Algorithmic Foundations of Data Science

He Sun

h.sun@ed.ac.uk

Room 5.03, Informatics Forum

THE UNIVERSITY of EDINBURGH
informatics

What is Data Science?

- Most people outside Edinburgh call *a similar subject* **computer science**.
- People around the world call *some fancier subject* **data science**.

BUT, if we follow the same rule,

- we should have called astronomy **telescope science**.
- we should have called biology **microscope science**.

The use of Natural Philosophy in history

From the ancient world to the 19th century, the term “natural philosophy” was the common term used to describe the practice of studying nature. It was in the 19th century that the concept of “science” received its modern shape with new titles emerging such as “biology” and “biologist”, “physics” and “physicist” among other technical fields and titles; …… Issac Newton’s book *Philosophiae Naturalis Principia Mathematica* (1687), whose title translates to "Mathematical Principles of Natural Philosophy", reflects the then-current use of the words "natural philosophy", akin to "systematic study of nature".

--- Wikipedia

My understanding about Data Science

- Data Science is at its rather early stage, the stage in which we couldn't even find a way to name it without using the word "Science".

- Looking back to history, the early stage of a research field is usually the moment when big ideas and breakthrough have occurred.

- Therefore, you should attend AFDS if you're willing to be a great scientist.

What should you attend AFDS? (cont.)

- Several topics covered in AFDS already have wide practical applications.

- Some more frontier topics covered in AFDS are excellent mathematical training for you, and could have big industrial impacts for you in five to ten years.

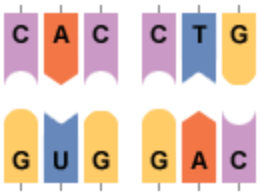
- *A good university education is not only to help you find a job next year, but also to ensure that you'll be highly-qualified in a decade.*

What will be covered in the course?

Here are some sample problems.

Problem 1: Streaming Algorithms

Background: big data in 2025



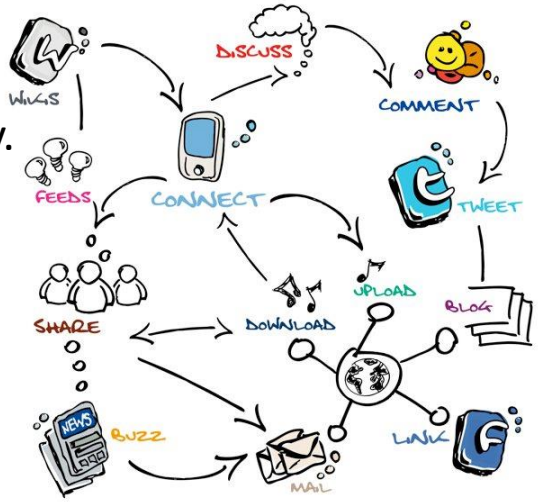
Genome sequences for many species are available: each megabytes to gigabytes in size.

There are billions of global mobile phone users.

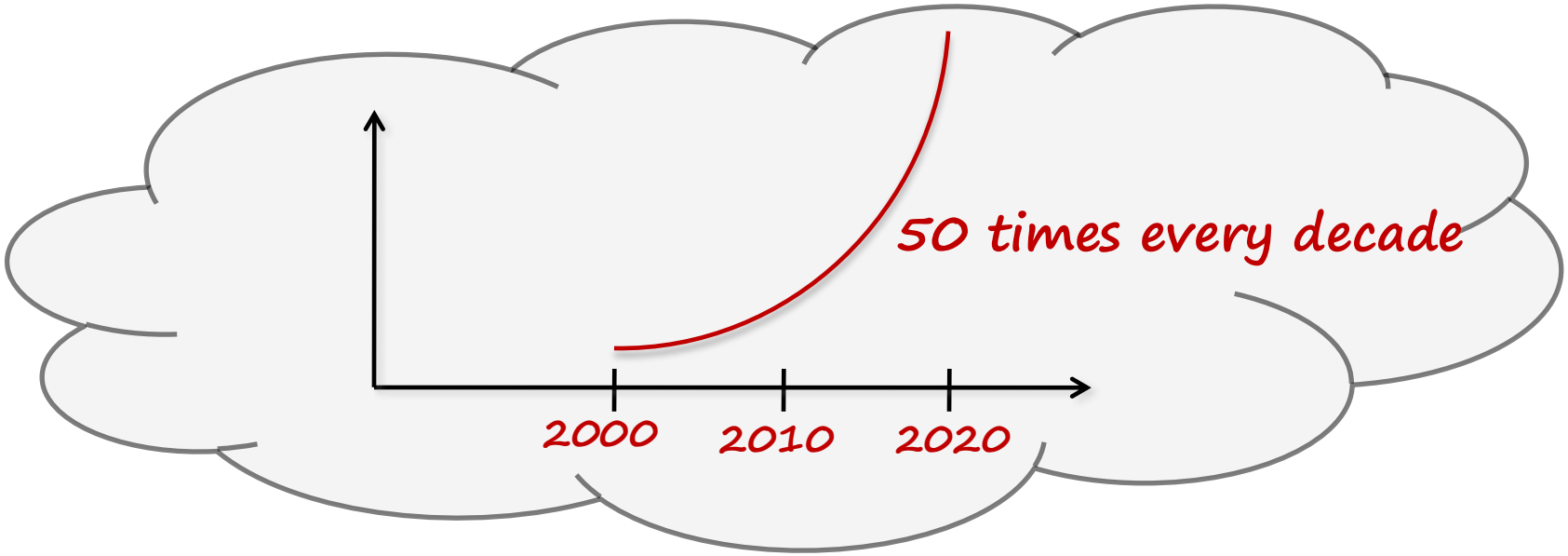
500 hours of videos uploaded per minute

Twitter produces over 500 million tweets per day.

There are about 3.07 billion monthly active users in Facebook.

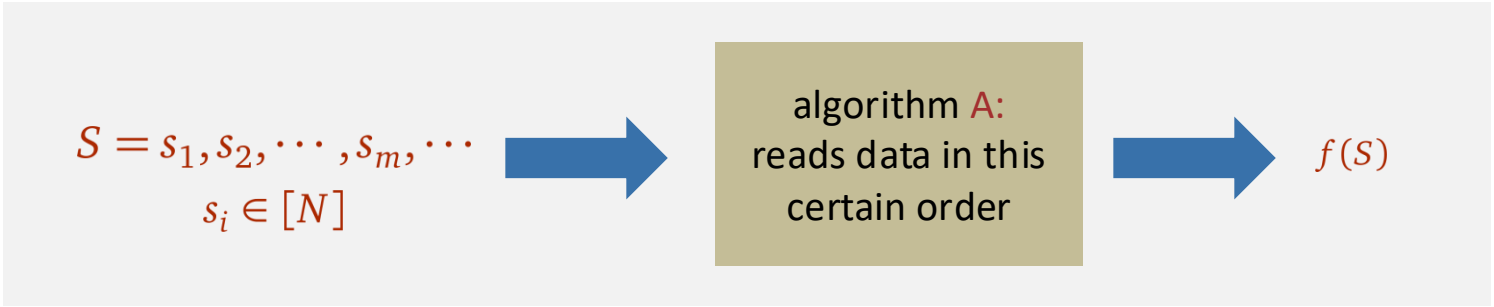


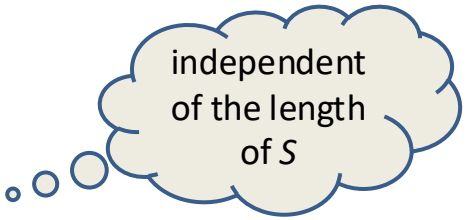
Background: big data in 21st century



- Storing the entire data is usually not possible anymore. Instead, we can only afford to store ***partial information*** of the input dataset.
- Good approximate solutions are usually sufficient for most practical applications!

Data streaming algorithms: model





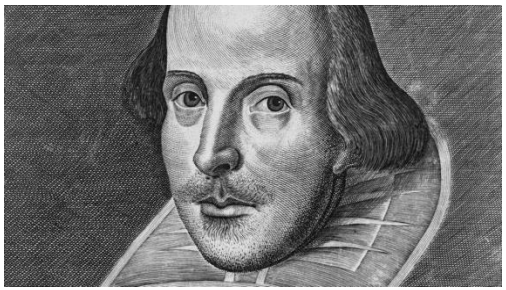
Space: Total space of algorithm *A* is $O(\text{poly}\log N)$.

Quick update: Processing time of every arriving item should be fast.

Image that the input is the IP addresses visiting UoE website.

Approximate guarantee: With probability 99%, the algorithm's output should be very close to the right answer, e.g., the error is within $\pm 10\%$.

Flajolet's Loglog counting algorithm



The complete works of Shakespear have 28,239 distinct words.

Philippe Flajolet
(1948 -2011)

```
ghfffghfghgghgggggghghheehfhfhhgghghghhfgffffhhhiigfhhffgfiihfhhh
igigighfgihfffghigihghigfhhgeegeghgghhhgghhfhidiigihighihehhhfgg
hfgighigffghdieghhhgggghhfghhfiiheffghghihifgggffihgihfggighgiiif
fjgfgjhhjiifhjgehgghfhhfhjhiggghghihigghhihihgiighgfhlgjfgjjjmfl
```

FIGURE 1. The LOGLOG Algorithm with $m = 256$ condenses the whole of Shakespeare's works to a table of 256 "small bytes" of 4 bits each. The estimate of the number of distinct words in this run is $n^\circ = 30897$ (the true answer is $n = 28239$), which represents a relative error of $+9.4\%$.

Link to an article about P. Flajolet:
https://rjlipton.wordpress.com/2011/03/27/philippe-flajolet-1948-2011/

Topics covered for streaming algorithms

- **Two central techniques for designing streaming algorithms: Sampling, and Sketching.**

- **Streaming algorithms for computing certain statistical information:**
  - The number of distinct elements
  - The number of occurrence of frequently occurring items

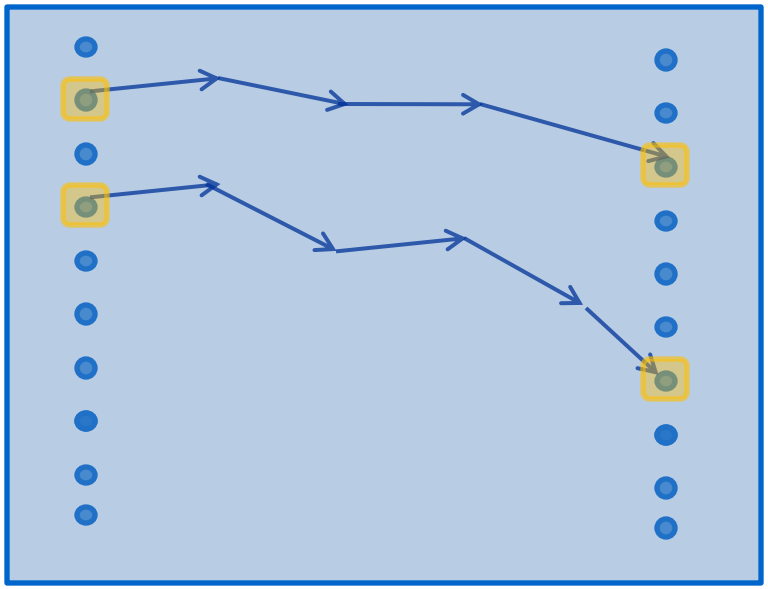
Problem 2: Expander & Super concentrators

Super concentrators

Given n cities in the north and n cities in the south, construct a highway network, such that for any k cities in the north and k cities in the south, there are k vertex disjoint paths.

"disjoint" => efficiency of transportation, no delay

# of edges ⇔ construction cost



Construct a network (directed graph) with n input nodes and n output nodes, such that for any K input nodes, and any K output nodes, there are K disjoint paths connecting them.

A complete bipartite graph is an example, but too "expensive".

For any n, there is a super concentrator with $28n$ edges.

Super concentrators (cont.)

" Finally, the super concentrators constructed by Valiant in the context of computational complexity established the fundamental role of expander graphs in computation. "

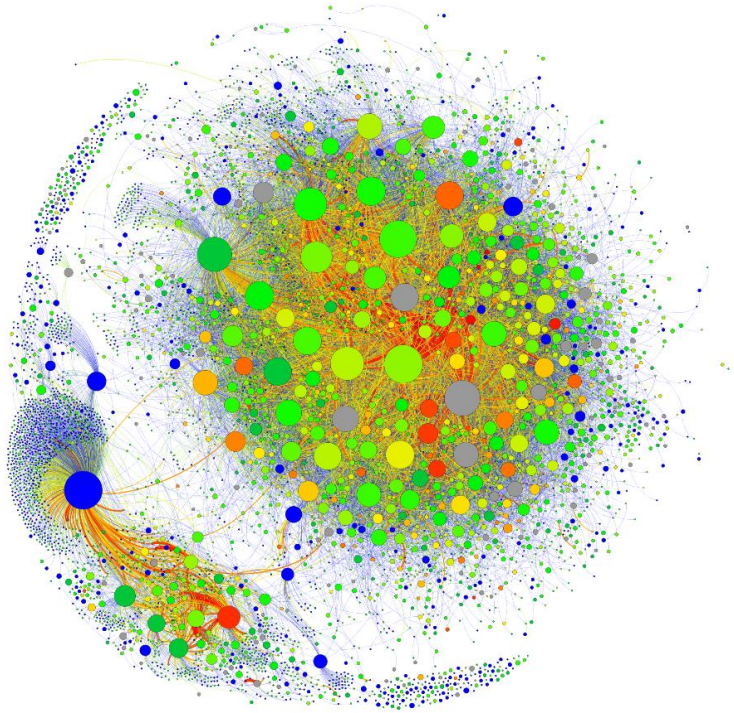
2010 ACM Turing Award Citation

Most of Valiant's work on super concentrators was done when he was at UoE.

In this course we will discuss various aspects of expander graphs, and their applications in Data Science.

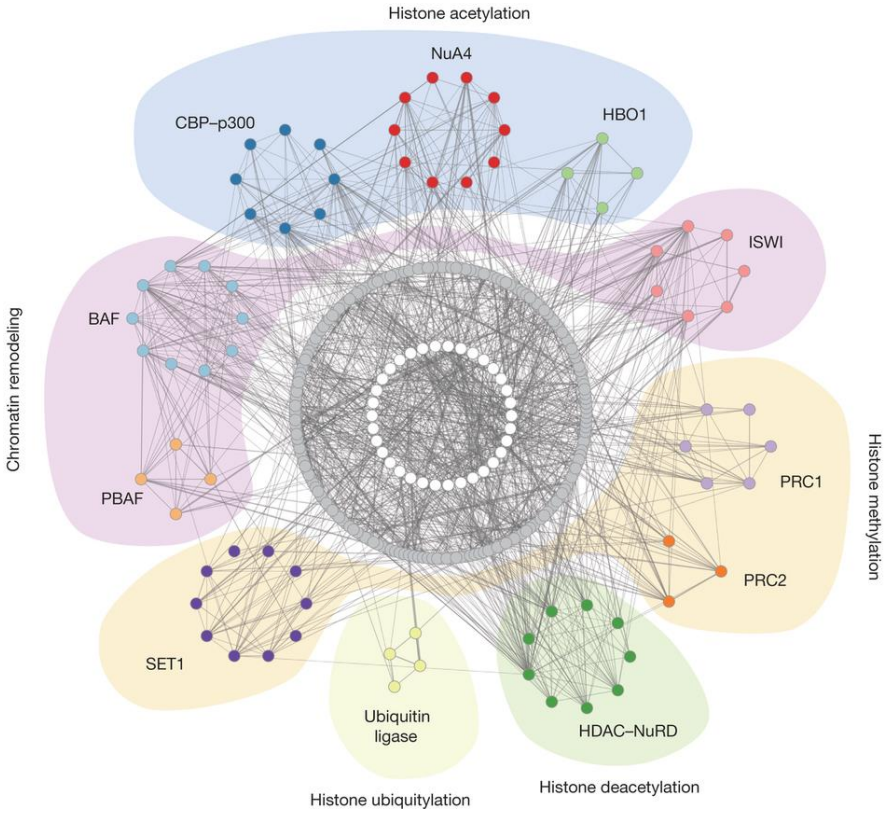
Problem 3: Graph Clustering

Applications of clustering



A follower and retweet network within 72 hours

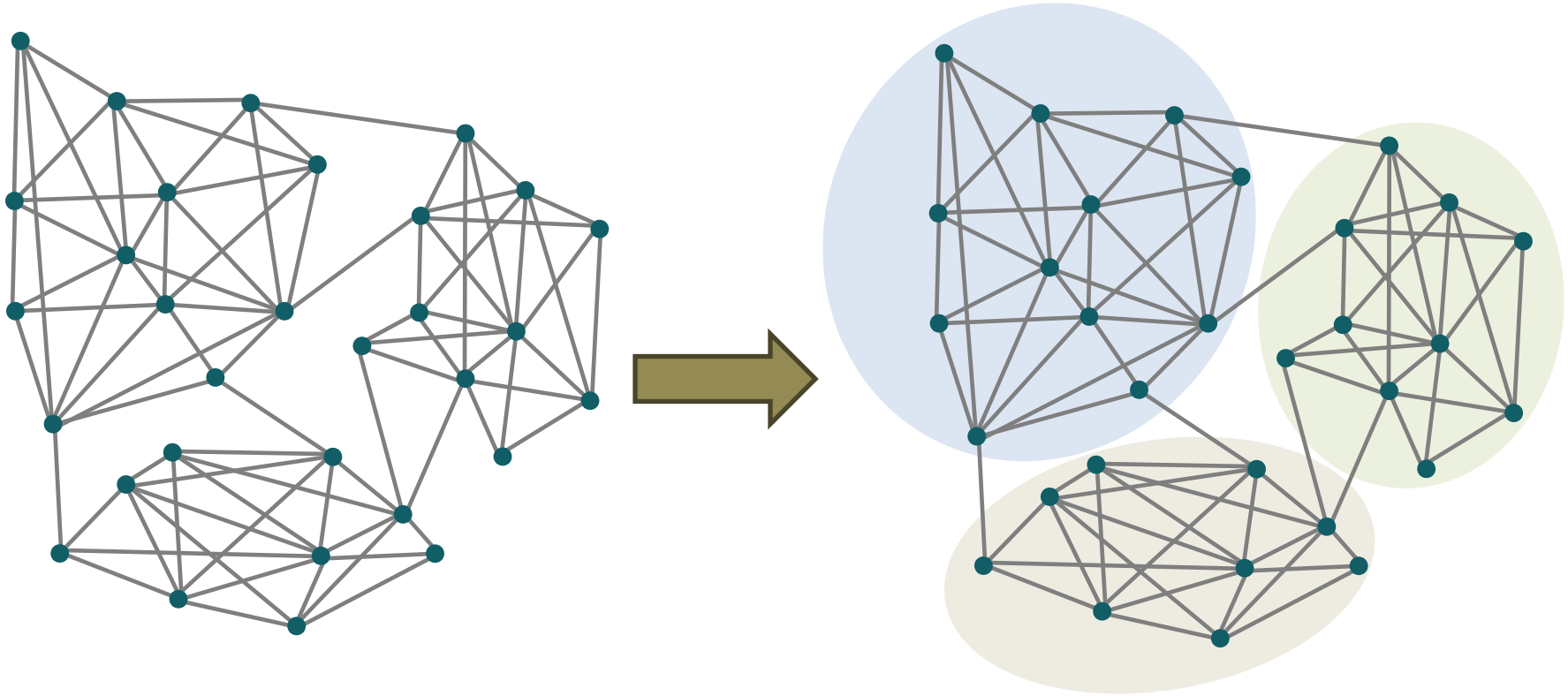
Segmentation in Computer Vision



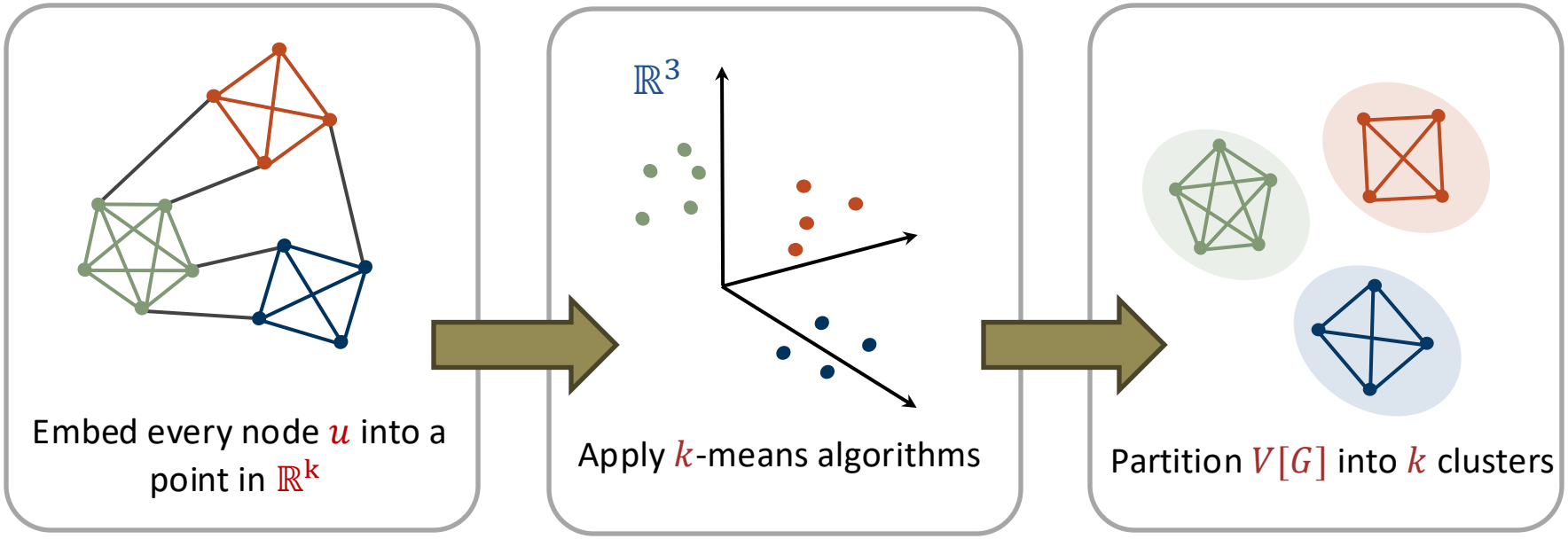
A protein–protein interaction network for the 425 human chromatin factors screened [Huang et al., 2013].

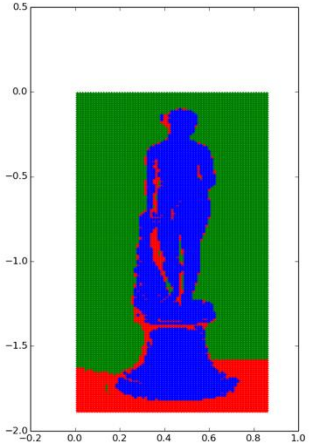
Graph clustering

Partition a graph into different clusters.



Spectral clustering





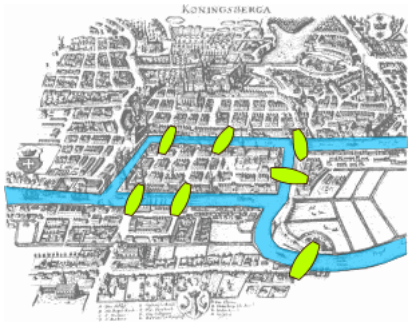
We will discuss how similar algorithms can be analysed.

A crucial theme around our discussion:

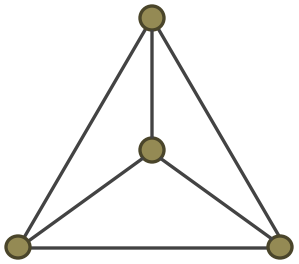
Relations among graphs, matrices, and geometric objects

Graphs, matrices, and geometric objects

Leonhard Euler
(1707-1783)



Seven Bridges of Königsberg
1736

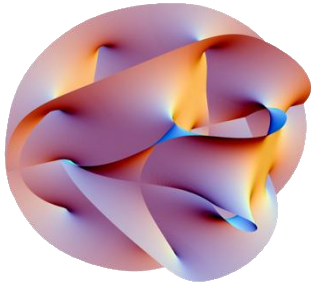


- Coloring
- matching
- Hamiltonian Cycles
- Spanning Trees

Since 1700s

$$L = \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix}$$

- Eigenvalues and Eigenvectors
- Rank, and kernel



- The isoperimetric problem
- Sobolev inequalities
- Heat equations

Construct a graph
≈ Construct a matrix
≈ Construct a geometric object