

# Introduction to Data Science with Python

WMASDS04

Week 11: Network Analysis

# Lecture Outlines

---

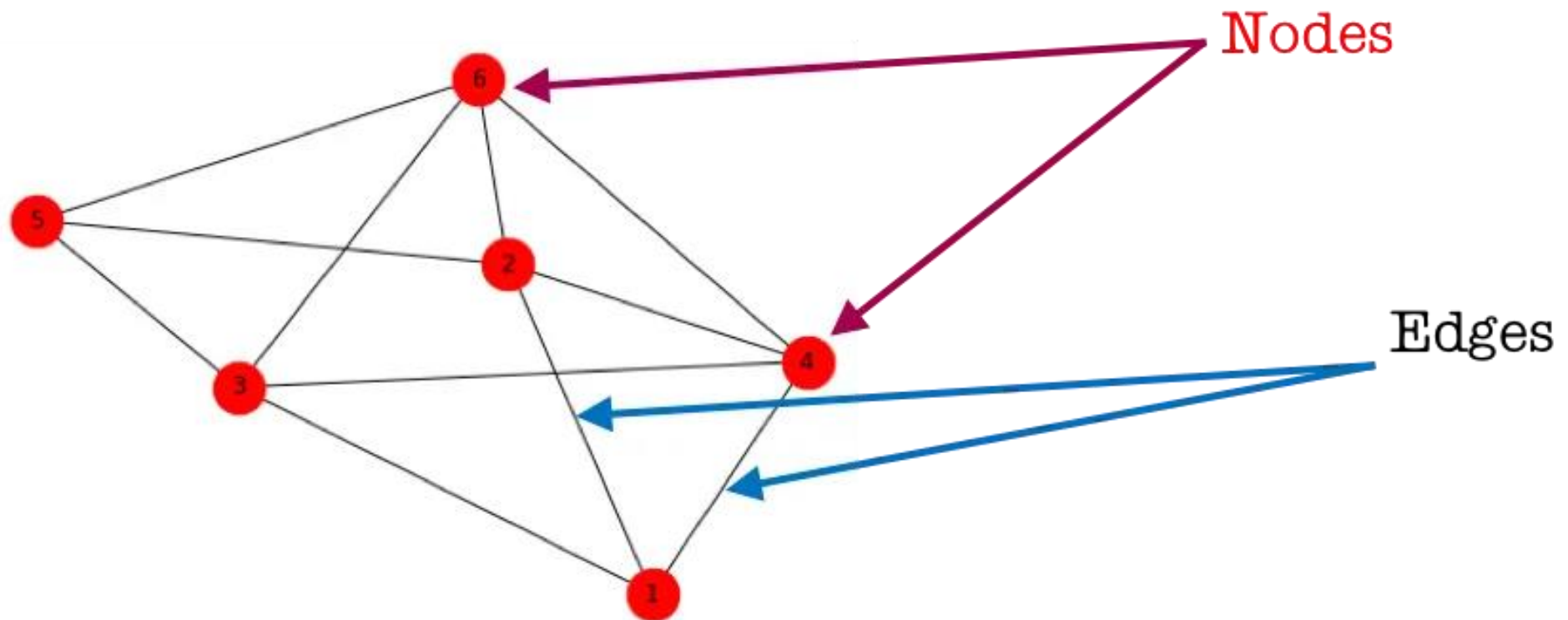
- Basic Definitions in Graphs,
- Social Network Analysis,
  - Scale free network and small world network
- Centrality,
- Ego-Networks,
- Community Detection.

# What is a Network?

---

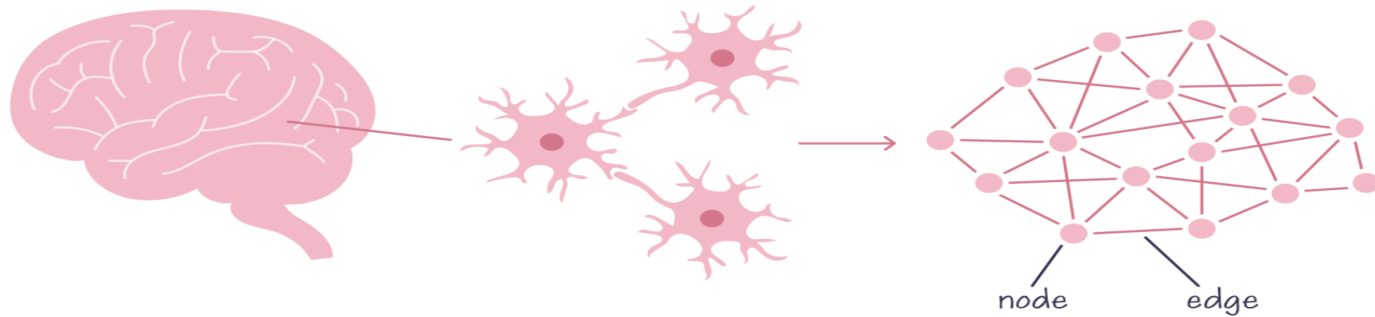
- A network refers to a structure representing a group of objects/people and relationships between them.
  - also known as a graph in mathematics.
- A network structure consists of nodes and edges.
  - nodes represent objects we are going to analyze
  - edges represent the relationships between those objects.
- For example,
  - in Facebook network, nodes are target users and edges are relationships such as friendships between users or group memberships.
  - In Twitter network, edges can be following/follower relationships.

# Example: Network of 6 Nodes

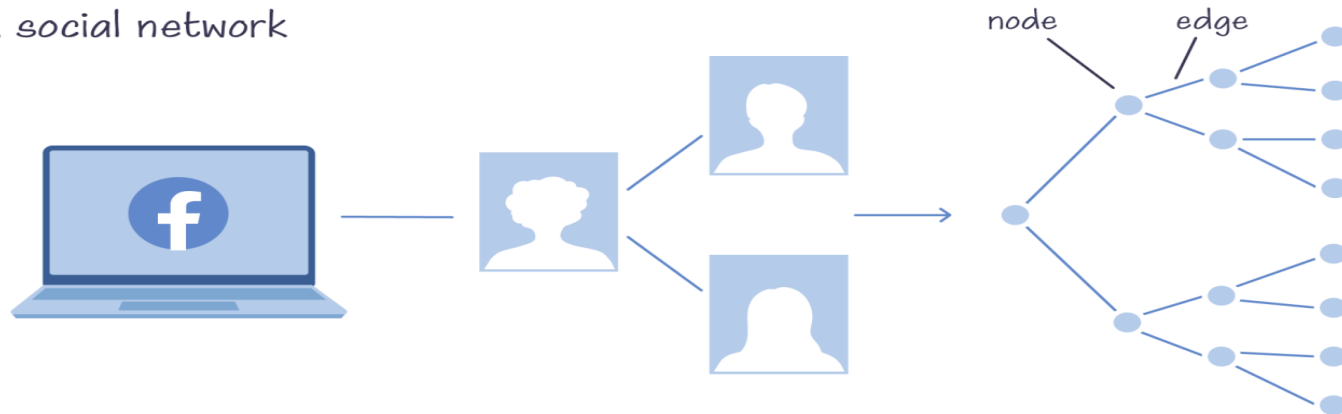


# Example: Networks

A neural network



A social network



# Why Network Analysis?

---

- It helps us in deep understanding the structure of a relationship in social networks, a structure or process of change in natural phenomena, or even the analysis of biological systems of organisms.
- Again, let's use the network of social media users as an example. Analyzing this network helps in
  - Identifying the most influent person/people in a group
  - Defining characteristics of groups of users
  - Prediction of suitable items for users
  - Other easy-to-understand examples are the Friend Suggestion function in Facebook or Follow Suggestion function in Twitter.

# Why do we need network analysis

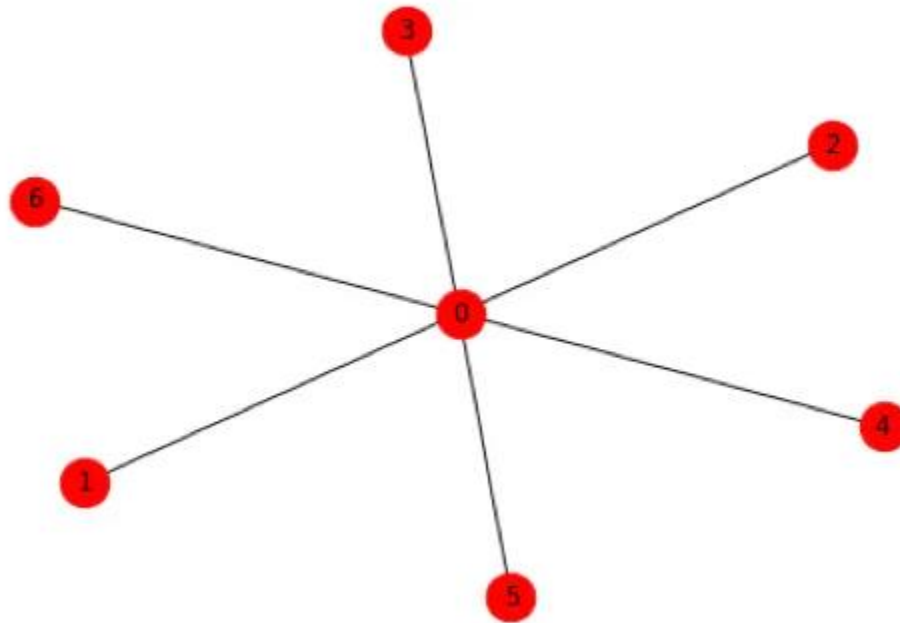
---

- Network analysis is a versatile tool that allows researchers, analysts, and decision-makers to uncover hidden patterns, gain insights into the structure of complex systems, and make informed decisions for optimization and problem-solving in various domains.
  - 1.Understanding Complex Systems:** Many real-world systems are complex and interconnected, and network analysis provides a way to comprehend the structure and dynamics of these systems. This understanding is crucial for making informed decisions and optimizing the functioning of these systems.
  - 2.Identifying Key Elements:** Network analysis helps identify key nodes (individuals, entities, or elements) within a network. These key nodes might be influential, have important connections, or play a critical role in the overall system. Understanding these key elements is valuable for strategic planning.
  - 3.Predicting Behavior:** By studying the patterns of connections and interactions in a network, one can make predictions about how the network will behave in the future. This predictive ability is useful in various fields, including epidemiology, finance, and social sciences.
  - 4.Optimizing Processes:** In many applications, such as supply chain management, transportation, and communication networks, network analysis helps optimize processes. It can identify bottlenecks, streamline pathways, and enhance overall efficiency.

# Example

---

• .

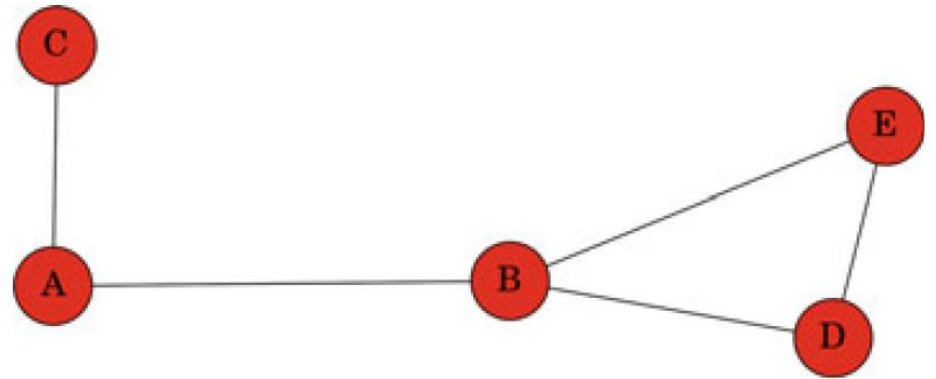




# Expected Output

---

**Fig. 8.1** Simple undirected labeled graph with 5 nodes and 5 edges



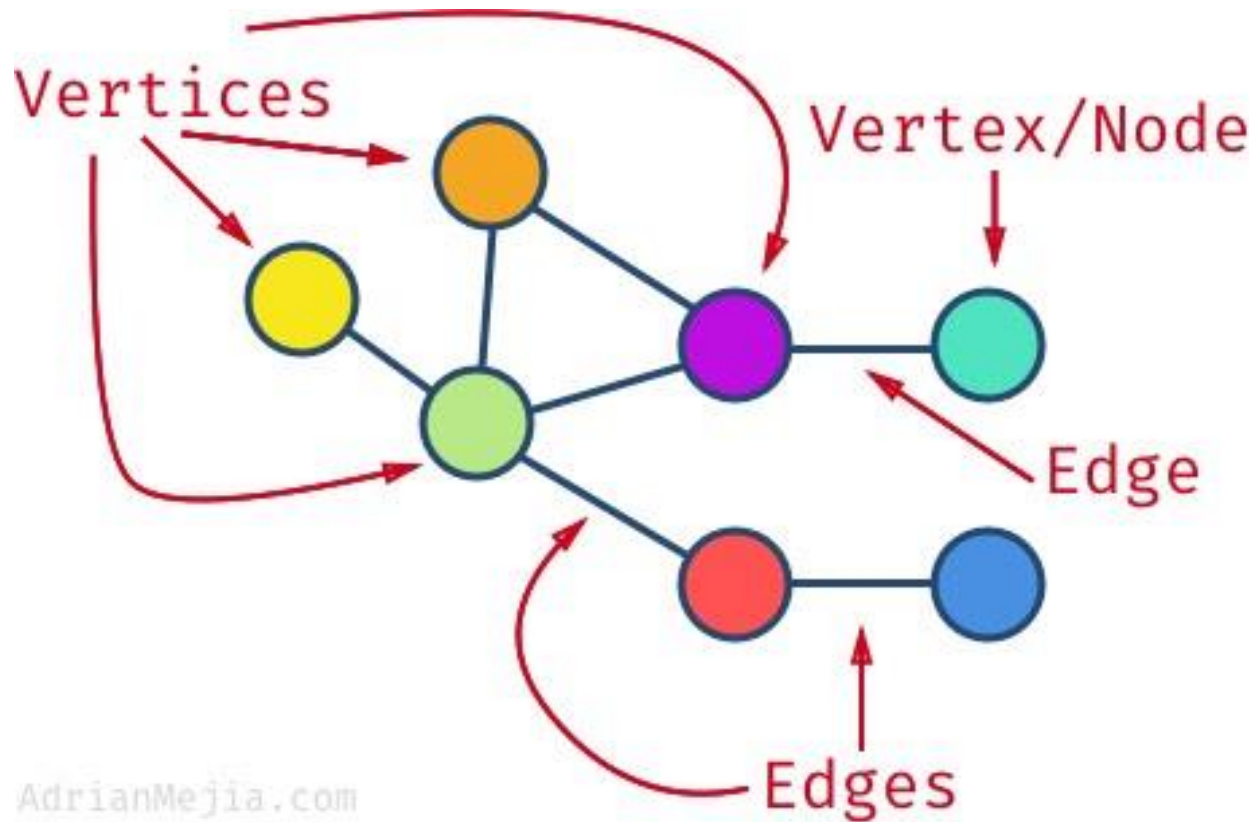
Activ

# Basic Definitions in Graphs

---

- **Graph:** is the mathematical term used to refer to a network.
- **Graph theory:** is the field that studies networks, and it provides the tools necessary to analyze networks.
- A graph is defined as a set of **nodes**, which are an abstraction of any entities (parts of a city, persons, etc.), and the connecting links between pairs of nodes called **edges** or relationships.
- **Node (or Vertex):** Represents a fundamental unit or point in a network. Nodes can represent entities such as individuals, locations, or any other discrete elements in the system being modeled.
- **Edge:** Represents a connection or relationship between two nodes (vertices).

# Graph



# Definitions

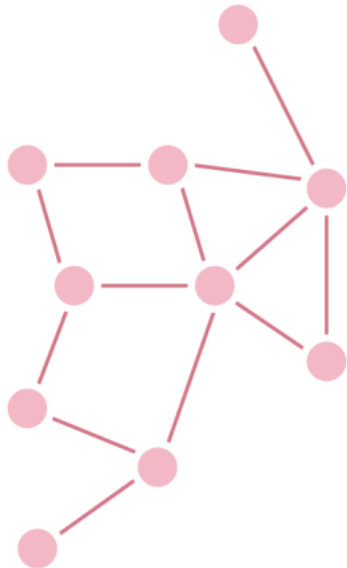
---

- A **directed edge** means that the edge points from one node to the other and not the other way round.
  - An example of a directed relationship is “a person knows another person”.
  - An edge has a direction when person A knows person B, and not the reverse direction if B does not know A.
  - **which is usual for many fans and celebrities.**
- An **undirected edge** means that there is a symmetric relationship.
  - An example is “a person shook hands with another person”; in this case, the relationship, unavoidably, involves both persons and there is no directionality.

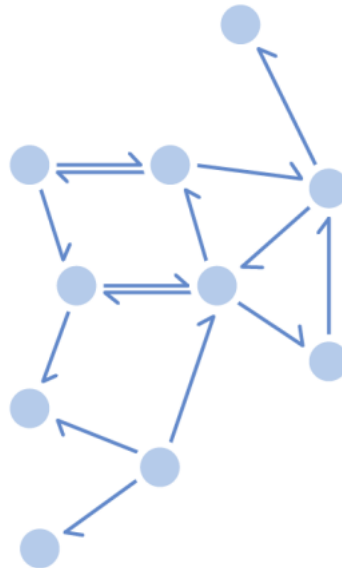
# Types of graphs

## Types of graphs

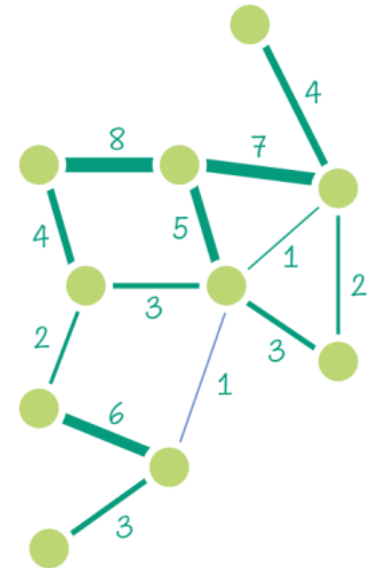
undirected



directed



weighted



# Definitions

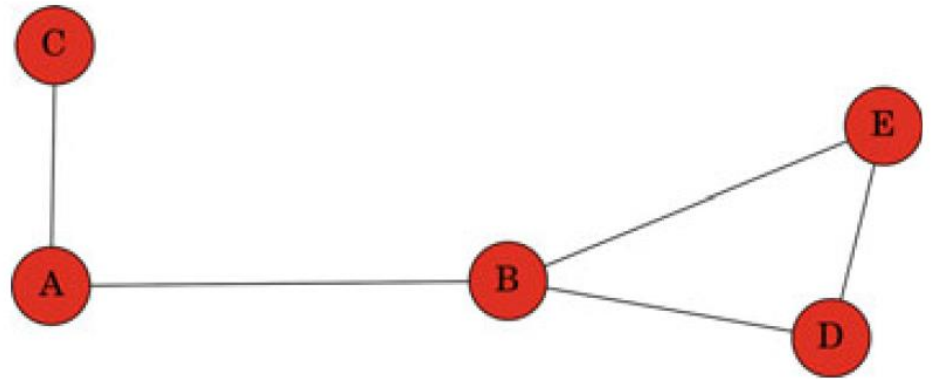
---

- The ***degree of a node*** is the number of edges that connect to it.
  - Figure 8.1 shows an example of an undirected graph with 5 nodes and 5 edges.
  - The degree of node C is 1, while the degree of nodes A, D and E is 2 and for node B it is 3.
  - If a network is directed, then nodes have two different degrees, the *in-degree*, which is the number of incoming edges, and the *out-degree*, which is the number of outgoing edges.
- We could add *strengths* or *weights* to the links between the nodes, to represent some real-world measure. In this case, the graph is called a *weighted graph*.
  - For instance, the length of the highways connecting the cities in a network.

# Expected Output

---

**Fig. 8.1** Simple undirected labeled graph with 5 nodes and 5 edges



Activ

# Definitions

---

- We define a ***path*** in a network to be a sequence of nodes connected by edges.
- The **shortest path** problem is the problem of finding a path between two nodes in a graph such that the length of the path or the sum of the weights of edges in the path is minimized.
  - In the example in Fig. 8.1, the paths (C, A, B, E) and (C, A, B, D, E) are those between nodes C and E.
  - This graph is unweighted, so the shortest path between C and E is the one that follows the fewer edges: (C, A, B, E)

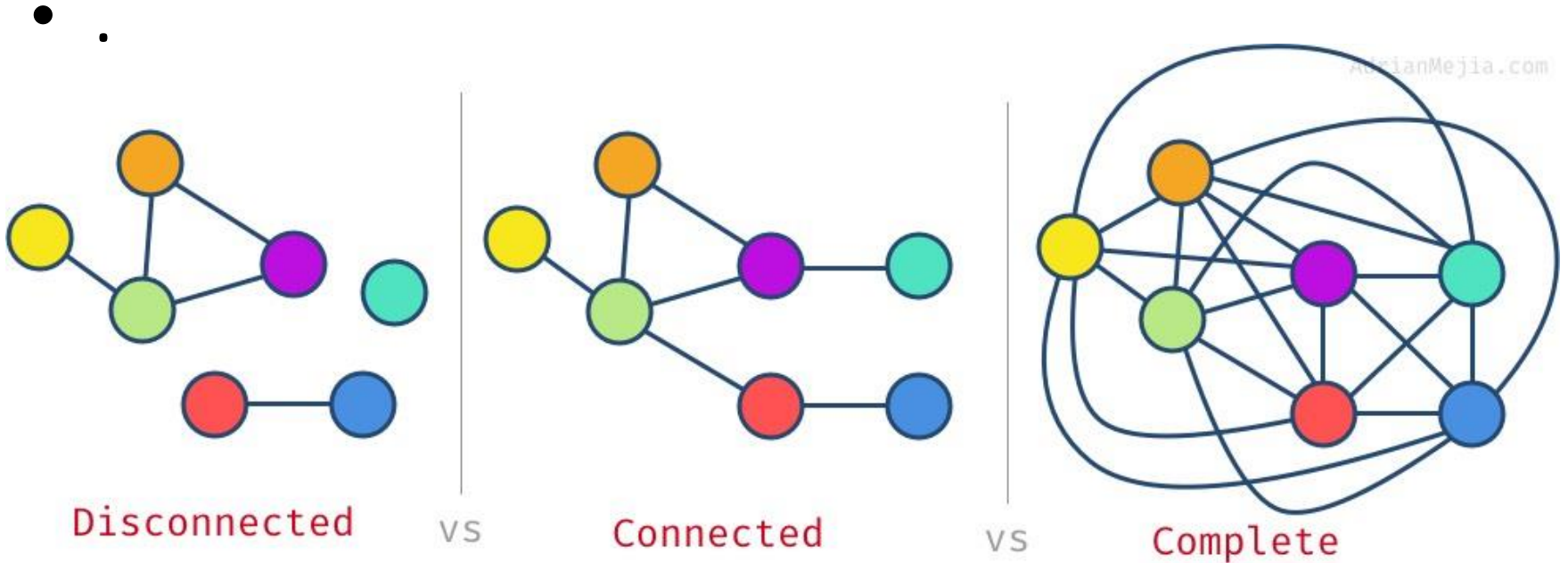


# Definitions

---

- A graph is said to be *connected* if for every pair of nodes, there is a path between them.
- A graph is *fully connected* or *complete* if each pair of nodes is connected by an edge.
- A *connected component* or simply a *component* of a graph is a subset of its nodes such that every node in the subset has a path to every other one.
  - In the example of Fig. 8.1, the graph has one connected component.
- A *subgraph* is a subset of the nodes of a graph and all the edges linking those nodes. Any group of nodes can form a subgraph.

# Types of graphs

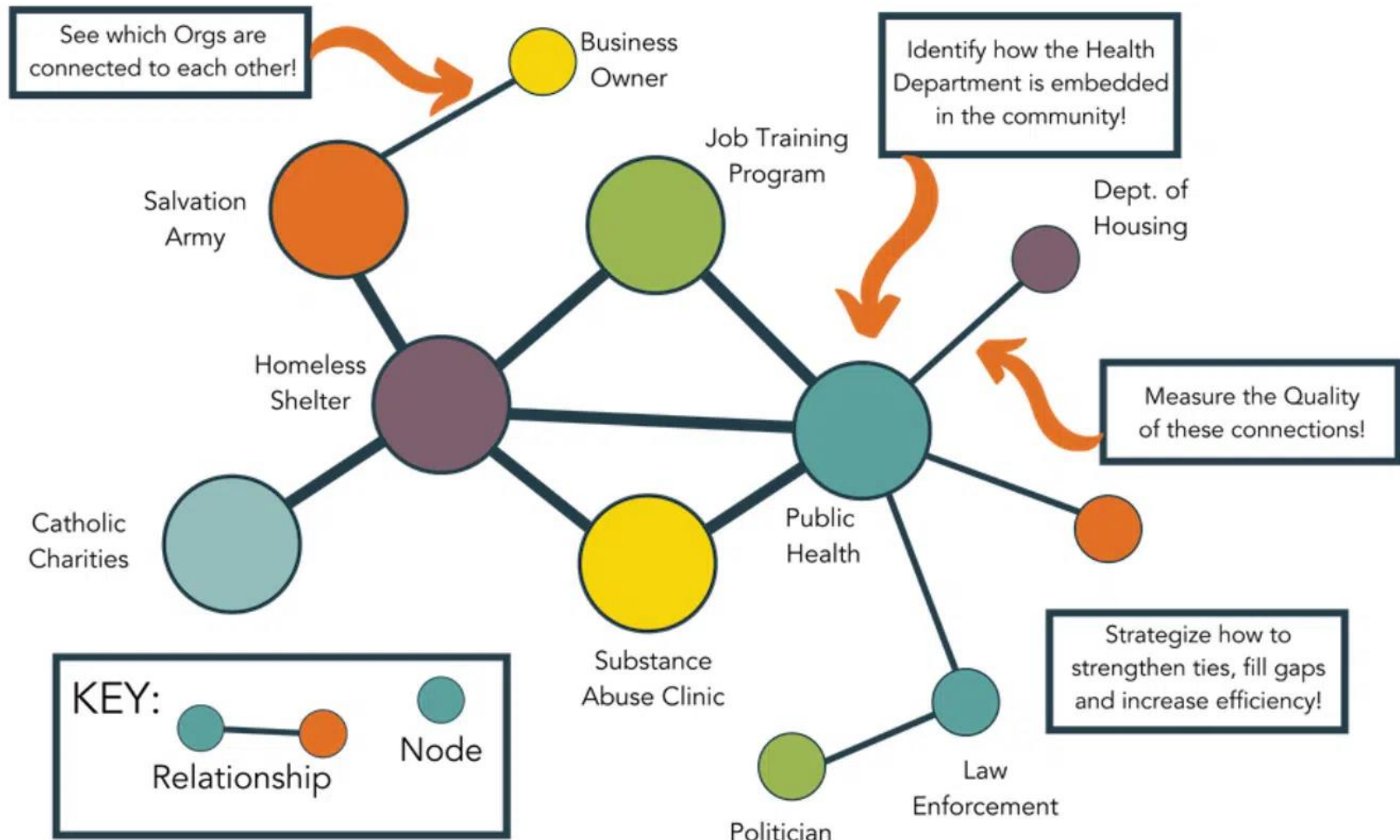


# Social Network Analysis

---

- Social Network Analysis (SNA) studies relationships between individuals or entities in a social system.
  - Social network analysis processes **social data structured** in graphs.
- Key concepts include nodes, edges, and metrics like degree, centrality, and clustering.
- It involves the extraction of several characteristics and graphics to describe the main **properties of the network**.
  - SNA visually represents networks, analyzes social structures, and identifies influential individuals or groups.
- Some general properties of networks, such as the shape of the network degree distribution (defined bellow) or the average path length, determine the type of network, such as a ***small-world network*** or a ***scale-free*** network.

# Example: Social Network Analysis



# Small World Network

---

- A small-world network is a type of graph in which most nodes are not neighbors of one another, but **most nodes can be reached from every other node in a small number of steps**.
- This is the so-called *small-world phenomenon* which can be interpreted by the fact that **strangers are linked by a short chain** of acquaintances.
- In a small-world network, people usually form communities or small groups where everyone knows everyone else. Such communities can be seen as complete graphs.
- In addition, most the community members have a few relationships with people outside that community.
- However, some people are connected to a large number of communities. These may be **celebrities** and such people are considered as the **hubs** that are responsible for the small-world phenomenon.
- Many small-world networks are also scale-free networks.

# Scale-free Network

---

- In a scale-free network the node degree distribution follows a power law (a relationship function between two quantities  $x$  and  $y$  defined as  $y = x^n$ , where  $n$  is a constant).
- The name *scale-free* comes from the fact that power laws have the same functional form at all scales, i.e., their shape does not change on multiplication by a scale factor.
- Thus, by definition, a scale-free network has **many nodes with a very few connections** and **a small number of nodes with many connections**.
- This structure is typical of the World Wide Web and other social networks.

# Example: Scale-free Network

---

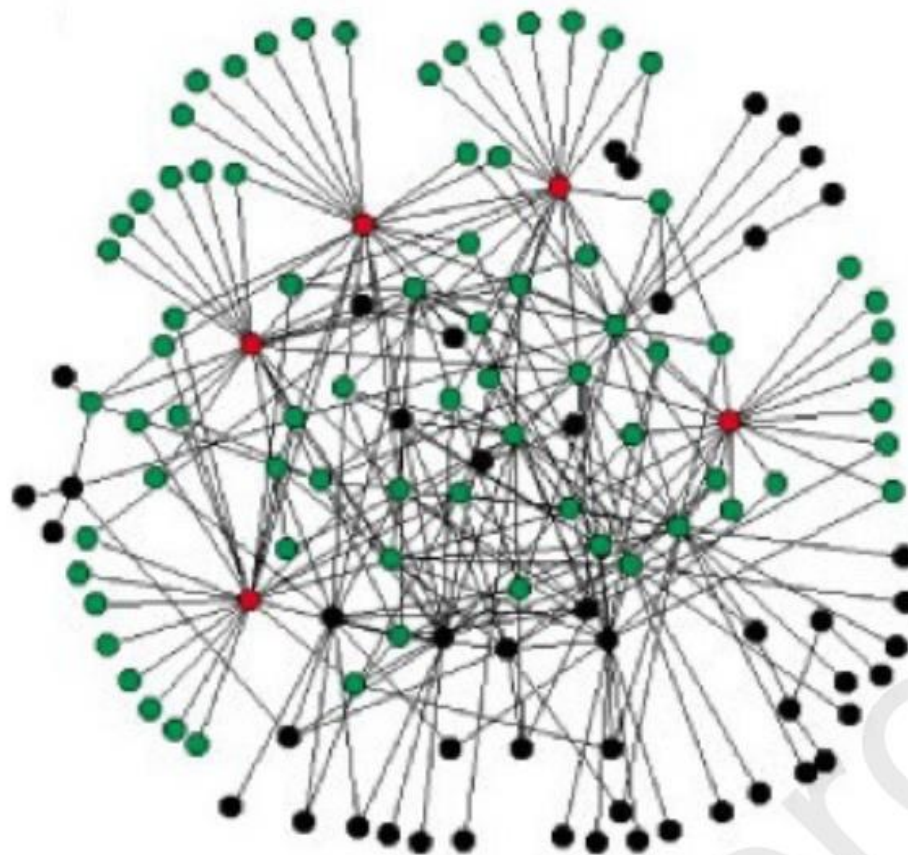
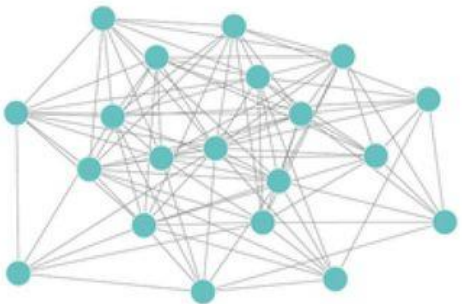
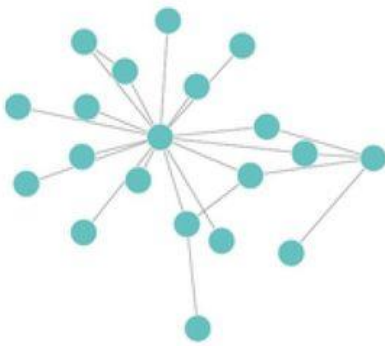
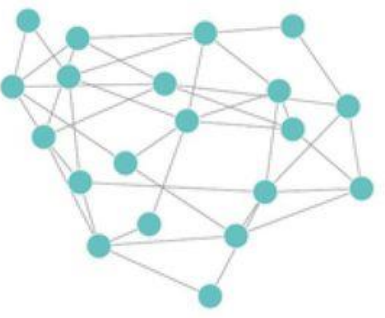


Figure 2: A scale-free network of 130 nodes. Five nodes with the biggest degree (red nodes) are in contact with a large fraction of other nodes, 60% of other nodes (green nodes) (Wang and Chen, 2003).

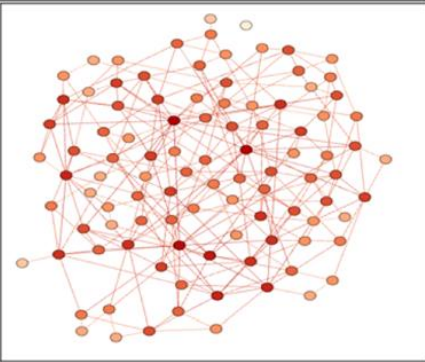
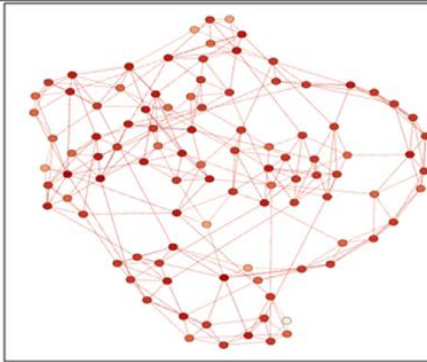
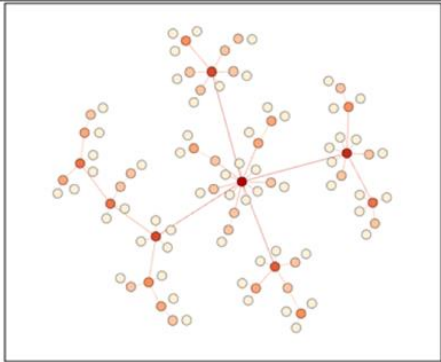
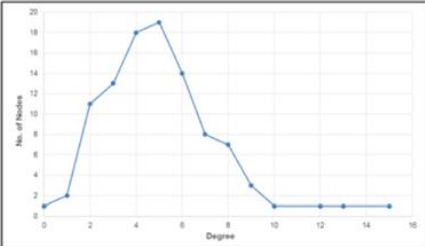
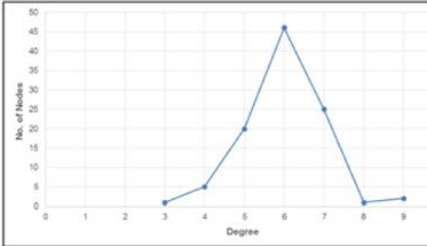
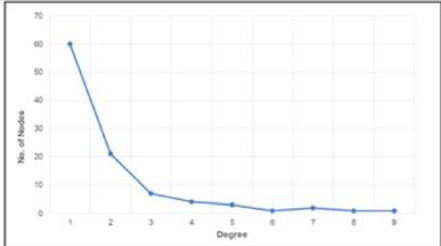


# Small-world and scale free network

	RANDOM	SCALE-FREE	SMALL-WORLD
Graph			
Degree distribution	Poisson distribution	Power-law distribution	Power-law distribution
Mini description	Nodes are randomly connected with probability $p$ , thus most nodes have equal number of degrees.	A small number of nodes has a high degree (i.e. are connected with many nodes), whereas the other nodes have low degree.	There is a short average path between most nodes, similar to a random, but they display higher clustering coefficients.



# Comparison of small-world and scale free networks.

	Random	Small-World	Scale-Free
Visualisation of the Topology			
Degree Distribution			
Robustness Characteristics	Responds similarly to both random and targeted attacks.		Resilient against random failures but very sensitive to targeted attacks.

# Example

---

- **World Wide Web (WWW):**
  - Websites exhibit scale-free properties with a few highly linked pages.
- **Scientific Citation Networks:**
  - Academic papers often form a scale-free network where a few influential papers receive a large number of citations.
- **Collaboration Networks:**
  - Networks of collaboration among scientists, artists, or authors often show scale-free patterns.
- **Movie Actor Collaboration Network:**
  - In the film industry, actors who frequently collaborate create a network with scale-free features.
- **Power Grids:**
  - The connections between power stations often display small-world characteristics, allowing efficient transmission of electricity.
- **Transportation Networks:**
  - Airports, highways, and subway systems may display small-world features, allowing quick connectivity between locations.
- **Epidemiological Networks:**
  - The spread of diseases in a population can be modeled as a small-world network.
- **Friendship Networks in Offline Communities:**
  - Beyond online social networks, friendships in physical communities may also show small-world characteristics.
- **Social Networks:**
  - Platforms like Facebook, Twitter, and LinkedIn display both small-world and scale-free characteristics.
- **Neuronal Networks:**
  - Neuronal connections in the brain often exhibit both small-world and scale-free characteristics.

# Basics in NetworkX

---

*NetworkX*<sup>1</sup> is a Python toolbox for the creation, manipulation and study of the structure, dynamics and functions of complex networks. After importing the toolbox, we can create an undirected graph with 5 nodes by adding the edges, as is done in the following code. The output is the graph in Fig. 8.1.

```
import networkx as nx
G = nx.Graph()
G.add_edge('A', 'B');
G.add_edge('A', 'C');
G.add_edge('B', 'D');
G.add_edge('B', 'E');
G.add_edge('D', 'E');
nx.draw_networkx(G)
```

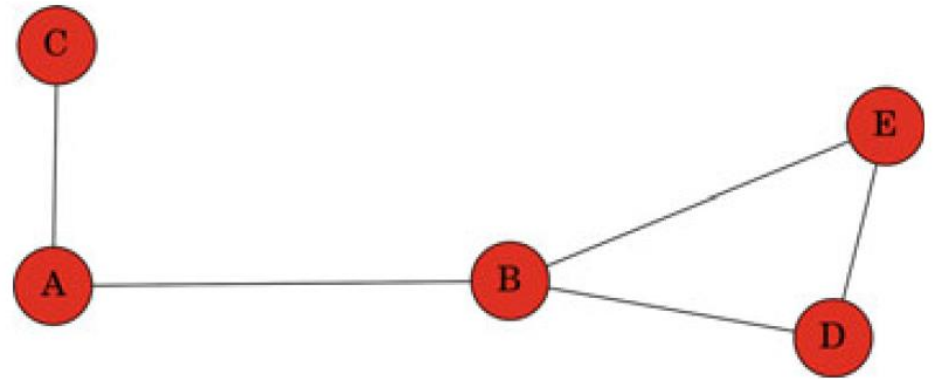
To create a directed graph we would use `nx.DiGraph()`.

Activate  
Go to Sett

# Expected Output

---

**Fig. 8.1** Simple undirected labeled graph with 5 nodes and 5 edges



Activ

# Centrality

---

- Who is the Important Person?
- A crucial application of network analysis is **identifying the important node** in a network. This task is called Measuring Network Centrality.
- In social network analysis, it can refer to the task of identifying the most influential member, or the representative of the group.

# Centrality

---

- The centrality of a node measures its relative importance within the graph.
- In practice, what centrality means will depend on the application and the meaning of the entities represented as nodes in the data and the connections between those nodes.
- Various measures of the centrality of a node have been proposed. We present four of the best-known measures:
  - degree centrality
  - betweenness centrality
  - closeness centrality, and
  - eigenvector centrality.
- The applications of centrality concepts in a social network include identifying the most influential people, the most informed people, or the most communicative people.
  - central nodes are probably more influential, have greater access to information, and can communicate their opinions to others more efficiently.

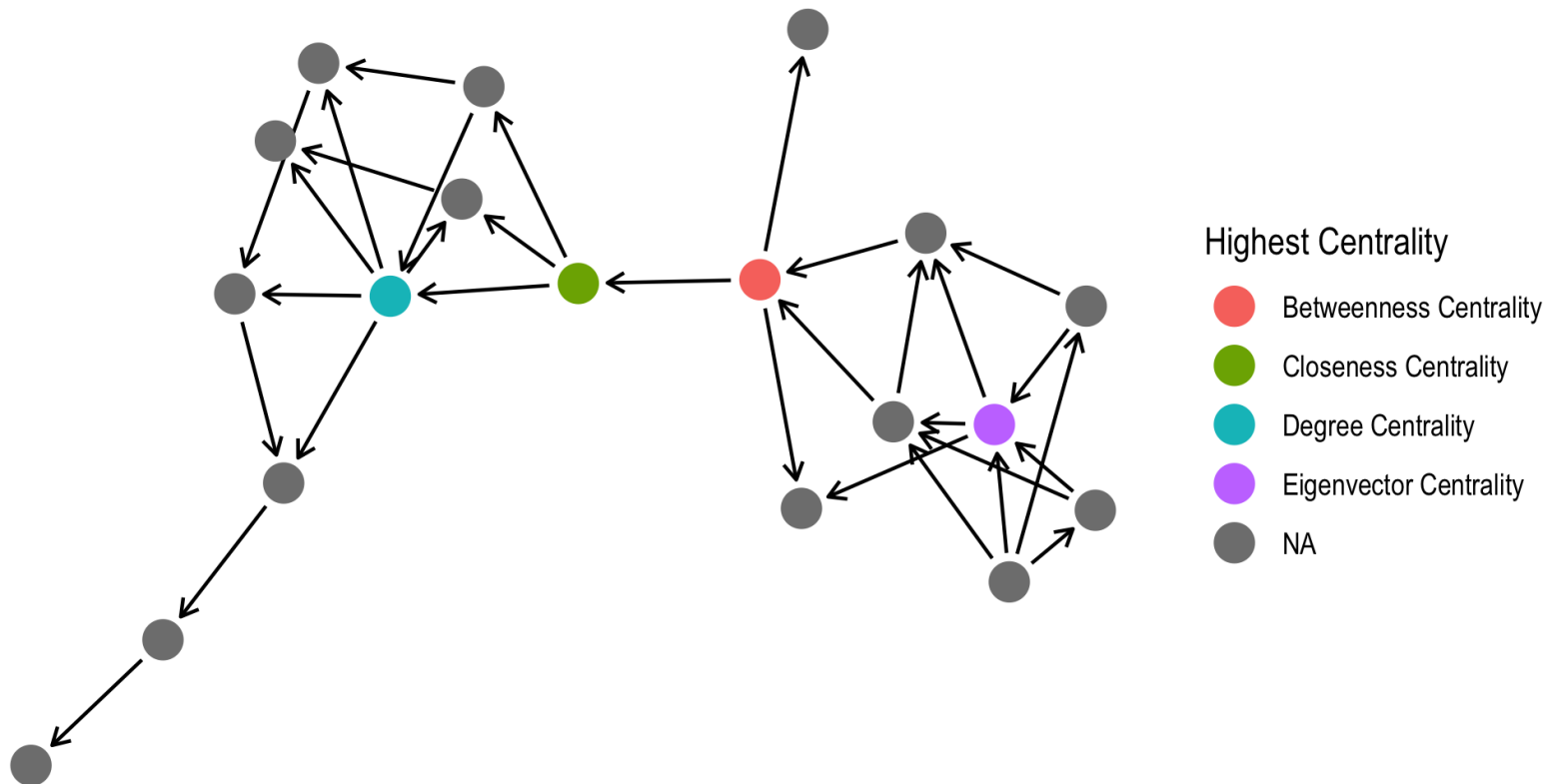
# Measures of Centrality

---

- Degree centrality
  - Defined as the number of edges of the node.
  - So the more ties a node has, the more central the node is.
- Betweenness centrality
  - quantifies the number of times a node is crossed along the shortest path/s between any other pair of nodes.
  - a public bus transportation network, the bus stop (node) with the highest betweenness has the most traffic.
- Closeness centrality
  - tries to quantify the position a node occupies in the network based on a distance calculation.
- Eigenvector centrality.
  - it measures the extent to which a node is connected to influential nodes.

# Measures of Centrality

## Variability of Centrality Measures





# Degree centrality vs betweenness

---

- Degree centrality depends only on the node's neighbors.
- Betweenness centrality depends on the connection properties of every pair of nodes in the graph, except pairs with the node in question itself.

# Clustering coefficient

---

- The clustering coefficient of a node in a network measures the degree to which its neighbors are connected to each other.
- For a node with  $k_i$  neighbors, the clustering coefficient  $C_i$  is given by:

$$C_i = \frac{2 \times E_i}{K_i \times (K_i - 1)}$$

Where:

$E_i$  is the actual number of edges between the neighbors of node  $i$ .

$k_i$  is the total number of neighbors of node  $i$ .

- The clustering coefficient ranges from 0 to 1, where:
- $C_i=0$  indicates that none of the neighbors of node  $i$  are connected to each other.
- $C_i=1$  indicates that all neighbors of node  $i$  are fully connected to each other.
- The global clustering coefficient for the entire network is often calculated by averaging the clustering coefficients of all nodes in the network.
- Clustering coefficients are essential in understanding the local connectivity patterns within a network and are a key property in characterizing small-world networks.

# Ego-Networks

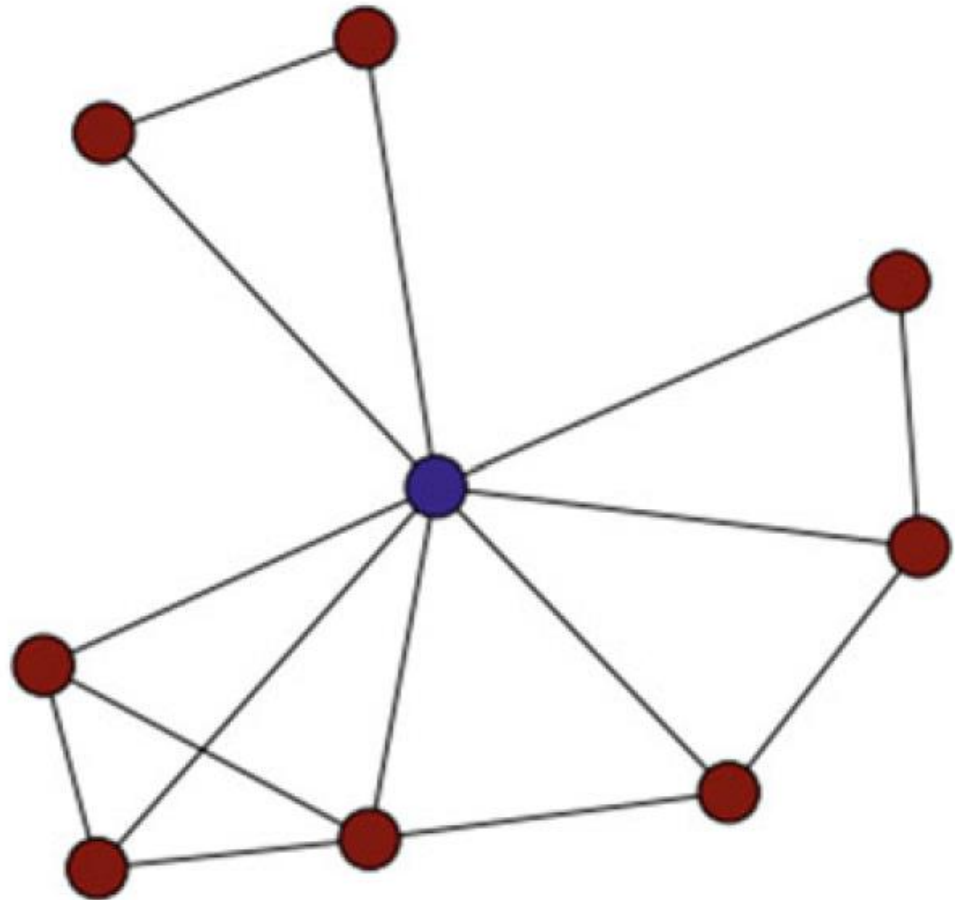
---

- An ego network, also known as a personal network, is the network of connections surrounding an individual node, referred to as the "ego."
  - In Facebook and LinkedIn, these are described as “your network
  - the network distance of 2 means that a person, C, is a friend of a friend of A
- Ego networks are useful for studying individual behavior, information flow, and the dynamics of social interactions within a smaller, more manageable context compared to the entire network.
  - **Family Members:** Your siblings, parents, and maybe cousins form a part of your ego network.
  - **Close Friends:** Individuals you frequently interact with and share personal experiences with would be part of your ego network.
  - **Colleagues:** People you work closely with, both professionally and perhaps on a personal level, contribute to your ego network.
  - **Neighbors:** Those who live in close proximity and with whom you have regular interactions form another layer of your ego network.
- **Knowing the size of an ego-network is important when it comes to understanding the reach of the information that a person can transmit or have access to.**

# Example: Ego-Network

---

**Fig. 8.11** Example of an ego-network. The *blue* node is the ego



# Community

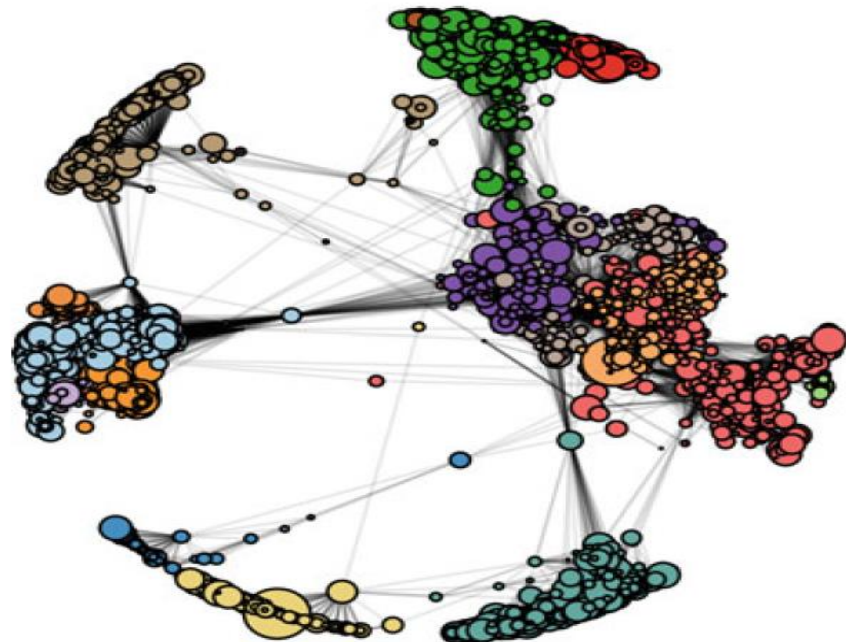
---

- A "community" refers to a group or subset of nodes within a network that are more densely connected to each other internally than to nodes outside the group.
  - High Internal Connectivity
  - Low External Connectivity
- Community Detection: The process of identifying and delineating communities within a network is known as community detection or clustering.
- Modularity: Modularity is a measure commonly used to assess the quality of a network partition into communities. It quantifies the degree to which the network is divided into modules with strong internal connections.
- Real-World Applications: social networks, biological networks, and technological networks.
  - For example, in a social network, communities may represent groups of friends with strong connections among themselves.

In [18]:

```
import community
partition = community.best_partition(fb)
print "#
communities found:", max(partition.values())
colors2 = [partition.get(node) for node in fb.nodes()]
nsize = np.array([v
for v in degree_cent_fb.values()])
nsize = 500*(nsize - min(nsize))/(max(nsize) - min(nsize))
nodes = nx.draw_networkx_nodes(
    fb, pos = pos_fb,
    cmap = plt.get_cmap('Paired'),
    node_color = colors2,
    node_size = nsize,
    with_labels = False)
edges = nx.draw_networkx_edges(fb, pos = pos_fb, alpha = .1)
```

**Fig. 8.14** The Facebook network drawn using the Spring layout and different colors to separate the communities found



# Modularity

---

- The modularity index ( $Q$ ) is a measure used to assess the quality of a network partition into communities. The interpretation of the modularity index values is as follows:
- **$Q = 0$ :**
  - Indicates that the network partition is not better than a random partition. There is little to no evidence of meaningful community structure.
- **$0 < Q < 0.3$ :**
  - Suggests a weak community structure. The network partition is better than random, but the communities are not well-defined.
- **$0.3 < Q < 0.7$ :**
  - Indicates a moderate to strong community structure. The network partition is considered meaningful, and communities are reasonably well-defined.
- **$Q > 0.7$ :**
  - Suggests a very strong community structure. The network partition is highly meaningful, and communities are well-separated.

---

THANK YOU