Structural properties of large networks

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

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

Wide variety in structure

- Unipartite networks, e.g., social networks
  - Weighted or unweighted
  - Directed or undirected

- Bipartite networks, e.g., movie-actor networks
- Tripartite networks, e.g., user-tag-item networks

- Temporal or time-varying networks

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
men	hispanic	23	308	114	38
men	white	26	46	599	68
men	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

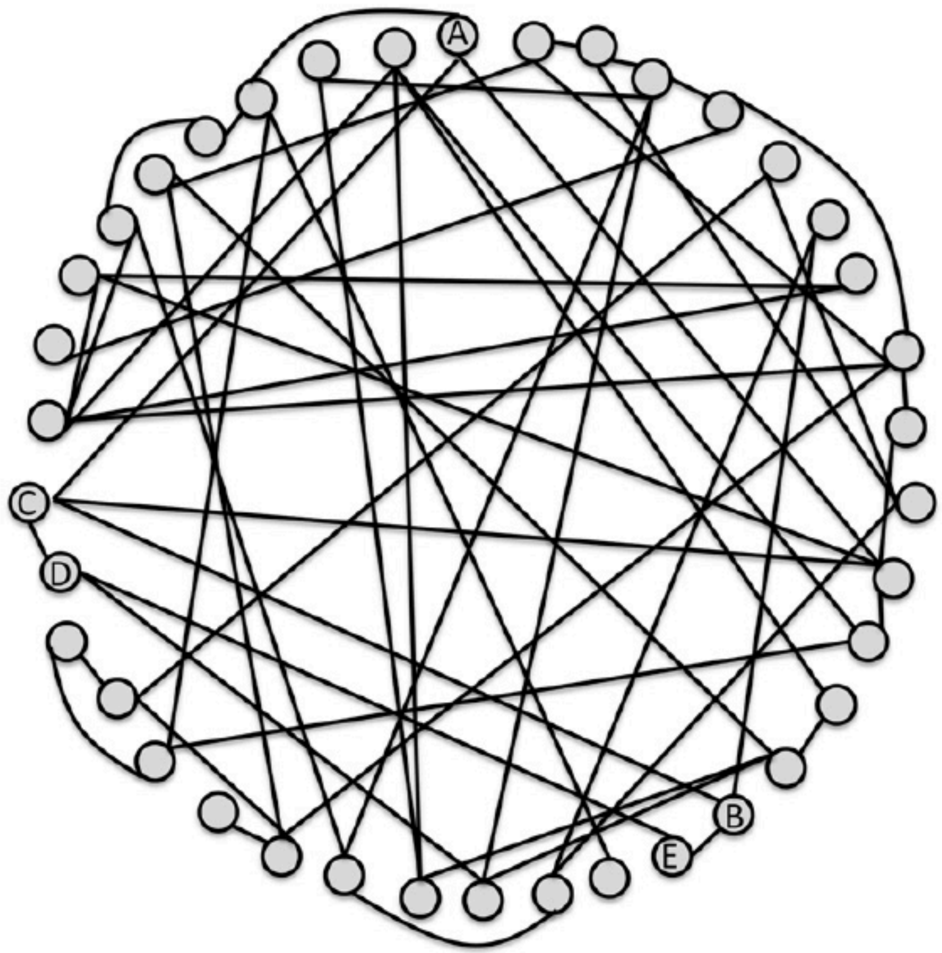
- Properties
  - Degree distribution: Poisson distribution
  - Clustering close to zero
  - Assortativity close to zero (no degree correlations)
  - Distance between any two nodes is usually low

- Real networks differ widely from random networks

Random networks

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

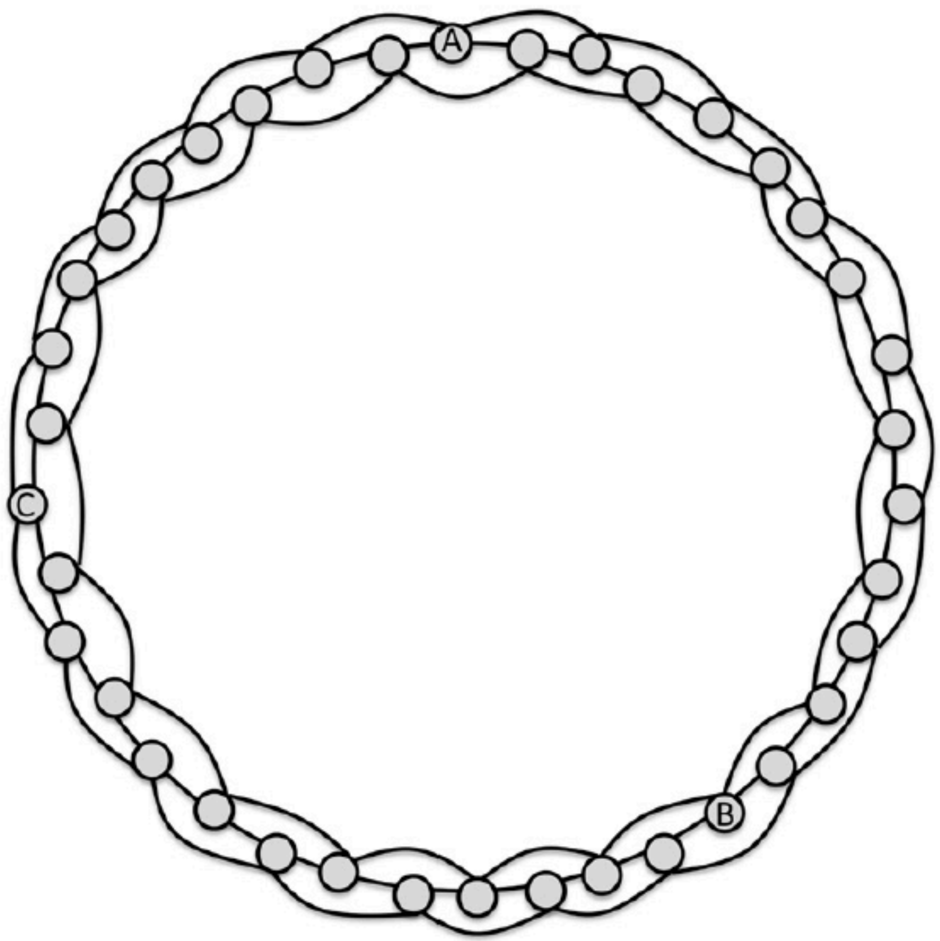
- What is the distance between A and B?
- What is the clustering coefficient of A, B?



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

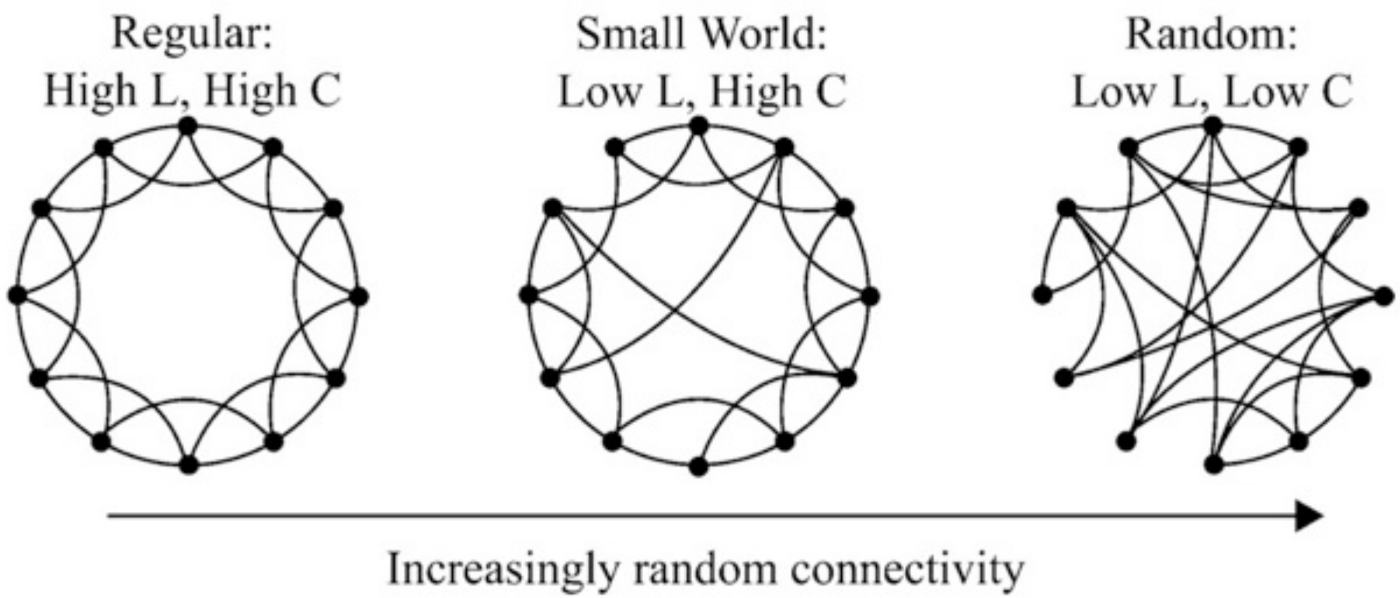
- Defined by Watts and Strogatz

- 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