Social Networks: Introduction

Saptarshi Ghosh
Department of CSE, IIT Kharagpur
Social Computing course, CS60017

Social networks in off-line world

- Social networks studied for several decades
  - Friendship networks among students of a school, members of a club, …
  - Collaboration networks among scientists, movie actors, …
  - Citation networks: scientists / papers referring to other scientists / papers

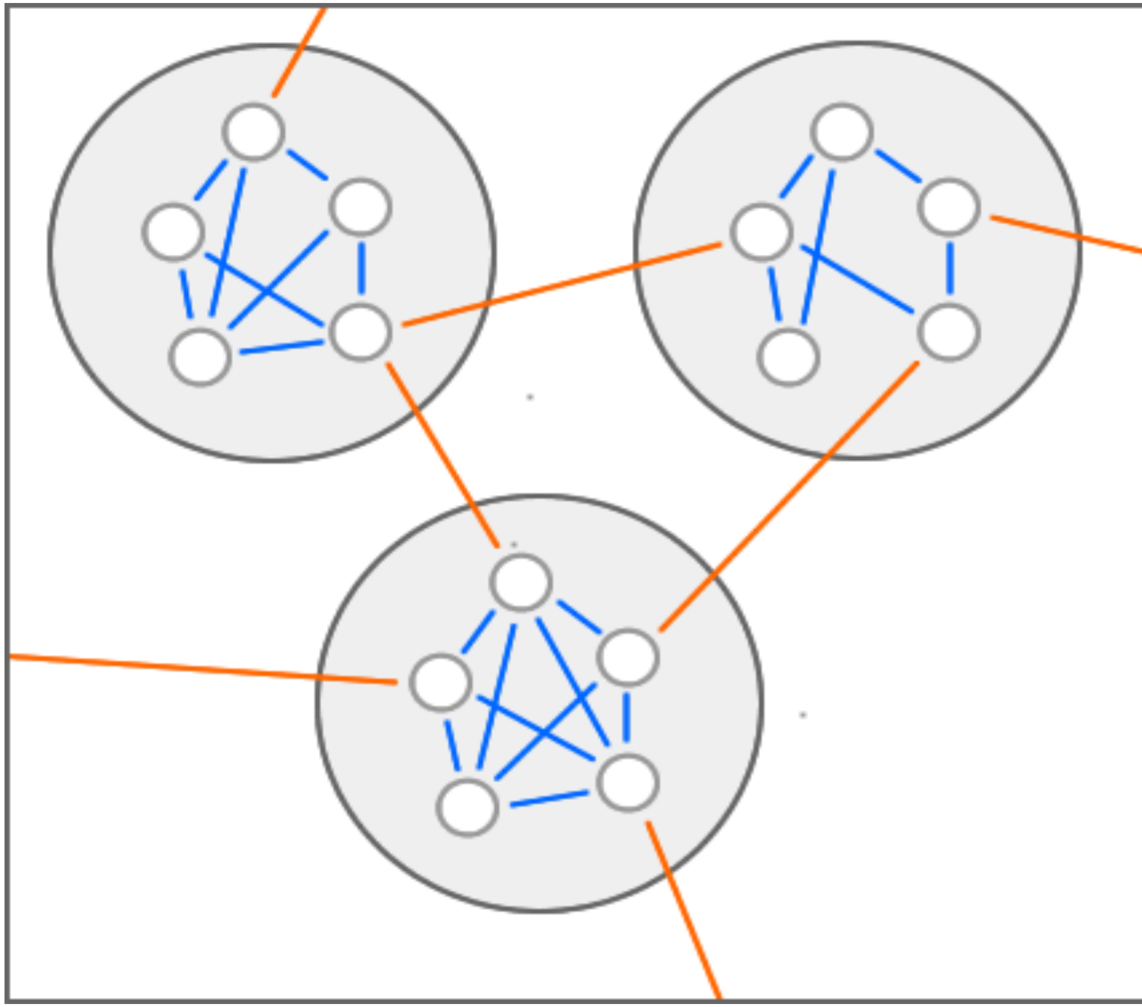
Sociological theories

- Several sociological theories developed
  - Homophily – birds of a feather flock together
  - Six degrees of separation - Milgram's experiments (1967)
  - Strength of weak ties (1973)
  - Spread of epidemics / conventions / news / rumors

Milgram's experiment in 1967

- Sent packets to people in Omaha, Nebraska and Wichita, Kansas
  - You need to get the packets to a specific person in Boston
  - If you know the recipient, send the packet directly to him
  - If not, think of a friend you know, who is likely to be closer to the recipient in Boston; sign your name to a roster, and send the packet to your friend
- Boston recipient examined the roster and saw how many steps it took for the letter to arrive
- 64 letters reached recipient, average number of links: between 5 and 6

Strength of ties



Group/Network

Group members, because of their frequent interaction, tend to think alike over time. This reduces the diversity of ideas, and in worst-case scenarios leads to "groupthink"

Weak Ties

Weak ties are relationships between members of different groups. They are utilized infrequently and therefore don't need a lot of management to stay healthy. They lead to a diversity of ideas, as they tie together disparate modes of thought.

Strong Ties

Strong ties are relationships between people who work, live, or play together. They are utilized frequently and need a lot of management to stay healthy. Over time, people with strong ties tend to think alike, as they share their ideas all the time.

Advent of online social networks

Advent of online social networks

- Among the most popular sites on today's Web

- Billions of users world-wide
  - Celebrities, media houses, politicians, commoners, …
  - Spammers, cyber-bullies, hatemongers, …

- Huge impact
  - Advertisers reach large population at minimal cost

OSN and researchers

- Huge data readily available
  - Volume – networks of billions of users, petabytes of user-generated content every day
  - Variety – text, image, speech, video, …
  - Velocity – thousands of posts / minute during major events
- Automated data collection rather than surveys

Multi-disciplinary research on OSNs

- Computer networks & distributed systems
- Sociology, social psychology, linguistics, …
- Network science, complex network theory
- Data mining, machine learning, information retrieval, natural language processing, …

Structural properties of large (social) networks

Large networks - examples

- Social networks
  - Friendship networks, collaboration networks among scientists / movie actors, communication networks (email or phone call), online social networks
- Information networks
  - Citations among research papers, the Web
- Technological networks
  - The Internet, electric power grid, transportation networks
- Biological networks
  - Genetic regulatory network, food web, neural networks

How to model social networks?

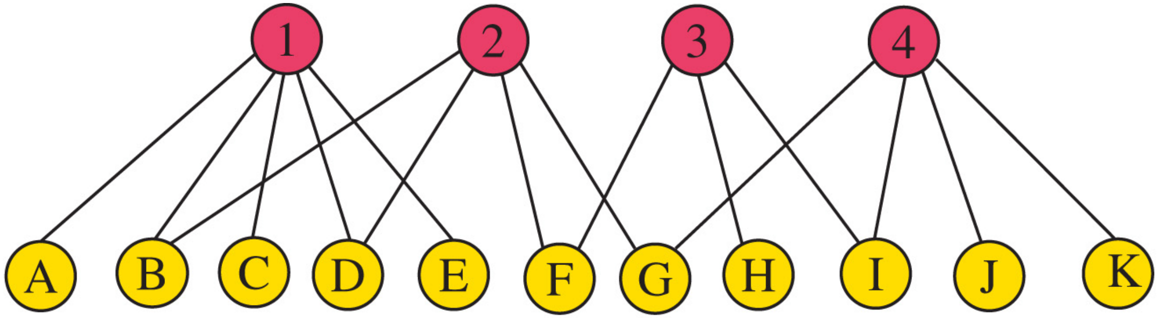
- Most common representation: a graph
  - Nodes: users,   edges: social links

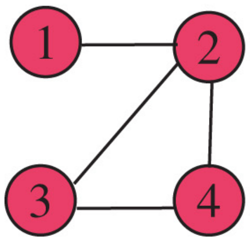
- Undirected networks: Facebook
- Directed networks: Twitter
- Weighted networks
  - Edge-weights usually measure "strength" of social link, e.g., number of interactions

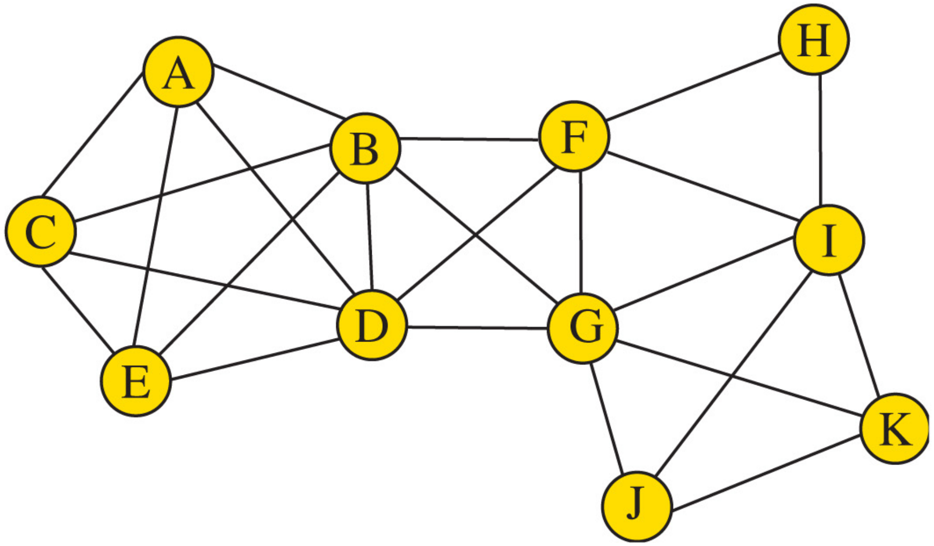
Graph models of OSNs

- Other varieties of networks
  - Networks among blogs, videos, …
  - Bipartite networks, e.g., viewer-video model of Youtube
  - Folksonomy: **Users** annotate **resources** with **tags**, modeled as tri-partite hypergraphs **[Cattuto, AI Communications 2007]**

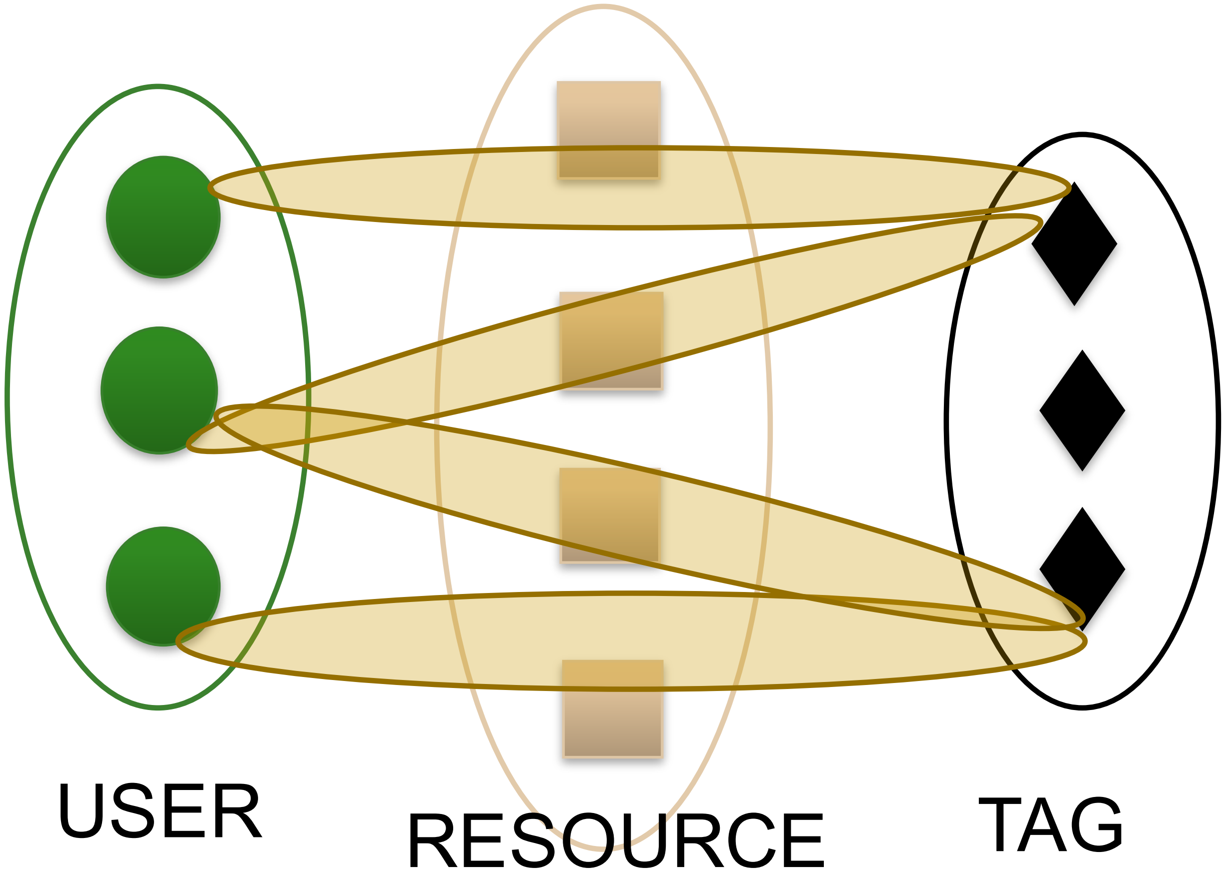
Bipartite networks and projections







Tri-partite model for folksonomies



How to study structure of large networks?

- Too large to visualize even by tools
  - Some popular network visualization tools: Gephi, Pajek, …

- Individual nodes do not have much significance w.r.t. structure or function

- Use statistical measures to describe structure

Topological properties of networks

Degree distribution

- p_k: fraction of nodes having degree k, $k = 0, 1, 2, \ldots$
  - Equivalently, probability that a randomly chosen node has degree k

- Cumulative degree distribution
  - Fraction of nodes having degree at least k

- Many real networks show
  - Power-law degree distribution: $p_k \sim k^{-a}$
  - Exponential degree distribution: $p_k \sim e^{-k/\gamma}$

Shortest distances between nodes

- L: mean shortest distance between any pair of nodes
- Diameter
  - Maximum shortest distance between any pair of nodes
- Effective diameter
  - A value such that 90% of the shortest distance between any pair of nodes is lower than this value

Shortest distances between nodes

- Many real large networks have very small L compared to the number of nodes
  - Typically L varies as *log(n)*, where *n* is #nodes

- Six degrees of separation – Milgram's experiment

- Even lower for online social networks like Facebook

Clustering / transitivity

- If node A is connected to B and B to C, is there a higher probability of A being connected to C?

- Measured by clustering coefficient [0, 1]

- CC for a node *n*
  - Among the pairs of neighbors of *n*, what fraction is connected between themselves?

Clustering / transitivity

- Clustering coefficient for a network:

$$C = \frac{3 \times \text{ number of triangles in the network}}{\text{number of connected triples of vertices}}$$

connected triple: a node with edges to an unordered pair of nodes

- Alternative definition of CC for a network: mean CC for all nodes
- What type of networks are likely to have high / low clustering coefficient?

Mixing patterns / assortativity

- A network usually has nodes of several different types
  - Do nodes of the same type connect to each other selectively?

- Example: mixing by race in San Francisco

		women			
		black	hispanic	white	other
men	black	506	32	69	26
	hispanic	23	308	114	38
	white	26	46	599	68
	other	10	14	47	32

Mixing patterns / assortativity

- Assortativity coefficient r (in [-1,1])
  - $r > 0$: assortative network
  - $r < 0$: disassortative network

- How to measure assortativity coefficient?
  - e_{ij} : fraction of all edges in the network, that connects a node of type i with a node of type j
  - e : matrix whose (i,j)-th element is e_{ij}

Mixing patterns / assortativity

		women			
		black	hispanic	white	other
men	black	506	32	69	26
	hispanic	23	308	114	38
	white	26	46	599	68
	other	10	14	47	32

Matrix *e* =

		women				
		black	hispanic	white	other	a_i
men	black	0.258	0.016	0.035	0.013	0.323
	hispanic	0.012	0.157	0.058	0.019	0.247
	white	0.013	0.023	0.306	0.035	0.377
	other	0.005	0.007	0.024	0.016	0.053
	b_i	0.289	0.204	0.423	0.084	

Topological properties of networks

- Definition of assortativity coefficient *r* (in [-1,1])

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i} = \frac{\mathrm{Tr}\,\mathbf{e} - \|\,\mathbf{e}^2\,\|}{1 - \|\,\mathbf{e}^2\,\|},$$

  - where $||x||$ means the sum of all elements of matrix x

- Degree assortativity – most commonly studied
  - E.g., do high (low) degree nodes connect to other high (low) degree nodes?

Different types of networks

Random networks

- Random network: Erdos-Renyi network
  - Take n nodes and connect each pair with probability p

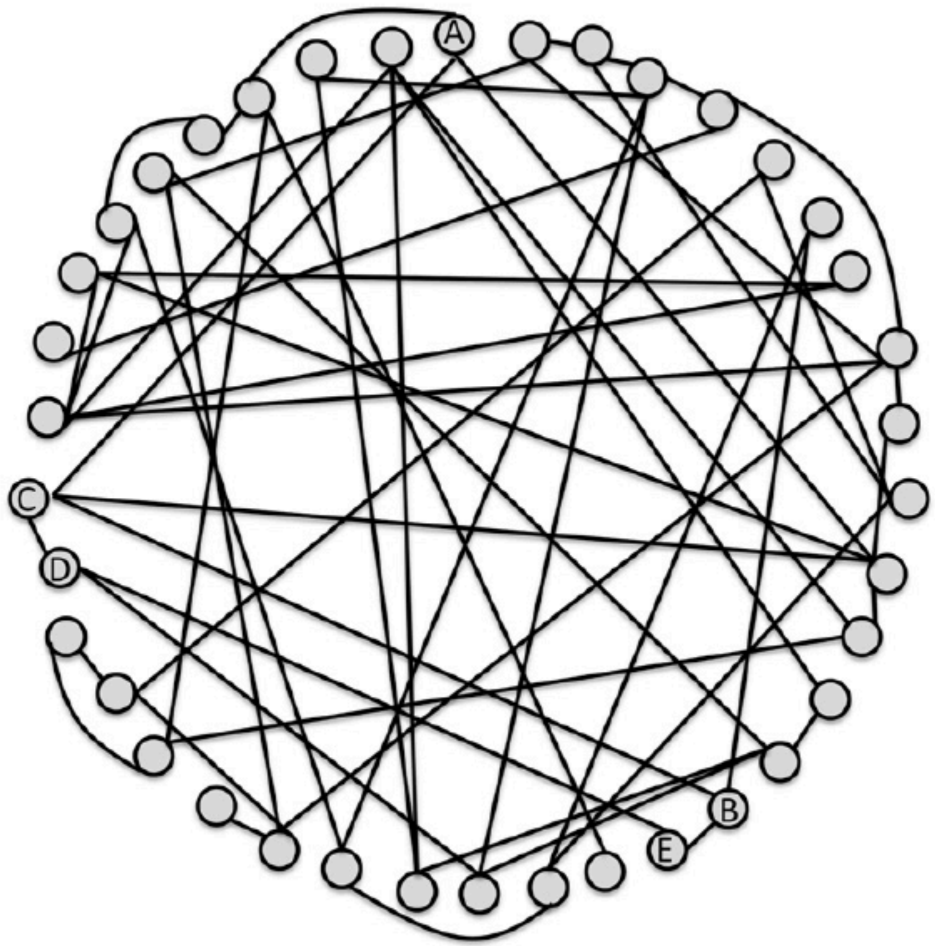
- Properties
  - Degree distribution: Poisson distribution
  - Clustering close to zero
  - Assortativity close to zero (no degree correlations)
  - Distance between any two nodes is usually low

- Real networks differ widely from random networks

Random networks

- Edges randomly connect the nodes
- A random graph with 36 nodes and 72 edges

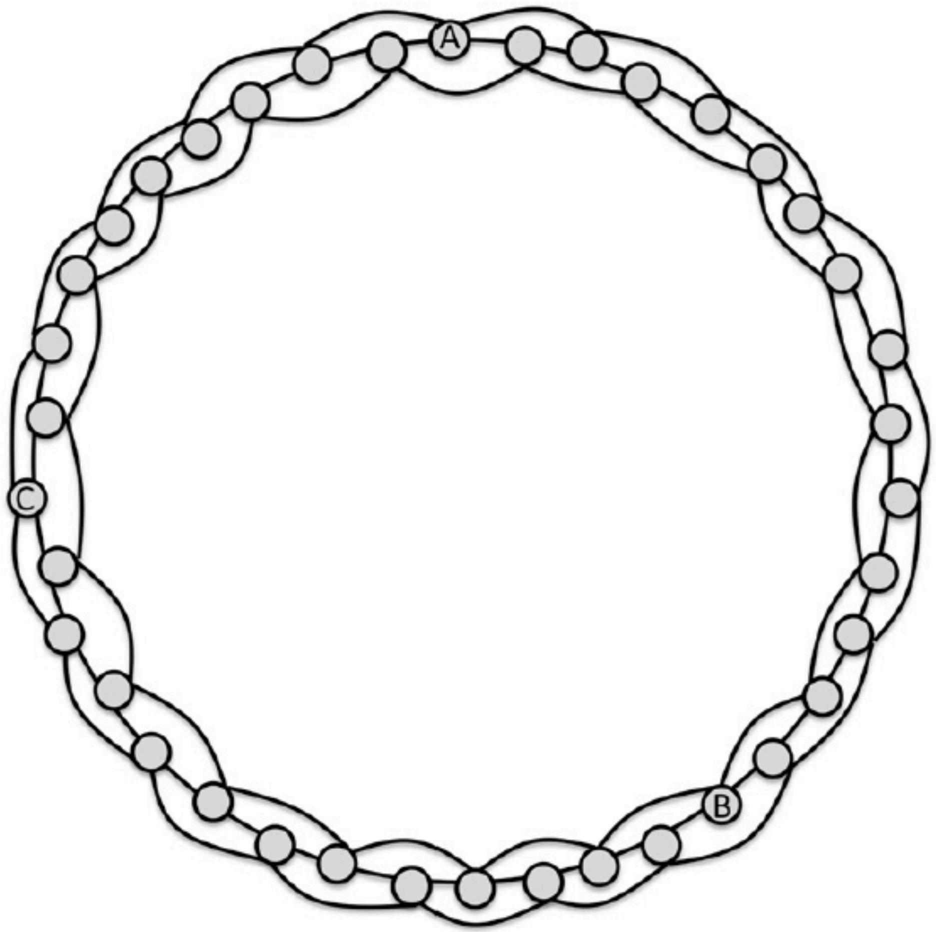
- What is the distance between A and B?
- What is the clustering coefficient of A, B?



Regular network

- Each node has a fixed number of neighbors
- A regular graph with 36 nodes and 72 edges

- What is the distance between A and B?
- What is the clustering coefficient of A, B?



Small world networks

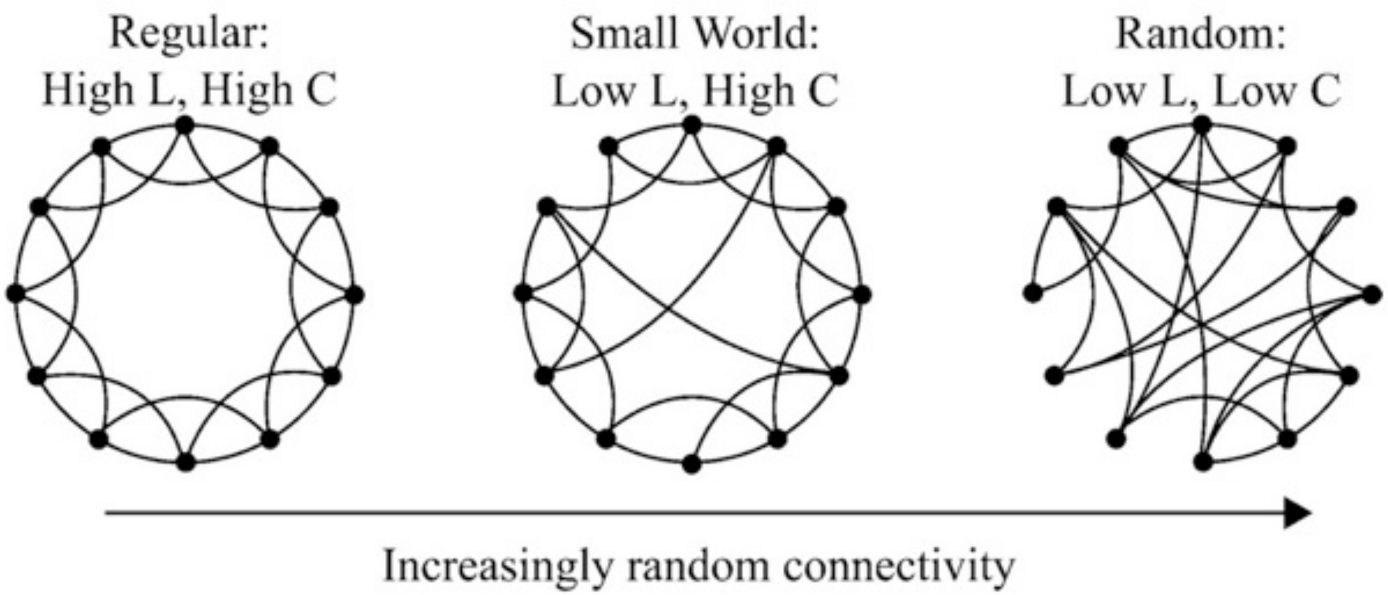
- Defined by Watts and Strogatz

- Informally
  - Most nodes are not neighbors of one another, but most nodes can be reached from every other node by a small number of hops or steps

- More formally
  - L ~ log *n* (average shortest path length is low)
  - High clustering coefficient

Small world networks

- Combination of regular graph and random graph
- Take a regular graph and randomly re-wire a few edges
- No significant impact on clustering (remains high)
- Shortest distances drop drastically



Social networks – Case study 1

Measurement and Analysis of Online Social Networks, Mislove et al., IMC 2007

One of the earliest measurement studies of OSNs

- Crawled data of four OSNs: Flickr, Orkut, Youtube, LiveJournal
- Used BFS crawls to crawl user profiles, links, …

- Observed properties for the social networks
  - Link symmetry – most links are reciprocated
  - Power law degree distributions (Orkut deviates)
  - In-degree highly correlated with out-degree
  - Average shortest path lengths between 4 and 6

Properties of social networks

- Assortativity coefficient
  - Flickr: 0.202, LiveJournal: 0.179, Orkut: 0.072
  - Youtube: -0.033
  - Web: -0.067, Internet: -0.189
- Social networks have a densely connected core
  - A relatively small strongly connected group of nodes that is necessary to keep the remainder of the network connected (relatively small diameter)
- Clustering coefficient of nodes falls with out-degree

Social networks – Case study 2

The Anatomy of the Facebook social graph,
Ugander et al., 2011

Facebook social network

- Undirected network
  - Nodes: users / accounts
  - Edges: friendship links

- Ugander et al., The Anatomy of the Facebook Social Graph, 2011
  - 721 million nodes
  - 68.7 billion friendship links

Results

- Degree distribution
  - Most users have < 200 friends, some have thousands
  - Not power-law
- Average pairwise distances
  - Neighborhood function N(h) of a graph: number / fraction of pairs of nodes (u, v) such that distance between u and v is at most h
  - Average distance between pairs of users: 4.7
- 99.9% of nodes in a single connected component

Results

- Clustering coefficients of nodes are typically high
  - For an average user with 100 friends, c = 0.14
  - Average c for users with degree k decreases with k

- Though Facebook graph sparse as a whole, it contains dense neighborhoods

Results

- Assortativity
  - Degree assortativity r = 0.226
  - Assortativity w.r.t. age: a random neighbor is most likely to be the same age as you; probability of friendship with older individuals falls off rapidly
  - Assortativity w.r.t. country: 84.2% of links are within countries → indicates community / modular structure based on geography

Social networks – Case study 3

What is Twitter, a Social Network or a News Media?, Kwak et al., WWW 2010

One of the first large-scale measurement studies on Twitter

- Crawled: 41.7 M users, 1.47 B links, tweets, trends
- Properties observed:
  - In-degree distribution is a power-law, but not the out-degree distribution
  - Only 22% links are reciprocal
  - Average path length 4.12, effective diameter 4.8
  - Reciprocated links exhibit homophily to some extent
- Twitter has characteristics of both a social network and a news media

Research issues on (social) networks

Sociological issues

- Sociological theories investigated on OSNs
  - Homophily, strength of weak ties [Grabowicz, Plos ONE, 2012]
  - Emergence and spread of conventions [Kooti, ICWSM 2012]

- OSNs different from offline SNs in some aspects
  - Almost zero cost of maintaining social links
  - Important users readily connect to many ordinary ones
  - Geographical distance does not matter

Explaining the network properties

- What nature of link-creation dynamics explain the empirically observed properties of OSNs?

- Several evolution models proposed
  - Global rules, e.g., Preferential Attachment **[Barabasi, Science 1999]**
  - Local rules, e.g., triangle closure **[Kleinberg, ICWSM 2010]**, random walk starting from a node **[Vasquez, PRE 2003]**
  - Biased PA, based on different types of users: inactive, linkers, inviters **[Kumar, KDD06]**
  - Co-evolution of social and content networks **[Singer, Making Sense of Microposts, 2012]**

Dynamic network properties

- Dynamic nature: how do properties of OSNs change with time?
  - Network density varies non-monotonically **[Kumar, KDD06]**
  - Assortativity varies non-monotonically **[Hu, Physics Letters A, 2009]**

- Models to explain temporal variation of properties

Link analysis

- Classification of social links
  - Strong and weak links (e.g. based on level of interaction) **[Wilson, EuroSys09][Valafar, WOSN09] [Xiang, WWW10]**
  - Some OSNs allow positive and negative links (friends and enemies)
- Variation of strength of links with time **[Viswanath, WOSN09]**

Centrality (importance) of nodes

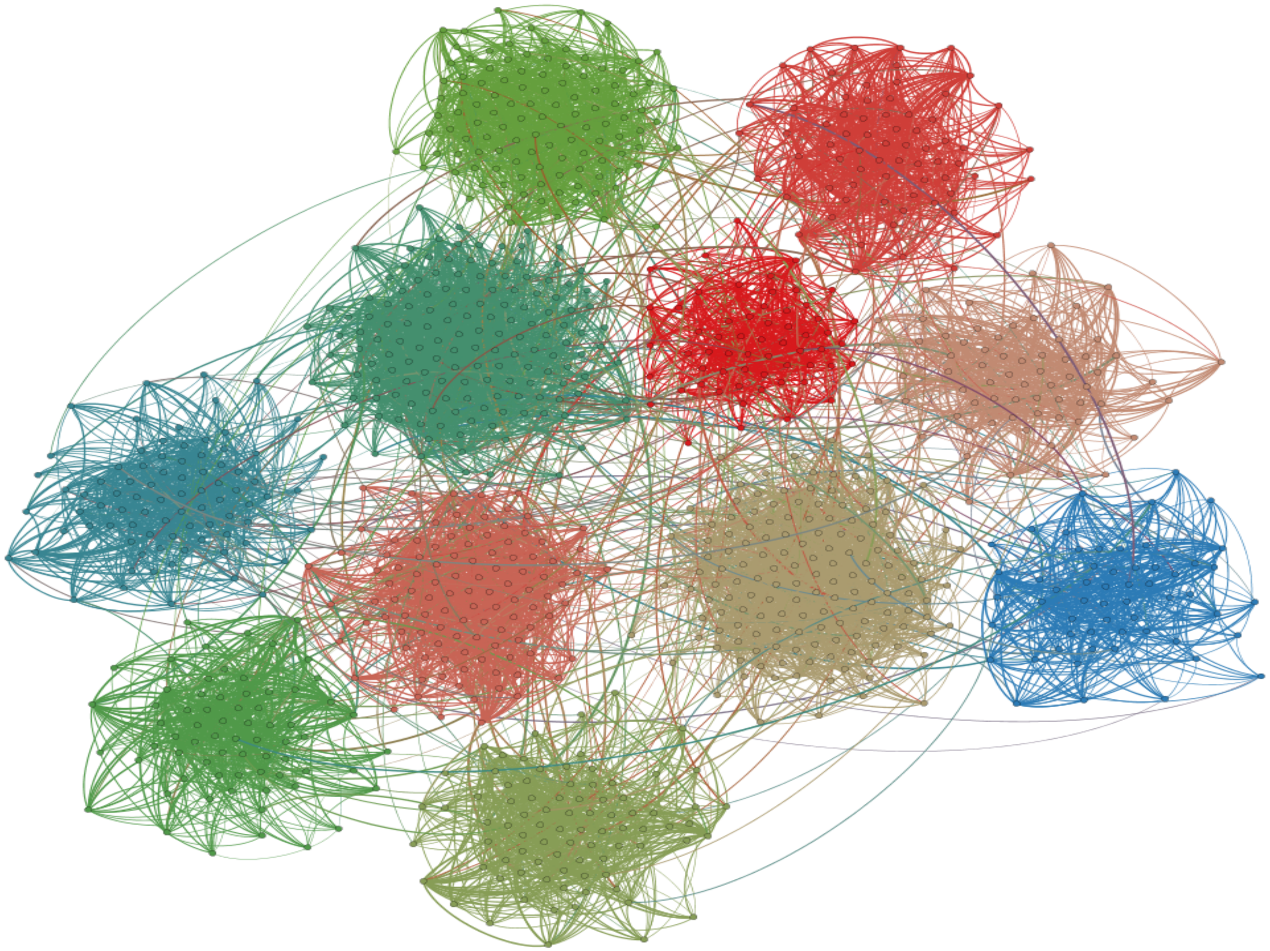
- How important is a node in a network?
  - How influential is a person in a social network?
  - How important is a website on the Web?

- Many proposed centrality metrics
  - Degree centrality
  - Closeness centrality
  - Betweenness centrality
  - Eigenvector centrality, PageRank

Community detection / clustering

- Identifying communities of ‘similar’ users
  - Traditionally, only rely on network structure: several algorithms **[Fortunato, Physics Reports 2010] [Leskovec, WWW10]**
  - Content can also be leveraged in case of OSNs

- Dynamic communities: how do communities change with time? **[Mitra, Computer Networks, 2012]**



Friendship network among students in a US school