GRAPHS, KNOTS AND SURFACES

JONATHAN A. HILLMAN

Curves, surfaces and their higher-dimensional analogues are the basic settings for studying problems depending on many variables. This part of the course is an introduction to how we may think about such objects, without using equations to describe them. The basic objects of interest are figures in the plane $\mathbb{R}^2$ or 3-space $\mathbb{R}^3$ which may be assembled from line segments and polygons.

We shall use simple ideas from arithmetic and algebra to help distinguish between various such objects, and to decide, for instance, when we need 3 dimensions rather than 2 to carry out our constructions. (We may even consider surfaces in 4-space $\mathbb{R}^4$ briefly).

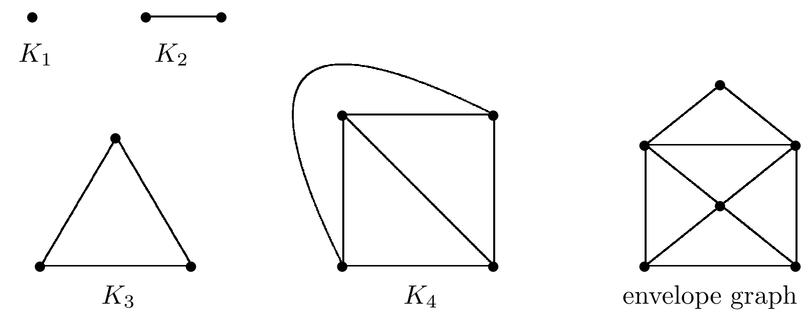
In particular, we shall often be counting elements of finite sets, and may use mathematical induction and some linear algebra.

1. Graphs

A *graph* G is a finite set of points (*vertices*) connected by *edges*, which meet only at the vertices. If the edge e connects the vertices P and Q, we shall say that P and Q are *adjacent* and that they are the *endpoints* of e.

We shall let V be the set of vertices and E the set of edges. Any graph $\widetilde{G}$ obtained by taking some subset $\widetilde{V} \subseteq V$ of the vertices and some subset of the edges connecting the vertices of $\widetilde{V}$ is called a *subgraph* of G. (We shall then write $\widetilde{G} \leq G$.)

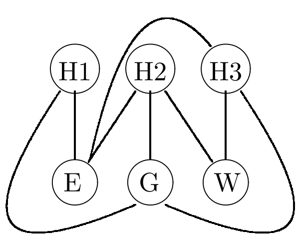
The complete graph on n vertices K_n is obtained by taking n points and joining every distinct pair of points by exactly one edge. Thus there are $\binom{n}{2}$ edges in all. Clearly $K_n \leq K_{n+1}$ for all $n \geq 1$.



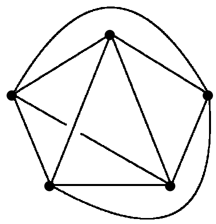
1. *Some planar graphs*

A graph is *planar* if it can be drawn in the plane without edges crossing.

The graphs in Figure 1 are all planar. We shall show later that K_5 is not planar. Since $K_n \leq K_{n+1}$ for all n it shall follow that K_n is planar if and only if $n \leq 4$. Another (possibly familiar?) nonplanar graph underlies the "Three Utilities" puzzle. Can you connect three houses to electricity, gas and water so that the supply lines never cross? Here there are six vertices (the houses and the utilities) and nine edges (the supply lines). (See Figure 2. Note that in the figure the first house has no water supply!)



three utilities



K_5

2. *Two non-planar graphs*

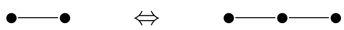
One of the purposes of this course is to show how we can use algebraic or numerical tests to decide geometric questions such as whether a given graph is planar.

A graph may have isolated points, and we allow loops and multiple edges connecting the same pair of points. However we assume every edge has at least one endpoint! (In particular, we do not consider a simple closed curve without vertices to be a graph.) We do not require that the edges be straight line segments.

This definition may be modified in various ways. The most important variants are

(1) to insist that there be no loops (edges with only one endpoint), and at most one edge connecting any given pair of vertices;
(2) to assume that each edge is oriented, i.e., we specify a direction along each edge;
(3) to allow infinite graphs.

We can always arrange that (1) holds, by adding new vertices at the midpoints of loops and multiple edges. This process is called *subdivision.* We may reverse this process to delete a vertex Q which is an endpoint of exactly two edges e, f whose other vertices P, R are distinct from Q, and replace the pair of edges e, f by a single edge e' from P to R. (We allow the possibility $P = R$.)



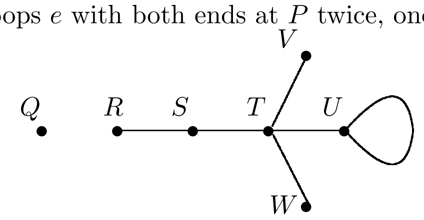
3. *Subdivision of an edge*

We shall want to consider graphs obtained from each other by iterated subdivision (or the reverse operation) as being equivalent. (In the language of topology, they are homeomorphic, i.e., "look the same".)

We shall not have time to say much about oriented graphs, and shall not consider infinite graphs at all.

Although not all graphs are planar, they can all be constructed in 3-space, and if there are no multiple edges we may then assume that all edges are straight line segments. It is enough to notice that we can choose points in $\mathbb{R}^3$ representing the vertices such that no four points lie in a common plane, for then no two edges connecting four distinct points can meet. This is clear if we have only three vertices, Given n points $P_1, \ldots, P_n$, they determine at most $\binom{n}{3}$ planes, and so we can always find points of $\mathbb{R}^3$ not in any of these planes. If no four of $P_1, \ldots, P_n$ lie in a common plane then adjoining a new point P_{n+1} which is not on this union of planes will give a set of $n+1$ points with this property. We take such a "maximally independent" set of points as our vertices, and join them by straight line segments as required.

A vertex P has *degree* $deg(P) = d$ if there are d edges which have P at one end. (We count loops e with both ends at P twice, once for each end).



4. *Vertices of degrees* 0, 1, 2, 3 *and* 4

Vertices of degree 0 are isolated points, not connected to any other part of the graph, and as such are not very interesting. Vertices of degree 1 are extreme points of the graph. Vertices of degree 2 may be created by subdivision or removed by the reverse operation. (See Figure 4.)

It is easy to see that $\Sigma_{P \in V} deg(P) = 2\#E$, since each edge has two endpoints.

2. Paths, connected, trees

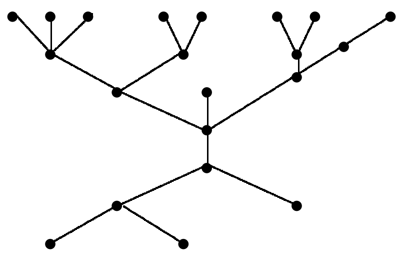
If P and Q are vertices of G a *path* from P to Q is a set of edges $e_1, \dots, e_n$ such that P is a vertex of e_1, each edge meets the next and Q is a vertex of e_n. The graph is *connected* if any two vertices are the ends of such a path. If G is connected the *distance* from P to Q is the number of edges in the shortest path from P to Q. (It is convenient to allow also paths of length 0, consisting of just a single vertex.)

In figure 4, the distance from R to U is 3. (The distance from Q to R is not defined.)

In general, the *path component* of a vertex P is the subgraph consisting of all vertices which can be reached by paths starting at P and all edges with endpoints such vertices. A subdivision of a connected graph remains connected, and subdivision does not change the number of components.

A *circuit* is a path which starts and finishes at the same point. If a circuit contains two consecutive edges e_i, e_{i+1} which are copies of the same edge then deleting these edges gives a shorter circuit. The circuit is *trivial* if repeating this process eventually gives a circuit of length 0.

A connected graph is a *tree* if it has no loops and no nontrivial circuits. (See Figure 5.)



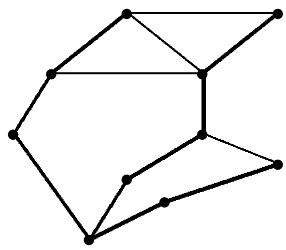
5. *A tree*

In a tree there is an unique shortest path connecting any two given vertices P and Q.

Theorem 1. *A tree T with more than one vertex has at least two vertices of degree 1.*

Proof. We induct on the number of vertices of T. The result is clearly true if T has two vertices. Suppose that T has at least three vertices, and choose an edge e of T, with endpoints P, Q, say. If we delete this edge (keeping the vertices) the resulting graph has two path components, T_P and T_Q, say. Each of these is a tree, and has fewer vertices than T. If each has more than one vertex then they each have at least two vertices of degree 1, by the inductive hypothesis. Replacing the edge e to recover T uses up at most two of these (four or more) vertices, and so T has at least two such vertices. If on the other hand one of them, T_P say, is reduced to a single point then P must have been a vertex of degree 1 in T. Since T_Q has more than one vertex, it has at least two vertices of degree 1 and the result again holds. □

Let G be a connected graph with n vertices, and suppose that T_k is a subgraph which is a tree with $k < n$ vertices. Since G is connected there is an edge e with one endpoint P in T_k and the other endpoint Q not in T_k. Let T_{k+1} be the subgraph obtained by adjoining the vertex Q and the edge e to T_k. Then T_{k+1} is a tree. In this way we see that every connected graph contains a tree which has all the vertices of G as its vertices. (See Figure 6 for an example.) Such maximal trees are usually not unique; we must choose which edges to include. (This nonuniqueness is evident already for $G = K_3$.)



6. *A connected graph with a maximal tree*

3. Euler characteristic and Eulerian circuits

Subdivision changes the total numbers of vertices $\#V$ and edges $\#E$. However each time we add a new vertex in this way each of these numbers increases by 1, and so the difference $\chi(G) = \#V - \#E$ does not change. This difference is called the *Euler characteristic* of G, and is historically the first example of a topological invariant. We shall meet an extension of this idea when we consider surfaces, and shall then use it to show that certain graphs are not planar.

Let T be a tree. Then $\chi(T) = 1$, i.e., T has one more vertex than edges. We again argue by induction on the number of vertices. It is clearly true if T has one vertex, for a tree with one vertex has no edges. In general, if T is a tree with n vertices and P is a vertex of degree 1 then deleting P and the unique edge with P as an endpoint gives a tree T' with $n-1$ vertices, and clearly $\chi(T) = \chi(T')$, since the numbers of vertices and edges have each been reduced by the same amount (1).

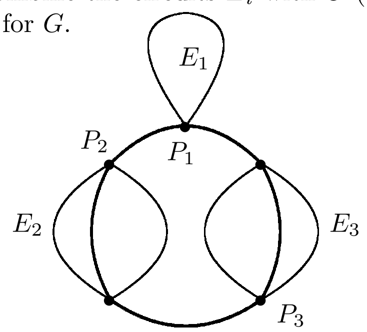
If G is connected then $1 - \chi(G)$ is the "number of independent circuits" in G.

An *Eulerian circuit* is a circuit which contains every edge exactly once. The first result in Graph Theory (and, more generally, in Combinatorial Topology) was due to Euler:

Theorem 2. *A graph G has an Eulerian circuit if and only if it is connected and every vertex has even degree.*

Proof. If G has a circuit which contains every edge then it must be connected. An Eulerian circuit may pass through a given vertex P more than once, but each time it does so it contributes 2 to the degree of P. Thus every vertex has even degree.

For the converse, we use induction on the number of edges. The result is clearly true if G has 0 edges (in which case G is empty or a single vertex). Choose a vertex P as a starting point, and move along edges. Eventually we must complete a circuit C by returning to P. If we delete the edges of this circuit the graph G breaks up into one or more subgraphs $G_1, \ldots, G_k$ with fewer edges, and in which each vertex still has even degree. Each of these has an Eulerian circuit E_i, by the inductive hypothesis. Suppose our starting point P is in G_1. We first go round the Eulerian circuit E_1, coming back to P, and then we continue along C until we meet a vertex not in G_1. We then go round the Eulerian circuit through this point, and so on. In this way we may combine the circuits E_i with C (as indicated in Figure 7) to get an Eulerian circuit for G. □



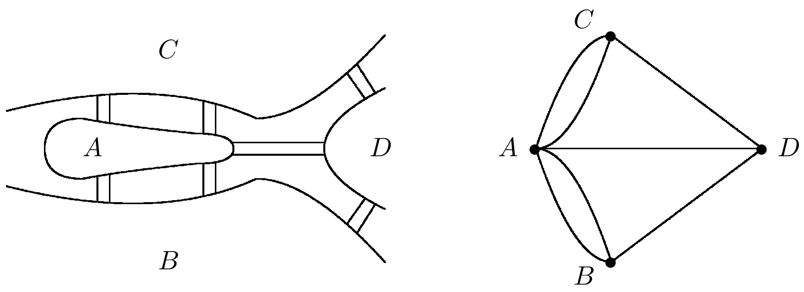
7. *Improving a circuit C to an Eulerian circuit (schematic)*

In this Figure C is the (approximate) circle drawn with heavier lines.

Example:

Can you draw the envelope graph without lifting your pen off the paper and without retracing any edges?

Euler's result was reputedly prompted by someone asking whether it was possible to walk around Koenigsberg, crossing all seven bridges exactly once. This was perhaps the first mathematical problem with a genuine topological flavour. The corresponding graph has four vertices, one for each piece of land and seven edges, one for each bridge. (See Figure 8.)



8. *The bridges of Königsberg – with associated graph*

The letters represent land: two rivers flow from right to left on either side of land D, A is an island in the combined river, while B and C are its left and right banks. The edges represent the bridges. (Note that the vertices all have odd degree.)

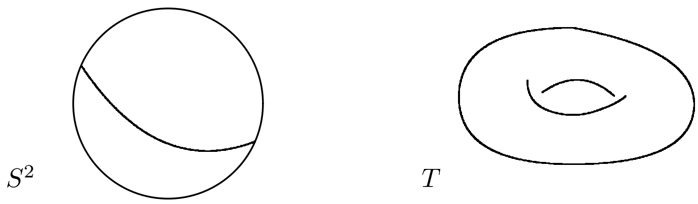
There is a corresponding notion focusing on vertices rather than edges. A *Hamiltonian circuit* is a circuit which passes through each vertex exactly once. It is a major open problem of practical consequence whether there is an efficient algorithm for finding Hamiltonian circuits in connected graphs. (This is known as the "Travelling salesman" problem. It arises in scheduling the movement of aircraft, tankers, data through computers, etc.) We shall say no more about Hamiltonian circuits.

4. Surfaces

A *surface* in 3-space is a subset that looks locally like the graph of a function of two variables. Typically, a surface may be described by a single equation $f(x, y, z) = 0$. For instance, the unit sphere is the set

$$S^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}.$$

of points (x, y, z) satisfying the equation $x^2+y^2+z^2-1 = 0$. This is a mathematical model of the surface of the solid ball of radius 1, in the ordinary sense of the word "surface". Similarly, the torus defined later is a model of the surface of a doughnut. (See Figure 9.)

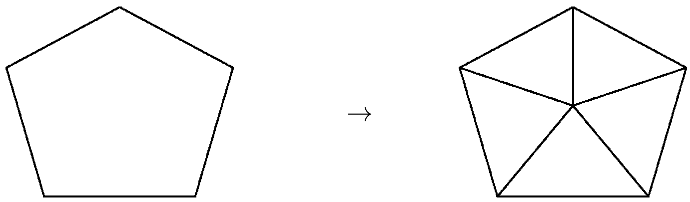


9. *The sphere S^2 and the torus T*

In practice it is often difficult to find simple equations, and in any case the equations describing a given surface are not very useful, and are far from unique. Moreover some of the objects which deserve to be considered as surfaces cannot be constructed in 3-space. We shall approach surfaces in a different way.

Definition. A *surface* $\mathcal{S}$ is a subset of some m-space $\mathbb{R}^m$ which may be assembled by gluing together polygons along their edges, so that at most two polygons meet along any given edge.

We shall usually assume that $m = 3$ or 4. We could insist that our basic building blocks be triangles, since any polygon with n sides is the union of n triangles, glued together two at a time. (See Figure 10.) Allowing general polygons simplifies some of our constructions. Note also that we shall often match up sides of the same polygon. On the other hand, we do not assume that all edges are used up in the matching process. The edges left over form the *boundary* $\partial\mathcal{S}$ of our surface.



10. *Dividing a polygonal face into triangles*

The plane itself is a union of polygons (in many ways). For instance, we may take all the squares whose vertices have integer coordinates. (See the treatment of the torus below.) However we shall concentrate on surfaces obtained from finitely many polygons. In this case the boundary is always a finite collection of circles. (This includes the possibility that the boundary is empty!)

For ease of drawing, we shall not insist that the sides of our polygons be straight line segments. (In the language of topology, our basic building blocks shall be homeomorphic to standard polygons). Note also that we shall often write "polygon" when we mean the sides of the polygon AND the 2-dimensional region of the plane that they enclose.

Definition. Two subsets X and Y of $\mathbb{R}^n$ are *homeomorphic* if there is a continuous function $f : X \to Y$ with a continuous inverse $g : Y \to X$.

We shall not elaborate on how we define "continuous" here. (The definitions you met in earlier years will suffice). It shall usually be clear how this definition applies.

Examples: The standard unit disc is the set

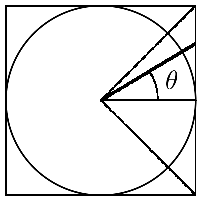
$$D^2 = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}.$$

Every convex polygon is homeomorphic to D^2. To see this, put a small copy of the disc at the centre of the polygon and expand radially to fill out the interior. Note that any such homeomorphism carries the circle

$$S^1 = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$$

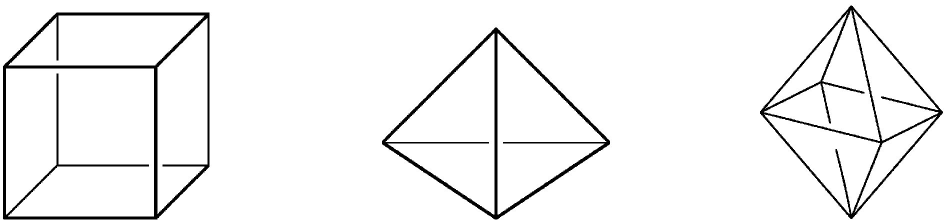
onto the boundary of the polygon. (Thus $\partial D^2 = S^1$.)

In particular, we may define a homeomorphism from the disc to the square $\square = \{(x, y) \in \mathbb{R}^2 \mid |x| \leq 1, |y| \leq 1\}$ by using polar coordinates (r, θ) and setting $f(r, \theta) = (r|\sec\theta|, \theta)$ if $-\frac{\pi}{4} \leq \theta \leq \frac{\pi}{4}$ or if $\frac{3\pi}{2} \leq \theta \leq \frac{5\pi}{4}$ and setting $f(r, \theta) = (r|\mathrm{cosec}\theta|, \theta)$ otherwise. As you can see, actually writing out a formula for a homeomorphism is tedious and not very enlightening, and we shall usually avoid doing so. (See Figure 11.)



11. *Expanding the unit disc*

Similarly, we may show that any convex solid in 3-space is homeomorphic to the solid 3-ball $B^3 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^3 \leq 1\}$. In particular, the boundary of such a solid is homeomorphic to the boundary of the solid ball, which is the standard unit sphere S^2. We shall use this without further comment to see that S^2 is a union of four triangles (boundary of the solid tetrahedron), six squares (boundary of the cube), etc. (See Figure 12.)

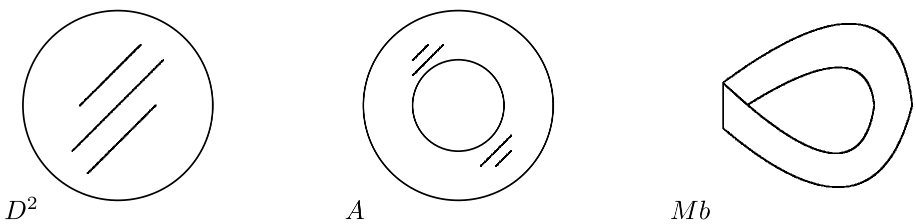


12. *Three regular solids: the cube, tetrahedron and octahedron*

Examples:

(1) The annulus $A = \{(x, y) \in \mathbb{R}^2 \mid \frac{1}{2} \leq x^2 + y^2 \leq 1\}$.
(2) The torus T.
(3) The Möbius band Mb.
(4) The Klein bottle Kb.
(5) The projective plane RP^2.

The standard disc and annulus are defined as subsets of the plane $\mathbb{R}^2$. The sphere, torus and Möbius band can all be constructed in 3-space $\mathbb{R}^3$, but the natural home for the Klein bottle and the projective plane is in 4-space $\mathbb{R}^4$! We shall show later how these surfaces may be "visualized", using time (or colour intensity) to indicate the extra dimension. (See Figure 13 for some of these surfaces.)

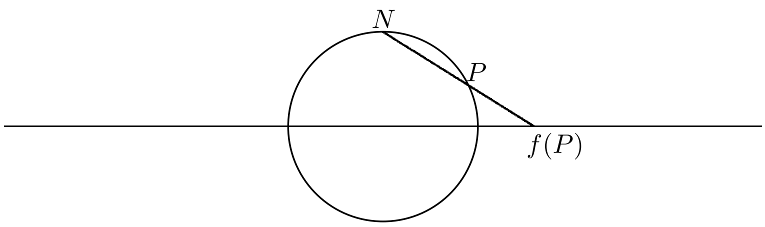


13. *Three surfaces with boundary*

Note that A has two boundary circles, whereas Mb and D each have just one. The other surfaces given so far have no boundary.

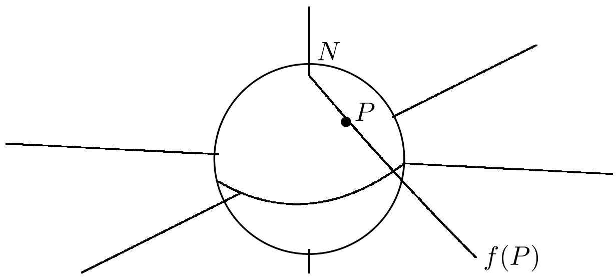
5. Stereographic projection

Let $N = (0, 1)$ be the "north pole" of the unit circle $S^1 \subset \mathbb{R}^2$. Then we may define a homeomorphism from the complement $S^1 - \{N\}$ to the real line $\mathbb{R}$ by setting $f(P)$ to be the point of intersection of the line from N through P with $\mathbb{R}$. It is easy to see that in fact $f(x, y) = \frac{x}{1-y}$. The inverse function sends $t \in \mathbb{R}$ to $f^{-1}(t) = (\frac{2t}{t^2+1}, \frac{t^2-1}{t^2+1})$. In particular, $f^{-1}(0) = (0, -1)$ is the "south pole" of the circle. Thus we may view the circle S^1 as being obtained from $F^{-1}(\mathbb{R})$ (a copy of the real line) by adding one more point (N). From the point of view of $\mathbb{R}$, N is very far from the origin; as $t \in \mathbb{R}$ moves towards $\pm\infty$ the corresponding point f^{-1} approaches N on S^1. We say that S^1 may be obtained by "adding a point at ∞" to $\mathbb{R}$, and write $S^1 = \mathbb{R} \cup \{\infty\}$. (See Figure 14.)



14. *Stereographic projection from the circle to the line*

A similar idea works in higher dimensions. Let $N = (0, 0, 1)$ be the north pole of S^2. Then stereographic projection from N identifies every point $P \in S^2 - \{N\}$ with the point of intersection of the line from N through P with the horizontal plane with equation $z = 0$. Once again, we may write down formulae for the stereographic projection and its inverse, but these are not important for our purposes. We again say that S^2 may be obtained by "adding a point at ∞" to $\mathbb{R}^2$, and write $S^2 = \mathbb{R}^2 \cup \{\infty\}$. (See Figure 15.)

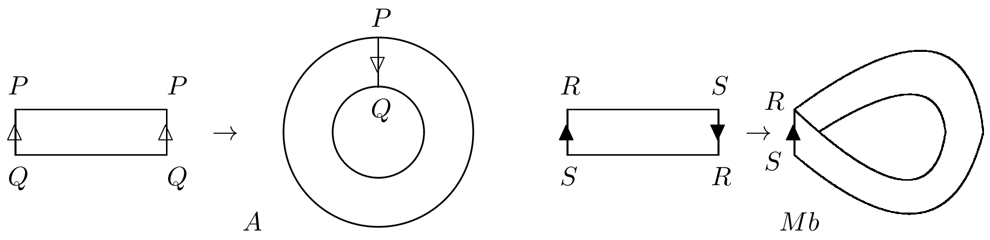


15. *Stereographic projection from the sphere to the plane*

Warning. This construction of S^2 is NOT the one used to construct the projective plane, where we add a *line* at infinity. We shall say more about the projective plane later.

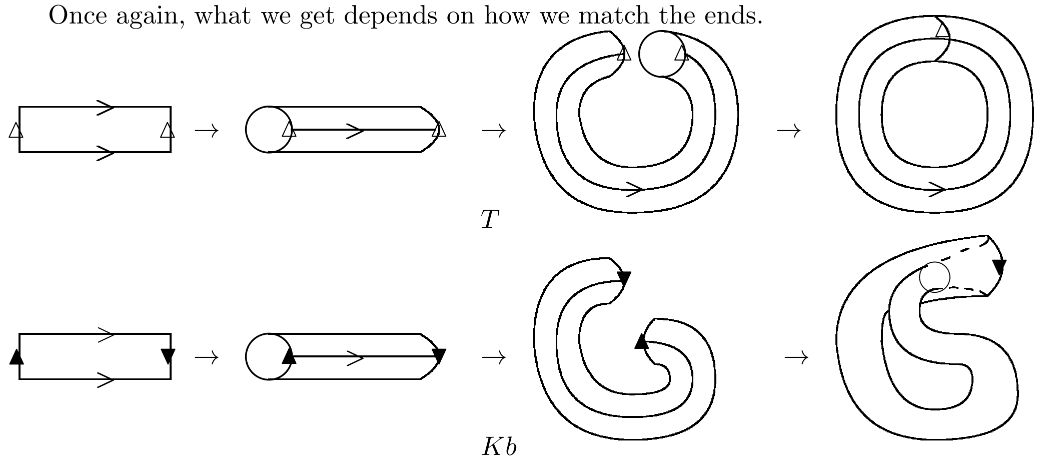
6. The other basic surfaces

The annulus A and Möbius band may each be constructed by gluing together one pair of ends of a rectangle. What we get depends on how we match the two ends. If you try to make one of these surfaces out of a piece of paper or some ribbon, you have to twist the paper/ribbon to get Mb. More precisely, an even number of half twists gives a copy of A, and an odd number gives a copy of Mb. Although we only get two types of surface in this way, the twisted versions are sitting in 3-space in distinct ways, and the twists cannot be removed without cutting and regluing. (See Figure 16.) Note also that A is homeomorphic to the cylinder $\{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1, 0 \leq z \leq 1\}$.



16. *Constructing the annulus and Möbius band*

The torus T and Klein bottle Kb may each be obtained by identifying both pairs of opposite sides of a rectangle. First we glue together top and bottom, to get a cylinder with two ends, and then glue the ends together, as indicated in Figure 17. Once again, what we get depends on how we match the ends.



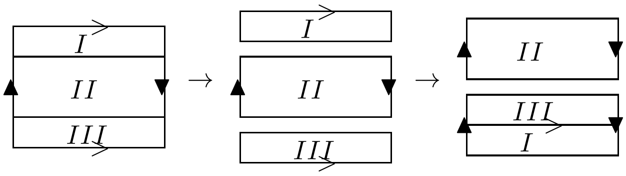
17. *Constructing the torus and Klein bottle*

A slick way of explaining the construction of T is to say that the abelian group $\mathbb{Z}^2$ of vectors with integer entries acts by translation on $\mathbb{R}^2$, so that (m, n) moves a point m units to the right and n units up. (Remember that $left = -right$ and $down = -up$!). We then identify points which are equivalent under this action, i.e., $(x, y) \sim (x', y')$ if $x' - x$ and $y' - y$ are integers. It is easy to see that every point is equivalent to a point (x, y) in the unit square (with $0 \leq x \leq 1$ and $0 \leq y \leq 1$), and that opposite sides are matched by horizontal or vertical translation.

If we cut the rectangle horizontally we see that each half of the rectangle closes up to an annulus (on gluing together the ends) and so the torus is the union of two annuli.

The Klein bottle Kb may also be constructed by identifying opposite sides of a rectangle. However this time we invert one of the sides (say, the left vertical side) before making the identification. In this case the group acting on $\mathbb{R}^2$ is generated by an integer translation in the vertical direction and a glide reflection with axis a unit horizontal vector.

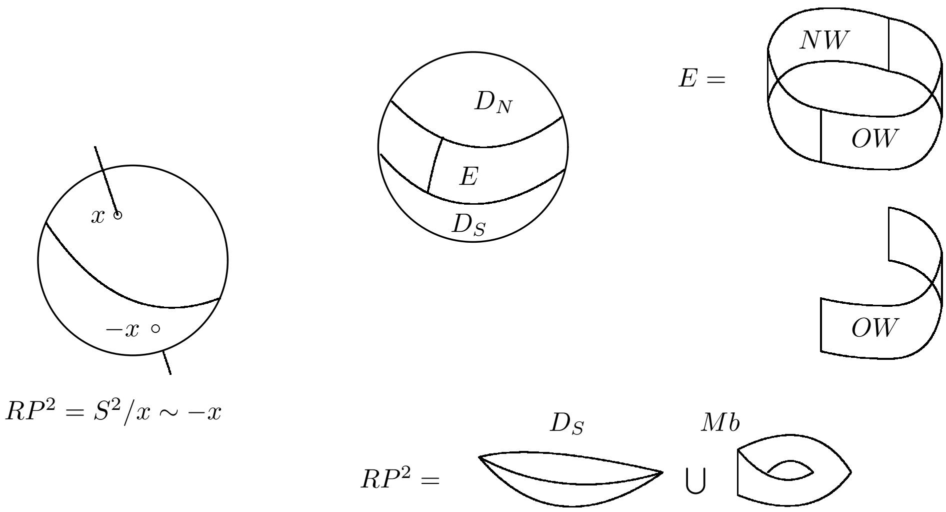
If we cut out a thin strip centered on the horizontal midline of the square we see that the square falls into three rectangles. The central one closes up to give a Möbius band, while after gluing the other two along the edges corresponding to the top and bottom of the original square we get a second Möbius band. Thus the Klein bottle is the union of two copies of Mb. (See Figure 18.)



18. *The Klein bottle decomposed into Möbius bands*

The projective plane is the space of lines through the origin in $\mathbb{R}^3$. Every such line meets S^2 in two antipodal points, $P = (x, y, z)$ and $-P = (-x, -y, -z)$,say. Thus we may obtain RP^2 from the sphere by identifying antipodal points. (We again have a natural group action. The group $\mathbb{Z}/2\mathbb{Z}$ of order 2 acts via $P \mapsto -P$.)

We may decompose S^2 into the union of an equatorial zone E, a northern region D_N and a southern region D_S, so that the antipodal map maps E to itself and swaps D_N with D_S. If moreover we divide the equatorial zone into an "old world" half OW, (from the 0 meridian to the international dateline) and a "new world" half, we see that the antipodal map swaps these two halves, and inverts the edge running along the 0 meridian. The image of D_S in RP^2 is a disc, while the image of OW there is a copy of Mb. Since every point of RP^2 is represented by some point from $D_S \cup OW$ we see that RP^2 is the union of Mb and a disc along their common boundary. (See Figure 19.)



19. *Constructing the projective plane*

(It is not easy to depict RP^2 convincingly. However I hope to persuade you that Kb is easy to "see", even though it cannot be constructred in 3-space. In fact one can make satisfactory models of Kb from chicken wire or even, with enough glass-blowing skill, from glass. All such physical models live in 3-space, and so will have apparent self-intersections.)

All other surfaces may be obtained from the surfaces we have described so far, by a process of "surface arithmetic".

7. Boundaries

The boundary of a surface constructed by identifying pairs of edges of polygons is the union of the unmatched edges. Initially, we have some (possibly large) collection of polygonal pieces, and before assembly of the pieces all the edges are unmatched. The edges of each polygon together meet head-to tail to form a circle. Whenever we identify a pair of edges of two distinct polygons we obtain one larger polygon, and the remaining edges again form a circle. More generally, if S and S' are two surfaces whose boundaries are nonempty families of circles and we glue S and S' together by identifying an edge in ∂S with an edge in $\partial S'$ the boundary of the resulting surface is still a collection of circles.

Note that if we glue together two sides which are part of the same boundary circle of a given surface (as in the construction of A from a rectangle), we may *increase* the number of boundary circles by 1.

8. Subdivision, Euler characteristic (χ), orientability

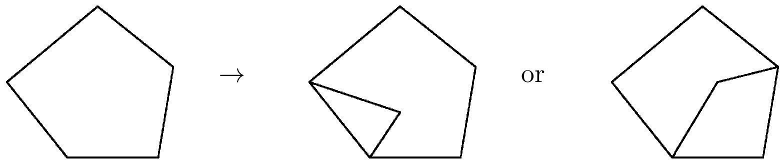
A decomposition of a surface $\mathcal{S}$ into a union of polygon pieces is called a *polygonal decomposition* of $\mathcal{S}$ and a decomposition into triangular pieces is called a *triangulation* of $\mathcal{S}$. The *vertices* and *edges* of the decomposition are the points on the surface corresponding to the vertices and edges of the polygons, while the *faces* are the regions corresponding to the interiors of the polygons. Let V, E and F be the sets of vertices, edges and faces of the triangulation.

As in the case of graphs, we may subdivide triangulations, by introducing new vertices and new edges. There are now three basic types of subdivision:

1. add a new vertex in the middle of an edge, and new edges connecting this vertex to the opposite vertices of the (one or two) triangles sharing this edge;
2. add a new vertex in the middle of a face, and at least one new edge connecting this vertex to some vertex of the face;
3. add a new edge joining two vertices of the same face.

(If you feel unhappy about creating "reentrant" polygons by the second type of subdivision, add new edges connecting the new vertex to several of the existing vertices of the face. See Figure 20.) We may similarly subdivide polygonal decompositions. Any two polygonal decompositions have a common subdivision, using all the vertices and edges of the given decompositions and enough new edges to ensure that the surface is divided into polygonal faces.

The number of vertices, edges and faces may be increased at will by subdivision. However the alternating sum $\chi(\mathcal{S}) = \#V - \#E + \#F$ does not change under either of the basic subdivisions, and thus is independent of the decomposition. This is the Euler characteristic for surfaces.



20. *Internal subdivision of a face*

Examples: $\chi(S^2) = 2$; $\chi(D^2) = 1$; $\chi(T) = 0$.

If A and B are surfaces which are subsets of a given surface then $\chi(A \cup B) = \chi(A) + \chi(B) - \chi(A \cap B)$, since the vertices, edges and faces in $A \cap B$ are "counted twice" in the sum $\chi(A) + \chi(B)$. We may also write this equation in the more symmetric form

$$\chi(A \cup B) + \chi(A \cap B) = \chi(A) + \chi(B).$$

Note that the *area* of a surface satisfies a similar equation. (From the right point of view χ and area are closely related.)

If $\partial\mathcal{S}$ is empty, then $2\#F = 3\#E$, since each face has three sides and each edge is common to two faces. (This is not correct if there are unmatched edges.)

If we assume that the edges are actually straight-line segments in $\mathbb{R}^3$ then each edge is determined by its endpoints, and so $\#E \leq \binom{\#V}{2}$. However this assumption usually can only be satisfied if we have a large number of triangles. (We may look more closely at the case of the torus later).

It is easy to see that subdivision does not change the number of circles forming the boundary.

The notion of orientability is perhaps the most subtle aspect of the study of surfaces. We shall given a working definition now, and an extended discussion later. A surface $\mathcal{S}$ is *nonorientable* if some subcollection of the polgygons in a polygonal decomposition of S together form a Möbius band. The surface is *orientable* if it is not nonorientable. It is again easy to see that orientability (or otherwise) is preserved under subdivision.

9. Classification of surfaces

Two surfaces are *(combinatorially) homeomorphic* if they have polygonal decompositions with a common subdivision. For example, if we subdivide the base of a tetrahedron by adding midpoints to the three sides and a midpoint to the face, and edges connecting the midpoints of the sides to the midpoint of the face we obtain a (distorted) cube: six quadrilateral faces meeting pairwise just as the six faces of the cube do.

A surface is *connected* if any two points on the surface are the endpoints of a continuous path on the surface. (Since we are considering surfaces with polygonal decompositions, we may assume any such path is a finite union of line segments in the faces, matching head-to tail at the edges). As with the case of graphs, subdivision of a polygonal decomposition of a surface does not disconnect a connected surface. (It is easy to see that a surface is connected if and only if the graph formed by the vertices and edges of some polygonal decomposition is connected.)

Every surface $\mathcal{S}$ is a finite union $\mathcal{S} = \cup_{i\in I}\mathcal{S}_i$ of pairwise disjoint connected surfaces, and for most purposes it is sufficient to consider these "connected components" separately.

We may now formulate our first version of the classification theorem for surfaces.

Theorem 3. *Two connected surfaces $\mathcal{S}$ and $\mathcal{S}'$ are homeomorphic if and only if $\chi(\mathcal{S}) = \chi(\mathcal{S}')$, $\mathcal{S}$ and $\mathcal{S}'$ have the same number of boundary circles and both are orientable or both are nonorientable.*

Moreover $\chi(\mathcal{S}) \leq 2$ and is even if $\mathcal{S}$ is orientable and $\partial\mathcal{S} = \emptyset$.

10. Cutting and pasting - reduction to standard form

We have defined a surface as a figure which admits a decomposition into polygons in a specific way. The decomposition is far from unique, as we are free to subdivide. We can however always arrange that a connected surface is represented by a single polygon with identifications of pairs of edges.

To see this, suppose that we have decomposed our surface $\mathcal{S}$ into a finite number of polygonal pieces, and that we have labeled all the edges so that we know which should be glued together (and how). If there is more than one polygon, there must be some pair with sides representing the same edge of the decomposition. Glue these two polygons together along just that pair of edges, to get one new, larger polygon. If we retain all the labels for the other edges it is clear that we can reconstruct our surface. This procedure has reduced the number of faces by one. Clearly we can repeat this process until there is just one face, which may have many sides.

We shall now be more specific about our labeling. Choose a vertex of the polygon as a starting point, and go round the polygon labeling the edges alphabetically $a, b, \ldots$ until you reach an edge which is to be glued to one already labeled. (The mathematical convention is to go counter-clockwise.) Suppose that in fact this first repeat is to be identified with the edge already labeled e, say. We give the new edge the label e or $\bar{e}$, depending on whether the edges are matched head-to-head and tail-to-tail, or head-to-tail. (You may also see e^{-1} used instead of $\bar{e}$. There is good reason for this.) Continuing, we label all the edges so that matched pairs have the same letter (but possibly different exponents). Of course, if there are more than 26 edges on the surface we may have to use other alphabets, but the principle is clear.

The gluing instructions may be written out as a "word"

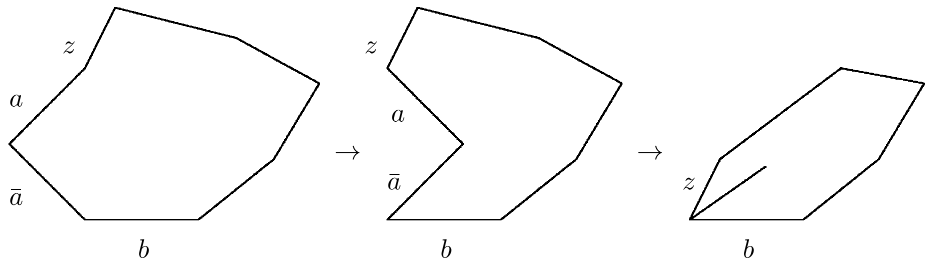
$$abc \ldots qc^{\pm 1} \ldots .$$

We shall show that by cutting and rearranging the polygon we can put the word into a standard form. We shall use capital letters like $U, V, W, \ldots$ to represent segments of our word between letters of interest.

We say that a side of the polygon represents a *free* edge of the surface if it is not matched with any other side. The free edges together form the boundary circles, and the number of free edges gives an upper bound on the number of circles. (Compare the usual representations of A and Mb by a rectangle.)

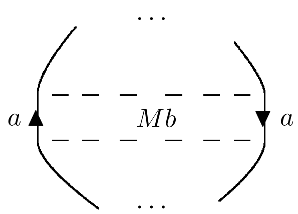
We may calculate the Euler characteristic of the surface from the polygon as follows: there is one face, represented by the polygon itself, one edge for each free edge and one edge for each pair of matched sides. (Thus there is one edge for each distinct letter used, so the word $ab\bar{a}c$ representing the annulus gives rise to 3 edges.) We must count the vertices carefully. In general, many of the vertices of the polygon will become identified after doing all possible gluings, and we must count *after* such identification, not before. This is not difficult, just careful book-keeping.

If a letter is immediately repeated in inverted form $\dots a\bar{a}\dots$ or $\dots\bar{a}a\dots$ then we may delete these two letters to get a shorter word representing the same surface. Geometrically, this corresponds to pushing the vertex between these sides inwards and then gluing the sides together. (See Figure 21). This process may lead to further such deletions: consider $a\bar{b}c\bar{c}ba$, which collapses in two steps to aa.



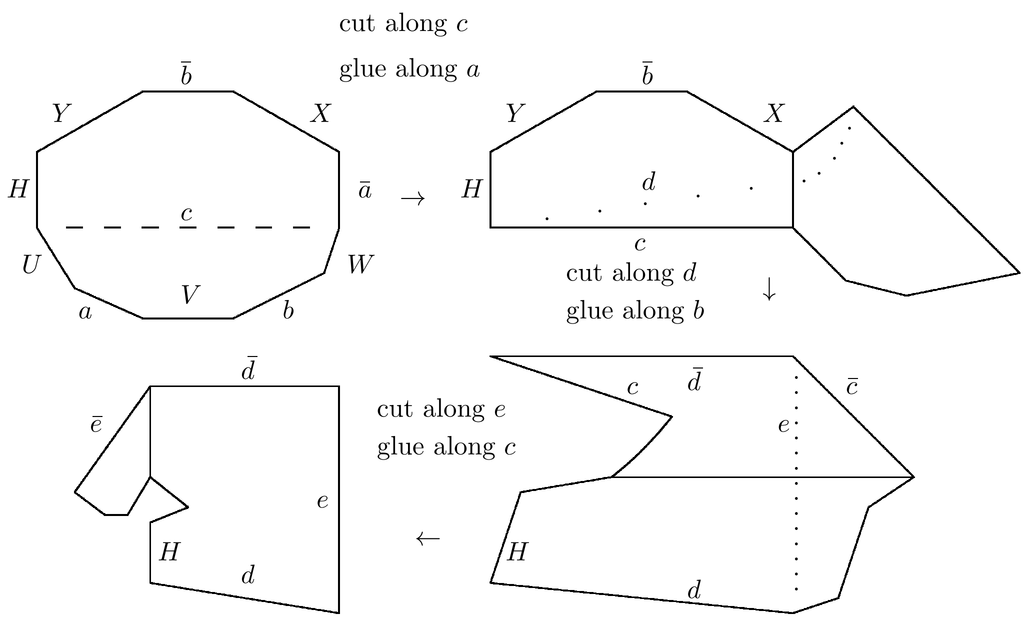
21. *Deleting* $a\bar{a}$ – other sides unchanged

If some label occurs twice with the same exponent, i.e., if the word contains a subword of the form $\dots aXa\dots$ then $\mathcal{S}$ must contain a copy of Mb, and so is nonorientable. (See Figure 22). We shall show later that if there are no such repetitions $\mathcal{S}$ is orientable.



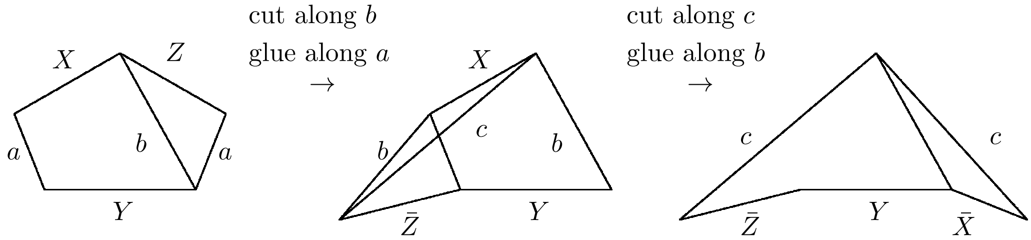
22. *The rectangular region closes up to a Möbius band on the surface*

We shall assume at first that there are no subwords $\dots aXa\dots$ (i.e., the surface $\mathcal{S}$ is orientable), there are no free edges and there are no redundant pairs $a\bar{a}$ or $\bar{a}a$. Our goal is to show that cutting and pasting gives a word of the "standard' form $ab\bar{a}\bar{b}cd\bar{c}\bar{d}\dots mn\bar{m}\bar{n}$. (The number of sides is then a multiple of 4, the number of edges is even, and there is just one vertex.) Pick a side a such that the segment $aZ\bar{a}$ with ends a and $\bar{a}$ has minimal length. Then Z must be nonempty (since there are no redundant pairs) and the letters appearing in Z must all be distinct (by minimality of our choice). Pick one such, say b, and write $aZ\bar{a}$ as $aVbW\bar{a}$. Then our word has the form $HUaVbW\bar{a}X\bar{b}Y$ for some subwords H, U, V, W, Y. (The first segment H shall represent a portion that we have already "improved" to a standard form.) We then cut and paste as indicated in Figure 23 to find a new polygon, with associated word $Hded\bar{e}UXWVY$. Note that the number of sides has not changed, but we may have created redundant pairs, which we can remove. After repeating this process atmost $\frac{n}{4}$ times (where n is the number of sides in the original polygon) we get a word $a_1b_1\bar{a_1}\bar{b_1}\dots a_gb_g\bar{a_g}\bar{b_g}$ for some $g \le \frac{n}{4}$. This is the "standard form" for orientable surfaces with no boundary.



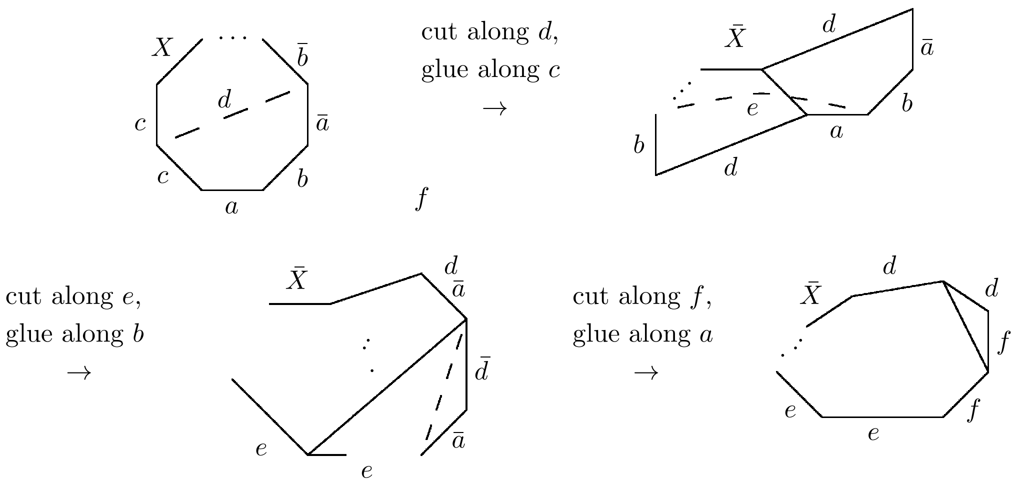
23. *Improving a word* $\ldots a \ldots b \ldots \bar{a} \ldots \bar{b}$ *to* $\ldots d e \bar{d} \bar{e} \ldots$

We next allow repetitions $\ldots aXa \ldots$. A shorter and simpler operation shows that we may replace $UaVaW$ by $Uc^2W\bar{Y}$ or $c^2W\bar{Y}U$. (See Figure 24. Here $\bar{Y}$ means that we reverse each side in Y and also reverse their order.) Eventually we use up all the repetitions, and obtain a word $a^2b^2 \ldots k^2W$, where every repetition of a letter in W is inverted. We may now use the first argument to replace W by a sequence of subwords of the form $de\bar{d}\bar{e}$. We have not finished yet!



24. *Improving a word* $XaYaZ$ *to* $\bar{X}cc \ldots$

A third cutting and pasting operation shows that any word of the form $Ha^2bc\bar{b}\bar{c}V$ may be replaced by one of the form $Hd^2e^2f^2V$. In the presence of at least one copy of Mb (or "crosscap"), every "handle" can be replaced by two crosscaps! (See Figure 25.) There is a nice picture that illustrates this, by "walking" one end of a handle around a Möbius band. (See Figure 27.)



25. *Handle plus crosscap* ($Xccab\bar{a}\bar{b}\dots$) + *three crosscaps* ($eeffdd\bar{X}\dots$)

The ultimate step is to deal with the free edges. This is easy. The free edges represent the boundary circles of our surface. We attach new polygons, one for each boundary circle, to use up these free edges. This reduces us to the cases already considered. Once we have recognised the resulting surface $\widehat{\mathcal{S}}$ we may recover $\mathcal{S}$ by making an appropriate number of punctures.

Here is the second version of our classification theorem:

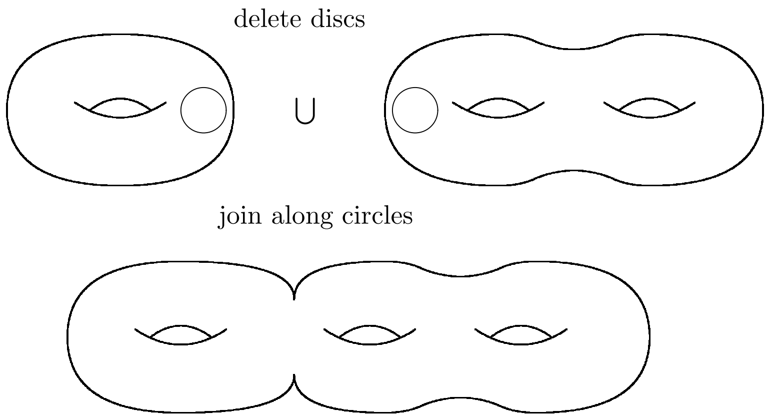
Theorem 4. *Every connected surface without boundary is represented by an essentially unique standard polygon, with corresponding word* $a_1b_1\bar{a}_1\bar{b}_1\dots a_gb_g\bar{a}_g\bar{b}_g$ *(if the surface is orientable) and* $a_1a_a\dots a_ca_c$ *(if the surface is nonorientable).*

The numbers g and c are determined by the Euler characteristic: $\chi(S) = 2-2g$ if S is orientable and $\chi(S) = 2-c$ if S is nonorientable.

11. Surface arithmetic

Let S and S' be connected surfaces without boundary. We form the *connected sum* $S\#S'$ by deleting a disc from each surface, to obtain "once-punctured" surfaces $S_o = S-D$ and $S'_o = S'-D$. These punctured surfaces each have a single boundary circle. Glueing them together along the boundary circles gives $S\#S'$. (See Figure 26.)

Suppose that that the discs deleted from each surface are faces in some polygon decomposition. Then all the vertices, all the edges and all the *other* faces of the original surfaces are used in the construction of the sum. Hence $\chi(S\#S') = \chi(S)+\chi(S')-2$. (The vertices and edges on the boundary circles are counted twice on the right hand side of this equation, once for each surface, but since a polygon has the same number of corners as sides this does not affect the alternating sum $\chi = V-E+F$.)



26. *Connected sum of torus T and pretzel $\#^2T$ to get $\#^3T$*

This construction appears to involve various choices: which discs to remove and how to glue the resulting boundary circles. It can be shown that the result is independent of any such choices. We shall write $\#^kS$ to mean the iterated (k-fold) connected sum of k copies of the surfaces S. In particular, $\#^2RP^2 = Kb$, since $RP^2_o = Mb$ and the union of two copies of Mb is the Klein bottle.

If $S = S^2$ is the unit sphere then $S_o = S^2 - D$ is just a disc, and so $S^2\#S = S$ for any surface S, since all we are doing is removing a disc and replacing it with another disc. It is easy to see that $S\#S' = S'\#S$, and that $(S\#S')\#S'' = S\#(S'\#S'')$ (commutativity and associativity of connected sum). Thus $\#$ behaves rather like ordinary addition of natural numbers, with S^2 acting like 0. However we cannot in general cancel like terms from an equation, since $T\#RP^2 = Kb\#RP^2 = \#^3RP^2$! (These surfaces are nonorientable and have the same Euler characteristic. See also Figures 26 and 27.)

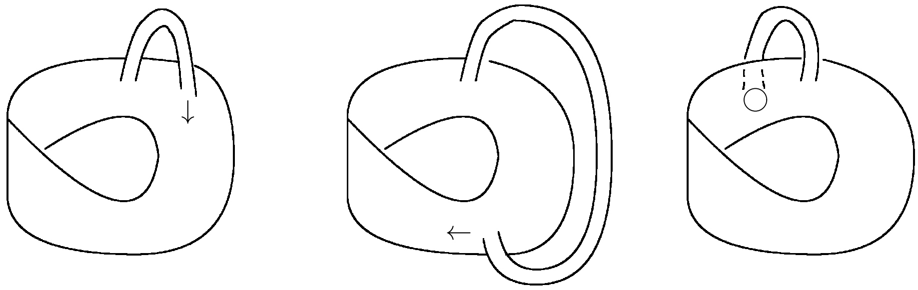
We may now give a third version of the classification theorem:

Theorem 5. *Every connected orientable surface without boundary is a sum $\#^gT$ of copies of the torus T, and every connected nonorientable surface is a sum $\#^cRP^2$ of copies of the projective plane RP^2.*

We interpret the "empty sum" of 0 copies of T as the 2-sphere: $S^2 = \#^0T$. The numbers g, c are uniquely determined by the Euler characteristics. In fact $\chi(\#^gT) = 2 - 2g$ and $\chi(\#^cRP^2) = 2 - c$. The symbols are traditional: "g" stands for *genus* and "c" for *crosscap number*.

Let S be an orientable surface, possibly with boundary, and let $\widehat{S}$ be the surface without boundary obtained from S by attaching discs along the boundary circles, to fill in the punctures. Then $\widehat{S} = \#^gT$ for some $g \geq 0$. We shall say that S has *genus* $g(S) = g$. Note that $g(S) = g(\widehat{S})$. Clearly $g(S) \geq 0$, and $g(S) = 0$ if and only if $S \subseteq S^2$.

We may extend the notion of connected sum to surfaces with boundary, provided we are careful not to use up any of the already existing boundary circles when we glue the punctured surfaces together. We then find that $D\#D = A = D_o$ and, more generally, $S\#D = S_o$ for any surface S. The full classification theorem for surface with boundary could be formulated in terms of connected sums of copies of T, RP^2 and D; we shall leave this as something for you to think about.



27. *"Walking" one leg of a handle around a Möbius band*

12. Orientation

This is potentially the hardest part of our discussion of surfaces. In this section we shall try to outline how our definition of "orientable" relates to other uses of the word.

On a 1-dimensional figure such as the real line or a circle there are two possible ways to move: left or right along the line, and clockwise or counterclockwise on the unit circle. The standard choice for $\mathbb{R}$ is to move from left to right, and for the circle to move counterclockwise. (Note that stereographic projection translates counterclockwise movement on S^1 into left-to-right movement on $\mathbb{R}$.)

The notion of *orientation* extends the idea of a "preferred direction" to higher dimensions. There are several ways to approach this.

12.1. **Orientability in terms of polygonal decompositions.** An orientation of a polygon is a choice of one of the two possible orientations of its boundary. These orientations determine orientations for each side of the the polygon, which we may indicate by an arrow on each edge. Note that all such arrows meet head-to-tail as we go round the polygon. Conversely such a compatible choice of orientations for the sides determines an orientation of the boundary.

Given two polygons $ABC\dots$ and $PQR\dots$, we may glue them together along a single edge (e.g., by gluing AB to PQ so that $A = P$ and $B = Q$) to obtain a single larger polygon. It is then easy to to see that the orientations of the remaining edges match head-to-tail (and so correspond to a well defined orientation for the new polgygon) if and only if the arrows on AB and PQ are *opposed*.

This leads us to say that a surface $\mathcal{S}$ represented by a union of polygons with gluing instructions is *orientable* if and only if we can consistently choose orientations for each polygon so that whenever two edges are matched (as for $AB = PQ$ above) the arrows on the two edges are opposed. (An orientation for $\mathcal{S}$ is such a consistent set of choices. If $\mathcal{S}$ is connected and orientable there are two possible orientations, since changing the orientation of any one polygon forces all the others to change.)

It is not hard to check that this condition is unchanged under subdivision, and that it is equivalent to the "no Möbius bands" condition.

12.2. **Orientations in higher dimensions.** An orientation for a real vector space V is determined by an ordered basis, and two ordered bases give the same orientation if the change of basis matrix has positive determinant.

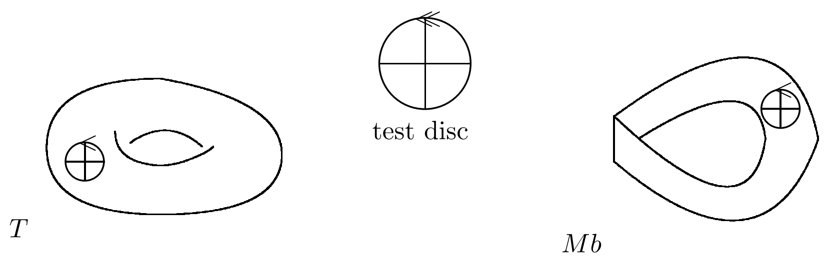
In $\mathbb{R}$ a basis is simply a choice of a nonzero number, and two bases determine the same orientation if and only if the ratio of these numbers is positive. (In terms of the informal discussion in the first paragraph, we move from 0 towards the basis, and thus move from left to right if the number chosen as the basis is positive.)

In general the standard orientation of $\mathbb{R}^n$ is the one corresponding to the standard basis vectors $\mathbf{e_1} = (1, 0, \dots, 0), \dots, \mathbf{e_n} = (0, \dots, 0, 1)$, in that order. (We shall only need $n = 1, 2$ or 3.)

In $\mathbb{R}^2$ we need two basis vectors. The standard choice is $\mathbf{e_1} = (1, 0)$ and $\mathbf{e_2} = (0, 1)$, in that order. (If we switch the order, say $\mathbf{f_1} = (0, 1)$ and $\mathbf{f_2} = (1, 0)$, then the change of basis matrix is $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, which has determinant -1.) Note that if we represent $\mathbf{e_1}$ and $\mathbf{e_2}$ by arrows of unit length at the origin, pointing along the axes, then their tips are on ∂D^2 and we move a quarter-turn counterclockwise from $\mathbf{e_1}$ to $\mathbf{e_2}$. (See Figure 28.)

28. *Orientation*

We shall use the disc marked in this way as a "test pattern". If we can move such a disc around a loop on the surface $\mathcal{S}$ in such a way that when we return to a starting point we have reversed the orientation of the disc we say that $\mathcal{S}$ is *nonorientable*. Otherwise it is orientable. The key example of a nonorientable surface is the Möbius band.



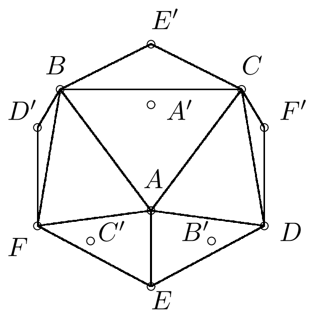
29. *Orientation – test disc*

If $\mathcal{S} \subset \mathbb{R}^3$ is a surface with empty boundary then it must be orientable. The reason is that, firstly, the surface must divide $\mathbb{R}^3$ into two regions; a bounded region inside the surface and an unbounded region outside. (Think of a beachball or car tyre). I hope this is plausible; it requires some work to give a satisfactory proof. A consequence is that at every point on the surface we have a well-defined outward normal vector. We can combine this with the orientation given by our test disc at any point on the surface to get an ordered basis for $\mathbb{R}^3$, $(\mathbf{b_1}, \mathbf{b_2}, \mathbf{b_3})$, say, with $\mathbf{b_3}$ representing the outward normal vector. (See Figure 29.) The change of basis matrix from this to the standard basis $(\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3})$, is just the matrix whose columns are $(\mathbf{b_1}, \mathbf{b_2}, \mathbf{b_3})$. As we move about the surface the basis changes continuously, and so the determinant of the change-of-basis matrix cannot suddenly change sign. Thus the test disc on the surface cannot flip over as we go round a loop. Hence such a surface cannot contain a copy of Mb.

This argument doesn't apply to the Möbius band as it has a boundary, and clearly does not enclose any region of 3-space. However it is enough to show that neither Kb nor RP^2 can be realized as surfaces in R^3; we need more room. (In fact 4 dimensions is enough.)

13. Regular Solids

A regular solid is a convex body in 3-space bounded by a finite number of polygons of the same shape, and such that the same number of polygons meet at each vertex. The classical examples are the tetrahedron, cube, octahedron, dodecahedron and icosahedron. (See Figures 12 and 30.) We shall outline why this is a complete list.



30. *The icosahedron*

The vertices A', B' and C' are "behind" the faces ABC, ADE and AEF, respectively, and the edges $A'B'$, $A'C'$, $A'D'$, $A'E'$, $A'F'$, $B'C'$, $B'D$, $B'E$, $B'F'$, $C'D'$, $C'E'$, $C'F$, $D'E'$ and $E'F'$ are not shown.

The icosahedron has 12 pentangular faces. The dodecahedron may be obtained from the icosahedron by taking the midpoints of all the faces as vertices, and joining any two such vertices by an edge if the corresponding faces of the icosahedron have a common edge. It has 20 triangular faces.

Let B be a regular solid with m faces, each face being an n-gon for some $n \geq 3$, and with $p \geq 3$ faces meeting at each vertex. The faces determine a polygonal decomposition of the boundary $\partial B = S^2$. We have $2E = nF$ (since every face has n sides and each edge is common to two faces), $pV = nF$ (since each face has n corners and each vertex is common to p faces) and clearly $F = m$. Therefore

$$2 = \chi(S^2) = V - E + F = m(\frac{n}{p} - \frac{n}{2} + 1).$$

Clearing denominators gives

$$4p = m(2p - (p-2)n).$$

Hence $(p-2)n < 2p$. Since $n \geq 3$ it quickly follows that $p < 6$. Hence $p = 3, 4$ or 5. Examining the possible factors of $4p$ leads quickly to the possibilities $(m, n, p) =$

$(4,3,3)$, $(6,4,3)$, $(8,3,4)$, $(12,5,3)$ or $(20,3,5)$. It is not hard to see that in each case the polygons must be assembled as in one of the classical regular solids.

14. Nonplanarity of K_n, $n \geq 5$

Recall that a graph G is planar if we can embed it in $\mathbb{R}^2$ (i.e., if we can draw it in the plane without crossings other than the vertices). It is convenient to modify this definition slightly. If G is planar then it embeds in $S^2 = \mathbb{R}^2 \cup \{\infty\}$. Conversely, if $G \subset S^2$, we may move it to miss the "north pole" $N = \infty$. Stereographic projection from N then embeds G in the plane. Thus G is planar if and only if it embeds in the 2-sphere S^2.

Let K_n be the complete graph on n vertices. We saw that K_n is planar if $n \geq 4$, and shall show now no other complete graph is planar. It shall suffice to show that K_5 is not planar, since $K_5 \subset K_n$ for all $n > 5$.

Suppose $K_5 \subset S^2$. Then the complement $S^2 - K_5$ breaks up into a finite number of regions, bounded by circuits of K_5. Since K_5 is connected these regions are all polygons, with a single boundary. (This would not be true for a disconnected graph, and is not always true for graphs on other surfaces.) This we have a polygonal decomposition of S^2, with $V = 5$ vertices, $E = \binom{5}{2} = 10$ edges and F faces. Since $\chi(S^2) = 2 = V - E + F$ we find that $F = 7$. Now each face must have at least 3 edges, since K_5 has no multiple edges, and each edge is common to two faces. Therefore $2E \geq 3F$, and so $20 > 21$. This is clearly wrong, and so K_5 cannot be planar.

A similar argument applies to the "Three Utilities" graph $K_{3,3}$. (This notation reflects the fact that it belongs to another standard family, the family of "complete bipartite graphs".) One of the basic results of topological graph theory is Kuratowski's Planarity Theorem, which asserts that a finite graph G is planar if and only if it has no subgraph which is homeomorphic to K_5 or to $K_{3,3}$.

15. The 5-Colour Theorem

Map colouring began as recreational mathematics. In 1852 F.Guthrie raised the question of how many colours were needed to colour a map so that no two neighbouring countries had the same colour. It is easy to see that in general at least 4 colours are needed. (Consider Botswana, Luxembourg or Paraguay, which each have three neighbours. See Figure 31.) Experimentation produced no examples requiring more than 4 colours. This became the "Four-Colour Conjecture". It was eventually confirmed in 1976, by a proof using several hundred hours of mainframe computer time (perhaps one hour on a PC now?), to check a large number of special cases. The only mathematicians who are happy with this proof are those with a vested interest in computation, but computation has been since used to decide other difficult problems, and so such techniques are becoming more acceptable. Surprisingly, the corresponding question for maps on more general surfaces has been settled by "pure thought", and the *Five* Colour Theorem is quite easy.

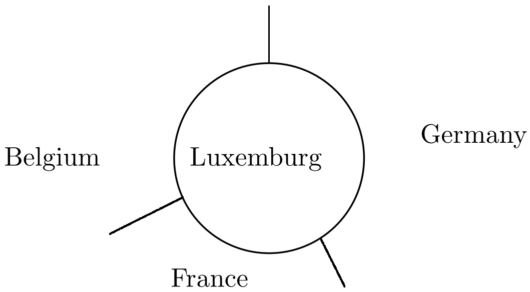
We shall show that every map on the 2-sphere can be coloured with at most 5 colours. The basic idea is to induct on the number of regions (i.e., countries or bodies of water).

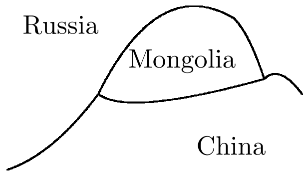
A *map* on a surface $\mathcal{S}$ is a graph embedded in $\mathcal{S}$. The map divides $\mathcal{S}$ into a finite number of connected regions. Two regions are neighbours if they share a common edge. The map is *special* if all the complementary regions are polygons. It then determines a polygonal decomposition of the surface, and we may use the Euler characteristic to restrict the possibilities.

We shall concentrate on the case $\mathcal{S} = S^2$. We first make some simplifying assumptions. (These do not hold for all maps of the Earth!)

(1) No region shares a border with itself. (This is obviously a reasonable assumption.)

(2) We may assume that no region has only two borders. (This condition fails to hold for Mongolia.) For if region A meets just regions B and C and we can 5-colour the map obtained by absorbing A into B or C then we may give A any colour different from the ones used for B and C. (This argument applies also if $B = C$.)





31. *Four colours are needed sometimes*

(3) We may assume no region entirely surrounds another. (I.e., there are no lakes or countries completely surrounded by another country, like the ACT or Lesotho!) For if we can 5-colour the map obtained by deleting the surrounded region we can then recolour the deleted region differently from its only neighbour. (See also condition (5) below.)

(4) We may assume that each vertex has degree 3, or, equivalently, at each vertex exactly 3 regions meet. (This condition fails to hold at one point in the USA: Colorado, Utah, Arizona and New Mexico meet at one point.) For if 4 or more regions meet at a vertex P we may expand P to a small disc. If the resulting map can be 5-coloured so can the original map.

(5) If we can 5-colour all special maps with at most n regions then we can 5-colour all maps with at most n regions. For if R is a region which is not polygonal then it must have at least two boundary circles, C and C', say. Consider two new maps formed by (i) annexing all regions on the other side of C; and (ii) annexing all regions on the other side of C'. These maps have fewer regions, and so each can be 5-coloured. We may assume the colour used for R is the same in both cases. Together these colourings show that the original map may be 5-coloured.

Clearly any map with at most 5 regions can be 5-coloured. (This is the basis for our induction.) Suppose that all maps with fewer than n regions and with all vertices of degree 3 can be 5-coloured, and let M be a special map with n regions and all vertices of degree 3. Let V and E be the numbers of vertices and edges of the corresponding polygonal decomposition of S^2. Then $2E = 3V$, since each edge has two ends and each vertex has degree 3, and

$$2 = \chi(S^2) = V - E + F = n - \frac{1}{3}E,$$

since there are n faces (regions). Let p_k be the number of regions with exactly k edges. Then $p_k = 0$ if $k \leq 2$ and so

$$n = \sum_{k\geq 3} p_k$$

while

$$2E = \sum_{k\geq 3} kp_k$$

(since each edge is common to two regions). Hence

$$2 = n - \frac{1}{3}E = n - \frac{1}{6}2E = \sum_{k\geq 3}(1 - \frac{k}{6})p_k = \frac{1}{2}p_3 + \frac{1}{3}p_4 + \frac{1}{6}p_5 + \sum_{k\geq 6}(1 - \frac{k}{6})p_k.$$

Since $\sum_{k\geq 6}(1 - \frac{k}{6})p_k \leq 0$ at least one of p_3, p_4 or p_5 must be nonzero.

If $p_3 > 0$ there is a triangular region R. The three neighbouring regions must all meet each other (since the vertices have degree 3), and so are all distinct. If we allow R to be annexed by one of these regions we get a (special) map with all vertices of degree 3 and with one less region. It may be 5-coloured, by the inductive hypothesis. Since only three colours are needed for the neighbours we can use one of the other two colours to recolour the region R differently.

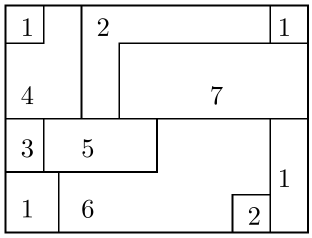
If $p_3 = 0$ but $p_4 > 0$ there is a quadrilateral region R, with four neighbours. These neighbours need not be all distinct. Annexing R by one of its neighbours gives a map with all vertices of degree 3 and one less region (which, however, may not be special). As before, we can 5-colour this map and then recolour the region R.

If $p_3 = p_4 = 0$ then there must be a pentagonal region. In this case annexing R by one of its neighbours won't help. However we note that there must be a pair of neighbours which have no common edge elswhere. (*WHY?) We combine these two neighbours and R into one large region, 5-colour the resulting map, and then recolour the region R. (Thus only 4 colours are needed for the five neighbouring regions.)

This completes the inductive step. Therefore all maps in which all vertices have degree 3 can be 5-coloured. Hence all maps can be 5-coloured, by the comments following condition (4). $\square$

For more general surfaces, we introduce the following notation. The *chromatic number* $\mu(M)$ of a map M on a surface $\mathcal{S}$ is the minimal number of colours need to colour the map, and the chromatic number $\mu(\mathcal{S})$ of the surface is the maximum number of colours needed for any map on $\mathcal{S}$. (Thus we have seen that $\mu(S^2) \leq 5$.) Heawood showed that for any surface $\mathcal{S}$ other than S^2 we have $\mu(\mathcal{S}) = [\frac{1}{2}(7 + \sqrt{49 - 24\chi(\mathcal{S})}]$, where $[x]$ means the greatest integer less than or equal to x. (He showed this number was an upper bound, but the task of constructing examples showing this formula to be correct was not completed until 1968. See *Map Color Theorem* by G.Ringel.)

In particular, if $\mathcal{S} = T$ Heawood's formula gives $\mu(T) = 7$. There is a map on the torus with 7 polygonal regions which all meet each other, and thus which needs 7 colours.



32. *A 7-colour map on the torus*

16. Knots

A *knot* is a simple closed curve in 3-space. In other words, it is a subset of $\mathbb{R}^3$ which is homeomorphic to the circle S^1. Two knots are *equivalent* if one may be continuously deformed to the other, so that no strands cross through each other at any intermediate stage.

In this course we do not want to get bogged down in technicalities about continuous functions, so we shall modify our definitions to have a more "combinatorial" flavour. A *polygonal* knot is one which is the union of finitely many straight line segments, meeting only at their endpoints. Two polygonal knots are *combinatorially equivalent* if we can get from one to the other by repeatedly

(1) adding a vertex to a line segment;
(2) replacing one line segment PQ by the other two sides of a triangle PQR which meeets the knot only along PQ; and
(3) the reverse of either of these types of move.

It is however usually easier to draw knots as smooth curves, and the figures look nicer, so we shall generally do this. We may obtain a model of a knot in the mathematical sense from one in the usual (sailor's) sense by splicing together the ends of a knotted rope.

Examples. The unknot U is the standard unit circle in the XY-plane.

A knot is *trivial* if it is equivalent to the unknot.

The trefoil knot 3_1 and its mirror image. (These are not equivalent, although we shall not attempt to prove this.)

The figure eight knot 4_1. It is equivalent to its mirror image.

The symbols 3_1, 4_1 refer to standard tables of the simpler knots, which are organized in terms of the number of *crossings* of the simplest possible diagram of an equivalent knot. We may assume that the knot K is positioned in $\mathbb{R}^3$ so that no line segment is vertical, and so that its *projection* (the shadow in the horizontal XY-plane) is a closed curve with a finite number of crossings, where two arcs of the diagram cross. We may assume also that there are no points where 3 or more arcs cross. We may recover the knot from such a diagram, provided we indicate at each crossing which arc corresponds to the "higher" line segment. We do this by making a gap in the arc corresponding to the lower line segment.

The effect of deforming a knot combinatorially (replacing a line segment by the other sides of a triangle) on the knot diagram may be described in terms of three basic operations on diagrams, the *Reidemeister moves.* Two knots are equivalent if they have diagrams which may be related by a sequence Reidemeister moves.

Any quantity defined from a knot diagram which is unchanged under each of the basic Reidemeister moves depends only on the combinatorial equivalence class of the knot, and so we may call such a quantity an *invariant* of the knot.

Note that if a knot is different from its mirror image, the standard tables include only one of each such pair. Thus they include only one 3-crossing knot.

It is often convenient to assume that our knots are oriented. An *orientation* for a knot is a choice of one of the two possible directions in which to go round the knot. For the unknot U in the XY-plane, the possibilities may be described as clockwise or counterclockwise (seen from above), but for more general knots such terms are inappropriate.

See *The Knot Book*, by C.C.Adams.

17. 3-colourable knots

A knot diagram D is 3-colourable if we can colour each of the arcs with one of three colours (say, Red, Blue or Green) so that

(1) at each crossing either all 3 colours are the same or all 3 are different; and
(2) all 3 possible colours are used.

It is not hard to check that if D' is obtained from D by a single Reidemeister move then D is 3-colourable if and only if D' is 3-colourable. This extends immediately to show that if two diagrams are related by a sequence of Reidemeister moves and one is 3-colourable then so is the other. Thus if a knot has one 3-colourable diagram all the diagrams of this knot are 3-colourable, and so this is a property of the knot.

Examples: the trefoil 3_1 is 3-colourable, but neither the unknot U nor the figure eight knot 4_1 are. Thus the trefoil knot is nontrivial.

We need a new idea to show that the figure eight knot is nontrivial.

18. The knot determinant

Given a diagram D of a knot K, with n crossings say, we shall define an $n \times n$ matrix. This matrix shall have the property that $\Sigma cols = 0$, and hence in particular its determinant is 0. In the special case that the diagram is *alternating*, the sum of the rows is also 0. In this case we may define the determinant of the knot as the absolute value of the determinant of *any* $(n-1) \times (n-1)$ submatrix (obtained by deleting a row and a column). It is a straightforward exercise in linear algebra (using the properties $\Sigma cols = 0$ and $\Sigma(rows) = 0$) to show that this number does not depend on which row and column are deleted.

For more general diagrams there is an additional complication.

The procedure.

1. Choose a starting point and an orientation for the knot, and label the arcs successively: $x_1, x_2 \dots, x_n$. (We count cyclically. Thus we interpret $n+1$ as 1, where necessary.)

2. Write down a $n \times n$ matrix M whose entries are 2s, -1s, 0s and possibly 1s as follows. On the diagonal put $M_{i,i} = -1$ for $1 \leq i \leq n$.

On the "superdiagonal" put $M_{i,i+1} = -1$ for $1 \leq i < n$. Put also $M_{n,1} = -1$.

If the j^{th} arc separates the i^{th} and $(i+1)^{st}$ arcs put $M_{i,j} = 2$.

If more than one of these rules apply to a given entry add each of the contributions. Thus if the i^{th} arc turns back to go underneath itself (so $j = i$) we put $M_{i,i} = +1 = -1 + 2$. If $j = i + 1$ we put $M_{i,i+1} = +1$. Likewise in the exceptional case of the 1-crossing diagram of the unknot, when all three arcs are the same: we get the 1×1 matrix $M = [-1 - 1 + 2] = [0]$.

All other entries are 0.

4. Delete any one row and any one column. The highest common factor of the determinants of the resulting $(n-1) \times (n-1)$ matrices is the determinant of the diagram. (Note that we can always delete the last row; we then have n determinants to evaluate. In the alternating case these agree up to sign.)

FACT. The determinant is unchanged if we modify the diagram by a Reidemeister move, and so is an invariant of the knot. Therefore we shall write $DET(K)$ for this invariant.

$DET(K)$ is clearly non-negative. It is always odd; in particular, it is nonzero. This is easily seen by using modular arithmetic: working *mod* (2). For $-1 \equiv 1$ and $2 \equiv 0$ *mod* (2), so if we delete the last row and column we get an upper triangular matrix with 1s down the diagonal. Hence $DET(K) \equiv 1$ *mod* (2). (This argument works also in the nonalternating case, since it is enough to show that one of the subdeterminants is odd.)

Examples. $DET(U) = 1$. This is easily seen if we use the ABC logo, which has two arcs. (We could use the 1-crossing diagram instead, if we accept the convention that the determinant of a 0×0 matrix is always $+1$.) The standard diagrams for the trefoil knot and for the figure eight knot are alternating, so we need compute only one determinant in each case. We find $DET(3_1) = 3$ and $DET(4_1) = 5$.

19. Seifert surfaces

The single most useful geometric property of knots is that they can be "spanned" by orientable surfaces. There is a very nice algorithm for constructing such a surface, due to H.Seifert.

1. Choose an orientation for the knot, and put arrows in the preferred direction on each arc segment between consecutive crossings.

2. Starting anywhere on the knot projection, follow the arrows around to a crossing point (over or under). Then jump onto the other arc at the crossing and continue in the preferred direction.

3. After finitely many such jumps, the starting point is reached again, and the route taken is a simple closed curve in the plane.

4. Start again at any point not on an arc segment already used.

5. At the end of this procedure you have a collection of disjoint simple closed curves in the plane, called *Seifert circles*. Some of these may be nested, i.e., one inside another. Each such circle bounds a disc. If one circle is inside another we stack the discs at different heights, so they are disjoint.

6. Atttach the discs together by half-twisted ribbons at each crossing point. There are two possible ways of giving a ribbon a half twist; choose the one similar to the type of crossing.

7. The resulting surface is connected, orientable and has one boundary circle – the original knot!

What else can we say about the surface? By the classification theorem, all we need to know is the Euler characteristic. Suppose that the diagram has n crossings, and there are c Seifert circles. The Seifert surface has a polygonal decomposition with $F = c + n$ faces (the c discs and n rectangular ribbons) and $V = 4n$ vertices (the corners of the rectangles). Finding the number of edges requires a little more thought: there are $4n$ edges coming from the rectangles, and another $2n$ edges coming from the arc segments between crossings. Thus $\chi(S) = 4n - 6n + c + n = c - n$.

Note that there is an "obvious" surface with boundary the trefoil knot; give a rectangular ribbon three half twists before joining the ends. However this surface is a Möbius band, and so is not orientable.

20. Genus and knot arithmetic

There are many of possible spanning surfaces, for we may always add extra handles, without changing the boundary knot. Adding handles increases the genus of the surface. It is natural to ask: given a knot, what is the minimal genus of a spanning surface. Knot theorists find it convenient to use this number rather than the Euler characteristic. (See the next section.)

The *genus* $g(K)$ of a knot K is the minimal genus of any connected spanning surface for K. (Recall that if S is a connected orientable surface with one boundary circle then $\chi(S) = 1 - 2g(S)$ and so $g(S) = \frac{1}{2}(1 - \chi(S))$.)

Examples. $g(U) = 0$; $g(3_1) = g(4_1) = 1$.

A knot K is trivial if and only if $g(K) = 0$. For then K bounds a disc D in 3-space. We may "flatten out" D and so K is equivalent to a circle in the plane, i.e., to the unknot.

We know the trefoil is nontrivial, and so $g(3_1) > 0$. The Seifert surface obtained from the standard 3-crossing diagram is a punctured torus, so $g(3_1) \leq 1$. Therefore $g(3_1) = 1$. (Similarly for the figure eight knot.)

The genus is particularly useful because it is *additive*. There is a natural notion of *knot sum*; tie two knots in succesion in the same piece of string before splicing the ends together. This idea can be given a precise mathematical formulation; we assume K and K' are *oriented knots*, with $K \subset \mathbb{R}^3_+$ (upper half 3-space, with $z \geq 0$) and $K' \subset \mathbb{R}^3_-$ (lower half 3-space, with $z \leq 0$) and $K \cap K'$ an arc α in the XY-plane ($z = 0$). We assume also that K and K' determine *oppsite* orientations of α. Then we obtain the sum by deleting the overlap α (except for its endpoints): $K\#K' = K \cup K' - \alpha$. Note that $K\#K'$ has a natural orientation compatible with those of K and K'.

It is is not hard to show that the sum is well-defined, and is quite easy to see that it has the following nice properties

(1) (identity element) $K\#U = K = U\#K$ for any knot K;
(2) (commutativity) $K\#K' = K'\#K$ for any two knots K, K';
(3) (associativity) $(K\#K')\#K'' = K\#(K'\#K'')$ for any three knots K, K', K''.

These properties are analogous to familar properties of the integers $\mathbb{Z}$ with respect to addition, or better the positive integers $\mathbb{Z}_{>0} = \{1, 2, \dots\}$ with respect to multiplication.

Examples. The reef knot is the sum of the trefoil 3_1 and its mirror image $r(3_1)$. The granny knot is the sum of two copies of the trefoil.

FACT. $g(K\#K') = g(K) + g(K')$. ("Additivity of the genus".)

It is very easy to show that $g(K\#K') \leq g(K) + g(K')$. Showing that these numbers are equal is not hard, but requires some care. The basic idea is to assume that $K \subset \mathbb{R}^3_+$ and $K' \subset \mathbb{R}^3_-$ and that S is a surface with boundary $K\#K'$. The intersection of S with the XY-plane is nonempty, since the knot $\partial S = K\#K'$ meets the XY-plane in two points. After juggling S slightly, if necessary, we may assume that S meets this plane in a finite collection of curves. These curves are either simple closed curves or arcs with endpoints on ∂S. Since any arc has two endpoints there is exactly one such arc. We then show that we can modify S to successively remove the simple closed curves, and that the modification does not increase $g(S)$.

An immediate consequence is that no two knots can ever cancel each other out. For if $g(K) > 0$ or $g(K') > 0$ then $g(K\#K') > 0$, and so if K or K' is nontrivial then so is $K\#K'$.

A deeper fact is that prime factorization holds. We say that a knot is *indecomposable* if it is nontrivial but is not the sum of two nontrivial knots. For instance, it follows from additivity of the genus that if $g(K) = 1$ then K is indecomposable. (For if $K = K_1\#K_2$ and $1 = g(K) = g(K_1) + g(K_2)$ then $g(K_1)$ or $g(K_2)$ must be 0.)

FACT. Every knot is a sum of indecomposable knots in an essentially unique way. (The factors are unique up to order).

The proof of this fact uses similar ideas to those used in showing that the genus is additive, but is harder.

This result is analoguous to the unique factorization of positive integers as products of prime numbers. Thus knot sum is better behaved then surface sum.

The standard knot tables give only prime knots, so although the reef knot or the granny knot have 6-crossing diagrams you will not find them in the standard tables. (Note also that if a knot is different from its mirror image, the tables include only one of each such pair.)

21. References

Map Color Theorem, G.Ringel ([517.521/46] in the Maths Library.)
The Knot Book, C.C.Adams

The following material was not discussed in 2006.

22. Links and linking numbers

Suppose that we have two disjoint knots K_1 and K_2 in 3-space. The linking number measures the number of times one knot passes through the other. To understand what "passes through" should mean, we consider some examples.

[See Figures]

There is one ingredient missing. We should specify the *orientations* of the component knots.

We may compute the linking number in several ways. One involves spanning surfaces. Let F_2 be an orientable surface in 3-space with boundary K_2. Define TOP and BOTTOM sides of F_2 by the following rule: if we stand a figure on the top side, beside the boundary K_2, facing along K_2 in the direction given by the orientation, then the surface is on the left. We may assume that K_1 passes through F_2 in finitely many places P. Assign a number $\varepsilon_P = \pm 1$ to each such point , using the rule $\varepsilon_P = +1$ if K_1 passes through from TOP to BOTTOM, and $\varepsilon_P = -1$ otherwise.

The other method refers directly to the crossings in a diagram. It is a straightforward algorithm, but it is perhaps less clear that the calculation has any geometric significance. We assume the knots oriented. At each point P where K_1 passes over K_2 we set $\varepsilon_P = \pm 1$ according to the rule illustrated in Figure ??

Using either approach, we define the linking number to be the sum of these terms: $\ell(K_1, K_2) = \Sigma \varepsilon_P$.

Note in particular that this rule is consistent with our earlier calculation for the most basic nontrivial link.

Properties: $\ell(K_1, K_2) = \ell(K_2, K_1)$. If we change the orientation of one of the knots we change the sign of the linking number. If we reflect the link we change the sign of the linking number. Thus if the orientations are not specified in advance we may get a well defined invariant by taking the absolute value of the linking number.

23. Other related topics

Knots and links in complete graphs

Visualizing knotted surfaces in 4-space

School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia
E-mail address: `jonathan.hillman@sydney.edu.au`