CSCI 347 Data Mining

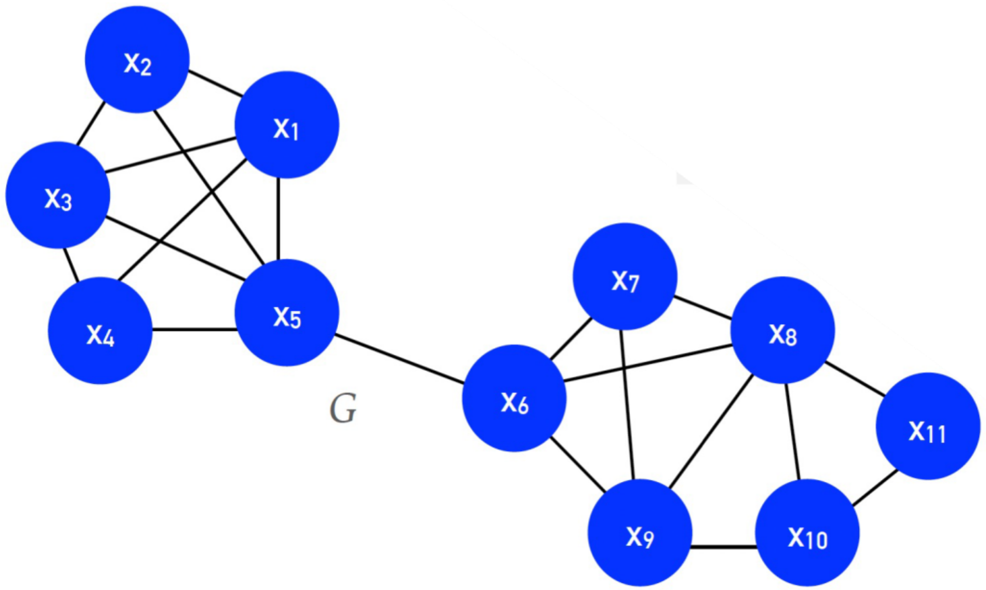
Graph Data

Graph Data

- Data instances are often not entirely independent
- they can be interconnected through various types of relationships.
- Graph data or networks are a data structure where instances are depicted as nodes, and the connections between these instances are represented by edges.

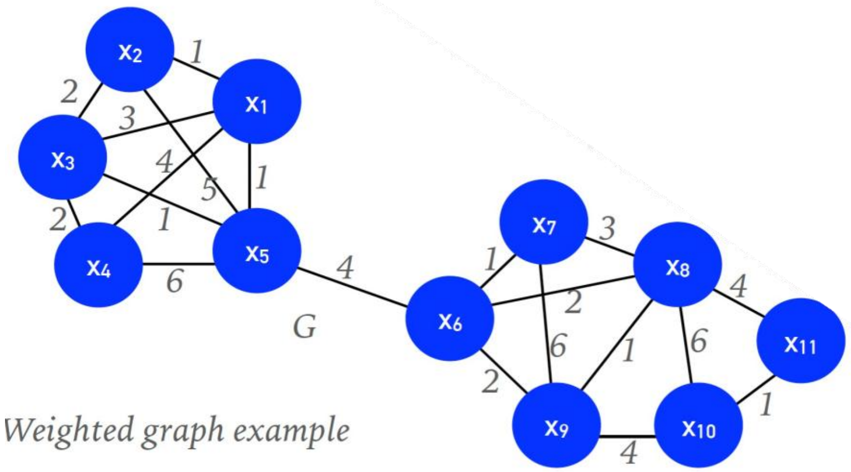
Graph Data

$$G = (V, E)$$
$$V = set\ of\ vertices$$
$$E \subseteq \{\{u, v\} : u, v \in V\}$$



Graph Data (Weighted graph)

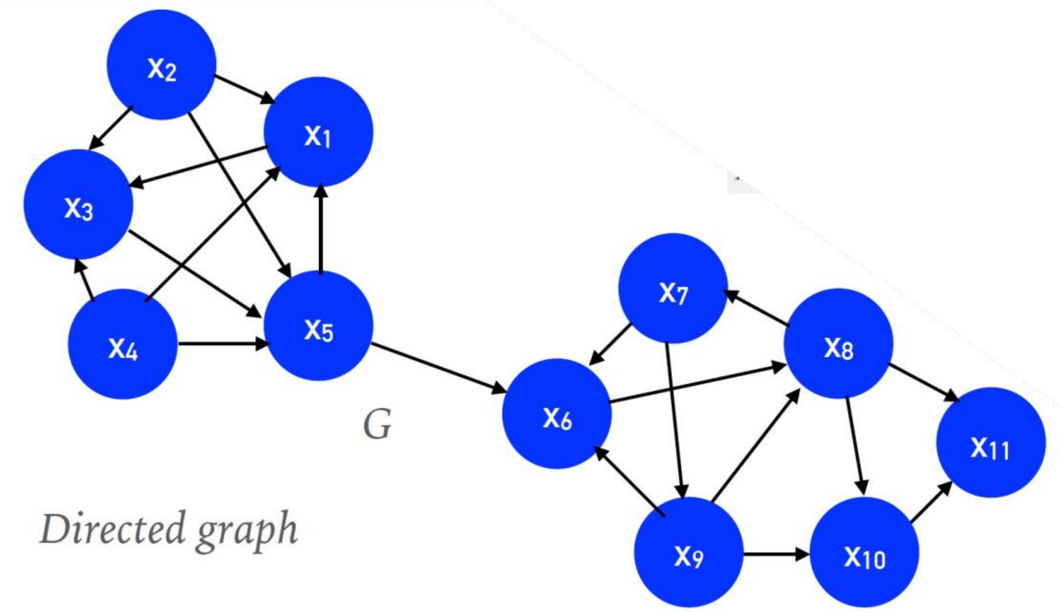
- $G = (V, E, w)$
- V = Vertices or Nodes
- E = Unordered pairs of vertices with weights (w_{ij})
- $w : E \rightarrow \mathbb{R}^{+}$ (usually positive real values)



Weighted graph example

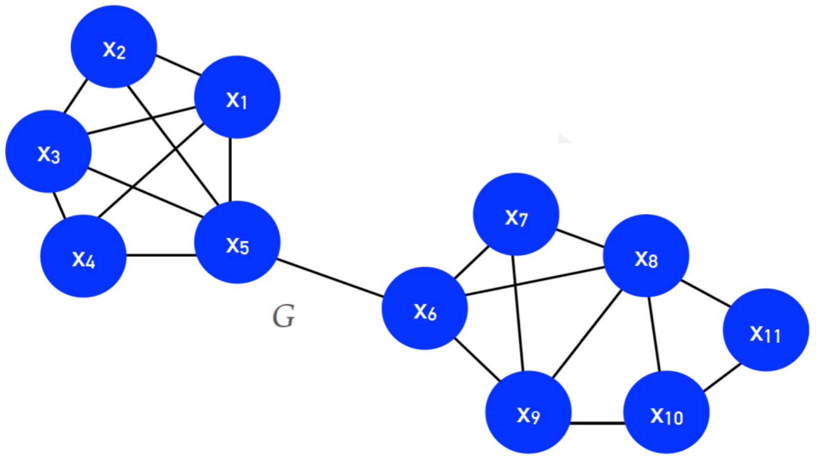
Graph Data (Directed Graph)

- $G = (V, E)$
- $V = Vertices\ or\ Nodes$
- $E \subseteq V \times V$ is the **ordered** pairs of vertices.



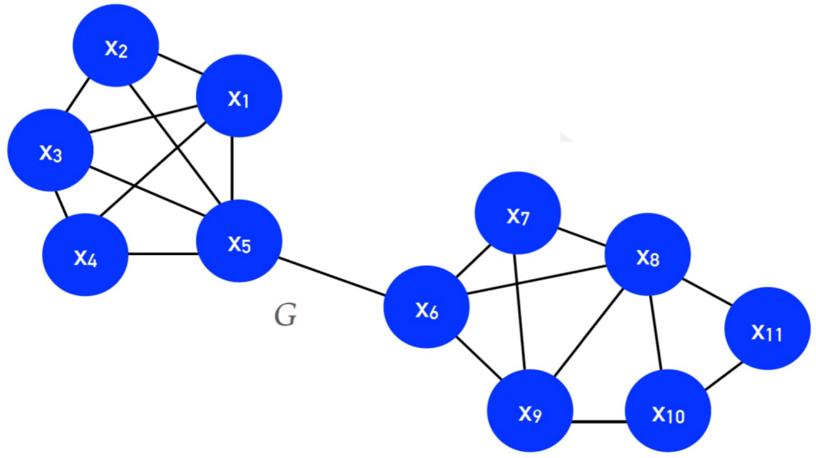
Graph Data

- $G = (V, E)$
- $V = Vertices\ or\ Nodes$
- $E = Unordered\ pairs\ of\ vertices$
- Simple graph = Undirected graph without loops
- Edge, $e = (v_i, v_j)$, $v_i\ and\ v_j$ are adjacent or neighbors.
- Order: $|V| = n$, Size: $|E| = m$



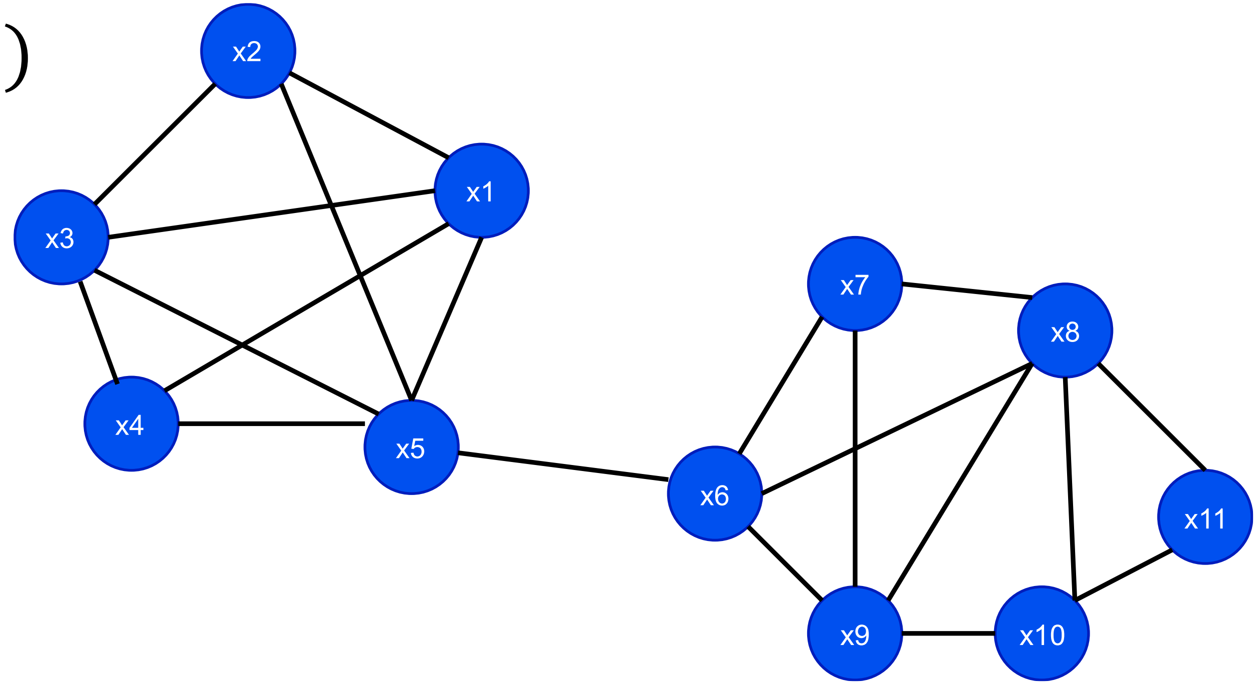
Graph Data

- $G = (V, E)$
- $V = Vertices\ or\ Nodes$
- $E = Unordered\ pairs\ of\ vertices$
- Simple graph = Undirected graph without loops
- Edge, $e = (v_i, v_j)$, $v_i\ and\ v_j$ are adjacent or neighbors.
- Order: $|V| = n$, Size: $|E| = m$
- A graph $H = (V_H, E_H)$ is called a subgraph of $G = (V, E)$, if $V_H \subseteq V$ and $E_H \subseteq E$ and the endpoints of edges in E_H is in V_H.



Degree of a node

- The degree of a node $v_i \in V$ is the number of edges incident with it and is denoted as $d(v_i)$ or just d_i .

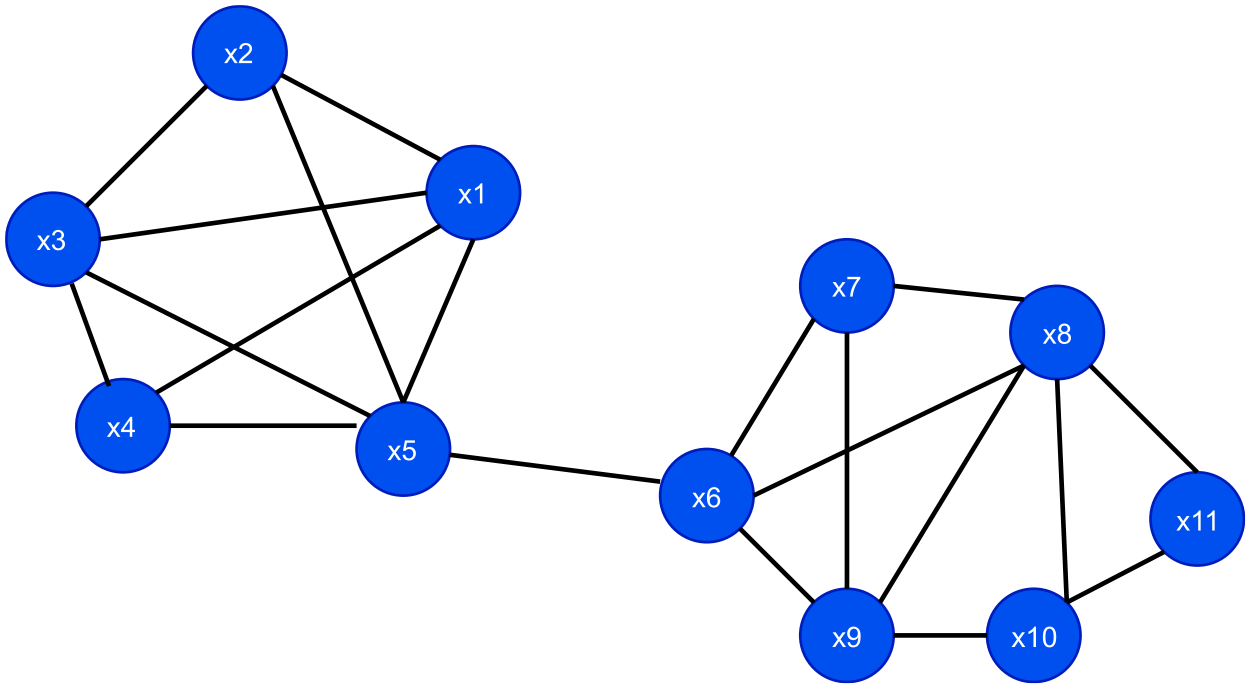


Degree of a node

- The degree of a node $v_i \in V$ is the number of edges incident with it and is denoted as $d(v_i)$ or just d_i .

What is the degree of $x9$?

1. 3
2. 1
3. 4
4. 8



Degree Distribution

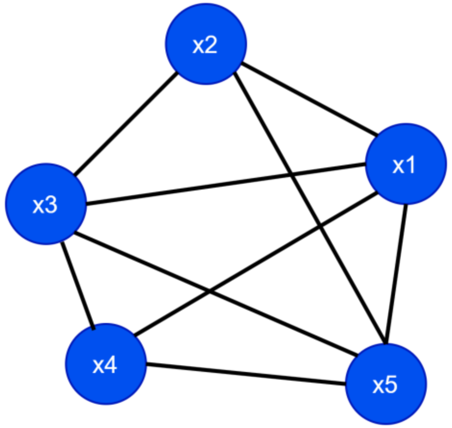
- Let N_k denote the number of vertices with degree k. The degree frequency distribution of a graph is given as $(N_0, N_1, \ldots, N_t)$
  - t is the maximum degree of a node in the graph.

Degree Distribution

- Let N_k denote the number of vertices with degree k. The degree frequency distribution of a graph is given as $(N_0, N_1, \dots, N_t)$
  - t is the maximum degree of a node in the graph.

What is the degree distribution of this graph?

1. $(4, 3, 4, 3, 4)$
2. $(3, 3, 4, 4, 4)$
3. $(0, 0, 0, 2, 3)$
4. $(2, 3)$



Degree Distribution

- The probability that a given node is of degree k is $\frac{N_k}{n}$.
- Suppose you have a random process of picking a node in a graph, and random variable X that assigns the degree of the picked node.

$$P(X = k) = \frac{N_k}{n}, n \text{ is the number of nodes}$$

Note: A random variable is a measurable function that assigns numerical values to outcomes of a random experiment.

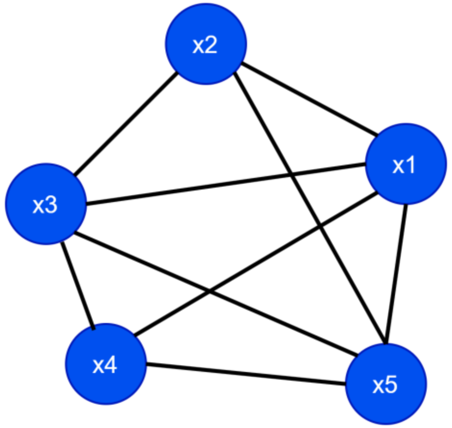
$$X : \Omega \to \mathbb{R}, \qquad \Omega \text{ is the sample space}$$

Degree Distribution

- The probability that a given node is of degree k is $\frac{N_k}{n}$.
- Suppose you have a random process of picking a node in a graph, and random variable X that assigns the degree of the picked node.

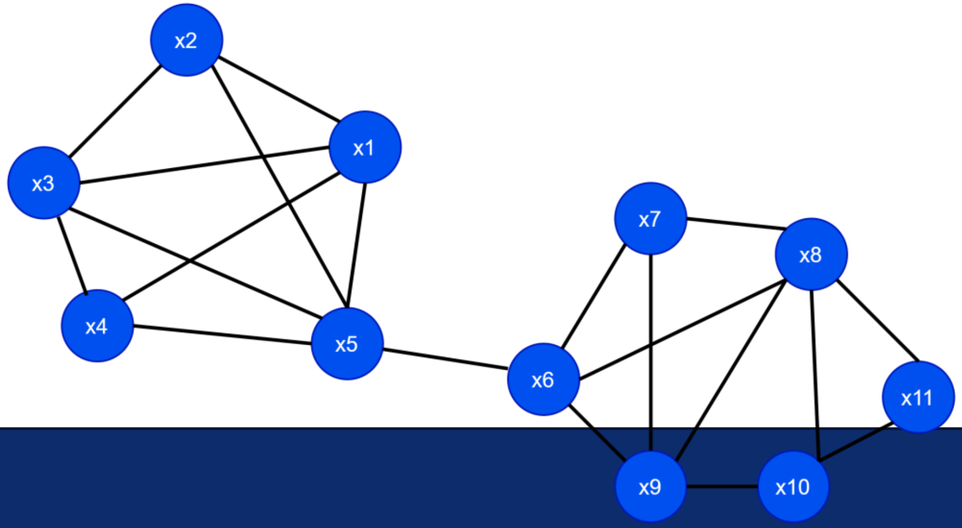
$$P(X = k) = \frac{N_k}{n}$$

- Given the node distribution $(0,0,0,2,3)$ what is the probability that a node is of degree 3?
    1. 0
    2. 2
    3. 3/5
    4. 3
    5. 2/5
    6. None of the above



Walk, Path, shortest path

- A walk in a graph G between nodes x and y is an ordered sequence of vertices, starting at x and ending at y.
  $Walk := < v_0, v_1, \ldots, v_t >, v_0 = x, v_t = y, \forall i \in [0..t-1]: (v_i, v_{i+1})\ exists$
- The length of the walk t, is the number of edges along the walk.
- A path is a walk with distinct vertices.
- A path of minimum length between nodes x and y is called a shortest path.

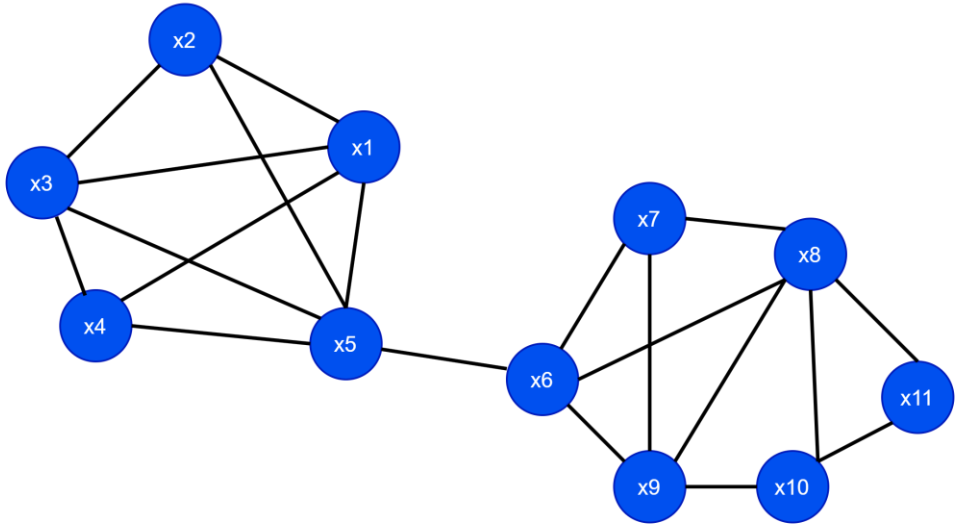


Walk, Path, shortest path

- A path of minimum length between nodes x and y is called a shortest path.

What is the length of the shortest path between $x2$ and $x10$?

1. 6
2. 3
3. 4
4. 1
5. 0

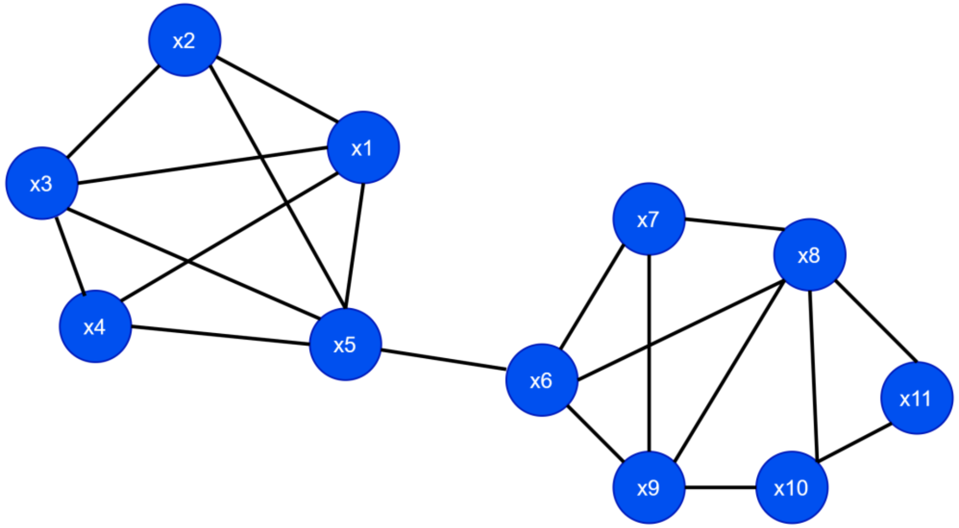


Connectedness

- Two nodes v_i and v_j are said to be connected if there exists a path between them.
- A graph is connected if there is a path between all pairs of vertices.
- A connected component, or just component, of a graph is a maximal connected subgraph.
  - **maximal** means that the subgraph cannot be extended any further while still maintaining the property of being connected.

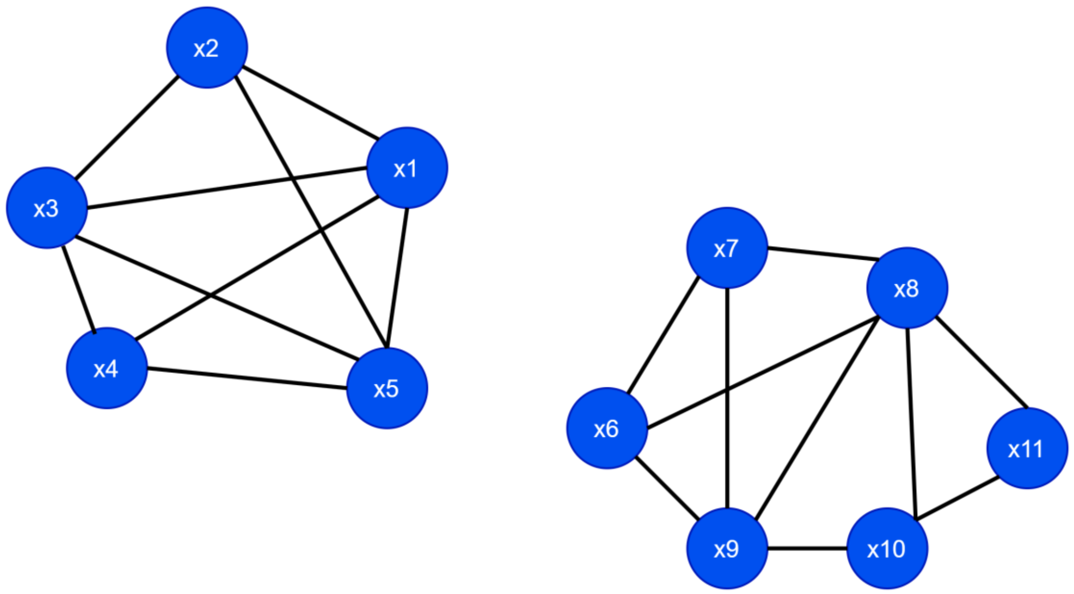
Connectedness

- Is this graph connected?
  - Yes
  - No



Connectedness

- Is this graph connected?
  - Yes
  - No



Adjacency Matrix

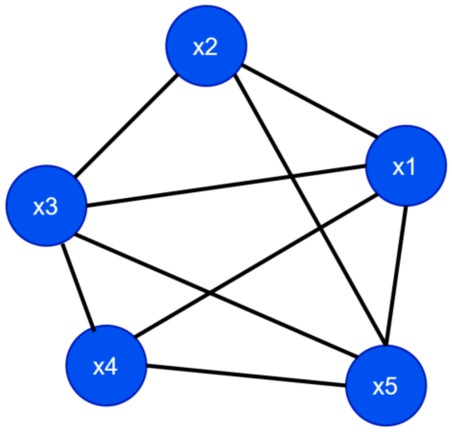
- A graph $G = (V, E)$, with $|V| = n$ vertices, can be conveniently represented in the form of an $n \times n$, symmetric binary adjacency matrix, A, defined as:

$$A(i,j) = \begin{cases} 1 & if\ v_i\ is\ adjacent\ to\ v_j \\ 0 & otherwise \end{cases}$$

- A weighted graph can be represented by $n \times n$ weighted adjacency matrix.

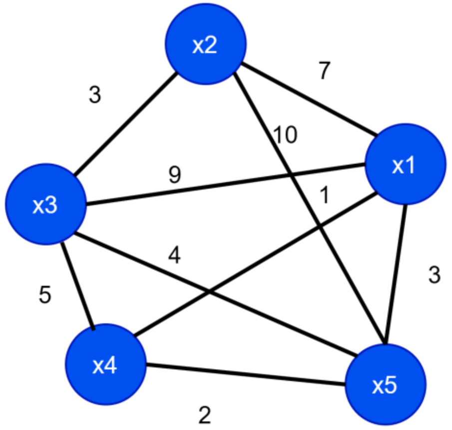
$$A(i,j) = \begin{cases} w_{ij} & if\ v_i\ is\ adjacent\ to\ v_j \\ 0 & otherwise \end{cases}$$

Adjacency matrix example.



$$
\begin{array}{c|ccccc}
 & x_1 & x_2 & x_3 & x_4 & x_5 \\
x_1 & 0 & 1 & 1 & 1 & 1 \\
x_2 & 1 & 0 & 1 & 0 & 1 \\
x_3 & 1 & 1 & 0 & 1 & 1 \\
x_4 & 1 & 0 & 1 & 0 & 1 \\
x_5 & 1 & 1 & 1 & 1 & 0
\end{array}
$$

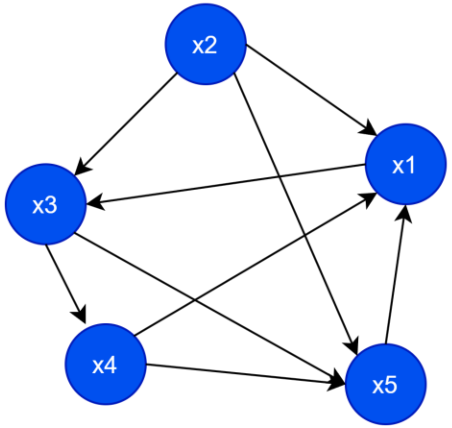
Adjacency matrix (weighted) example.



$$
\begin{array}{cccccc}
 & x_1 & x_2 & x_3 & x_4 & x_5 \\
x_1 & 0 & 7 & 9 & 1 & 3 \\
x_2 & 7 & 0 & 3 & 0 & 10 \\
x_3 & 9 & 3 & 0 & 3 & 4 \\
x_4 & 1 & 7 & 5 & 0 & 2 \\
x_5 & 3 & 10 & 4 & 2 & 0
\end{array}
$$

Adjacency matrix: directed graph

- In a directed graph adjacency matrix is not symmetric.

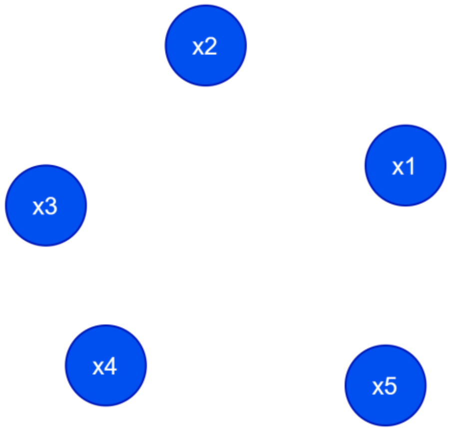


$$
\begin{array}{c|ccccc}
 & x_1 & x_2 & x_3 & x_4 & x_5 \\
x_1 & 0 & 0 & 1 & 0 & 0 \\
x_2 & 1 & 0 & 1 & 0 & 1 \\
x_3 & 0 & 0 & 0 & 1 & 1 \\
x_4 & 1 & 0 & 0 & 0 & 1 \\
x_5 & 1 & 0 & 0 & 0 & 0
\end{array}
$$

Graphs from Data Matrix

- Given a dataset in the form of a matrix, can we create a graph?

$$
\begin{array}{cccc}
 & X_1 & X_2 & X_3 \\
x_1 & 0.2 & 1 & 12.3 \\
x_2 & 1.3 & 4 & 89.23 \\
x_3 & 5.6 & 5 & 56.1 \\
x_4 & 4.5 & 7 & 47.3 \\
x_5 & 7.3 & 12 & 45.23
\end{array}
\xrightarrow{\quad ? \quad}
$$



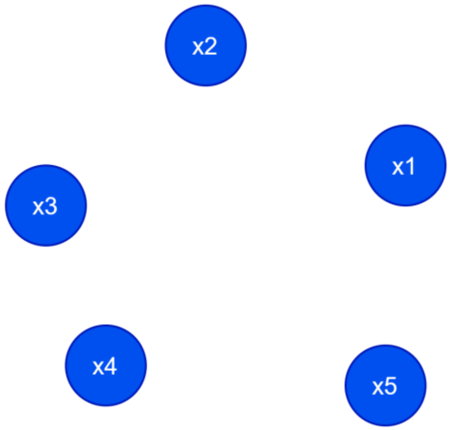
But what can we do about the edges?

Graphs from Data Matrix

- Given a dataset in the form of a matrix, can we create a graph?

	X_1	X_2	X_3
x_1	0.2	1	12.3
x_2	1.3	4	89.23
x_3	5.6	5	56.1
x_4	4.5	7	47.3
x_5	7.3	12	45.23

?



How about we using a similarity measure and then use the similarity measure as the edge weights?

How to create a graph from matrix?

- Define a weighted graph $G = (V, E)$.

$$V = \{v_i \mid v_i \text{ represents the entity } x_i\}$$

$$w_{ij} = sim(x_i, x_j)$$

$$sim(x_i, x_j)$$

$$\text{represents the similarity between points } x_i \text{ and } x_j$$

Gaussian similarity

$$w_{ij} = sim(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$$

$$\sigma \text{ is the spread parameter.}$$

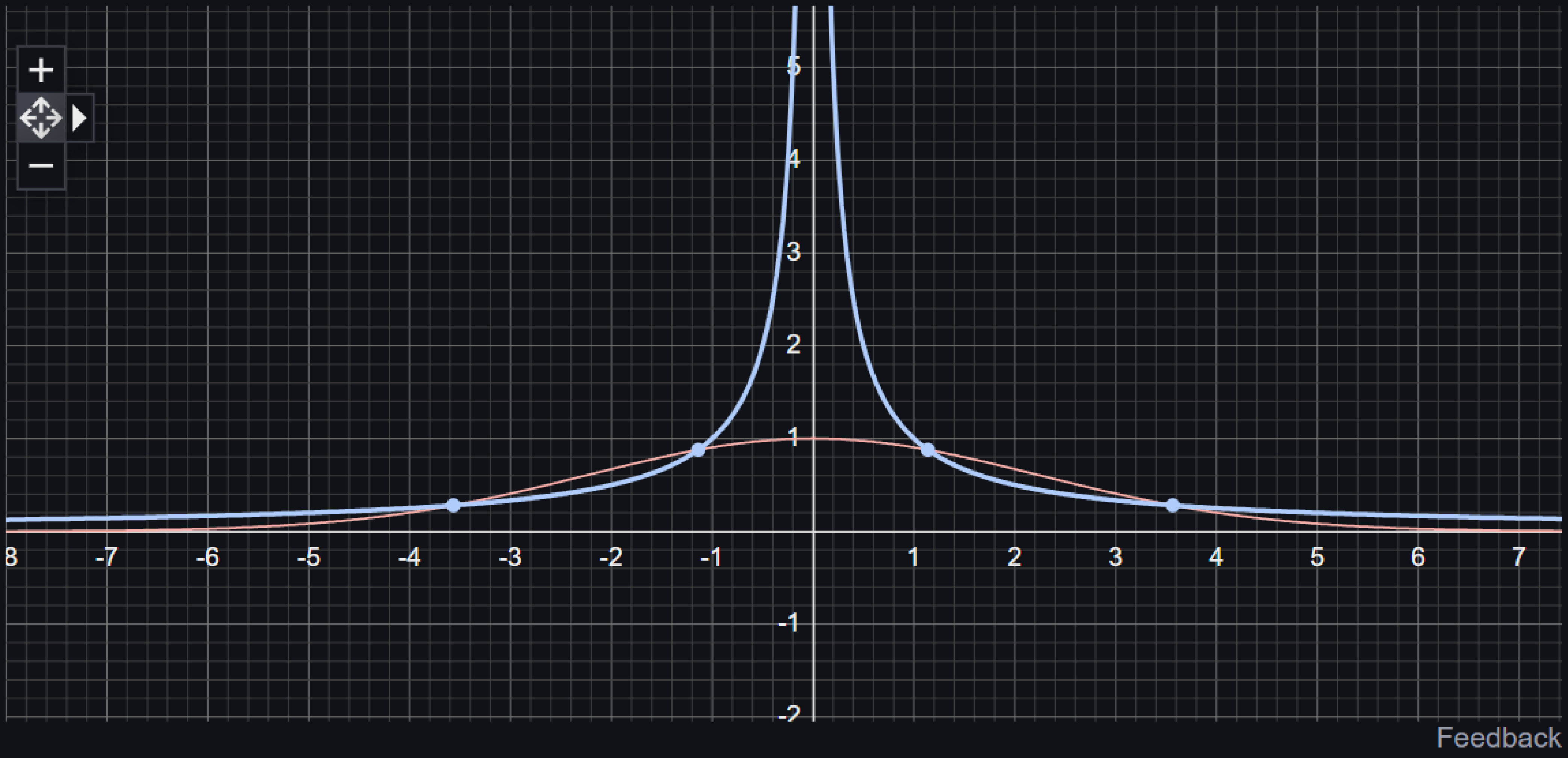
Gaussian similarity

- Similarity is defined as being inversely related to the Euclidean distance.
- If two vectors are far apart, then we say it's less similar.
  - Therefore, we can put lower weight between them.
- But why do we use this?
  - We can use something like $\frac{1}{\|x_i - x_j\|}$

Gaussian similarity

- Exponential Decay
  - The similarity measure w_{ij} decays smoothly and asymptotically to 0 as the distance increases, ensuring that distant points contribute very little but not abruptly.
  - It is bounded between 0 and 1, which simplifies interpretation and normalization in algorithms.
- Inverse distance
  - $\frac{1}{\|x_i - x_j\|}$ decays too slowly as the distance increases, leading to non-negligible contributions from distant points.
  - It has an unbounded range $(0, \infty)$, which can create numerical instability and make it harder to interpret.

Graph for $1/|x|$, $e^{-|x|^2/10}$



Gaussian similarity

- Handling zero distance:
  - When handling $\|x_i - x_j\| = 0$, w_{ij} simplifies to $e^0 = 1$, but if we use $\frac{1}{\|x_i - x_j\|}$, then mathematically this is not defined.
- Sensitivity control
  - The parameter σ allows you to control the sensitivity to distance.
    - Smaller σ, similarity decays quickly.
    - Larger σ, similarity decays slowly.
    - We can tweak the graph by changing this parameter.
  - $\frac{1}{\|x_i - x_j\|}$ does not have this property.

Gaussian similarity

- Robustness to outliers.
  - Gaussian similarity drops off quickly for large distances, effectively ignoring the outliers.
  - $\frac{1}{\|x_i - x_j\|}$, even if far away, can have disproportionately large effect to slow decay of $\frac{1}{d}$.

Graphs from Data Matrix

- Gaussian similarity with $\sigma = 25$

$$e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$$

$$\begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 1 & 12.3 \\ x_2 & 1.3 & 4 & 89.23 \\ x_3 & 5.6 & 5 & 56.1 \\ x_4 & 4.5 & 7 & 47.3 \\ x_5 & 7.3 & 12 & 45.23 \end{array}$$

	x1	x2	x3	x4	x5
x1	0	0.008709	0.207862	0.359302	0.366178
x2	0.008709	0	0.409967	0.241611	0.196716
x3	0.207862	0.409967	0	0.936019	0.87281
x4	0.359302	0.241611	0.936019	0	0.970737
x5	0.366178	0.196716	0.87281	0.970737	0

Graphs from Data Matrix

- Gaussian similarity with $\sigma = 50$

$$e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$$

	X_1	X_2	X_3
x_1	0.2	1	12.3
x_2	1.3	4	89.23
x_3	5.6	5	56.1
x_4	4.5	7	47.3
x_5	7.3	12	45.23

	x1	x2	x3	x4	x5
x1	0	0.30549	0.675218	0.774221	0.777899
x2	0.30549	0	0.800179	0.701098	0.665978
x3	0.675218	0.800179	0	0.983606	0.966562
x4	0.774221	0.701098	0.983606	0	0.992603
x5	0.777899	0.665978	0.966562	0.992603	0

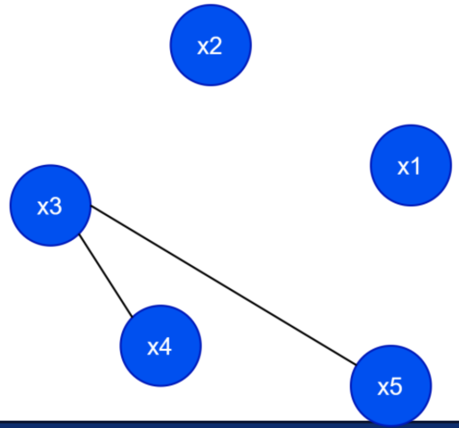
Creating a graph from matrix

$\tau = 0.94$

	X_1	X_2	X_3
x_1	0.2	1	12.3
x_2	1.3	4	89.23
x_3	5.6	5	56.1
x_4	4.5	7	47.3
x_5	7.3	12	45.23

	x1	x2	x3	x4	x5
x1	0	0.30549	0.675218	0.774221	0.777899
x2	0.30549	0	0.800179	0.701098	0.665978
x3	0.675218	0.800179	0	0.983606	0.966562
x4	0.774221	0.701098	0.983606	0	0.992603
x5	0.777899	0.665978	0.966562	0.992603	0

	x1	x2	x3	x4	x5
x1	0	0	0	0	0
x2	0	0	0	0	0
x3	0	0	0	1	1
x4	0	0	1	0	1
x5	0	0	1	1	0



$$sim(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$$

$$A(i,j) = \begin{cases} 1 & if\, sim(x_i, x_j) \geq \tau \\ 0 & otherwise \end{cases}$$

Iris Similarity Graph: Gaussian Similarity

$\sigma = \frac{1}{\sqrt{2}}$, edge exist if and only if $w_{ij} \geq 0.777$
Order: $|V| = n = 150$, size: $|E| = m = 753$

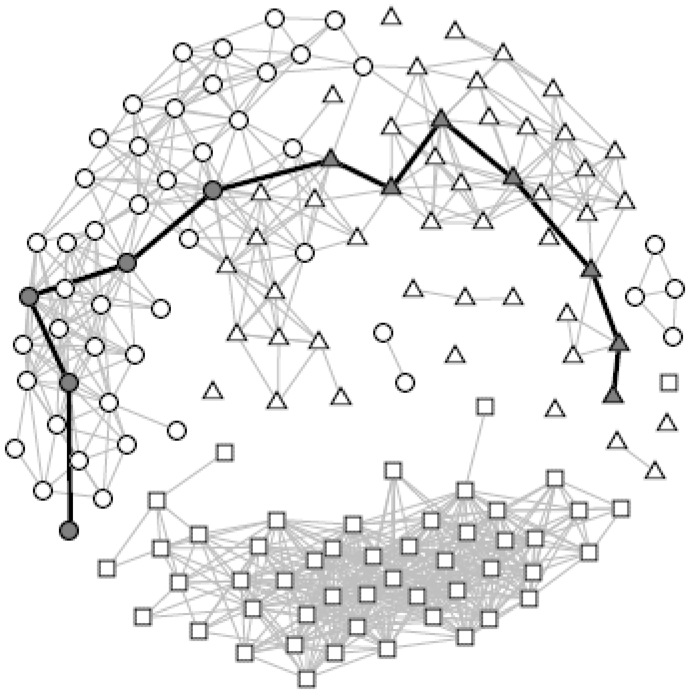


Figure 4.2: Iris Similarity Graph