(Social) Network/Graph Analysis

- Major Classes of Graphs
- Network Descriptive Analytics
  - Local, Large-scale & Global features
  - Degree Distribution
  - Clustering coefficient & Transitivity
- Network Connectivity Analytics
  - Small World Phenomenon
  - Strong/Weak Ties and Bridges
- Network Centrality Analytics
  - Structural vs Functional Importance
  - Node Centrality Measures
- Large Scale Network Structure
  - Homo/Heterogeneous Mixing
  - Homophily and Heterophilly
  - Social Influence/Selection
  - Modularity and Communities
- Communities in Graphs
  - Clustering Nodes
  - Finding Communities
- Graph Representation Learning
  - Node2Vec & Graph2Vec
  - Convolution NN

Imdad ullah Khan

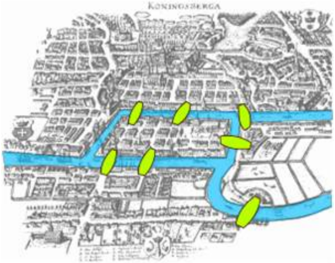
Graph Analytics

Depending on the domain of graphs and applications the area is also called

Network Analysis, Link Analysis, Social Network Analysis

Modeling, formulating and solving problems with graphs

Use tools from graph theory, linear algebra (algebraic graph theory) and algorithms for data analysis problems modeled with graphs



One of the earliest graph analysis: Euler argued that there is no way to tour the city of Konigsberg (now Kaliningrad) crossing each of the 7 bridges exactly once

Graph Analytics

Rather than individual data points or the global structure of the datasets

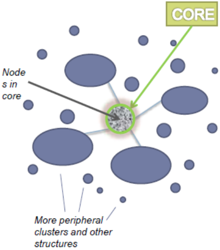
Graph Analysis focuses on pairwise interaction between data objects

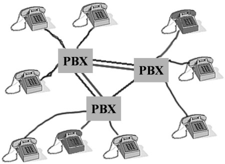
Allows to examine how pairwise interaction of entities in a network determine the behavior or function of an individual entity, groups of entities or the whole system

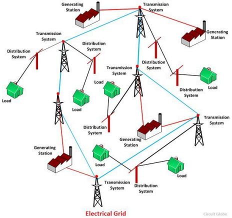
Graphs are everywhere: Six major classes of networks

Technological Networks

- Internet (Autonomous Systems connected with BGP connections)
- Telecom Network (telephone devices connected with wires or wireless)
- Power Grid (generating stations/loads, transmission line)



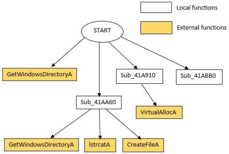


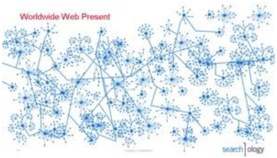


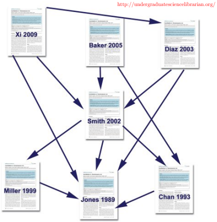
Graphs are everywhere: Six major classes of networks

Information Networks

- Software (functions connected with function calls)
- The Web Graph (webpages and hyperlinks)
- Documents (Research manuscripts, citations)



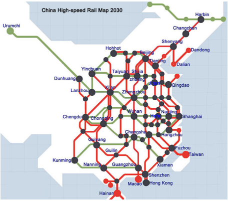


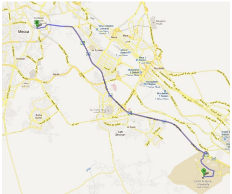


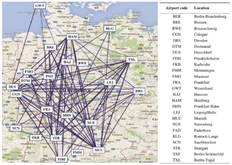
Graphs are everywhere: Six major classes of networks

Transportation Networks

- Railway System (train station and railroad tracks)
- Highway network (Intersection, road segments)
- Air Transportation (Airports, non-stop flight)







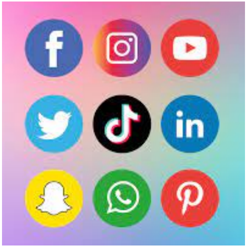
Airport code	Location
BER	Berlin-Brandenburg
BRE	Bremen
BWE	Braunschweig
CGN	Cologne
DRS	Dresden
DTM	Dortmund
DUS	Dusseldorf
FDH	Friedrichshafen
FKB	Karlsruhe
FMM	Memmingen
FMO	Muenster
FRA	Frankfurt
GWT	Westerland
HAJ	Hanover
HAM	Hamburg
HHN	Frankfurt Hahn
LEJ	Leipzig/Halle
MUC	Munich
NUE	Nuremberg
PAD	Paderborn
RLG	Rostock-Laage
SCN	Saarbruecken
STR	Stuttgart
SXF	Berlin-Schönefeld
TXL	Berlin-Tegel

Graphs are everywhere: Six major classes of networks

Social Networks

- Social Network (people, friendship/acquaintance/coworker)
- Online Social Network (people, friendship or co-worker relation)

Social Network



Online Social Nework

Graphs are everywhere: Six major classes of networks

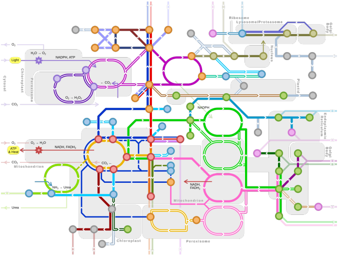
Biological Networks

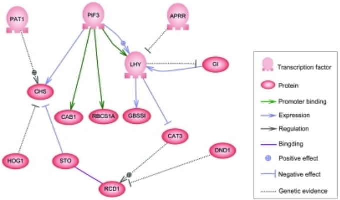
Represents interactions between biological units

ecological, evolutionary, physiological, metabolic, gene regulatory network

Most genes and proteins play a role through interactions with other proteins, genes, and biomolecules

Analyzed to understand the origin and function of cellular components, treatments for diseases, determine comorbidities and risk factors

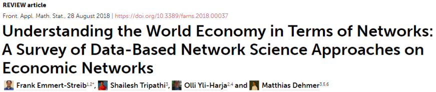




Graphs are everywhere: Six major classes of networks

Economic Networks

Business, companies, governments interacting via credit and investment, trade relations, supply chain

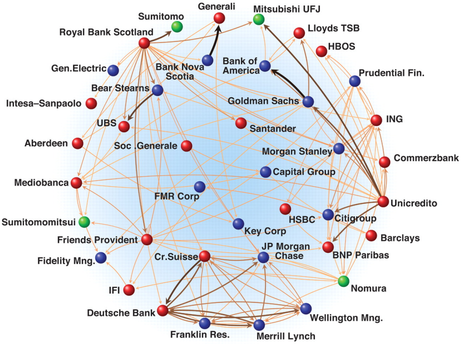


REVIEW article

Front. Appl. Math. Stat., 28 August 2018 | https://doi.org/10.3389/fams.2018.00037

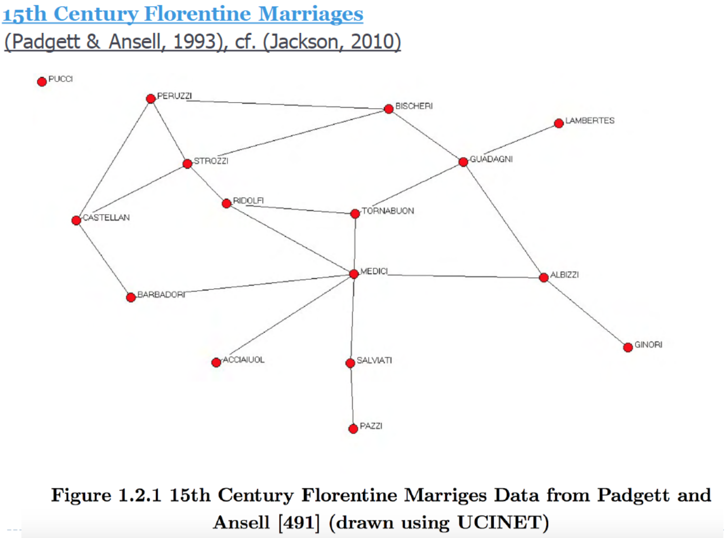
Understanding the World Economy in Terms of Networks: A Survey of Data-Based Network Science Approaches on Economic Networks

Frank Emmert-Streib[1,2*], Shailesh Tripathi[3], Olli Yli-Harja[2,4] and Matthias Dehmer[3,5,6]



F. Schweitzer et.al. (2009) Economic Networks: The New Challenges

Graphs are everywhere



15th Century Florentine Marriages

(Padgett & Ansell, 1993), cf. (Jackson, 2010)

Figure 1.2.1 15th Century Florentine Marriges Data from Padgett and Ansell [491] (drawn using UCINET)

Types of Graphs

Generally, Graphs are denoted as $G = (V, E)$

- V is a set of vertices ▷ aka nodes, points
- E is set of edges $E \subseteq V \times V$ ▷ aka links, lines

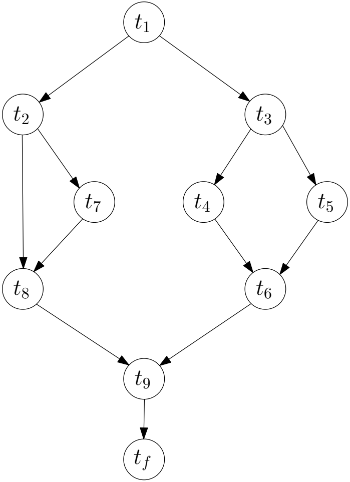
Graphs Types, Notation and Terminology

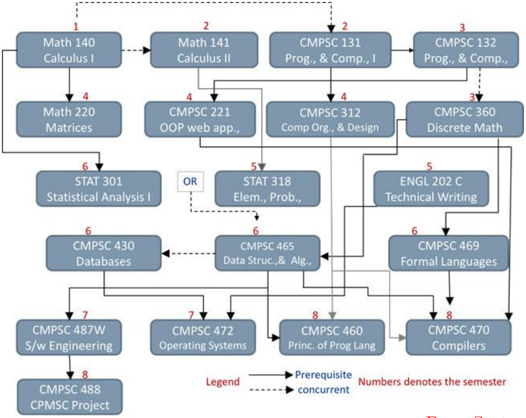
- Undirected Graphs
- Directed Graphs
- Weighted Graphs
- Pseudographs
- Multigraphs
- (Edge/Node) labeled graphs

- Adjacency
- Adjacency Matrix
- Adjacency List
- Incidence Matrix
- Weighted Adjacency List
- Density
- Degree, Out/In Degree

Directed Acyclic Graphs

In many applications graphs are DAGs



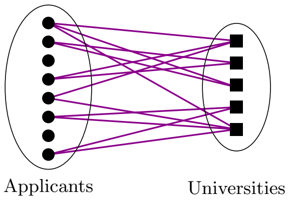


1
Math 140 Calculus I
2
Math 141 Calculus II
2
CMPSC 131 Prog., & Comp., I
3
CMPSC 132 Prog., & Comp.,
4
Math 220 Matrices
4
CMPSC 221 OOP web app.,
4
CMPSC 312 Comp Org., & Design
3
CMPSC 360 Discrete Math
6
STAT 301 Statistical Analysis I
OR
5
STAT 318 Elem., Prob.,
5
ENGL 202 C Technical Writing
6
CMPSC 430 Databases
6
CMPSC 465 Data Struc.,& Alg.,
6
CMPSC 469 Formal Languages
7
CMPSC 487W S/w Engineering
7
CMPSC 472 Operating Systems
8
CMPSC 460 Princ. of Prog Lang
8
CMPSC 470 Compilers
8
CMPSC 488 CPMSC Project
Legend
Prerequisite
concurrent
Numbers denotes the semester

Penn State

Bipartite Graphs

In many applications the graphs are bipartite

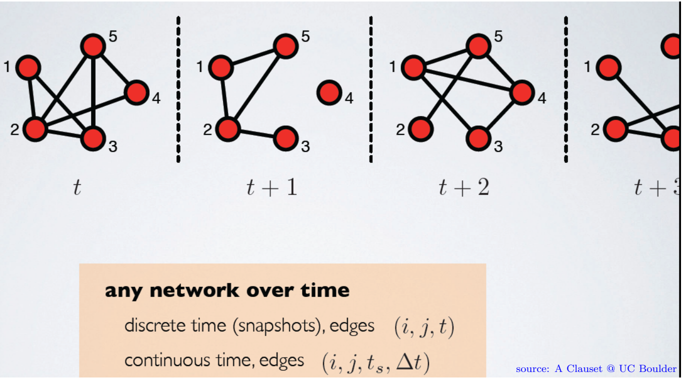


- Actors & Movies
- Artists & Albums
- Authors & Papers
- Users & Online groups
- Words & Documents
- Users & Checkins locations
- Metabolites & Reactions

Evolving, Dynamic, and Temporal Graphs

Graphs may be dynamic

e.g. new users on Twitter, new friendship on Facebook, new call in call graphs



any network over time

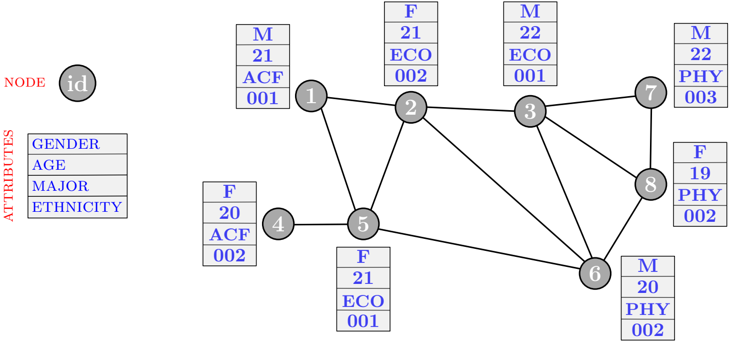
discrete time (snapshots), edges (i, j, t)

continuous time, edges $(i, j, t_s, \Delta t)$

source: A Clauset @ UC Boulder

Attributed Graphs

- Each element (vertex/edge) has associated properties
- It can be directed/undirected



Network Analysis: Applications

Network Analysis: Applications

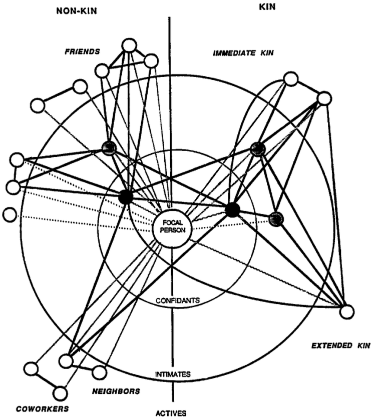
- Online Social Networks
  - Identify groups (communities) and group interactions
  - Find influencers in community
  - Extract topics of interests
- Biology (Biological Entities Networks)
  - Discover unknown relationships (disease to disease etc.)
  - Exploratory data analysis and Anomaly detection
- Geo Information System (Smart Cities)
  - Coverage analysis
  - Traffic flow, congestion estimation, routing
  - Failure impact analysis
- Reasoning (Predictive Maintenance)
  - Predict the next state given the current (and previous states)
  - Compute the probability of sequence of events

Network Analysis: Applications

- Computer Science: webgraph, Internet, Information dissemination
- Life Sciences: Protein-Protein Interaction Network, Foodchain in ecosystem
- Businesses Analyze and improve communication flow within and between organizations
- Advertisers and Marketers Figure out the most influential people in a Social Network and rout message through them
- Security and Law Enforcement: Identify criminal networks from traces (call-logs), find identify key players in such networks
- Find unusual patterns in the flow of money across interconnected Banking networks to identify fraudulent transactions, money laundering, terror financing
- Online Social Networks: Find communities and interest groups, recommend friends
- Mobile Network Operators: Optimize network structure to enhance QoS, analyze Cell towers to ensure maximum coverage

Social Network Analysis

- Network Perspective of Society
- How the Individuals', communities' and society's behavior is influenced by their social connectivity

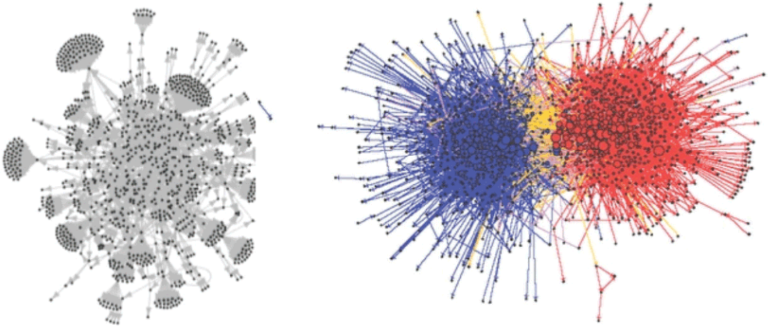


An early use of network analysis in sociology. This diagram of the 'ego-network' shows varying tie strengths in concentric circles-Wellman 1998

Social Network Analysis

- Network Perspective of Political discourse

source: Ulicny, Kokar, Matheus, (2010)



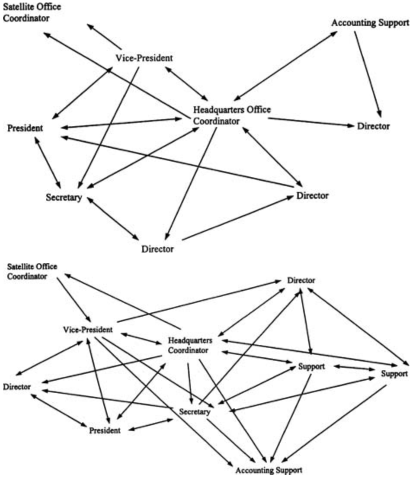
(a) Malaysian Sopo blogosphere

(b) US political blogosphere (Adamic & Glance, 2005[4])

A visualization of Malysia and US blogosphere (nodes are blogs and edges are links to blogs). Left reveals importance/credibility/popularity of blogs, while the right visual clearly show that bloggers are more likely to link to bloggers with the same party affiliations, forming two dense clusters with little interaction with the other cluster

Social Network Analysis

- Communication within an organization

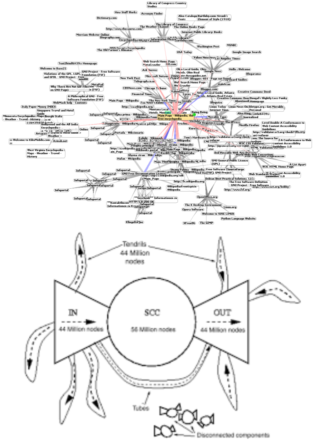


Intra-organization communication before and after implementation of a content management system (Garon et.al. (1997)

Web Structure

- Broder et.al. (2000) SCC analysis of webgraph (AltaVista index)
- Study replicated for larger recent webgraphs reveal similar structure

- $\sim 200m$ pages, $\sim 1.5b$ links
- bow-tie structure (macroscopic)
- grouping of SCC's
- CORE: a giant SCC ($\sim 56m$) nodes
- IN: can reach CORE (unidirectional)
- OUT: can be reached from CORE
- TENDRILS:
  - reachable from IN cannot reach CORE
  - can reach OUT not reachable from CORE
- TUBES: both types of tendrils
- DISCONNECTED COMPONENTS



The bow-tie structure of the web (A. Broder et.al (2000))

Early Network Analysis in Social Science

Linton Freeman (1996), Some Antecedents of Social Network Analysis

Network analysis is applied in Educational Psychology, Child Development, Sociology, Anthropology, Political Science, Information Science

> Society is not a mere sum of individuals. Rather, the system formed by their association represents a specific reality which has its own characteristics... The group thinks, feels, and acts quite differently from the way in which its members would were they isolated. If, then, we begin with the individual, we shall be able to understand nothing of what takes place in the group
>
> Émile Durkheim, The Rules of Sociological Method (1895)

Dukheim defined "social facts"- a phenomenon that is created by interactions of individuals but it is independent of any individual.

Network Descriptive Analytics

Network Descriptive Analytics

What is the structure of the Network? How does the network look like?

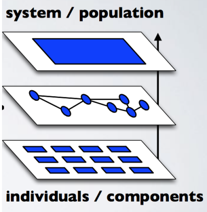
- What is the magnitude of the graph?
- How are the edges organized?
- How do vertices differ?
- Does Network Location matter?
- Are there underlying patterns

- What process shape these networks
- What is the underlying reason for this structure
- How can we exploit the structural features of the network

Network Descriptive Analytics

Describe features of the network

- Local-Level Features
- Large scale-features
- Global Features



How does the network structure shape the network function?

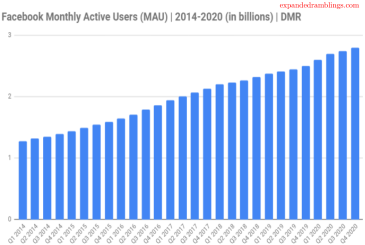
Network Descriptive Analytics

Describe features of the network

- Order, size, density
- Average degree, degree sequence, degree distribution
- Vertex positions and centrality
- Shortest loop density (triangles)
- Connectivity of the network
- Shortest path, radius, diameter
- Small world graphs

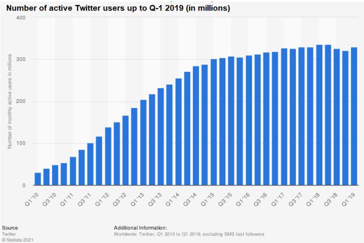
Network Descriptive Analytics: Magnitude

Order of Graph: Number of Nodes or vertices in the Network



expandedramblings.com

Facebook Monthly Active Users (MAU) | 2014-2020 (in billions) | DMR



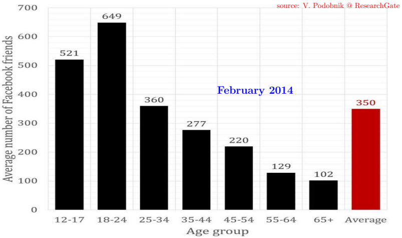
Number of active Twitter users up to Q-1 2019 (in millions)

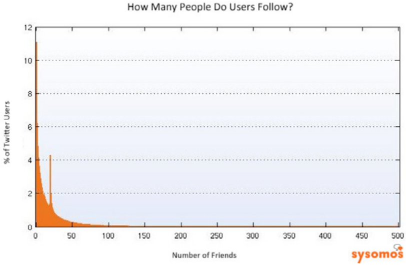
Source
Twitter
© Statista 2021

Additional Information:
Worldwide; Twitter; Q1 2010 to Q1 2019; excluding SMS fast followers

Network Descriptive Analytics: Magnitude

Size of Graph: Number of Edges in the Network



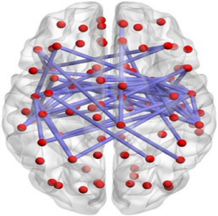


Network magnitudes have direct impact on the storage, computation, and communication, visualization complexity of the network

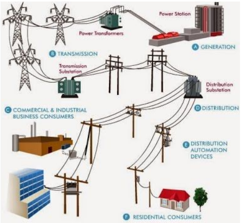
Magnitude of Networks

Big Graph Data Analytics or Massive graph analysis deals with graphs on millions of nodes and billions of edges

World Wide Web



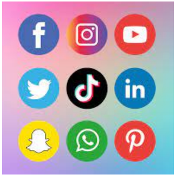
Biological Network



Power Network

Social Network

The Internet



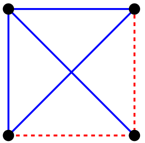
Online Social Nework

Density of Networks

Density of Graph: Ratio of Number of Edges present in the network to number of edges possible in (simple) network

Suppose $G = (V, E)$, $|V| = n$, $|E| = m$

$$d(G) = \frac{m}{\binom{n}{2}}$$



$\mathbf{d(G) = \frac{4}{6}}$

- A single parameter to compare connectivity of two graphs
- Very useful in comparing subgraphs - Clusters are dense subgraphs
- A clique has density 1, an independent set has density 0

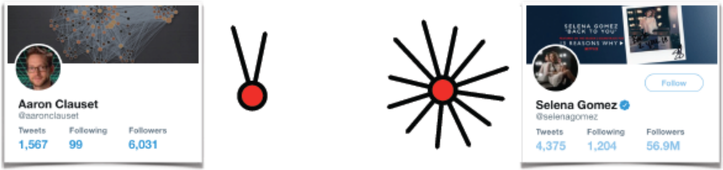
Degrees and Average Degree

Degree of a vertex v : is number of edges adjacent to v, $d(v) = |N(v)|$

Recall the notion of in-degree and out-degree, in and out-neighborhood in digraphs

In a bipartite graph $G = (A, B, E)$ degree of $v \in A$ is usually normalized by $|B|$

source: Aaron Clauset @ UC Boulder



Aaron Clauset
@aaronclauset

Tweets	Following	Followers
1,567	99	6,031

Selena Gomez
@selenagomez

Tweets	Following	Followers
4,375	1,204	56.9M

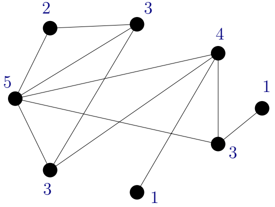
Average degree is the average degree over all vertices

$$d_{av}(G) = \frac{\sum_{v \in V} d(v)}{n} = \frac{2m}{n} = 2d(G)$$

Exploring Degree Distribution

Degree Distribution

$Pr(k)$: Probability that a randomly chosen vertex has degree k



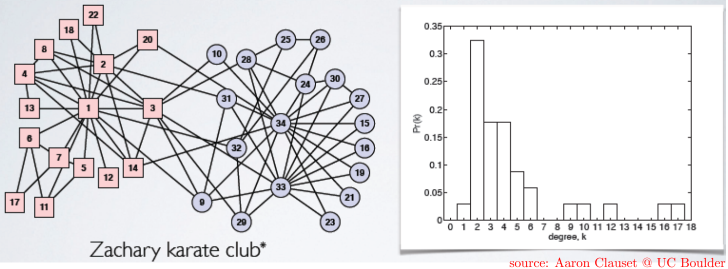
k	$Pr(k)$
1	$^2/_8$
2	$^1/_8$
3	$^3/_8$
4	$^1/_8$
5	$^1/_8$
6	0
7	0

- Degree distribution is used to determine the appropriate synthetic graph generation, to explain the observed structural patterns
- Typically real-wold graphs have right-skewed degree distributions

Exploring Degree Distribution

Degree Distribution

$Pr(k)$: Probability that a randomly chosen vertex has degree k



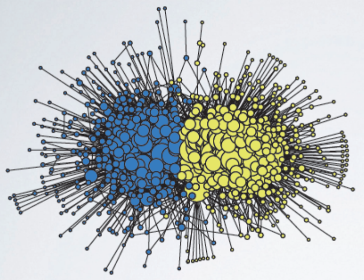
source: Aaron Clauset @ UC Boulder

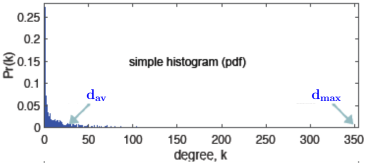
Exploring Degree Distribution

Degree Distribution

$Pr(k)$: Probability that a randomly chosen vertex has degree k

Simple pdf: $Pr(deg = k)$ vs. k



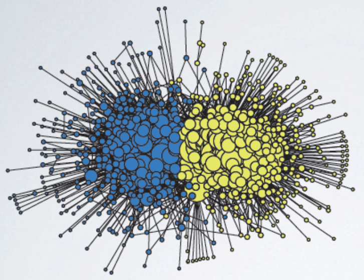


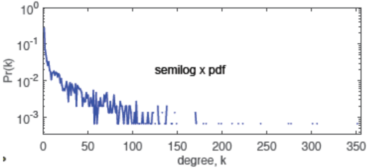
Exploring Degree Distribution

Degree Distribution

$Pr(k)$: Probability that a randomly chosen vertex has degree k

Semilog x pdf: $\log_{10} Pr(deg = k)$ vs. k



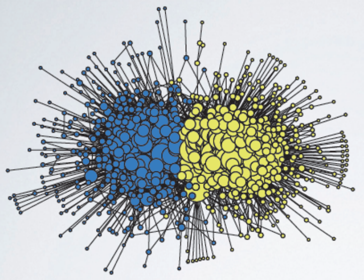


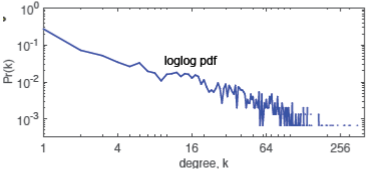
Exploring Degree Distribution

Degree Distribution

$Pr(k)$: Probability that a randomly chosen vertex has degree k

loglog x pdf: $\log Pr(deg = k)$ vs. $\log k$





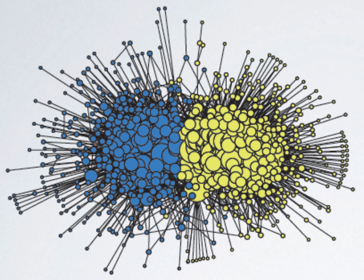
log plots are good for variables with high variance (in one or both dimensions)

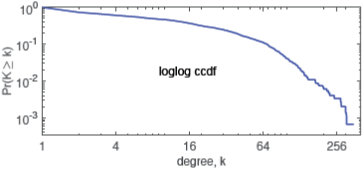
Exploring Degree Distribution

Degree Distribution

$Pr(k)$: Probability that a randomly chosen vertex has degree k

complementary cdf: $Pr(deg \geq k) = \sum_{i=k}^{n} Pr(deg = j)$ vs. k





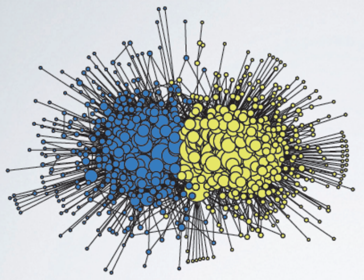
Complementary cdf is monotonic, smoother than pdf and reveals more info

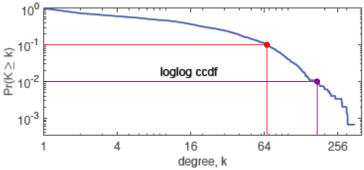
Exploring Degree Distribution

Degree Distribution

$Pr(k)$: Probability that a randomly chosen vertex has degree k

complementary cdf: $Pr(deg \geq k) = \sum_{i=k}^{n} Pr(deg = j)$ vs. k





90% vertices have degree ≤ 67 (accounting for 53% all edges)

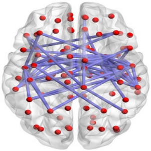
1% vertices have degree ≥ 169 (accounting for 90% all edges)

Complementary cdf is monotonic, smoother than pdf and reveals more info

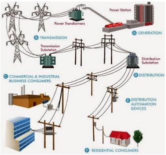
Exploring Degree Distribution

Nearly all real-world networks have a heavy-tailed degree distribution

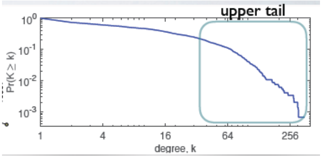
World Wide Web



Biological Network

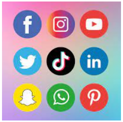


Power Network

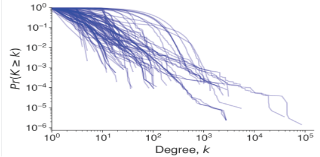


Social Network

The Internet

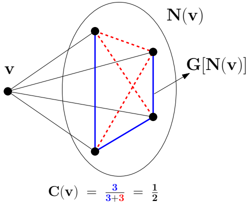


Online Social Nework



Clustering Coefficient

Clustering Coefficient of v $C(v) = \frac{|E(G[N(v)])|}{\binom{d(v)}{2}}$



$E(G[N(v)])$ is edges in graph induced by $N(v)$

Clustering coefficient of a graph $C(G)$ is defined to be the average clustering coefficient overall vertices

In some text $C(G)$ is defined as

$$C(G) = \frac{3 \times t(G)}{\sum_v \binom{d(v)}{2}}$$

This is sum of numerators and sum of denominators in $C(v)$ as $t(G)$ is the number of triangles in G. This latter quantity is also called transitivity

Transitivity

- Transitivity is the overall probability for the network to have adjacent nodes interconnected
- Reveals existence of tightly connected communities (clusters/cliques)
- Complex networks and notably small-world networks often have a high transitivity and a low diameter

	No. of existing triangles ($t0$)	No. of possible triangles ($t1$)	Transitivity ($t0/t1$)
A	0	4	0
B	1	4	0.25
C	2	4	0.5
D	3	4	0.75
E	4	4	1

Network Connectivity Analytics

Graph Connectivity

Recall the concepts about graph connectivity

- Paths, Walks, Cycles
- Connected and reachable vertices
- Connected graphs, connected components and strongly connected components

Connectivity Analytics

- Highly connected nodes (Nodes with high in/out-degree)
- Graph robustness (How easy to break the graph by removing a few nodes/edges "build-in redundancy")
  - Connectivity coefficient
    - Minimum number of nodes needed to remove to disconnect a graph (useful in network fragility analysis and social media advertisement)
  - Connectivity
    - Node X is reachable from node Y, or node Y is reachable from node X
  - Strong connectivity
    - Node X is reachable from node Y **AND** node Y is reachable from node X (high degree nodes make the network more vulnerable)

Connectivity Analytics

- Fully connected graph
  - Each node has edges to all other nodes in the graph
- Clique
  - It is a subset of vertices of an undirected graph such that every two distinct vertices in the subset are adjacent
- Terminal node
  - A node with no outgoing edges
- Unreachable node
  - A node with no ingoing edges
- Hub vs Authorities
  - Hub: A node with highest in-degree
  - Authorities: A node with highest out-degree
  - Example: Social networks (Talkers vs Listeners)

Shortest Path (Geodesic Distance)

Geodesic distance $d(u,v)$: distance between u and v is the length of the shortest path b/w u and v

Underlying assumption that things being equal communication takes place using shortest paths

Average distance is the average shortest path over all pairs of vertices

$$\ell_G \;=\; \frac{\sum_{u,v \in V} d(u,v)}{\binom{n}{2}}$$

- A measure of network cohesion, efficiency of communication
- Indicates how far apart any two nodes are on average

Radius and Diameter of a Network

Measures of graph connectivity (reachability) compare it network density

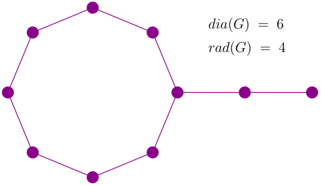
Network diameter: The longest geodesic distance between a pair of vertices

$$dia(G) = \max_v \max_u d(v,u)$$

Network radius: The minimum maximum distance of any vertex

$$rad(G) = \min_v \max_u d(v,u)$$

Denser networks are likely to have small diameter and vice versa



$dia(G) = 6$

$rad(G) = 4$

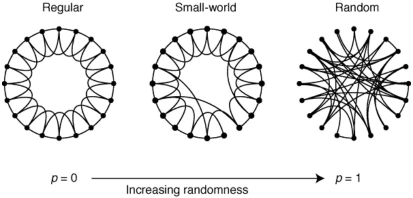
Small World Phenomenon

Real-world networks though are very sparse yet their diameters are typically small - Small world phenomenon

A small world network is one in which most pairs of nodes are not adjacent (sparse) but they have larger clustering coefficients and pairs are reachable in a few hops (low diameter)

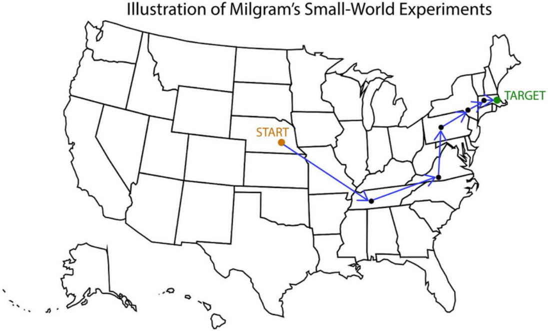
Small world graphs are in-between random graphs and regular graphs

- Regular graphs: All nodes have equal degrees
- Erdös-Renyi graph $G(n, p)$: n nodes and a pair is adjacent with prob. p
- $G(n, m)$ graphs: Randomly from all graphs on n nodes and m edges



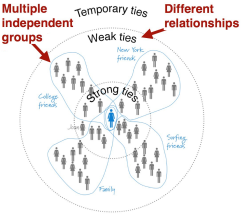
Small World: Milgram's Experiment

- Letters were given to people in Omaha NB to be sent to a target in Boston, MA
- Instructed to pass on to someone they knew on first name basis
- Average lengths of successful chain was about 6
- Many did not reach and many reached via the same intermediaries

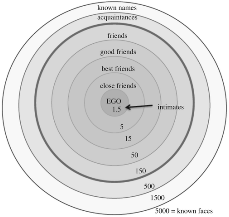


Strength of Weak Ties

In many networks all edges are not the same



source: towardsdatascience.com



https://royalsocietypublishing.org/doi/10.1098/rspa.2020.0446

- Structure of human egocentric social networks
- Number of people included in each circle increases, but the frequency of contact and emotional closeness declines, with each layer
- The outermost layer (5000) was identified by face recognition experiment (Num of faces that can be recognized as known by sight)
- Biological networks: tie strength based on biochemical interaction
- Computer networks: based on link bandwidth

Strength of Weak Ties

Granovetter,'s "The strength of weak ties": "Most job seekers (study subjects) found jobs through an acquaintance (weak tie), rather than a close friend (strong tie)"

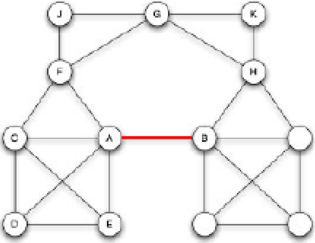
- Information at end-points of a strong tie is nearly identical
  ▷ frequent synchronization
- weak tie could help communication of novel information
  ▷ rare synchronization
- Acquaintance can more likely inform of "new" job opportunities

There are some vertices (and edges) that act as bridges between network segments, they are important for communication and explain the small-world phenomena in many network

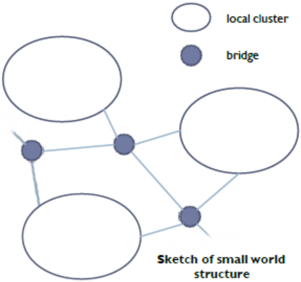
Bridges

There are some edges that act as bridges between network segments, they are important for communication and explain the small-world phenomena in many networks

An edge (i, j) is a local bridge if i and j have no friends in common



source: Frank Dignum @ Umea University



Preferential Attachment

- A mechanism in which a quantity (e.g. wealth, credit, degree) is distributed among objects according to how much they already have
- aka *rich gets richer, early bird advantage, cumulative advantage*

source: Dr. Giorgos Cheliotis @ NUS

Popularity

We want to be associated with popular people, ideas, items, thus further increasing their popularity, irrespective of any objective, measurable characteristics

Also known as 'the rich get richer'

Quality

We evaluate people and everything else based on objective quality criteria, so higher quality nodes will naturally attract more attention, faster

Also known as 'the good get better'

Mixed model

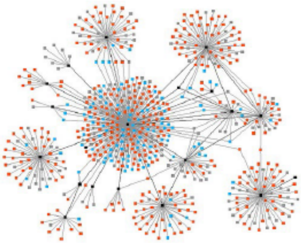
Among nodes of similar attributes, those that reach critical mass first will become 'stars' with many friends and followers ('halo effect')

May be impossible to predict who will become a star, even if quality matters

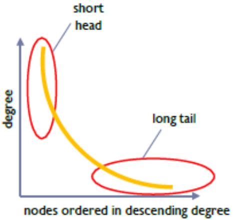
Preferential Attachment

In networks generated via preferential attachment process a great majority of new edges are incident to nodes with an already high degrees - degrees of these nodes increase disproportionately

- Results in network with few very high and majority low degrees nodes
- These networks have long-tailed degree distribution
- Tend to have small-world structure
- Transitivity and strong/weak tie characteristics are not necessary to explain small-world structure



source: Dr. Giorgos Cheliotis @ NUS



Barabási–Albert model

- Generate random networks using preferential attachment process
- WWW, citation networks, the Internet, and some OSN have long-tail degree distribution ▷ BA model tries to explain them

Initialize with a complete graph on m_0 nodes

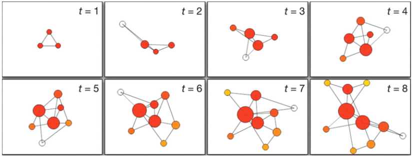
Each new node has $m \leq m_0$ edges (dangling)

A dangling edge is adjacent to an existing node v_i with probability p_i

$$p_i = \frac{d(v_i)}{\sum_j d(v_j)}$$

High degree nodes (rich nodes) quickly accumulate more edges (get richer)

Barabassi: (2009) Scale-Free Networks: A Decade and Beyond



- degree distribution $P(k) \sim k^{-3}$ ▷ Power law with scale parameter 3

Network Centrality Analytics

Centrality Measures

- Importance (functional) role of network players is often related to their (structural) position in the network
- It can significantly differ from presumption about importance e.g. fathers, mothers, executives, teachers, ...
- Centrality analytics undertakes quantitative social network analysis to determine types of actors and find key players

Structural and functional (dynamic) importance in essence are $f : V \mapsto \mathbb{R}$

- Can be used to identify influential actors
- Robustness and vulnerability of network
- Determine exposure of nodes to disease or their role in immunization
- Study of spread and countering epidemic

Vertex Importance Analysis

- Vulnerability
  - Node x belongs to minimum nodes set that, if removed from the graph, the graph is disconnected
    - Example: Removal of x will cause a high disruption in the network
- Network Centralization (graph-level measure)
  - Measure of variation of centrality score amongst network nodes

Centrality Measures

Measuring dynamic importance (influence) is expensive (may be impossible)

It is generally correlated with structural importance

To avoid the expensive computation and extensive domain knowledge we approximate functional importance with structural importance

Structural positions of nodes in a network are called centrality measures

In digraphs (e.g. Twitter or Web) they are called prestige

Node Centrality in Graphs

Node centrality

- Degree centrality: how many neighbors a node has

$$C_d(v) \;:=\; deg(v)$$

- Closeness centrality: how “close” a node is to other nodes

$$C_{close}(v) \;:=\; \frac{1}{\sum\limits_{u \neq v \in V} d_G(v,u)}$$

- Betweenness centrality: how often a nodeis on the shortest paths

$$C_{bw}(v) \;:=\; \sum_{s,t \neq v \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

$\sigma_{st}(v)$: number of shortest paths between s and t through v
σ_{st}: number of shortest paths between s and t

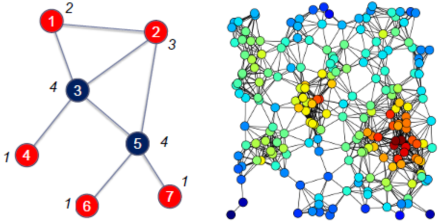
- Eigenvector centrality: Value of eigenvector at corresponding coordinate

$\mathbf{x}[i]$, $\mathbf{x}$: eigen vector correspond. to leading eigenvalue

Centrality Measures: Degree Centrality

Degree centrality: How many nodes can this node reach directly?

$$C_d(v) \ := \ deg(v)$$



- In a digraph we often use in-degree
- In Twitter or Web graph amounts to nodes popularity or influence
- Useful to determine important nodes for spreading information and influencing others in their immediate neighborhood

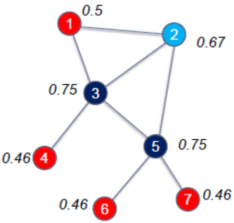
Centrality Measures: Closeness Centrality

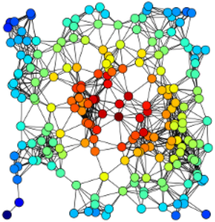
Closeness centrality: How fast can this node reach other nodes?

$$C_{close}(v) \;:=\; \frac{1}{\sum_u d(v,u)} \qquad \text{or} \qquad C_{close}(v) \;:=\; \frac{|V|}{\sum_u d(v,u)}$$

Assuming communication happens via shortest paths only, high closeness centrality nodes can reach other nodes the easiest- A measure of reach

To compare graph of varying orders one usually normalize





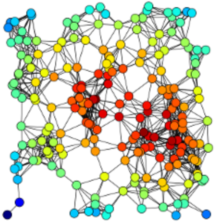
Centrality Measures: Harmonic Centrality

Harmonic centrality: How fast can this node reach other nodes?

$$H(v) \ := \ \sum_{u} \frac{1}{d(v,u)}$$

- Flips the sum and reciprocal
- ${}^{1}\!/_{d(x,y)} = 0$ if x and y are not connected

source: Wikipedia



Harmonic Centrality

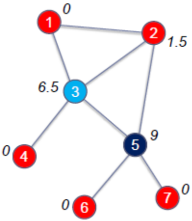
Centrality Measures: Betweenness Centrality

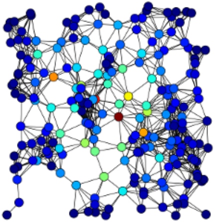
Betweenness centrality: likelihood of node to be on communication path

$$C_{bw}(v) \; := \; \sum_{s,t \neq v} \frac{\lambda_{st}(v)}{\lambda_{st}}$$

- $\lambda_{st}(v)$: Number of shortest path between s and t via v
- λ_{st}: Number of shortest path between s and t

Assuming communication happens via shortest paths only, high betweenness centrality nodes are critical for information flow





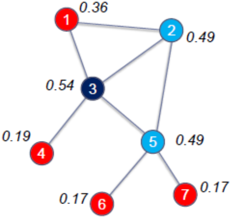
Centrality Measures: Eigenvector Centrality

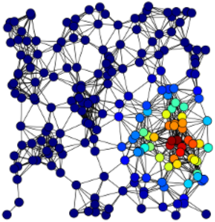
Eigenvector centrality: the node's connectivity to "well-connected" nodes?

Proportional to sum of eigencentralities of neighbors

$$\mathbf{c}(v) \;:=\; \frac{1}{\lambda} \sum_{u \in N(v)} \mathbf{c}(u)$$

- λ is a constant ▷ leading eigen value
- Computed as $A\mathbf{c} = \lambda\mathbf{c}$ ▷ A is the adjacency matrix





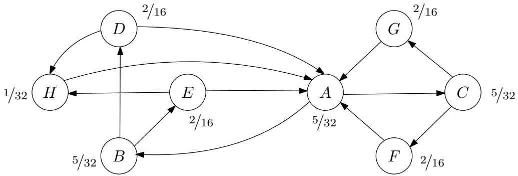
Centrality Measures: Pagerank Centrality

Eigenvector centrality: the node's connectivity to "well-connected" nodes?

Proportional to weighted sum of pagerank of out-neighbors

$$\mathbf{c}(v) \; := \; \alpha \sum_{u \in N^-(v)} \frac{\mathbf{c}(u)}{deg^+(u)} + \frac{1-\alpha}{|V|}$$

- α is the damping factor
- Probability of a random walker to visit the node



Centrality Measures:Clustering Coefficient

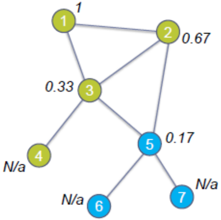
Clustering Coefficient: which nodes in the graph tend to cluster together

$$C(v) = \frac{|E(G[N(v)])|}{\binom{d(v)}{2}}$$

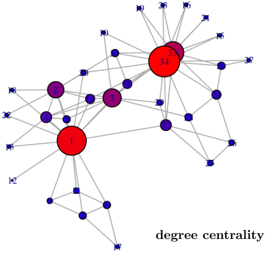
- $E(G[N(v)])$ is edges in graph induced by $N(v)$

number of triangles around a node v (friendships b/w v's friends)

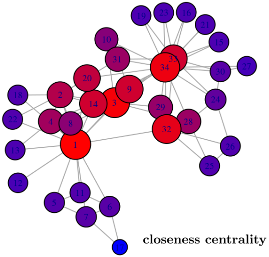
Closely related to **transitivity** of a graph - ratio of observed number of closed triangles/triplets and max possible number of closed triplets



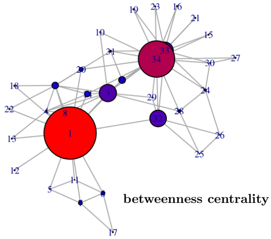
Node Centrality in Graphs



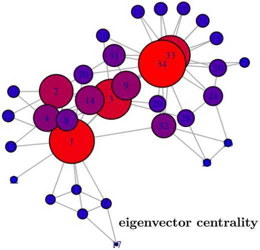
degree centrality



closeness centrality



betweenness centrality



eigenvector centrality

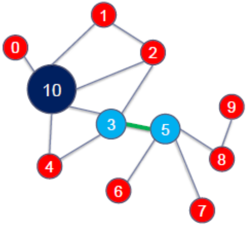
source: D. Petrov, Y. Dodonova, A. Shestakov (2015)

Caveats about Centrailities

- No single “right” centrality measure, each gives a different perspective
- Each centrality measure is a proxy of an underlying network process
- Unrealistic or irrelevant process lead to unrealistic centrality
- Centrality is used as graph EDA to gain insights about structure
- “Enhanced metrics” exist for graphs with more “features” (e.g. directed, weighted edges)
- Notion of centralities can be extended to edges

Identifying sets of key players

Importance of individual nodes may not reveal much

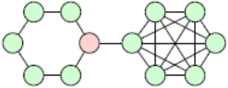


node 10 is most central, but node 3 and 5 together are more critical for network connectivity (the edge (3, 5) is a bridge)

Large-Scale Network Structure

Large-scale network structure

- Vertex level structural measures are local-do not reveal network shape
- Network level structural measures (somewhat) reveal the shape of the network e.g. average degree, clustering coefficient, radius, diameter
- They are unstable statistics (sensitive to non-uniformity)

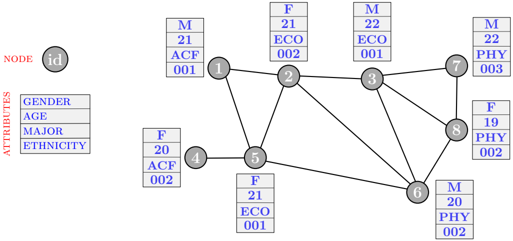


Nearly all value of clustering coefficient comes from the clique

- Vertex level measures (distributions thereof) provide some insights into structural heterogeneity, e.g. degree/centralities distribution
- Don't reveal organization patterns, do high degree nodes form cliques?
- Graph measures over aggregate, vertex measures over disaggregate

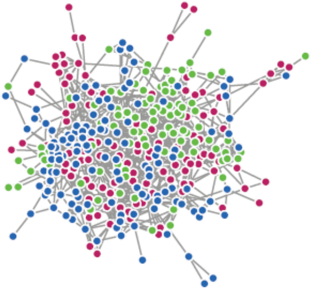
Heterogeneity in Networks

- Nodes in a social networks can have attributes like gender, age, political affiliation, interests, ...
- Node in power network could be the specification of components
- Biological Network nodes can have structural and functional properties of proteins
- What is network's organizational pattern of structural heterogeneity?
- Quantify the tendency of vertices with similar characteristics to be found close to each other in network (or the lack of this property)

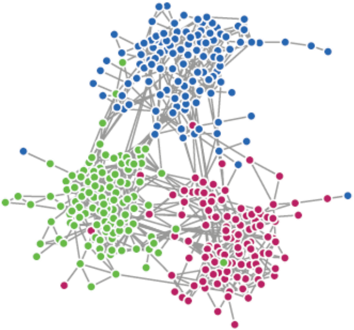


Mixing

- Mixing: a key concept in understanding large-scale structure
- Reveal whether vertices mix differently with some types of vertices than with others
- Homogeneous and heterogeneous mixing



homogeneous

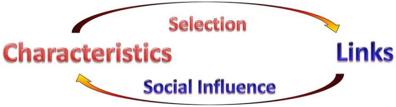


heterogeneous

Social Influence and Social Selection

Two important phenomena in Sociology

- Social Selection: Individual's attributes drive the interaction with others
- Social Influence: Interactions among people shape people's attributes



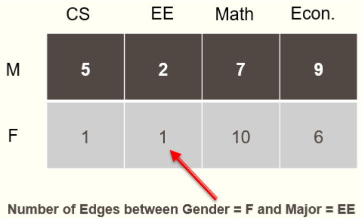
- Nodes characteristics and network structure are highly interlinked

Homogeneous and Heterogeneous Mixing

Mixing Matrix: Summary of Interconnection two attributes **a** and **b** is

- A row/column corresponding to each possible value of attribute **a**/**b**
- $(i, j)^{th}$ entry of $M_{(\mathbf{a},\mathbf{b})}$ is the number of edges connecting nodes with attribute value a_i of **a** to nodes with attribute value b_j of **b**

$$M_{(\mathbf{a},\mathbf{b})}(i, j) = |\{(u, v) \in E : \mathbf{a}(u) = a_i \text{ AND } \mathbf{b}(v) = b_j\}|$$

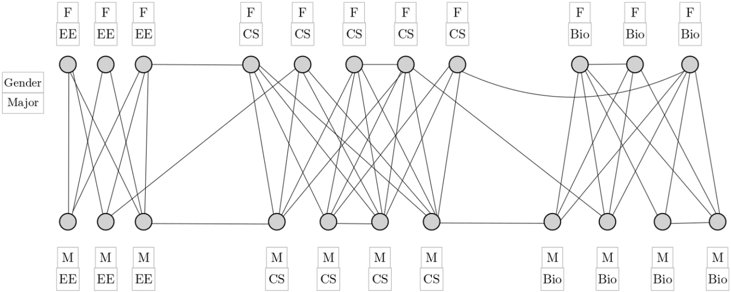


	CS	EE	Math	Econ.
M	5	2	7	9
F	1	1	10	6

Number of Edges between Gender = F and Major = EE

Homophily and Hetrophily

- Homophily: Connections among nodes having same attribute values ▷ assortative mixing
- Heterophily: Connections among nodes having different attribute ▷ disassortative mixing



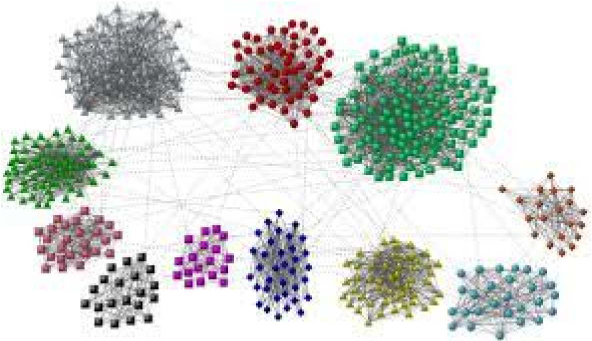
- 'MAJOR' attribute is homophilic
- 'GENDER' attribute heterophilic

Clusters in Graphs

Modular or community structure in graphs: homogeneous building blocks of the larger heterogeneous structure

Cluster: A dense subgraph within a graph

- Cohesion: Nodes are more similar to other nodes in the same cluster
- Separation: Nodes in a cluster are dissimilar to nodes in other clusters



Clusters in Graphs

Modular or community structure in graphs: homogeneous building blocks of the larger heterogeneous structure

Cluster: A dense subgraph within a graph

- Cohesion: Nodes are more similar to other nodes in the same cluster
- Separation: Nodes in a cluster are dissimilar to nodes in other clusters

Analytical questions:

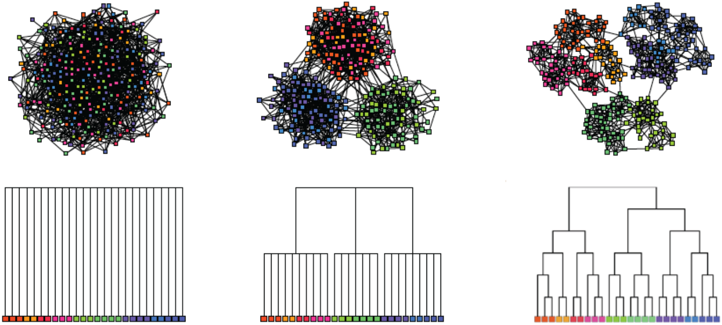
- Discover communities
- Describe interaction with a community
- Describe interaction between communities
- How a community emerged/dissolved
- Which communities are stable
- Predict whether a node will migrate to another community

Communities in Graphs

Community structure alone is a single level of organization reveal little information about structure within and between communities

Hierarchical community structure

e.g. CS department inside SSE inside LUMS inside Lahore ...



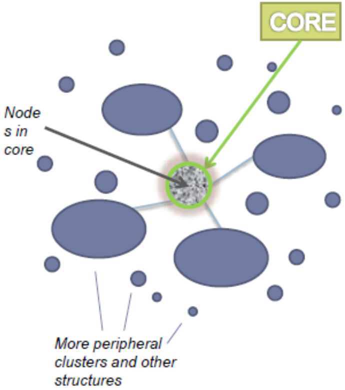
homogeneous

modular

hierarchical

Core-periphery Structure in Graphs

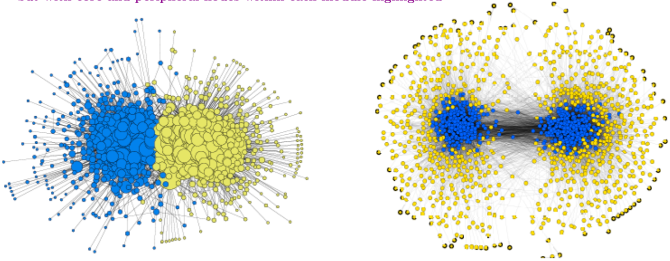
- A common large-scale structure in graphs is a core-periphery pattern
- One or more densely connected groups of nodes called cores
- Surrounded by a more diffuse, weakly connected and large periphery
- There could be small cluster-like groups in periphery too



Core-periphery Structure in Graphs

- The network could still be modular
- Where each module have a local core-periphery structure
- Recall the bow-tie structure of web graph is actually this pattern
- Cores can be identified visually ▷ Small networks
- It can be examined by whether highly centrality nodes are connected to other central nodes, i.e. centralities used as proxy for core nodes

A modular division of the classic political blogs network on the left, and the same network on the right but with core and peripheral nodes within each module highlighted



modular

core-periphery within modules

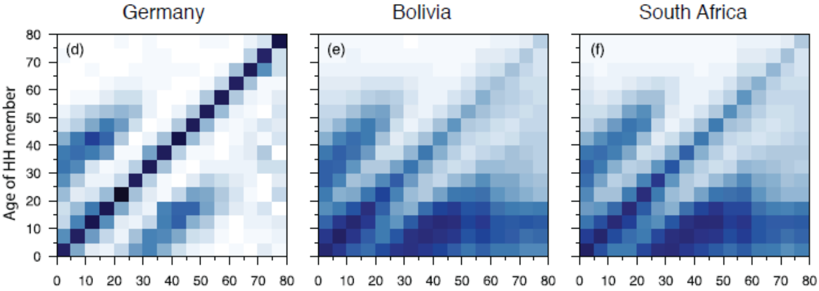
Linear Hierarchy Patterns

When nodes are ordered into a linear hierarchy (each node has a level)

Nodes at level k tend to connect only to those at levels $k \pm \Delta$

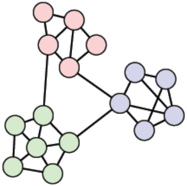
An ordered pattern can be see in such graphs

source: Prem, Cook, & Jit (2017) Profecting social contact matrices in 152 countries using surveys and demographic data

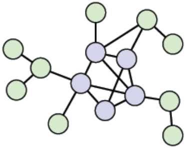


Social interactions among people are ordered with respect to people's ages

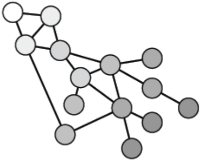
Large Scale Patterns in Graphs



modular*



core-periphery



ordered

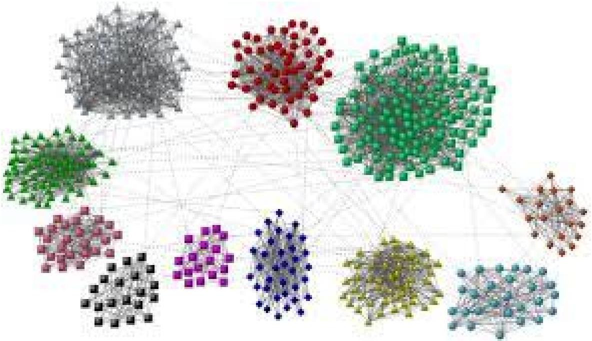
Modularity and Community Structure in Graphs

Communities in Graphs

Modular or community structure in graphs: homogeneous building blocks of the larger heterogeneous structure

Cluster: A dense subgraph within a graph

- Cohesion: Nodes are more similar to other nodes in the same cluster
- Separation: Nodes in a cluster are dissimilar to nodes in other clusters



Communities in Graphs

Modular or community structure in graphs: homogeneous building blocks of the larger heterogeneous structure

Cluster: A dense subgraph within a graph

- Cohesion: Nodes are more similar to other nodes in the same cluster
- Separation: Nodes in a cluster are dissimilar to nodes in other clusters

Analytical questions:

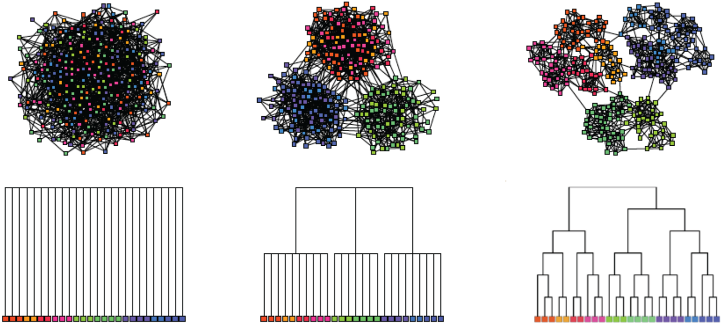
- Discover communities
- Describe interaction with a community
- Describe interaction between communities
- How a community emerged/dissolved
- Which communities are stable
- Predict whether a node will migrate to another community

Communities in Graphs

Community structure alone is a single level of organization reveal little information about structure within and between communities

Hierarchical community structure

e.g. CS department inside SSE inside LUMS inside Lahore ...



homogeneous

modular

hierarchical

Finding Communities

Recall two classes of clustering algorithms

- Hierarchical
  - Agglomerative or divisive
  - Requires a similarity measure between (sets of) objects ▷ for merging
- Partition-Based
  - Point assignment or density based algorithm
  - Require vector representation of objects
  - And a proximity measure on the vector space

For clustering nodes of a graph we derive a proximity measure from the provided association between nodes ▷ **adjacency**

Similarity must be inferred from the adjacency relaations between nodes, i.e proximity measure must relfect this structural property of network

Note: In spectral clustering we derive similarity graphs from the provided distance measure

k-Partition Problem

- Given a set of n points, $\mathcal{P} \subset \mathbb{R}^m$ and $k \in \mathbb{Z}$, number of clusters
- Assume Euclidean distance measure over $\mathcal{P}$ $\triangleright$ ℓ_p, cosine can be used
- For a subset $C_i \subseteq \mathcal{P}$, denote by $\mathbf{c}_i$ the **centroid** of C_i

$$\mathbf{c}_i := \frac{1}{|C_i|} \sum_{x \in C_i} x$$

- Centroid is the arithmetic mean of m-dim vectors (coordinate-wise mean)
- Goodness of a k-partition $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ is measured by

$$\text{sum of squared error,} \quad SSE(\mathcal{C}) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mathbf{c}_i\|^2$$

also called Within SSE

- **Problem:** Find a k-partition $\mathcal{C}^*$ of $\mathcal{P}$ with minimum SSE
- Brute force approach (try all $\binom{n}{k}$ partitions) is not feasible

Finding Communities: Similarity between nodes

Suppose nodes of $G = (V, E)$ are labeled $v_1, \dots, v_n$. v_i is referred to as i

Let A be the adjacency matrix, i.e. $A_ij = A(i,j) = 1 \leftrightarrow (v_i, v_j) \in E$

- Jaccard Index ▷ treating A's rows as sets

$$s_J(i,j) \;=\; \frac{N(i) \cap N(j)}{N(i) \cup N(j)} \;=\; \frac{\sum_k A_{ik}A_{kj}}{\sum_k (A_{ik} + A_{kj})}$$

- Consine Similarity ▷ treating A's rows as vectors

$$s_{cos}(i,j) \;=\; \frac{\sum_k A_{ik}A_{kj}}{\sqrt{\sum_k A_{ik}^2}\sqrt{\sum_k A_{jk}^2}} \;=\; \frac{d(i,j)}{\sqrt{d_i d_j}}$$

  - $d(i,j) \;=\; |N(i) \cap N(j)|$ called co-degree of v_i and v_j

Finding Communities: Similarity between nodes

Suppose nodes of $G = (V, E)$ are labeled $v_1, \ldots, v_n$. v_i is referred to as i

Let A be the adjacency matrix, i.e. $A_i j = A(i,j) = 1 \leftrightarrow (v_i, v_j) \in E$

- Euclidean distance ▷ treating A's rows as bit-strings

$$d(i,j) \;=\; \sum_k (A_{ik} - A_{jk})^2$$

- Normalized Euclidean distance

$$d(i,j) \;=\; \frac{\sum_k (A_{ik} - A_{jk})^2}{d(i) + d(j)} \;=\; 1 - 2\frac{d(i,j)}{d(i) + d(j)}$$

  - Adjust denominator appropriately when $d(i)$'s, or $d(i,j)$'s can be 0

Finding Communities: Similarity between sets of nodes

Suppose nodes of $G = (V, E)$ are labeled $v_1, \ldots, v_n$. v_i is referred to as i

Let A be the adjacency matrix, i.e. $A_ij = A(i, j) = 1 \leftrightarrow (v_i, v_j) \in E$

Let $X, Y \subseteq V$ and $s(i, j)$ be a similarity measure between nodes

- Single link ▷ tend to make small clusters

$$S_{XY} = \min_{u \in X, v \in Y} s(u, v)$$

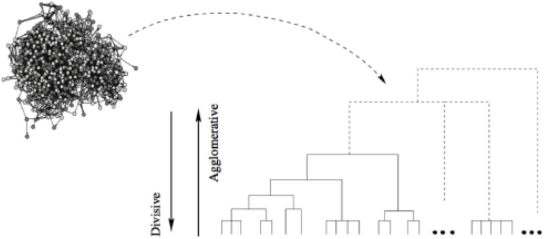
- Complete link ▷ tend to make large clusters

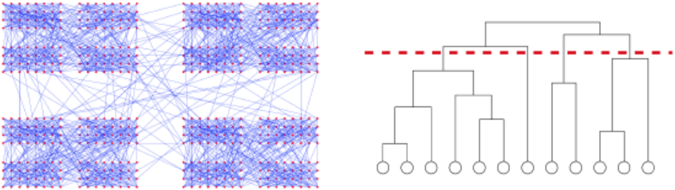
$$S_{XY} = \max_{u \in X, v \in Y} s(u, v)$$

- Average link ▷ regular

$$S_{XY} = \frac{\sum_{u \in X, v \in Y} s(u, v)}{|X| \times |Y|}$$

Agglomerative Graph Clustering





Agglomerative Graph Clustering

- Given a graph $G = (V, E)$
- A similarity measure between nodes
- A similarity measure sets of nodes

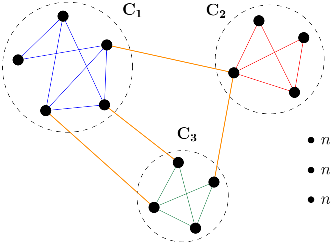
Algorithm : Generic Agglomerative Clustering ($\mathcal{G}$)

```
Initialize with each node as a cluster in $\mathcal{C}$
while stopping criterion is not met do
    Choose the best pair of clusters $(C_i, C_j) \in \binom{\mathcal{C}}{2}$
    $C_m \leftarrow \textsc{merge}(C_i, C_i)$
    $\mathcal{C} \leftarrow \mathcal{C} \setminus \{C_i, C_j\} \cup \{C_m\}$
    Recompute similarity between the merged cluster and others
```

Quantifying Cluster Quality

Broadly, we want dense (more intra-cluster edges) and well-separated (few inter-cluster edges) clusters

- Let $G = (V, E)$, $|V| = n$, and C a subset of nodes, $|C| = n_c$
- $G[C]$: the subgraph induced by C
- $G[C, \overline{C}]$: bipartite graph between C and $\overline{C}$
- $m_c = |E(G[C])|$ number of edges inside C
- $f_c = |\{(u, v) \in E : u \in C, v \notin C\}|$ number of frontier edges



- $n_{c_1} = 5,\ m_{c_1} = 7,\ f_{c_1} = 3$
- $n_{c_2} = 4,\ m_{c_2} = 5,\ f_{c_2} = 2$
- $n_{c_3} = 4,\ m_{c_3} = 5,\ f_{c_3} = 3$

Quantifying Cluster Quality

Goodness of a cluster

- Intra-cluster density of C aka internal density

$$\delta_{int}(C) \;=\; {}^{m_c}\!/\!\binom{n_c}{2}$$

- Inter-cluster density of C aka cut ratio

$$\delta_{ext}(C) \;=\; {}^{f_c}\!/_{n_c(n-n_c)}$$

- Conductance of C, fraction of total edge volume pointing outside C

$$conductance(C) \;=\; {}^{f_c}\!/_{2m_c+f_c}$$

- Modularity of C, difference b/w m_c and expected internal edges in random graph with same degree distribution

$$Q(C) \;=\; {}^{1}\!/_{4m}(m_c - E[m_c])$$

Quantifying Clustering Quality

All above were goodness of a community, for the whole network, compute their weighted average

$$metric(G) \; = \sum_{C \in communities(G)} \frac{n_C}{n} \times metric(C)$$

Community Detection Algorithms

[Fortunato, 2010]

- Girvan-Newman algorithm
- Modularity optimization algorithms
  - Greedy
    - Hierarchical: join clusters leading to largest increase in modularity [Newman, 2004]
    - Clauset algorithm: fast version using nice data structures that exploit sparsity [Clauset et al., 2004]
    - Louvain algorithm [Blondel et al., 2008]
  - Spectral algorithms [Newman, 2006]
  - and many others
- Graph partitioning algorithms
- Clique percolation method

Graph Representation

Graph Representation

- Node2Vec
- Graph2Vec
- Convolution NN
- Embedding
- Graph Clustering and Classification