Academic Year 2023/2024

Social Media Analytics (SMA)

Network and Graph Theory

Part 1 – Basic Definitions

Marco Viviani

University of Milano-Bicocca

Department of Informatics, Systems, and Communication

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

Network and network types

- The term **network** commonly refers to a set of interconnected objects (or entities)

- Network types:
  - Technological networks
  - Information networks
  - Biological networks
  - Social networks

Technological networks

- **Technological networks**: man-made networks that are designed for the distribution of goods, resources, services
  - They have developed mainly in the last century and form the backbone of modern technological societies

- The most famous example is the **Internet**, the global network that connects computers and other information systems through electrical, optical and wireless technologies

- Other examples of technology networks include power grids, transportation grids, shipping and distribution grids, and telephone networks

Information networks

- **Information networks**: These are man-made networks made up of data and information linked together in some way

- The best-known example is the **World Wide Web** (WWW)
  - The WWW is a network in which the vertices are made up of Web pages made up of text, images or other information
  - Users can navigate from one page to another using hyperlinks, which are the edges that connect the vertices

- Other examples of information networks are e-mail communications networks and citation networks (example next slide)

Marco Viviani

Università degli Studi di Milano-Bicocca / DISCo
Verified email at unimib.it - Homepage
Social Computing Social Media Analytics Information Retrieval Text Mining
Trust and Reputation Syste...

FOLLOW

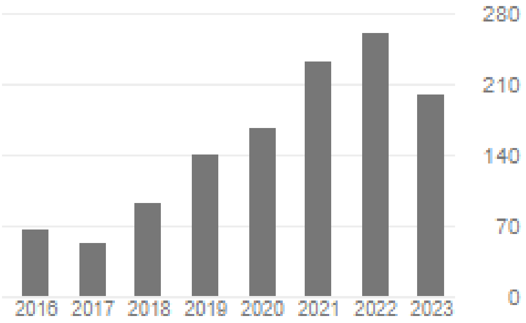
TITLE	CITED BY	YEAR
Credibility in social media: opinions, news, and health information—a survey M Viviani, G Pasi Wiley interdisciplinary reviews: Data mining and knowledge discovery 7 (5 ...	220	2017
Overview of the CLEF eHealth evaluation lab 2020 L Goeuriot, H Suominen, L Kelly, A Miranda-Escalada, M Krallinger, Z Liu, ... International Conference of the Cross-Language Evaluation Forum for European ...	132	2020
A survey on user modeling in multi-application environments M Viviani, N Bennani, E Egyed-Zsigmond 2010 Third International Conference on Advances in Human-Oriented and ...	86	2010
WoLMIS: A labor market intelligence system for classifying web job vacancies R Boselli, M Cesarini, S Marrara, F Mercorio, M Mezzanzanica, G Pasi, ... Journal of intelligent information systems 51, 477-502	80	2018
Feature analysis for fake review detection through supervised classification J Fontanarava, G Pasi, M Viviani 2017 IEEE international conference on data science and advanced Analytics ...	70	2017
Security and Trust in Online Social Networks B Carminati, E Ferrari, M Viviani Synthesis Lectures on Information Security, Privacy, & Trust 4 (3), 1-120	58	2013
A WOWA-based aggregation technique on trust values connected to metadata E Damiani, SDC di Vimercati, P Samarati, M Viviani Electronic Notes in Theoretical Computer Science 157 (3), 131-142	52	2006
Surveilling COVID-19 Emotional Contagion on Twitter by Sentiment Analysis C Crocamo, M Viviani, L Famiglini, F Bartoli, G Pasi, G Carrà European Psychiatry, 1-21	45	2021
LOOKER: a mobile, personalized recommender system in the tourism domain based on social media user-generated content S Missaoui, F Kassem, M Viviani, A Agostini, R Faiz, G Pasi	41	2019

GET MY OWN PROFILE

Cited by

VIEW ALL

	All	Since 2018
Citations	1539	1102
h-index	20	18
i10-index	39	29



Public access

VIEW ALL

2 articles 16 articles
not available available
Based on funding mandates

Co-authors

VIEW ALL

Gabriella Pasi
Università degli Studi di Milano B...

Ernesto Damiani
Khalifa University

nadia bennani

Biological networks

- **Biological networks**: networks that represent interaction patterns between biological elements

- Examples of biological networks:
  - Biochemical networks: networks representing molecular-level interaction patterns and control mechanisms in biological cells
  - Neural networks: networks that exist in the human brain and in the central nervous systems of animals
  - Ecological networks: networks of ecological interactions between species

Social networks

- **Social networks**: networks where vertices represent people or groups of people and the connections between them represent some form of social interaction, such as friendship

- Examples of social structures:
  - Friendship and knowledge networks
  - Meme spreading
  - Collaboration charts
  - Kinship
  - Disease transmission
  - Sexual relationship networks

Network theory

- **Network theory** is the study of structures that represent symmetrical or asymmetrical relationships between discrete objects

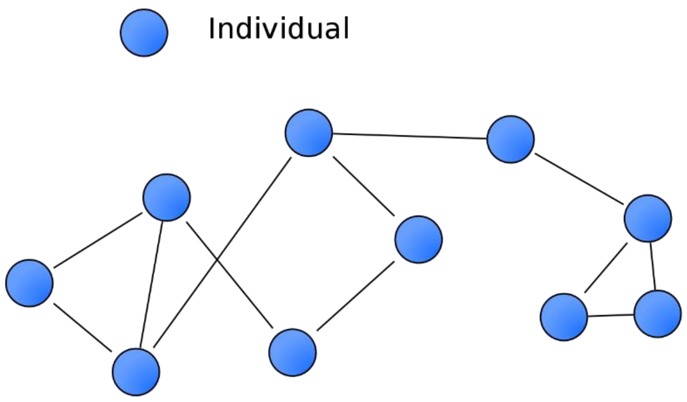
- In computer science, network theory is a part of **graph theory**: a network can be defined as a graph of nodes (vertices) interconnected through links (edges). Both vertices and edges can have attributes (e.g., names, weights, …)

Social Network Analysis (SNA)

- **Social Network Analysis (SNA)** is the process of investigating social structures through the use of network and graph theory

- Multiple graph representations:
  - Algebraic representation
  - Graphical representation
  - Matrix representation
  - List representation

Sociometry and sociograms



- The **sociogram** is the original tool, conceived by Jacob L. Moreno (1889 - 1974) in the context of sociometry, to graphically represent the configuration of social relations within a group

- The term remained in use even later, with the application of graph theory in the analysis of (online) social networks, to indicate that it is a **social network graph**

Moreno, Jacob Levy. "Who shall survive? Foundations of sociometry, group psychotherapy and socio-drama." (1953)

Graph theory

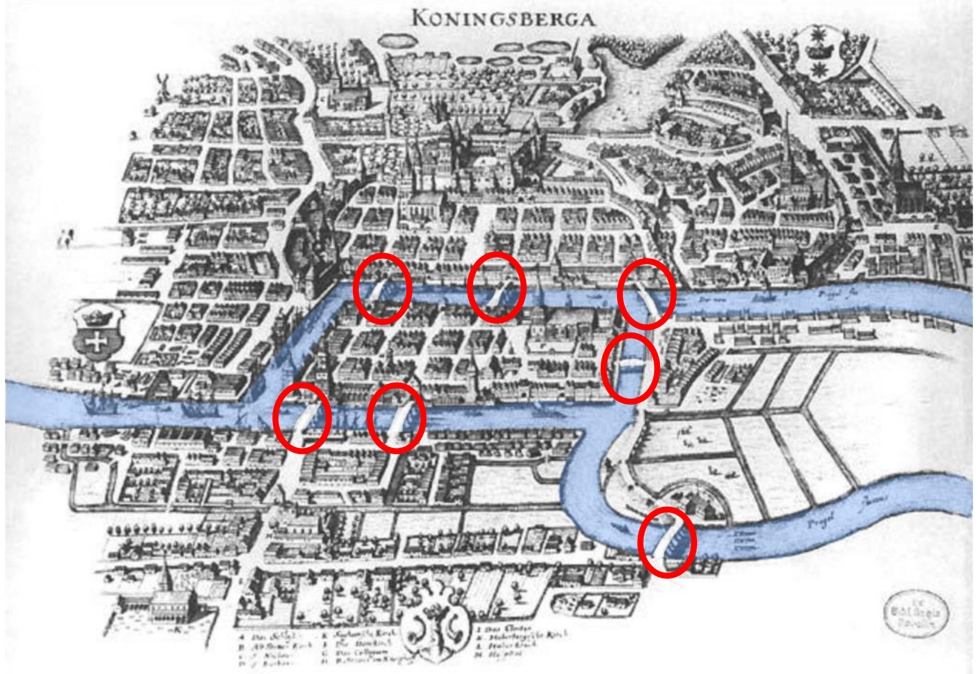
Origins, theoretical aspects, definitions

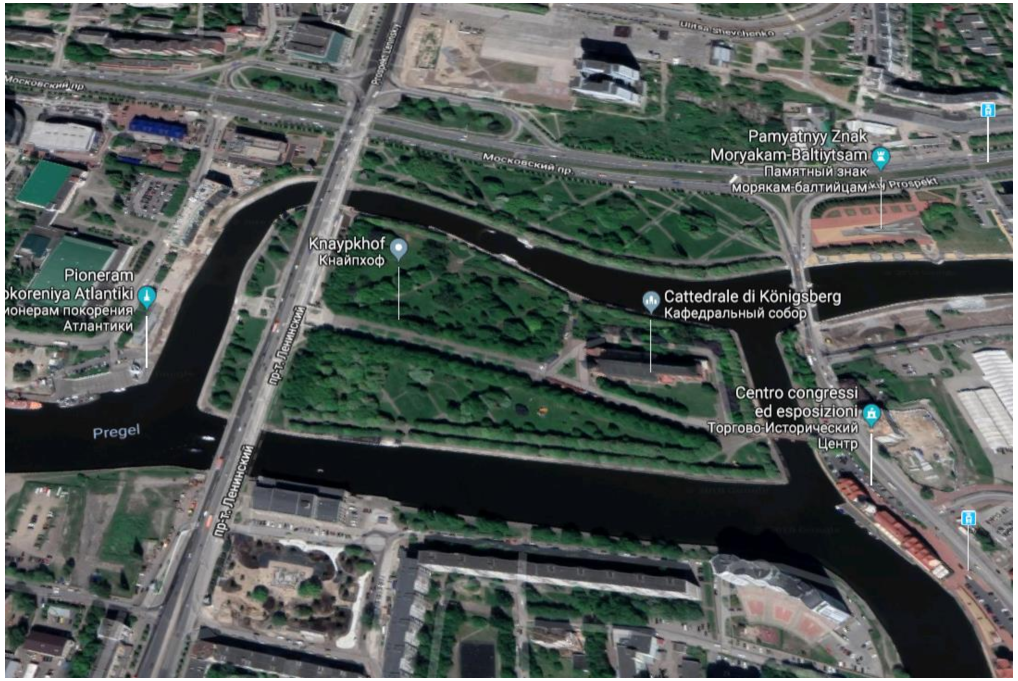
Origins

- **Graph theory** was introduced by the Swiss mathematician Euler (Leonhard Euler) (1707 - 1783)

The problem of the Königsberg bridges

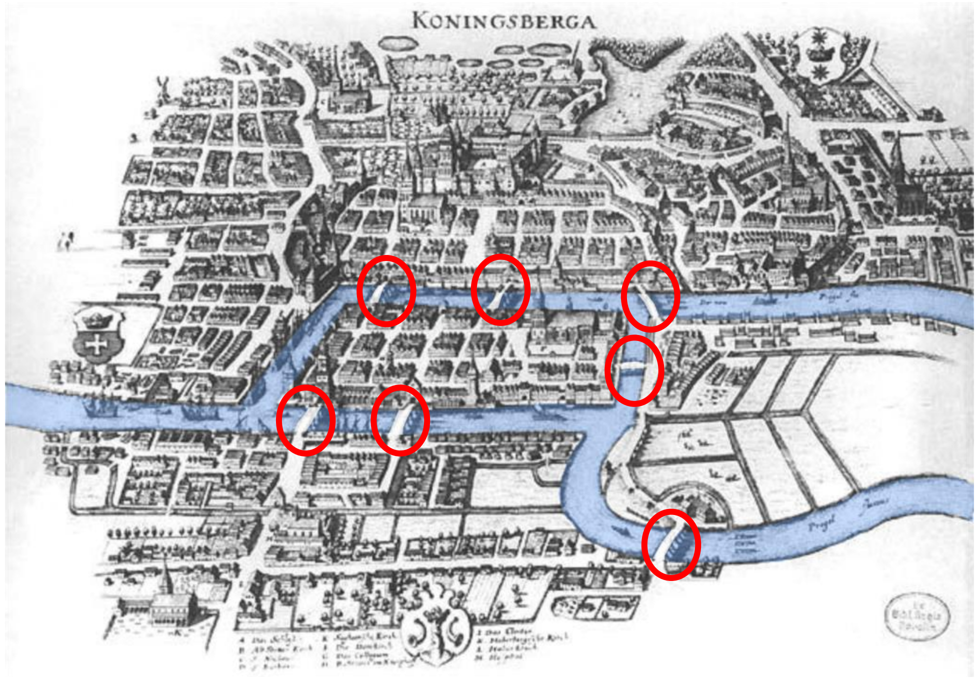
- Euler formulated using graphs the problem of the bridges of Königsberg (a city at the time Prussian, now Russified in Kaliningrad)

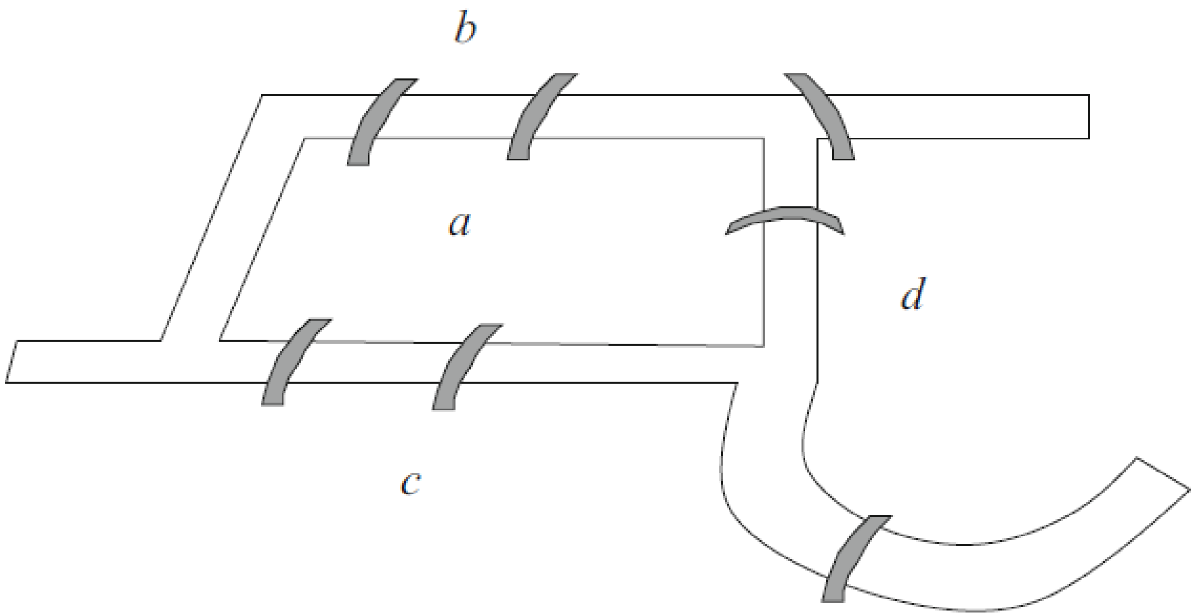




Description of the problem

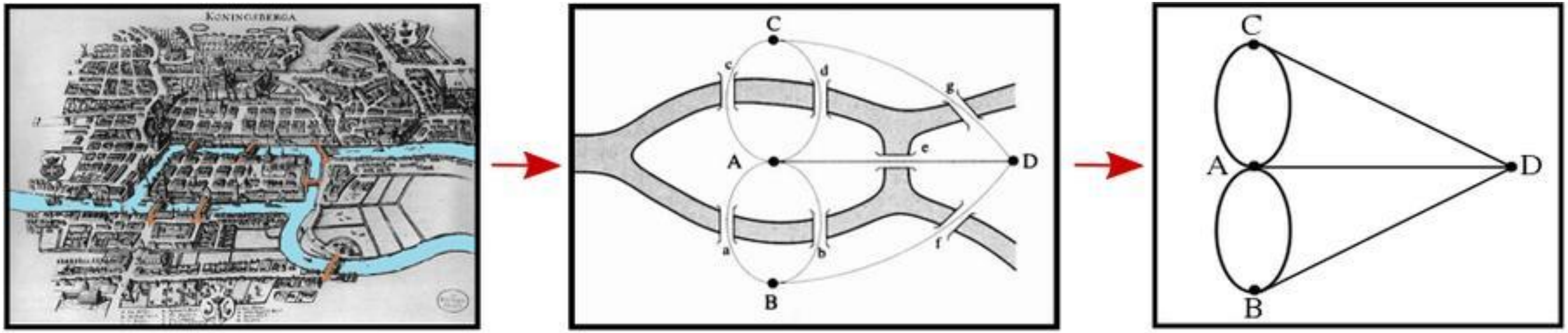
- From any part of the city, is it possible to take a walk-in order to cross all the bridges once and only once?



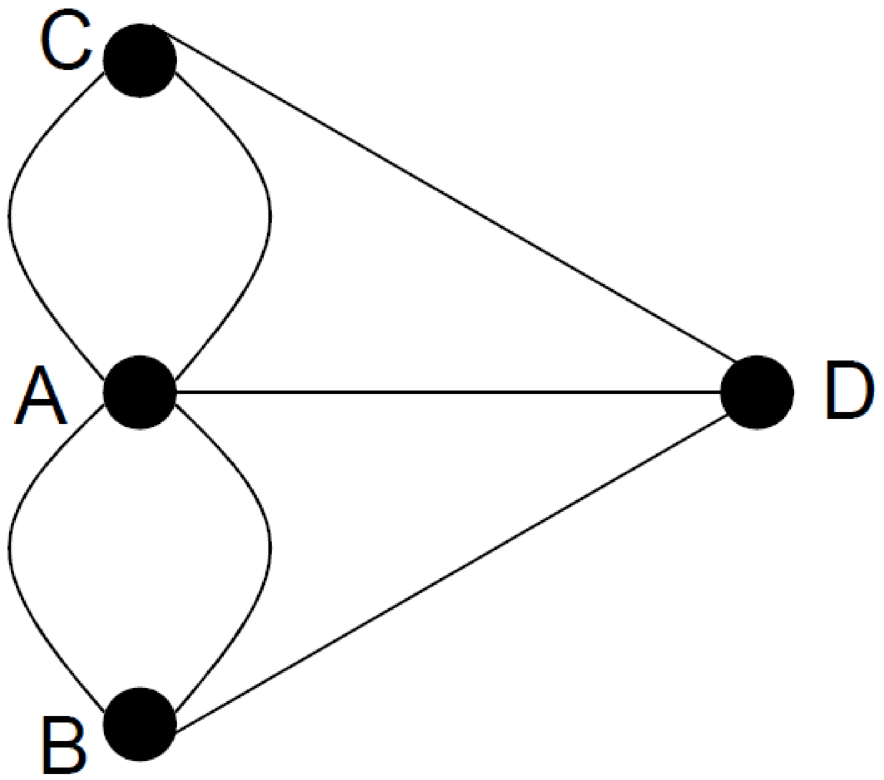


Modeling the problem

- If the land areas are associated with points (nodes or vertices) and the bridges are associated with line sections (arcs or edges) the Königsberg bridge problem is modeled by the graph:



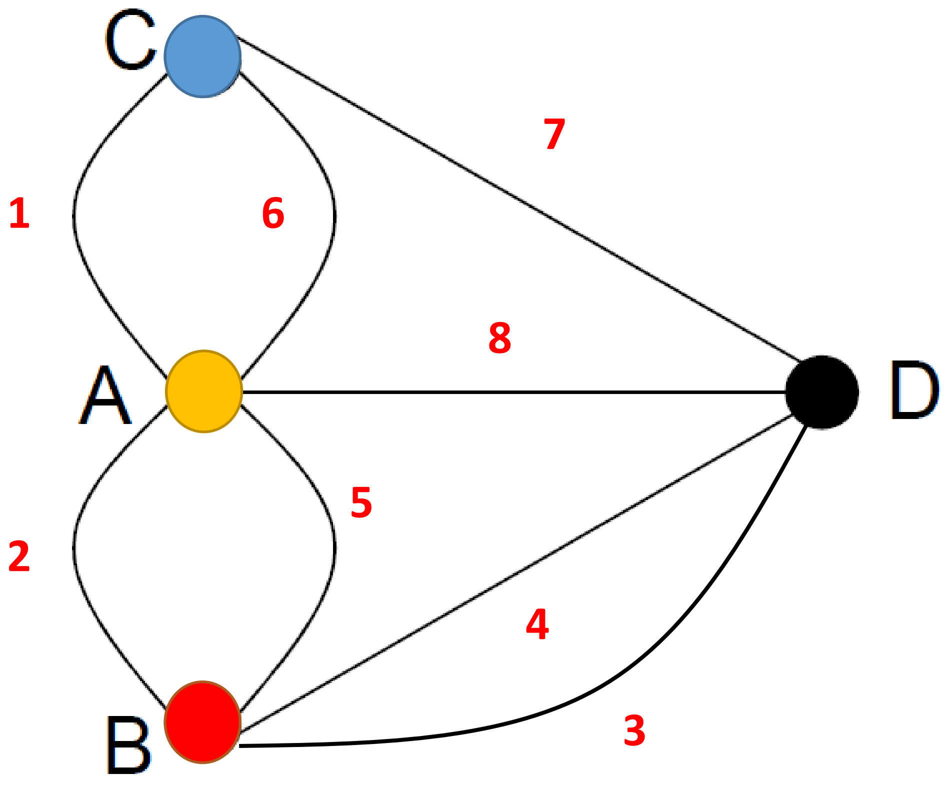
Problem
Outcome



- Euler used this graph to establish that it is **impossible** to find the required path, with the bridges thus distributed

Problem

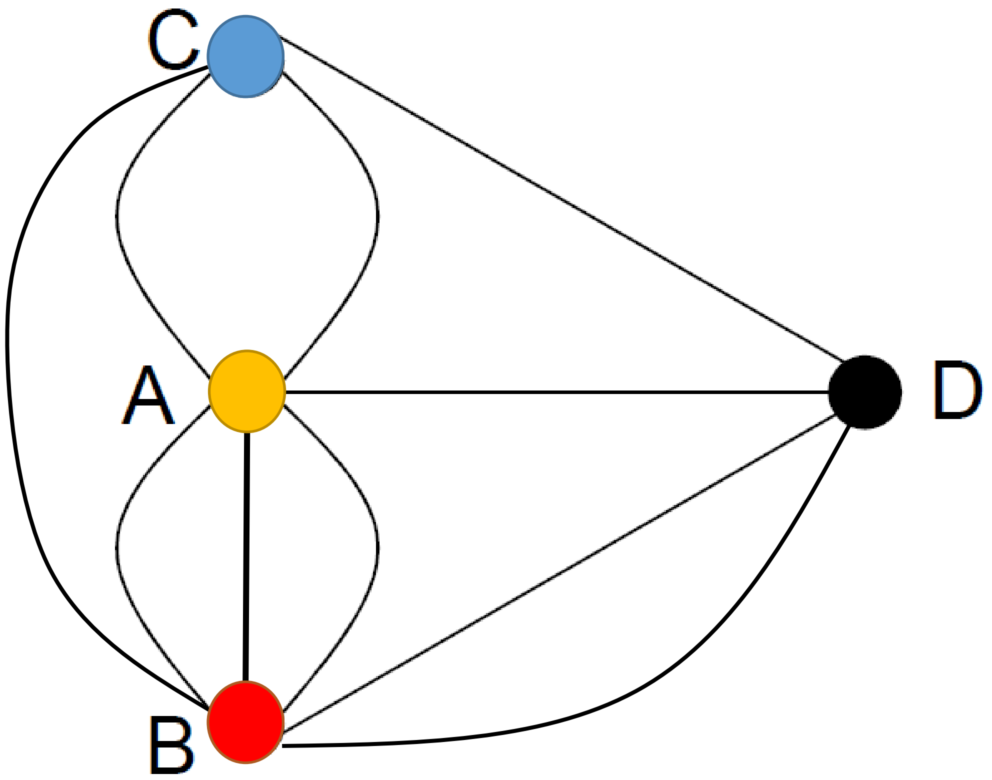
Solution



Addition of the eighth "bridge"

- Instead, it is possible:
  - If the number of incident arcs in each node is **even** (there are 0 nodes with odd number of incident arcs)
  - Or if **only two nodes** have an odd number of incident edges
- The **blue** node is the starting point, the **orange** node is the arrival point

Modified problem



- **Exercise**: find the path that pass only once for each bridge and back to the point of departure

- To solve the problem formulated this way it is necessary to add two more "bridges"

Formal definition of graph

- A **graph** is a pair $G = (V, E)$ of sets such that $E \subseteq [V]^2$; hence, the elements of E are 2-element subsets of V

- The elements of V are the **vertices** (or nodes or points) of the graph G, the elements of E are its **edges** (or arcs)

Sets of vertices and edges

Insiemi dei vertici e degli archi

- The **set of vertices** of a graph G is denoted as:

$$V(G) = \{v_1, v_2, \dots, v_n\}$$

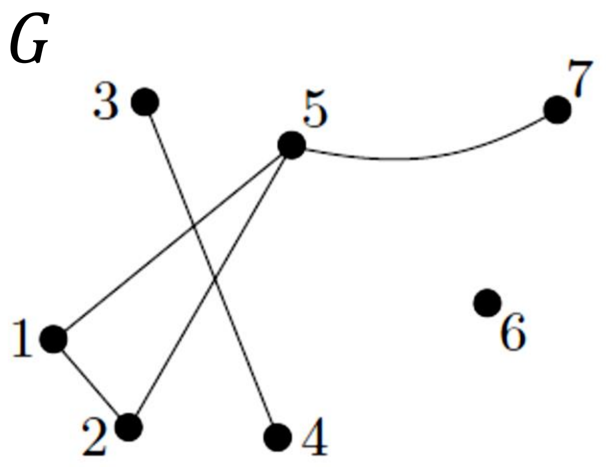
- The set of **edges** of a graph G is denoted as:

$$E(G) = \{e_1, e_2, \dots, e_m\}$$

- We denote with $e = \{a, b\}$ (or $e = (a, b)$) an edge between any two vertices a and b

Representation of a graph

- The usual way to **represent a graph** is to draw a point for each vertex and join two of these points with a line if the two corresponding vertices form an edge



The graph in the figure has as a set of vertices:

$$V(G) = \{1, 2, \dots, 7\}$$

It has as a set of edges:

$$E(G) = \{\{1,2\}, \{1,5\}, \{2,5\}, \{3,4\}, \{5,7\}\}$$

Or, with alternative notation:

$$E(G) = \{(1,2), (1,5), (2,5), (3,4), (5,7)\}$$

Order and size of a graph

Ordine e dimensione di un grafo

- The number of vertices (the cardinality of the vertex set) of a graph represents its **order**, denoted as:

$$|G| = |V(G)| = n$$

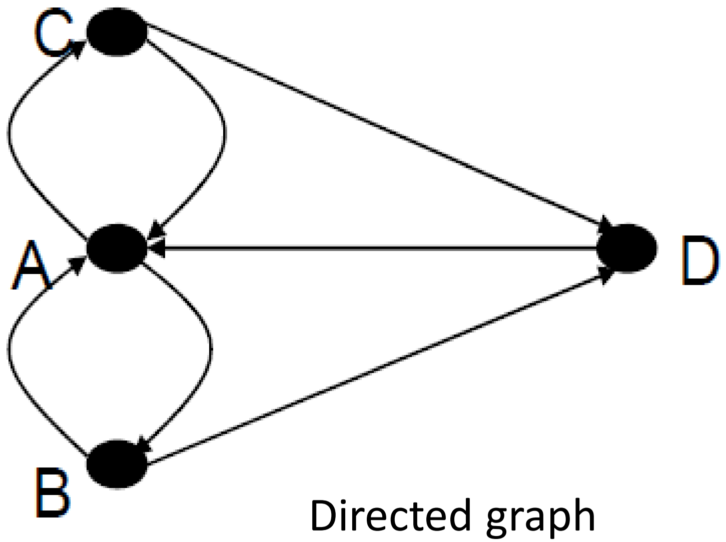
- The number of edges (the cardinality of the edge set) of a graph represents its **size**, denoted as:

$$||G|| = |E(G)| = m$$

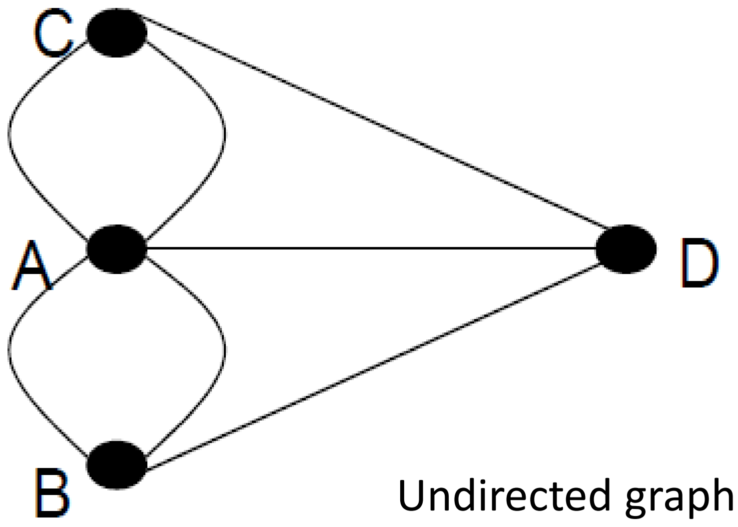
Directed and undirected graphs

Grafi orientati e non orientati

- A graph is said to **directed** if the edges have an orientation, i.e., all the edges are directed from one vertex to another; otherwise, it is said to be **undirected**, i.e., all the edges are bidirectional

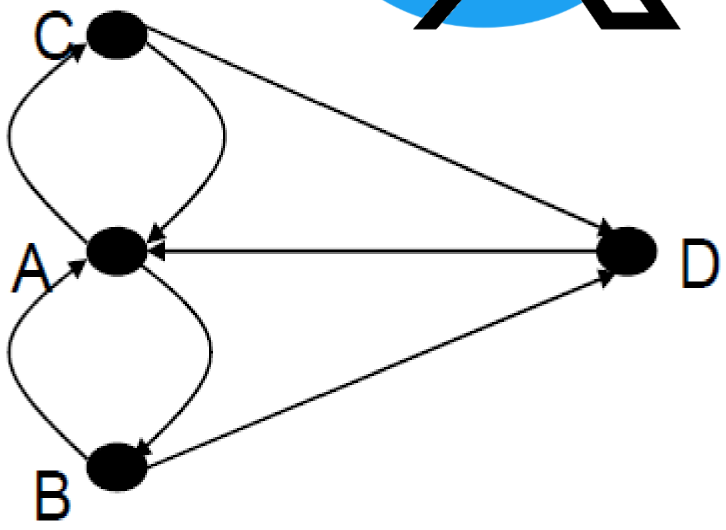


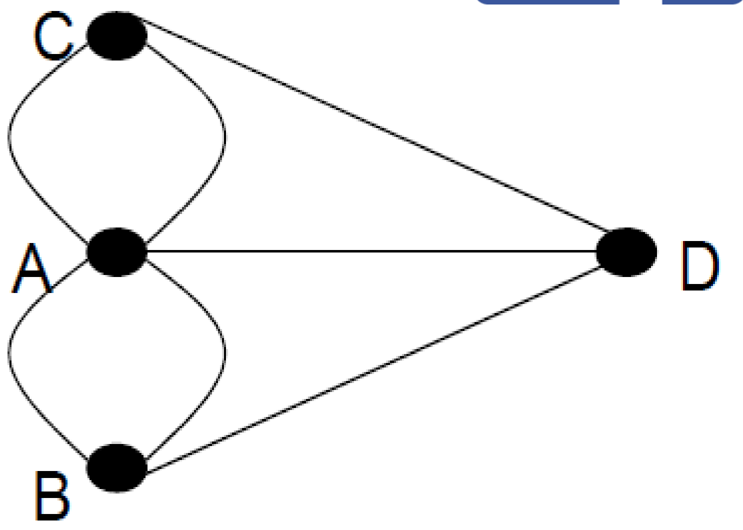
Directed graph



Undirected graph

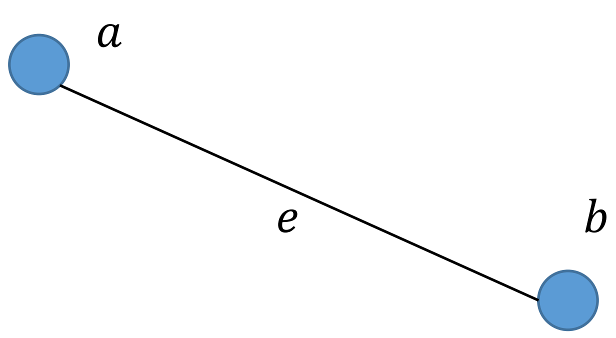
Directed and undirected graphs in social media





Undirected graphs
Definitions – 1

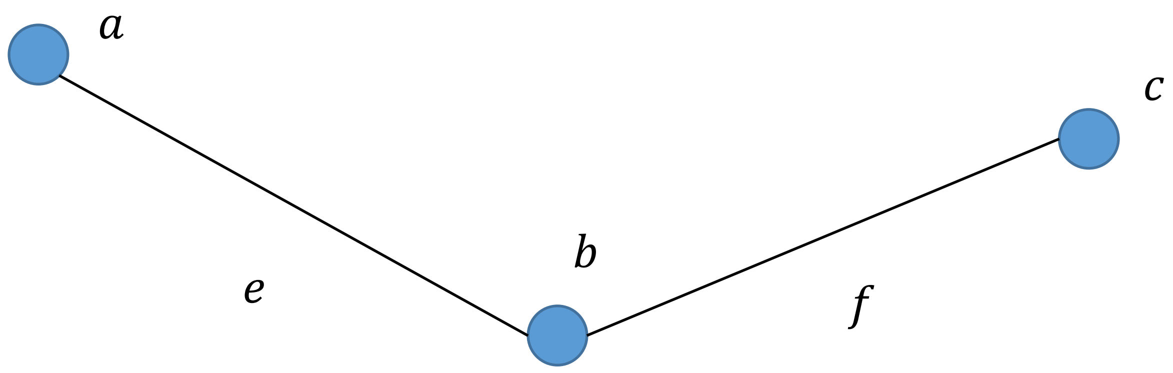
- Given the edge $e = (a, b)$
  - a, b are called the extreme vertices of e
  - e is called the incident edge (*arco incidente*) in a and b
  - a, b are called adjacent vertices (*vertici adiacenti*)
  - a is called neighbor of b in G and vice versa



Undirected graphs

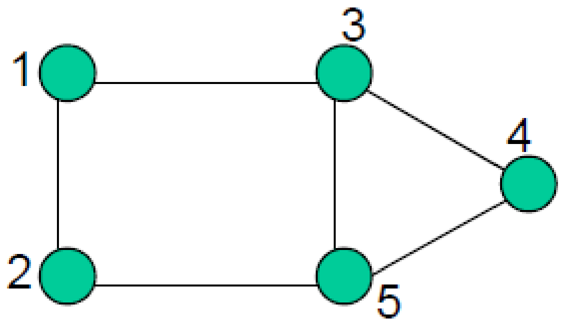
Definitions – 2

- Given the edges $e = (a, x)$ and $f = (y, c)$
    - Two edges e, f are said to be **adjacent** if they have a common vertex, i.e., $x = y = b$.



Undirected graphs
Definitions – 3

- Given the vertex a in G
  - The neighbourhood (*intorno*) of a, denoted as $N_G(a) = N(a)$, is the set of vertices adjacent to a
  - The star (*stella*) of a, denoted as $s_G(a) = s(a)$, is the set of edges incident in a



G

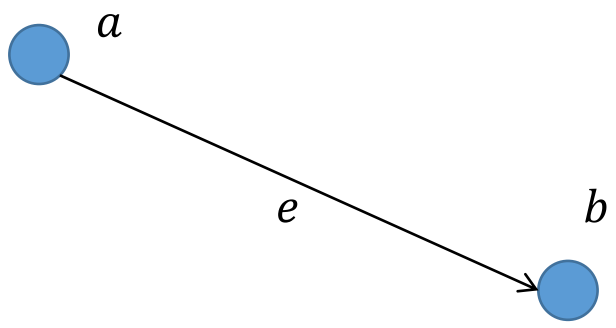
$$N_G(1) = N(1) = \{2, 3\}$$

$$s_G(3) = s(3) = \{(1,3), (3,4), (3,5)\}$$

Directed graphs

Definitions – 1

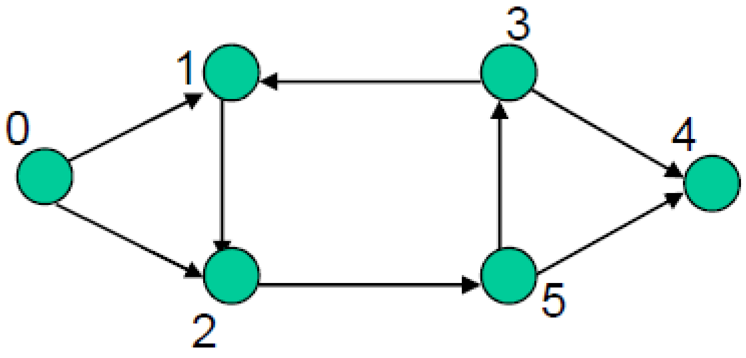
- Given a directed edge $e = (a, b)$
  - e is called outgoing edge (*arco uscente*) from a (tail vertex/*vertice coda*)
  - e is called ingoing edge (*arco entrante*) in b (head vertex/*vertice testa*)
  - a is said to be direct predecessor (*predecessore diretto*) of b
  - b is said to be the direct successor (*successore diretto*) of a



Directed graphs

Definitions – 2

- Given a vertex a of a directed graph G
  - $E^+(a)$ is the set of outgoing edges from a
  - $E^-(a)$ is the set of incoming edges in a
  - a vertex with only incoming edges is called sink vertex (*pozzo*)
  - a vertex with only outgoing edges is called source vertex (*sorgente*)



$E^+(5) = \{(3,5),(4,5)\}$

$E^-(2) = \{(0,2),(1,2)\}$

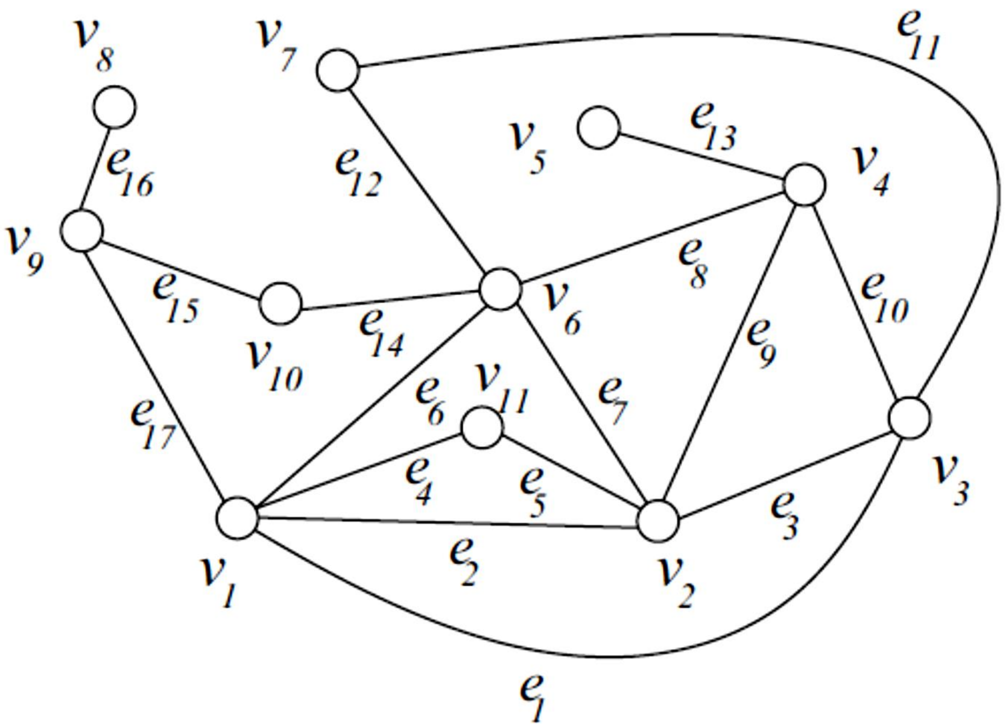
What is vertex 0?

What is vertex 4?

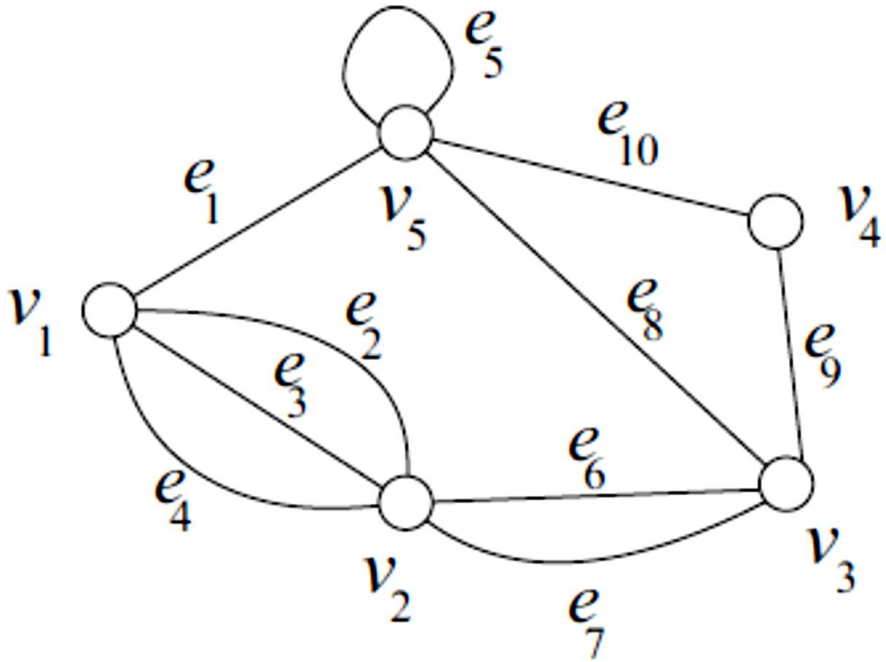
Simple graph and multigraph

- A **loop** (*cappio*) is an edge whose extreme vertices are the same vertex
- **Multiple edges** (*archi multipli*) are (different) edges with the same pair of extreme vertices
- If a graph G has no loops and no multiple edges, G is called a **simple graph** (*grafo semplice*)
- If not, it is called a **multigraph** (*multigrafo*)

Simple graph and multigraph (Example)



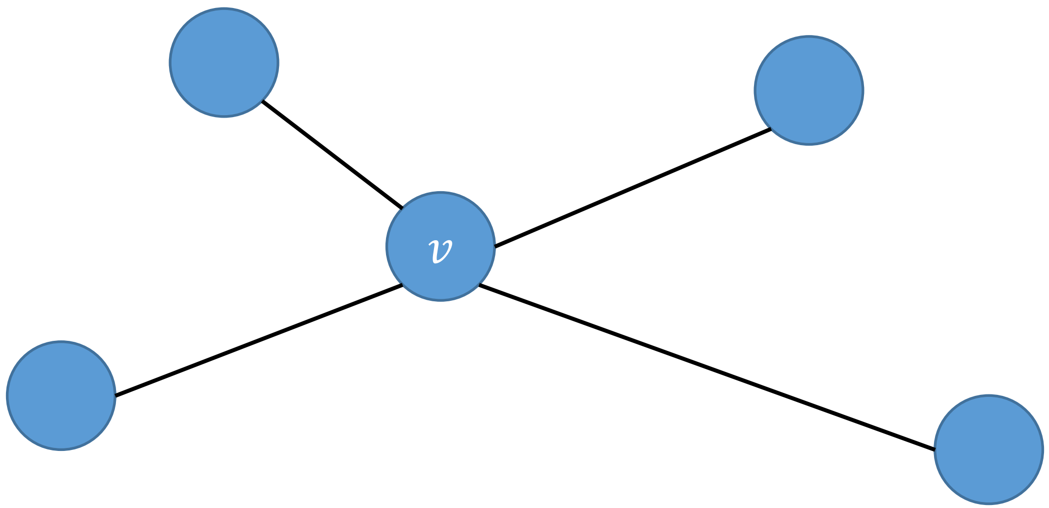
Simple graph



Multigraph

Degree of a vertex
Undirected graphs

- Generally speaking, the **degree of a vertex** (*grado di un vertice*) is the number of edges incident to a node



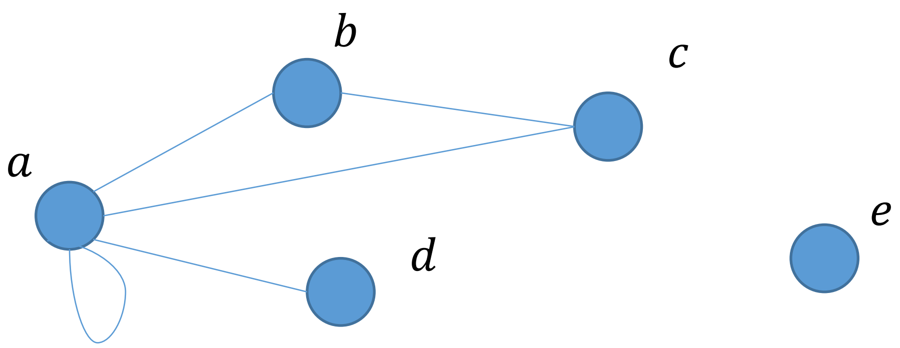
Degree definition

Undirected graphs

- Let $G = (V, E)$ be a graph. Il **degree** $d_G(v) = d(v)$ of a vertex v è is the number $|E(v)|$ of edges incident at v; this is equivalent to saying that it is the number of neighbors of v. Each loop on v is counted twice

- A vertex of degree 0 is called an **isolated vertex** (*vertice isolato*)

$d(a) = 5$



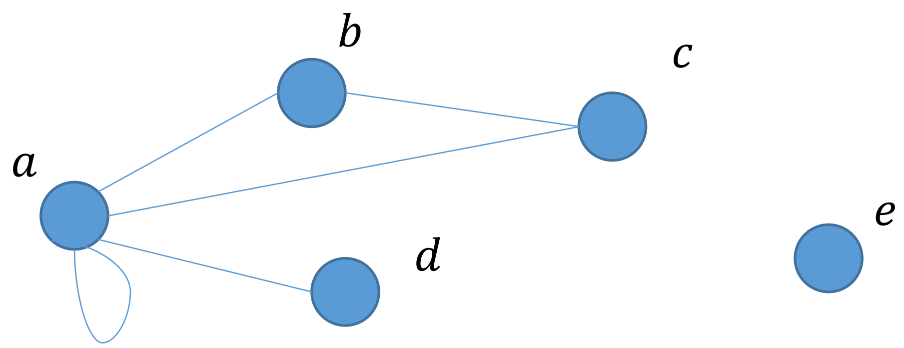
$d(e) = 0$ (isolated vertex)

Minimum and maximum degree of G

Undirected graphs

- The value $\delta(G) := \min\{d(v) | v \in V\}$ is the **minimum degree** (*grado minimo*) of G

- The value $\Delta(G) := \max\{d(v) | v \in V\}$ is the **maximum degree** (*grado massimo*) of G



$$\delta(G) = 0$$

$$\Delta(G) = 5$$

Average degree of G
Undirected graphs

- The value

$$d(G) := \frac{1}{|V|} \boxed{\sum_{v \in V} d(v)}$$

is called the **average degree** (*grado medio*) of G

- It follows that:

$$\delta(G) \leq d(G) \leq \Delta(G)$$

Handshaking lemma
Undirected graphs

- Let $G = (V, E)$ be a graph constituted by $|E(G)| = m$ edges. It follows that:

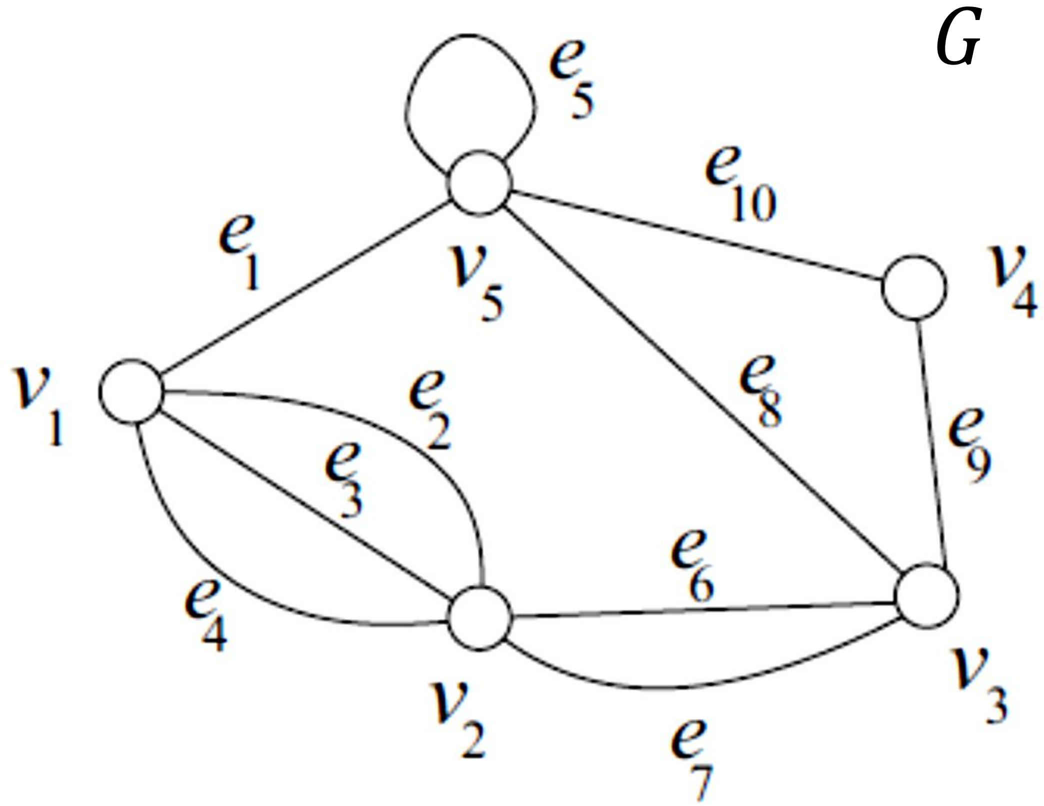
$$\boxed{\sum_{v \in V} d(v)} = 2|E(G)| = 2m$$

- **Proof**: every edge (which is not a loop) is incident at exactly two distinct vertices. It is therefore counted twice. Also, each loop is already counted twice. This gives the result shown in the formula

Example

Undirected graphs

- $d(v_1) = ?$ 4
- $d(v_5) = ?$ 5
- $\delta(G) = ?$ 2
- $\Delta(G) = ?$ 5
- $d(G) = ?$ 4

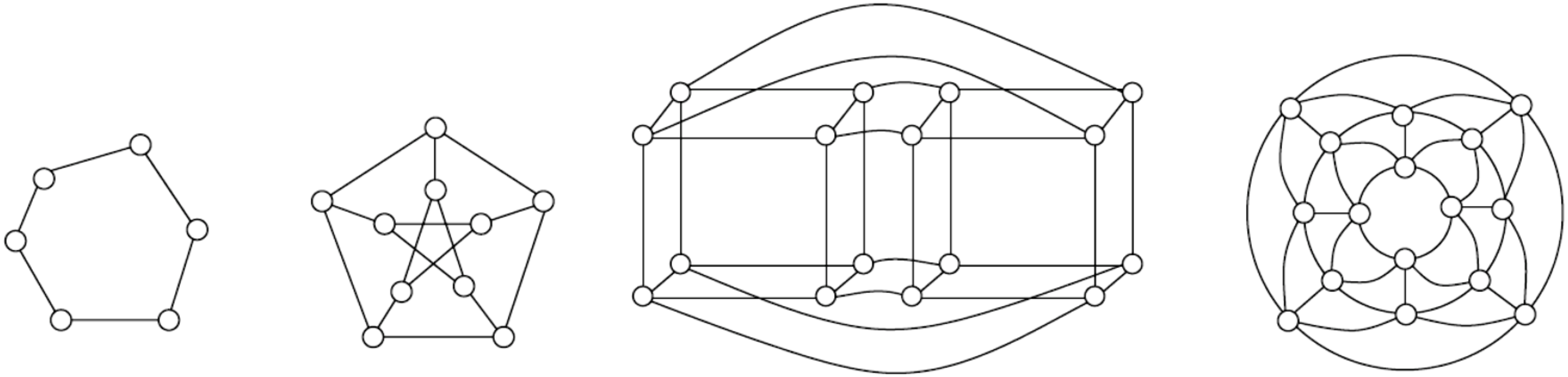


Regular graphs

- If all vertices of G have the same degree k, then G is called **k-regular**, o **regular of degree k,** or, simply, **regular**

- Graphs made up of only isolated vertices (they have an empty set of edges) are called 0-regular, or **null graphs**

Examples of regular graphs

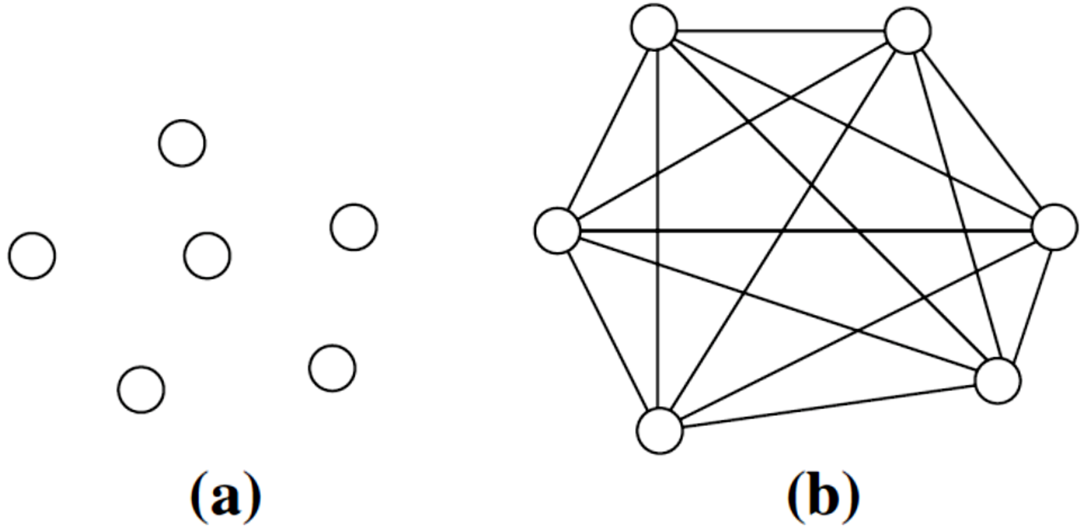


(a) (b) (c) (d)

(a) A 2-regular graph (a simple cycle), **(b)** a 3-regular graph (Petersen graph), **(c)** a 4-regular graph (4-dimensional hypercube), and **(d)** a 5-regular graph (a doughnut graph)

Complete graphs

- A graph in which each pair of distinct vertices is adjacent is called a **complete graph**

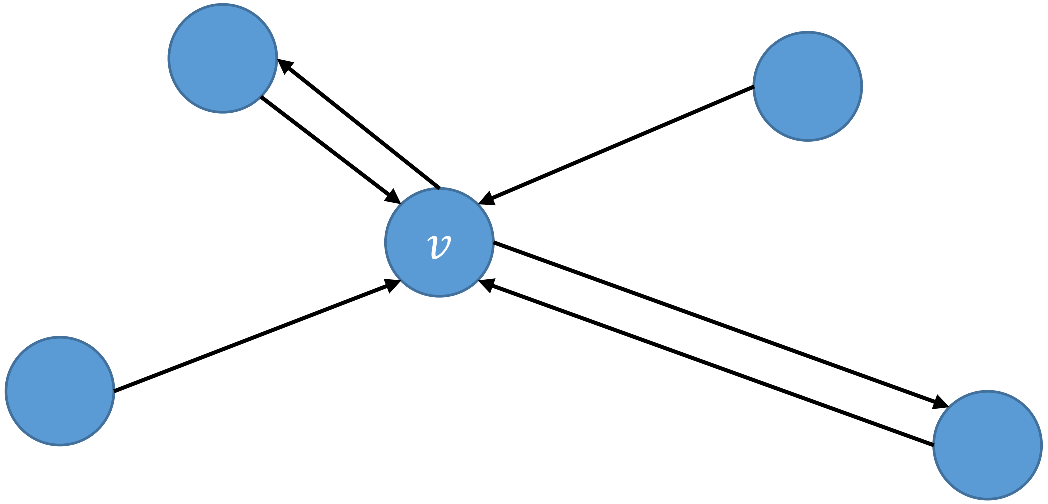


(a) A null graph N_6 with six vertices, **(b)** a complete graph K_6 with six vertices

In-degree
Directed graphs

- The **in-degree** (*grado in entrata*) $d_{in}(v)$ of a vertex v is the number of edges arriving at the v vertex (incoming edges)

$$d_{in}(v) = 4$$

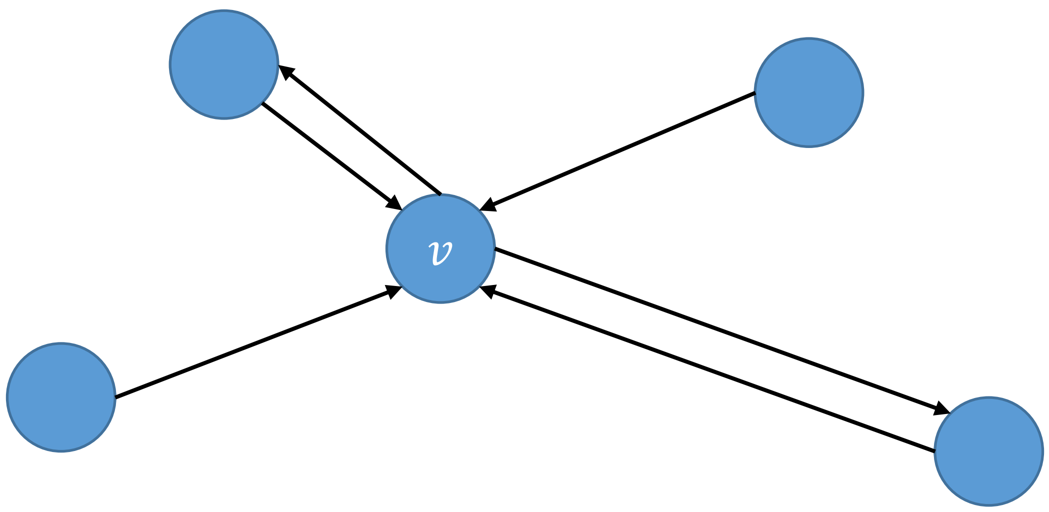


Out-degree

Directed graphs

- The **out-degree** (*grado in uscita*) $d_{out}(v)$ of a vertex v is the number of edges starting from the vertex v (outgoing edges)

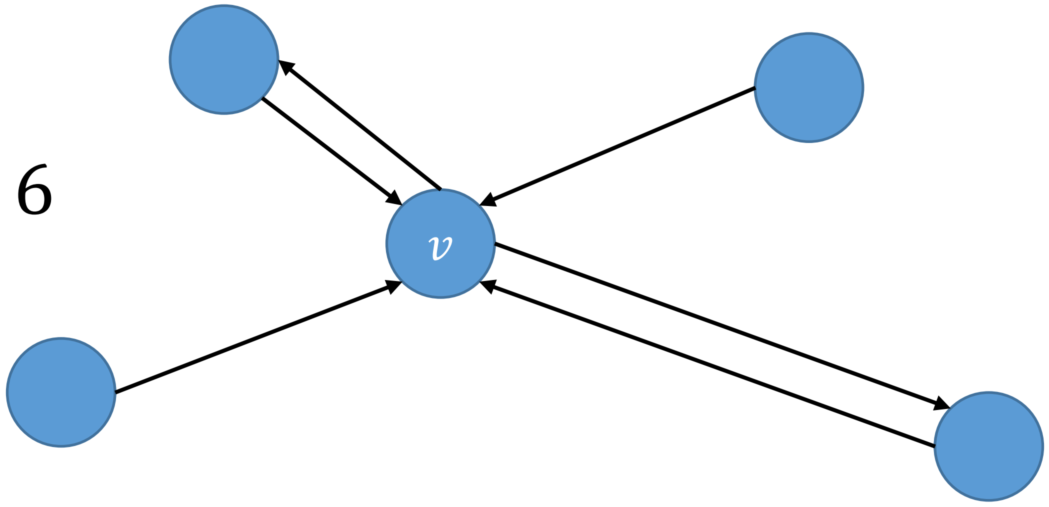
$$d_{out}(v) = 2$$



Degree of a vertex
Directed graphs

- The degree $d(v)$ of a vertex v is the sum of the number of its incoming and outgoing edges

$$d(v) = 4 + 2 = 6$$



Subgraph
Definition

- A **subgraph** (*sottografo*) of a graph $G = (V, E)$ is a graph $G' = (V', E')$ such that $V' \subseteq V$ and $E' \subseteq E$

- It is possible to obtain subgraphs of a graph G by removing some vertices and/or edges from G

Subgraph
Edge deletion

- Let e be an edge of G. We denote by $G - e$ the graph obtained by deleting the edge e from G

- More generally, if F is a set of edges of G, we denote by $G - F$ the graph obtained by deleting all edges in F from G

- The deletion of edges does not involve further modifications to the graph

Subgraph

Vertex deletion

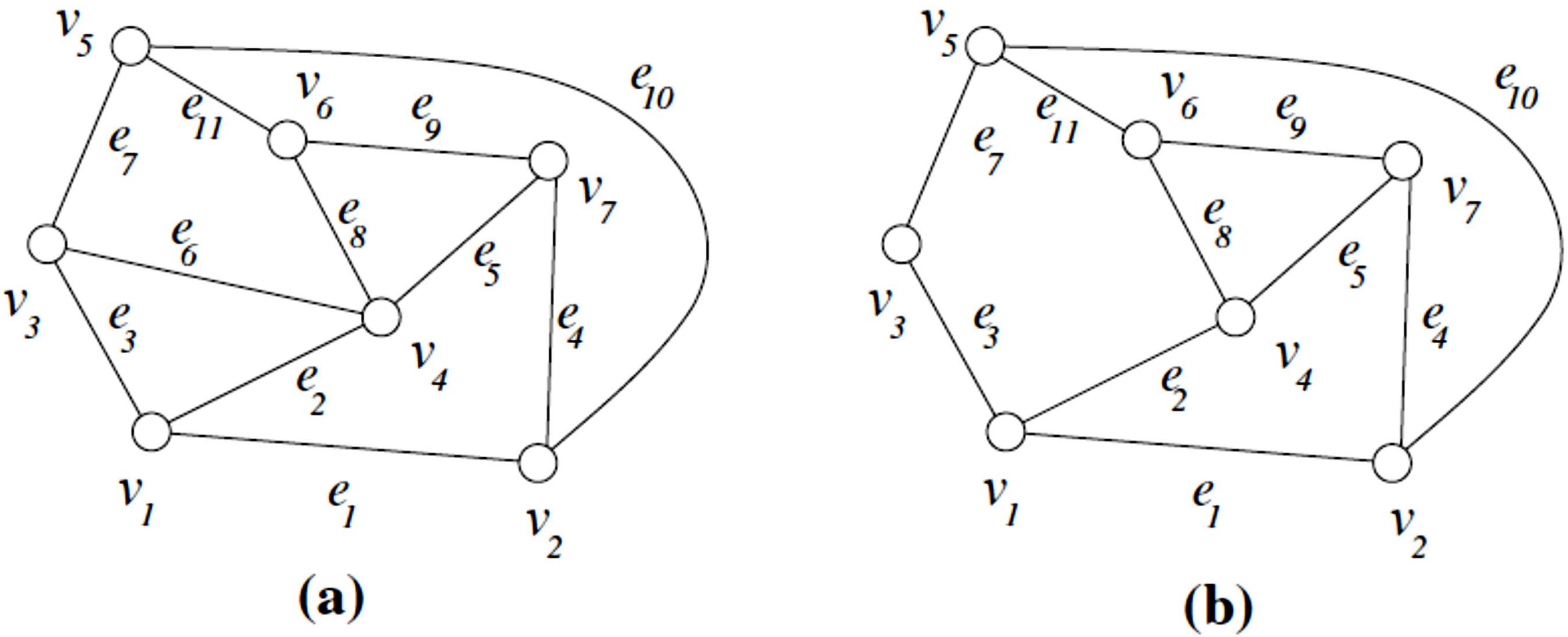
- Similarly, we can define the deletion of a vertex v from a graph. However, deleting a vertex v from a graph G also requires deleting the edges incident to v in G

- Let v be a vertex of a graph G. We denote by $G - v$ the graph obtained by eliminating the vertex v and all its edges incident by G

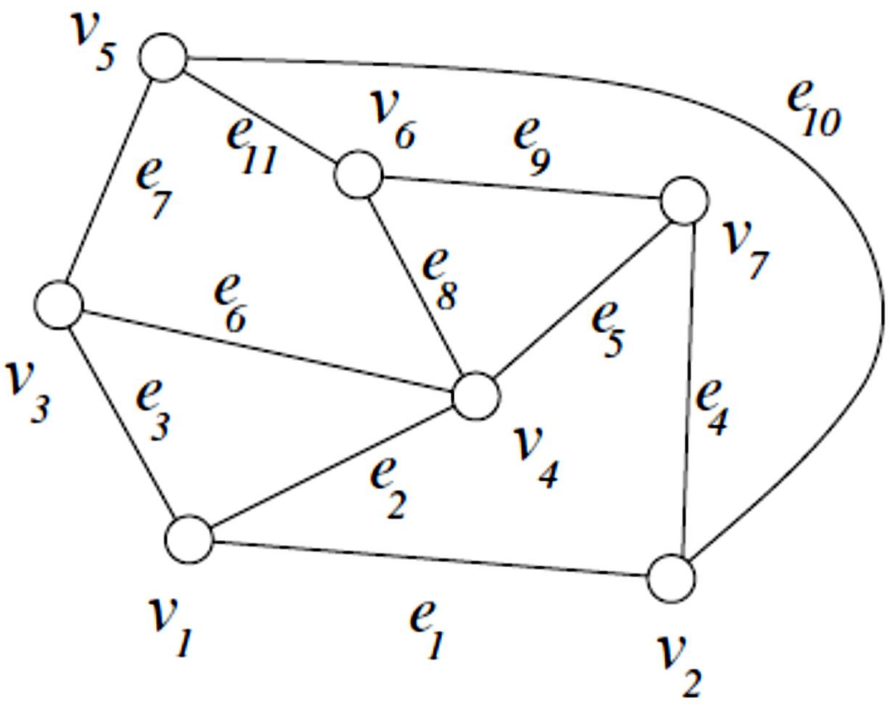
- More generally, if W is a set of vertices of G, we denote by $G - W$ the graph obtained by deleting the vertices in W (and all incident edges) from G

Examples of subgraph

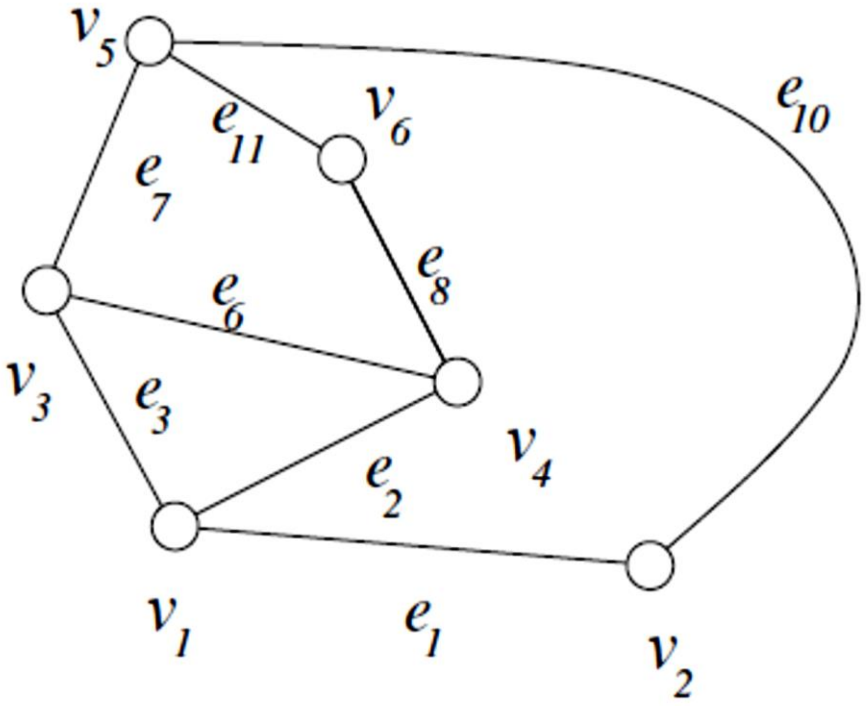
Deleting an edge (e_6)



Examples of subgraph
Deleting an edge (v_7)



(a)

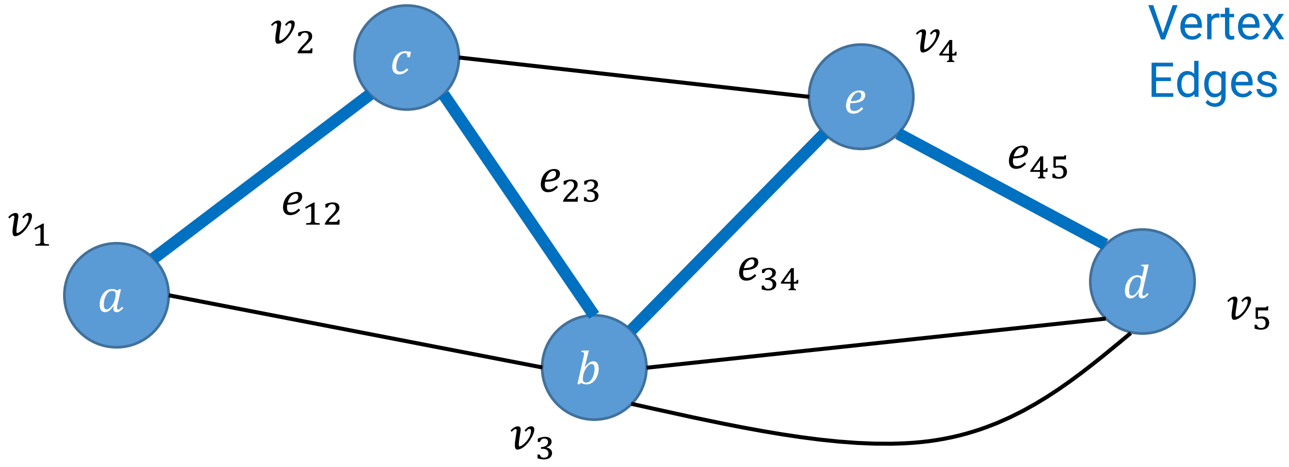


(c)

Walk

Cammino

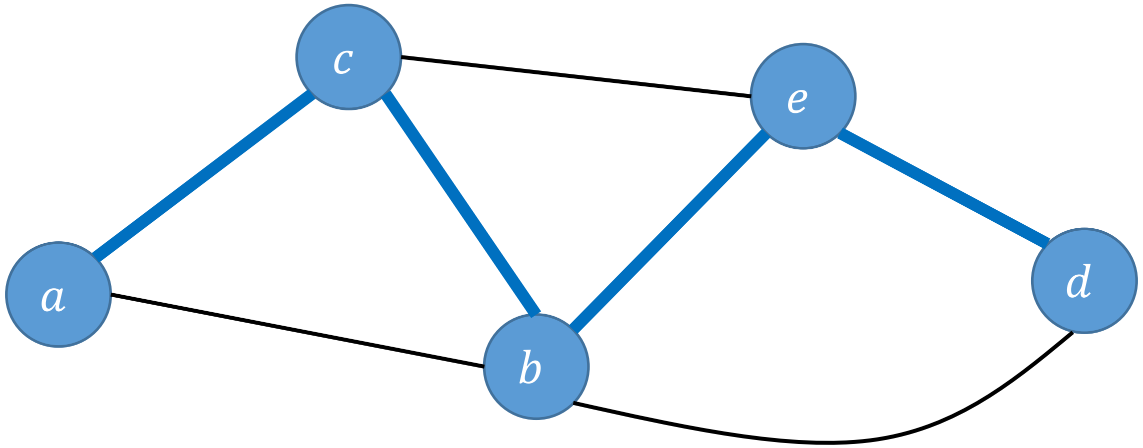
- Given a graph $G = (V, E)$ wit $V = \{v_1, v_2, \dots, v_n\}$ and $E = \{e_1, e_2, \dots, e_m\}$ a **walk** is a finite or infinite alternating sequence of vertices and edges $\omega = \{v_1, e_{12}, v_2, e_{23}, \dots, v_{k-1}, e_{(k-1)k}, v_k\}$ where v_{i-1} and v_i are adjacent



Vertex can be repeated
Edges can be repeated

Walk on a simple graph

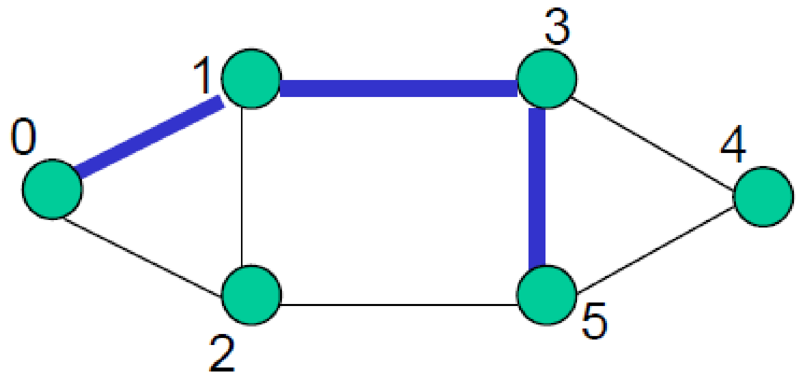
- On a simple graph (without loops and multiple arcs) a walk can be indicated as a sequence of vertices

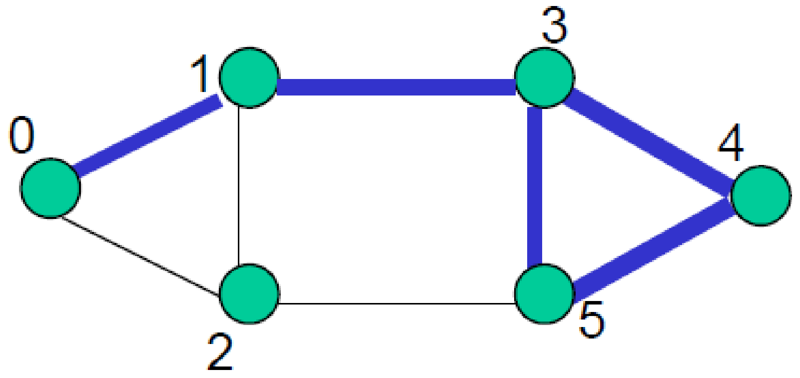


- Example: $w = \{a, c, b, e, d\}$

Walk types

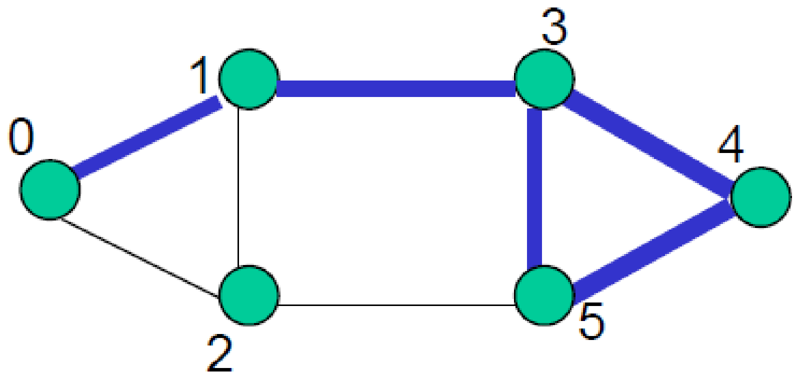
- A walk is called simple if the edges and vertices of the walk are all distinct (no repeated vertices or edges in the walk); otherwise it is said not simple



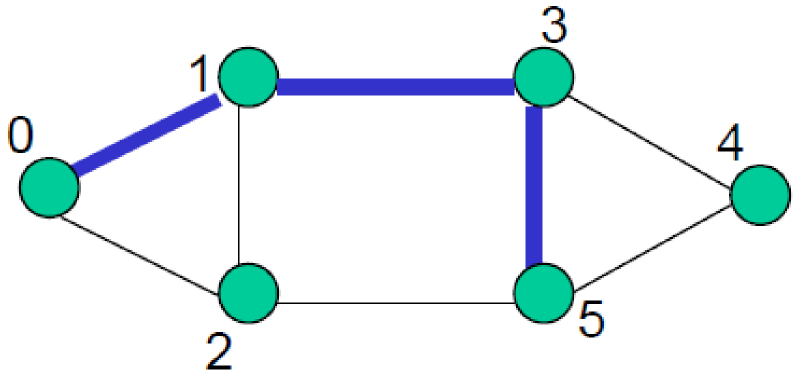


In particular…

- A **trail** is a walk in which all edges are distinct

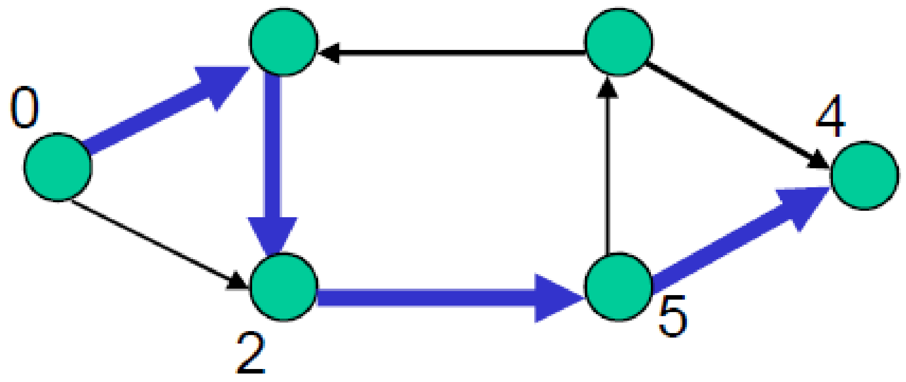


- A **path** is a trail in which all vertices (and therefore also all edges) are distinct



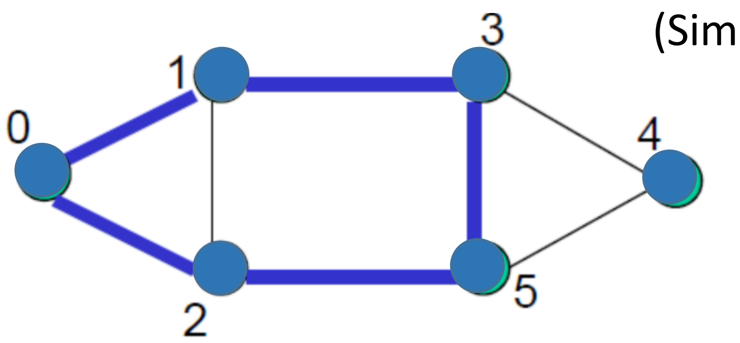
Directed walk

- A walk is said to be **directed** if for each edge $e = (a, b)$ in the walk, the vertex a is the tail of e and the vertex b is the head of e.

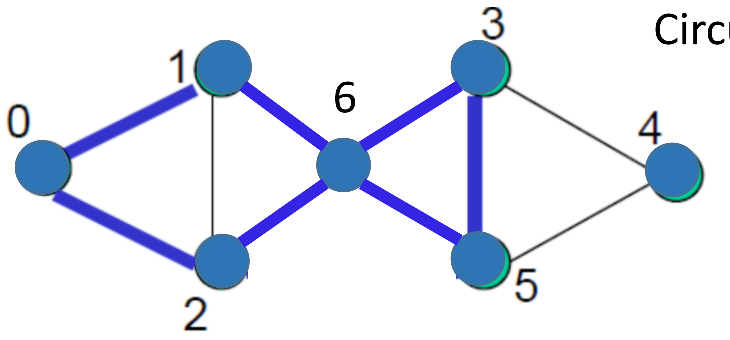


Closed walk, cycle, circuit

- A **closed walk** is a walk where the extreme vertices coincide
- A simple and closed walk is called a (simple) **cycle**
- A **circuit** can be a closed walk that allow repetition of vertices but not of edges

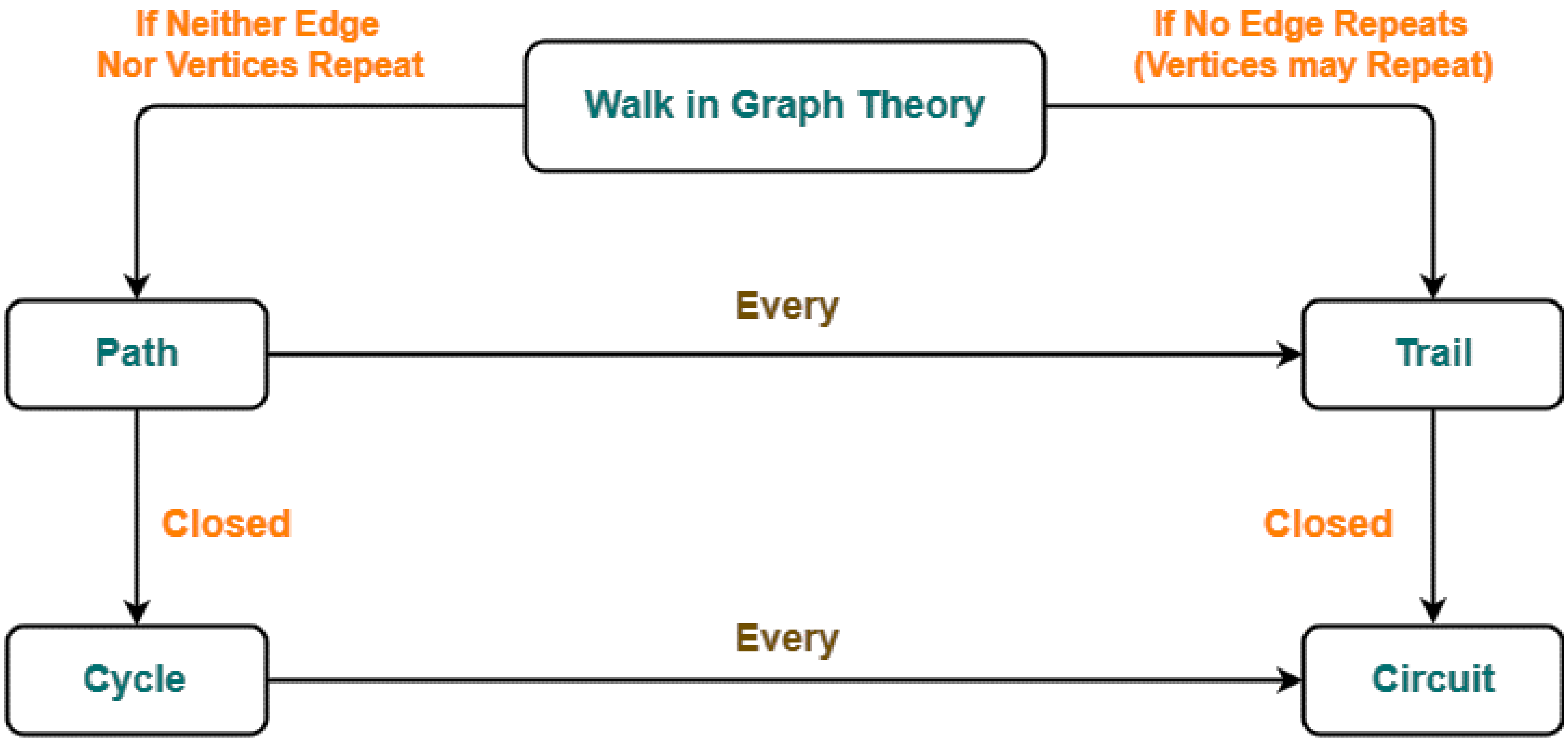


(Simple) cycle



Circuit

To recap…

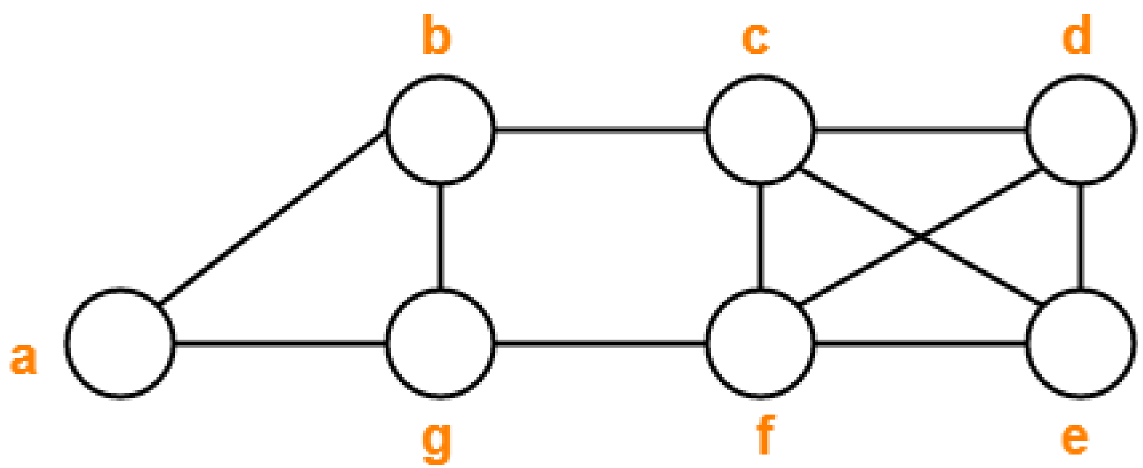


If Neither Edge
Nor Vertices Repeat
Walk in Graph Theory
If No Edge Repeats
(Vertices may Repeat)
Every
Path
Trail
Closed
Closed
Every
Cycle
Circuit

Exercise

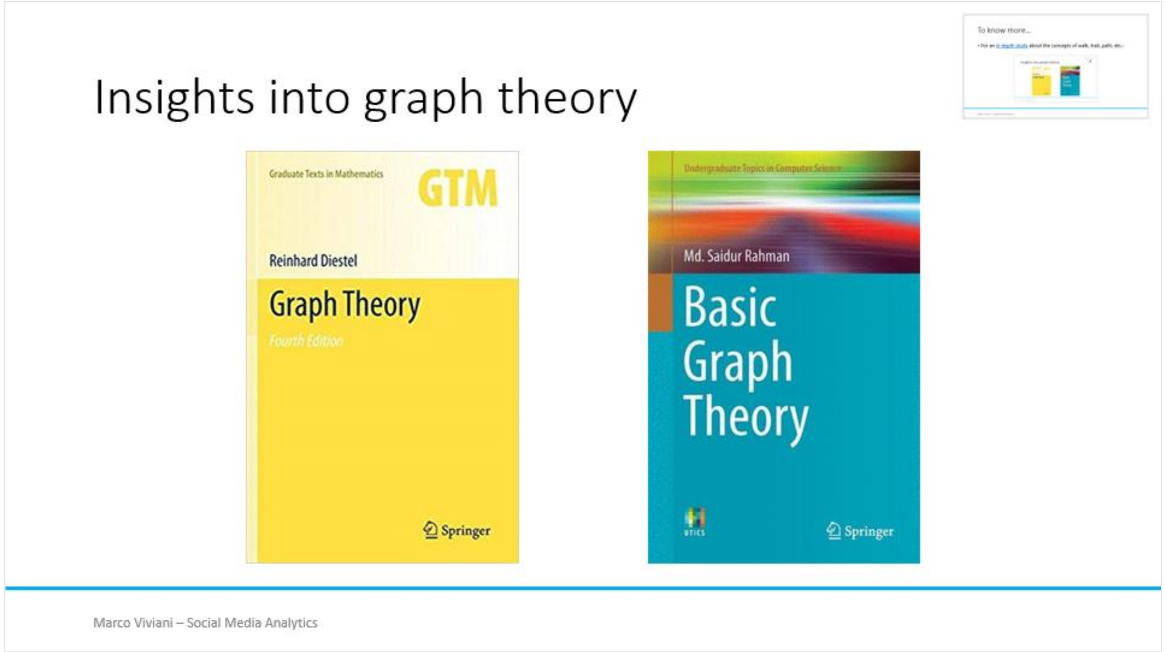
- Consider the graph in the figure. For those sequences of vertices that are walks, decide whether they are a path, a trail, a cycle or a circuit.

  - a , b , g , f , c , b — *Trail*
  - b , g , f , c , b , g , a — *Walk*
  - c , e , f , c — *Cycle*
  - c , e , f , c , e — *Walk*
  - a , b , f , a — *Not a walk*
  - f , d , e , c , b — *Path*
  - b, g, f, c, e, d, c, b — *Circuit*



To know more…

- For an in-depth study about the concepts of walk, trail, path, etc.:



Concepts related to walks

Some definitions

- **Length of a walk**
  - Number of edges crossed
- **Distance** between two nodes a e b
  - Length of the shortest path between a and b

Some metrics

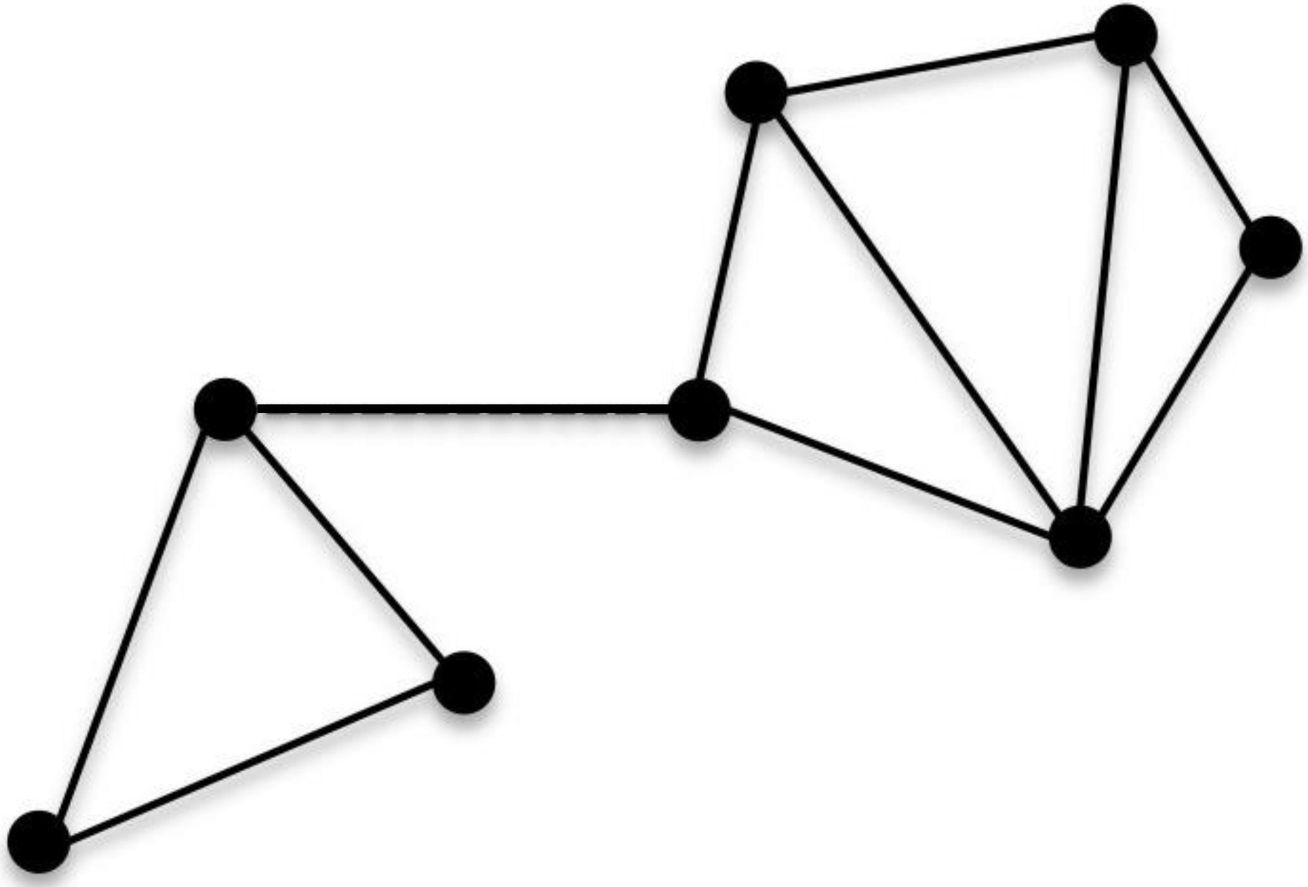
- **Eccentricity** of a vertex v: $\epsilon(v)$
  - Maximum distance of the vertex from any other vertex
- **Radius** of the graph: $rad(G)$
  - Minimum eccentricity in the graph
- **Diameter** of the graph: $diam(G)$
  - Maximum eccentricity in the graph

Connected graph

Valid for both undirected and directed, «but»...

- A graph is said to be **connected** when there is a path between each pair of vertices
- In a connected graph, there are **no unreachable vertices**
- A graph with only one vertex is connected

Connected graph (Example)



Connectivity

Valid for both undirected and directed, «but»…

- **Connectivity** is one of the basic concepts of graph theory. It is indicated by con $\kappa(G)$ and measures the minimum number of elements (vertices or edges) that must be removed to disconnect the graph

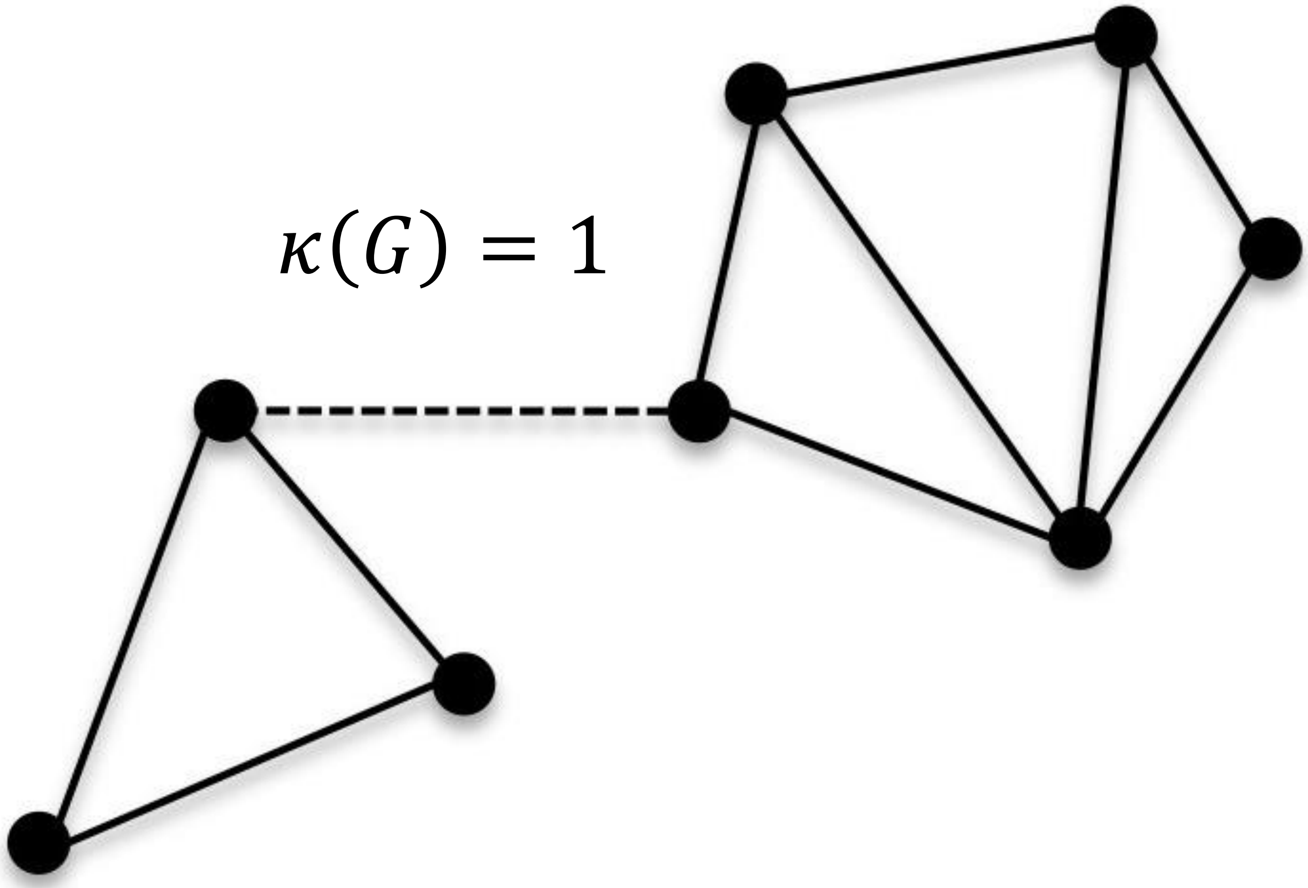
- The measure of connectivity is closely related to the theory of network flow problems

- The connectivity of a graph is an important measure of its resilience as a network

Connectivity

Example: Edge deletion on undirected graphs

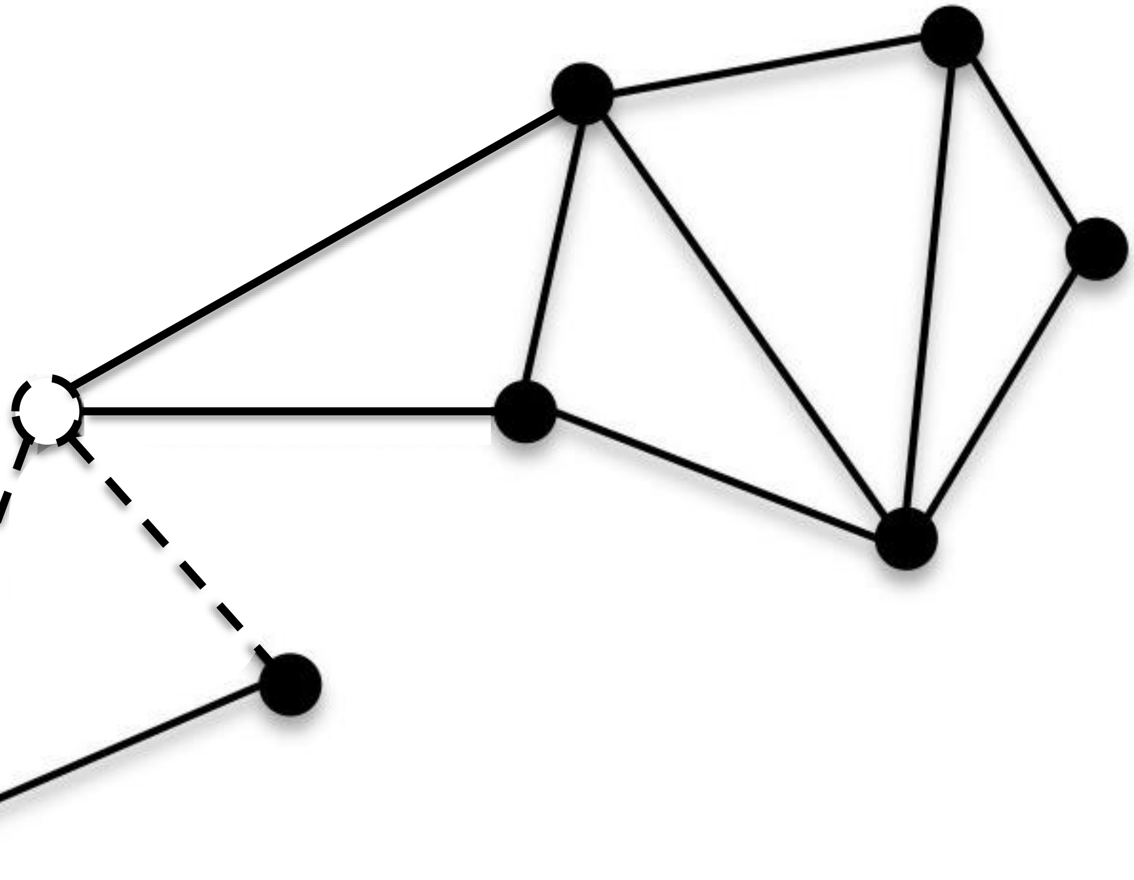
$$\kappa(G) = 1$$



Connectivity

Example: Vertex deletion on undirected graphs

$\kappa(G) = 1$



Connected component

Valid for both undirected and directed, «but»…

- In cases where a graph is not completely connected, it can be divided into **connected components**.

- A connected component is a **subset of vertices** within the graph such that there is a path between any pair of vertices within the same component, but there are no paths between vertices in different components.

Bridges and articulation points

Undirected graphs

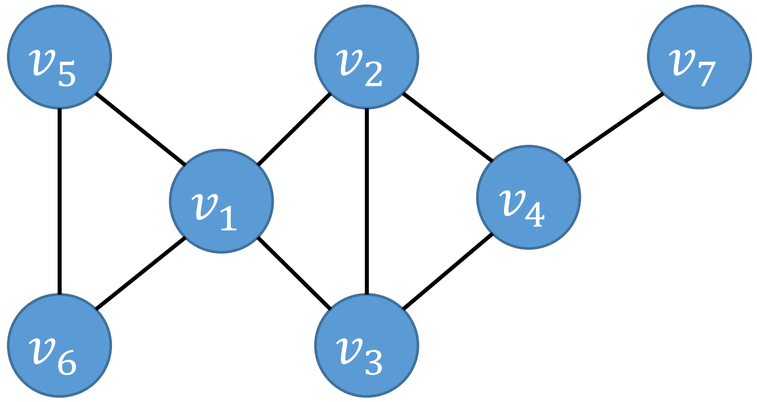
- An **articulation point** (*punto di articolazione*) is a vertex whose removal disconnects a component of the graph
  - In other words, an edge whose removal increases the number of connected components in the graph

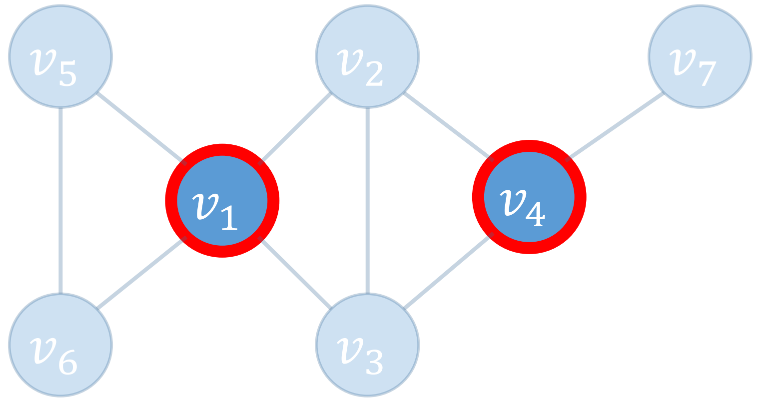
- A **bridge** (*ponte*) is an edge whose removal disconnects a component of the graph
  - In other words, an edge whose removal increases the number of connected components in the graph

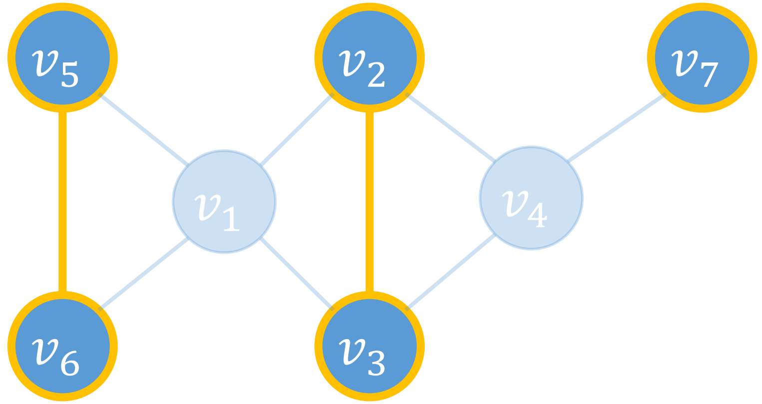
Articulation point

Example on an undirected graph

- Articulation points?
  - v_1, v_4
- Graph components?
  - $v_5, v_6, \{v_5, v_6\}$
  - $v_2, v_3, \{v_2, v_3\}$
  - v_7

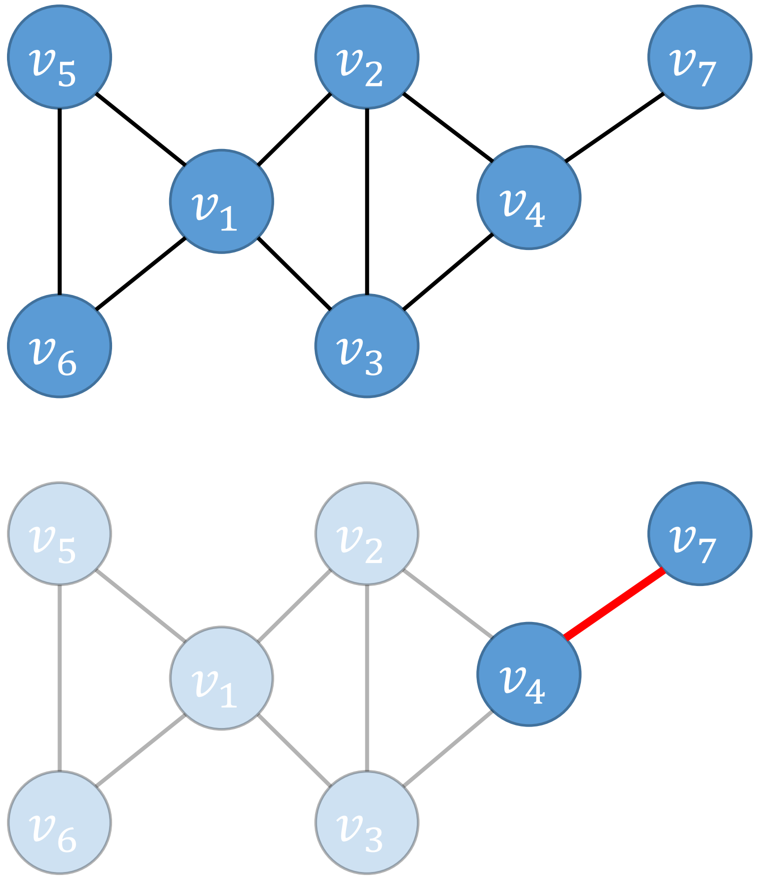






Bridge
Example on an undirected graph

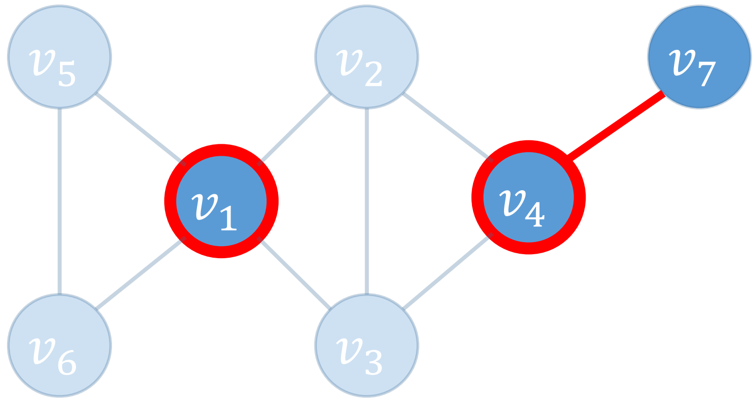
- Is there any bridge in the graph?
  - $\{v_4, v_7\}$



Biconnectivity
Undirected graphs

- When $\kappa(G) = 2$
  - No single edge or vertex removal disconnects the graph
  - No network failure points compromise the network itself

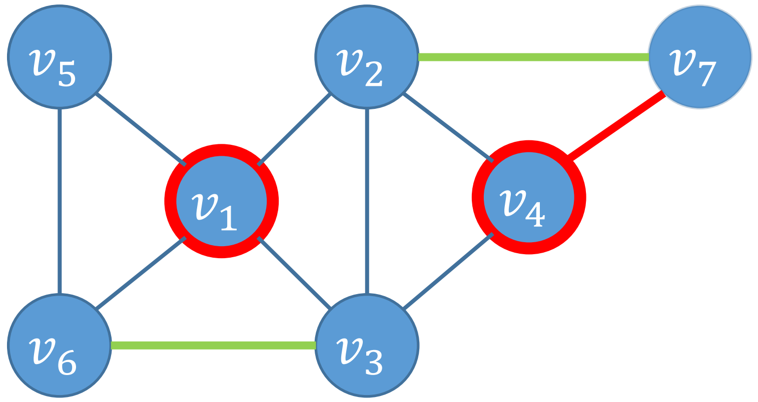
- The graph of the example is NOT biconnected:
  - A bridge: $\{v_4, v_7\}$
  - Two articulation points: v_1, v_4



Obtaining biconnectivity

Example on an undirected graph

- To resolve the articulation point v_1
  - $\{v_3, v_6\}$
- Resolves the articulation point v_4
- Alternative to the bridge $\{v_4, v_7\}$
  - $\{v_2, v_7\}$



Connectivity

Directed graphs

- The concept of **connectivity** is more nuanced because edges have directions
  - Two vertices are said to be connected if there exists a directed path from one vertex to the other

Strongly Connected Component (SCC)

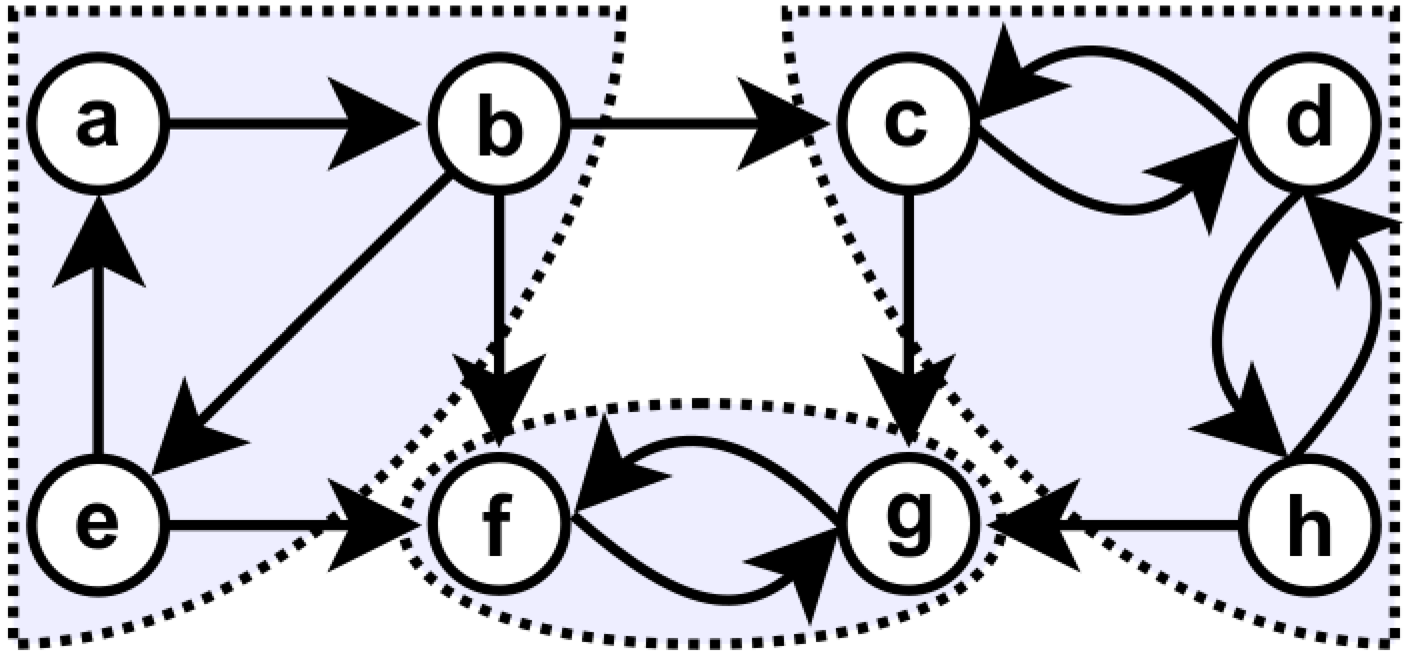
Directed graphs

- **Strongly Connected Components** (SCCs) are a key concept in directed graphs
  - An SCC is a subset of vertices in which every vertex is reachable from every other vertex by following directed paths

- A directed graph can have **multiple SCCs**
  - The graph as a whole is strongly connected if it consists of a single SCC
  - If a directed graph has more than one SCC, it is considered weakly connected if there is an undirected path between any pair of vertices when ignoring edge directions

Strongly Connected Component (SCC)

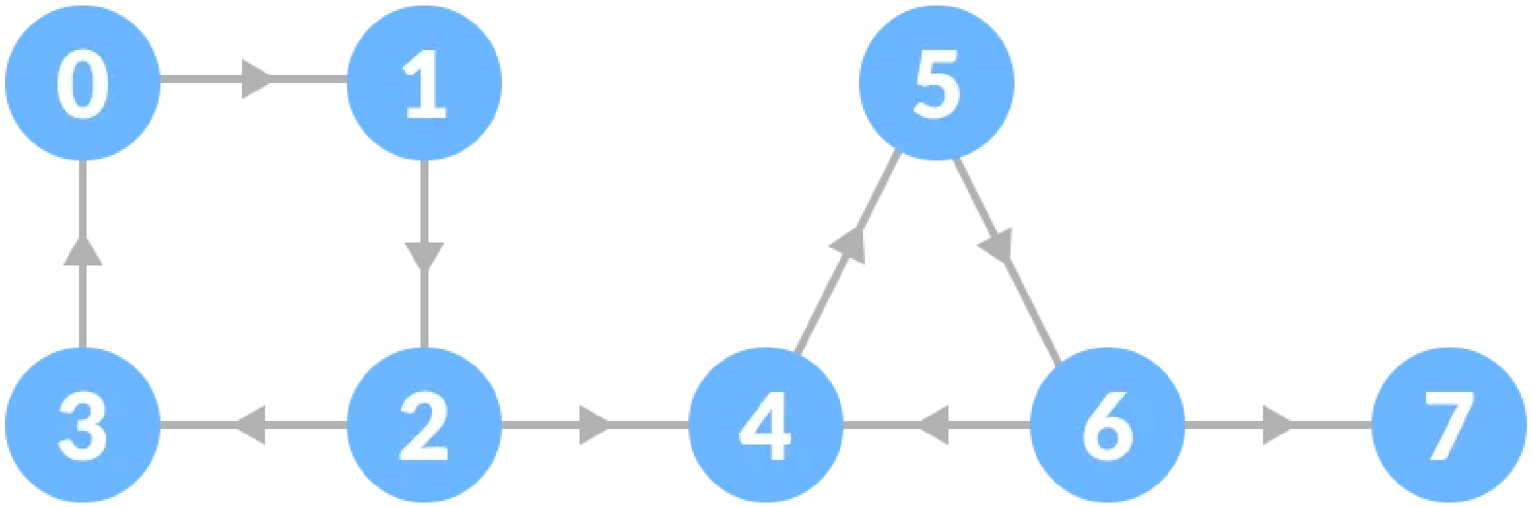
An example on directed graphs



Strongly Connected Component (SCC)

An exercise on directed graphs

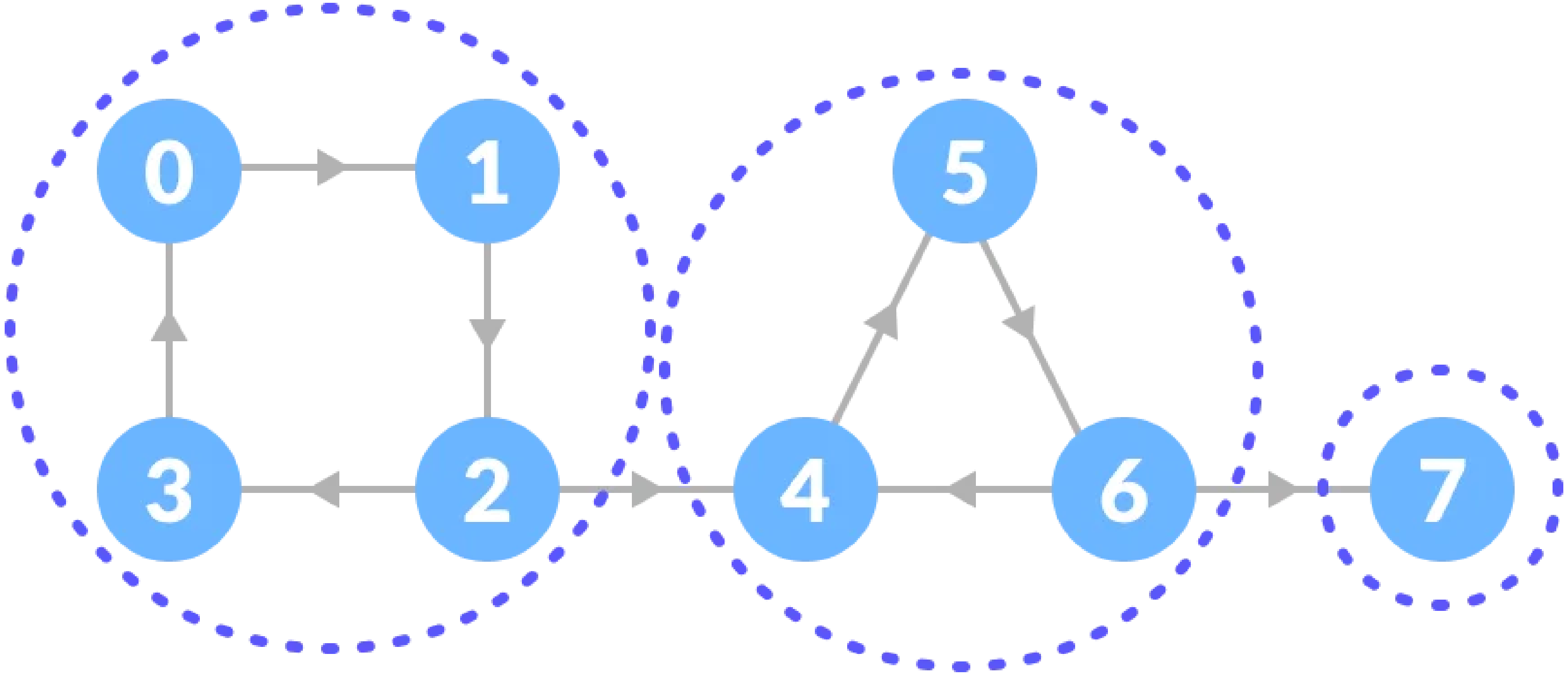
- How many SCCs?



Strongly Connected Component (SCC)

An exercise on directed graphs

- How many SCCs? → **3**



Disconnected graph
Directed graphs

- A directed graph is **disconnected** if:
  - There are subsets of vertices (SCCs) where you can reach any vertex from any other vertex within the same SCC.
  - However, there are no paths that allow you to traverse from one SCC to another. This lack of connectivity between SCCs makes the graph disconnected.

Articulation points and bridges
Directed graphs

- **Articulation points** (or vertices)
  - Often referred to as “cut vertices”
  - These are vertices whose removal would increase the number of SCCs in the graph
  - Articulation vertices are crucial for understanding the resilience and control of directed networks

- **Bridges**
  - Edges whose removal would increase the number of strongly connected components (SCCs) in the graph
  - Removing a bridge can break an existing SCC into smaller SCCs
  - Identifying bridges in directed graphs can help understand the stability and flow of information or processes in systems represented by these graphs

Insights into graph theory