature*. This is a technical notion, and requires a bit more explanation.

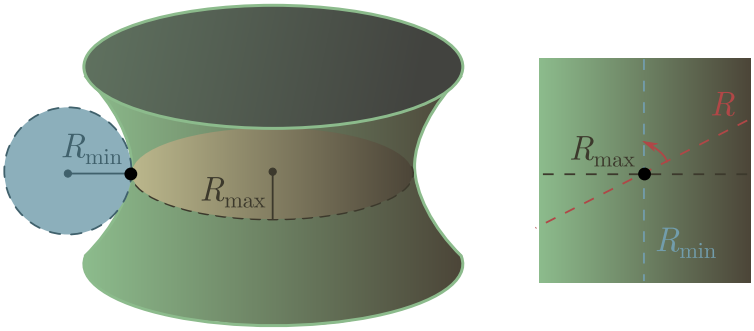


Figure 34: *Left.* A surface, with principal "snug" circles. *Right.* Straight slices through a point.

Suppose we have a two-dimensional surface like the one in Fig. 34 (left). If we take various straight slices through the black point (shown in Fig. 34 (right)), each will give rise to a radius of curvature, i.e. the radius of a circle which fits "snugly" onto the curve at that point, and which is perpendicular to the surface. As we rotate the red slice in Fig. 34 (right), the radius of curvature R will vary, producing a maximum value R_{max} and minimum value R_{min}. The reciprocals $1/R_{\text{max}}$ and $1/R_{\text{min}}$ are called the *principal curvatures*. Note that a radius curvature can be *negative* if it is outside the surface, as in Fig. 34 (left).[27]

The *mean curvature* H is defined as the sum of principal curvatures:

$$H = \frac{1}{R_{\text{max}}} + \frac{1}{R_{\text{min}}}. \tag{19}$$

A *constant mean curvature (CMC)* surface is one where the mean curvature H is the same everywhere on the surface. Notice that, as in Fig. 34 (right), it is always the case that the principle curvatures R_{max} and R_{min} are measured along orthogonal slices. We call this the *orthogonal circle theorem*.[28] To generalize the constant curvature rule from planar bubble networks, we take bubble surfaces to be CMC. You can explore some of the physics behind this in Exercise 5.2.

Exercise 5.1. *Spheres are CMC.*

Show that a sphere of radius R has constant mean curvature $H = 2/R$. *Hint.* The slice normal to the sphere at any point is a great circle.

Exercise 5.2. *Young-Laplace II.* ▲

In Exercise 4.7, we saw the Young-Laplace law (13) for a bubble wall:

$$\Delta P = \frac{2\sigma}{R},$$

for $\Delta P = P_{\text{out}} - P_{\text{in}}$ and R the radius of curvature of the wall.

[27] Sometimes, outside and inside aren't well-defined, so you just make an aribtrary choice, and attach a minus sign to any circles which are outside. The sign of curvature depends on this choice.

[28] Unfortunately, it would take us too far afield to prove it here.

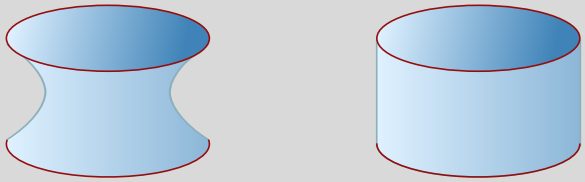
(a) Viewing a bubble wall as a surface, argue that $1/R_{\max} = 0$.

(b) Using the orthogonal circle theorem, deduce that $H = 1/R$ is the mean curvature of the wall. Hence, the Young-Laplace law can be written

$$\Delta P = 2\sigma H. \tag{20}$$

This turns out to be the correct form for an arbitrary surface!

(c) If we dip two identical circular bubble blowers in soap film (red below), the surface that results is typically like the one below left, rather than a cylinder:



Give a qualitative explanation, using (20) and mean curvature.

(d) Using Exercise 5.1, what is the *smallest* spherical bubble that can form in the atmosphere? Atmospheric pressure is $P = 10^5$ N/m^2 and the surface tension of soapy water is $\sigma = 7 \times 10^{-2}$ N/m. Can a spherical bubble form in space?

5.2 Plateau's laws

Finally, we have to generalize the "minsym" principle to three dimensions. The "no kink" requirement means that we cannot have two walls meeting at an angle, since that would introduce a kink, and if there is no angle, they may as well be count as part of the same wall. Thus, we must have at least three walls meet at any junction of walls. According to the minsym principle, precisely three faces should meet (minimum) separated by 120° (symmetry), as in Fig. 35 (left). In fact, this is exactly what we need to get the 120° rule in a planar bubble network, since the walls are secretly two-dimensional and vertical oriented between the plexiglass plates (Fig. 35 (middle)).

The edge along which three bubble walls meet is called a *Plateau border*. These borders themselves can intersect! Minsym requires us to figure out the minimum number to avoid kinks, and the most symmetric arrangement thereof. Clearly, we need at least three, since otherwise we can arrange a junction of three walls with a kink, as in Fig. 35 (right).

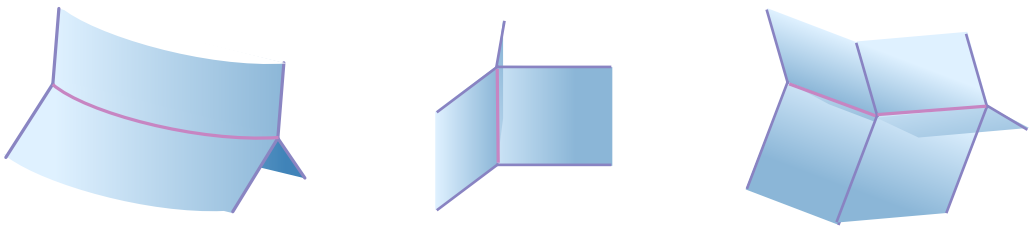


Figure 35: *Left.* Three faces meeting at a border. *Middle.* Vertical bubble walls. *Right.* A kink.

Can we have exactly three? It's not hard to see that the three sets of three faces cannot be connected smoothly, simply because we have an odd number of faces! You can check the details in Exercise 5.3. This exercise also shows that it *is* possible to connect the faces smoothly for four sets of Plateau borders. Thus, the minsym principle suggests that *precisely* four Plateau borders should meet in the most symmetric arrangement. Symmetry is maximized by shooting out the Plateau borders *tetrahedrally*, i.e. from the center towards the corners of a regular tetrahedron (Fig. 36). If you like vectors, you can play around with the geometry in Exercise 5.4.

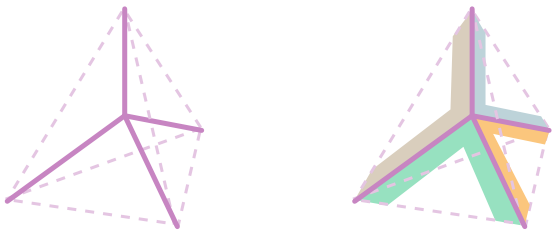


Figure 36: *Left.* Four Plateau borders meeting tetrahedrally. *Right.* Smoothly connected walls.

Having defined constant mean curvature, and worked out the implications of the minsym principle in three dimensions, we are finally in a position to state the laws Plateau discovered [23]:

Box 5.3. *Plateau's laws.*

1. *No kinks.* The faces in a soap film are smooth.
2. *Constant curvature.* Any face has constant mean curvature.
3. *Minsym I.* Three faces always meet at a Plateau border, separated by 120°.
4. *Minsym II.* Plateau borders always meet tetrahedrally at a vertex.

These are empirical observations about bubbles. While the constant curvature condition follows from the Young-Laplace law (Exercise 5.2), and Minsym I from the 120° rule, it is not at all obvious that a tetrahedral arrangement of Plateau borders minimizes area. Minimizing subject to what constraints? (Feel free to have a guess now.) Is every configuration satisfying Plateau's laws a local minimum, subject to these constraints? And does every such locally minimal solution satisfy Plateau's laws? (These are harder to figure out without a doctorate in math.) Read on to find out!

Exercise 5.3. *Plateau borders.*

Let's check we need four Plateau borders in order to smoothly connect walls. Consider some number of Plateau borders meeting at a vertex. Each border has three associated bubble walls.

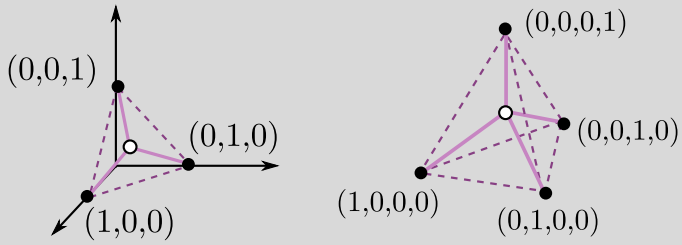
(a) Explain why "no kinks" requires each wall to connect smoothly to another.

(b) Argue that this is impossible for an odd number of borders meeting at a node.

(c) Show explicitly it is possible for four Plateau borders to smoothly connect. *Hint.* Add the two remaining walls in Fig. 36 (right).

Exercise 5.4. *Simplices.* 🛥

The equilateral triangle and the tetrahedron are part of a family of symmetric shapes called *simplices*. We can describe them using vector analysis.



(a) We can embed the vertices of an equilateral triangle in three dimensions as

$$\Delta_3 = \{(1,0,0), (0,1,0), (0,0,1)\}.$$

Why is this a maximally symmetric arrangement of three points?

(b) The *center* of the triangle is just the average of the vertices. Show that the vectors from center to vertices have length $\sqrt{2}$ and are given by

$$V_3 = \left\{\tfrac{1}{3}(-2,1,1), \tfrac{1}{3}(1,-2,1), \tfrac{1}{3}(1,1,-2)\right\}.$$

(c) Using the formula

$$\theta = \cos^{-1}\left(\frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{u}||\mathbf{v}|}\right),$$

check that the vectors in V_3 make angle $120^\circ = \cos^{-1}(-1/2)$ with each other.

(d) We can embed the tetrahedron in *four* dimensions as

$$\Delta_4 = \{(1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1)\}.$$

Show that the vectors from center to vertex have length $\sqrt{3}$, and make angles

$$\theta = \cos^{-1}\left(-\frac{1}{3}\right) \approx 109.5^\circ.$$

The tetrahedron is just a higher-dimensional version of an equilateral triangle! We can continue in this fashion, defining the *n-simplex* Δ_n as a maximally symmetric arrangement of n points in n dimensions:

$$\Delta_n = \{(1,0,\ldots,0), (0,1,\ldots,0), \ldots, (0,\ldots,0,1)\}.$$

(e) Check that the distance from the center of Δ_n to each vertex is $\sqrt{n}$, and that any two such vectors make an angle

$$\theta = \cos^{-1}\left(-\frac{1}{n}\right).$$

As n gets large, confirm these vectors are almost orthogonal.

(f) Extrapolate the minsym principle to higher-dimensional foams. In other words, if the universe has n dimensions, make a guess at Plateau's laws.

5.3 Bubbles and wireframes

As you might have guessed, Plateau's laws are related to the *three-dimensional* version of the planar bubble configuration problem outlined in Box 4.1. Instead of enclosing areas $A_1, A_2, \ldots, A_n$, we want to enclose volumes $V_1, V_2, \ldots, V_n$ with a bubble film of minimal surface area. As before, we allow for empty pockets and split bubbles. We state the optimization problem as follows:

> **Box 5.4.** *The Bubble Configuration Problem.*
>
> Find the connected bubble film of smallest area enclosing volumes $V_1, V_2, \ldots, V_n$, allowing air pockets and split bubbles.

Surfaces with bubbles of fixed volume, and which locally minimize area, also satisfy Plateau's laws, as mathematican Jean Taylor proved in her 1976 tour-de-force [28].[29] The converse is not true, since we can find bubbles satisfying Plateau's laws that are not stable (Fig. 38).

The planar bubble configuration problem (Box 4.1) is a special case of the three-dimensional bubble configuration problem, where we put the foam between plates. This implies that local minima satisfy the bubble network laws (Box 5.1), since these are simply Plateau's laws in the case where bubble walls are vertical, and because they are vertical, there are no vertices at which Plateau borders intersect. And since planar bubbles are hard, three-dimensional bubbles are hard! Physically speaking, we expect that foams will take longer and longer to converge to a minimum, or answer the wrong question, if we force them to compute for us.

Even when Nature does make conjectures, they can be bewilderingly hard to prove. The simplest example is a single bubble of volume V. Experience suggests that a lone bubble is always spherical, as in Fig. 37 (left). The corresponding conjecture is that the area-minimizing surface of volume V is a sphere. This is the three-dimensional version of the isoperimetric theorem for circles in §4.4.

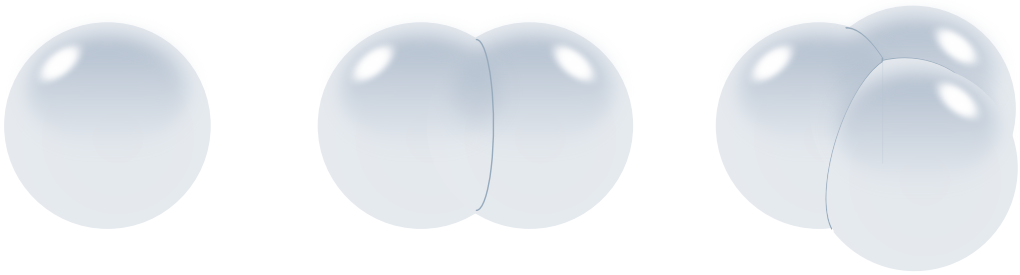


Figure 37: *Left.* A single spherical bubble. *Middle.* The standard double bubble. *Right.* The standard triple bubble.

The proof is remarkably similar. We split the surface into many small patches of area $A_1, A_2, \ldots, A_k$, with mean curvature $H_1, H_2, \ldots, H_k$. If these areas are "pushed out" a distance

[29] I want to point out that the 120° rule was more or less proved as soon the problem was stated. It took over 100 years for the teatrahedral rule to go from empirical observation to mathematical proof. It's much harder!

$t_1, t_2, \ldots, t_k$ normal to the surface, the volume and area change as[30]

$$\Delta V = A_1 t_1 + A_2 t_2 + \cdots + A_k t_k$$
$$\Delta A = A_1 t_1 H_1 + A_2 t_2 H_2 + \cdots + A_k t_k H_k.$$

If the wobbling preserves volume, then $\Delta V = 0$, and if area is locally minimized, then $\Delta A = 0$ as before. We can then repeat our argument word for word to conclude that $H_1 = H_2 = \cdots = H_k$. Mean curvature is the same everywhere, and we have a CMC surface![31]

In the plane, there was exactly one way to have a smooth curve with constant radius of curvature. In three dimensions, there are all sorts of exotic CMC surfaces. But it turns out that the sphere is the *only* CMC surface that enclose a finite volume, as proved by Aleksandr Aleksandrov (1912–1999) in 1958 [2].[32] The same argument we gave in Exercise 4.8 shows that empty pockets and splitting bubbles will not help. Thus, we have proved the isoperimetric theorem in three dimensions: the sphere is the surface of smallest area enclosing a given volume V.

The next simplest problem is two bubbles of equal volume V. Again, Nature seems to prefer the "standard double bubble", with two spheres fused at a single Plateau border (Fig. 37 (middle)) over its non-standard competitors. One of these competitors is the "donut" configuration, where a single bubble is squeezed into the shape of an apple core by a donut-shaped ring around the outside (Fig. 38). This bubble satisfies Plateau's laws, but turns out to be unstable, and jiggling the donut will cause it to collapse into the standard double bubble [27]. It wasn't until 2002 that the standard double bubble was shown to be minimal [19]. Similarly, we often observe the standard triple bubble for three cells of volume V (Fig. 37 (right)). No one knows if this is truly minimal, so it remains the *triple bubble conjecture*.

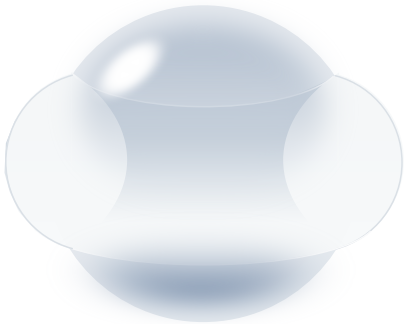


Figure 38: An unstable double bubble, consisting of an apple core wrapped in a donut.

If there is a three-dimensional bubble configuration problem, it stands to reason there should be a three-dimensional minimal network problem. In the minimal network problem, Box 2.1), we had to find a network of shortest length connecting some set of fixed nodes. A node is a *zero-dimensional* object—it has no extent at all! If we raise the number of dimensions of the configurable object, going from a one-dimensional graph to a two-dimensional surface, perhaps we should raise the dimensions of the fixed object, going from fixed points to *fixed curves*. The suggests the following task:

[30] It's not hard to see that volume changes this way, but area is the tricky one, and I won't prove it here.

[31] This proof actually generalizes to higher dimensions, where the mean curvature is $H = 1/R_1 + 1/R_2 + \cdots + 1/R_n$ for n mutually orthogonal principal curvatures.

[32] Well, almost. If the surface is allowed to intersect itself, there is an odd three-lobed donut called the *Wente torus*.

Box 5.5. *Wireframe problem.*

Given some fixed curves $C_1, C_2, \ldots, C_n$ in three-dimensional space, find a soap film of minimal area that connects them.

These fixed curves are called *wireframes*, since physically speaking, we can implement them with twisted pieces of wire. Dunking wire into soapy water gives soap bubbles something to hold onto, and as with plexiglass and screws, we have an analogue computer to solve our problem for us. We give two very beautiful examples in Fig. 39: the *catenoid*, a surface forming between two rings, and the *tesseract* formed when we dip a wireframe cube. The cube creates a second, slightly puffed out[33] inner cube, and then connects corresponding corners with Plateau borders.

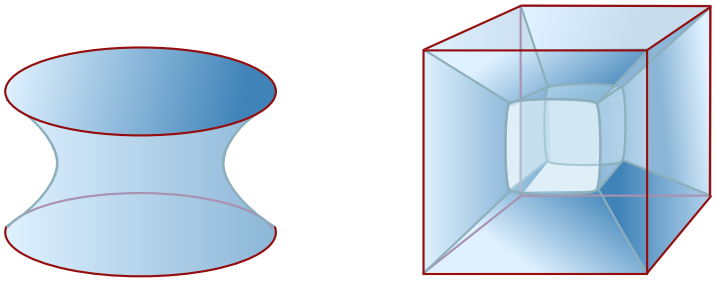


Figure 39: *Left.* The catenoid, a surface with mean curvature zero, which forms between identical wireframe rings. *Right.* The tesseract formed from a wireframe cube.

Since this generalizes the minimal network problem (the screws are a particularly boring type of wireframe), the wireframe problem is NP-hard. We can dip some arbitrarily complicated piece of wire into the soap, but when we pull it out, it may take a very long time—longer than the age of the universe in some cases—for the soap bubbles to converge on a stable solution. Or it will solve a different problem altogether. This is yet another physical prediction!

But it's not obvious there is a solution at all. If we dunk some random wireframe into the water, the bubble film that connects them must satisfy Plateau's laws, except along the wire itself, in the same way that minimal networks satisfy the 120° rule at a hub but not at a fixed node. But do Plateau's laws always allow a solution? Perhaps we can defeat Nature by giving it some wacky curve it cannot connect with soap film. The intuitive physical argument is that we can simply dip our wireframe in and see what comes out. But as we've just argued, for a complicated enough problem, it may take a very, very, very long time to converge. And I find an argument less convincing if I am guaranteed to die before it successfully terminates! The question of the *existence* of a solution to the wireframe task is called *Plateau's problem*. In the 1930s, mathematicians Jesse Douglas (1897–1965) and Tibor Radó (1895–1965) independently showed these solutions always exist [8, 25]. Even if takes longer than the lifetime of universe, Nature will eventually get there.

5.4 Space-filling foams

In the last section, we will consider the infinite cell version of the bubble configuration problem (Box 5.4). This is like honeycomb on the plane, but now we wish to partition an infinite number

[33] To ensure borders meet tetrahedrally.

of equal volumes, rather than an infinite number of equal areas. For this partition to minimize the surface area per cell, we know it must satisfy Plateau's laws. In the same way that the 120° had powerful for the structure of bubble networks, the tetrahedral law has similar implications for the structure of three-dimensional foams. To explore these, we first need to extend Euler's formula (6) to include multiple bubbles.

In a finite configuration of soap bubbles, let N denote the number of vertices (joining Plateau borders), E the number of Plateau borders, F the number of bubble faces (which meet at a border), and C the number of enclosed bubble cells. As before, we will count the region outside the bubble configuration as a cell as well. Recall from §4.2 that Euler's formula also applies to polyhedron like the cube. This has only two cells, the inside and the outside. But we can divide up this internal cell into multiple internal cells by adding inner walls.

If there are C cells altogether, there are $C-1$ internal cells, and $C-2$ "extra" internal cells compared to a regular polyhedron. For each extra cell, we can remove an internal face so that it forms a single cell with one of its neighbours. This leaves something we can flatten into a planar graph, with $F' = F - (C-2)$ faces, and hence by Euler's formula

$$N - E + F' = 2.$$

Rearranging gives Euler's "foamula":

$$N - E + F = 2 + (C-2) = C. \tag{21}$$

Equation (21) is generally true for a polyhedron with extra internal cells, whether or not it also satisfies Plateau's laws.

Using this formula, along with Plateau's laws, you can show that bubble faces tend to have *less* than six sides. More precisely, if F_{avg} is the average number of faces per bubble cell, E_{avg} the average number of edges around the boundary of a cell, and e_{avg} the average number of edges per face, you can show in Exercise 5.5 that

$$F_{\text{avg}} = \frac{1}{3}E_{\text{avg}} + 2 = \frac{12}{6 - e_{\text{avg}}}. \tag{22}$$

Since F_{avg} is positive, it follows that $e_{\text{avg}} < 6$, i.e tend to be "sub-hexagons".

Exercise 5.5. *Sub-hexagonal faces.* ▲

(a) From Plateau's fourth law and the handshake lemma (1), argue that $E = 2N$.

(b) Let F_{avg} denote the average number of faces per cell and E_{avg} the average number of edges per cell. Show that

$$F_{\text{avg}} = \frac{2F}{C}, \quad E_{\text{avg}} = \frac{3E}{C}.$$

Hint. You may assume that, like in a bubble network, a face in a bubble foam always has different cells on either side.

(c) From (21), deduce the relation between average number of edges and faces:

$$3F_{\text{avg}} - E_{\text{avg}} = 6.$$

This is analogous to the result $3F - E = 6$ for bubble networks.

(d) Let e_{avg} be the average number of edges per face. Derive the relation

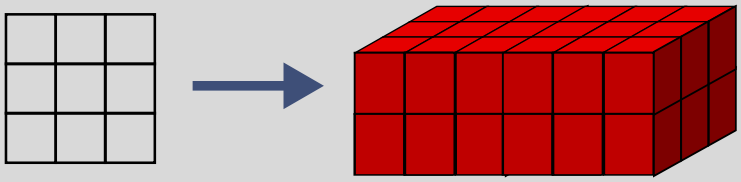
$$F_{\text{avg}} = \frac{12}{6 - e_{\text{avg}}}.$$

Hint. Write E_{avg} in terms of F_{avg} and e_{avg}. Don't forget to handshake!

Since we are discussing *averages*, they continue to make sense even if the foam is infinite! The simplest infinite foams are those in which each bubble is the same, so the bubbles forms a *space-filling tessellation*. This is the three-dimensional version of the plane tessellations we saw earlier in Fig. 29. First, we will rule out a simple scheme based on plane tessellations.

Exercise 5.6. *Prisms.*

One way to tessellate space is to take a regular tessellation of the plane, then extend it in the perpendicular direction to form a layer of prisms (as below). We can then stack these layers on top of each other to tessellate space.



Explain why no prism-based tessellation satisfies Plateau's laws.

Our next step is to consider the three-dimensional analogue of regular polygons, the *Platonic solids* (Fig. 40). These are polyhedra whose faces are identical regular polygons. Could any of these describe an infinite soap foam? Of these solids, only the cube can tessellate space by itself, but since this is a prism-based tessellation (it is an extruded version of the rectangular tiling of the plane), Exercise 5.6 rules it out. We need to work harder!

But before we move on, we should mention that the Platonic solids have a beautiful alternative interpretations as *regular tessellations of the sphere*. (They are not the only regular tesselllations, as you can check in Exercise 5.7.) If we "inflate" the polyhedron into a sphere, the faces will form regular polygons. We show a few of these on the upper row of Fig. 41. And in fact, the solids with three edges meeting at a node—tetrahedron, cube, dodecahedron—obey the 120° rule. If we "flatten" them in the same way we did the cube[34] (Fig. 26) we get the bubble networks shown on the bottom row of Fig. 41. The first is the standard triple bubble we encountered in Fig. 32. It is remarkable that this pattern, which so beautifully exhibits the 120° rule, also shows up as a bubble configuration, the tetrahedron, and a spherical tessellation!

Like Exercise 4.6, the optimal tessellation now depends on the cell size. But we can ask if these regular tessellations minimize the total amount of wax required for a spherical honeycomb with given

[34]Technically, we have drawn a very special flattening called the *stereographic projection*. This is what the vertices would look like to an observer positioned on top of the sphere, or if a lantern at the same point cast the shadows of edges onto the plane.

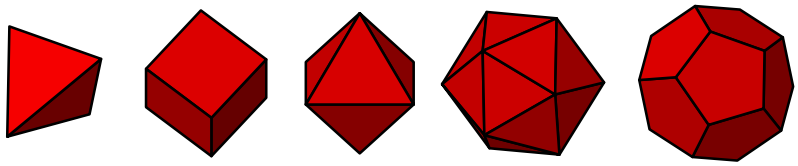


Figure 40: From left to right: tetrahedron, cube, octahedron, icosahedron, dodecahedon.

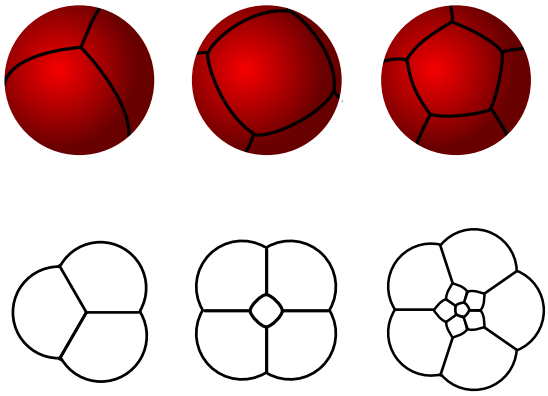


Figure 41: The tetrahedron, cube, and dodecahedron as tessellations of a sphere (above) and bubble networks (below).

cell size. The "spherical honeycomb conjecture" is proved for the dodecahedron and tetrahedron [18, 10], but apparently remains a conjecture for the cube. And in case you're wondering, our hexagonality argument from §4.3 does not apply simply because the sphere does not allow the network to get large!

Exercise 5.7. *Platonic solids.*

In this exercise, we'll classify the regular tessellations of the sphere. Recall Euler's formula $N - E + F = 2$.

(a) Suppose each face of the tessellation has a edges, and each node joins up with b edges. Show that

$$2E = Na = Fb.$$

(b) Using Euler's formula, deduce that

$$E = \frac{2ab}{2(a+b) - ab}. \tag{23}$$

(c) Since E is a whole number, so is the RHS of (23). We will find all possible solutions! To begin with, argue that we can interchange the roles of a and b, so a tessellation with a edges per face and b edges per node also gives a tessellation with b edges per face and a edges per node. These tessellations are said to be *dual*.

(d) If $a = 2$, what are the possible values of b? Draw the corresponding patterns on the sphere and their duals.

(e) The denominator of (23) must be positive. Show that this implies

$$a < \frac{2b}{b-2},$$

and hence there are no solutions for $b \geq 6$ and $a \geq 3$.

(f) The numerator in (23) is even, for a, b whole numbers. Argue that, in order for E to be a whole number, at most one of a and b can be odd.

(g) Finally, using part (f), conclude that the remaining solutions that only five combinations of a and b are allowed for $3 \leq a, b \leq$. Check that each of these gives a Platonic solid.

Let's return to the problem of space-filling foams. None of the Platonic solids creates a tessellation satisfying Plateau's laws, so we must turn to the the "semi-regular" polyhedra: the 13 *Archimedean solids*, whose vertices all look alike, and the 13 *Catalan solids*, whose faces all look alike. The faces of Archimedean solids are regular polygons, while the Catalan solids do not. Rather than describe these all in detail, we skip to the shortlist which can tessellate space:

- the *rhombic dodecahedron*, a Catalan solid with twelve rhombic faces;
- the *truncated octahedron*, an Archimedean solid we get by snipping off an octahedron's corners.

These are shown in Fig. 42, including the "snipping" of a single corner of the octahedron.

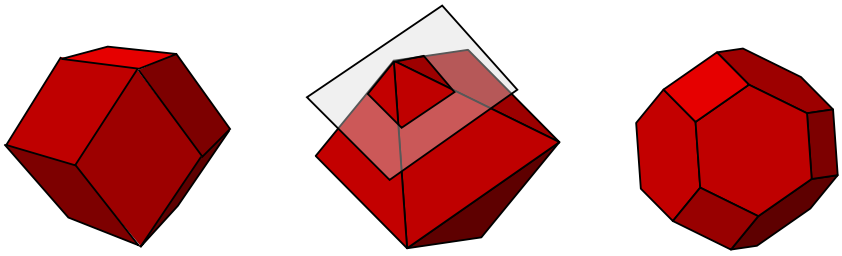


Figure 42: *Left.* The rhombic dodecahedron, with 12 rhombic faces. *Middle.* Snipping off the corner of an octahedron. *Right.* Doing this for all six corners gives the truncated octahedron.

We can take each face of the rhombic dodecahedron and extrude it to form a pyramid, so that each rhombic face is replaced by four triangular faces. The result is called a *stellated rhombic dodecahedron*, with "stellated" meaning "star-like". It is also called *Escher's solid*, since it features in Escher's marvellous lithograph *Waterfall* (Fig. 43). Remarkably, the extrusions interlock in such a way that Escher's solid continues to tessellate space. Perhaps it should be called a "testellation"![35] We can now use (22) to eliminate all but one candidate.

[35] Admittedly, I've included Escher's solid mainly to justify this pun.

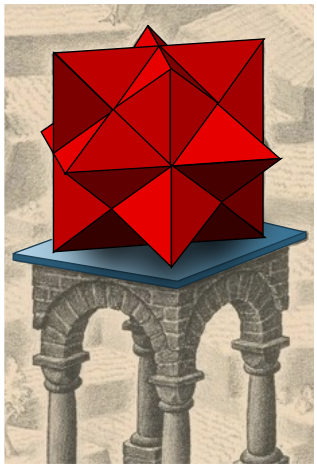


Figure 43: Stellated rhombic dodecahedron. Adapted from *Waterfall* (1961), M. C. Escher.

Exercise 5.8. *The Kelvin structure.*

We now have three candidates for space-filling foam:

- the rhombic dodecahedron, with two rhombic faces;
- Escher's solid, with 48 triangular faces; and
- the truncated octahedron, with eight hexagonal faces and six squares.

Show that only the truncated octahedron satisfies (22).

This truncated octahedron tessellation (Fig. 44) is called the *Kelvin structure*, after physicist William Thomson, 1st Baron Kelvin (1824–1907), who conjectured it was the most efficient way to separate equal volume cells.[36] Kelvin's conjecture, often called the *Kelvin problem*, is the three-dimensional version of the honeycomb conjecture: are there any *irregular*, equal-volume tessellations of space which are more efficient than the Kelvin structure? Put differently, what structure should four-dimensional bees store their honey in?

Like the hexagonal lattice, the Kelvin structure is the only regular tessellation which is locally minimal. But the space of possibilities is much richer in three dimensions than in two, and in 1993, Denis Weaire and Robert Phelan discovered they could improve on the Kelvin structure by weaving together two funny-shaped cells of equal volume [30]. The arrangement is called the *Weaire-Phelan structure*, shown in Fig. 45. The tessellation uses two different cell types:

- an irregular dodecahedron A_0, with twelve pentagonal faces; and
- a 14-hedron A_2 with two hexagonal and twelve pentagonal faces.

We will explain the notation below.

While we have (cautiously) extolled the virtues of soap bubble computers, Weaire and Phelan took the amusing approach of *simulating* foams on a regular digital computer. They were using the

[36] Note that we need to bend the edges a little to ensure they meet at $\theta \approx 109.5°$, in accord with Plateau's laws.

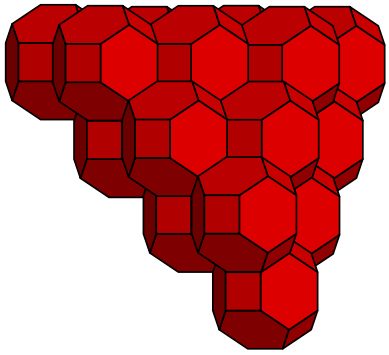


Figure 44: The Kelvin structure, made by tiling truncated octahedra.

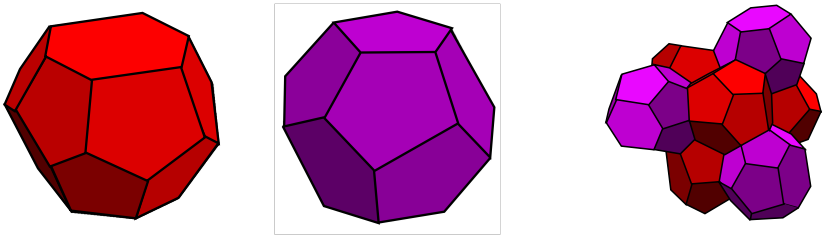


Figure 45: *Left.* The 14-hedron A_2. *Middle.* The irregular dodecahedron A_0. *Right.* A chunk of the Weaire-Phelan structure.

software of the physical universe, but not the hardware! The problem is that real soap films are finnicky, and it is experimentally challenging to arrange equal-volume bubbles.[37] Even when you ask it to solve the correct problem, it often returns the Kelvin structure!

But Nature knew about this structure long before the ingenious monkeys it evolved. In 1931, chemists noticed that layer of tungsten[38] formed by electrolysis have an odd chemical structure; a couple of years later, the same structure was observed in chromium silicide Cr_3Si. More spectacularly, in 1953, chemists observed this same arrangement in vanadium silicide V_3Si, which is a *superconductor*[39] when cooled to low enough temperatures. Chemists F. C. Frank and J. S. Kasper began formally studying this odd structure and its cousins, together called *tetrahedrally closed-packed (TCP)* structures [12, 13]. The original TCP structure occurring in vanadium silicide is the *A15 phase*. It turns out this is precisely the arrangement you get if you put an atom at the center of each polyhedron in the Weaire-Phelan structure!

Above, we introduced the 12- and 14-sided polyhedra A_0 and A_2. It turns out there are *four* polyhedra appearing in the TCP structures studied by Frank and Kasper, pictured below (Fig. 46). Each has twelve pentagonal faces, and either 0, 2, 3 or 4 hexagonal faces, sp we label them according to the number of hexagonal faces they have. No one has classified all the combinations that are possible with these TCP polyhedra, though it seems there are an infinite number of possibilities. You can explore the constraints from graph theory in Exercise 5.9. Whether any of these possibilities beats Weaire-Phelan is an open problem! See [27] for further discussion.

[37] Weaire discusses some of these experimental challenges in his entertaining commentary [29].
[38] The metal light bulk filaments are made from.
[39] This means it exhibits no electrical resistance.

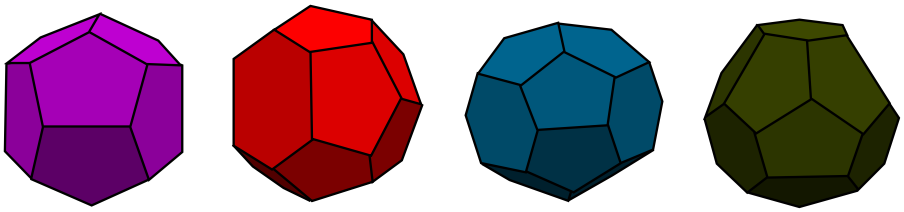


Figure 46: The polyhedra A_0, A_2, A_3 and A_4 appearing in TCP structures.

Exercise 5.9. *TCP structures.*

(a) Suppose we can build a tessellation out of the TCP polyhedra A_0, A_2, A_3, A_4 in the ratio

$$a_0 : a_2 : a_3 : a_4.$$

Using (22), what are the constraints on the possible ratios?

(b) The Weiare-Phelan structure (A15 phase) interleaves A_0 and A_2 polyhedra. What is their ratio?

(c) The *Z phase* has A_0, A_2 and A_3 in the ratio $a_0 : 2 : 2$. What is a_0?

(d) The *C15 phase* interleaves A_0 and A_4 polyhedra only. What is the ratio?

(e) Finally, show that we can write *any* ratio for a TCP structure as a combination of A15, C15 and Z ratios.

There is a second route to the Weaire-Phelan structure through chemistry. Instead of placing atoms at the centre of polyhedra, we can place them *vertices* where Plateau borders join. The arrangement corresponding to the Weaire-Phelan is then called the *Type I clathrate structure*, occuring in *clathrate hydrates*. "Clathrate" is from the Latin *clathratus*, meaning "with bars", while "hydrate" refers to water, since clathrate compounds are tiny, elaborate cages of ice which trap light gases like oxygen, methan, carbon dioxide, and so forth. The gas molecules provide just enough structural support to stop the cage from collapsing into conventional ice or water!

Clathrates are found in all sorts of strange places, from the deep ocean floor to the outer solar system. Ironically, they provide a vast but non-renewable energy source (since clathrates can trap natural gases like methane), as well as a possible means of controlling climate change (since clathrates can trap carbon dioxide). That there is natural and fairly tight chain of associations leading from railways to superconductors, trans-Neptunian objects and climate change is little short of miraculous.

6 Conclusion

> "Have you guessed the riddle yet?" the Hatter said, turning to Alice again. "No, I give it up," Alice replied: "what's the answer?" "I haven't the slightest idea," said the Hatter. "Nor I," said the March Hare. Alice sighed wearily. "I think you might do something better with the time," she said, "than waste it in asking riddles that have no answers."

We now know why a soap bubble is like a railway. But on our journey to the answer, we passed the eerie shadows of riddles in the mist which are not only unanswered, but in some cases, will *never* be answered as a matter of physical principle. What is the shortest rail network connecting all the cities of North America? We could run all the computers in the world (digital or analogue) until the sun explodes, without success. We have a better chance of finding out the best way for hyperbolic bees to store their honey, beating the Weaire-Phelan structure with exotic tessellations, or proving the optimality of a bubble array.

Did the Hare and the Hatter waste Alice's time? And has all this discussion of bees and bubbles been a railroad to nowhere? Asking unanswerable riddles is not a waste of time when it leads us to new insights and new ways of understanding the old riddles, the right questions to ask and how to ask them. I hope that we have learnt a little in this direction. But more importantly, like Carroll's unanswerable condundrum, it teaches us to wonder at the world in its manifold secret connections, which we cannot always articulate, but know are present and bind the whole together.

References

[1] Aaronson, S. NP-complete problems and physical reality. *SIGACT News 36*, 1 (Mar. 2005), 30–52.

[2] Aleksandrov, A. D. Uniqueness theorems for surfaces in the large. *V. Vestnik Leningrad University 13* (1958).

[3] Arora, S. Polynomial time approximation schemes for Euclidean Traveling Salesman and other geometric problems. *J. ACM 45*, 5 (Sept. 1998), 753–782.

[4] Chung, F., and Graham, R. A new bound for Euclidean Steiner minimal trees. *Annals of the New York Academy of Sciences 440* (12 2006), 328 – 346.

[5] Courant, R., and Robbins, H. *What is Mathematics?: An Elementary Approach to Ideas and Methods*. Oxford University Press, 1941.

[6] Darwin, C. *On the Origin of Species by Means of Natural Selection*. Murray, London, 1859. or the Preservation of Favored Races in the Struggle for Life.

[7] Di Giosia, L., Habib, J., Hirsch, J., Kenigsberg, L., Li, K., Pittman, D., Petty, J., Xue, C., and Zhu, W. Optimal monohedral tilings of hyperbolic surfaces. *arXiv e-prints* (Nov. 2019), arXiv:1911.04476.

[8] Douglas, J. Solution of the problem of Plateau. *Transactions of the American Mathematical Society 33*, 1 (1931), 263–321.

[9] Dreyer, D., and Overton, M. Two heuristics for the Euclidean Steiner tree problem. *Journal of Global Optimization 13*, 1 (Jan. 1998), 95–106.

[10] Engelstein, M. The least-perimeter partition of a sphere into four equal areas. *Discrete and Computational Geometry 44* (2010), 645–653.

[11] Foisy, J., Alfaro, M., Brock, J., Hodges, N., and Zimba, J. The standard double soap bubble in $\mathbf{r}^2$ uniquely minimizes perimeter. *Pacific J. Math. 159*, 1 (1993), 47–59.

[12] Frank, F. C., and Kasper, J. S. Complex alloy structures regarded as sphere packings. I. Definitions and basic principles. *Acta Crystallographica 11*, 3 (Mar 1958), 184–190.

[13] Frank, F. C., and Kasper, J. S. Complex alloy structures regarded as sphere packings. II. Analysis and classification of representative structures. *Acta Crystallographica 12*, 7 (Jul 1959), 483–499.

[14] Garey, M. R., Graham, R. L., and Johnson, D. S. The complexity of computing steiner minimal trees. *SIAM Journal on Applied Mathematics 32*, 4 (1977), 835–859.

[15] Gilbert, E. N., and Pollak, H. O. Steiner minimal trees. *SIAM Journal on Applied Mathematics 16*, 1 (1968), 1–29.

[16] Glassner, A. Soap bubbles. 2 [computer graphics]. *IEEE Computer Graphics and Applications 20*, 6 (2000), 99–109.

[17] Hales, T. C. The Honeycomb Conjecture. *arXiv Mathematics e-prints* (June 1999), math/9906042.

[18] Hales, T. C. The honeycomb problem on the sphere, 2002.

[19] Hutchings, M., Morgan, F., Ritoré, M., and Ros, A. Proof of the double bubble conjecture. *Annals of Mathematics 155*, 2 (2004).

[20] Jarník, V. On a certain problem of minimization. *Práce Moravské Přírodovědecké Společnosti 6* (1930).

[21] Jarník, V., and Kössler, M. On minimal graphs containing n given points. *Časopis pro pěstování matematiky a fysiky 63* (1934).

[22] Melzak, Z. On the problem of Steiner. *Canadian Mathematical Bulletin 4*, 2 (1961), 143–148.

[23] Plateau, J. *Statique Expérimentale Et Théorique Des Liquides Soumis Aux Seules Forces Moléculaires*. Gauthier-Villars, 1873.

[24] Prim, R. C. Shortest connection networks and some generalizations. *The Bell Systems Technical Journal 36*, 6 (1957).

[25] Rado, T. On Plateau's problem. *Annals of Mathematics 31*, 3 (1930), 457–469.

[26] Schönhage, A. On the power of random access machines. In *Automata, Languages and Programming* (Berlin, Heidelberg, 1979), H. A. Maurer, Ed., Springer Berlin Heidelberg, pp. 520–529.

[27] Sullivan, J. M. *The Geometry of Bubbles and Foams*. Springer Netherlands, Dordrecht, 1999, pp. 379–402.

[28] Taylor, J. The structure of singularities in soap-bubble-like and soap-film-like minimal surfaces. *Annals of Mathematics 103*, 2 (1976), 489–539.

[29] Weaire, D. Kelvin's foam structure: a commentary. *Philosophical Magazine Letters 88*, 2 (2008), 91–102.

[30] Weaire, D., and Phelan, R. A counter-example to Kelvin's conjecture on minimal surfaces. *Philosophical Magazine Letters 69*, 2 (1994), 107–110.

[31] Wichiramala, W. Proof of the planar triple bubble conjecture. *Journal für die reine und angewandte Mathematik 2004*, 567 (2004), 1 – 49.