Social and Technological Network Analysis

Lecture 1: Networks, Random Graphs and Small World Models

Dr. Cecilia Mascolo

UNIVERSITY OF CAMBRIDGE

About Me

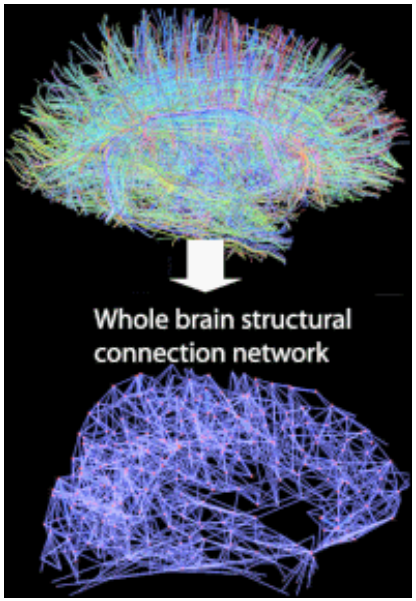
- Reader in Mobile Systems
  - NetOS Research Group
- Research on Mobile, Social and Sensor Systems
- More specifically,
  - Human Mobility and Social Network modelling
  - Opportunistic Mobile Networks
  - Mobile Sensor Systems and Networks

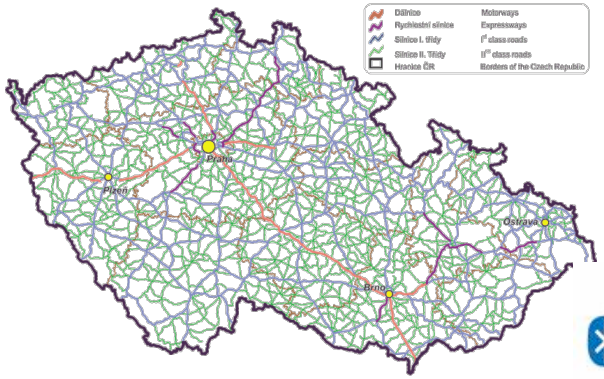
twitter

foursquare

UNIVERSITY OF CAMBRIDGE

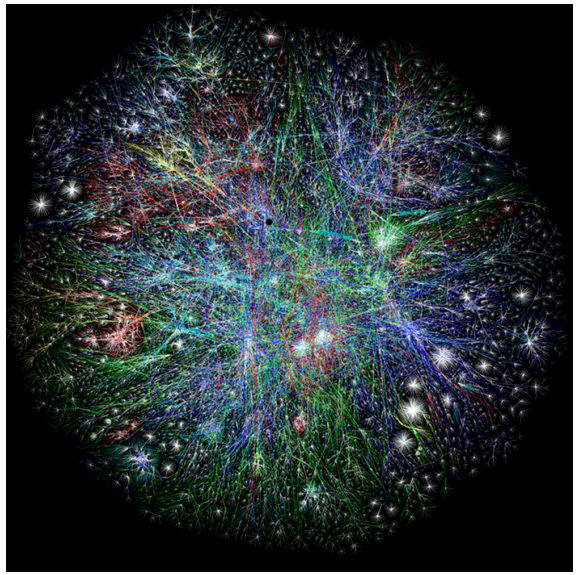
Networks are Everywhere





facebook

Bluetooth



UNIVERSITY OF CAMBRIDGE

Facebook Friendship Network

UNIVERSITY OF CAMBRIDGE

What Kind of Networks?

- Who talks to whom?
- Who is friend with whom?
- What leads to what?
- Who is a relative of whom?
- Who eats whom?
- Who sends messages to whom?

In This Course

- We will study the models and metrics which allow us to understand these phenomena.
- We will show analysis over large datasets of real social and technological networks.

List of Lectures

- Lecture 1: Networks and Small World Properties
- Lecture 2: Weak Ties and Community Detection
- Lecture 3: Structure of the Web and Power Laws
- Lecture 4: The Internet and Robustness
- Lecture 5: Search and PageRank
- Lecture 6: Information Cascades on Networks
- Lecture 7: Epidemic Dissemination on Networks
- Lecture 8: Practical Lecture on Handling Datasets

UNIVERSITY OF CAMBRIDGE

Assessment

- All information on the course page: http://www.cl.cam.ac.uk/~cm542/teaching/2010/stna2010.html

- **One report** (of approximately 1,500 words) on one assigned research paper. The report is due on **14th February** (Happy Valentine!) and it is worth 40% of the final mark.
- The second assignment will consist of **analysis of an assigned dataset according to some indicated network measures** using NetworkX: the analysis should be reported in a document of about 1,500 words where the results are commented and justified. This should be handed in by **18th March** and will be worth 60% of the final mark.

UNIVERSITY OF CAMBRIDGE

Structure of First Report

- **The first report should be approximately 1,500 words.** The report will contain two parts of about **750 words each**:
  - Critical analysis of the papers including, possibly, comparisons and references to other material presented in the course or found by the student and comments on how solid the result obtained are (e.g., comments on the evaluation methods or on the analysis applied can be included);
  - Discussion of possible future research ideas in the area.

Choices

- First assignment: send email immediately with a choice of paper **(and a backup)** from the website [assignment is first come first served].
  - **Do this in the next TWO days.**

- Second assignment: same. At that point you will receive an email with a link to the dataset.
  - **Do this by 11th February.**

UNIVERSITY OF CAMBRIDGE

In This Lecture

- We will introduce:
  - Networks/graphs
  - Basic network measures
  - Random Graphs
  - Small World Model
  - Examples

A Network is a Graph

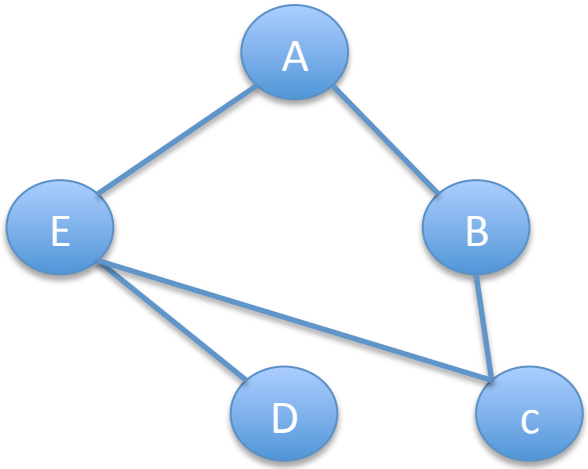
A **graph** G is a tuple (V,E) of a set of vertices V and edges E. An edge in E connects two vertices in V.

A **neighbour set** N(v) is the set of vertices adjacent to v:

$$N(v) = \{u \in V \mid u \neq v, (v,u) \in E\}$$

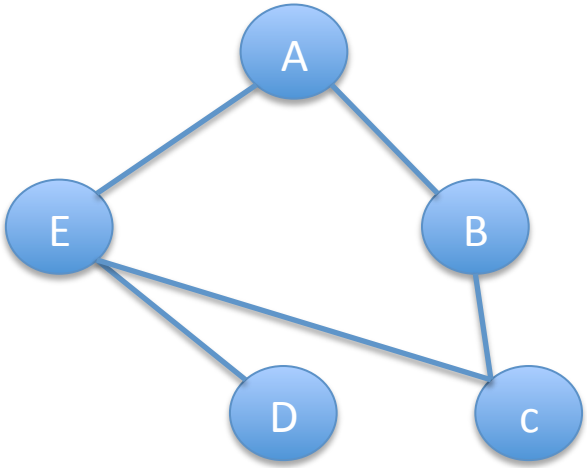
Node Degree

- The **node degree** is the number of neighbours of a node
- E.g., Degree of A is 2

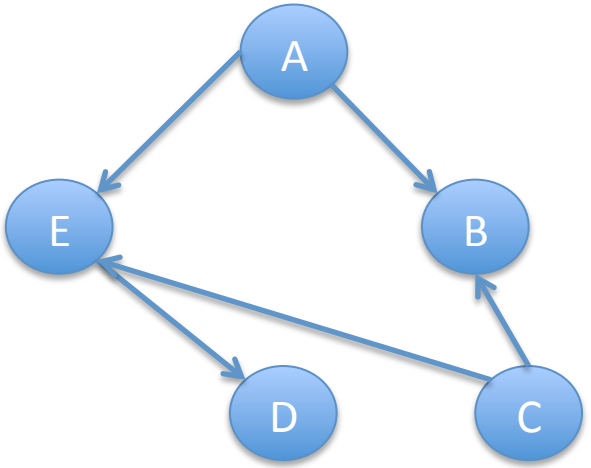


UNIVERSITY OF CAMBRIDGE

Directed & Undirected Graphs



Undirected Graph



Directed Graph

UNIVERSITY OF CAMBRIDGE

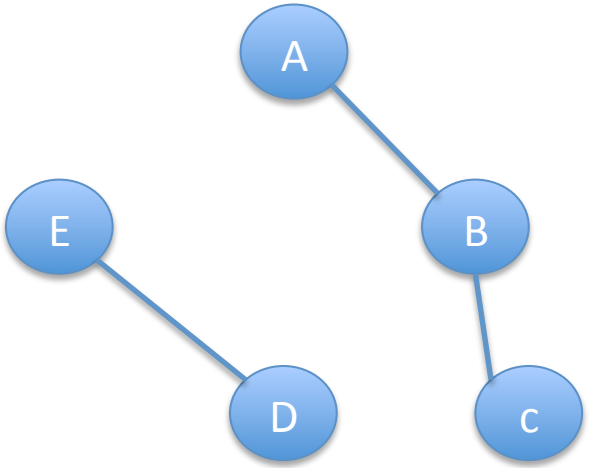
Paths and Cycles

- A **path** is a sequence of nodes in which each pair of consecutive nodes is connected by an edge.
  - If graph is directed the edge needs to be in the right direction.
  - E.g. A-E-D is a path in both previous graphs
- A **cycle** is a path where the start node is also the end node
  - E.g. E-A-B-C is a cycle in the undirected graph

Connectivity

- A graph is **connected** if there is a path between each pair of nodes

- Example of **disconnected** graph



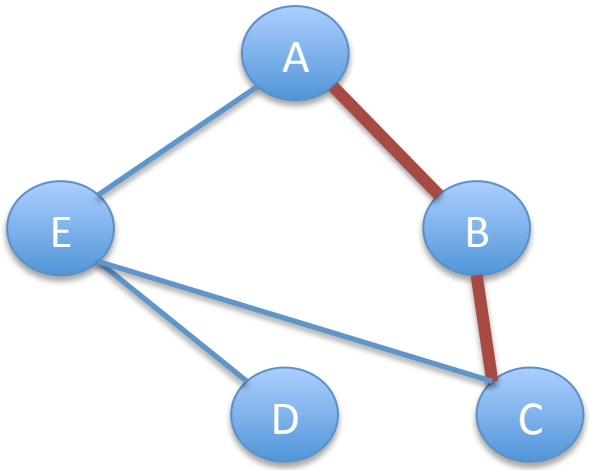
UNIVERSITY OF CAMBRIDGE

Components

- A **connected component** of a graph is the subset of nodes for which each of them has a path to all others (and the subset is not part of a larger subset with this property)
  - Connected components: A-B-C and E-D
- A **giant component** is a connected component containing a significant fraction of nodes in the network
  - Real networks often have one unique giant component

Path Length/Distance

- The **distance** (d) between two nodes in a graph is the length of the shortest path linking the two graphs.
- The **diameter** of the graph is the maximum distance between any pair of its nodes.

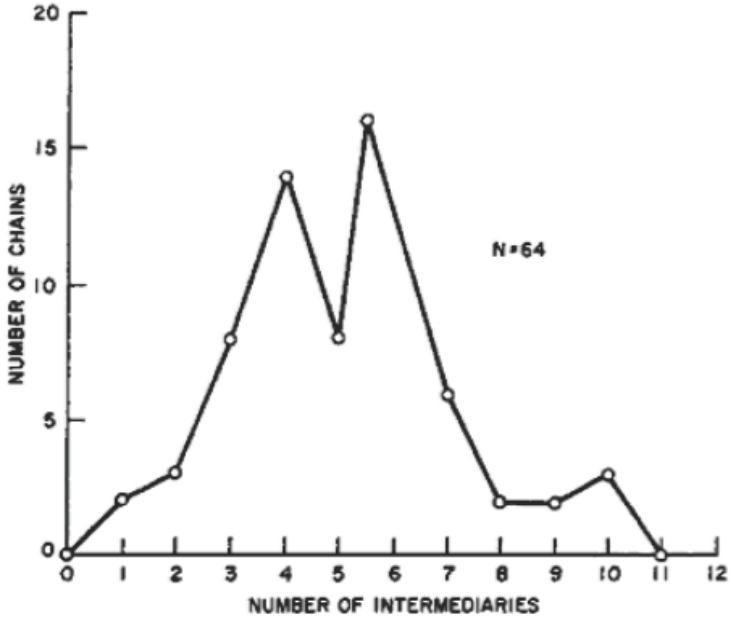


D(A,C)=2

UNIVERSITY OF CAMBRIDGE

Small-world Phenomenon
Milgram's Experiment

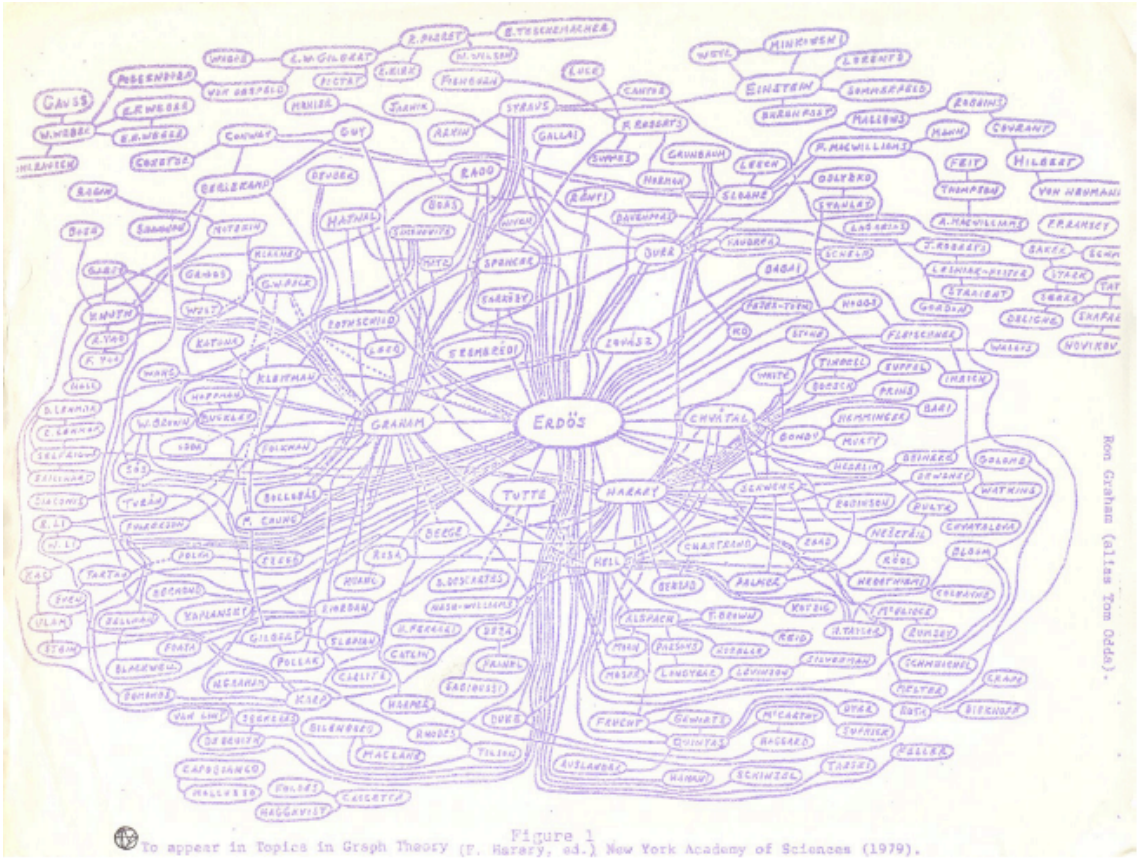
- Two random people are connected through only a few (6) intermediate acquaintances.
- Milgram's experiment (1967) shows the known "six degrees of separation":
  - Choose 300 people at random
  - Ask them to send a letter through friends to a stockbroker near Boston.
  - 64 successful chains.



UNIVERSITY OF CAMBRIDGE

Erdos Number

- Erdos Number: distance from the mathematician (most people are 4-5 hops away) based on collaboration.



UNIVERSITY OF CAMBRIDGE

Bacon Number

- A network of actors who costarred in a movie.
- Most actors are no more than ~3 hops from Kevin Bacon.
- One very obscure movie was at distance 8.

UNIVERSITY OF CAMBRIDGE

Random Graphs

- First way to model these networks:

- **Erdos-Renyi Random Graph** [Erdos-Renyi '59]:

G(n,p): graph with n vertices where an edge exists with independent random probability $0<p<1$ for each edge.

Random Graph Model

- For each node n1, an edge to node n2 exists with probability p.
- Degree distribution is **binomial.**
- The probability of a node to have degree k:

$$P(k_i = k) = C_{N-1}^{k} p^{k} (1-p)^{N-1-k}$$

- Where $C_{N-1}^{k} = \binom{N-1}{k}$

- Expected Degree of a node: $(N-1)p \approx Np$

Random Graphs Properties

- For large N this is approximated by the Poisson distribution with

$$P(k) \approx e^{-Np} \frac{(Np)^k}{k!} = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$

Random Graph Diameter

- The **diameter** of random graph and the **average path length** of the graph have been demonstrated to be:

$$d = \frac{\ln(N)}{\ln(pN)} = \frac{\ln(N)}{\ln(\langle k \rangle)} \approx l_{rand}$$

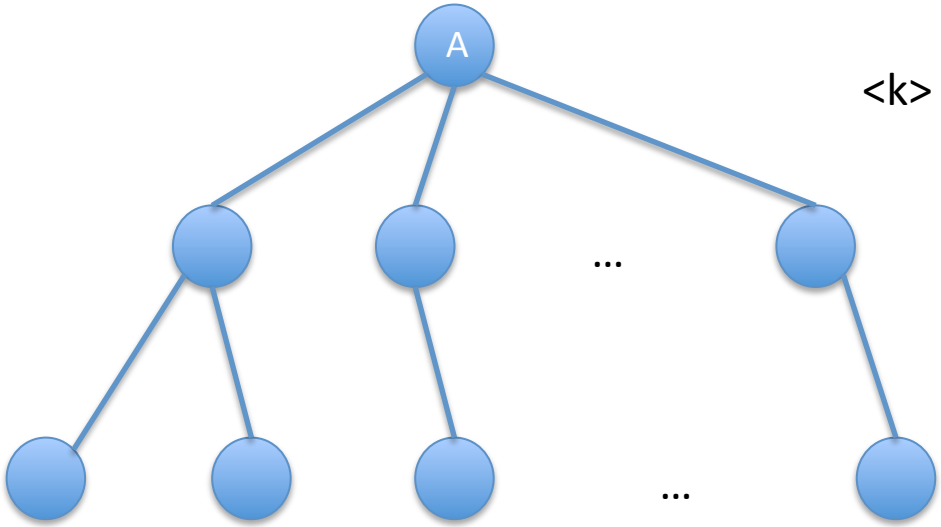
- The average distance between two nodes is quite small wrt to the size of the graph.

UNIVERSITY OF CAMBRIDGE

Relationship of <k> and connectivity

- <k> = average degree (np)
- If <k> < 1 disconnected network
- If <k> > 1 a giant component appears
- If <k> >= ln(N) graph is totally connected

Random Graph Diameter: An Intuition



<k> new connections

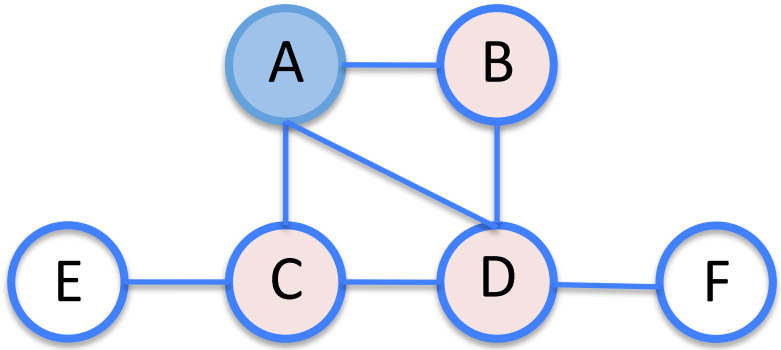
<k> new connections per each new node

- The nodes at distance l from A will be ~ $\langle k \rangle^l$

UNIVERSITY OF CAMBRIDGE

Clustering Coefficient

- The **clustering coefficient** defines the proportion of A's neighbours (N(A)) which are connected by an edge (are friends)



UNIVERSITY OF CAMBRIDGE

Clustering Coefficient of a Random Graph

- The clustering coefficient of a random graph is

$$C_{rand} = p = \frac{\langle k \rangle}{N}$$

- The probability that 2 neighbours of a node are connected is equal to the probability that 2 random nodes are connected
- Is this mirroring the clustering coefficient of real networks?

Clustering Coefficient of Real Networks

- From [Watts and Strogatz, 1998]
- Characteristic path length and clustering coefficient for some real networks and for random networks with same number of nodes and average number of edges per node.

- Aim is to check if random graphs can model real networks.

UNIVERSITY OF CAMBRIDGE

Real Networks vs Random Networks

- Film Actors: actors in movies together
- Power grid: the network of the electricity generators
- C. elegans: network of neurons of a worm
- **L is comparable while C is very different**

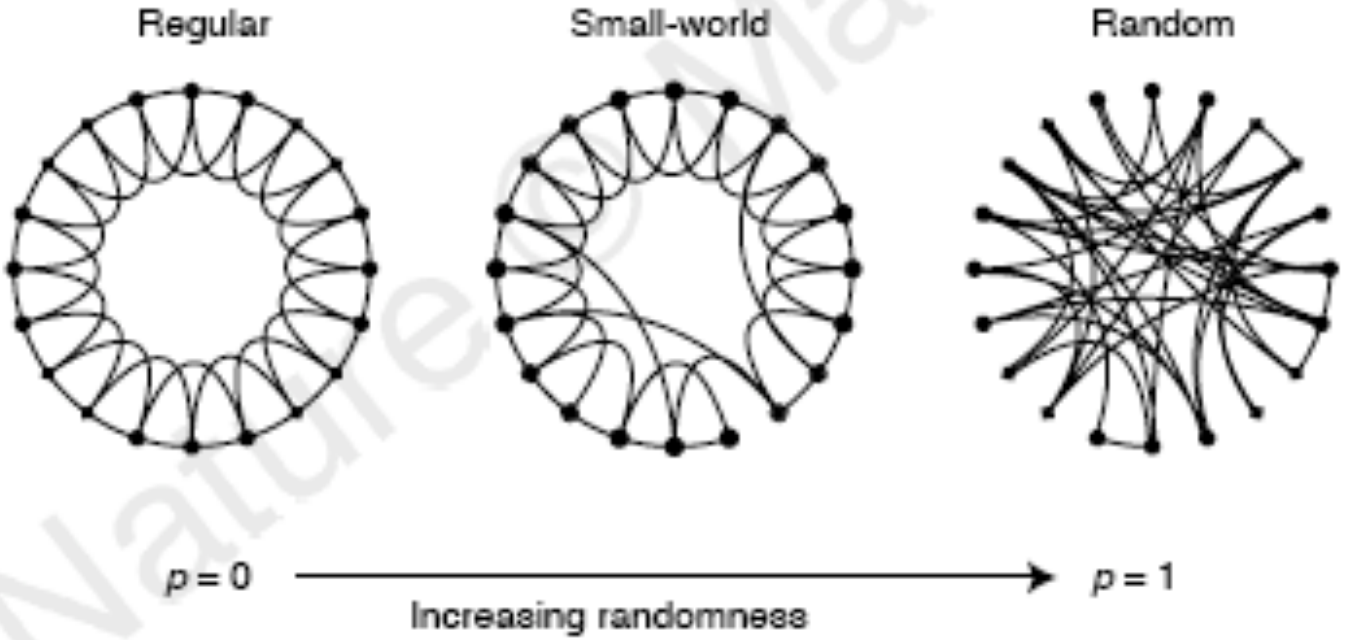
	L_{actual}	L_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power grid	18.7	12.4	0.080	0.005
C. elegans	2.65	2.25	0.28	0.05

UNIVERSITY OF CAMBRIDGE

Small World Model

- Watts & Strogatz built a model which was able to capture these characteristics.
- Start with regular lattice
  - Increase a probability p of “rewiring” a node to another node.
  - When p very high the lattice would become a random graph.

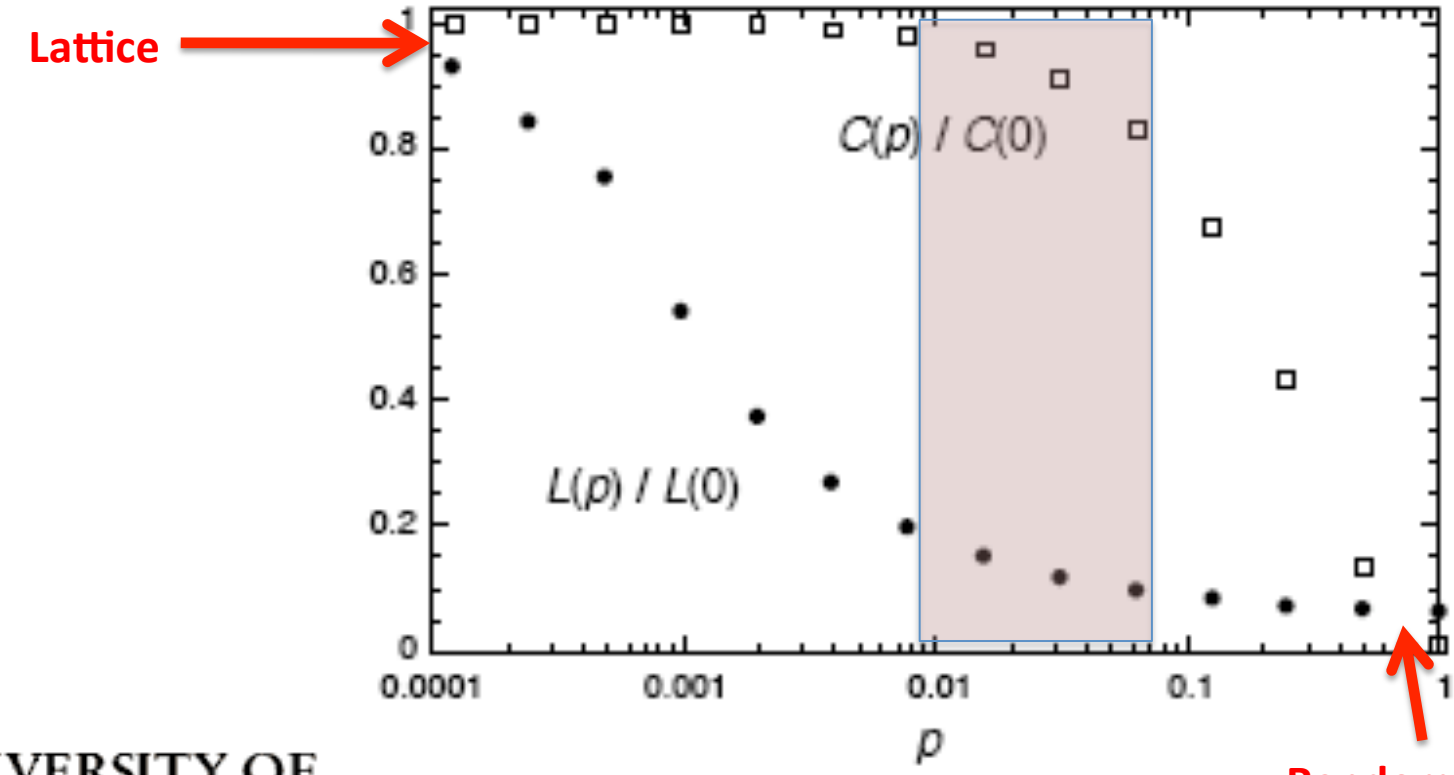
Small World Model (2)



UNIVERSITY OF CAMBRIDGE

How are L and C in this model?

- There is a zone where C is high and L is low
- These are small world networks



UNIVERSITY OF CAMBRIDGE

Other Real Networks Examples

Network	Size	$\langle k \rangle$	ℓ	ℓ_{rand}	C	C_{rand}	Reference	Nr.
WWW, site level, undir.	153 127	35.21	3.1	3.35	0.1078	0.00023	Adamic, 1999	1
Internet, domain level	3015–6209	3.52–4.11	3.7–3.76	6.36–6.18	0.18–0.3	0.001	Yook *et al.*, 2001a, Pastor-Satorras *et al.*, 2001	2
Movie actors	225 226	61	3.65	2.99	0.79	0.00027	Watts and Strogatz, 1998	3
LANL co-authorship	52 909	9.7	5.9	4.79	0.43	1.8×10^{-4}	Newman, 2001a, 2001b, 2001c	4
MEDLINE co-authorship	1 520 251	18.1	4.6	4.91	0.066	1.1×10^{-5}	Newman, 2001a, 2001b, 2001c	5
SPIRES co-authorship	56 627	173	4.0	2.12	0.726	0.003	Newman, 2001a, 2001b, 2001c	6
NCSTRL co-authorship	11 994	3.59	9.7	7.34	0.496	3×10^{-4}	Newman, 2001a, 2001b, 2001c	7
Math. co-authorship	70 975	3.9	9.5	8.2	0.59	5.4×10^{-5}	Barabási *et al.*, 2001	8
Neurosci. co-authorship	209 293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabási *et al.*, 2001	9
E. coli, substrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000	10
E. coli, reaction graph	315	28.3	2.62	1.98	0.59	0.09	Wagner and Fell, 2000	11
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06	Montoya and Solé, 2000	12
Silwood Park food web	154	4.75	3.40	3.23	0.15	0.03	Montoya and Solé, 2000	13
Words, co-occurrence	460.902	70.13	2.67	3.03	0.437	0.0001	Ferrer i Cancho and Solé, 2001	14
Words, synonyms	22 311	13.48	4.5	3.84	0.7	0.0006	Yook *et al.*, 2001b	15
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998	16
C. Elegans	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998	17

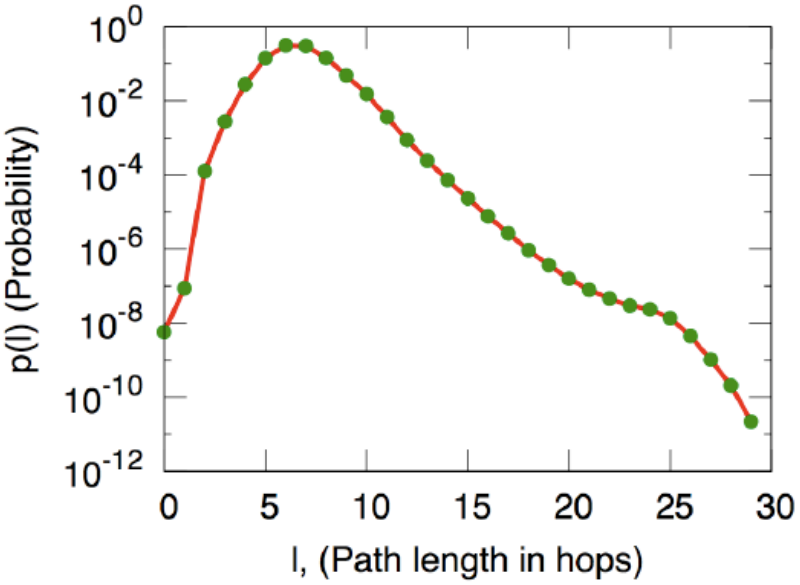
Analysis of Messenger Network

- [Leskovec and Horvitz 2008] analyzed a large dataset of the Microsoft Messenger.
- Communication Network contained 180 million users and 1.3 billion conversations in 1 month.
- Buddy Network contained 240 million users.
- *99.9% users belonged to a connected component.*

UNIVERSITY OF CAMBRIDGE

Analysis of a Messenger Network

- Average shortest path is 6.6 (confirming Milgram's study).
- Although some longer paths up to 29.
- Average clustering coefficient is quite high: 0.137.



UNIVERSITY OF CAMBRIDGE

Summary

- We have introduced graphs definitions and measures.
- Random graphs are a first examples of models for networks.
- Small world network models are able to capture a good quantity of real networks
  - They have characteristic path length comparable to random networks.
  - But much higher clustering coefficient.

References

- Material from Chapter 1, 2 and 20 of
  - **D. Easley, J. Kleinberg. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, 2010.**
- R. Albert, A. Barabasi. Statistical Mechanics of Complex Networks. Reviews of Modern Physics (74). Jan. 2002.
- S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex Networks: Structure and Dynamics Physics Reports 424 (2006) 175 .
- Watts, D.J.; Strogatz, S.H. (1998). "Collective dynamics of 'small-world' networks." *Nature* 393 (6684): 409–10.

UNIVERSITY OF CAMBRIDGE