CSCI 347 Data Mining

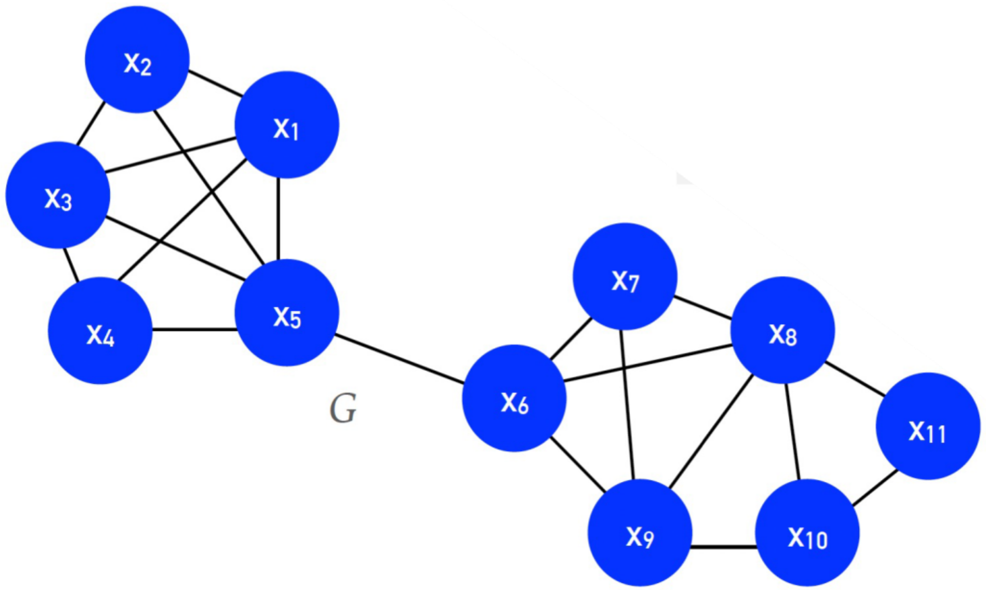
Graph Data

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.

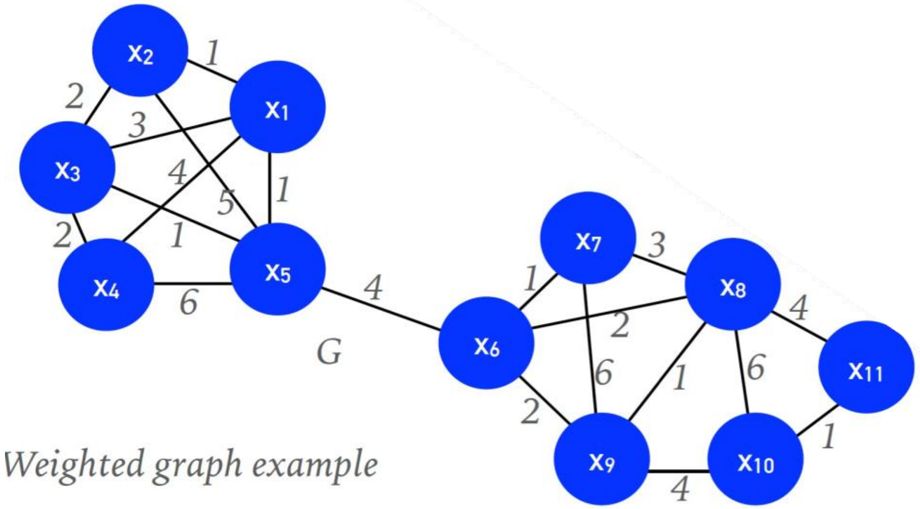
Graph Data

$$G = (V, E)$$
$$V = set\ of\ vertices$$
$$E \subseteq V \times V, is\ the\ set\ of\ vertices\ in\ the\ graph$$



Graph Data (Weighted graph)

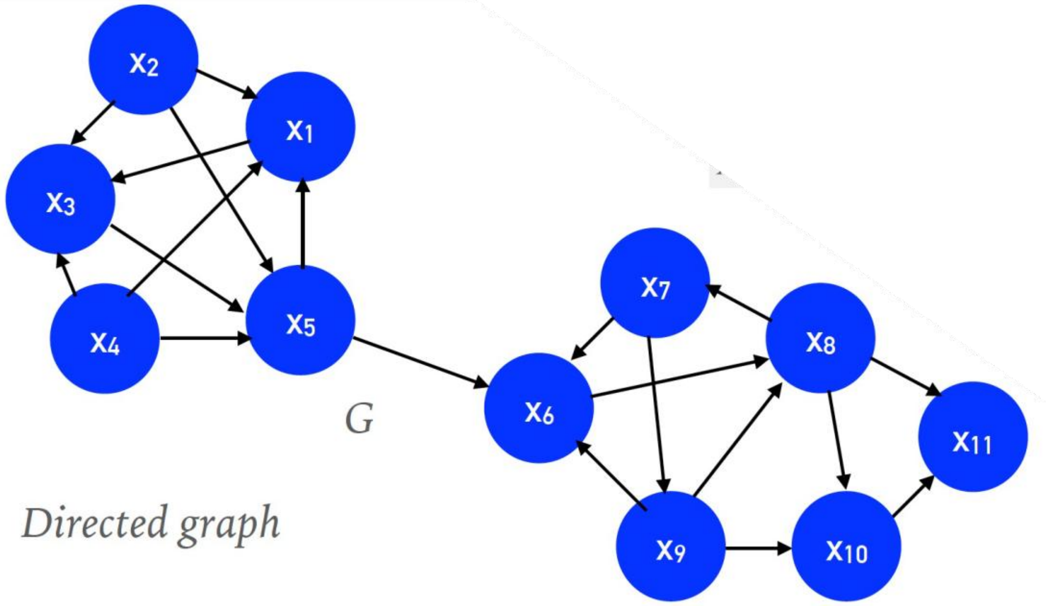
- G = (V, E)
- V = Vertices or Nodes
- E = Unordered pairs of vertices with different weights (w_{ij})



Weighted graph example

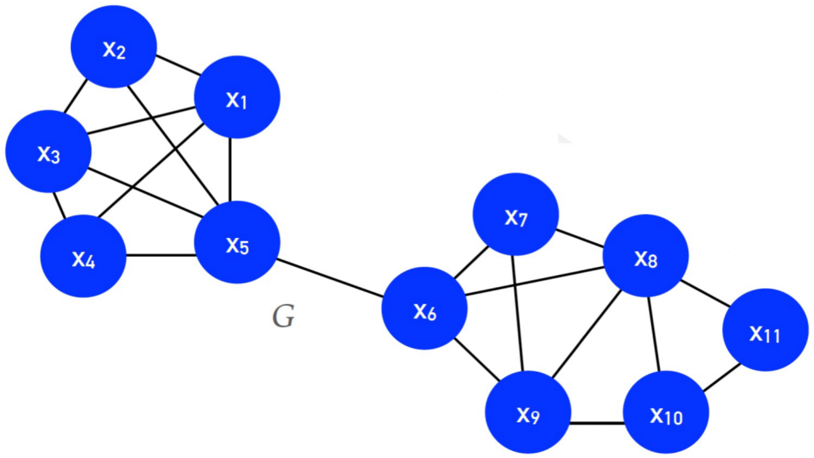
Graph Data (Directed Graph)

- G = (V, E)
- V = Vertices or Nodes
- E = **ordered** pairs of vertices.



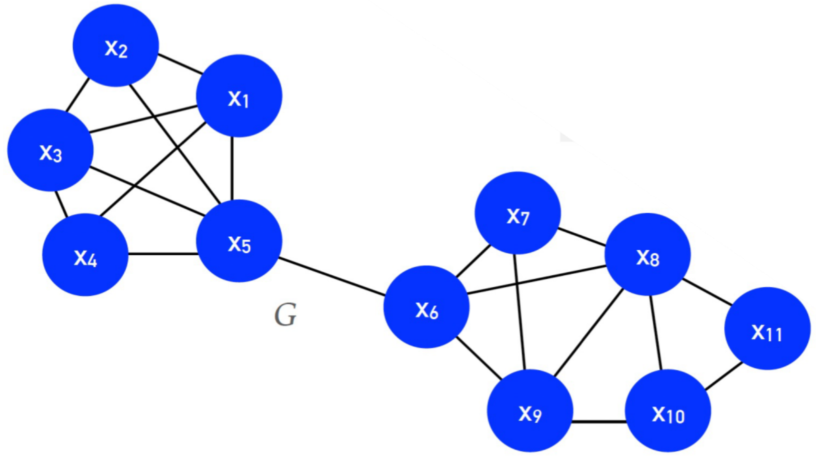
Graph Data

- $G = (V, E)$
- $V = Vertices\ or\ Nodes$
- $E = Unordered\ pairs\ of\ vertices$
- 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: $|E| = m$



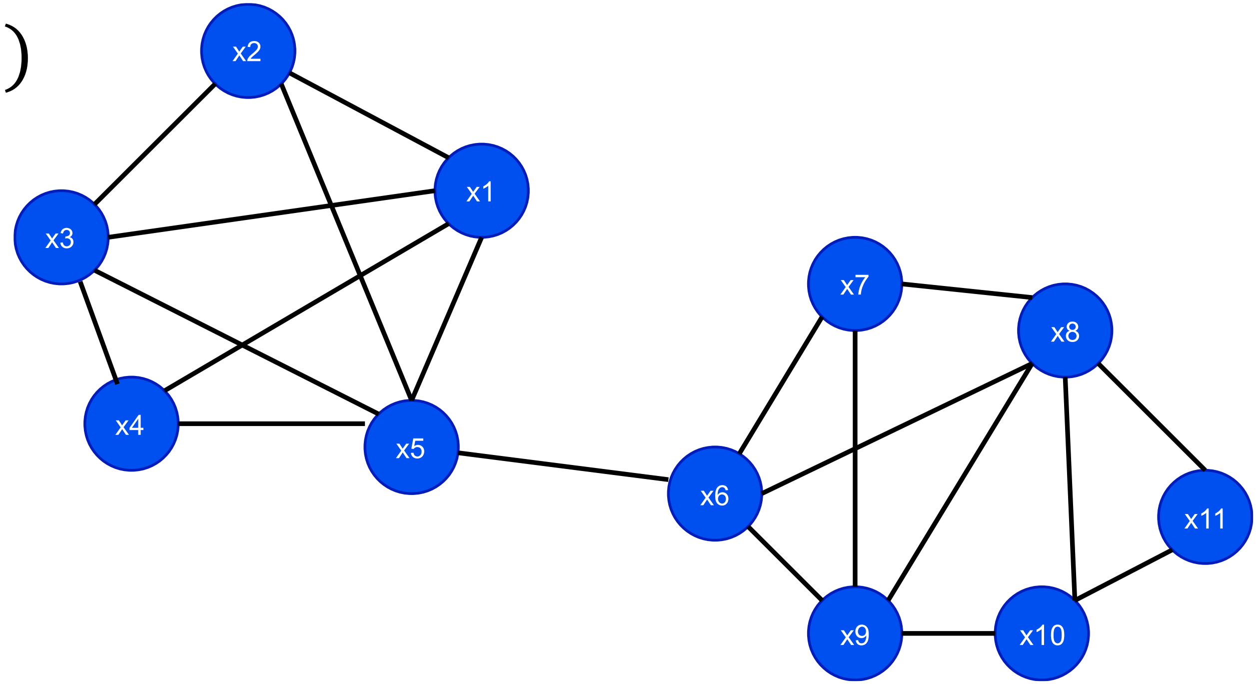
Graph Data

- $G = (V, E)$
- $V = Vertices\ or\ Nodes$
- $E = Unordered\ pairs\ of\ vertices$
- 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 $E_H \subseteq E$.



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 .

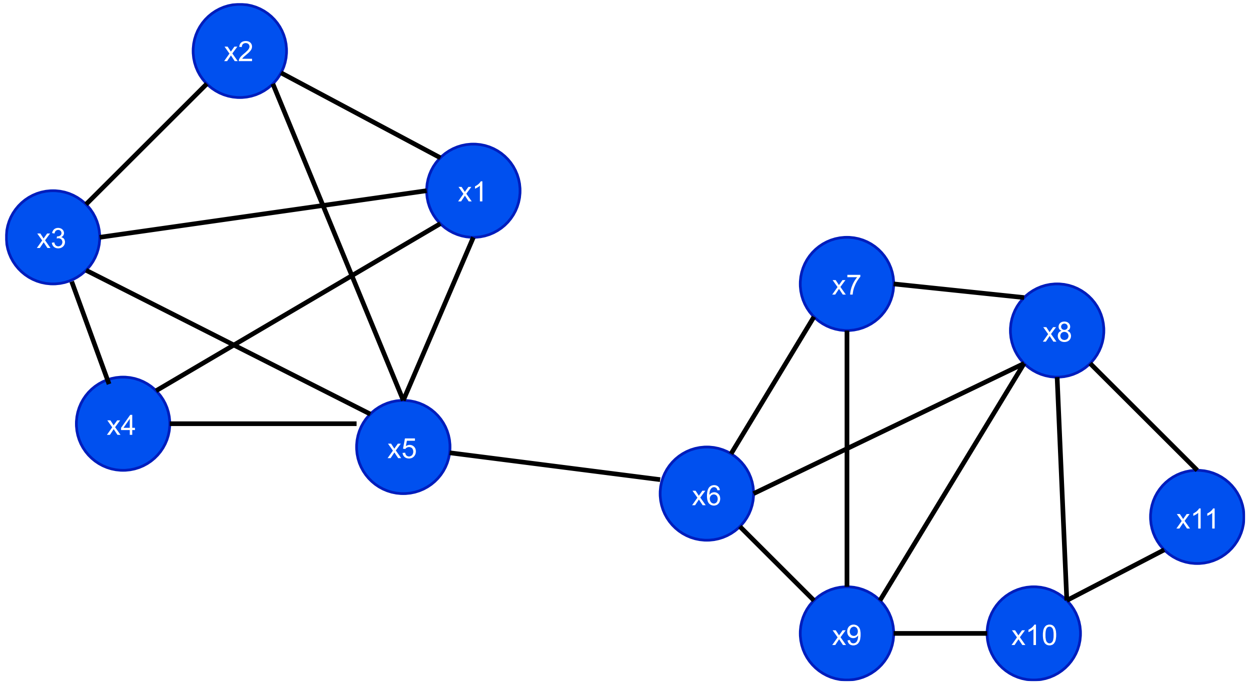


Degree of a node

- 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

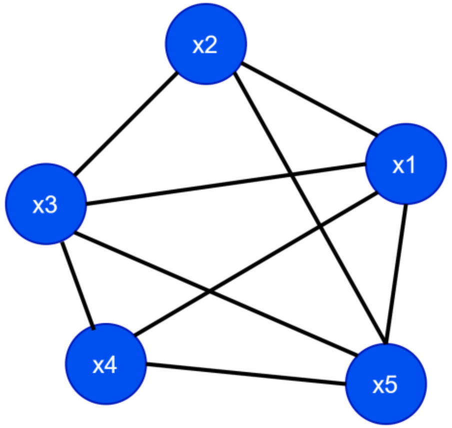
- Let N_k denote the number of vertices with degree k. The degree frequency distribution of a graph is given as $(N_0, N_1, \ldots, N_t)$
  - t is the maximum degree of a node in the graph.

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.

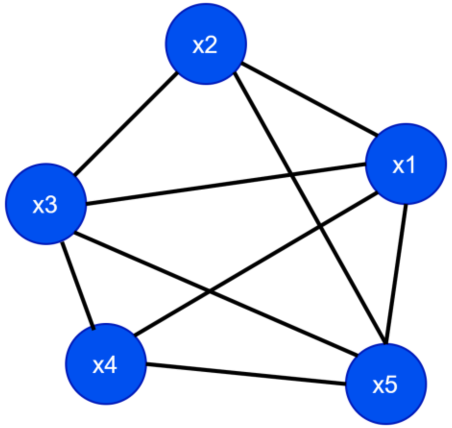
$$P(X = k) = \frac{N_k}{n}$$

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.

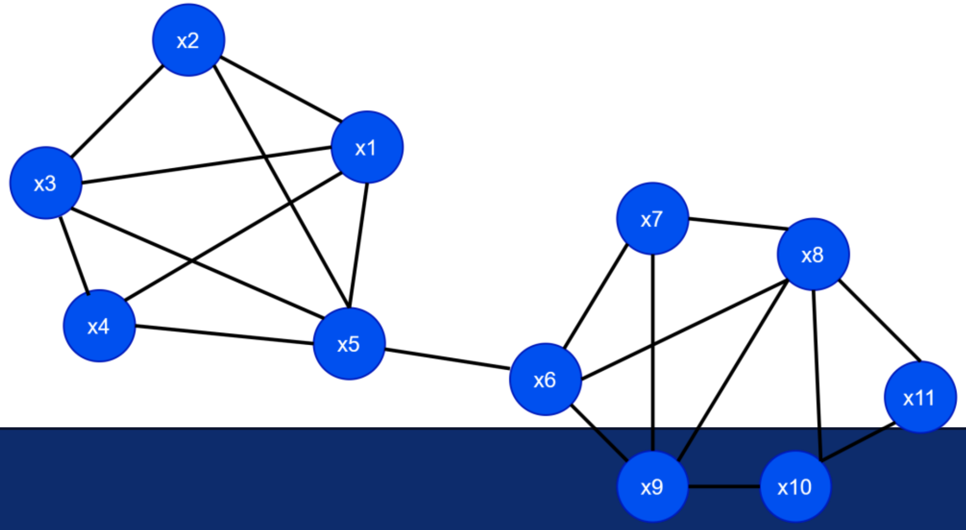
$$P(X = k) = \frac{N_k}{n}$$

- 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.

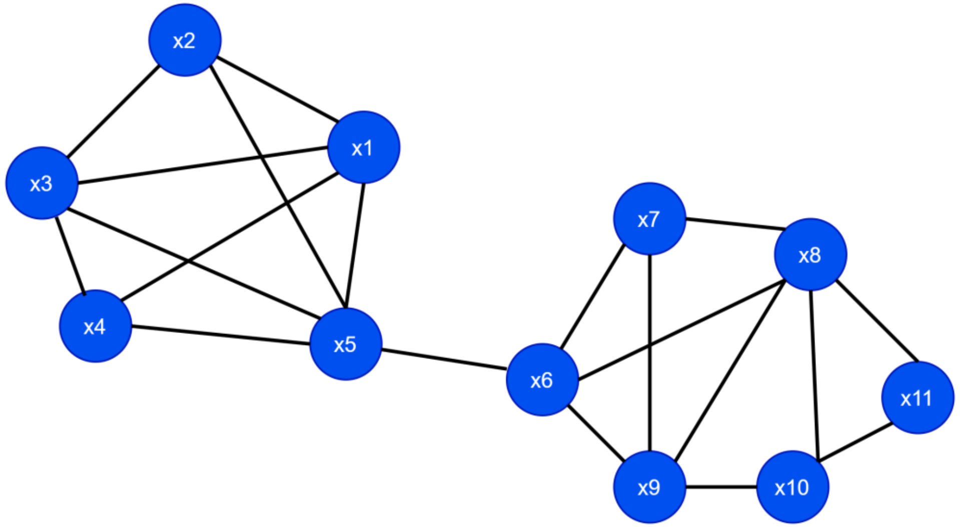


Walk, Path, shortest path

- 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

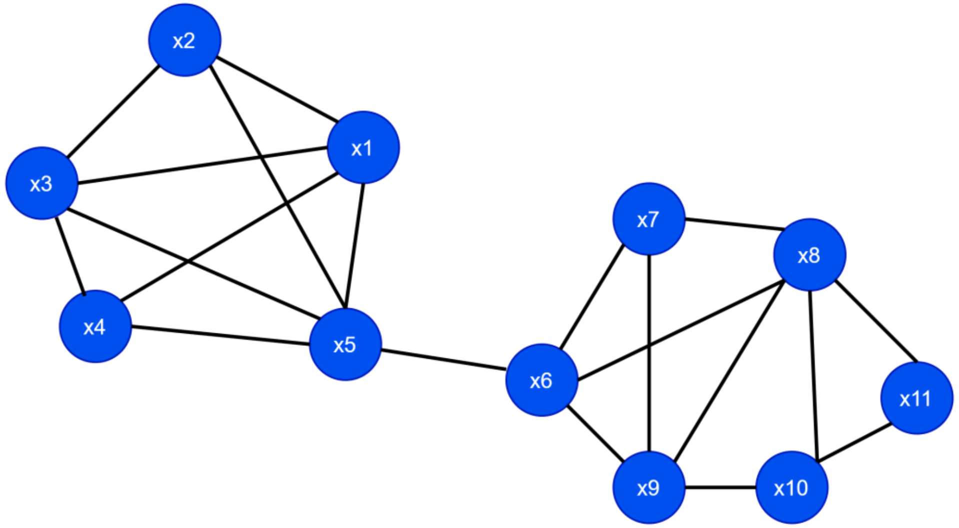


Connectedness

- 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.

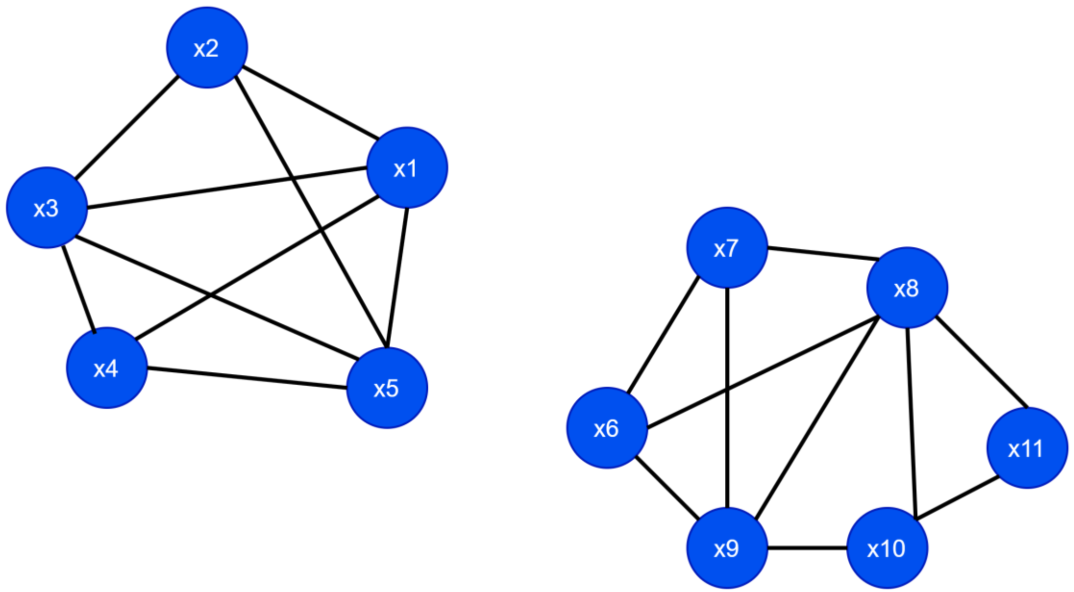
Connectedness

- Is this graph connected?
  - Yes
  - No



Connectedness

- Is this graph connected?
  - Yes
  - No



Adjacency Matrix

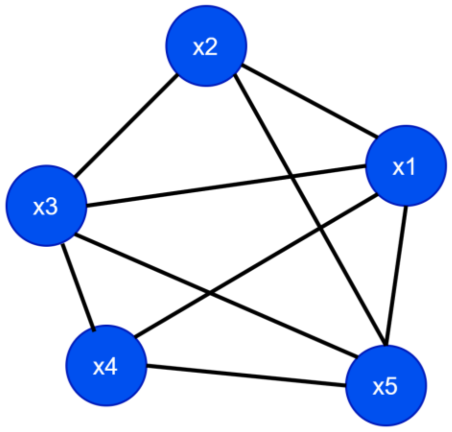
- A graph $G = (V, E)$, with $|V| = n$ vertices, can be conveniently represented in the form of an $n \times n$, symmetric binary adjacency matrix, A, defined as:

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

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

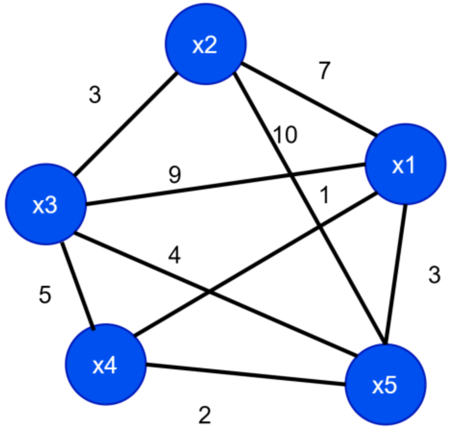
$$A(i,j) = \begin{cases} w_{ij} & if\ v_i\ is\ adjacent\ to\ v_j \\ 0 & otherwise \end{cases}$$

Adjacency matrix example.



$$
\begin{array}{cccccc}
 & x_1 & x_2 & x_3 & x_4 & x_5 \\
x_1 & 0 & 1 & 1 & 1 & 1 \\
x_2 & 1 & 0 & 1 & 0 & 1 \\
x_3 & 1 & 1 & 0 & 1 & 1 \\
x_4 & 1 & 0 & 1 & 0 & 1 \\
x_5 & 1 & 1 & 1 & 1 & 0
\end{array}
$$

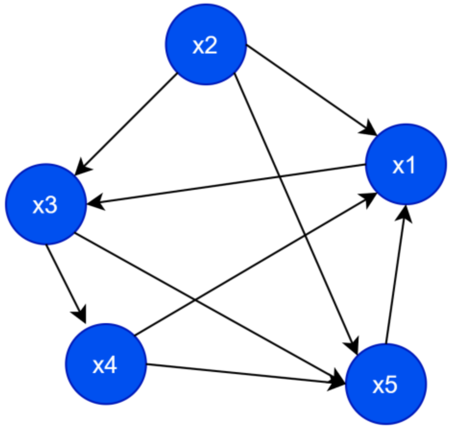
Adjacency matrix (weighted) example.



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

Adjacency matrix: directed graph

- In a directed graph adjacency matrix is not symmetric.



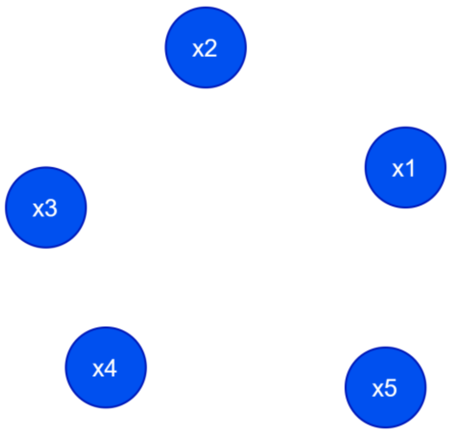
$$
\begin{array}{c|ccccc}
 & x_1 & x_2 & x_3 & x_4 & x_5 \\
x_1 & 0 & 0 & 1 & 0 & 0 \\
x_2 & 1 & 0 & 1 & 0 & 1 \\
x_3 & 0 & 0 & 0 & 1 & 1 \\
x_4 & 1 & 0 & 0 & 0 & 1 \\
x_5 & 1 & 0 & 0 & 0 & 0
\end{array}
$$

Graphs from Data Matrix

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

	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

?

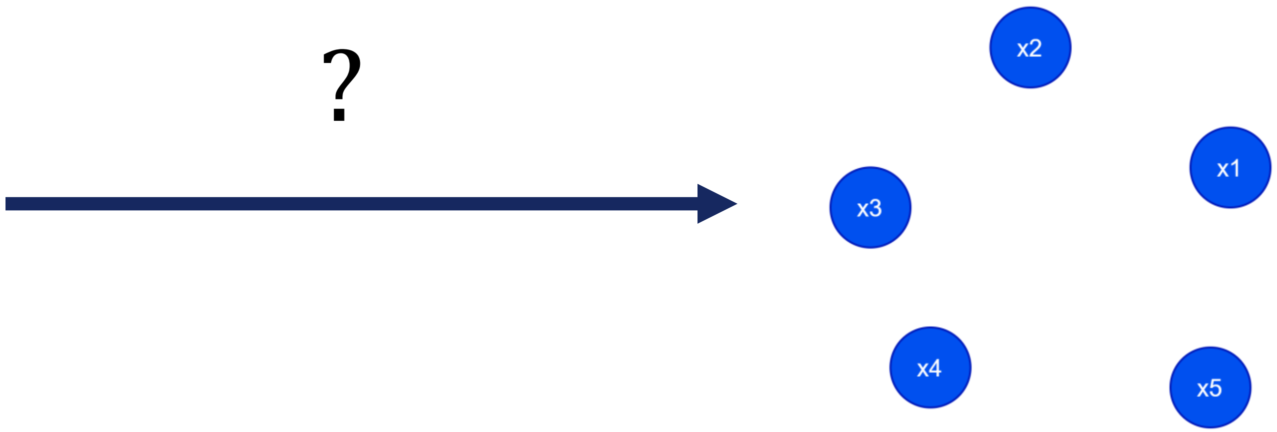


But what can we do about the edges?

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

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

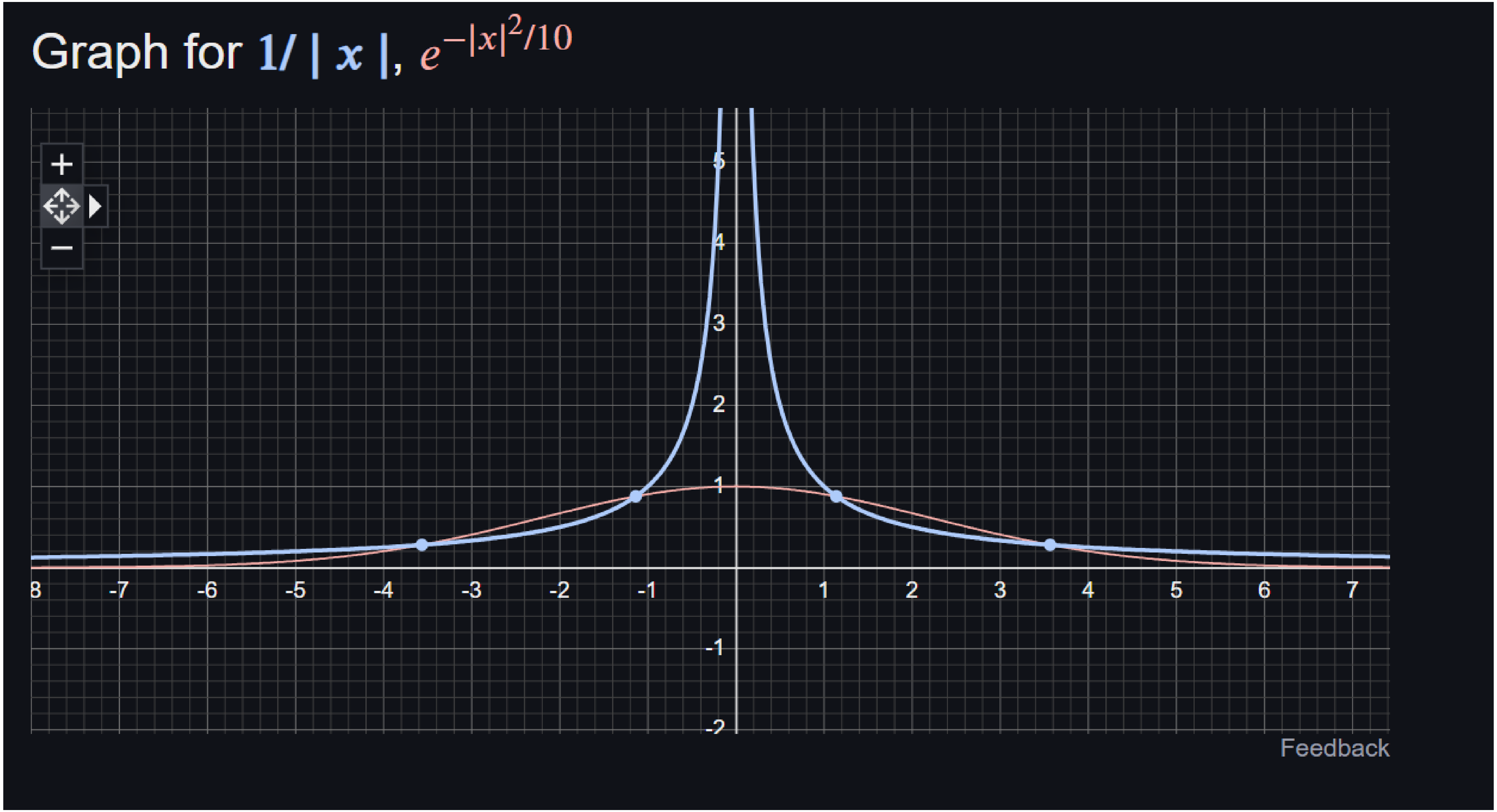
$$\sigma \text{ is the spread parameter.}$$

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 $1/|x|$, $e^{-|x|^2/10}$
+
−
5
4
3
2
1
-1
-2
8
-7
-6
-5
-4
-3
-2
-1
1
2
3
4
5
6
7
Feedback

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 $\sigma = 25$

	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

0	0.008709	0.207862	0.359302	0.366178
0.008709	0	0.409967	0.241611	0.196716
0.207862	0.409967	0	0.936019	0.87281
0.359302	0.241611	0.936019	0	0.970737
0.366178	0.196716	0.87281	0.970737	0

Graphs from Data Matrix

- Gaussian similarity with $\sigma = 25$

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

	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 $\sigma = 50$

	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

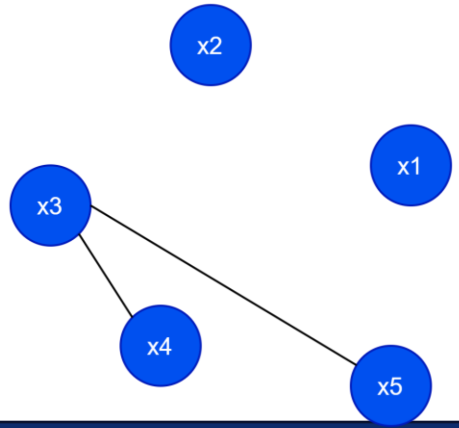
Creating a graph from matrix

$$\tau = 0.94$$

	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: $|E| = m = 753$

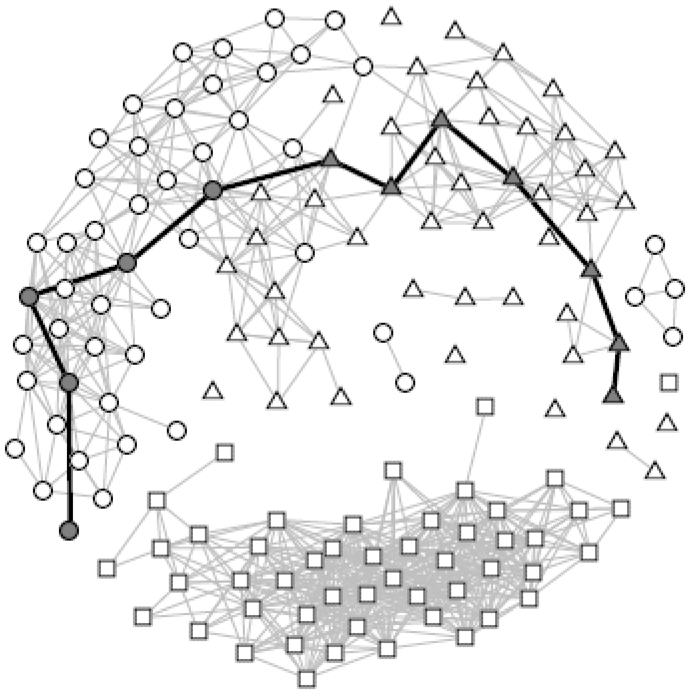


Figure 4.2: Iris Similarity Graph

Topological graph attributes

- **Topological graph attributes** refer to properties of a graph that describe its structure and connectivity without considering specific geometric or spatial embedding.
- We say a graph attribute is:
  - Local if they apply only to a single node.
  - Global if they refer to the entire graph.

- Degree
  - The degree of a node $v_i \in G$ is defined as:

$$d_i = \sum_j A_{ij}$$

  - Clearly, this is a local attribute.
  - The corresponding global attribute for the entire graph G is the average degree:

$$\mu_d = \frac{\sum_i d_i}{n}$$

- We can generalize this to weighted and directed graphs as well.
  - The in-degree of a node $v_i \in G$ is defined as:

$$id(v_i) = \sum_j A(j,i)$$

$$od(v_i) = \sum_j A(i,j)$$

  - The average indegree and average outdegree can be obtained by summing them up and dividing by n.

$$\mu_{indeg} = \frac{\sum_i id(v_i)}{n}, \mu_{outndeg} = \frac{\sum_i od(v_i)}{n}$$

Average path length

- The average path length, also called the characteristic path length, of a connected graph is given as:

$$\mu_L = \frac{\sum_i \sum_{j>i} d(v_i, v_j)}{\binom{n}{2}} = \frac{2}{n(n-1)} \sum_i \sum_{j>i} d(v_i, v_j)$$

- For directed graphs,

$$\mu_L = \frac{1}{n(n-1)} \sum_i \sum_j d(v_i, v_j)$$

For **disconnected graphs** , the average is taken over only the connected pairs of vertices.

Eccentricity

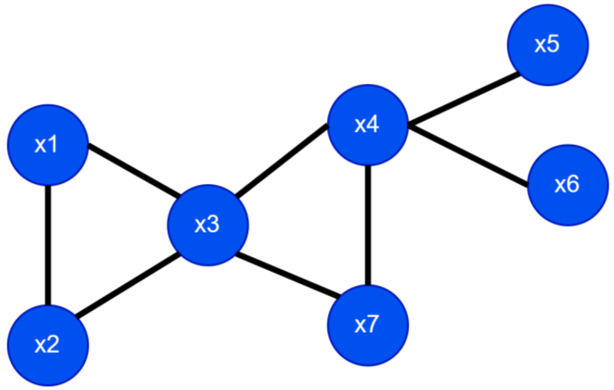
- The eccentricity of a node v_i is the maximum distance from v_i to any other node in the graph:

$$e(v_i) = \max_j \{d(v_i, v_j)\}$$

- The less eccentric a node is more central it is in the graph.

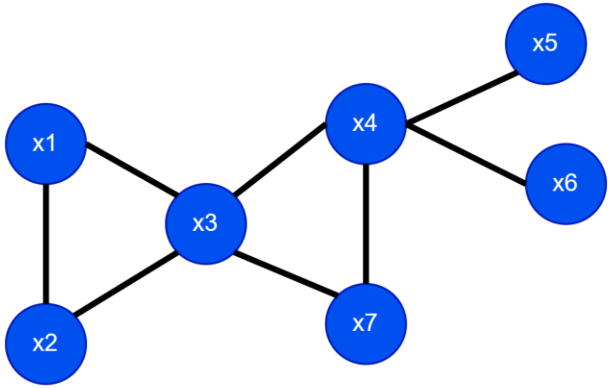
Example:

- What is the eccentricity of $x3$?
  A. 2
  B. 3
  C. 4
  D. 5



Example:

- What is the eccentricity of $x6$?
  A. 2
  B. 3
  C. 4
  D. 5



Radius and Diameter of a graph G

- The radius of a connected graph, denoted $r(G)$, is the minimum eccentricity of any node in the graph.

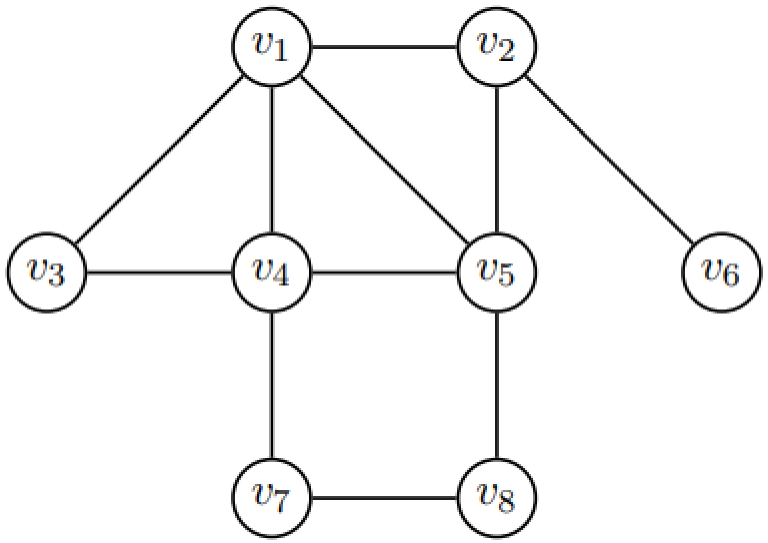
$$r(G) = \min_i \{e(v_i)\} = \min_i \left\{\max_j \{d(v_i, v_j)\}\right\}$$

- The diameter, denoted $d(G)$, is the maximum eccentricity of any vertex in the graph

$$d(G) = \max_i \{e(v_i)\} = \max_{ij} \{d(v_i, v_j)\}$$

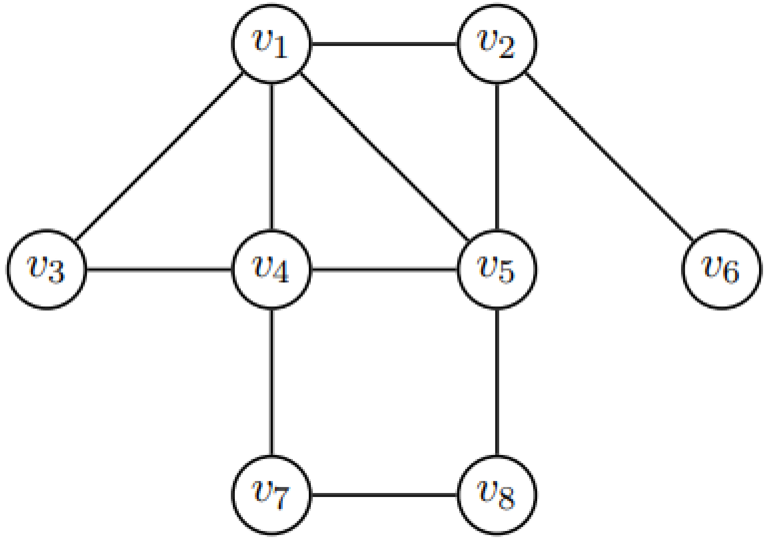
- For a disconnected graph, values are computed over the connected components of the graph.
- The diameter of a graph G is sensitive to outliers.

- What is the radius and diameter of the following graph?
  1. 2
  2. 3
  3. 4
  4. 5



(a)

- What is the radius and diameter of the following graph?
    1. 2
    2. 3
    3. 4
    4. 5



(a)

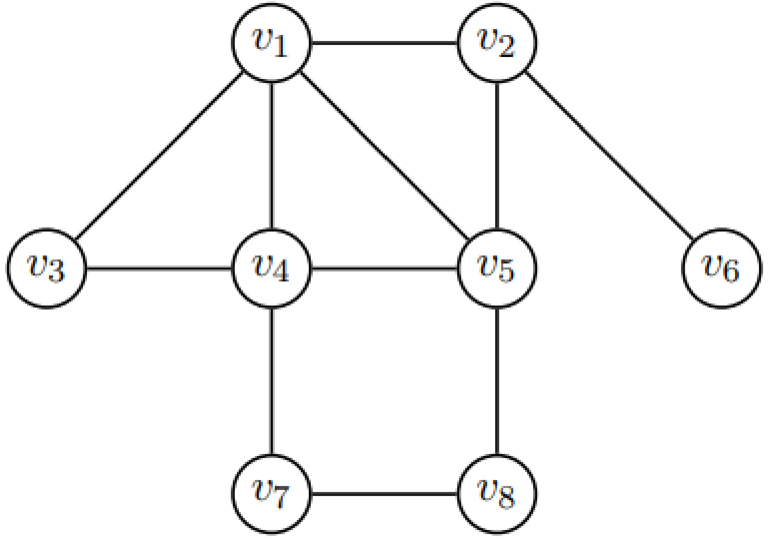
Clustering coefficient of a node

- The clustering coefficient of a node v_i is a measure of the density of edges in the neighborhood of v_i.
- Let $G_i = (V_i, E_i)$ be the subgraph induced by the neighbors of vertex v_i. Note that $v_i \notin V_i$, since we assume that G is simple. Let $|V_i| = n_i$ be the number of neighbors of v_i, and $|E_i| = m_i$ be the number of edges among the neighbors of v_i. The clustering coefficient of v_i is defined as:

$$C(v_i) = \frac{\#\ of\ edges\ in\ G_i}{maximum\ number\ of\ edges\ in\ G_i} = \frac{m_i}{\binom{n_i}{2}} = \frac{2 \cdot m_i}{n_i(n_i - 1)}$$

What is the clustering coefficient of v_3?

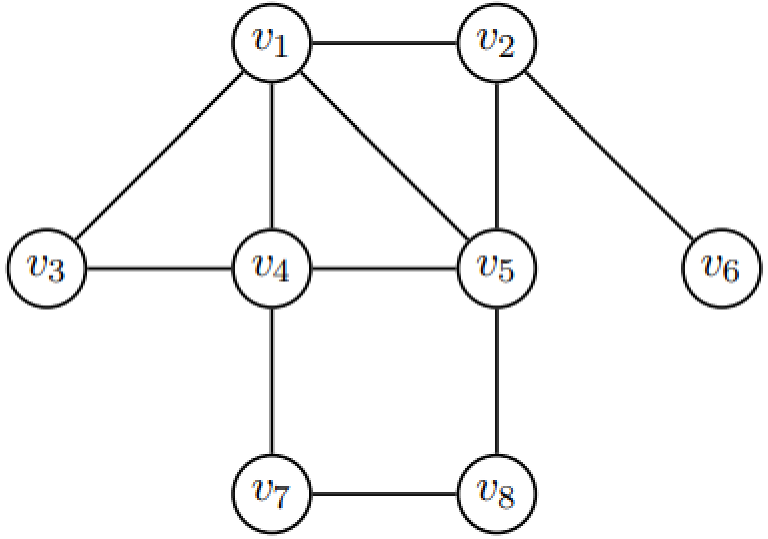
1. 1
2. 1/3
3. 1/2
4. 1/6
5. 6/10



(a)

What is the clustering coefficient of v_5?

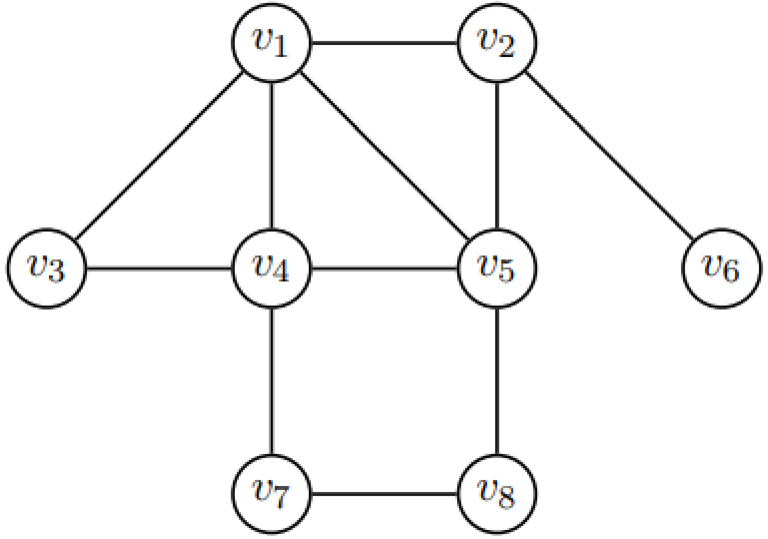
1. 1
2. 1/3
3. 1/2
4. 1/6
5. 6/10



(a)

What is the clustering coefficient of v_6?

1. 1
2. 1/3
3. 1/2
4. 1/6
5. 6/10
6. Not defined



(a)

Clustering coefficient of a node

- Note: Clustering coefficient of a node is not defined for nodes with degree less than 2.
- Therefore, if we need, we can consider 0 for the clustering coefficient for nodes with degree less than 2.

Clustering coefficient of a node

- How to interpret this value?
  - $C(v_i) = 1 \rightarrow$ All neighbors are connected to each other.
  - $C(v_i) = 0 \rightarrow$ None of the neighbors are connected to each other.
  - A higher clustering coefficient indicates a more tightly knit local community around the node.
- Example:
  - Fraud Detection in Financial Networks
    - In a bank transaction network, accounts are nodes, and transactions form edges.
    - Fraud rings often have a high clustering coefficient since fraudulent accounts transfer money within a small, well-connected group.

Clustering coefficient of a graph

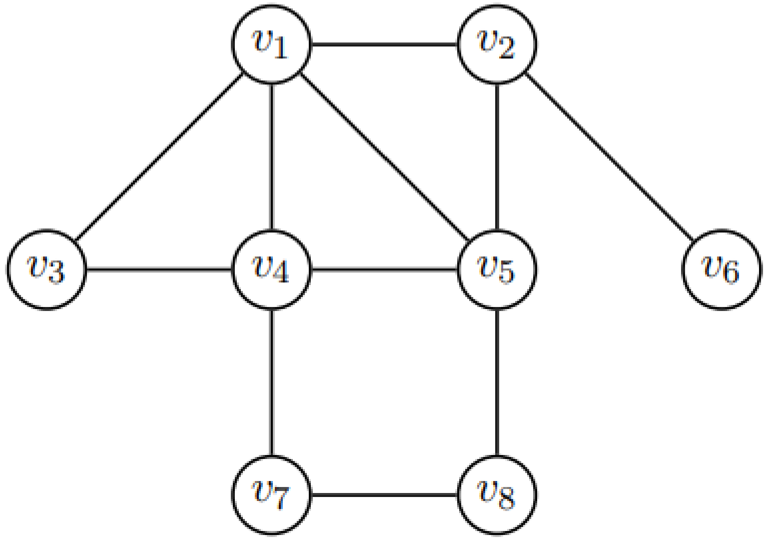
- The clustering coefficient of a graph G is simply the average clustering coefficient over all the nodes, given as:

$$C(G) = \frac{1}{n} \cdot \sum_{i} C(v_i)$$

Clustering coefficient of a graph

- The clustering coefficient of a graph G is simply the average clustering coefficient over all the nodes, given as:

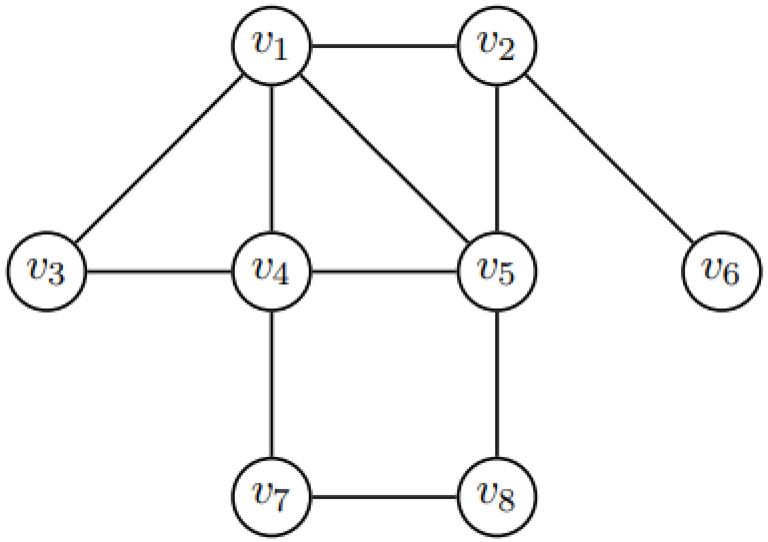
$$C(G) = \frac{1}{n} \cdot \sum_i C(v_i)$$



(a)

What is the clustering coefficient of this graph?

1. 3/8
2. 3/16
3. 5/16
4. 1/2



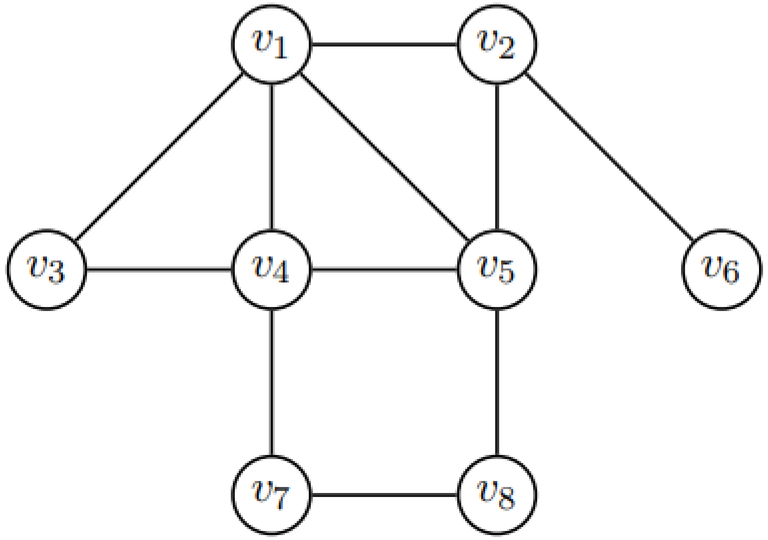
(a)

$$C(G) = \frac{1}{8} \cdot \left(\frac{3}{\binom{4}{2}} + \frac{1}{\binom{3}{2}} + 1 + \frac{2}{\binom{4}{2}} + \frac{2}{\binom{4}{2}} + 0 + 0 + 0 \right) = \frac{2.5}{8} = \frac{5}{16}$$

Clustering coefficient of a graph

- The clustering coefficient of a graph G is simply the average clustering coefficient over all the nodes, given as:

$$C(G) = \frac{1}{n} \cdot \sum_{i} C(v_i)$$



(a)

What is the clustering coefficient of this graph?

1. 3/8
2. 3/16
3. <u>5/16</u>
4. 1/2

Some examples

- Social networks
  - High clustering → tight friend groups, communities, social circles.
  - Low clustering → more broadcast-style or hub-and-spoke relationships.
- Example
  - A group of grad students in the same lab: very high clustering.
  - Twitter/X followers of a celebrity: low clustering (followers don't know each other).
- Insight
- Helps distinguish community-based social interaction from influence-based networks.
- High clustering often correlates with trust, repeated interaction, and social reinforcement.

Some examples

- Epidemiology and information spread
- High clustering can slow down global spread even if local spread is fast.
- Example
  - A disease spreads rapidly inside households or dorms (high clustering) but jumps between communities less frequently.
- Insight
  - Vaccinating a few “bridge” nodes between clusters can be more effective than random vaccination.

Some examples

- Biological networks
- High clustering often indicates functional modules.
- Example
  - Proteins involved in the same biological pathway tend to interact with each other, forming dense clusters.
- Insight
  - High local clustering can signal biological function.
  - Nodes with unusually low clustering inside a dense region may be regulatory or bridging proteins.

Some examples

- Infrastructure and robustness
- Clustering coefficient unravels details about redundancy and fault tolerance.
- Example
  - Power grids or communication networks with high clustering have local backup routes.
- Insight
  - High clustering improves local robustness.
  - But excessive clustering without long-range links can hurt global efficiency.

Efficiency

- The efficiency for a pair of nodes v_i and v_j is defined as $\frac{1}{d(v_i, v_j)}$
- If v_i and v_j are not connected, then efficiency between these two vertices is $\frac{1}{\infty} \approx 0.$
- Smaller the distance between two nodes, these nodes are efficient in communicating between them.
- The efficiency of a graph is defined as G is defined as the average efficiency of over all pairs of nodes.

$$\frac{2}{n \cdot (n-1)} \cdot \sum_{i} \sum_{j>i} \frac{1}{d(v_i, v_j)}$$

Efficiency

- What could be the maximum efficiency value of a Graph G and what kind of graph achieves this value?
- Maximum efficiency value is 1 and complete graphs have this property.

Local Efficiency

- The local efficiency for a node v_i is defined as the efficiency of the subgraph G_i induced by the neighbors of v_i.
- Since $v_i \notin G_i$, the local efficiency is an indication of fault tolerance.
- How efficient the communication between the neighbors of v_i, if v_i is removed.

- Intuition
  - Global efficiency answers: "If I pick two data points at random, how easily can information flow between them?"
  - Local efficiency answers: "If one data point disappears, can its neighbors still communicate efficiently?"
- These attributes reveal information about connectivity, robustness, redundancy and structure of the graph.

Measure of centrality

- The notion of centrality is used to rank the vertices of a graph in terms of how **"central"** or important they are.
- A centrality can be formally defined as a function:

$$c: V \to \mathbb{R}$$

$$\textit{that induces a total order on } V$$

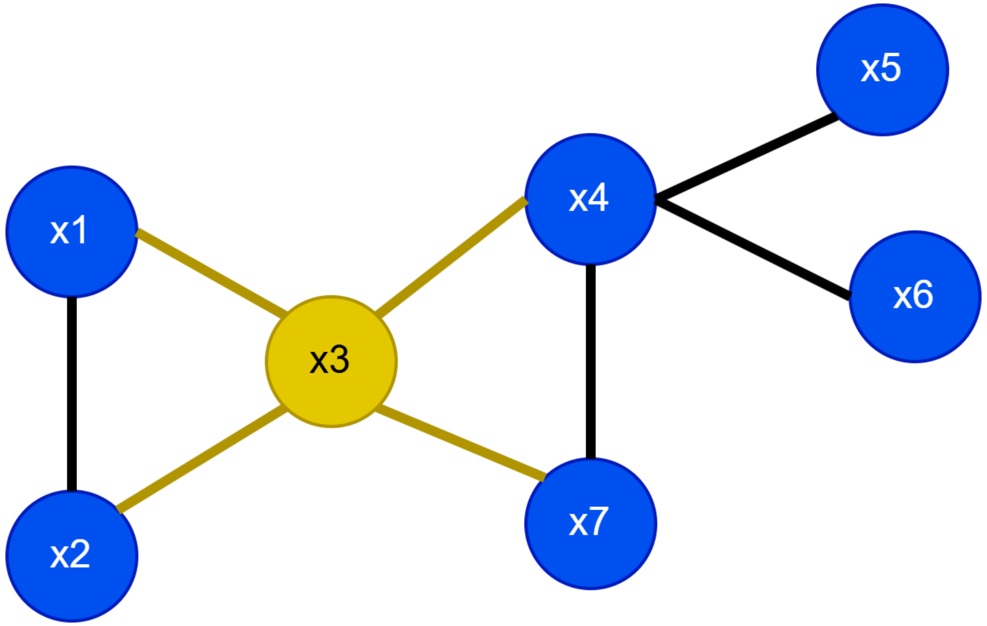
- We say that v_i is at least as central as v_j if $c(v_i) \geq c(v_j)$.
- A set has a total order if it has a partial order and every pair of elements in the set are comparable.

Degree centrality

- The simplest notion of centrality is the degree d_i of a vertex v_i.
- the higher the degree the more important or central the vertex.
- For directed graphs, one may further consider the indegree centrality and outdegree centrality of a vertex.

Degree centrality

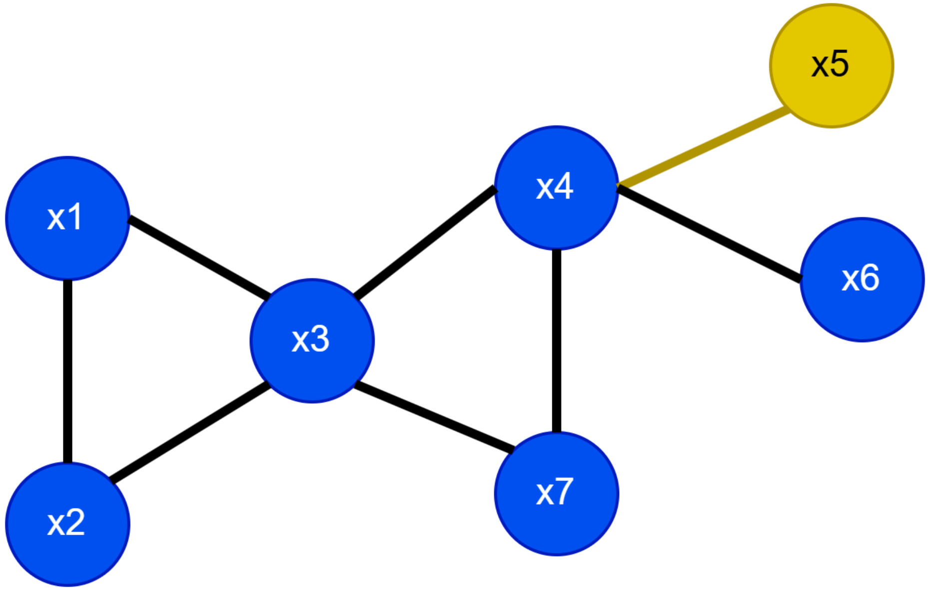
$$d(v_i) = \sum_{j}^{n} A(i,j)$$



$$d(x_3) = \sum_{1}^{n} A(3,j) = 4$$

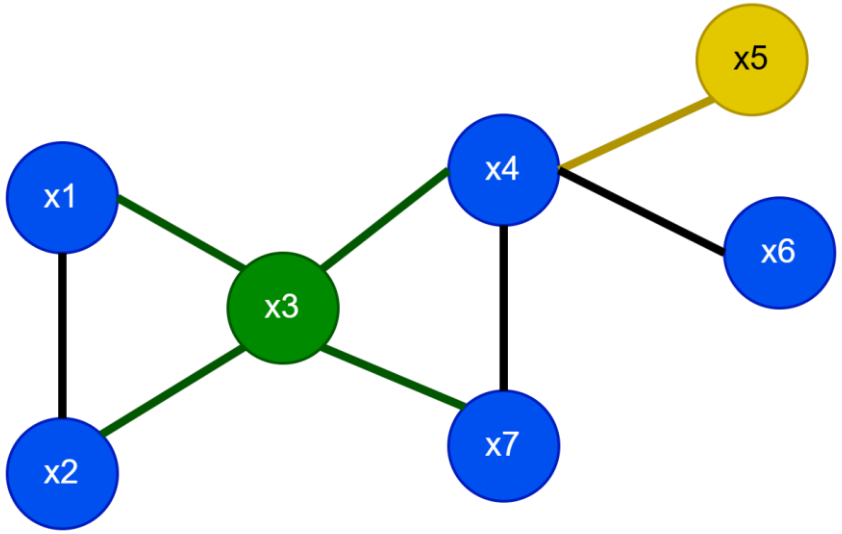
Degree centrality

$$d(v_i) = \sum_{j}^{n} A(i,j)$$



$$d(x_5) = \sum_{1}^{n} A(5,j) = 1$$

Degree centrality



$$d(x_3) = \sum_{1}^{n} A(3,j) = 4$$

$$d(x_5) = \sum_{1}^{n} A(5,j) = 1$$

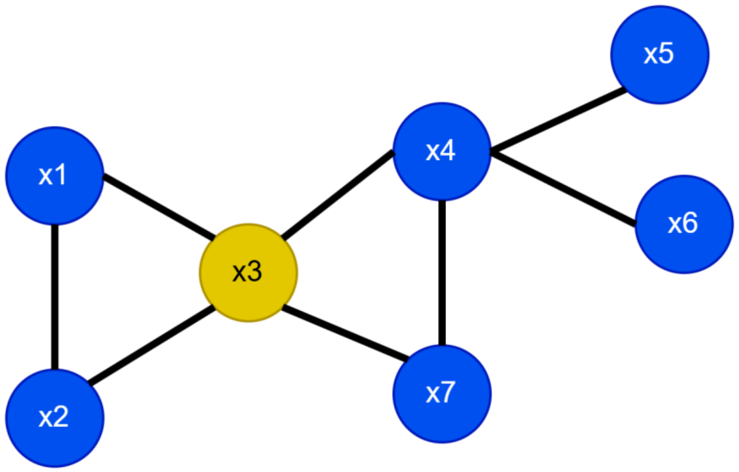
$x3$ has higher importance than $x5$ according to degree centrality.

Eccentricity Centrality

- Eccentricity centrality is thus defined as follows

$$c(v_i) = \frac{1}{e(v_i)} = \frac{1}{\max_j\{d(v_i, v_j)\}}$$

- The less eccentric a node is the more central it is.



What is the eccentric centrality of $x3$?

1. ½
2. 1/3
3. 4
4. 2
5. 3

Eccentric centrality

- A node v_i that has the least eccentricity, i.e., for which eccentricity is equal to the graph radius $e(v_i) = r(G)$, is called a ***center node.***
- A node that has the highest eccentricity, i.e., for which eccentricity equals to the graph diameter, $e(v_i) = d(G)$, is called a ***periphery node.***

Closedness centrality

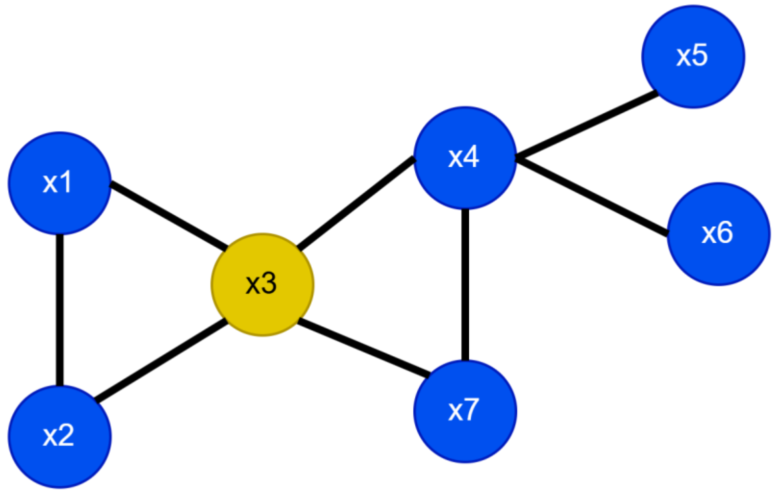
- The closeness centrality uses the sum of all the distances to rank how central a node is.

$$c(v_i) = \frac{1}{\sum_j d(v_i, v_j)}$$

A node v_i with the smallest total distance, $\sum_j d(v_i, v_j)$is called the ***median node***.

- Why is this median node important?
  - locate a facility that minimizes the distance to all other points.

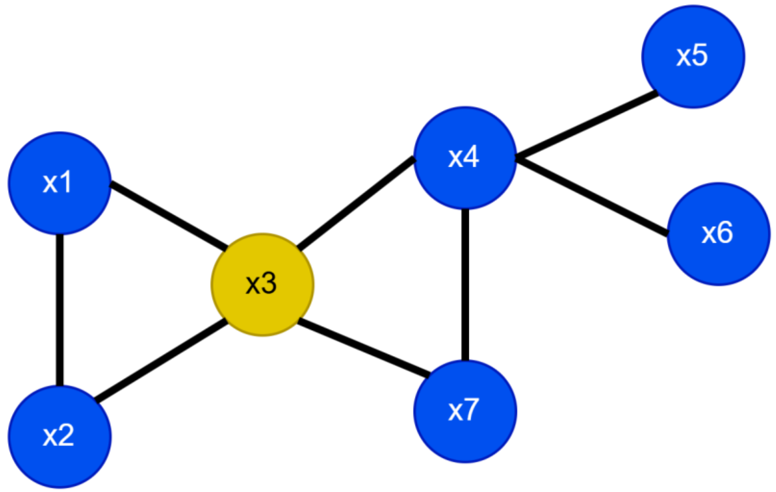
Closedness centrality



What is the **closedness centrality** of $x3$?

1. 1/8
2. 1/3
3. 3
4. 4
5. 8

Eccentricity centrality



What is the **eccentricity centrality** of $x3$?

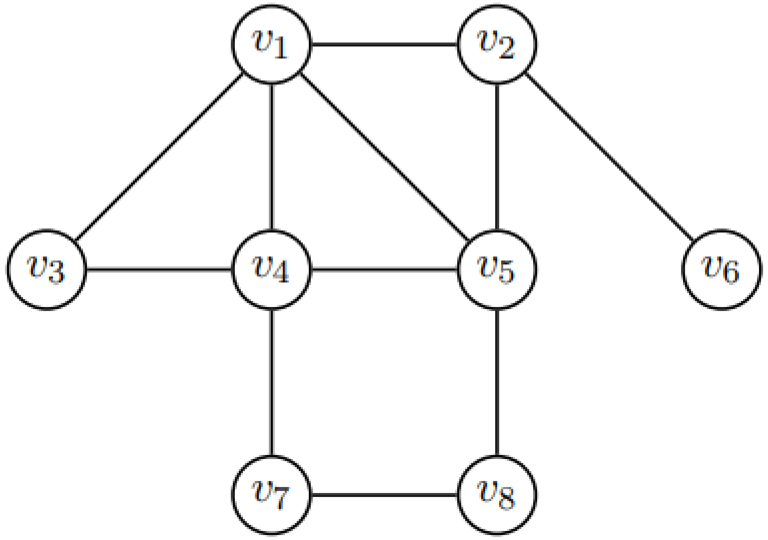
1. ½
2. 1/3
3. 4
4. 2
5. 3

Betweenness centrality

- For a given vertex v_i the betweenness centrality measures how many shortest paths between all pairs of vertices include v_i.
- This gives an indication as to the central "**monitoring**" role played by v_i for various pairs of nodes.

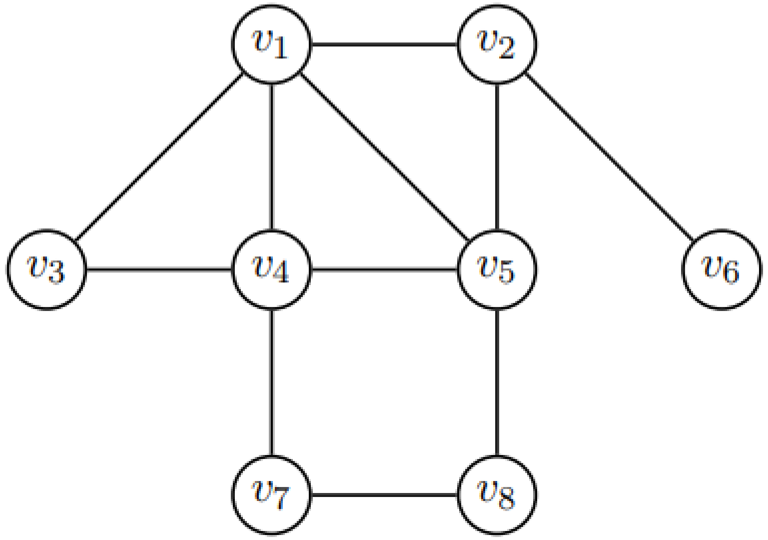
Betweenness centrality

- Let η_{jk} denote the number of shortest paths between vertices v_j and v_k, and let
- $\eta_{jk}(v_i)$ denote the number of such paths that include or contain v_i, then the fraction of paths through v_i is denoted as
$$\gamma_{jk}(v_i) = \frac{\eta_{jk}(v_i)}{\eta_{jk}}.$$
- If the two vertices v_j and v_k are not connected, we assume $\gamma_{jk} = 0$.
$$c(v_i) = \sum_{j \neq i} \sum_{\substack{k \neq i \\ k > j}} \gamma_{jk}(v_i)$$



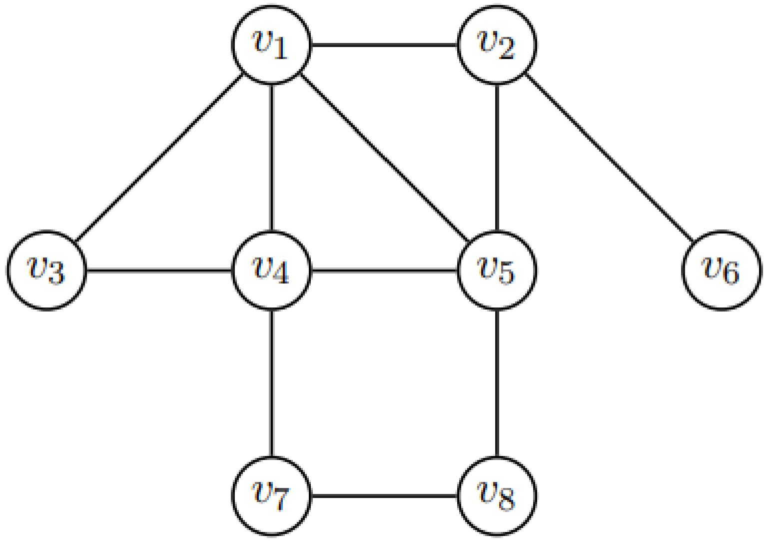
(a)

What is the betweenness centrality centrality of v_5?



(a)

	Shortest paths	Shortest paths through v_5 $\eta_{jk}(v_5)$
η_{12}	1	0
η_{13}	1	0
η_{14}	1	0
η_{16}	1	0
η_{17}	1	0
η_{18}	1	1
η_{23}	1	0
η_{24}	2	1
η_{26}	1	0
η_{27}	3	2
η_{28}	1	1
η_{34}	1	0
…		

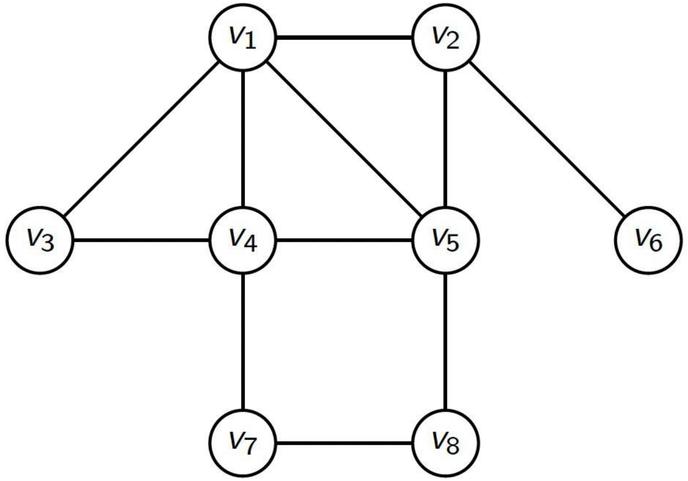


(a)

What is the betweenness centrality centrality of v_5?

$$c(v_5) = \gamma_{18} + \gamma_{24} + \gamma_{27} + \gamma_{28} + \gamma_{38} + \gamma_{46} + \gamma_{48} + \gamma_{67} + \gamma_{68}$$

$$c(v_5) = 1 + \frac{1}{2} + \frac{2}{3} + 1 + \frac{2}{3} + \frac{1}{2} + \frac{1}{2} + \frac{2}{3} = 6.5$$



Centrality	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
Degree	4	3	2	4	4	1	2	2
Eccentricity	0.5	0.33	0.33	0.33	0.5	0.25	0.25	0.33
$e(v_i)$	2	3	3	3	2	4	4	3
Closeness	0.100	0.083	0.071	0.091	0.100	0.056	0.067	0.071
$\sum_j d(v_i, v_j)$	10	12	14	11	10	18	15	14
Betweenness	4.5	6	0	5	6.5	0	0.83	1.17