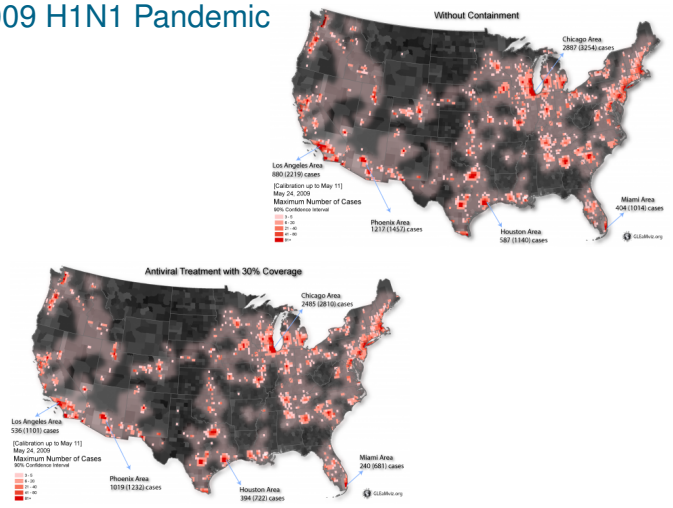
Topic 1: Introduction and Network Models

Matthew Johnson

matthew.johnson2@durham.ac.uk

2009 H1N1 Pandemic

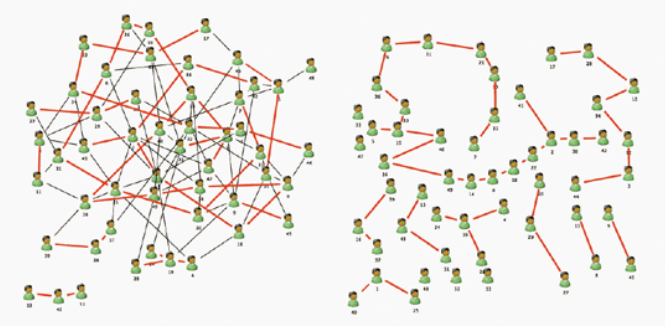


Questions for Computer Scientists

- How do we define the contact network for a disease?
- How can we model how a disease spreads through a network?
- Can we quantify and predict the effect of preventative action?

Case Study: `http://tinyurl.com/2009h1n1pandemic`

Social Networks

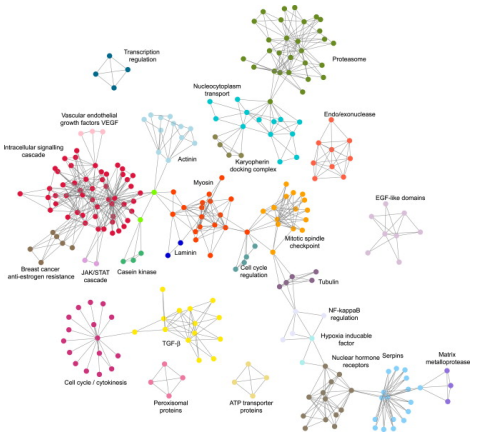


More Questions for Computer Scientists

- How do we decide which person is most important in the network?
- Or who is the most influential who we should target (as an advertiser)?
- Is this equivalent to asking which webpages a search engine should promote?

Page, Brin, Motwani and Winograd, The PageRank Citation Ranking: Bringing Order to the Web (1999).
`dbpubs.stanford.edu:8090/pub/1999-66`

Proteins



Yet more questions

- How can we find communities in the network of interacting proteins?
- Can we match these structures to the network's function?
- Should we assume that communities do not overlap?

Barabasi and Oltvai, Network biology: understanding the cell's functional organization (2004). `http://tinyurl.com/BarabasiOltvai`

And . . .

- What is the *small world effect* and how can we recognize it?
- Why might we want in distributed data storage systems?

Liu, Mackin, Antonopoulos, Small World Architecture for Peer-to-Peer Networks. `https://tinyurl.com/smallworldP2P`

Science of Networks

Networks are found everywhere in nature and society. Their components — the nodes and links — are diverse:

- in social networks they are people and their friendships;
- in the web they are pages and hyperlinks;
- in protein networks they are macromolecules and their biochemical interactions;

But the structure and the evolution of networks in various domains are often similar. We will see that we can uncover specific universal network properties, and explore their consequences.

Science of Networks

- Quantitative and mathematical: we will use techniques from graph theory (but also probability, information theory, statistics . . .)
- Computational: networks are typically very large and we need to implement our algorithms to explore them.
- Data driven: we will test what we find on real data.

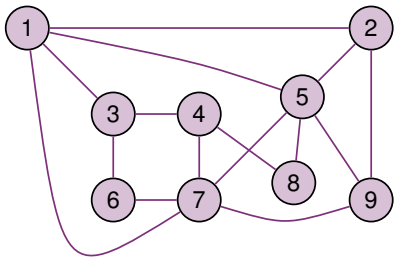
Course Arrangements

- 10 one-hour lectures.
- Books: see duo for suggestions.
- Assessment:
  - Coursework only: deadline is 2pm on 10 February 2022.
- Office hour: Mondays 10.30am–11.30am

Course Content

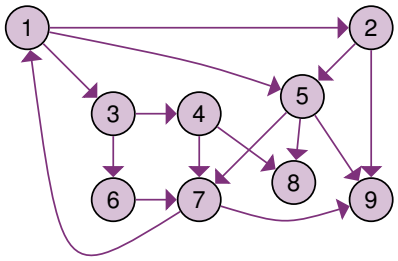
- Introduction and network models
- The small world effect and searching in networks
- Important nodes and communities
- Epidemics

Graphs



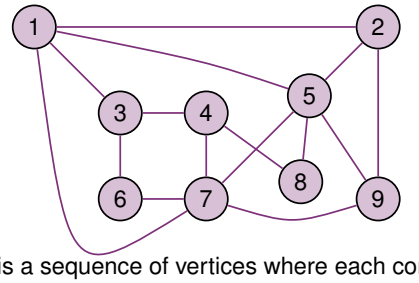
- A graph G is a pair of sets: vertices V and edges E. Each edge joins a pair of vertices.
- The neighbours of a vertex v are vertices joined to v by an edge.
- The degree of v is the number of neighbours it has.

Directed Graphs



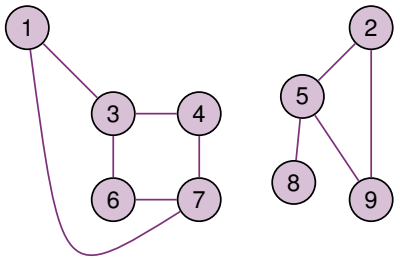
- Graphs can be directed (we will assume they are undirected unless stated)
- A vertex u is an out-neighbour of a vertex v if there is an edge from v to u.
- The out-degree of v is the number of out-neighbours it has.

Paths and Cycles



- A path is a sequence of vertices where each consecutive pair is joined by an edge (that points in the right direction if the graph is directed.
- A cycle is a path that starts and ends in the same vertex.
- The distance between a pair of vertices is the length of the shortest path between them. The diameter of a graph is the maximum distance between a pair of vertices.

Connectivity



- A graph is connected if every pair of vertices is joined by a path. Otherwise it is disconnected.
- A connected component is a maximal set of vertices such that each pair is joined by a path.

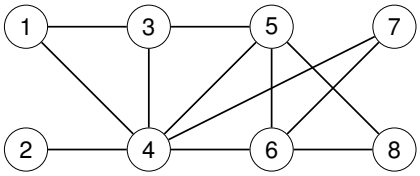
Exercise

- A vertex X is pivotal for a pair of distinct vertices Y and Z if X lies on every shortest path from Y to Z.
- Find
  1. a graph if which every vertex is pivotal for at least one pair of vertices;
  2. a graph if which every vertex is pivotal for at least two pairs of vertices;
  3. a graph (which has at least 4 vertices) which contains a vertex X that is pivotal for every pair of vertices (not including pairs that contain X).

Revision: breadth-first search

- Input: a graph $G = (V, E)$ and a source vertex s.
- Aim: to find the distance from s to each of the other vertices in the graph.
- Idea: send out a wave from s.
  - The wave first hits vertices at distance 1
  - Then the wave hits vertices at distance 2
  - and so on

Example



- Initialization: create queue Q containing source vertex.
- while the queue is not empty
  - remove first vertex v from the queue
  - add undiscovered neighbours of v to queue; distance is 1 greater than to v

Counting shortest paths

- Can we adapt breadth-first search so that it also counts the number of shortests paths from the source to every other vertex?
- Suppose there is one unit of flow between each pair of vertices u and v that is divided equally between all shortest paths between u and v. Can we determine how much of this flow goes through each other vertex?

Network Models

A model of a network typically states its size and how links are created.

Why use models?

- make comparisons
  - how big (or how dense) does a network need to be to be connected?
  - how many links do we need to be resilient to failure, or resistant to attack?
- make predictions
  - how will a network evolve as it grows?
  - what affect will actions on the network have (will our vaccination policy work)?
- explain observations
  - why do social networks have short average path lengths?
  - why do we find clusters in networks?

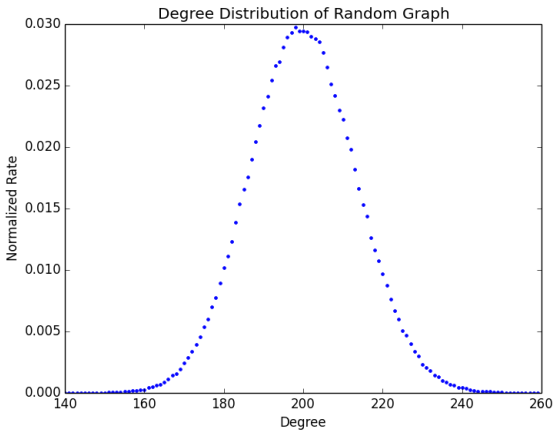
Random Graphs: Erdős-Rényi

The simplest random model of graphs was introduced by Erdős and Rényi in 1960. In fact, there are two variants:

- $G(n, p)$: there are n vertices and each pair is connected with probability p.
- $G(n, m)$: there are n vertices and m edges picked at random.

We look at the first.

Properties of $G(n, p)$: degree distribution



Properties of $G(n, p)$: degree distribution

What is the probability that a vertex v is linked to exactly k other vertices?

Suppose the vertices of G are v and $u_1, u_2, \ldots, u_{n-1}$. What is the probability that v is linked to $u_1, u_2, \ldots, u_k$?

The probability that v is linked to a vertex is p. The probability that v is not linked to a vertex is $1 - p$. So the probability that v is linked to $u_1, u_2, \ldots, u_k$ is

$$p^k(1-p)^{n-1-k}$$

So that is the probability that v is linked to a particular set of k other vertices ...

... and the number of different sets of k vertices (that don't include v) is

$$\binom{n-1}{k}$$

So the probability that v has degree k is

$$\binom{n-1}{k} p^k (1-p)^{n-1-k},$$

the binomial distribution Binomial$(n-1, p)$.

Expected Degree

The expected degree is

$$E[d] = \sum P[d = k].k = (n-1)p$$

For the distribution, we must ask how close to the expected value are actual degrees likely to be?

Chernoff Bound: if X is Binomial(n, p) and $\epsilon \leq 2/3$, then

$$P[|X - E[x]| \geq \epsilon E[X]] \leq 2\exp(-\frac{1}{3}\epsilon^2 E[X])$$

That is, the probability that X deviates from $E[X]$ is exponentially small; the degree concentrates on the mean.

Properties of $G(n, p)$: giant component

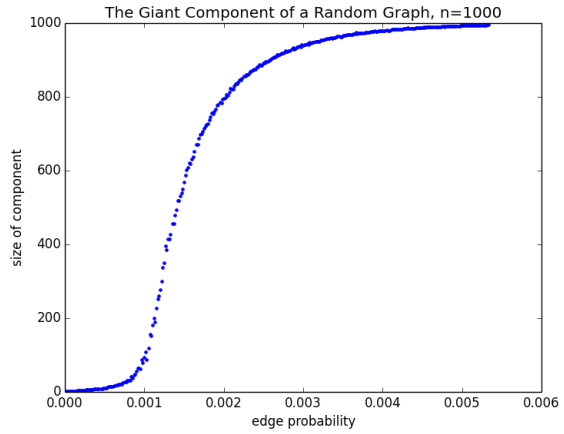
What happens when p is varied (for fixed n)?

When $p = 0$, the graph is empty: there are no edges.

When $p = 1$, the graph is complete: every pair of vertices is joined by an edge.

The giant component of a random graph is the connected component with the greatest number of vertices. What happens to the size of the giant component when p varies?

Properties of $G(n, p)$: giant component



Properties of $G(n, p)$: giant component

It appears that at when there is less than one edge per vertex on average, the components are all tiny (this is surely unsurprising).

What is perhaps more unexpected is that once the average number of edges per vertex is more than one, we soon find that most of the vertices belong to the same component.

The rapid change from one extreme state to the other is known as a phase change.

Properties of $G(n, p)$: local clustering coefficient

The clustering coefficient of a vertex measures the extent to which the neighbourhood of the vertex forms a complete graph.

For vertex i, let k_i be its degree and let e_i be the number of edges that join its neighbours. Then the clustering coefficient of i is

$$C_i = \frac{2e_i}{k_i(k_i - 1)}.$$

What is the local clustering coefficient of a vertex in a random graph?

For a random graph, each pair of a vertex i's k_i neighbours is linked with probability p. So we expect that there are

$$\binom{k_i}{2} p = pk_i(k_i - 1)/2$$

edges in its neighbourhood. And so we expect that the clustering coefficient of a random graph is p. Much smaller than observed in real networks.

Properties of $G(n, p)$: diameter

The distance between a pair of vertices in a graph is the length of a shortest path between them.
The diameter of a graph is the maximum distance (over all pairs).

How can we find the diameter of a random graph?

Let k be the average degree (and we know $k = np$ and that all nodes have degree close to k). As the clustering coefficient is small, the topology of a random graph is tree-like.

- a vertex has k neighbours,
- and there are about k^2 vertices at distance 2,
- and k^3 vertices at distance 3,
- and so on ...

Properties of $G(n, p)$: diameter

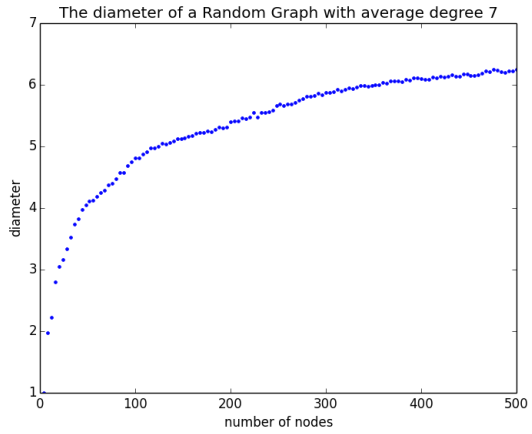
So if d is the maximum distance in the graph, we have

$$n = 1 + k + k^2 + \cdots + k^d = \frac{k^{d+1} - 1}{k - 1}$$

which is approximately k^d. Rearranging we find

$$d = \frac{\log n}{\log k}.$$

Properties of $G(n, p)$: diameter



Average path lengths in networks

(degrees and path lengths are averages)

network	nodes	links	degree, k	path length, d	$\frac{\log N}{\log k}$
internet	192,244	609,066	6.34	6.98	6.59
web	325,729	1,497,134	4.60	11.27	8.32
power grid	4,941	6,594	2.67	18.99	8.66
scientists	23,133	186,936	8.08	5.35	4.81
citations	449,673	4,707,958	10.47	11.21	5.55
e coli	1,039	5,802	5.84	2.98	4.04
yeast	2,018	2,930	2.90	5.61	7.14

Network Models

- It seems that random graphs are not always a good model for real networks.
- Most networks we know evolved over time so, even if we are looking only at a static snapshot, we might expect this to be reflected in the definition of our model.

Preferential Attachment

- A growing network exhibits preferential attachment if new nodes are more likely to connect to existing nodes that are already well-connected.
- The model we will use is sometimes known as the Barabási-Albert (BA) model. In this model, the network begins with m_0 nodes and then
  - at each time step a new node is created with $m \leq m_0$ links to existing nodes;
  - the probability that the new node links to an existing node v depends on the degree of v.

Note that the BA model leaves some details open. How are the original nodes connected? Are the links from a new node all added together or one by one? Do we allow parallel links?

A history of Prefential Attachment

The idea of Preferential Attachment emerged independently many times.

- György Pólya, Mathematician 1923
  Developed the urn model: an urn contains black and white balls, one is taken out and replaced with an additional ball of the same colour.
- George Udmy Yule, Statistician 1925
  Developed the Yule process to explain the numbers of species per genus of flowering plants
- Robert Gibrat, Economist 1931
  Used preferential attachment, calling it proportional growth to explain why large firms grow faster.
- George Kinsley Zipf, Economist 1941
  Used preferential attachment to explain the fat tailed distribution of wealth in society.

A history of Prefential Attachment continued

- Herbert Alexander Simon, Political Scientist 1955
  More fat tails: explained the distribution of city sizes, word frequencies or the number of papers published by scientists.
- Derek de Solla Price, Physicist 1968
  Explained citation statistics calling it cumulative advantage
- Robert Merton, Sociologist 1976
  Applied preferential attachment to sociology, coining the term Matthew effect

The Barabási-Albert network was proposed in 1999.