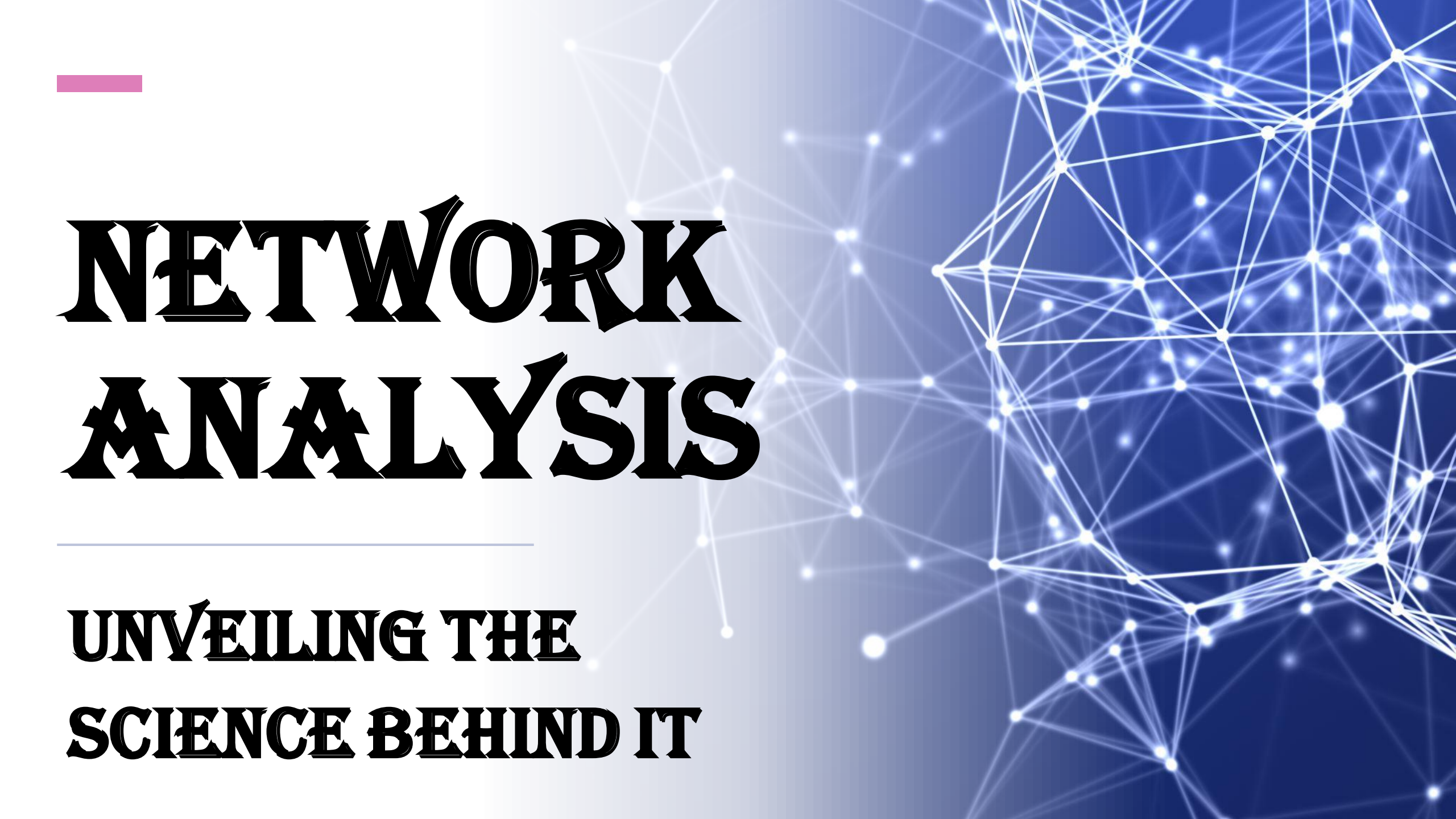
# NETWORK ANALYSIS

## UNVEILING THE SCIENCE BEHIND IT



# INTRODUCTION

# NETWORK ANALYSIS

- Network Analysis concerns itself with the formulation and solution of problems that have a network structure; such network structure is usually captured in a **graph**
- Instead of focusing on individuals and their attributes, Network analysis centers on relations between individuals, groups or social institutions
- Euler, a famous mathematician used a graph in 1736 to prove that there is no path that crosses only one bridge of the city

# NETWORK SCIENCE

*"The study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena"*

- Network science is an academic field which studies complex networks
- The field draws on theories and methods including graph theory from mathematics, statistical mechanics from physics, data mining and information visualization from computer science, inferential modeling from statistics, and social structure from sociology

# APPLICATIONS

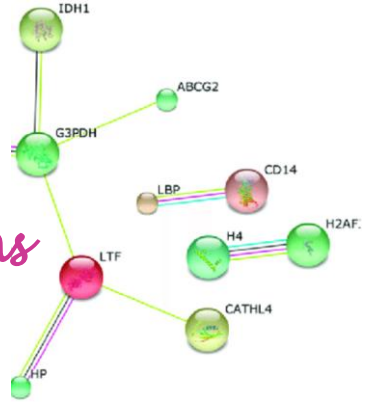
# APPLICATION FIELDS & DOMAINS

## Web Traffic

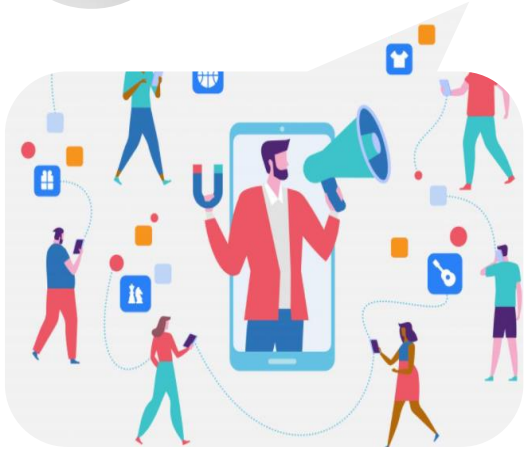
## Information Dissemination

## App Feeds

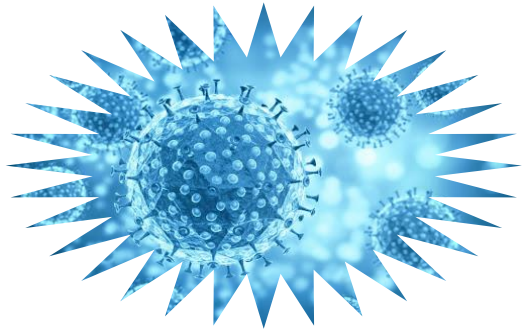

## Protein Interactions



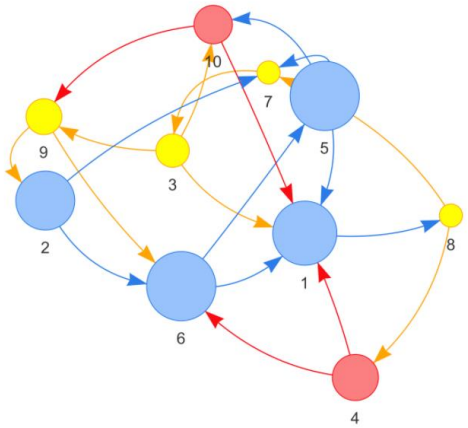

## Virality



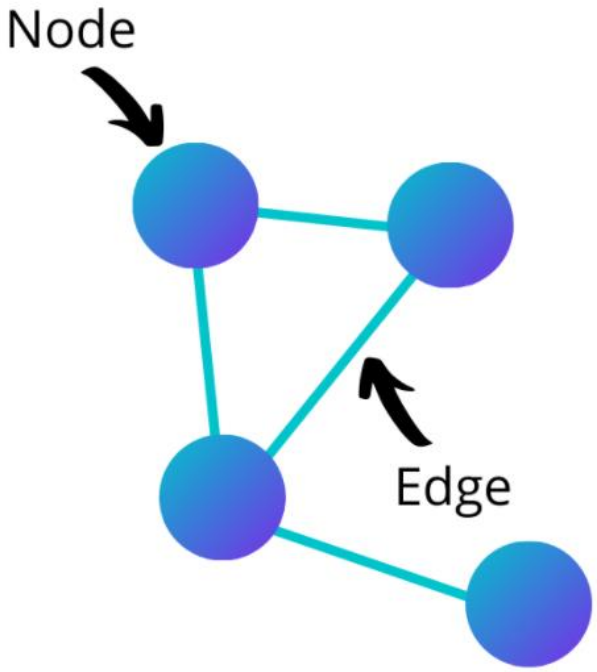
## Disease Transmission



# BASIC CONCEPTS



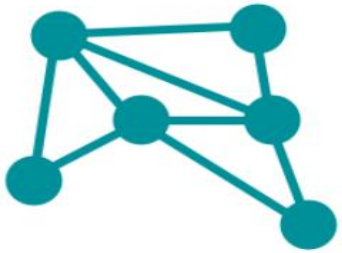
# TERMINOLOGY




Undirected Edges

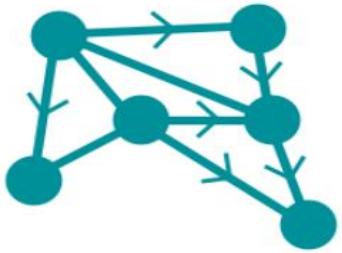
Directed Edges


Undirected Graph

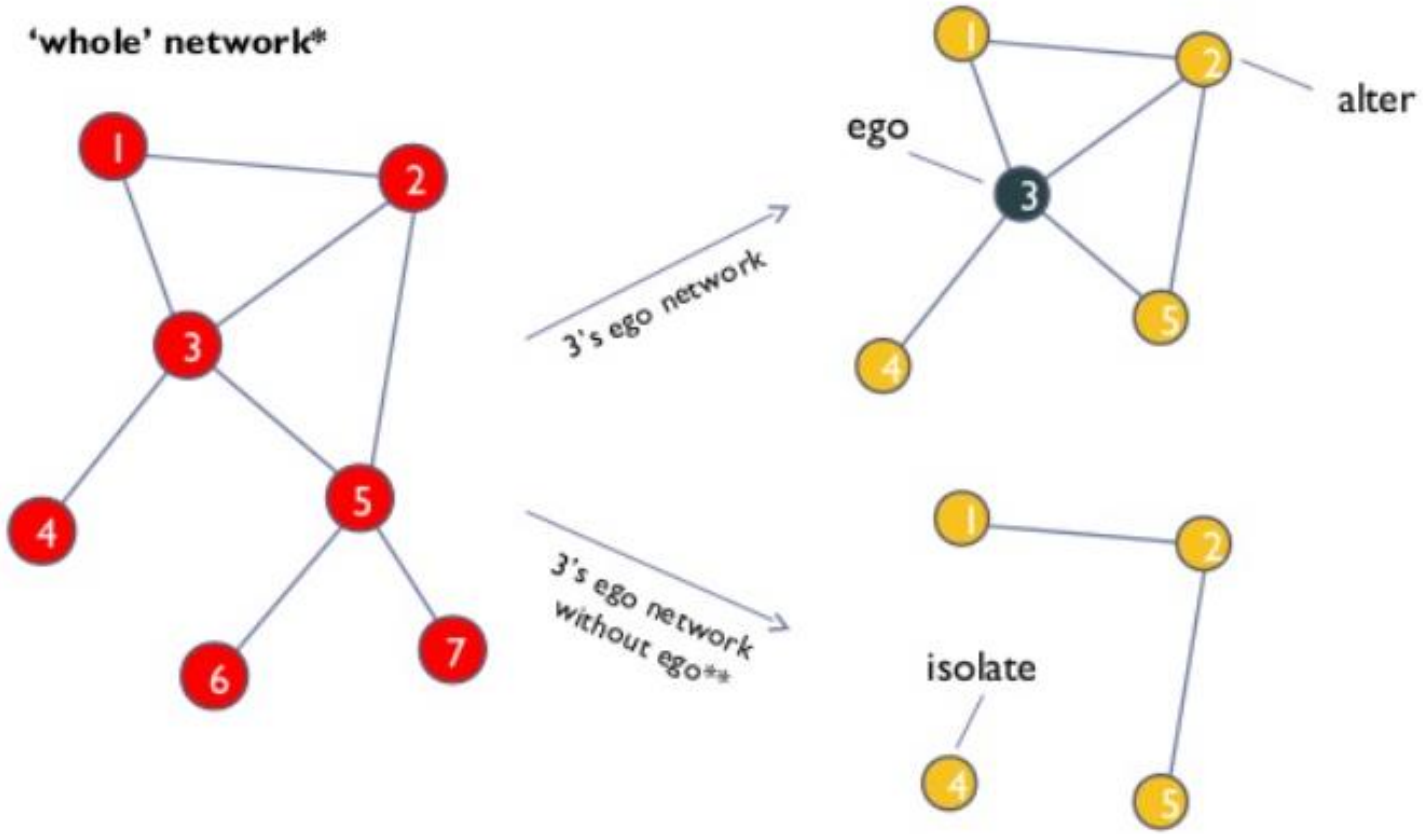


Directed Graph

aka Digraph



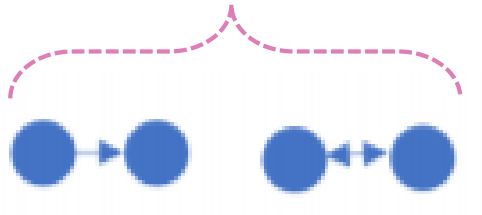
# EGO & ALTER


'whole' network*
1
2
3
4
5
6
7
3's ego network
ego
alter
3's ego network without ego**
isolate

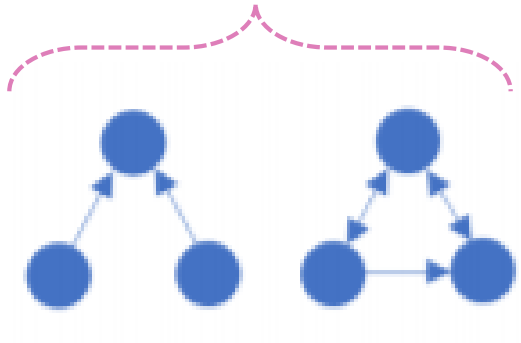
# NETWORK MOTIFS

*"**Network motifs** are recurrent and statistically significant subgraphs or patterns of a larger graph"*

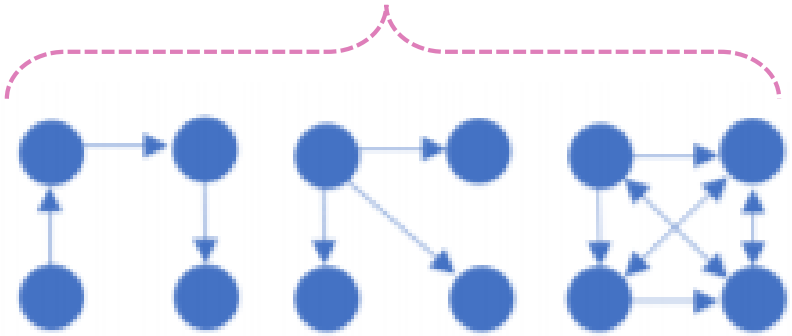
**Dyad**



**Triad**

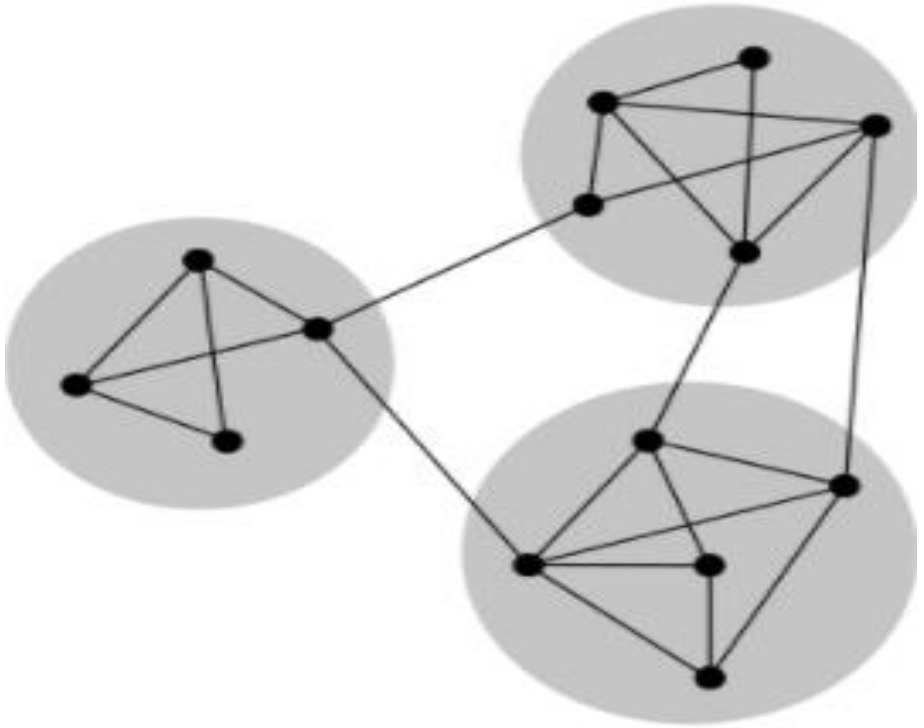


**Tetrads**



# COMMUNITY

*"A subset of nodes within the graph such that connections between the nodes are denser than connections with the rest of the network"*



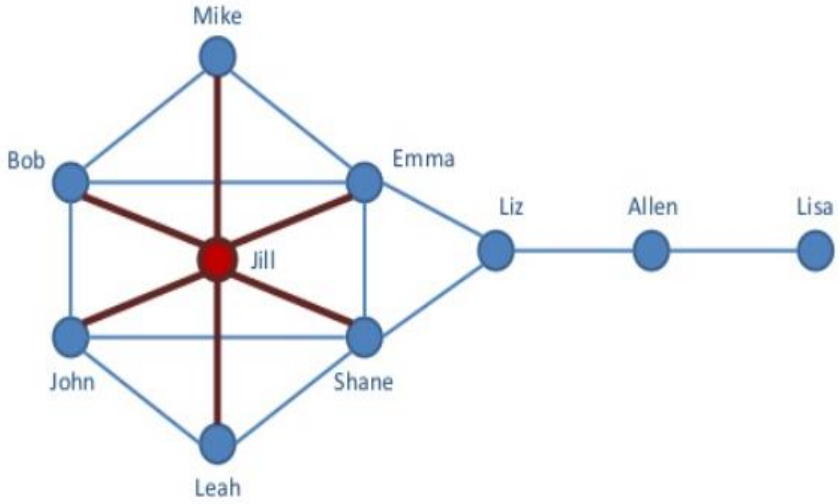
# CENTRALITY

# DEFINITION

- Centrality measures are a vital tool for identifying structurally important actors
- Centralization measures the extent to which the ties of a given network are concentrated on a single actor or group of actors
- Algorithms of graph theory cut through noisy data, revealing parts of the network that need attention using the Centrality measure

# DEGREE CENTRALITY

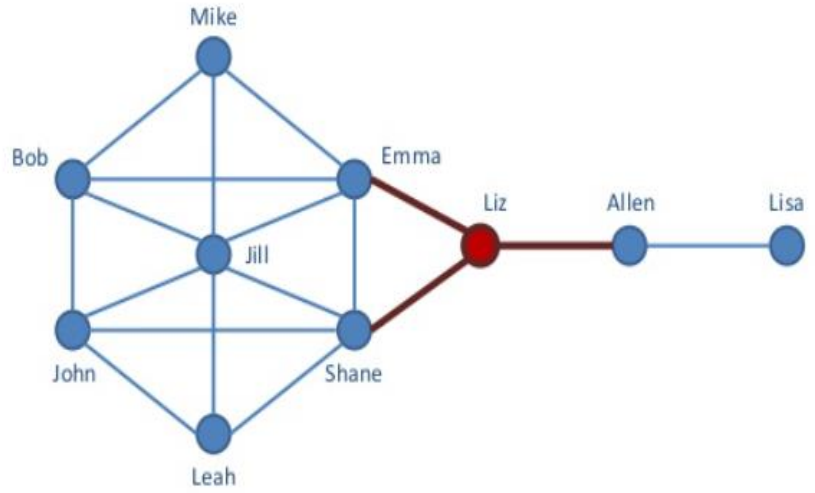
- Degree centrality assigns an importance score based simply on the number of links held by each node
- Used for finding very connected individuals, popular individuals, individuals who are likely to hold most information or individuals who can quickly connect with the wider network




What is the number of Connections?

# BETWEENNESS CENTRALITY

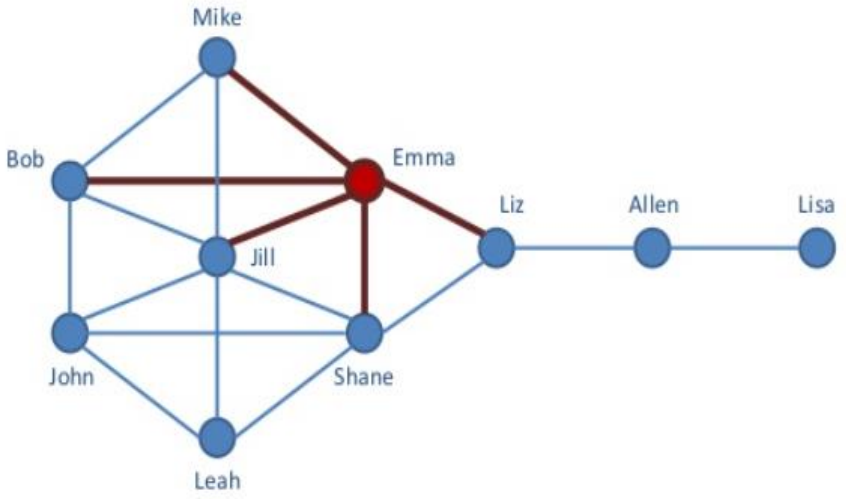
- Betweenness centrality measures the number of times a node lies on the shortest path between other nodes
- It shows which nodes are ‘bridges’ between nodes in a network
- Identifies all the shortest paths and then counts how many times each node falls on one




Which node has the most control over flow between nodes?

# CLOSENESS CENTRALITY

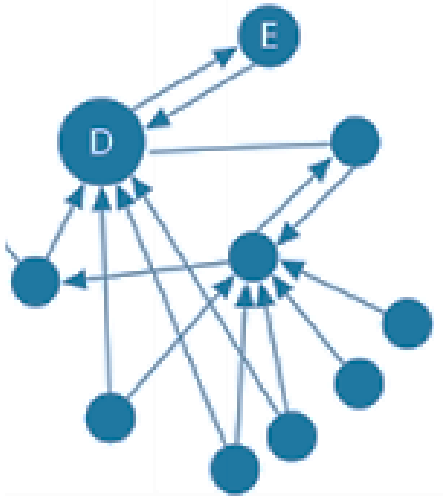
- Closeness centrality scores each node based on their 'closeness' to all other nodes in the network
- This measure calculates the shortest paths between all nodes, then assigns each node a score based on its sum of shortest paths




*Which node can most easily reach all other nodes in the graph?*

# EIGENVECTOR CENTRALITY

- Eigen Centrality measures a node's influence based on the number of links it has to other nodes in the network
- It goes a step further by considering how well connected a node is, and how many links their connections have
- Eigen Centrality can identify nodes with influence over the whole network, not just those directly connected to it

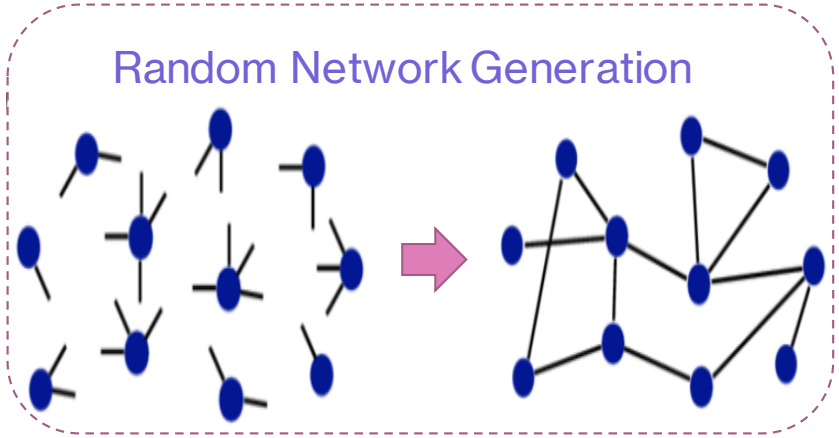



Which node is most important?

# MODULARITY

# DEFINITION

- Modularity (Q) is the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random
- Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules




$$Q = \frac{1}{(2m)} \sum_{vw} \left[ A_{vw} - \frac{k_v k_w}{(2m)} \right] \delta(c_v, c_w) = \sum_{i=1}^{c} (e_{ii} - a_i^2)$$

# DERIVATION - 1

Consider a graph with **n** nodes, **m** edges, **A** adjacency matrix

- Let the total number of stubs in the network be $l = \sum_u k_u = 2m$
- Expected number of edges between node v and w is $A_{vw} - \frac{k_v k_w}{2m}$
- Summing over all node pairs gives the equation for modularity, Q

$$Q = \frac{1}{(2m)} \sum_{vw} \left[ A_{vw} - \frac{k_v k_w}{(2m)} \right] \delta(c_v, c_w) = \sum_{i=1}^{c} (e_{ii} - a_i^2)$$

# DERIVATION - 2

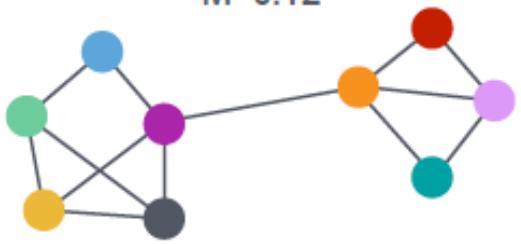
$$Q = \sum_{c \in C} \left[ \frac{\Sigma_{\text{in}}^{c}}{2m} - \left( \frac{\Sigma_{\text{tot}}^{c}}{2m} \right)^{2} \right]$$

$\Sigma_{\text{in}}^{c}$ — the sum of the weights from all internal edges of community $c_i$

$\Sigma_{\text{tot}}^{c}$ — the sum of the weights from edges incident to any vertex in $c_i$

# EXAMPLES

Negative Modularity
M=0.12



Single Community
M=0

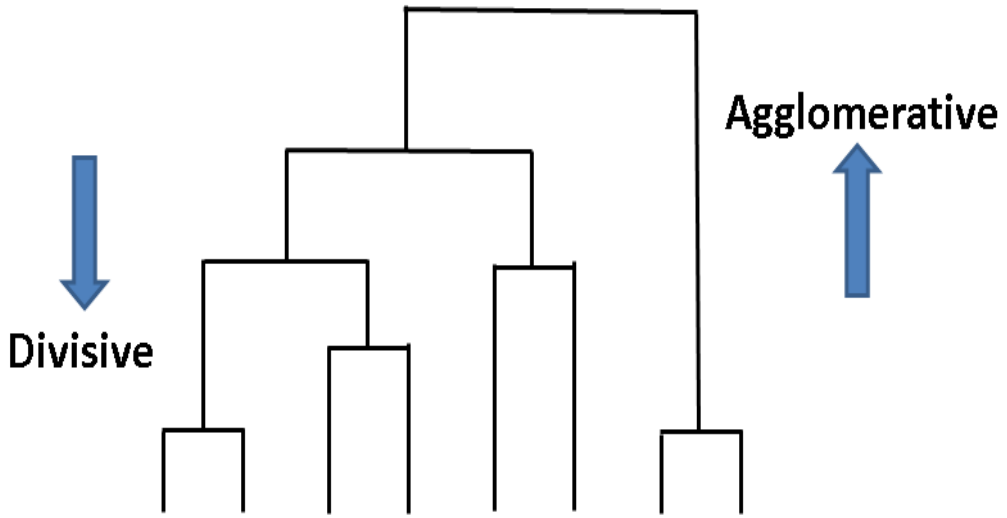
Suboptimal Partition
M=0.22

Optimal Partition
M=0.41

# COMMUNITY DETECTION

# AGGLOMERATIVE & DIVISIVE

- **Agglomerative:** This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy
- **Divisive:** This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy



# AGGLOMERATIVE - LOUVAIN

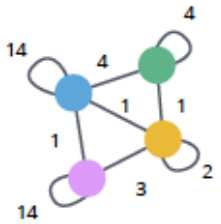
## Pass 1

Step 0

Choose a start node and calculate the change in modularity that would occur if that node joins and forms a community with each of its immediate neighbors.

Step 1

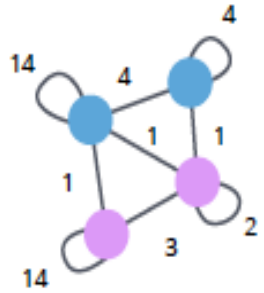
The start node joins the node with the highest modularity change. The process is repeated for each node with the above communities formed.




Step 2

Communities are aggregated to create super communities and the relationships between these super nodes are weighted as a sum of previous links. (Self-loops represent the previous relationships now hidden in the super node.)

## Pass 2



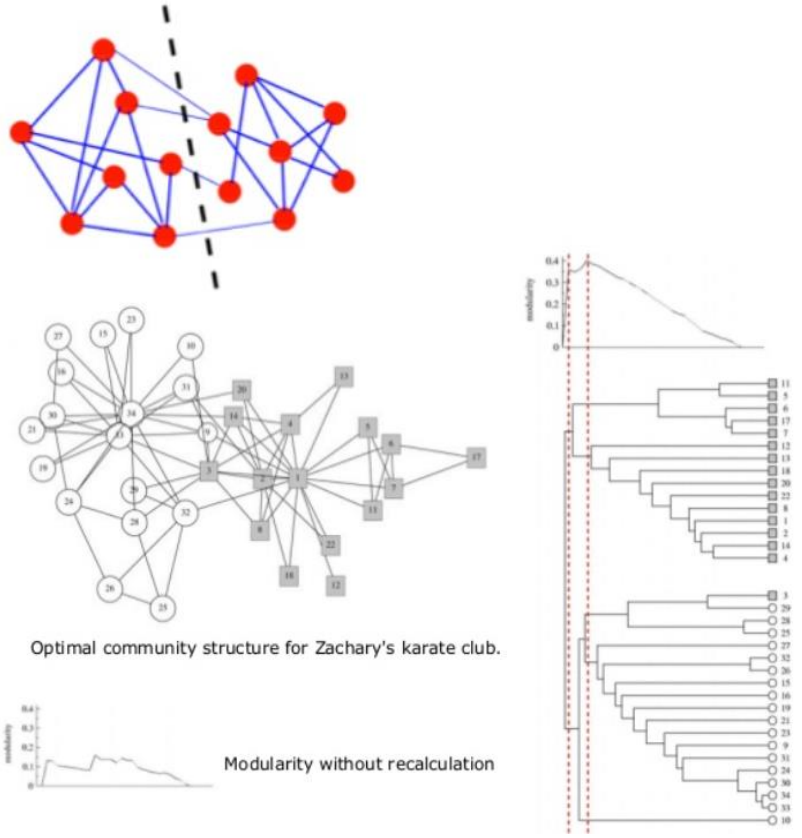

Step 1


Step 2

Steps 1 and 2 repeat in passes until there is no further increase in modularity or a set number of iterations have occurred.

# DIVISIVE – NEWMAN GIRVAN

1. The betweenness of all existing edges in the network is calculated first.
2. The edge(s) with the highest betweenness are removed.
3. The betweenness of all edges affected by the removal is recalculated.
4. Steps 2 and 3 are repeated until no edges remain.



# GOODNESS MEASURES

# MEASURES FOR GOOD COMMUNITIES

- **Separability** captures the intuition that good communities are well-separated from the rest of the network
- **Density** builds on intuition that good communities are well connected
- **Cohesiveness** characterizes the internal structure of the community. A good community should be relatively hard to split a community into two
- **Clustering coefficient** is a measure of the degree to which nodes in a graph tend to cluster together

# SCORING METRICS

# BASED ON INTERNAL CONNECTIVITY

- **Edges inside** : number of edges within a community S
- **Internal density:** $f(S) = \frac{m_S}{n_S(n_S-1)/2}$
- **Average degree :** $f(S) = \frac{2m_S}{n_S}$
- **Fraction over median degree :** Fraction of nodes of S that have internal degree higher than median
- **Triangle Participation Ratio** : Fraction of nodes in S that belong to a triad

# BASED ON EXTERNAL CONNECTIVITY

- **Expansion :** Number of edges per node that point outside community
- **Cut Ratio :** Fraction of existing edges (out of all possible edges) leaving the cluster

# BASED ON BOTH INTERNAL AND EXTERNAL CONNECTIVITY

- **Conductance** : Fraction of total edge volume that points outside the community
- **Normalized Cut**: $f(S) = \frac{c_S}{2m_S+c_S} + \frac{c_S}{2(m-m_S)+c_S}$
- **Maximum/Average Out Degree Fraction :** Maximum/Average fraction of edges of a node in S that point outside
- **Flake Out Degree Fraction :** Fraction of nodes in S that have fewer edges pointing inside than to the outside of the cluster

THANK

YOU