# Non-euclidian data and Graphs

Raoul Grouls, 18 juni 2024

# What is a vectorspace?

Let $V$ be a set, let $F$ be a field equipped with addition and multiplication

We define binary operations “+” on $V$, denoted $V \times V \rightarrow V$ , and “.” on $F \times V$ denoted $F \times V \rightarrow V$

A **vectorspace** satisfies $\forall c, d \in F, \forall u, v, w \in V$

Closure under addition: $u + v \in V$

Closure under multiplication: $c \cdot v \in V$

Addition (+):

1. Commutative: $u + v = v + u$
2. Associative: $(u + v) + w = u + (v + w)$
3. Identity: $u + 0 = 0 + u = u$
4. Inverse: There exists an element (-1) such that: $u + (-1)u = 0$

Multiplication (.):

1. Compatibility: $(cd)u = c(du)$
2. Distributivity: $c(u + v) = cu + cv$
3. Distributivity: $(c + d)u = cu + du$
4. Identity: $1 \cdot u = u$

# What is a metric?

For $\forall x, y, z$ :

1. Non-negativity: $d(x, y) \leq 0$
2. Identity of indiscernibles: e$d(x, y) = 0$ if and only if $x = y$.
3. Symmetry: $d(x, y) = d(y, x)$
4. Triangle inequality: d(x, y) + d(y, z) ≥ d(x, z)

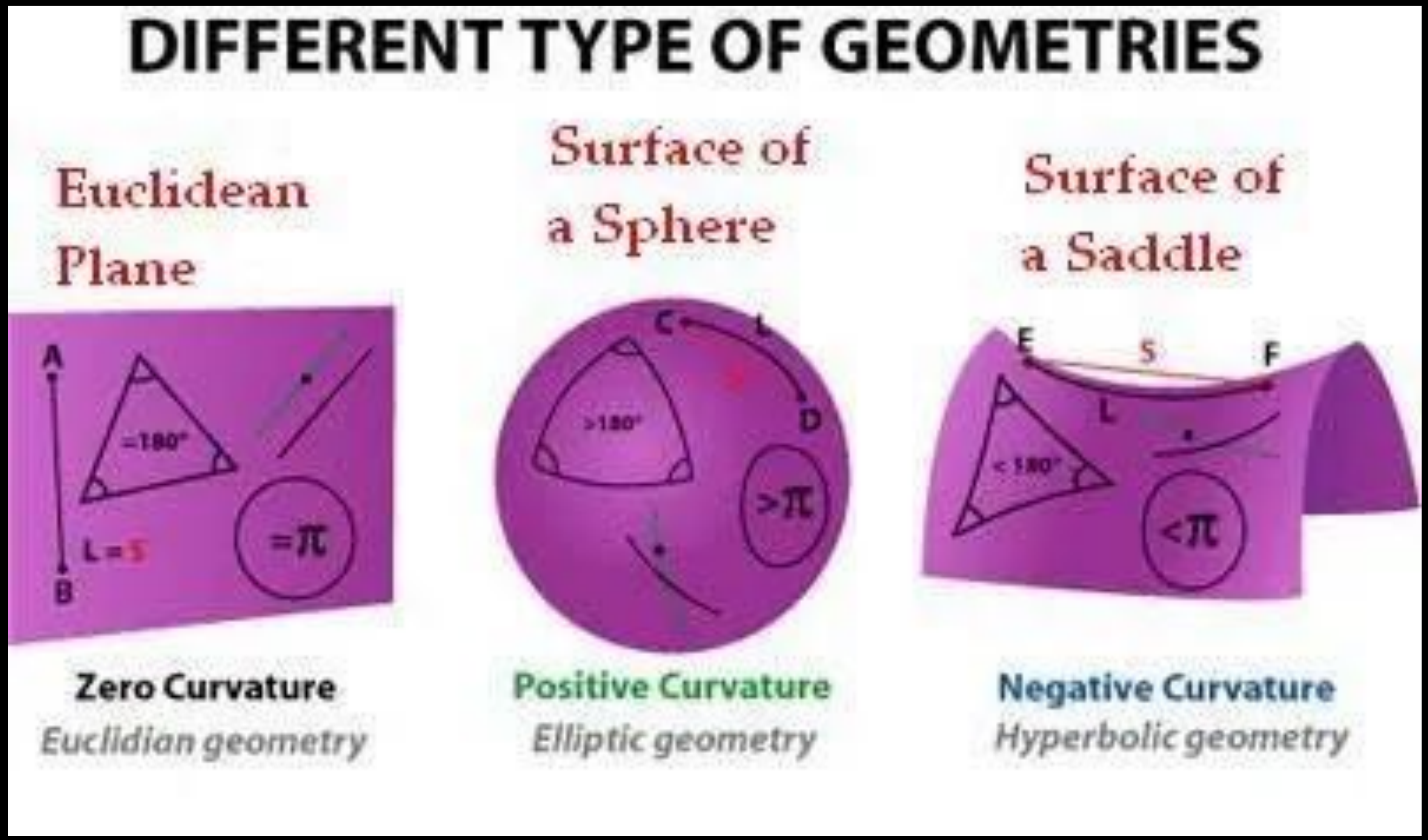
# What is Euclidian geometry?

Euclidian data follows the axioms of euclidian geometry

1. A straight line may be drawn between any two points.
2. Any terminated straight line may be extended indefinitely.
3. A circle may be drawn with any given point as center and any given radius.
4. All right angles are equal.
5. For any given point not on a given line, there is exactly one line through the point that does not meet the given line

# What is Non-Euclidian geometry?

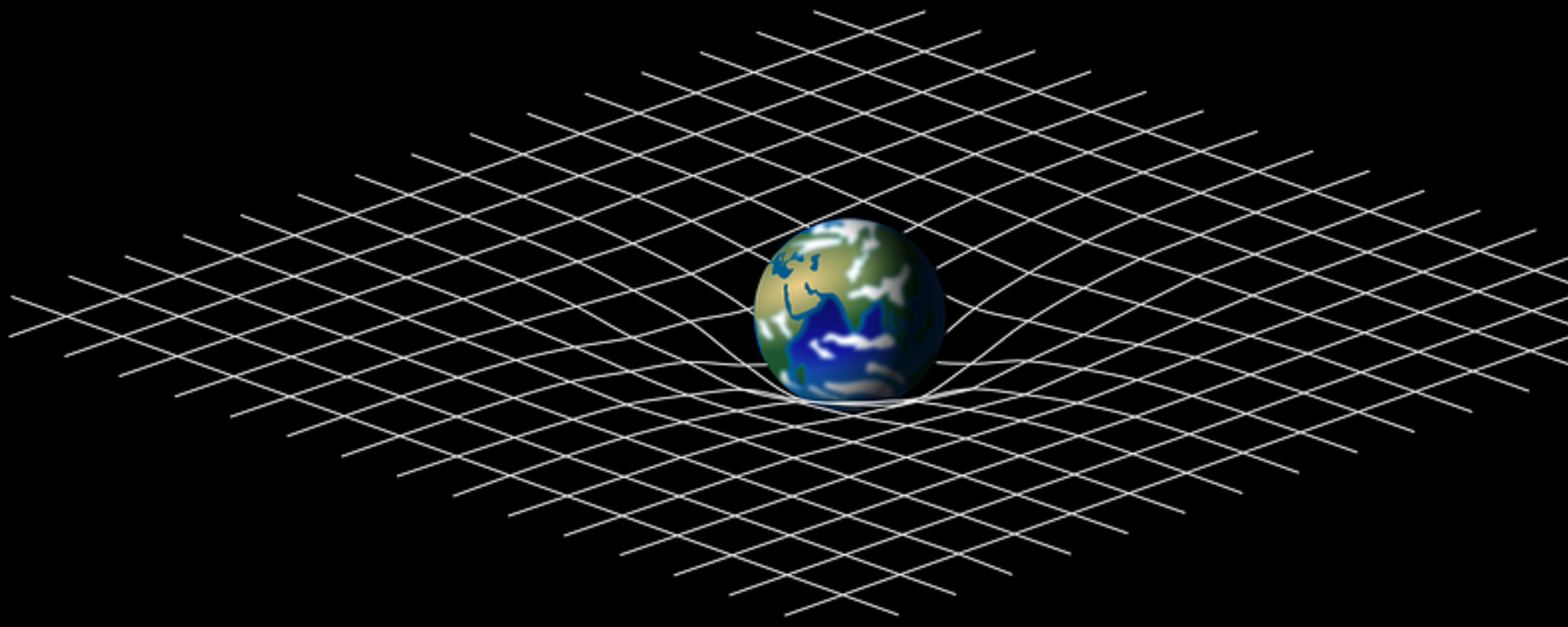
## Rejection of the parallel postulate



First attempts at challenging the parallel postulate are by Ibn al-Haytham in the 11th century.

Early 19th century, the parallel postulate was rejected as “apriori true”

- 1823/1832, Bolyai’s (Hungarian) father writes in 1820 ““You must not attempt this approach to parallels. […] I have traversed this bottomless night, which extinguished all light and joy in my life”, but in 1823 Bolyai writes back “I have created a strange new universe”. It is published in 1932 by his father.
- 1829 “*A Concise Outline of the Foundations of Geometry*” by Lobachevsky (Russian)
- 1848 Bolyai learns that Lobachevsky has published a similar piece. Their work is the basis for “hyperbolic geometry”
- 1905 Poincare describes his disk model of hyperbolic space and suggests that space might be hyperbolic.
- 1915 Einstein publishes “The field equations of gravitation”, describing space as non-euclidian.

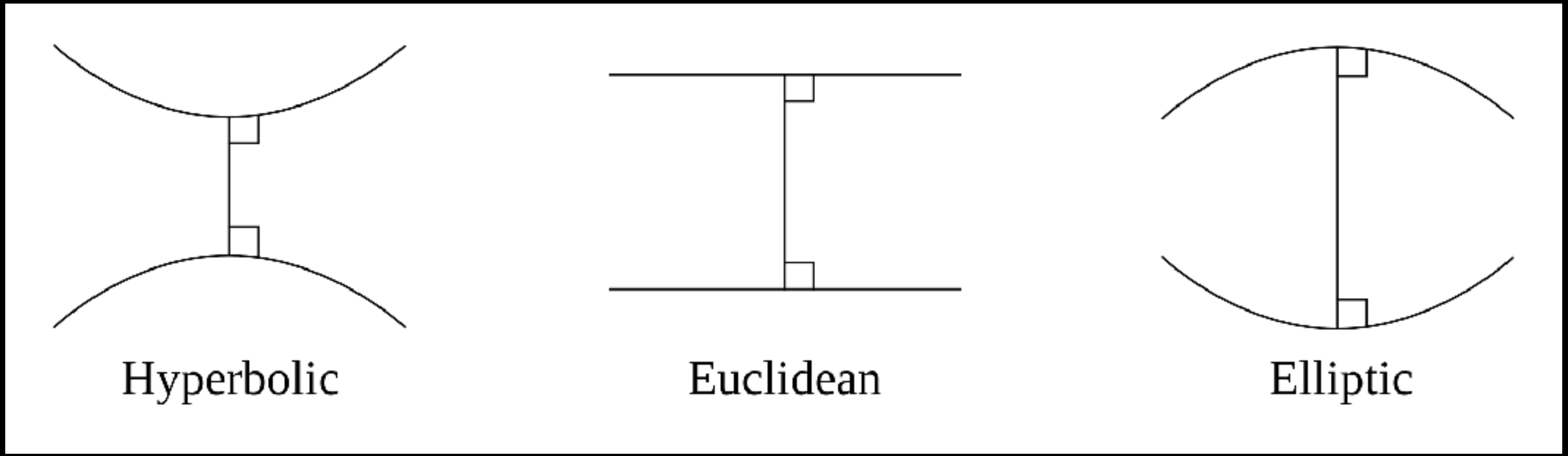


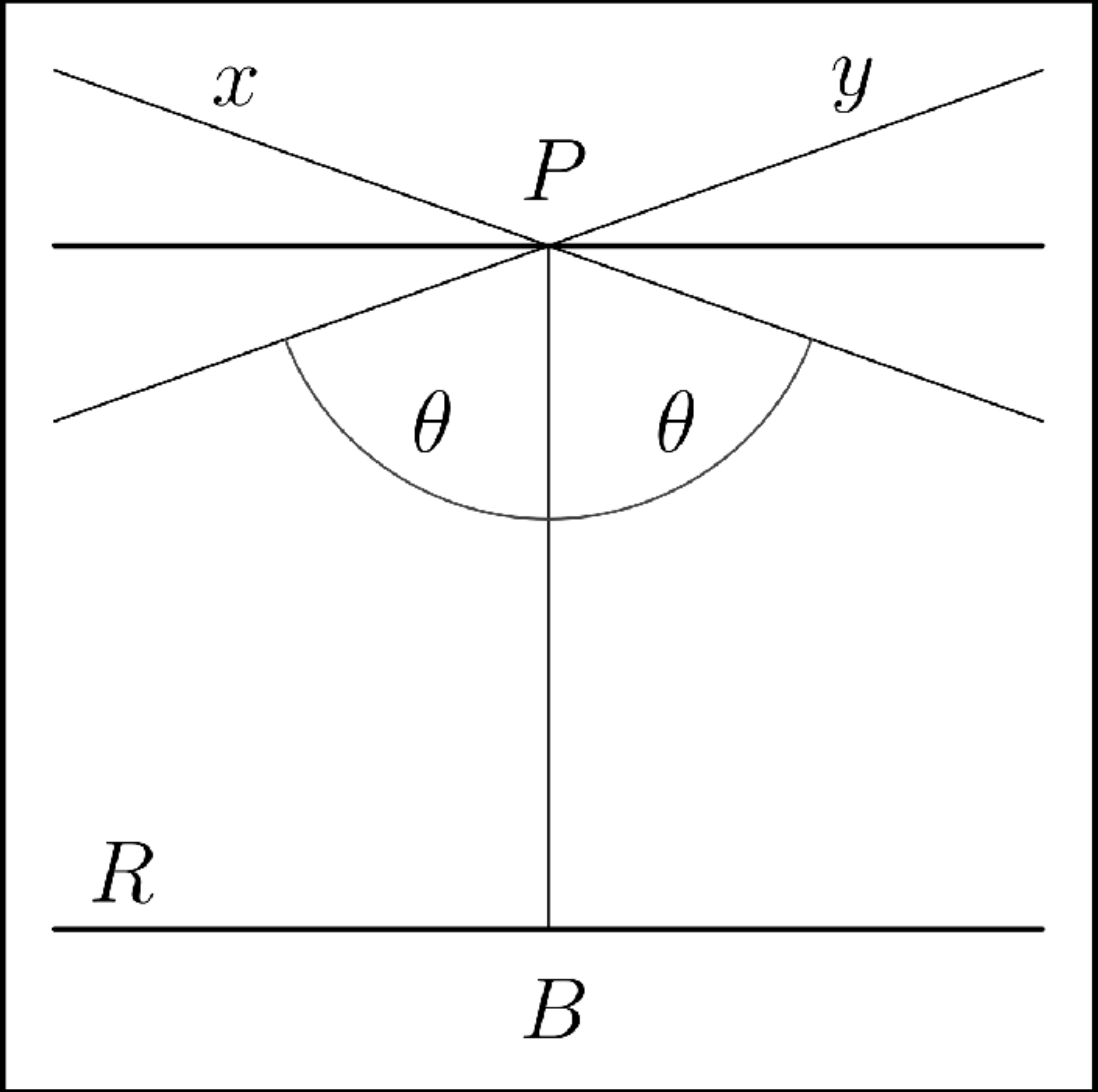
# What is Non-Euclidian geometry?

## Rejection of the parallel postulate

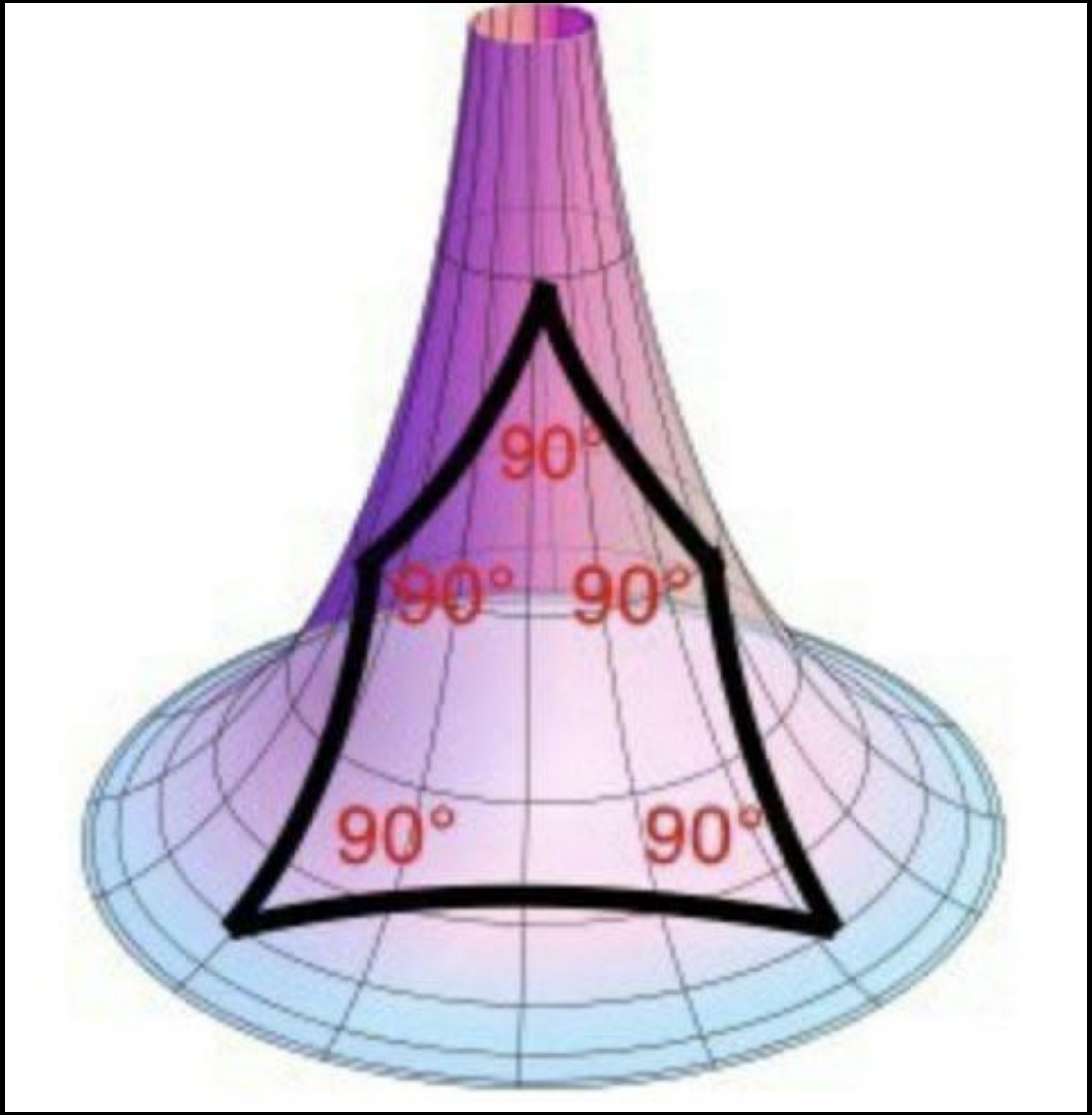
Bolyai-Lobachevskian geometry:

For any given line *R* and point *P* not on *R*, in the plane containing both line *R* and point *P* there are at least **two distinct lines** through *P* that do not intersect *R*.





## What is Non-Euclidian geometry?



Among other things, this gives us five sides squares.

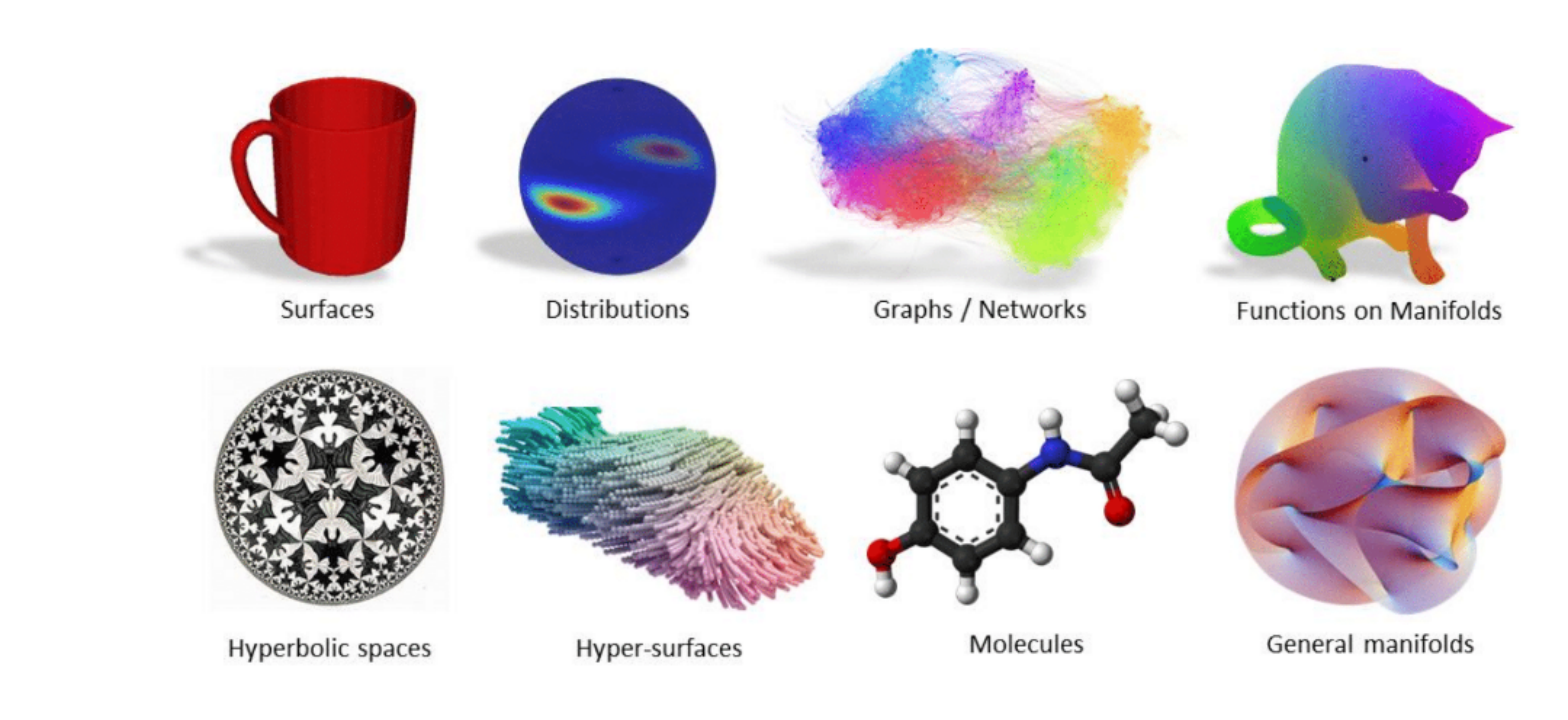
Actually, our space *is* hyperbolic due to the gravity of the Sun.

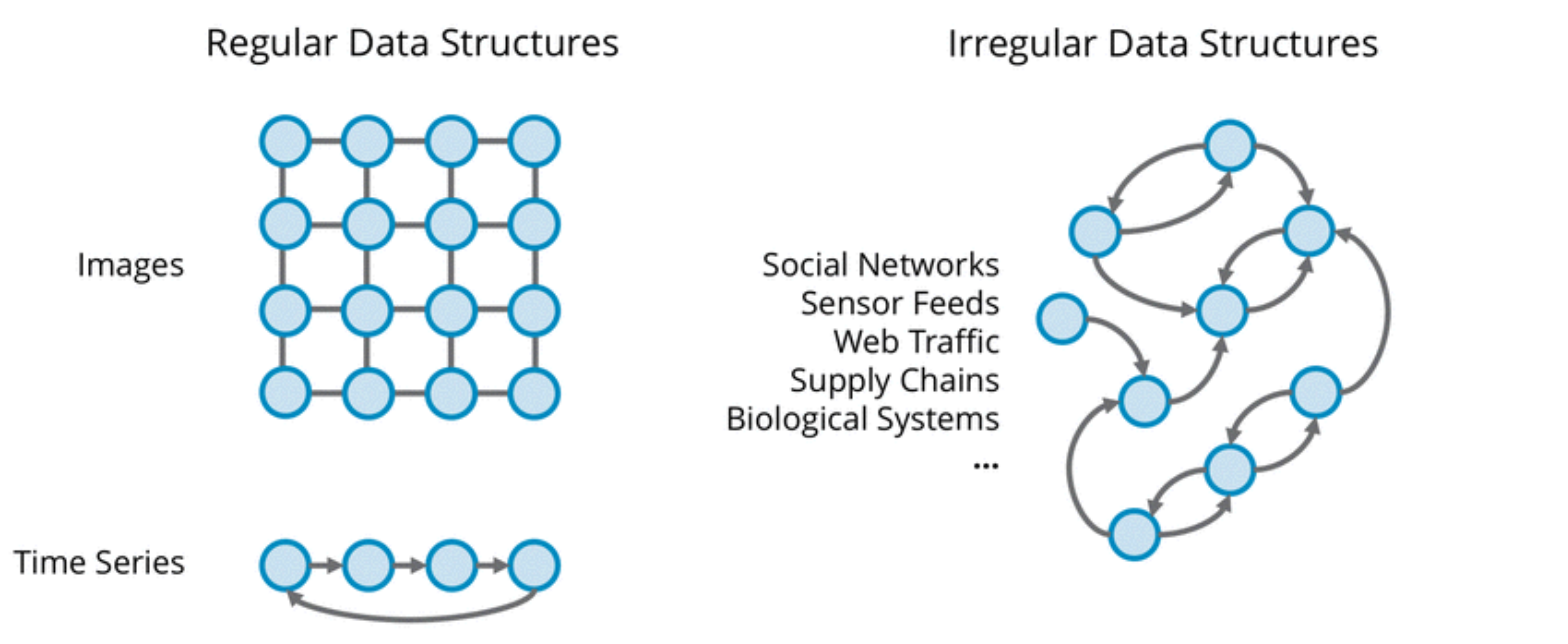
This is a very nice explanation https://youtu.be/n7GYYerlQWs

# What is Non-Euclidian data?

- Vectorspaces like $\mathbb{R}^d$ have a metric (distance measure).
- But not all data follows these principles. For example:
    - There are hyperbolic vectorspaces where the parallel postulate does not hold. There is still a metric (and a vectorspace).
    - Some data doesnt even has a good notion of distance, or is irregular (eg with holes). These are no longer vectorspaces, but topologies.

# What is Non—Euclidian data?





Regular Data Structures

Images

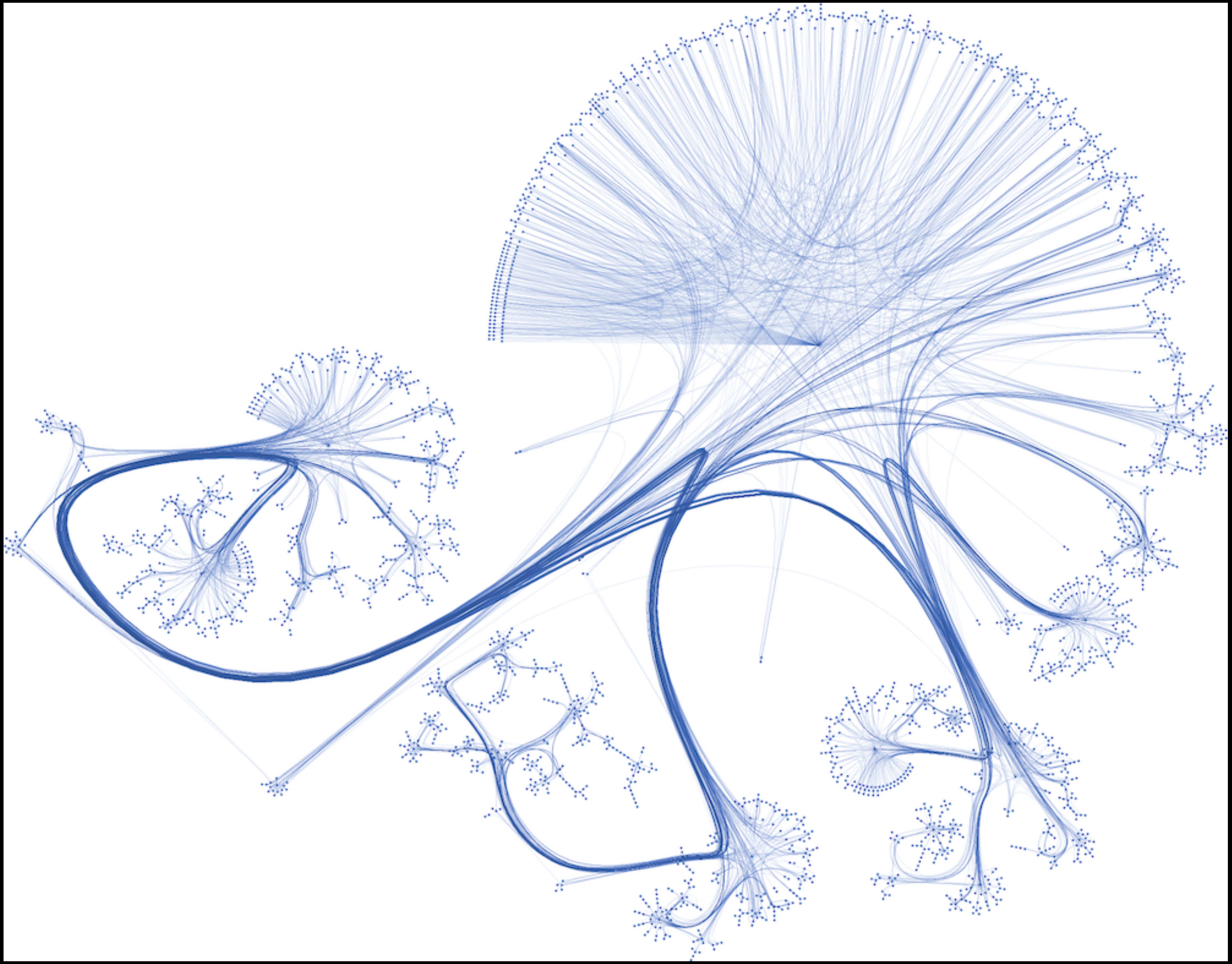
Time Series

Irregular Data Structures

Social Networks
Sensor Feeds
Web Traffic
Supply Chains
Biological Systems
...

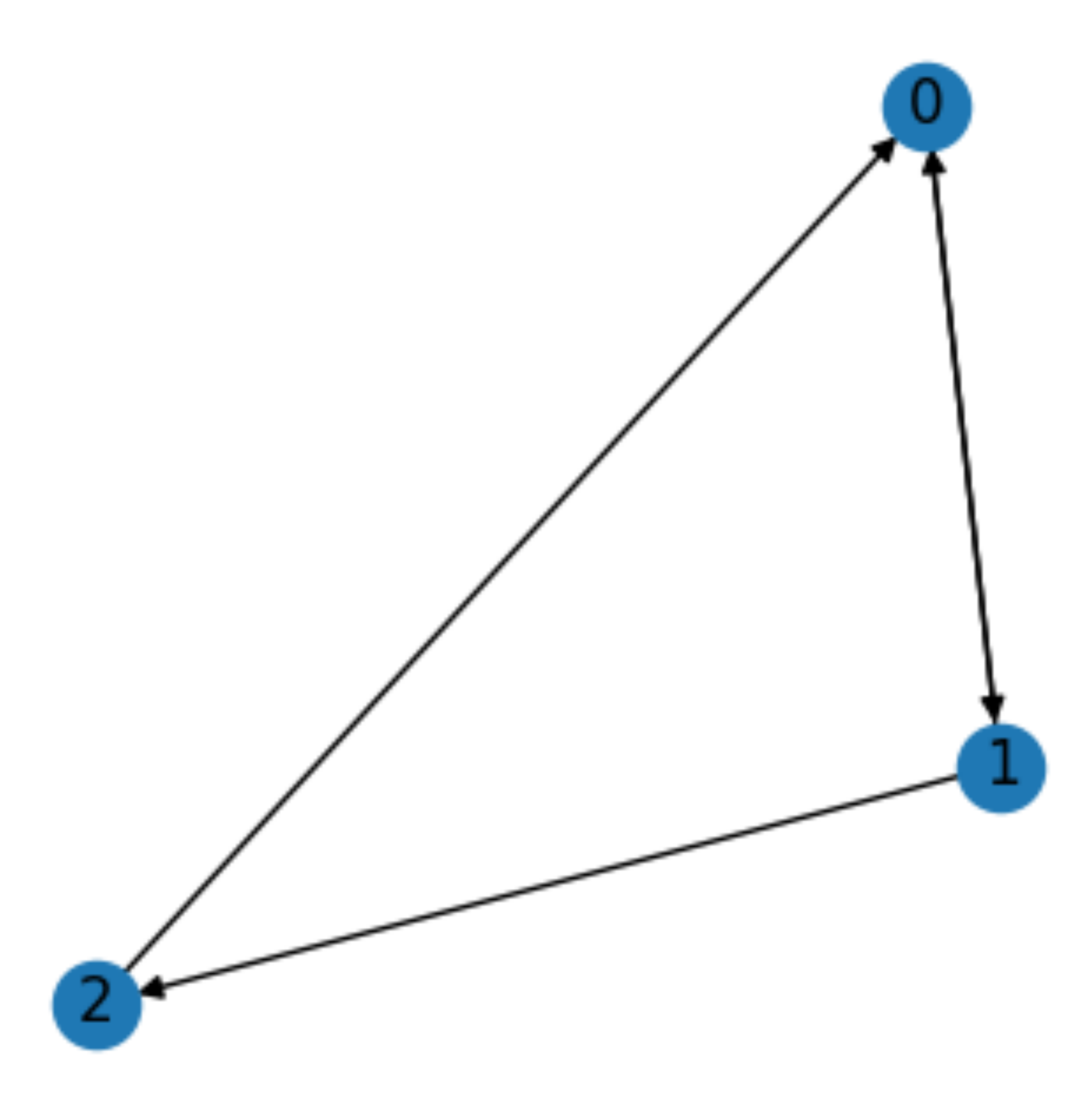
# Cora dataset

- 2708 scientific publications
- classified into one of seven classes.
- The citation network consists of 5429 links.
- Each publication in the dataset is described by a 0/1-valued word vector indicating the absence/presence of the corresponding word from the dictionary. The dictionary consists of 1433 unique words



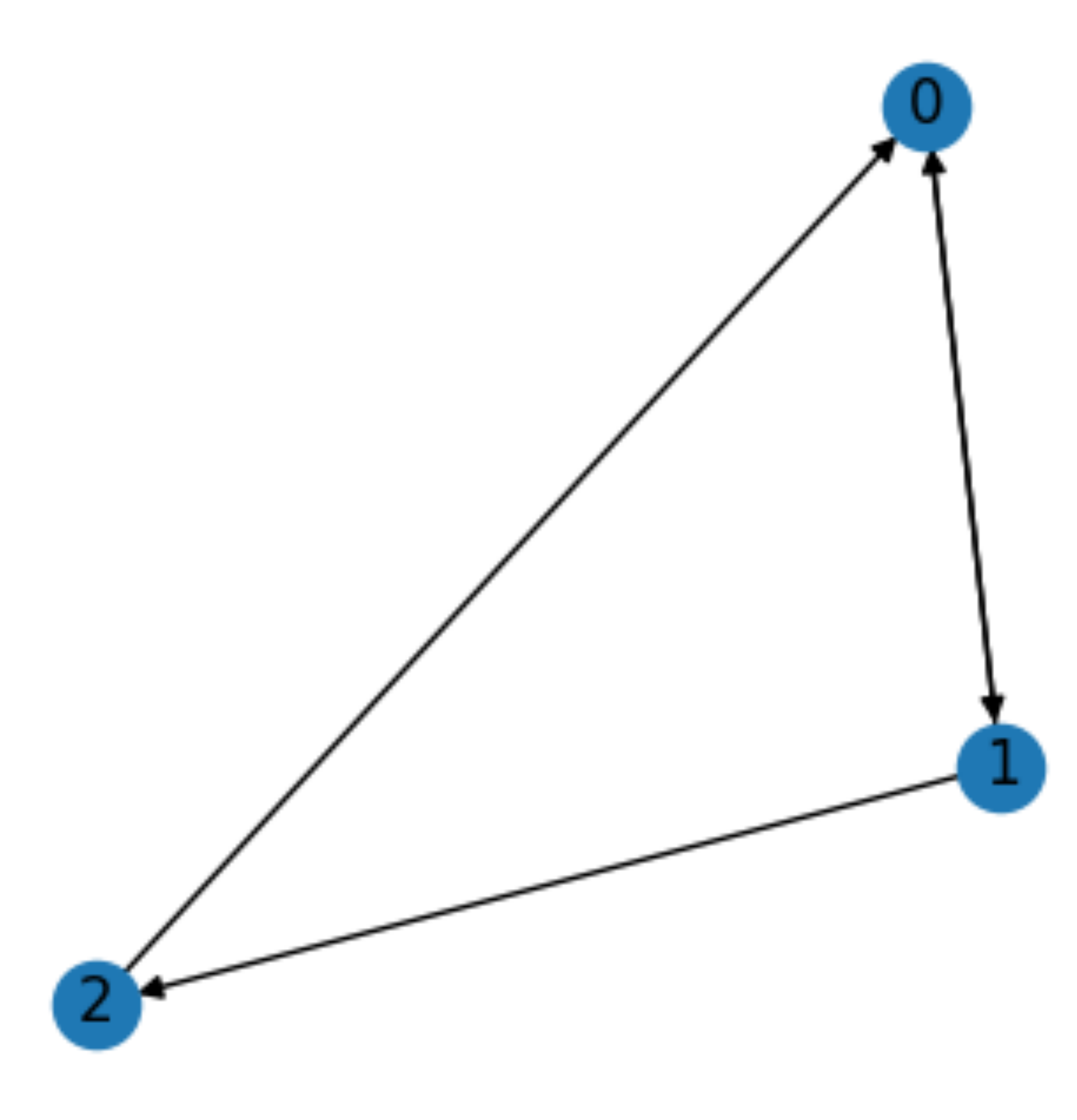
# What is a Graph?

- A Graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ is defined by:
    - A set of nodes $\mathcal{V} = \{v_1, \ldots, v_n\}$
    - A set of edges between nodes $\mathcal{E} = \{(v_i, v_j) \mid v_i, v_j \in \mathcal{V}\}$



# What is a Graph?

- The adjecency matrix $A$ has a 1 on every position where there is and edge: $A[i,j] = 1$ if $e_{i,j} \in \mathscr{E}$
- Excercise: let’s draw A for this graph!



# Some statistics on graphs

- Degree: the number of edges for a node $d_u = \Sigma_{v \in V} A[u, v]$
- The degree matrix $D$ has on each diagonal element the degree of the note: $D[i, i] = d_i$

# Some statistics on graphs

The Laplacian matrix is defined as

$$L = D - A$$

Among other things, it can tell you if there are groups of nodes who are all connected to each other but not much to others in the network.

It's like spotting cliques in your group

# Some statistics on graphs

- Betweenness centrality is the sum of the fraction of all shortest paths through v:

$$cb(v) = \Sigma_{s,t \in V} \frac{\sigma(s,t \mid v)}{\sigma(s,t)}$$

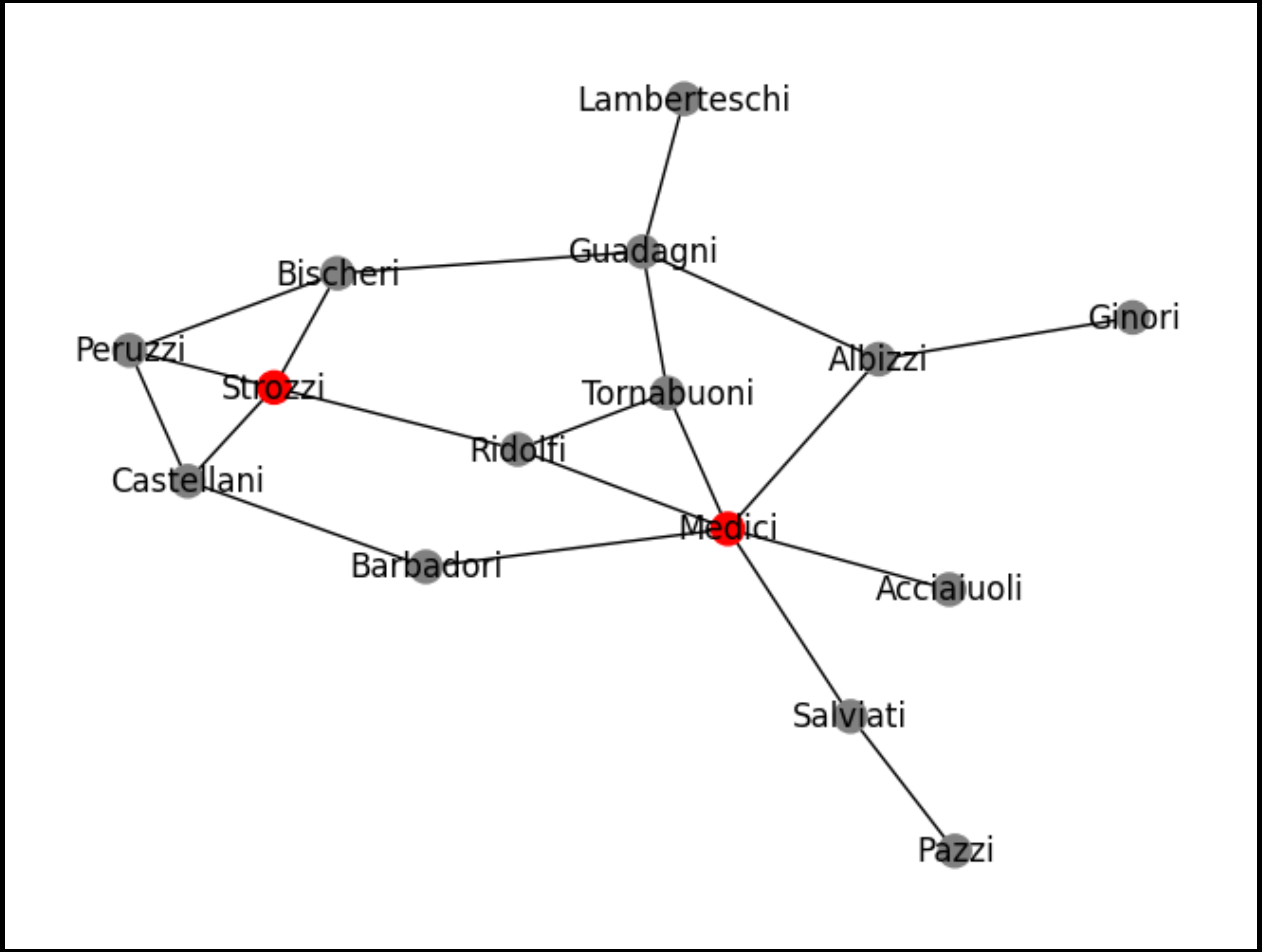
- with $\sigma(s,t)$ the number of shortest (s,t) paths and $\sigma(s,t \mid v)$ the paths through v.

# Florentine Families

Renaissance Florentine families around 1430, collected by John Padgett from historical documents.

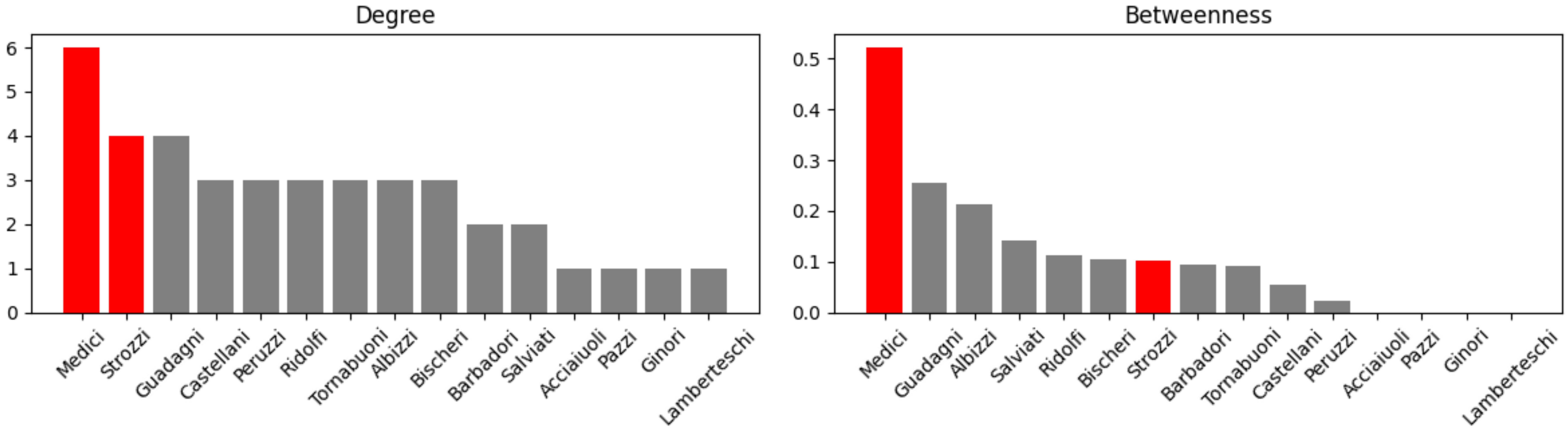
The graph shows marriage alliances.

The data include families who were locked in a struggle for political control. Two factions were dominant in this struggle: one revolved around the Medicis, the other around the Strozzis.



# Florentine Families

The Strozzi famililiy has a high degree but much lower betweenness centrality.
Betweenness is much more pronounced for Medici



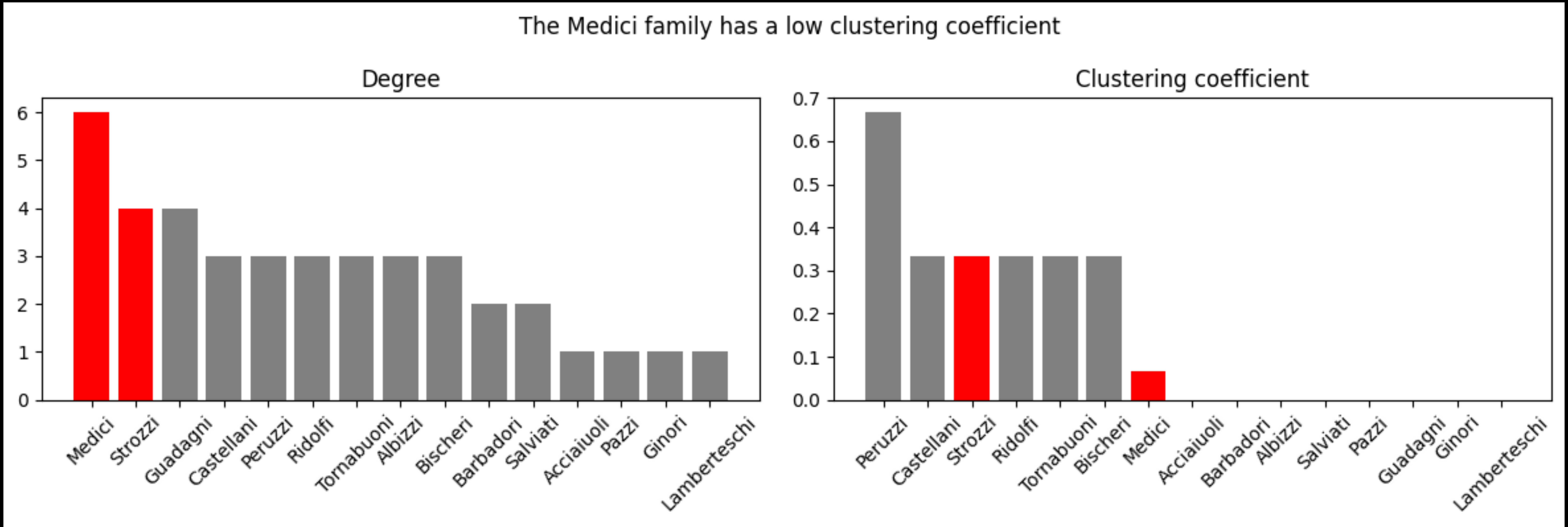
Degree

Betweenness

# Some statistics on graphs

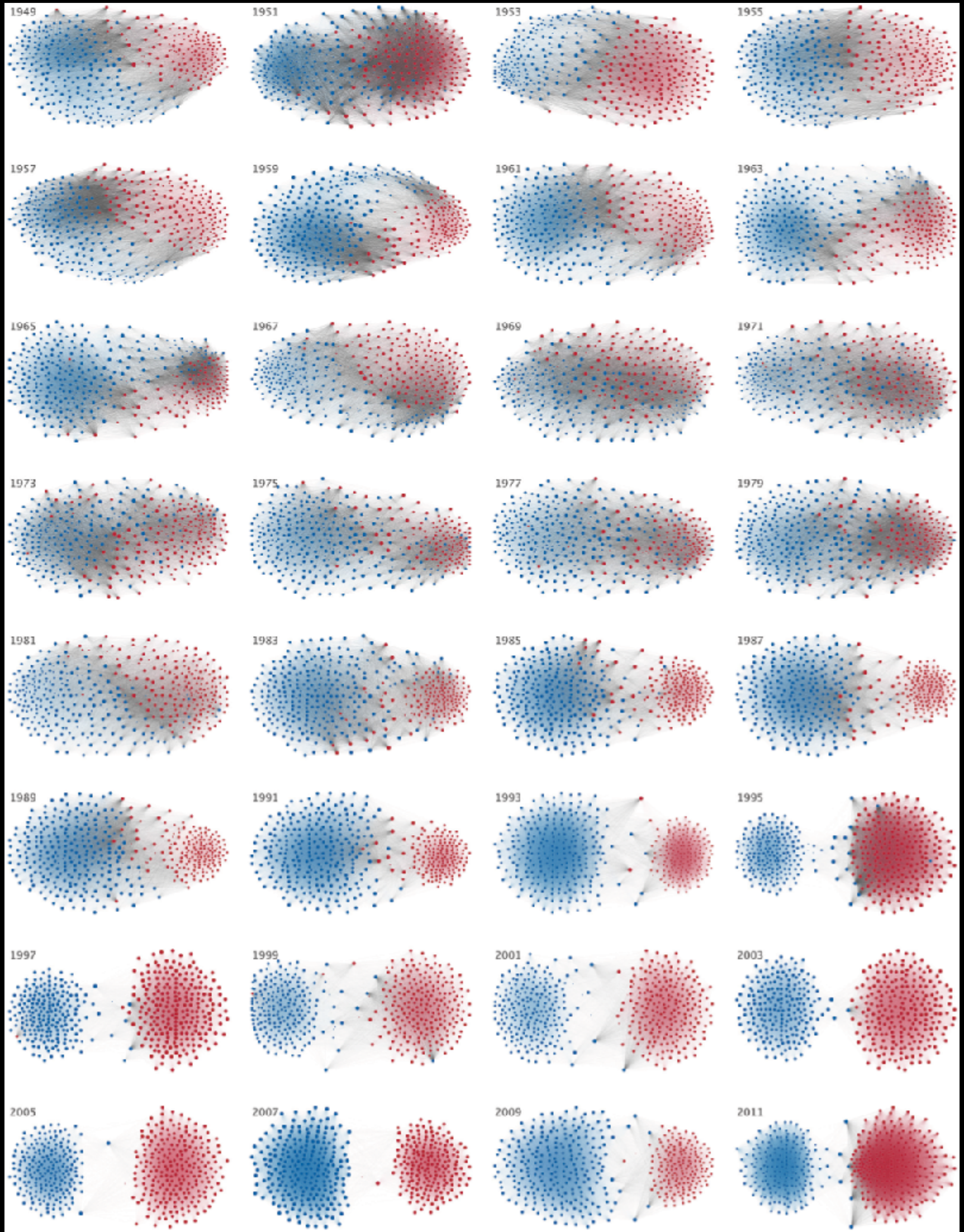
- Triangles: if your friend are also friends
- Clustering coefficient: the fraction of possible triangles.
- This can be a very relevant metric: e.g. there is a correlation between (lack of) triades and depression!

# Florentine Families



# See US Congress polarize over the past 60 years

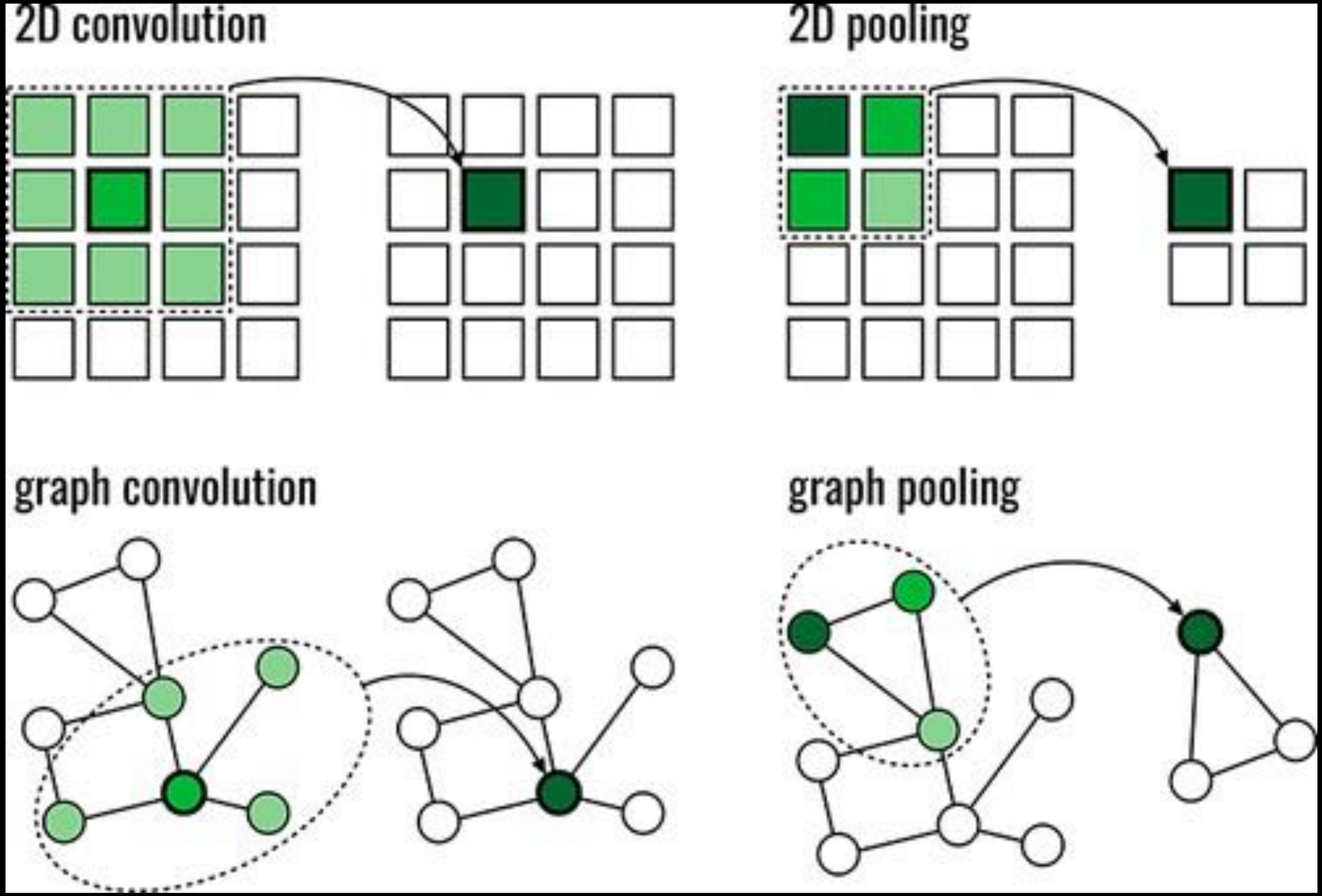
See how likely the House of Representatives' Democrats (in blue) and Republicans (in red) are to vote with their own party, or to cross party lines.



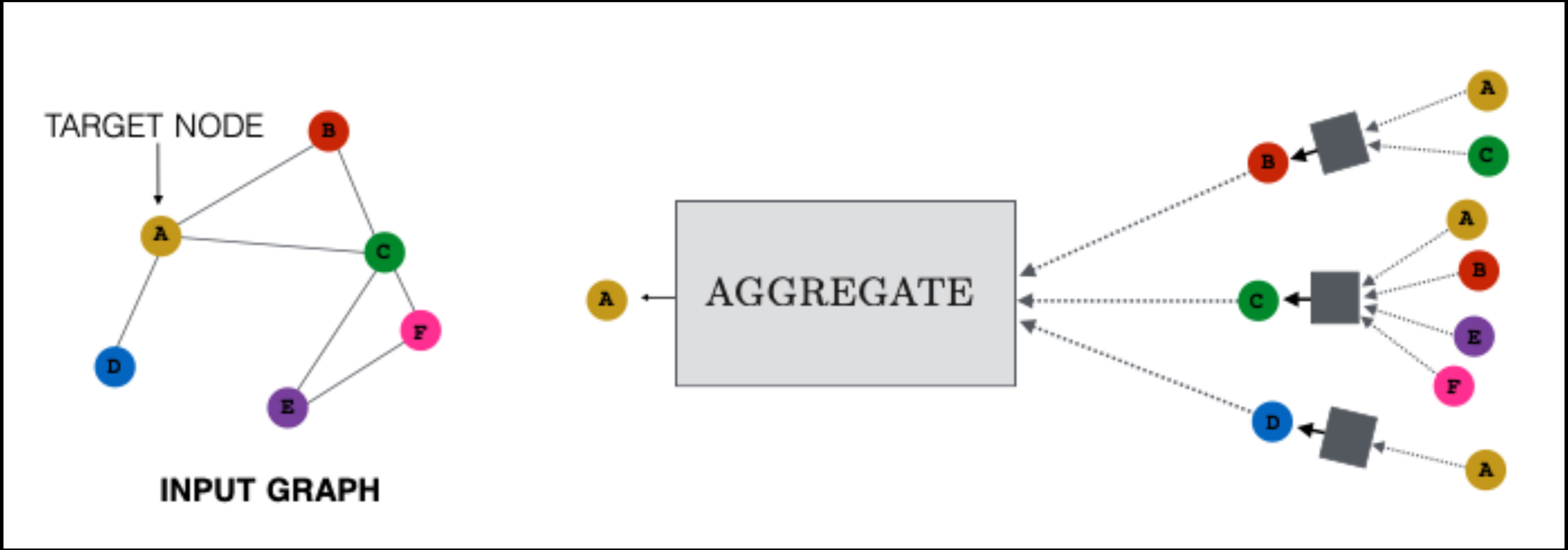
# Issues with machine learning on graphs

- Size and shape: graphs can have wildly different shapes. Unlike images, we can just resize.
- Isomorphism: We can make multiple adjecency matrices for the same graph, by reordering the nodes. However, they represent the same graph. We dont want that to matter.

# Solution: Graph convolutions



# Graph convolutions



# Graph convolutions

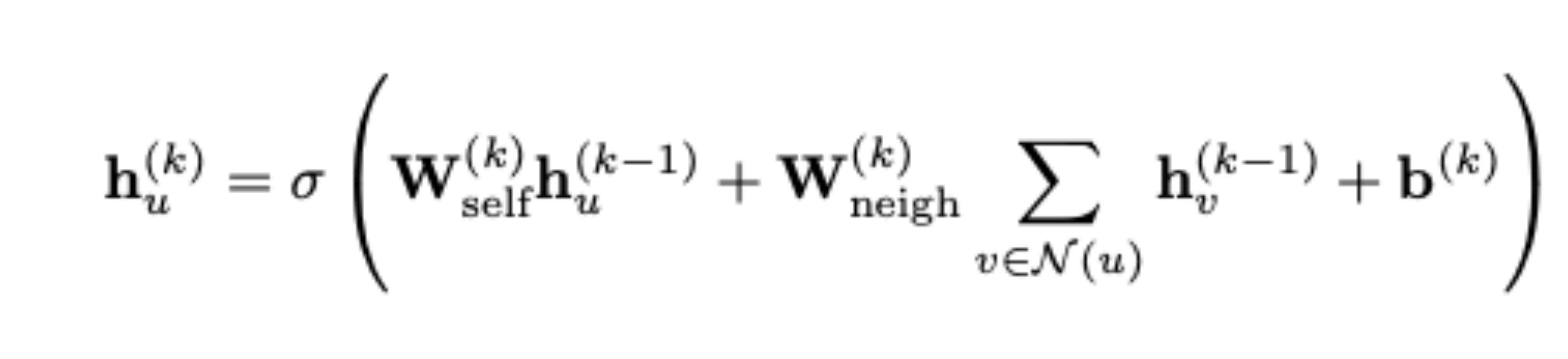
- At each iteration, every node aggregates information from its local neighborhood
- After k convolutions, each node contains information from its k-hop neighborhood

# Graph convolutions

## Basic message passing

- $h_u^{(k)}$ is the embedding of node u after k convolutions.

- $W_{self}, W_{neigh} \in \mathbb{R}^{d_k \times d_{k-1}}$ are trainable weights



$$\mathbf{h}_u^{(k)} = \sigma\left(\mathbf{W}_{\text{self}}^{(k)}\mathbf{h}_u^{(k-1)} + \mathbf{W}_{\text{neigh}}^{(k)} \sum_{v \in \mathcal{N}(u)} \mathbf{h}_v^{(k-1)} + \mathbf{b}^{(k)}\right)$$