Birkbeck
UNIVERSITY OF LONDON

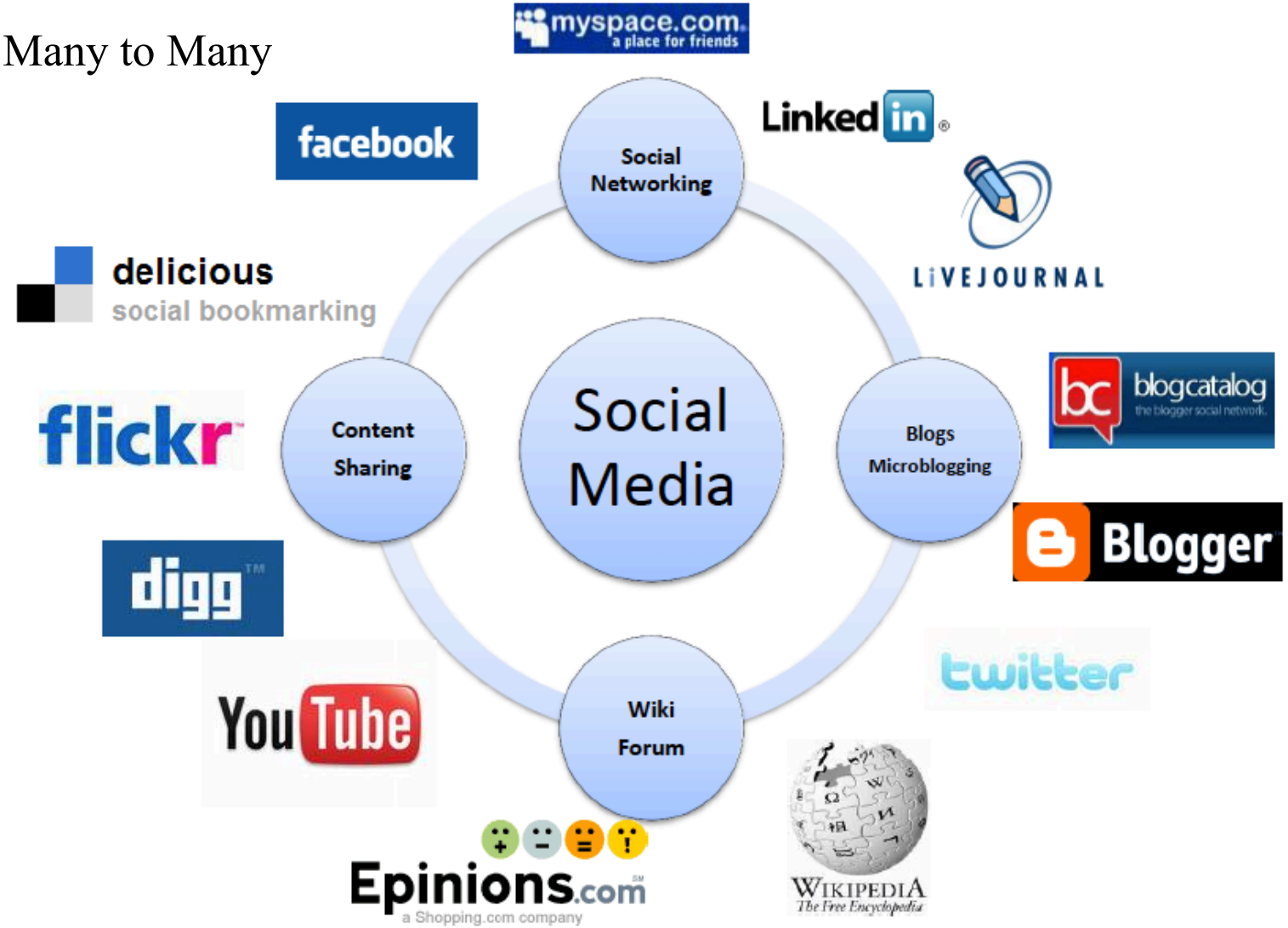
Big Data Analytics

Session

Social Networks

Social Media

- Many to Many



Characteristics of Social Media

- "Consumers" become "Producers"
- Rich User Interaction
- User-Generated Contents
- Collaborative Environment
- Collective Wisdom
- Long Tail

(2010)

Broadcast Media
Filter, then Publish

→

Social Media
Publish, then Filter

Top 20 Websites at USA

1	Google.com	11	Blogger.com
2	Facebook.com	12	msn.com
3	Yahoo.com	13	Myspace.com
4	YouTube.com	14	Go.com
5	Amazon.com	15	Bing.com
6	Wikipedia.org	16	AOL.com
7	Craigslist.org	17	LinkedIn.com
8	Twitter.com	18	CNN.com
9	Ebay.com	19	Espn.go.com
10	Live.com	20	Wordpress.com

40% of websites are social media sites

Social Network and Media

Birkbeck
UNIVERSITY OF LONDON

- Social Network
  - The networks formed by individuals
  - Graph
  - Node
  - Link

- Social Media
  - Social network + media
  - Media = content of twitter, tag, videos, photos

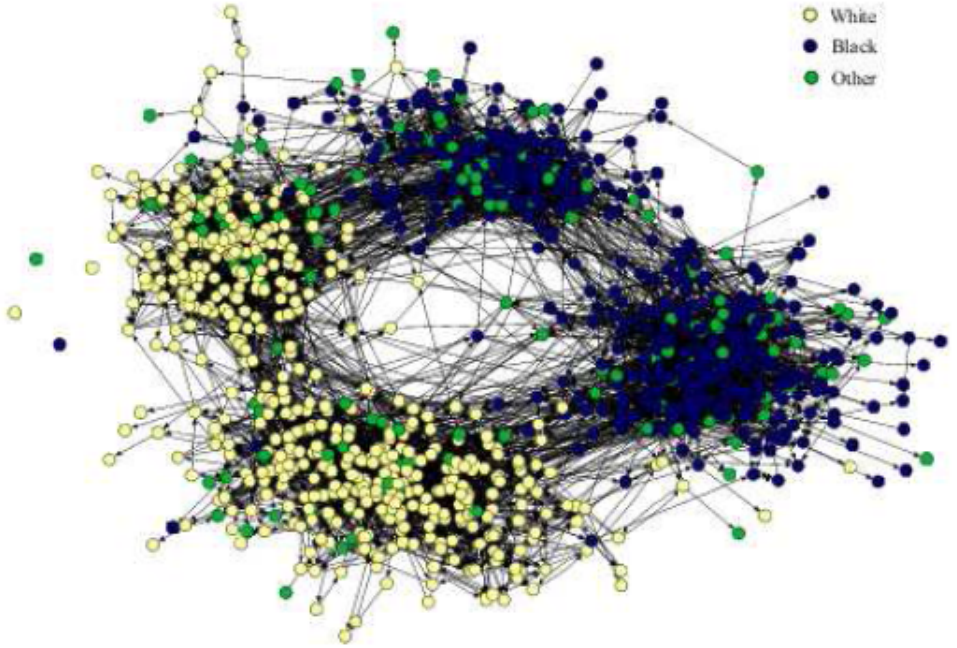
What is a social network?

Birkbeck UNIVERSITY OF LONDON

- Facebook
- LinkedIn
- ….
- The network of your friends and acquaintances
- Social network is a graph **G=(V,E)**
  - V: set of users
  - E: connections/friendships among users

Social Networks

- Links denote a social interaction
  - Networks of acquaintances
  - Collaboration networks
    - actor networks
    - co-authorship networks
    - director networks
  - phone-call networks
  - E-mail networks
  - IM networks
  - Bluetooth networks
  - Home pages
  - Blog networks



Data analysis for social networks

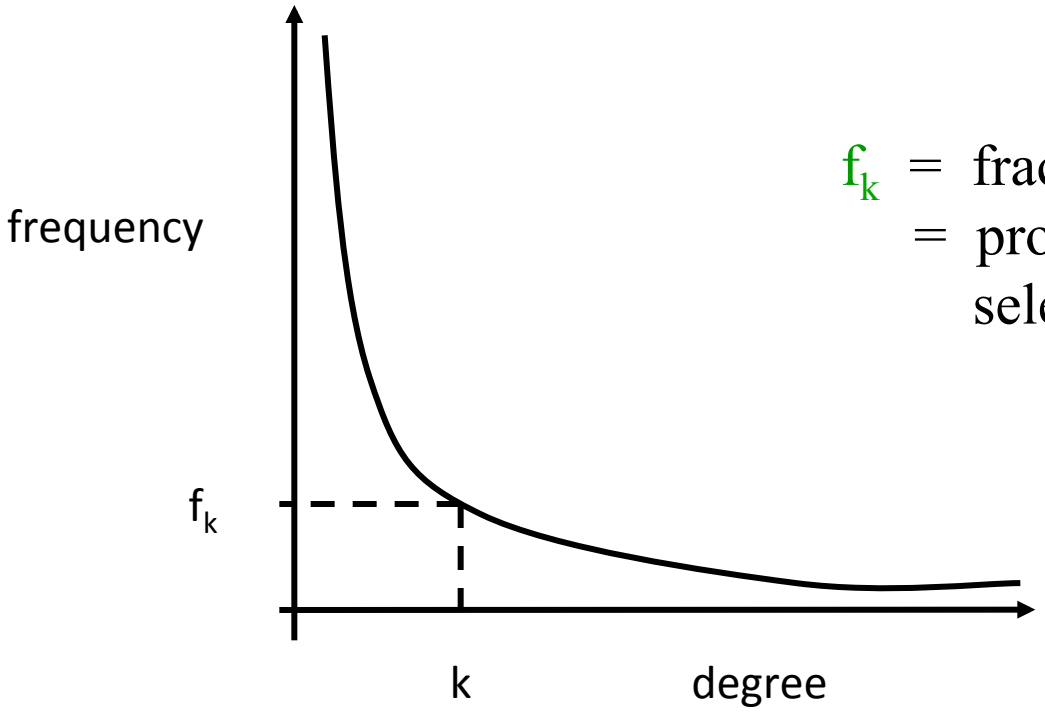
Birkbeck
UNIVERSITY OF LONDON

- Measure characteristics of social networks
  - How many hops apart are two random Facebook users
- Design models that capture the generation process of network data
  - Generate graphs with the same properties as real social network graphs
- Algorithmic problems related to
  - Information propagation
  - Advertising
  - Expertise finding
  - Privacy

Measuring Networks

- Degree distributions
- Small world phenomena

Degree Distributions



f_k = fraction of nodes with degree k
= probability of a randomly selected node to have degree k

- Problem: find the probability distribution that best fits the observed data

Power-Law Distributions

Birkbeck UNIVERSITY OF LONDON

- The degree distributions of most real-life networks follow a power law

$$p(k) = Ck^{-\alpha} = C/k^{\alpha}$$

- Right-skewed/Heavy-tail distribution
  - there is a non-negligible fraction of nodes that has very high degree (hubs)
  - scale-free: no characteristic scale, average is not informative

- The probability that any node is connected to *k* other nodes is proportional to $1/k^{\alpha}$

Power-Law Distributions

- The degree distributions of most real-life networks follow a power law

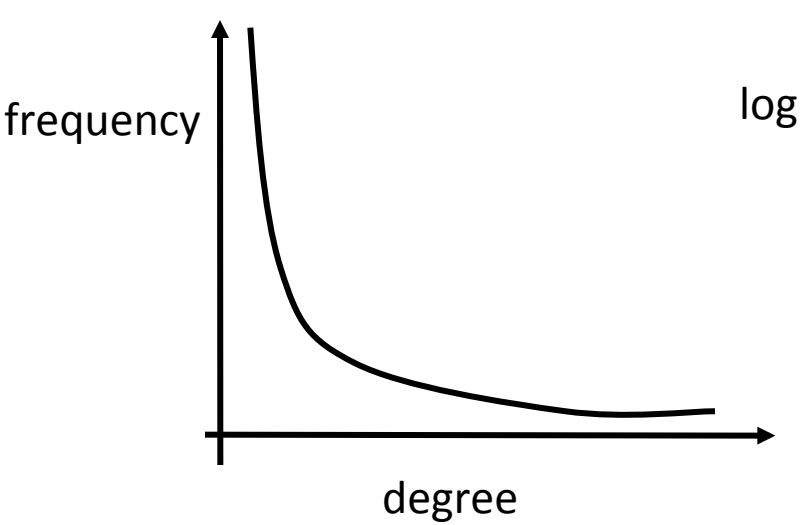
$$p(k) = 1/k^2$$

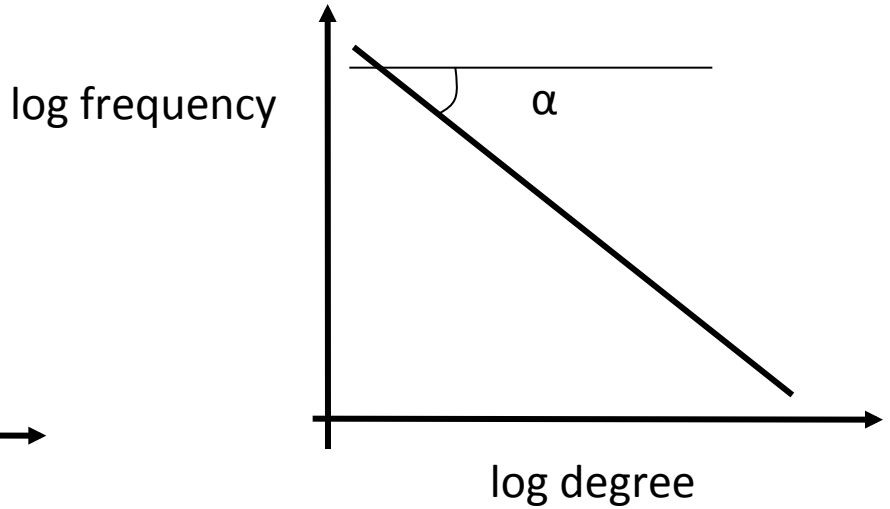
$$P(1) = 1$$
$$P(2) = 1/4$$
$$P(3) = 1/8$$
$$P(4) = 1/16$$
$$P(5) = 1/25$$
$$\ldots$$

Power-law signature

- Power-law distribution gives a line in the log-log plot

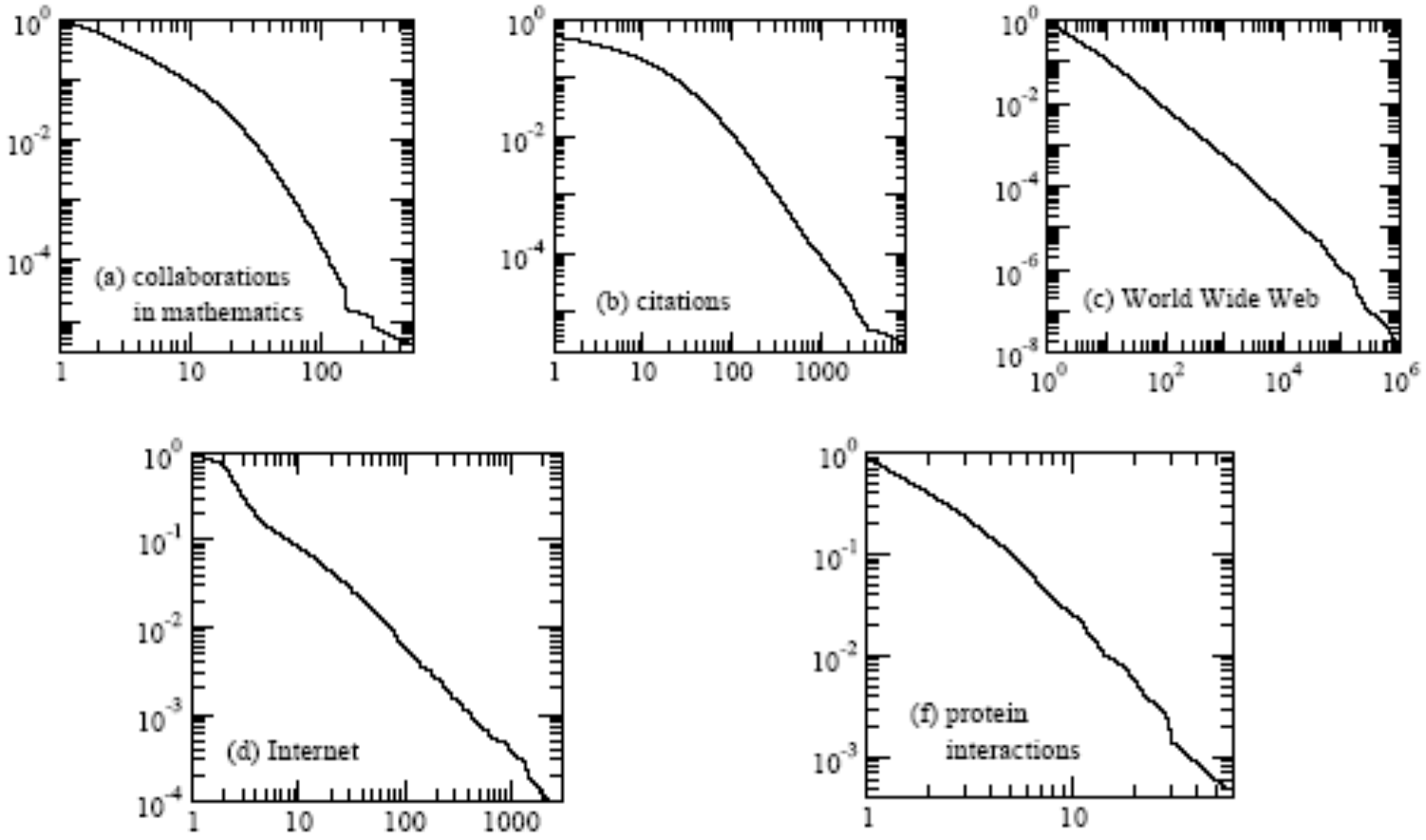
$$\log p(k) = -\alpha \log k + \log C$$





- α: power-law exponent (typically $2 \leq \alpha \leq 3$)

Examples



Taken from [Newman 2003]

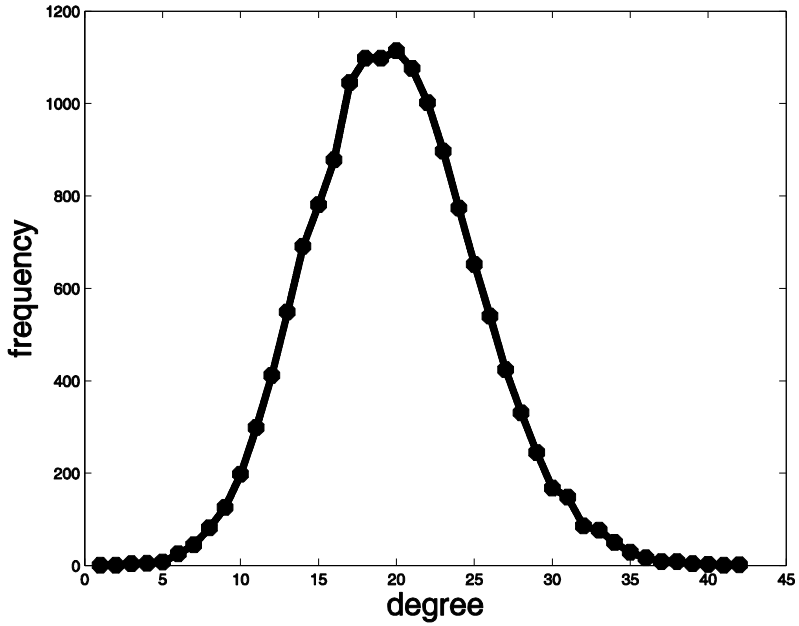
The basic random graph model

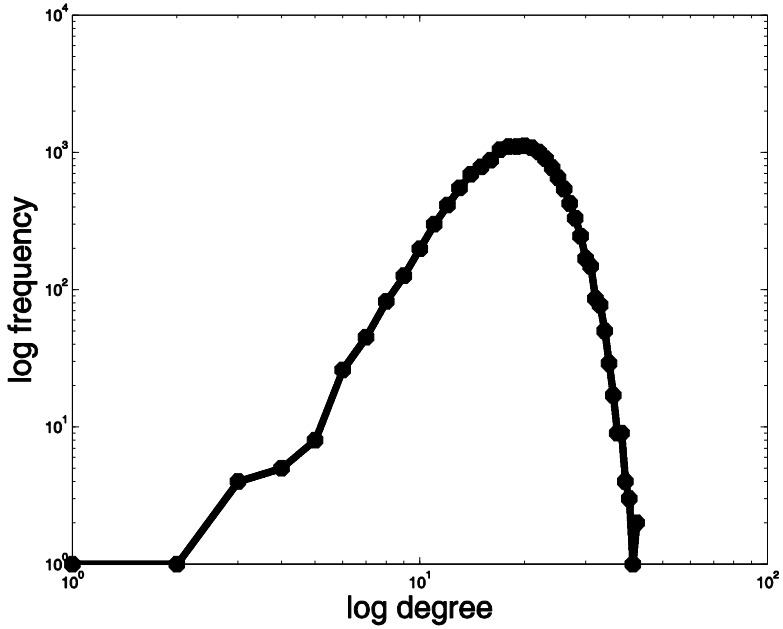
Birkbeck UNIVERSITY OF LONDON

- The measurements on real networks are usually compared against those on "random networks"

- The basic $G_{n,p}$ (Erdös-Renyi) random graph model:
  - n : the number of vertices
  - $0 \leq p \leq 1$
  - for each pair (i,j) generate the edge (i,j) independently with probability p

A random graph example





Average/Expected degree

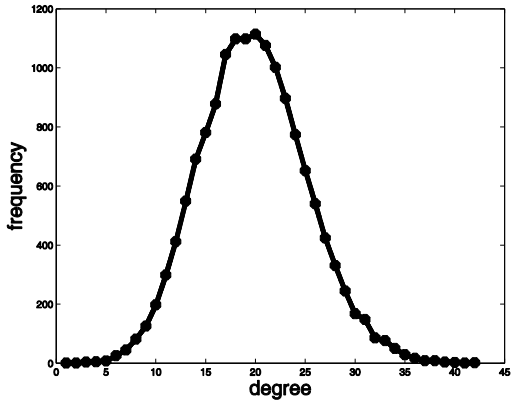
Birkbeck UNIVERSITY OF LONDON

- For random graphs the expected node degree is: $z = np$
  - n: number of nodes
  - p: probability of an edge

- For power-law distributed graphs the expected node degree is
  - constant, if $\alpha \geq 2$
  - diverges, if $\alpha < 2$

Maximum degree

- For random graphs, the maximum degree is highly concentrated around the average degree z



- For power law graphs the maximum degree is:

$$z_{\max} \approx n^{1/(\alpha-1)}$$

  - where α is the exponent of the power law

The small-world experiment

- Milgram 1967
- Picked 296 people at random from Omaha, Nebraska, Wichita, and Kansas
- Asked them to get a letter to a stockbroker in Boston
- Rule: they could bypass the letter through friends they knew on a first-name basis
- How many steps does it take?
  - Six on average

The small-world experiment

Birkbeck
UNIVERSITY OF LONDON

- 64 chains completed
  - 6.2 average chain length
  - thus “six degrees of separation”
- Critique
  - Several times people refused to forward the package
  - People had no knowledge of the topology of the network, hence the package may not follow the shortest path

Six Degrees of Kevin Bacon

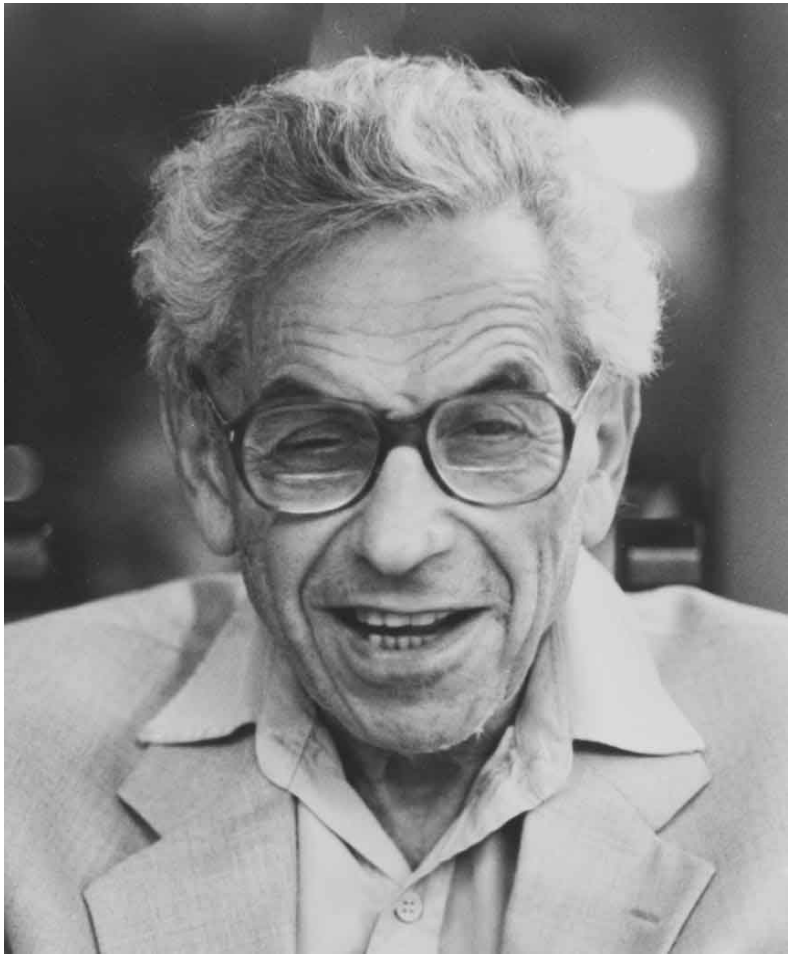
Birkbeck UNIVERSITY OF LONDON

- Bacon number:
  - Create a network of Hollywood actors
  - Connect two actors if they co-appeared in some movie
  - Bacon number: number of steps to Kevin Bacon

- As of Dec 2007, the highest (finite) Bacon number reported is 8

- Only approx 12% of all actors cannot be linked to Bacon

Erdos numbers?



Measuring the Small World Phenomenon

Birkbeck UNIVERSITY OF LONDON

- d_{ij} = shortest path between i and j
- Diameter:

$$d = \max_{i,j} d_{ij}$$

- Characteristic path length:

$$\ell = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij}$$

- Harmonic mean:

$$\ell^{-1} = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij}^{-1}$$

What is a network model?

- Informally, a network model is
  - a process (randomised or deterministic) for generating a graph
- Models of static graphs
  - input: a set of parameters Π and the size of the graph n
  - output: a graph $G(\Pi,n)$
- Models of evolving graphs
  - input: a set of parameters Π and an initial graph G_0
  - output: a graph G_t for each time t

Families of Random Graphs

Birkbeck UNIVERSITY OF LONDON

- A deterministic model D defines a single graph for each value of n (or t)

- A randomised model R defines a probability space $\langle G_n, P \rangle$ where G_n is the set of all graphs of size n, and P a probability distribution over the set G_n (similarly for t)
  - we call this a family of random graphs R, or a random graph R

Erdös-Renyi Random Graphs

Birkbeck UNIVERSITY OF LONDON

- The $G_{n,p}$ model
  - input: the number of vertices n, and a parameter p, $0 \leq p \leq 1$
  - process: for each pair (i,j), generate the edge (i,j) independently with probability p
- The $G_{n,m}$ random model:
  - process: select m edges uniformly at random

Random graphs degree distributions

Birkbeck UNIVERSITY OF LONDON

- The degree distribution follows a binomial distribution

$$p(k) = B(n;k;p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Assuming $z=np$ is fixed, as $n \to \infty$, $B(n,k,p)$ is approximated by a Poisson distribution

$$p(k) = P(k;z) = \frac{z^k}{k!} e^{-z}$$

- Highly concentrated around the mean, with a tail that drops exponentially

Random Graphs and Real Life

Birkbeck
UNIVERSITY OF LONDON

- A beautiful and elegant theory studied exhaustively
- Random graphs had been used as idealized network models
- Unfortunately, they don't capture reality…

Barabasi-Albert Model

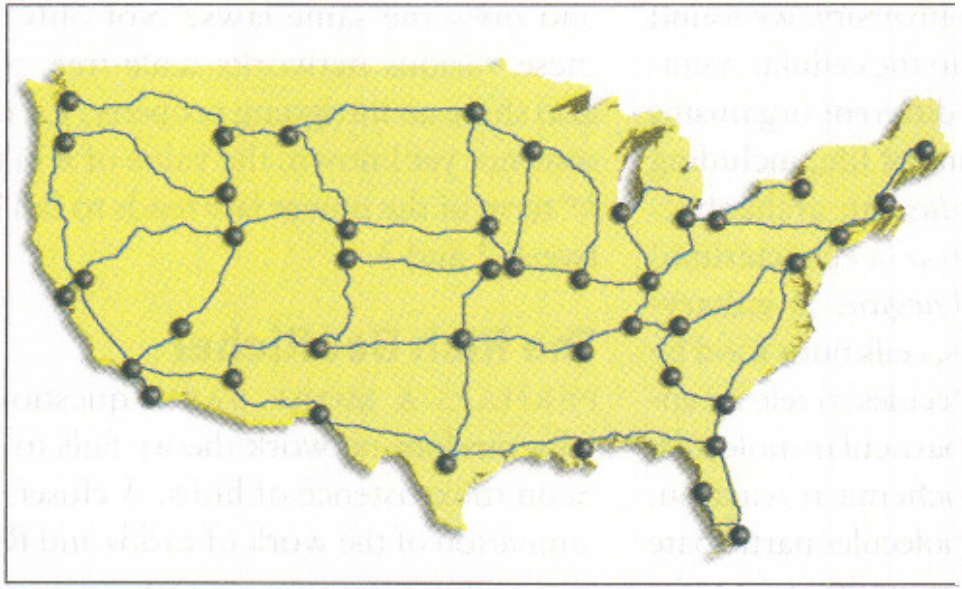
Birkbeck UNIVERSITY OF LONDON

- The BA model (undirected graph)
  - input:
    - some initial subgraph G_0 and
    - m: number of edges per new node
  - the process:
    - nodes arrive one at the time
    - each node connects to m other nodes selecting them with probability proportional to their degree
    - if $[d_1,\ldots,d_t]$ is the degree sequence at time t, then node t+1 links to node i with probability $\frac{d_i}{\sum_i d_i}$
- Results in power-law with exponent $\alpha = 3$

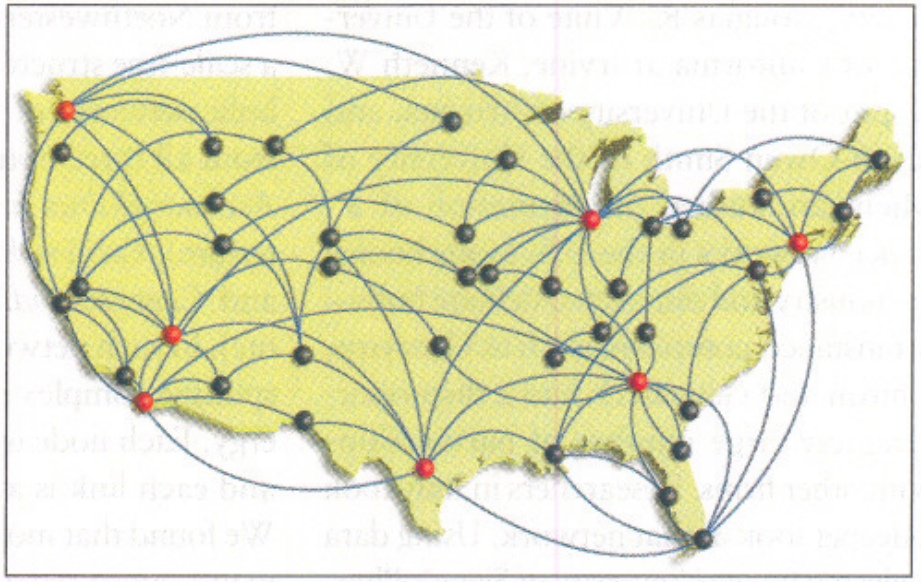
Barabasi-Albert model

Birkbeck
UNIVERSITY OF LONDON

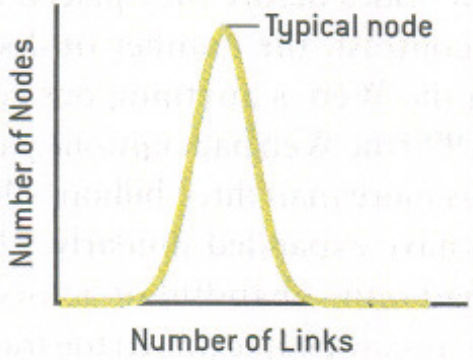
Random Network



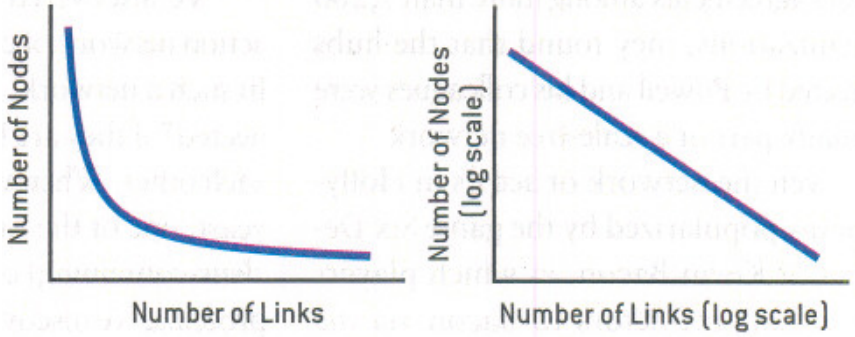
Scale-Free Network



Bell Curve Distribution of Node Linkages



Power Law Distribution of Node Linkages



Epidemic Processes

Birkbeck
UNIVERSITY OF LONDON

- Viruses, diseases
- Online viruses, worms
- Fashion
- Adoption of technologies
- Behavior
- Ideas

Example: Ebola virus

- First emerged in Zaire 1976 (now Democratic Republic of Kongo)
- Very lethal: it can kill somebody within a few days
- A small outbreak in 2000
- From 10/2000 – 01/2009 173 people died in African villages

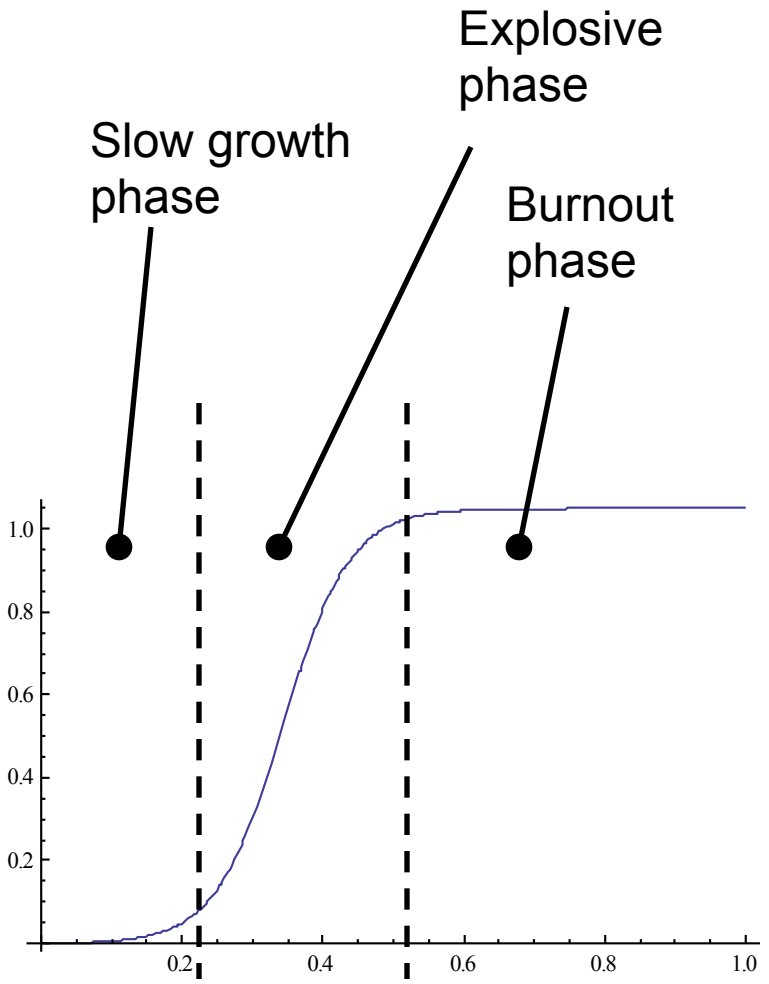
Example: Melissa computer worm

- Started on March 1999
- Infected MS Outlook users
- The user
  - Receives email with a word document with a virus
  - Once opened, the virus sends itself to the first 50 users in the outlook address book
- First detected on Friday, March 26
- On Monday had infected >100K computers

The Bass model

- Introduced in the 60s to describe product adoption
- Can be applied for viruses
- F(t): Ratio of infected at time t
- p: Rate of infection by outside
- q: Rate of contagion

$$F(t) = \frac{1 - e^{-(p+q)t}}{1 + \frac{q}{p} e^{-(p+q)t}}$$



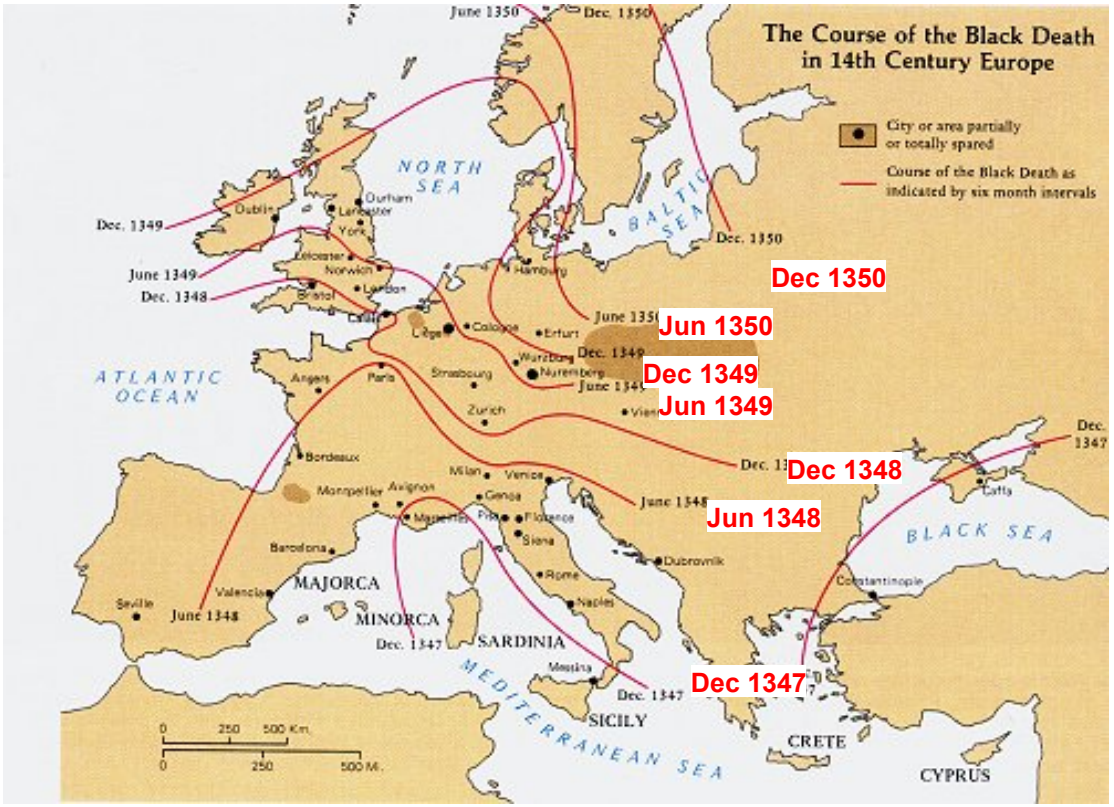
Network Structure

- The Bass model does not take into account network structure

- Let’s see some examples

Example: Black Death (Plague)

- Started in 1347 in a village in South Italy from a ship that arrived from China
- Propagated through rats, etc.



Example: Mad-cow disease

Birkbeck
UNIVERSITY OF LONDON

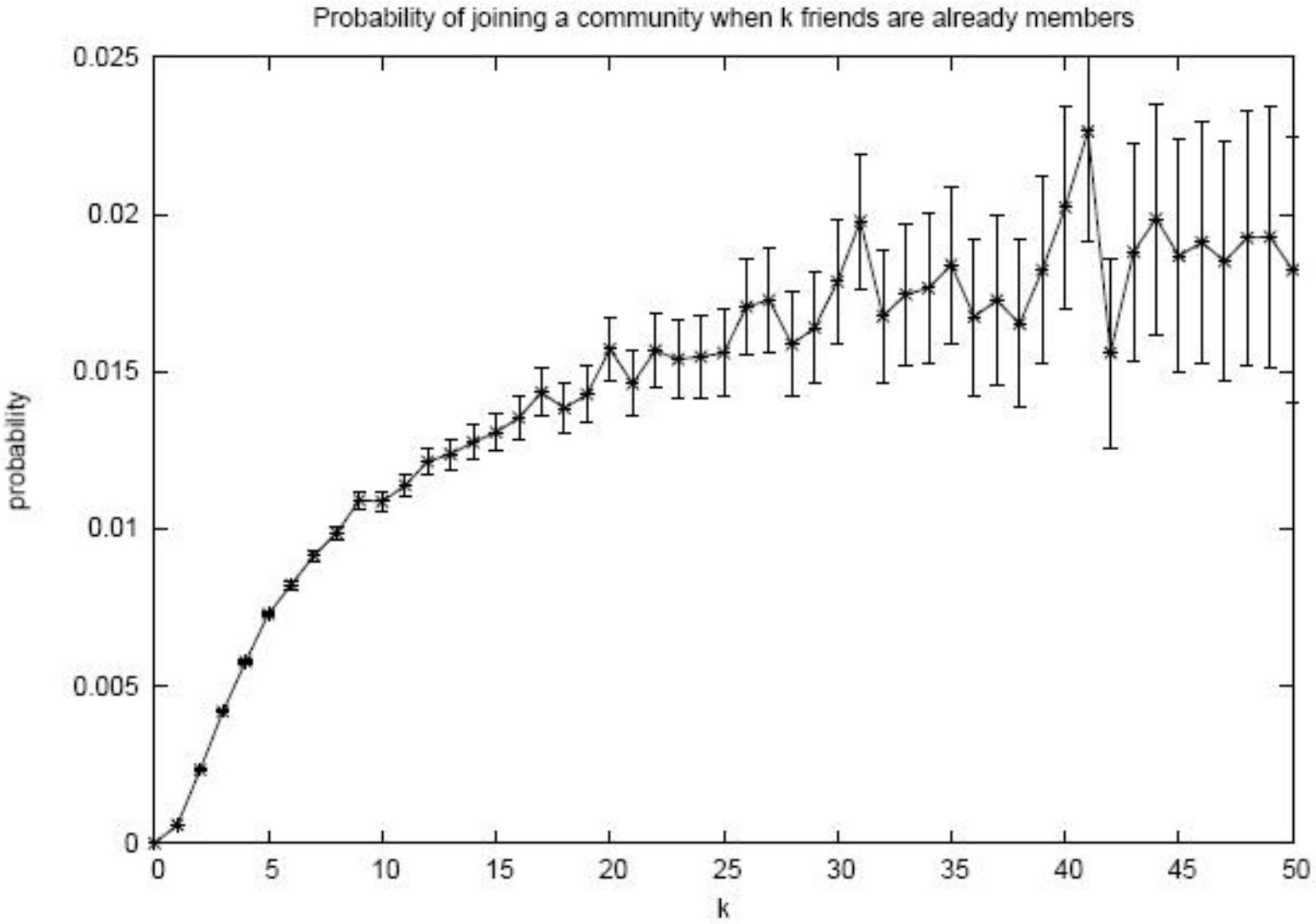
- Jan. 2001: First cases observed in UK
- Feb. 2001: 43 farms infected
- Sep. 2001: 9000 farms infected

- Measures to stop:
  - Banned movement, killed millions of animals

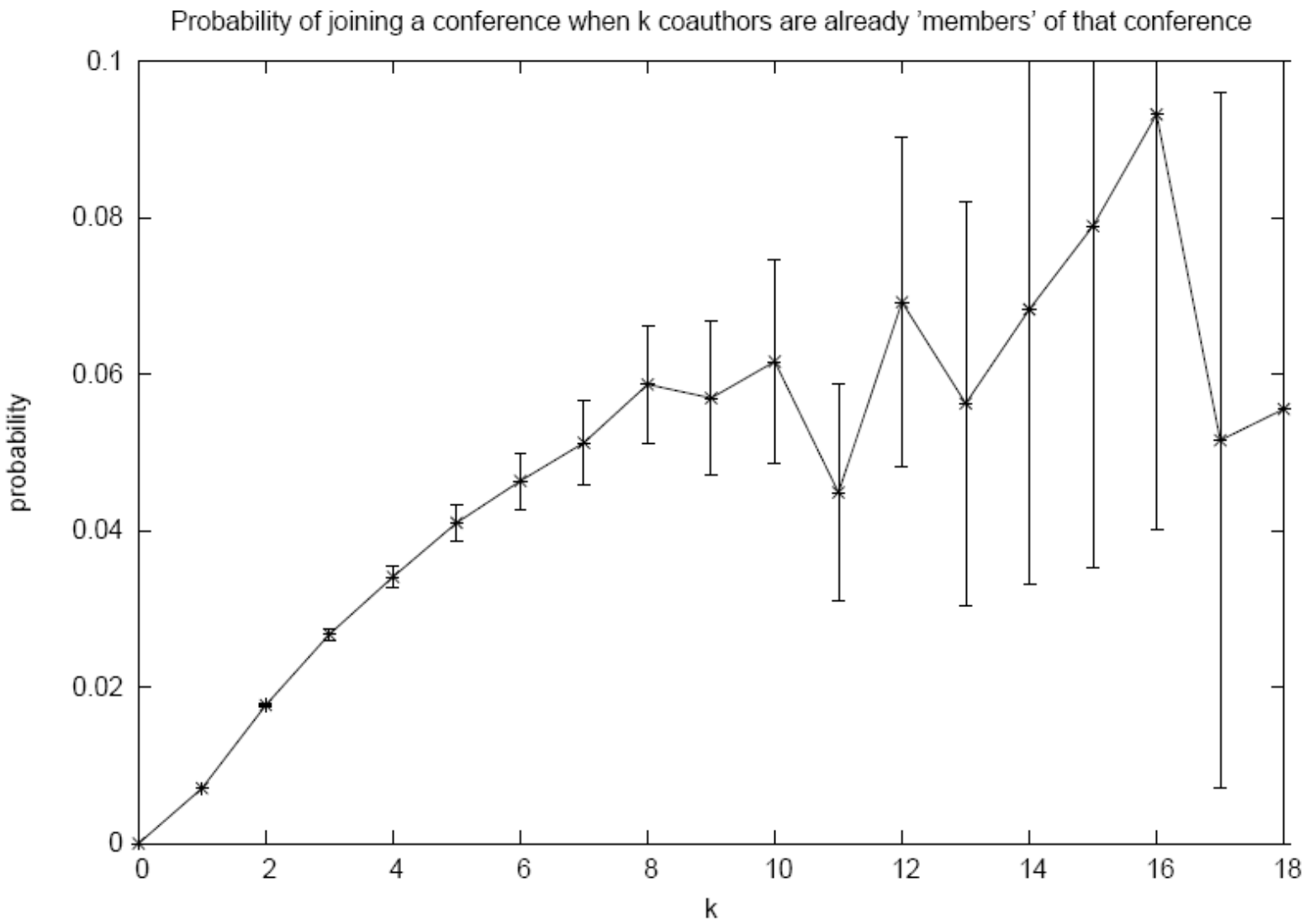
Network Impact

- In the case of the plague it is like moving in a lattice
- In the mad cow we have weak ties, so we have a small world
  - Animals being bought and sold
  - Soil from tourists, etc.
- To protect:
  - Make contagion harder
  - Remove weak ties (e.g., mad cows, HIV)

Example: Join an online group



Example: Publish in a conference



Example: Obesity Study

Christakis and Fowler, "The Spread of Obesity in a Large Social Network over 32 Years", New England Journal of Medicine, 2007.

- Data set of 12,067 people from 1971 to 2003 as part of Framingham Heart Study

- Results
  - Having an obese friend increases chance of obesity by 57%
  - obese sibling ! 40%, obese spouse ! 37%

Models of Influence

Birkbeck
UNIVERSITY OF LONDON

- We saw that often decision is correlated with the number/ fraction of friends
- This suggests that there might be influence:
  - the more the number of friends, the higher the influence
- Models to capture that behavior:
  - Linear threshold model
  - Independent cascade model

Linear Threshold Model

- A node v has threshold $\theta_v \sim U[0,1]$
- A node v is influenced by each neighbor w according to a *weight* b_{vw} such that

$$\sum_{w \text{ neighbor of } v} b_{v,w} \leq 1$$

- A node v becomes active when at least (weighted) θ_v fraction of its neighbors are active

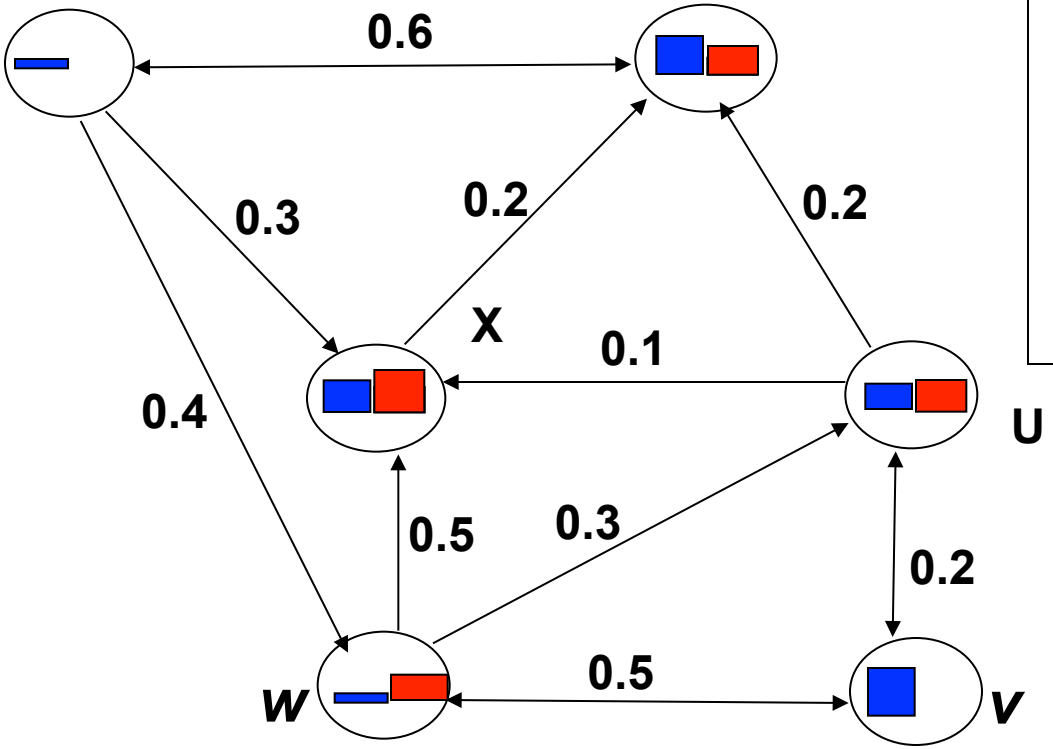
$$\sum_{w \text{ active neighbor of } v} b_{v,w} \geq \theta_v$$

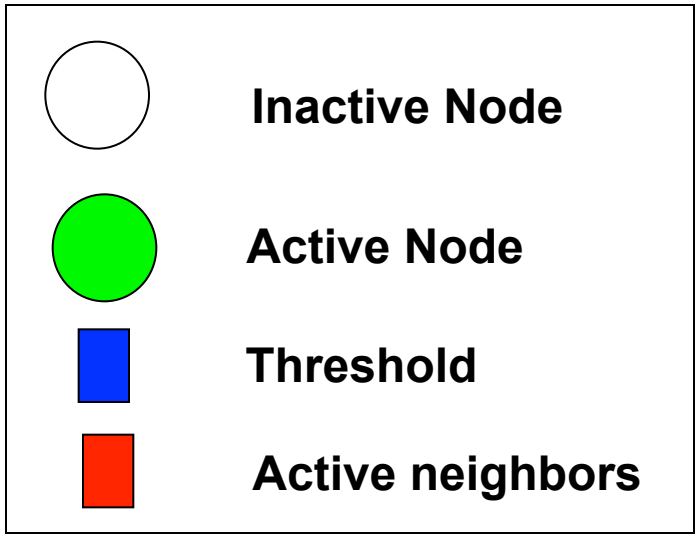
Linear Threshold Model

Birkbeck UNIVERSITY OF LONDON

- Given a choice of thresholds, and an initial set of active nodes (with all other nodes inactive)
- The diffusion process unfolds deterministically in discrete *steps*.
- At each step t:
  - all nodes that were active in step t-1 remain active
  - nodes for which the total weight of their active neighbors is at least θ_v are activated

Example





Stop!

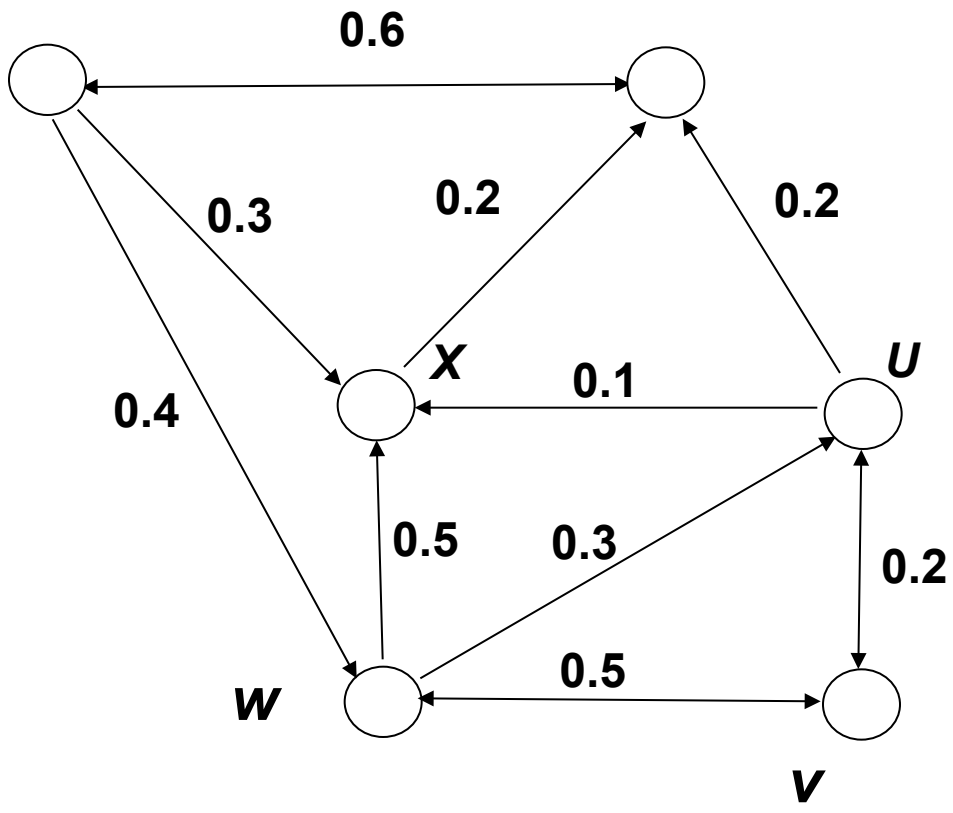
Independent Cascade Model

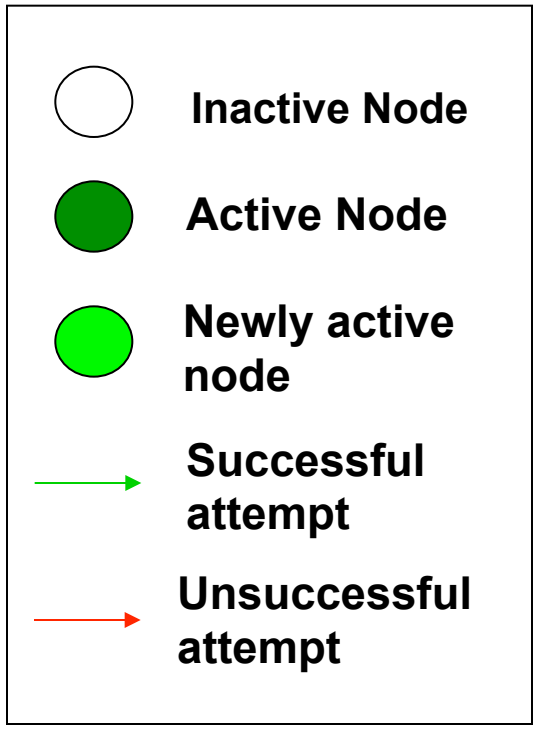
Birkbeck UNIVERSITY OF LONDON

- When node v becomes active, it has a single chance of activating each currently inactive neighbor w

- The activation attempt succeeds with probability p_{vw}

Example





Stop!

Optimization problems

Given a particular model, there are some natural optimization problems:

1. How do I select a set of users to give coupons to in order to maximize the total number of users infected?
2. How do I select a set of people to vaccinate in order to minimize influence/infection?
3. If I have some sensors, where do I place them to detect an epidemic ASAP?

Influence Maximization Problem

Birkbeck UNIVERSITY OF LONDON

- Influence of node set S: f(S)
  - expected number of active nodes at the end, if set S is the initial active set
- Problem:
  - Given a parameter *k* (budget), find a *k*-node set *S* to maximize f(S)
  - Constrained optimization problem with f(S) as the objective function

f(S): properties

- Non-negative (obviously)
- Monotone: $f(S+v) \geq f(S)$
- Submodular:
  - Let N be a finite set
  - A set function $f: 2^N \mapsto \Re$ is submodular *iff*

$$\forall S \subset T \subset N, \forall v \in N \setminus T,$$
$$f(S+v) - f(S) \geq f(T+v) - f(T)$$

Bad News

Birkbeck UNIVERSITY OF LONDON

- For a submodular function *f*:
  - if *f* only takes non-negative values
  - and is monotone
  - Fnding a *k*-element set *S* for which *f(S)* is maximized is an **NP-hard** optimization problem
- Hence:
  - It is NP-hard to determine the optimum for influence maximization for both independent cascade model and linear threshold model

Good News

Birkbeck UNIVERSITY OF LONDON

- We can use a Greedy Algorithm!
  - Start with an empty set S
  - For k iterations:
    - add node v to S that maximizes $f(S+v) - f(S)$

- How good (bad) is this?
  - Theorem: The greedy algorithm is a $(1 - 1/e)$ approximation
  - The resulting set S activates at least

$$(1- 1/e) > 63\%$$

of the number of nodes that any size-k set S could activate

Birkbeck
UNIVERSITY OF LONDON

Written Exam

Material

- Mining frequent itemsets and association rules:
  - What is support and confidence
  - What is the Apriori principle
  - What is the usefulness of closed and maximal itemsets
- Data Representation
  - Curse of dimensionality
  - Singular Value Decomposition and PCA

Material

- Classification
  - Decision trees
  - Perceptron and Support Vector Machines
  - Ensemble learners: boosting, bagging, stacking, randomization, random forests
- Clustering:
  - K-means and K-medoids
  - Hierarchical Clustering
- Evaluation:
  - Accuracy, precision, recall
  - Confusion matrix
  - ROC curve

Material

- Ranking
  - Pagerank and HITS
  - Convergence issues
  - Relation to linear algebra
  - Directed vs. undirected graphs
- Social Networks:
  - Network models
  - Influence maximization algorithms

Exam Structure

- Closed books: no books, notes, calculators! Only pens/ pencils and your brain ☺
- Two parts:
  - Part A (50%): multiple choice – True/False
  - Part B (50%): longer answers
- Total score:
  - Part A + Part B + Quiz Bonus
  - **To PASS, total score >= 60/100**

Exam Structure

- Part A:
  - Confidence 1 or 2
  - If answer correct: + Confidence
  - If answer wrong: – Confidence
  - If no confidence indicated then it is assumed to be 0
  - If Part A < 0 then you will receive no marks for this part. In other words, negative points will not be carried over to Part B.

Exam Structure

- Part B:
  - As concise as possible
  - Short and to-the-point answers are preferable to long answers