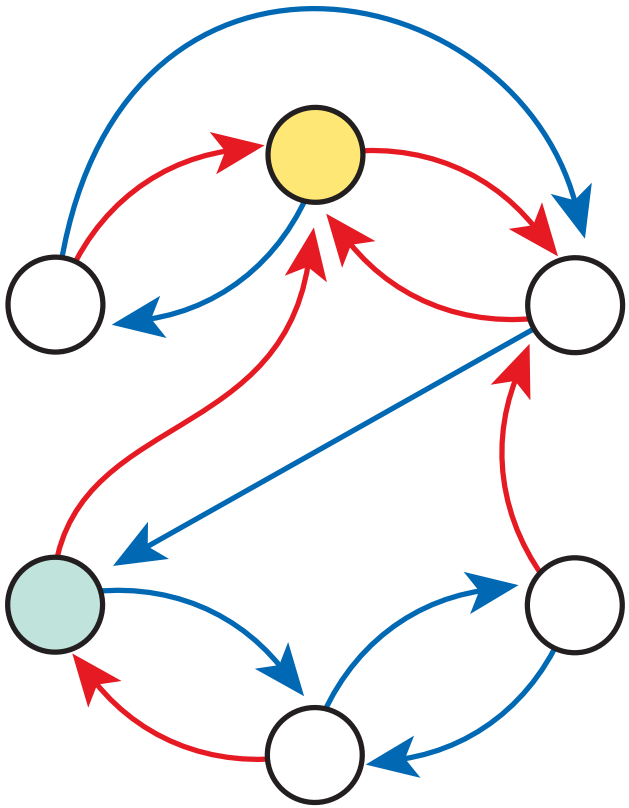
The material in these slides is not examinable

In Spring 2008 a Discrete Maths student called Matthew Kitchen asked me about an article in *The Telegraph* that quoted Marcus du Sautoy:

> *You have a map marked with lots of cities (represented by dots) and a set of roads running between these cities (the roads are then lines joining the dots). The roads are all one-way roads so you can only go along them in one direction.*
>
> *Suppose you live in one of the cities on this map. I'm in another city. You want to give me instructions which will get me from my city to your city. But you don't know where I am. Is there a way to colour the roads and a set of instructions like "Take a red road followed by a blue road followed by another blue road followed by a yellow road" so that where ever I am in the map, the set of instructions will always get me to your city?*

The story explained that Avraham Trahtman, a Russian emigre who had moved to Israel, had resolved[1] the Road Colouring conjecture in the affirmative, proving that such directions always exist.

[1] A. Trahtman (2007), The Road Colouring Problem, http://arxiv.org/abs/0709.0099



Here is an example with two kinds of road: following the sequence of edges (read left-to-right)

`Red Red Blue Red`

will get you from any vertex to the yellow one.

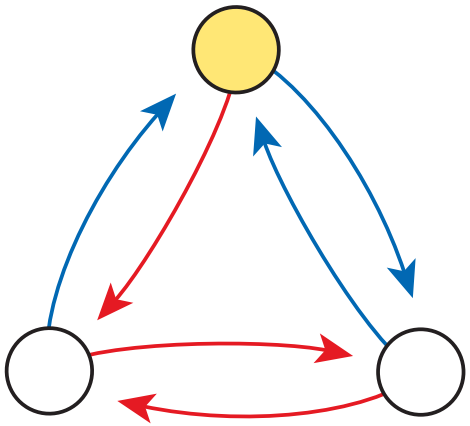
Further, as the graph is strongly connected, it's possible to make up (somewhat longer) universal directions to any other vertex:

`Red Red Blue Red Red Blue`

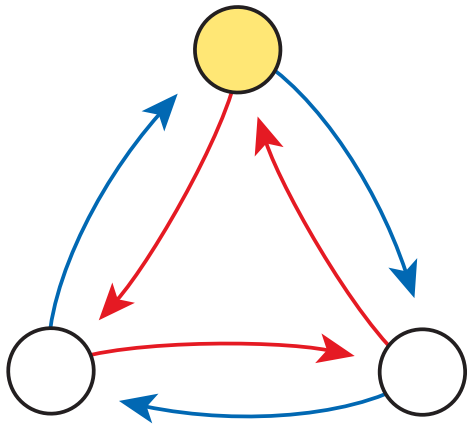
are universal directions for the minty green vertex.

Not all edge colourings work

The existence of universal directions depends on both the graph and the colours of the edges:



With edges coloured as above, the sequence `Red Blue` will get us from any vertex to the yellow one.



Here there are no universal directions.

The Road Colouring Problem was a conjecture due to Adler and Weiss[2] that arose in an entirely different part of mathematics. Their version actually uses the term "Road colouring", but an equivalent formulation is:

Conjecture (The Road Colouring Problem)

Let $G(V, E)$ be a directed, strongly-connected graph. If

(i) *all vertices share a common out-degree, $\deg_{out}(v) = k \ \forall v \in V$ and*

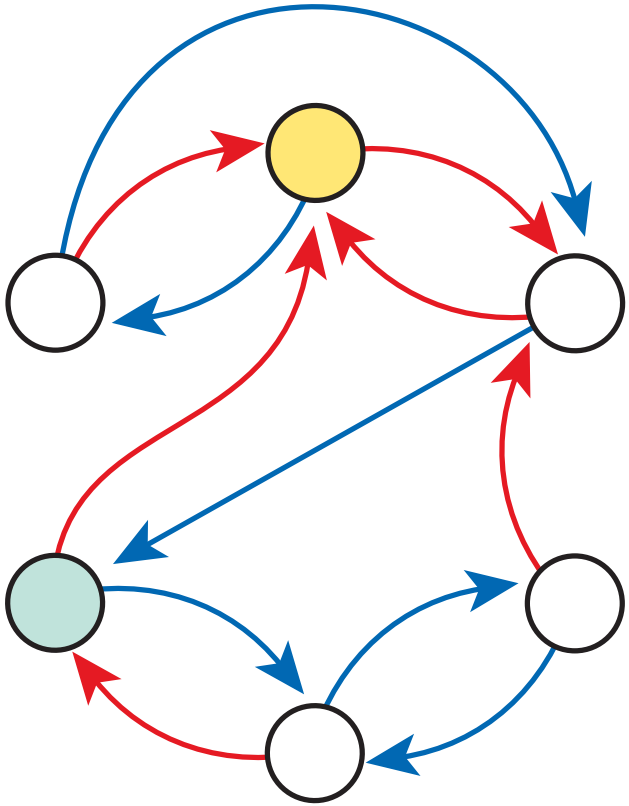
(ii) *the lengths of the cycles in G have greatest common divisor 1,*

then there is an edge-colouring that allows one to give a universal set of directions (in the form of a sequence of edge labels) that reach some distinguished vertex $v_\star$ from any starting vertex.

Such an edge colouring is called a *synchronizing colouring* and the set of directions is its *synchronizing word*.

[2] It's implicit in R. L. Adler and B. Weiss (1970), Similarity of automorphisms of the torus, *Memoirs of the AMS*, **98** and is stated explicitly in R. L. Adler, W. Goodwyn and B. Weiss (1977), Equivalence of Topological Markov Shifts, *Israel J. Maths.*, **27**, 46–63.

Our initial example came from Adler, Goodwyn and Weiss's 1977 paper.



Clearly:

- G is strongly-connected;
- G has uniform out-degree $k = 2$;
- G has cycles of length 2 and 3, so it satisfies the GCD condition.

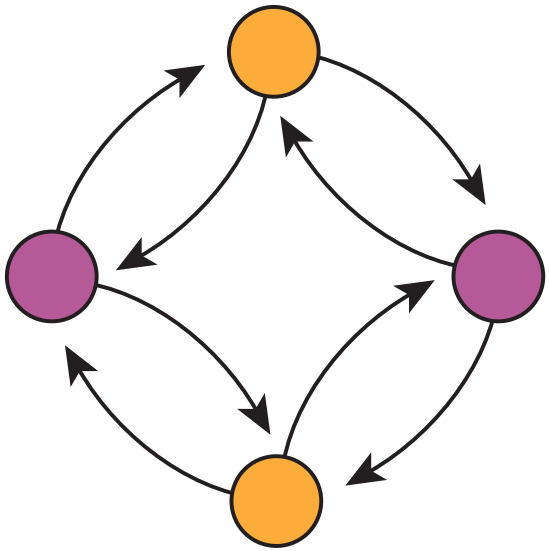
And, as we've seen, following a sequence of edges coloured

`Red Red Blue Red`

will get you from any starting vertex to the yellow one.

The GCD condition: illustrating the issue

The graph at left, which has cycles of length 2 and 4, cannot have a synchronizing colouring.



- All edges that start at an orange vertex point to a purple one, and all those that start at a purple vertex point to an orange one.
- Thus any walk of even length that starts on an orange vertex must also finish on an orange one.
- Any walk of even length that starts on a purple vertex must finish on a purple one.
- Walks of odd length go either orange-to-purple, or purple-to-orange.

Thus no matter how we colour the edges, there is no set of directions (what sort of length—even or odd—could it have?) that can send both an orange vertex and a purple one to the same destination.

The sickness on the previous slide is an example of a well-understood[3] phenomenon:

Theorem

If $G(V, E)$ is a strongly-connected digraph then the following two properties are equivalent

(i) *The lengths of the cycles in G have greatest common divisor $d > 1$.*

(ii) *One can partition the vertex set of G into d disjoint subsets*

$$V = V_0 \cup V_1 \cup \cdots \cup V_{d-1}$$

in such a way that every edge points from one group to the next. That is $(u, v) \in E$ and $u \in V_j \Rightarrow v \in V_{j+1}$, where the subscripts are taken $\bmod d$, so $V_d = V_0$

Such a graph is called *periodic* and, as the previous example illustrates, a periodic graph can't have a synchronizing word.

If the lengths of a graph's cycles have greatest common divisor $d = 1$, the graph is called *aperiodic*.

[3] I learned about it from J.P. Jarvis and G.R.Shier (2000), Graph-Theoretic Analysis of Finite Markov Chains, Chapter 13 in *Applied Mathematical Modeling: a multidisciplinary approach*, D.R. Shier and K.T. Wallenius, eds., Vol. **15** in *Discrete Mathematics and Its Applications*, Chapman & Hall / CRC.

Another problem, not obviously involving graphs . . .

Ribonucleic acid (RNA) is a family of molecules whose members play a great many roles[4] in the chemistry of living cells. RNA is similar to its more famous sister molecule, DNA (deoxyribonucleic acid), in that it consists of long chains composed of 4 subunits called *bases*. For RNA the names of the bases are usually abbreviated `U`, `C`, `G`, and `A` and they can be combined in any order, so a mathematician can, if she likes, think of an RNA molecule as a string of letters such as

`AGUCAGUGAGCA`

These slides are available on Blackboard and at `https://bit.ly/3glVwmj`

[4] If you want to learn more about RNA and its amazingly rich biochemistry, I recommend a book by (Manchester Uni's own) Terry Brown, *Genomes 3*, Garland Scientific (2006). The previous edition, *Genomes 2*, is freely available online at https://bit.ly/Genomes2.

Sequencing RNA in the 1960's

In the early days of RNA research biochemists could sequence accurately only rather short pieces of RNA, but were interested in much longer chains. They addressed the problem by chopping up the longer chains into smaller fragments that they could sequence directly. These collections of smaller fragments are called *digests* of the original molecule and one way to sequence a long chain is to compare and combine the results of two different sorts of digest.

These slides are available on Blackboard and at `https://bit.ly/3glVwmj`

Two different sorts of digest

Our first enzyme cuts the RNA sequence after every G, while the second cuts after both U and C:

$$\texttt{AGUCAGUGAGCA} \xrightarrow{\text{G-digest}} \texttt{AG UCAG UG AG CA}$$

$$\texttt{AGUCAGUGAGCA} \xrightarrow{\text{UC-digest}} \texttt{AGU C AGU GAGC A}$$

Of course, the digests don't come so neatly organized: in a real scientific problem we'd just get two jumbles of short fragments such as

G-digest: AG AG CA UG UCAG
UC-digest: A C AGU AGU GAGC

and have to work out how to rearrange them in a sensible order.

These slides are available on Blackboard and at https://bit.ly/3glVwmj

A thought experiment: digesting the digests

Imagine attacking each fragment from the G-digest with the UC-enzyme. They'd break down as follows:

$$\texttt{AG} \xrightarrow{\text{UC-digest}} \texttt{AG} \qquad \texttt{UCAG} \xrightarrow{\text{UC-digest}} \texttt{U C AG}$$

$$\texttt{UG} \xrightarrow{\text{UC-digest}} \texttt{U G} \qquad \texttt{AG} \xrightarrow{\text{UC-digest}} \texttt{AG} \qquad \texttt{CA} \xrightarrow{\text{UC-digest}} \texttt{C A}$$

The small pieces produced in this way are called *extended bases*.

Note that we'd get the same set of extended bases if we'd started with the fragments of the UC-digest and attacked then with the G-enzyme.

These slides are available on Blackboard and at `https://bit.ly/3glVwmj`

We'll mainly be interested in those cases, such as UCAG $\xrightarrow{\text{UC-digest}}$ U C AG, where a second digest produces two or more extended bases: we'll use them to draw a very helpful graph.

The results of the G-digest were

AG AG CA UG UCAG

so the double-digests that yield two or more pieces are

$$\text{CA} \xrightarrow{\text{UC-digest}} \text{C A}$$

$$\text{UCAG} \xrightarrow{\text{UC-digest}} \text{U C AG}$$

$$\text{UG} \xrightarrow{\text{UC-digest}} \text{U G}$$

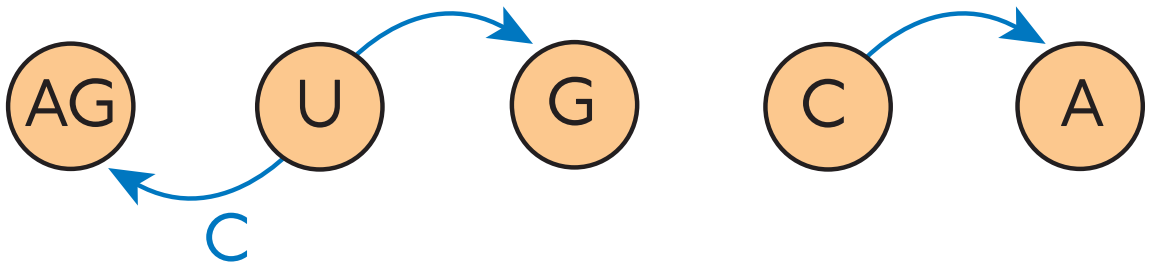
These slides are available on Blackboard and at https://bit.ly/3glVwmj

Starting from the double-digests of the G-fragments

$$\texttt{CA} \xrightarrow{\text{UC-digest}} \texttt{C A} \qquad \texttt{UG} \xrightarrow{\text{UC-digest}} \texttt{U G} \qquad \texttt{UCAG} \xrightarrow{\text{UC-digest}} \texttt{U C AG}$$

we build a graph which has:

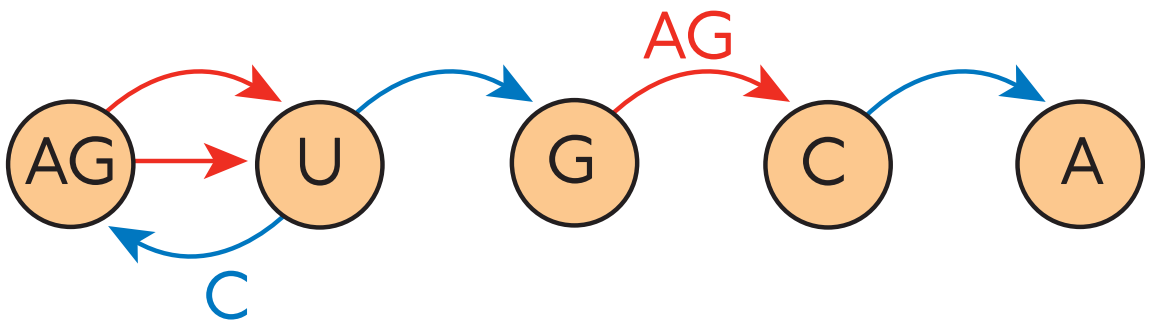
- a vertex for each extended base that appears on the right-hand side of one of the double-digests;
- a directed edge (a bit like a bridge with a one-way street on it) connecting the first and last extended base in each of the double-digests and
- for those double-digests where we get three or more extended bases, an edge label.



We now produce a second set of double-digests by attacking the original UC-fragments—`A C AGU AGU` and `GAGC`—with the G enzyme. The ones that break into two or more extended bases turn out to be

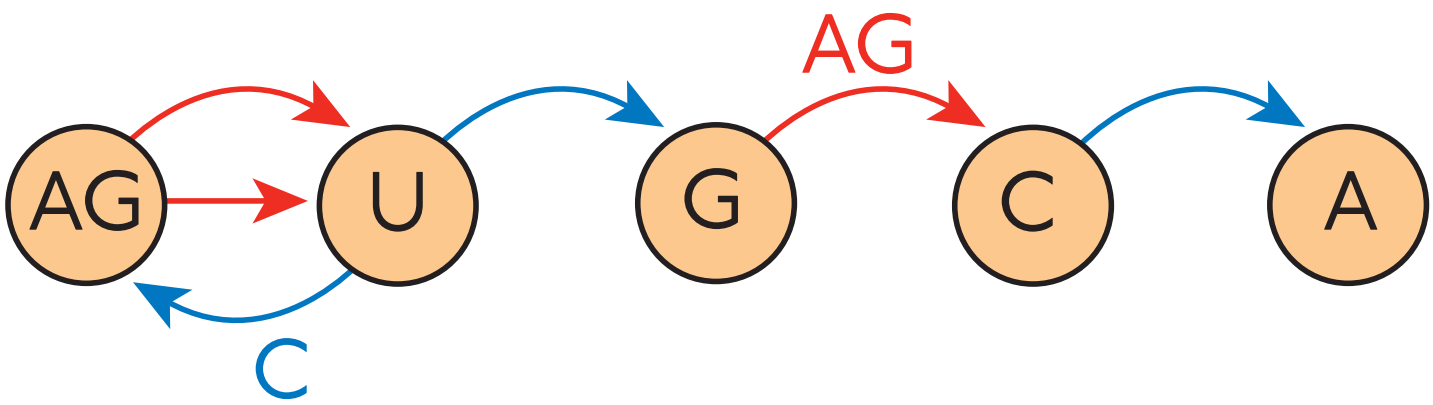
$$\texttt{AGU} \xrightarrow{\text{G-digest}} \texttt{AG U} \text{ (twice)} \qquad \texttt{GAGC} \xrightarrow{\text{G-digest}} \texttt{G AG C}$$

and if we then add the corresponding edges and labels to our graph we end up with:



where the edges and labels arising from the second group of double-digests are shown in red.

These slides are available on Blackboard and at `https://bit.ly/3glVwmj`



There's exactly one (directed) Eulerian trail through this graph and, if we write down the labels of its vertices and edges in the order we encounter them we get

$$\texttt{AG} \cdot \texttt{U} \cdot \texttt{C} \cdot \texttt{AG} \cdot \texttt{U} \cdot \texttt{G} \cdot \texttt{AG} \cdot \texttt{C} \cdot \texttt{A} = \texttt{AGUCAGUGAGCA}$$

which is the original sequence!

These slides are available on Blackboard and at https://bit.ly/3glVwmj

Theorem (Directed Eulerian Tours)

Let $G(V, E)$ be a strongly-connected, directed multigraph. Then the following statements are equivalent:

(i) *G has an Eulerian tour.*

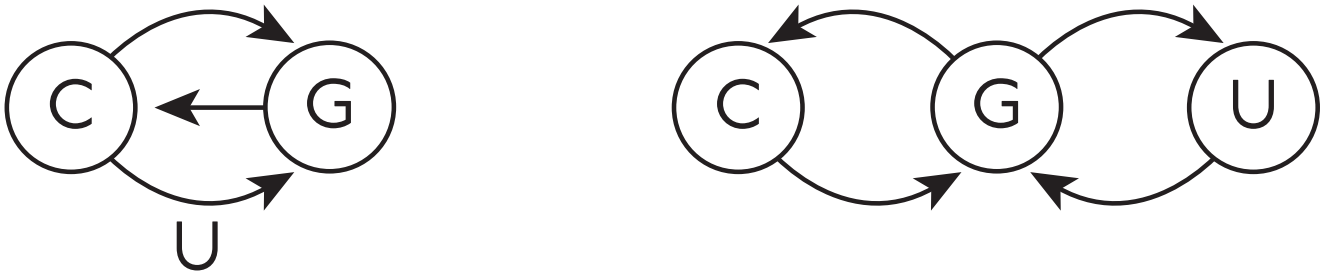
(ii) *For all vertices $v \in V$ we have $\deg_{in}(v) = \deg_{out}(v)$.*

(iii) *The edge set of G can be partitioned into cycles.*

The proof of this theorem is very similar to that of the undirected version covered in the vides, slides and lecture notes.

These slides are available on Blackboard and at `https://bit.ly/3glVwmj`

Some graphs have multiple tours or trails



The graph at left above contains two Eulerian trails:

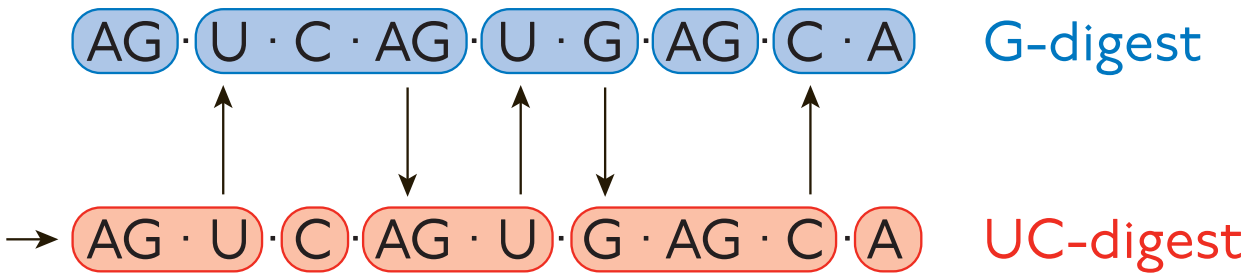
C·G·C·UG and C·UG·C·G.

Both give the same sets of digests. Similarly, the graph on the right has two Eulerian tours

G·C·G·U·G and G·U·G·C·G.

These are the shortest RNA sequences whose digests are ambiguous, in the sense that they produce the same graph as another, distinct sequence.

These slides are available on Blackboard and at https://bit.ly/3glVwmj



The diagram above shows two views of our example sequence, one with G-fragments shown in red rounded boxes and another with the UC-fragments in blue boxes: in both cases, the extended bases are shown separated by dots.

- It's only the larger fragments—those containing two or more extended bases—that contribute to the graph.
- These large fragments overlap in a very particular way: the extended base that marks the end of a large blue fragment also marks the beginning of the next large red one.
- The endpoints of large fragments are the vertices of our graph, while the graph's directed edges, along with their labels, come from the large fragments themselves.

These slides are available on Blackboard and at `https://bit.ly/3glVwmj`

Further Reading

I first learned about this application from a textbook

> Edgar G. Goodaire and Michael M. Paramenter (2006). *Discrete Mathematics with Graph Theory*, 3rd Edition, Prentice Hall. ISBN 0-13-167995-3.

whose authors give lots of related exercises. Early research papers about the Eulerian path approach to RNA sequencing include

> George Hutchinson (1969), Evaluation of Polymer Sequence Fragment Data Using Graph Theory, *Bulletin of Mathematical Biophysics*, **31**:541–562. DOI: 10.1007/BF02476636

> Robert W. Holley *et al.* (1965), Structure of a Ribonucleic Acid, *Science*, **147**:1462–1465. DOI: 10.1126/science.147.3664.1462

while the following article presents a more recent graph-theoretic approach to problems arising in whole-genome sequencing.

> P. E. C. Compeau,*et al.* (2011), How to apply de Bruijn graphs to genome assembly, *Nature Biotechnology*, **29**:987–991. DOI: 10.1038/nbt.2023