Networks I

Modelling Complex Systems

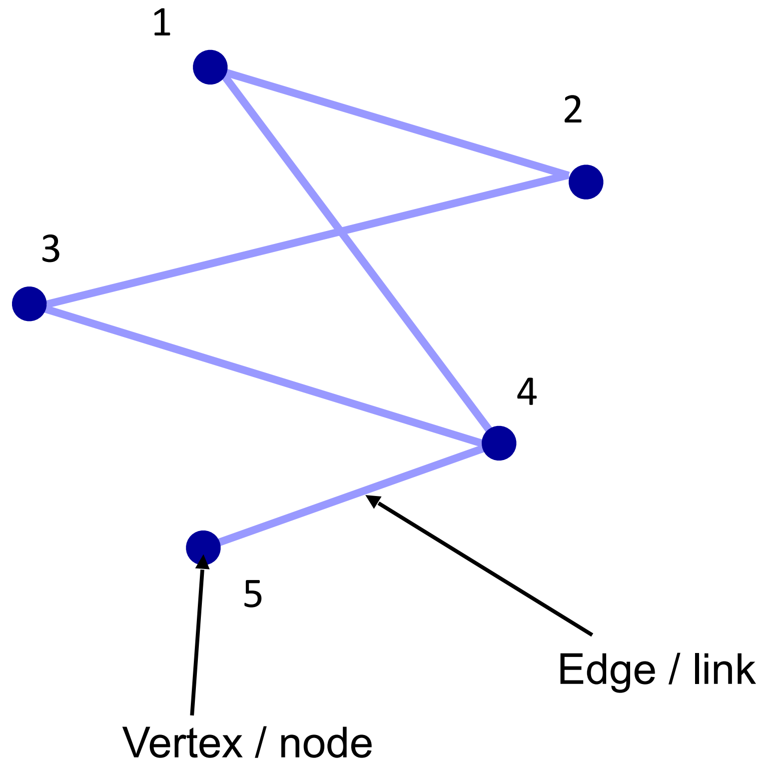
Some of this lecture is adapted from:
Albert and Barabasi, Reviews of Modern Physics 74 (2002)
M. Barthelemy, Physics Reports 499 (2011)
Newman, Networks (2011)
-previous slides of David Sumpter.

Networks

- Things with connections

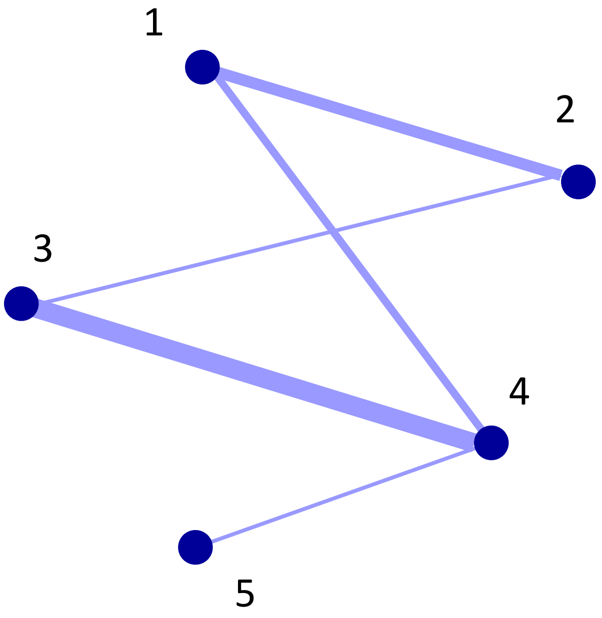
Networks

- Things with connections
- Or, “real life” graphs



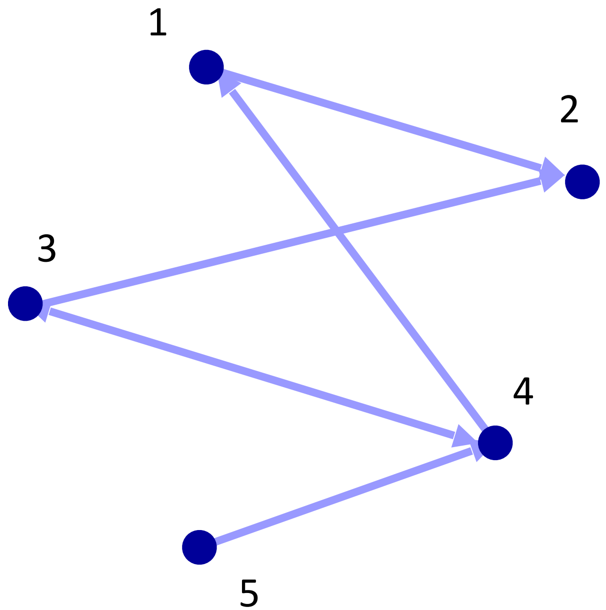
Networks

Can be weighted or unweighted



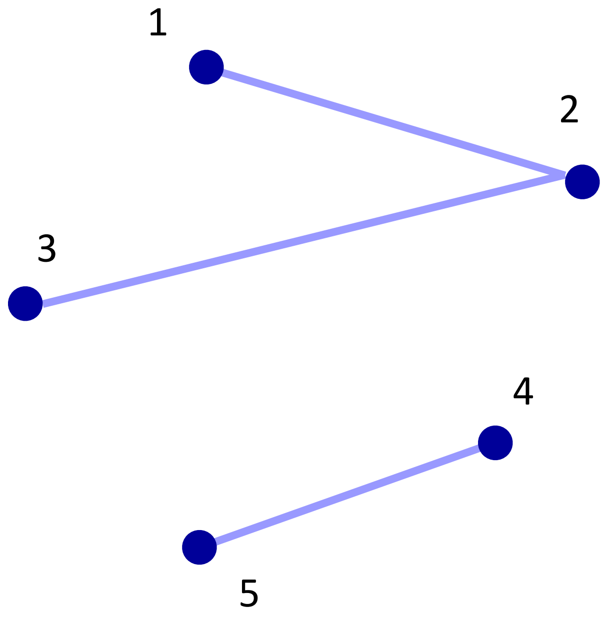
Networks

Can be directed or undirected



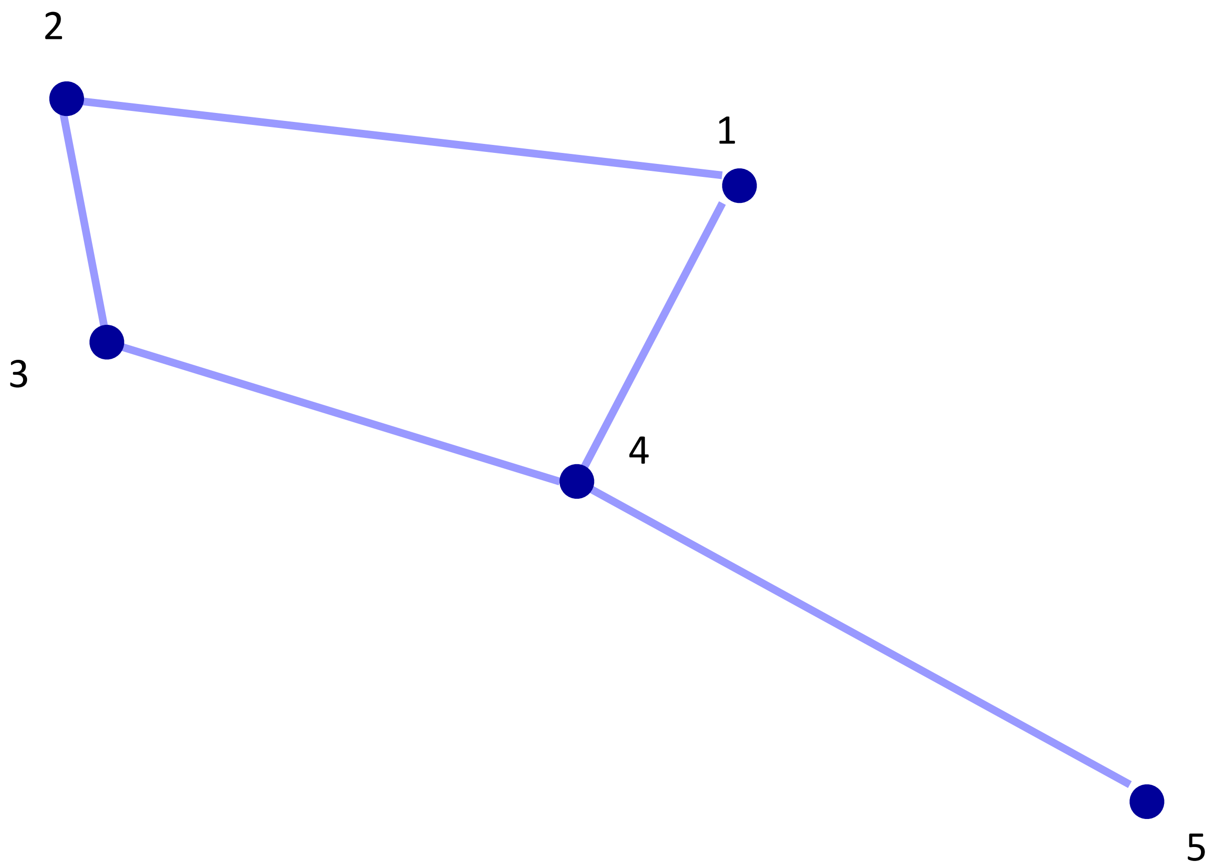
Networks

Can be connected or disjoint



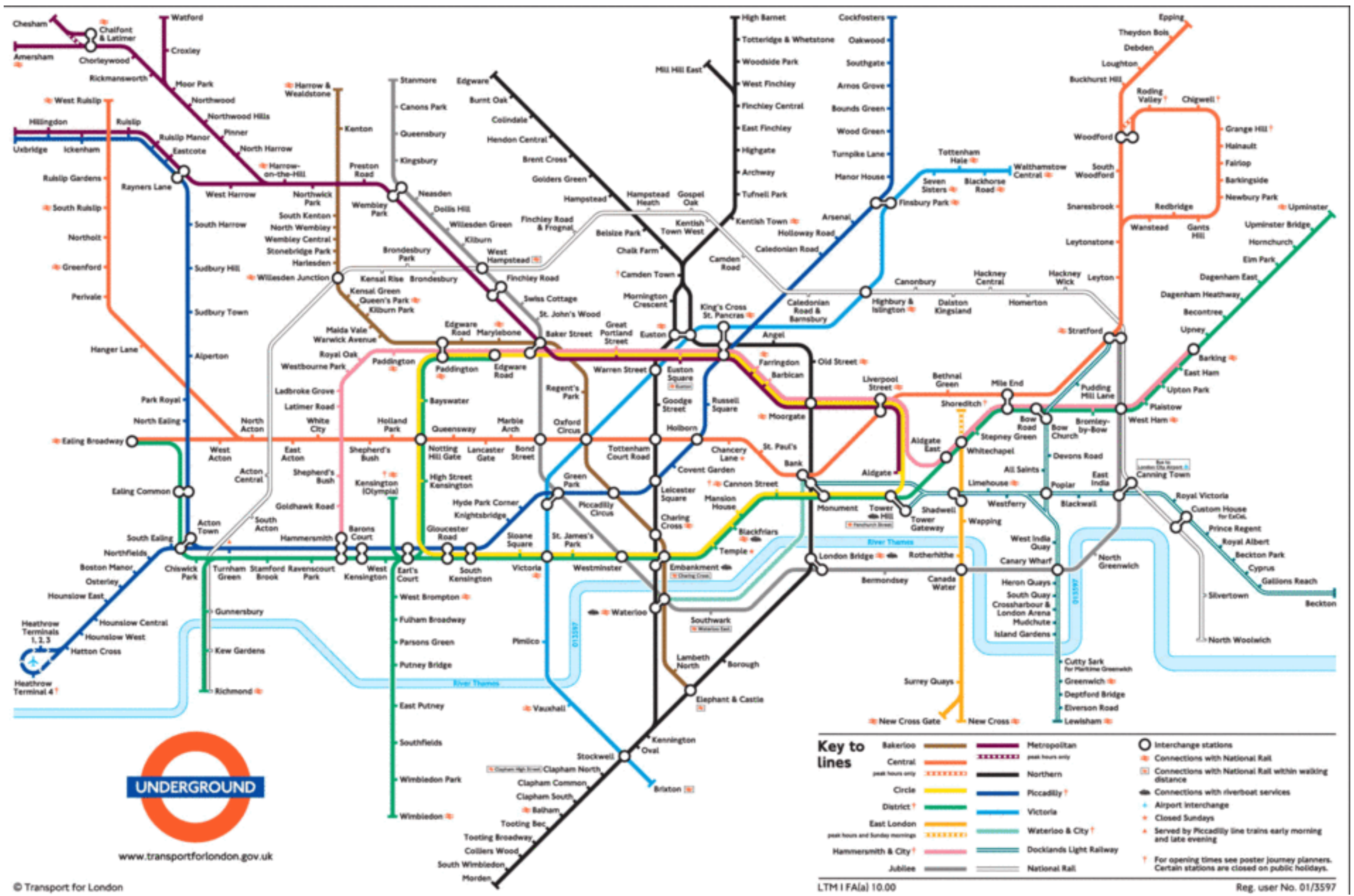
Networks

Can be planar or non-planar



Real-world Networks

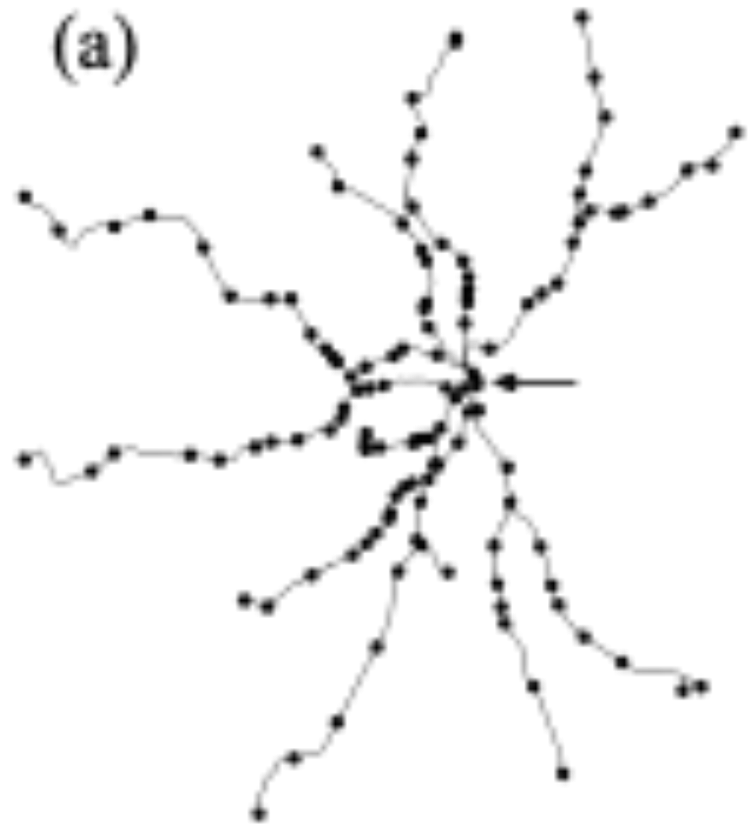
Planned networks

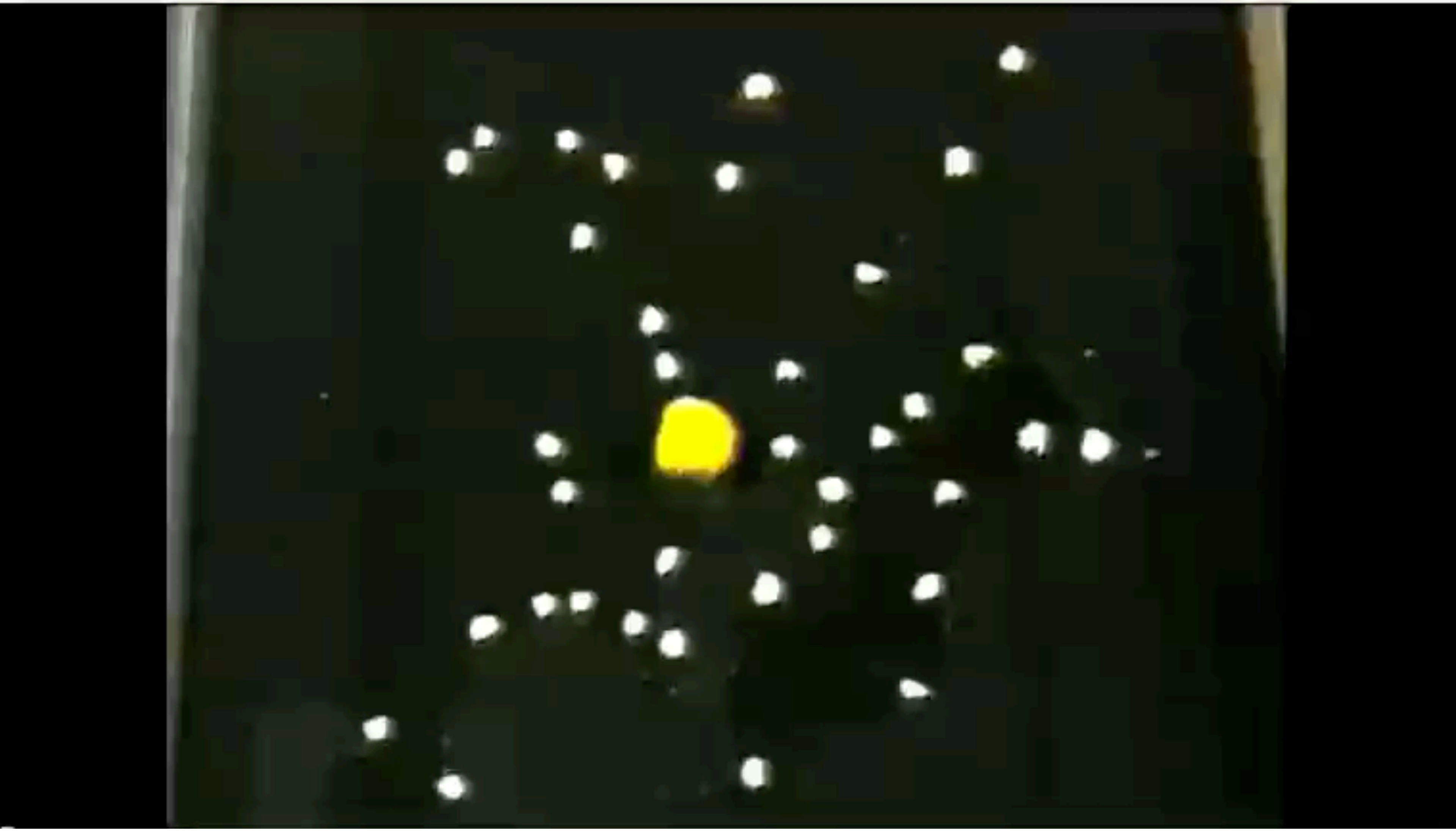


Real-world Networks

Commuter rail network in Boston area.

Physical and planar.





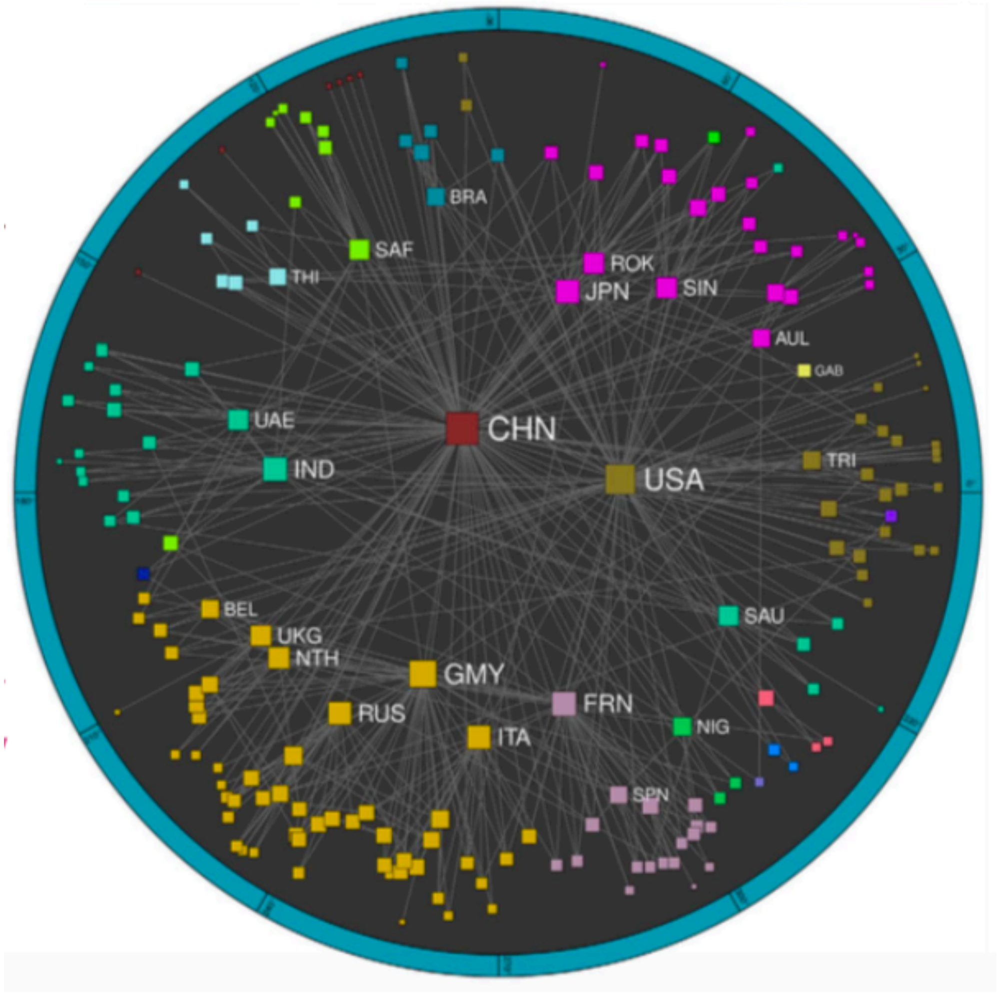
Toshi Nakagaki and co-workers

Real-world Networks

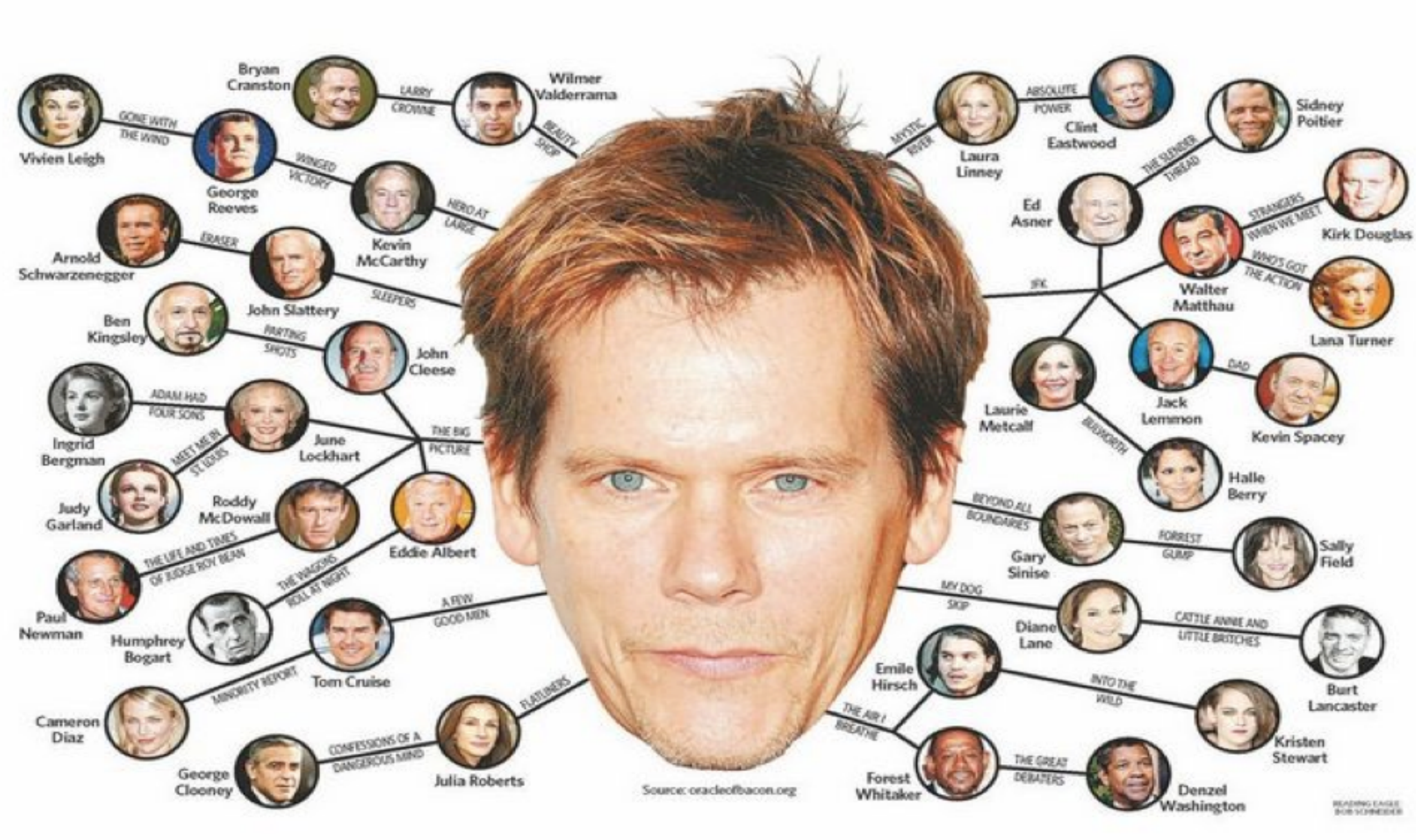
Slime mould

Tokyo Engineers

Real-world Networks



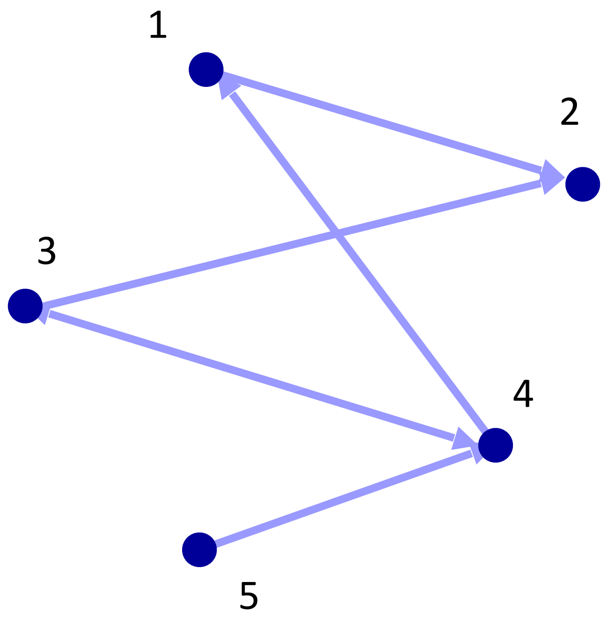
Real-world Networks



Representing Networks

directed

source	destination	weight
1	2	1
4	1	1
3	2	1
3	4	1
4	3	1
5	4	1



n=5 nodes

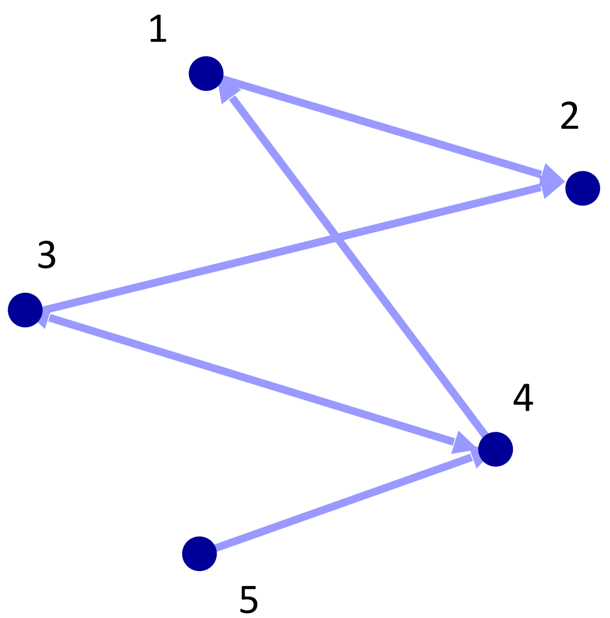
Representing Networks

- Adjacency matrix A_{ij}

directed

Destination

		j=1				j=5
	i=1	0	1	0	0	0
		0	0	0	0	0
Source		0	1	0	1	0
		1	0	1	0	0
	i=5	0	0	0	1	0



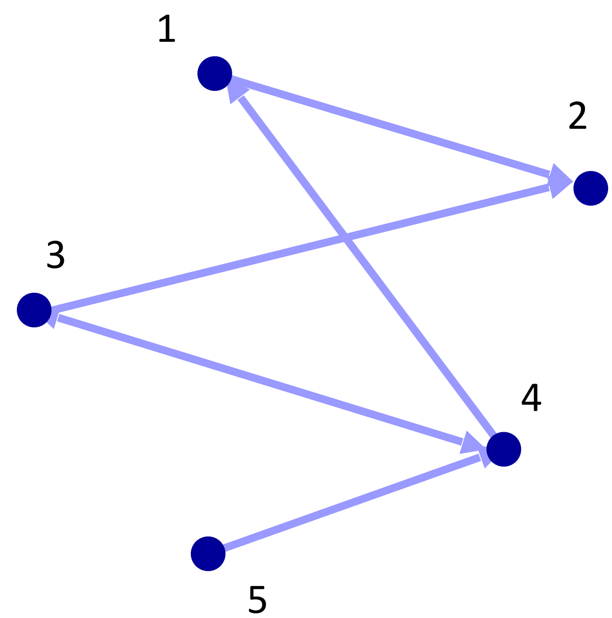
Representing Networks

- Adjacency matrix A_{ij}

Destination

	j=1				j=5	out-degree
i=1	0	1	0	0	0	1
	0	0	0	0	0	0
Source	0	1	0	1	0	2
	1	0	1	0	0	2
i=5	0	0	0	1	0	1
In-degree →	1	2	1	2	0	

directed



Another handy property: $(A^n)_{ij}$ tells us whether you can go from i to j in n steps

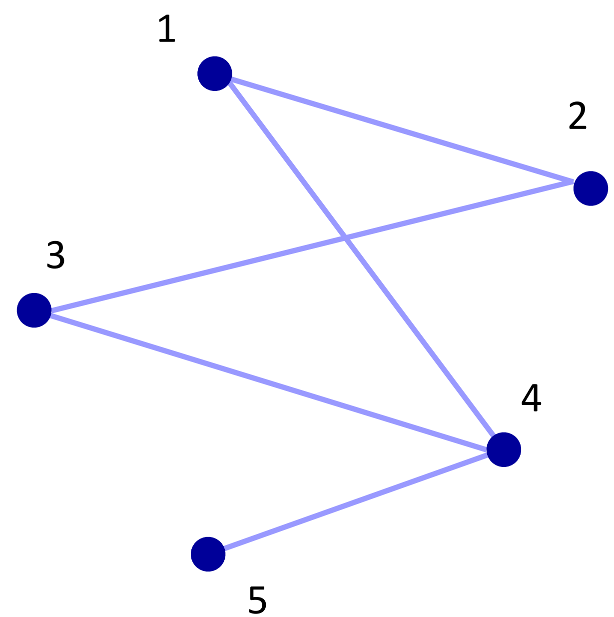
Representing Networks

- Adjacency matrix A_{ij}

Destination

		j=1				j=5	degree
	i=1	0	1	0	1	0	2
		1	0	1	0	0	2
Source		0	1	0	1	0	2
		1	0	1	0	1	3
	i=5	0	0	0	1	0	1
degree →		2	2	2	3	1	

undirected



Another handy property: $(A^n)_{ij}$ tells us whether you can go from i to j in n steps

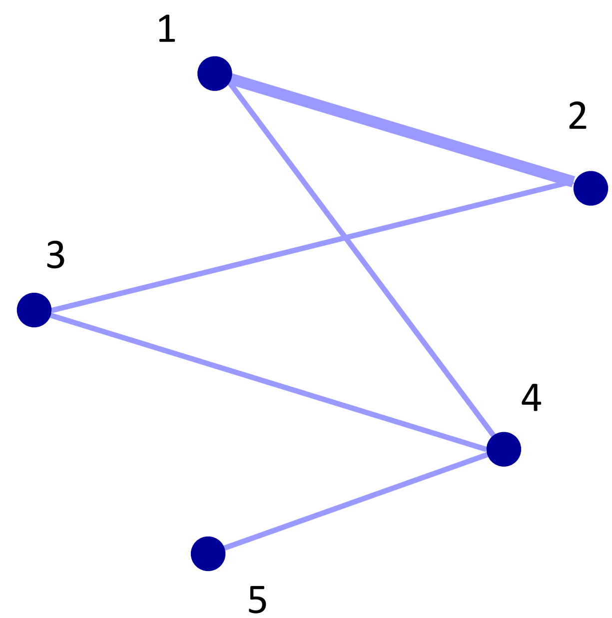
Representing Networks

- Adjacency matrix A_{ij}

Destination

Source	j=1				j=5	degree
i=1	0	1.5	0	1	0	2
	1.5	0	1	0	0	2.5
	0	1	0	1	0	2
	1	0	1	0	1	3
i=5	0	0	0	1	0	1
degree	2	2.5	2	3	1	

weighted



Another handy property: $(A^n)_{ij}$ tells us whether you can go from i to j in n steps

Other networks

- Hypergraph
- Multi-layer Network
- Temporal Network

Five (of many) network measures

- Average degree
- Degree distribution
- Mean path length
- Clustering coefficient
- Maximum modularity/
  Community partitions

…….

Degree and average degree

The in in and out degrees are

$$k_i^{in} = \sum_{j=1} A_{ij} \qquad\qquad k_i^{out} = \sum_{i=1} A_{ij}$$

The average degree is

$$\mathrm{c} = \frac{1}{n}\sum_{i,j} A_{ij}$$

same for in and out degree

Degree distribution

How many people follow you on Twitter.

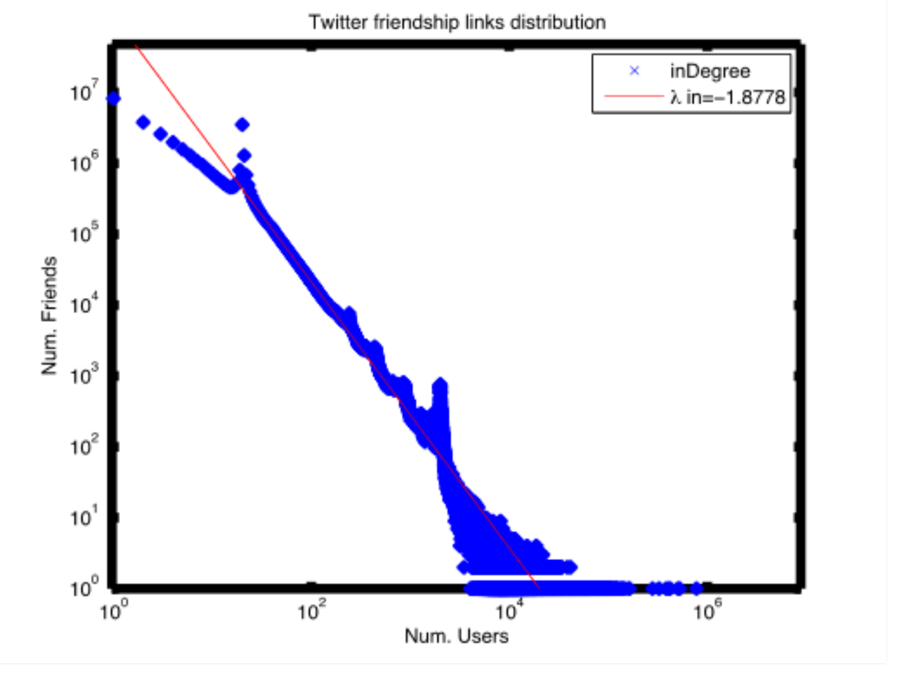


Figure 2. Incoming degree distribution of Twitter's network. As the figure shows, there are a few users with an enormous degree (number of followers). On the contrary, the majority of them have less than 100 followers.

Degree distribution $p(k)$ tells us how the connectedness varies between nodes

Degree distribution

How many people you follow on Twitter.

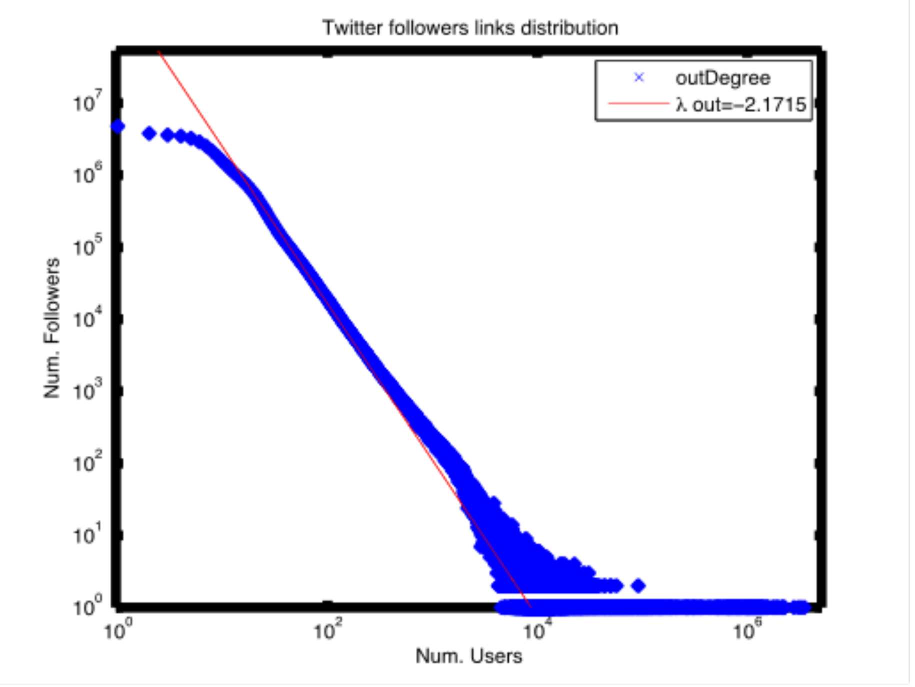


Figure 1. Outgoing degree distribution of Twitter's network. As the figure shows, there are a few users with an enormous degree (number of friends). On the contrary, the majority of them have just at most 1000 friends.

Degree distribution $p(k)$ tells us how the connectedness varies between nodes

Mean path length

- Find shortest path between all pairs i,j
- The mean path length *l* is the mean of each
- Measures degrees of separation

(**Diameter** = longest path length)

Distance between two random individuals

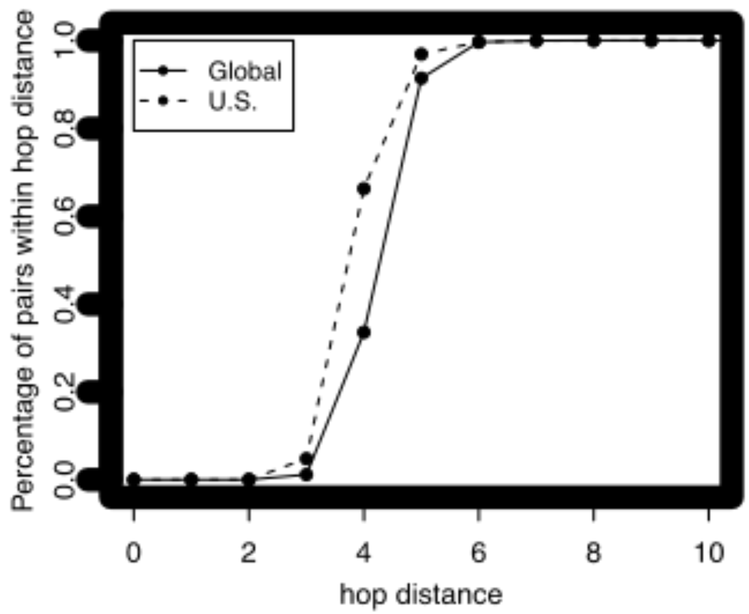
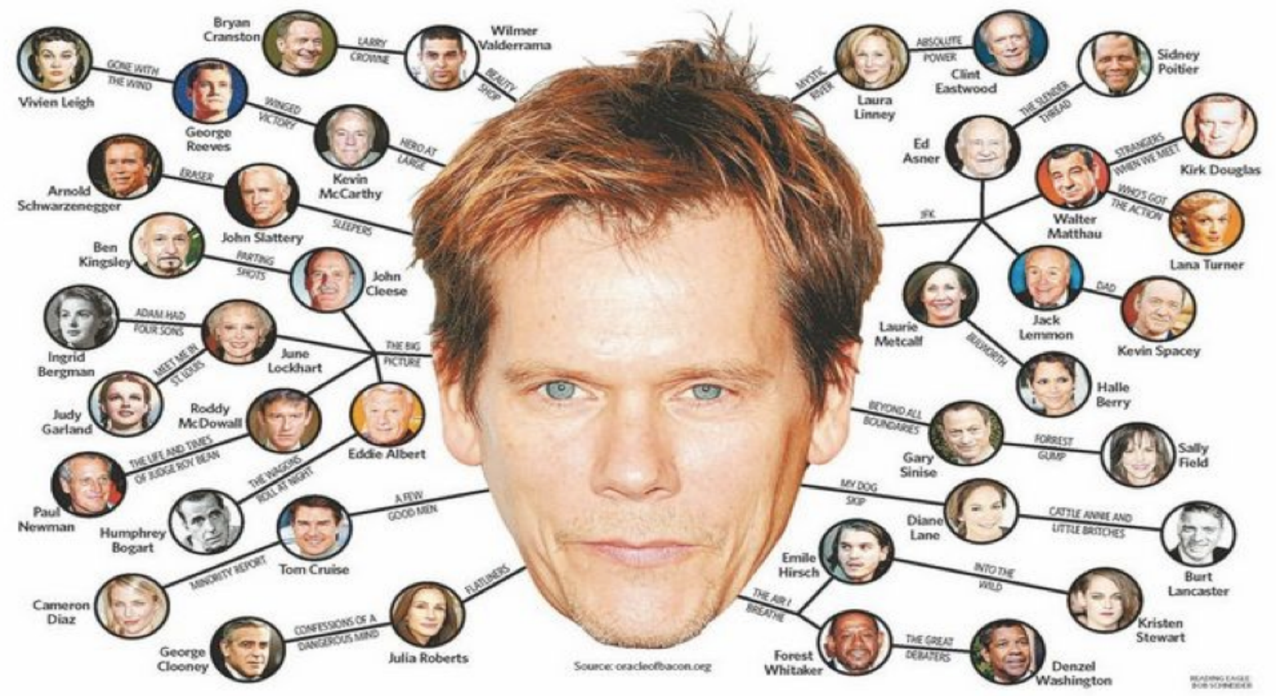
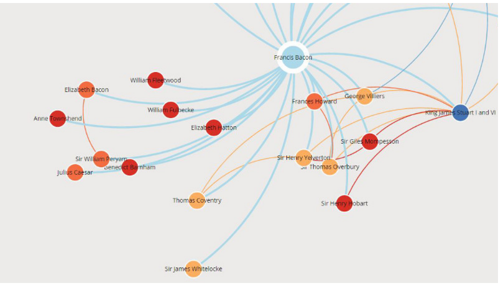


Figure 2. Diameter. The neighborhood function $N(h)$ showing the percentage of user pairs that are within h hops of each other. The average distance between users on Facebook in May 2011 was 4.7, while the average distance within the U.S. at the same time was 4.3.

Mean path length





Networks - community partition

Communities of interest

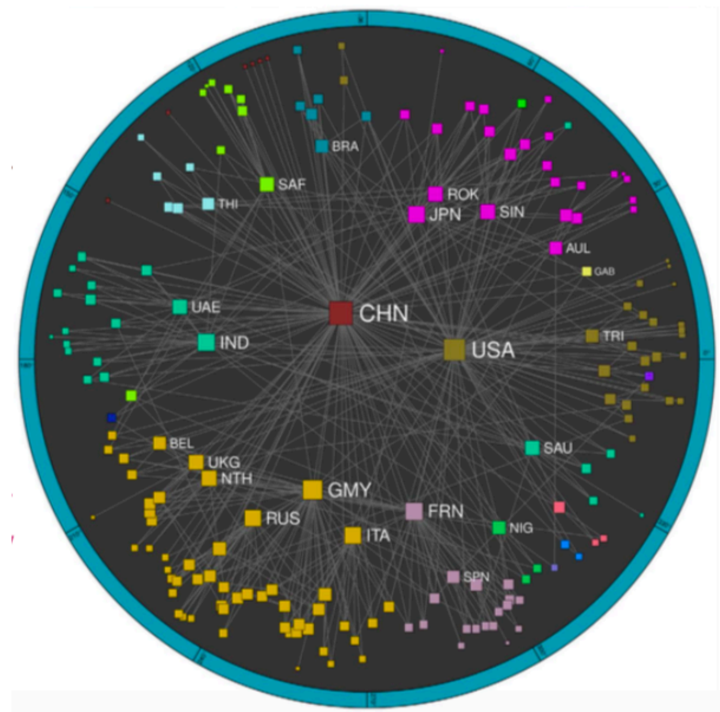
Network: nodes are countries, weight of each link is volume of trade between countries.

Garcia-Pérez 2016

USA, Canada, Bahamas, Haiti, Dominican Republic, Jamaica, Grenada, Mexico, Honduras, Venezuela, Peru

China, North Korea, Gambia, Sierra Leone, Togo, South Sudan

Japan, South Korea, Taiwan, Singapore, Sri Lanka, Philippines, New Zealand, Fiji, Papua New Guinea

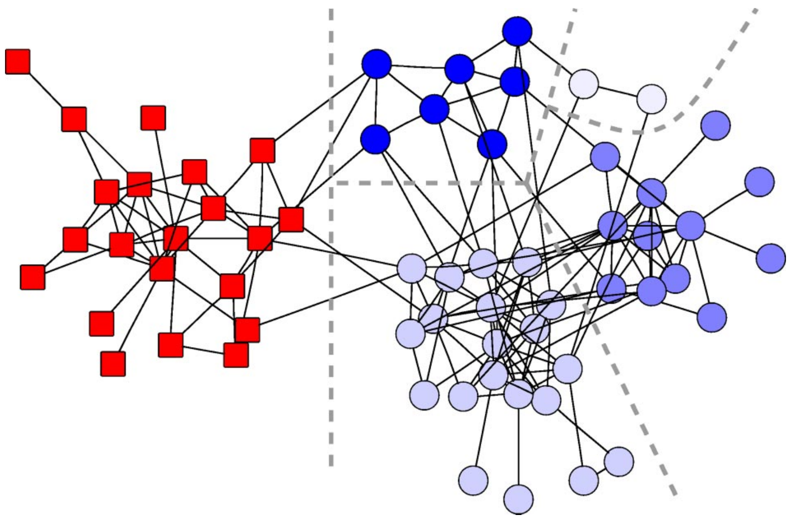


Garcia-Pérez, 2016

Networks - community partition

Communities of interest

Network: dolphins of doubtful sound, NZ, links between dolphins ‘often’ seen together.



Lusseau PhD Thesis,
Newman & Girvan, Finding and evaluating community structure in networks, *Phys Rev E*, 2004

Networks - community partition

As stepping stone: - analyse use of language in climate change debate

Network: links between blogs on climate change

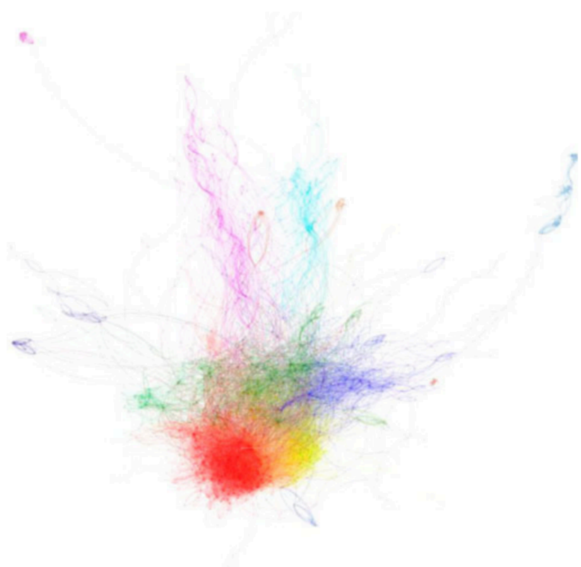


Figure 1. The network of climate change blogs, colored by community.

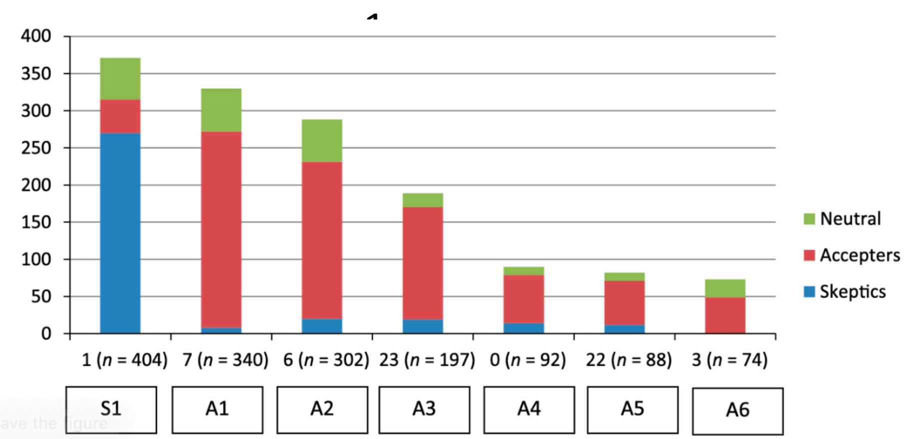


Figure 3. The distribution of skeptical, accepting, and neutral blogs in the seven largest among the central groups of blogs concerned with climate change.

Elgesem, Steskal & Diakopoulos 2015

Networks - community partition

As stepping stone: - analyse use of language in climate change debate

Network: links between blogs on climate change

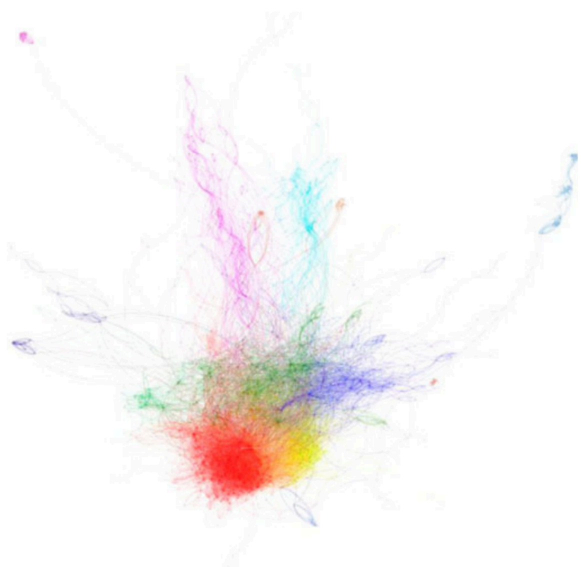


Figure 1. The network of climate change blogs, colored by community.

Table 5. The top 15 collocates around "climate" in communities 1 (skeptic), 23 (accepter), and 7 (accepter) computed with the point-wise mutual information metric.

Top collocates of "CLIMATE" in the skeptical community S1	Top collocates of "CLIMATE" in the accepter community A3	Top collocates of "CLIMATE" in the accepter community A1
1 CLIMATE	1 DENIERS	1 POPPIN
2 SKEPTICS	2 SKEPTICS	2 DENIERS
3 ALARMISM	3 CLIMAT	3 SKEPTICS
4 DENIERS	4 DECADAL	4 OBAMA
5 IPCC	5 CONTRARIANS	5 WWW
6 DECADAL	6 OBAMA	6 EU'S
7 ALARMISTS	7 NOAA'S	7 CLIMATE
8 CLIMAT	8 AGW	8 YVO
9 CHANGE	9 WWW	9 NOAA'S
10 INTERGOVERNMENTAL	10 DENIER	10 WILDFIRES
11 OBAMA	11 CLIMATE	11 CHANGE'S
12 ANTHROPOGENIC	12 VAPOR	12 IPCC
13 AGW	13 ANTHROPOGENIC	13 ALARMISM
14 IPCC'S	14 ALARMISM	14 PACHAURI
15 WARMING	15 CONTRARIAN	15 DENIER

Reference corpus: The British National Corpus, approximately 100 million words.

Elgesem, Steskal & Diakopoulos 2015

Mathematics of community partitions

Define a score! "Modularity" **<- most popular measure, but not universal**

$$q^*(G) = \max_{\mathcal{A}} q_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \frac{e(A)}{m} - \frac{\mathrm{vol}(A)^2}{4m^2}$$

$0 \leq q^*(G) \leq 1$

near 1 - high extent of community structure

near 0 - lack of community structure

Edge contribution/Coverage

$$q^E_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \frac{e(A)}{m}$$

Degree tax

$$q^D_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \frac{\mathrm{vol}(A)^2}{4m^2}$$

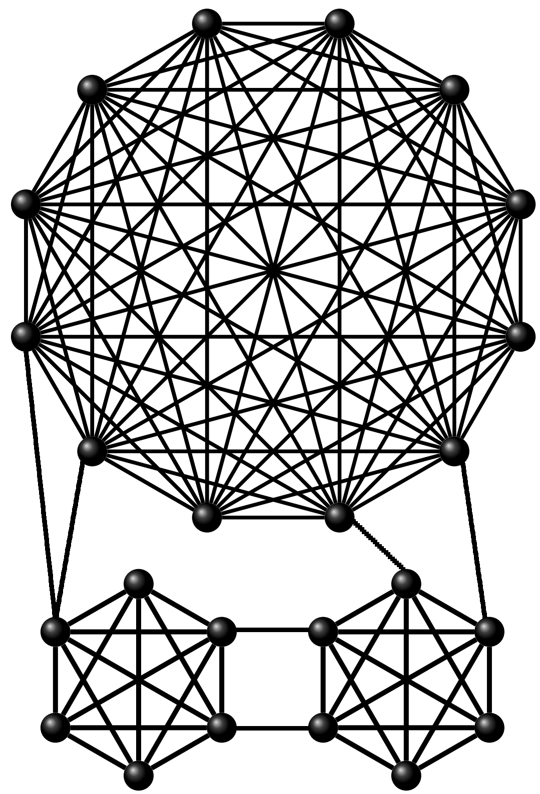
e(A) - number of edges in part/community A

vol(A) - sum of degrees in part/community A

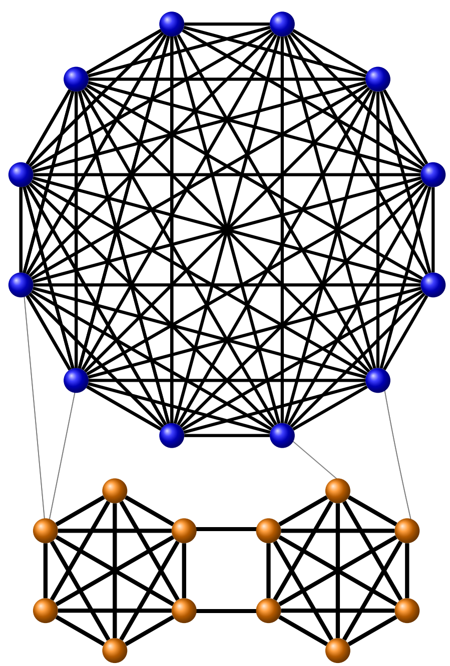
m - total number of edges in the graph

Newman & Girvan, Finding and evaluating community structure in networks, *Phys Rev E*, 2004

Networks - community partition

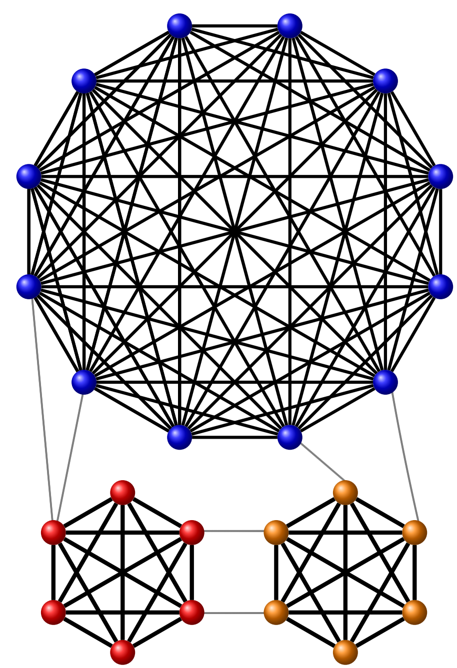


Networks - community partition



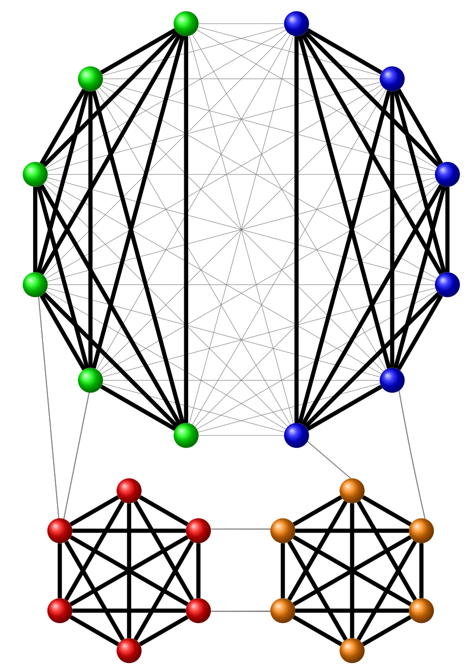
$q^{E}_{\mathcal{A}_1} = 0.96, \quad q^{D}_{\mathcal{A}_1} = 0.56$

$q_{\mathcal{A}_1} = 0.40$



$q^{E}_{\mathcal{A}_2} = 0.94, \quad q^{D}_{\mathcal{A}_2} = 0.50$

$q_{\mathcal{A}_2} = 0.44$



$q^{E}_{\mathcal{A}_3} = 0.59, \quad q^{D}_{\mathcal{A}_3} = 0.29$

$q_{\mathcal{A}_3} = 0.30$

Modelling Networks with (random) graphs

- Lattice graphs
- Erdos-Renyi random graph/Binomial random graph
- Chung-Lu random graph (omitted)
- Configuration model
- Preferential attachment model
- Geometric random graph
- Random hyperbolic graph/KPKVB model

How well does the behaviour of each model replicate that in real networks?

Recap-
Five (of many) network measures

- Average degree
- Degree distribution
- Mean path length
- Clustering coefficient *
- Maximum modularity/
  Community partitions

What values do these take in real networks?

Real networks

TABLE I. The general characteristics of several real networks. For each network we have indicated the number of nodes, the average degree $\langle k \rangle$, the average path length ℓ, and the clustering coefficient C. For a comparison we have included the average path length ℓ_{rand} and clustering coefficient C_{rand} of a random graph of the same size and average degree. The numbers in the last column are keyed to the symbols in Figs. 8 and 9.

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

Degree and average degree

The in in and out degrees are

$$k_i^{in} = \sum_{j=1} A_{ij} \qquad k_i^{out} = \sum_{i=1} A_{ij}$$

The average degree is

$$\mathrm{c} = \frac{1}{n} \sum_{i,j} A_{ij}$$

same for in and out degree

Degree distribution

How many people follow you on Twitter.

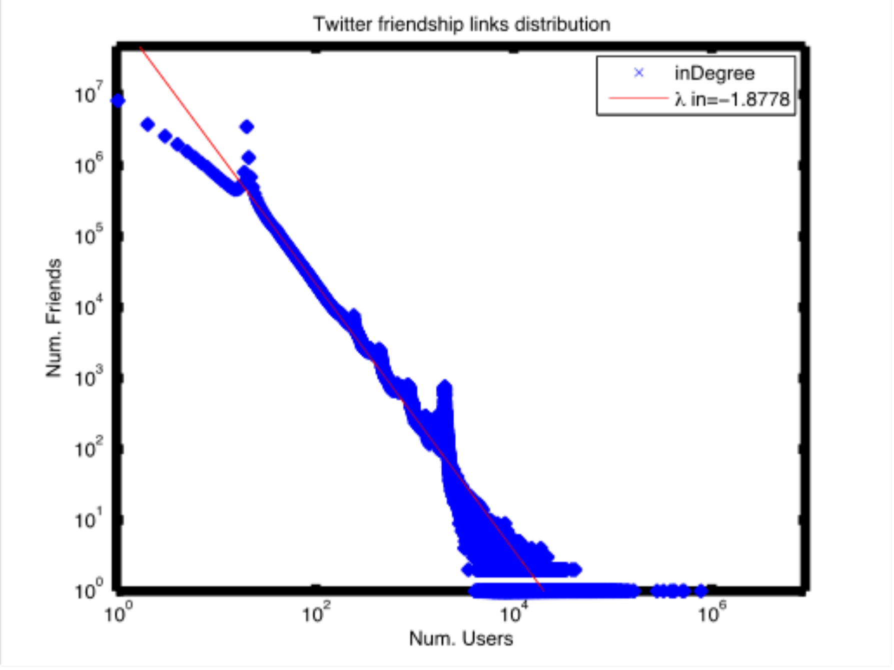


Figure 2. Incoming degree distribution of Twitter's network. As the figure shows, there are a few users with an enormous degree (number of followers). On the contrary, the majority of them have less than 100 followers.

Degree distribution $p(k)$ tells us how the connectedness varies between nodes

Degree distribution

How many people you follow on Twitter.

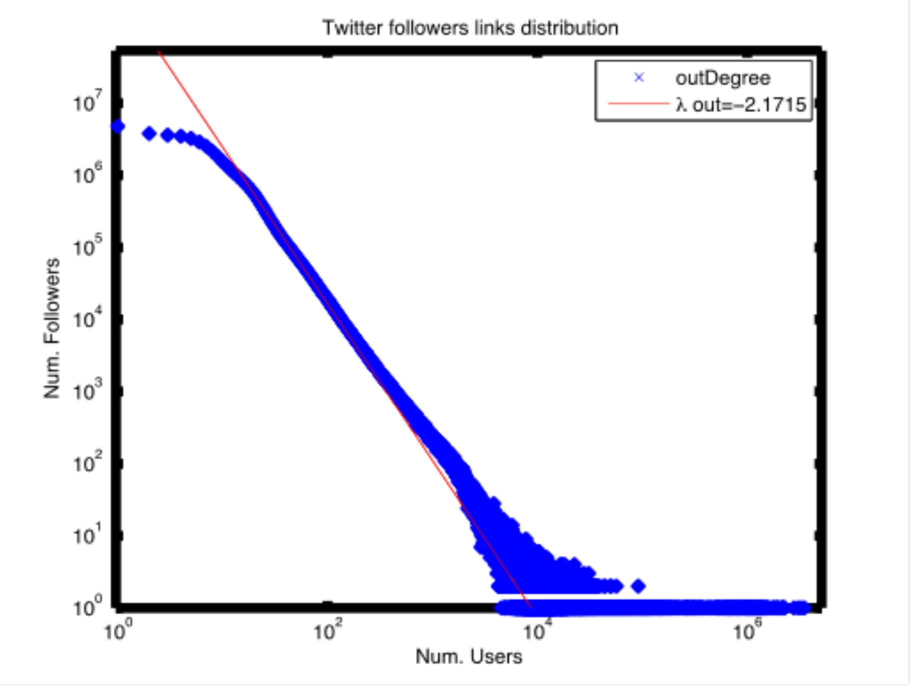


Figure 1. Outgoing degree distribution of Twitter's network. As the figure shows, there are a few users with an enormous degree (number of friends). On the contrary, the majority of them have just at most 1000 friends.

Degree distribution $p(k)$ tells us how the connectedness varies between nodes

Degree distribution

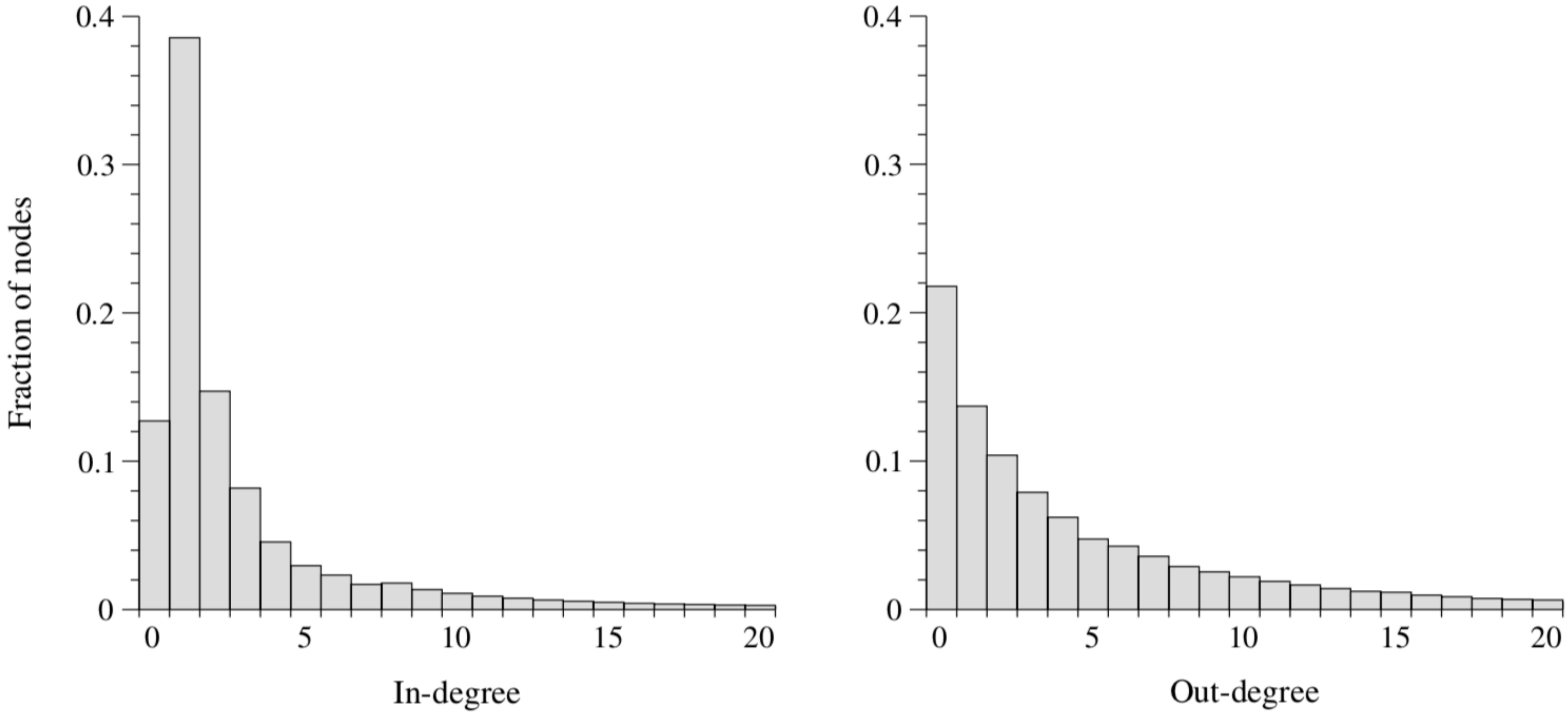


Figure 10.4: The degree distributions of the World Wide Web. Histograms of the distributions of in- and out-degrees of pages on the World Wide Web. Data are from the study by Broder *et al.* [84].

Degree distribution power law - $p(k)$ = k^{-\alpha}

Newman 'Networks' 2018

Modelling Networks with (random) graphs

- Lattice graphs
- Erdos-Renyi random graph/Binomial random graph
- Chung-Lu random graph
- Configuration model
- Preferential attachment model
- Geometric random graph
- Random hyperbolic graph/KPKVB model

How well does the behaviour of each model replicate that in real networks?

Lattice networks

- All internal nodes have the same degree
- High C   (~ constant)
- High mean path length (increases as $n^{1/d}$)

Erdös-Rényi Random graph

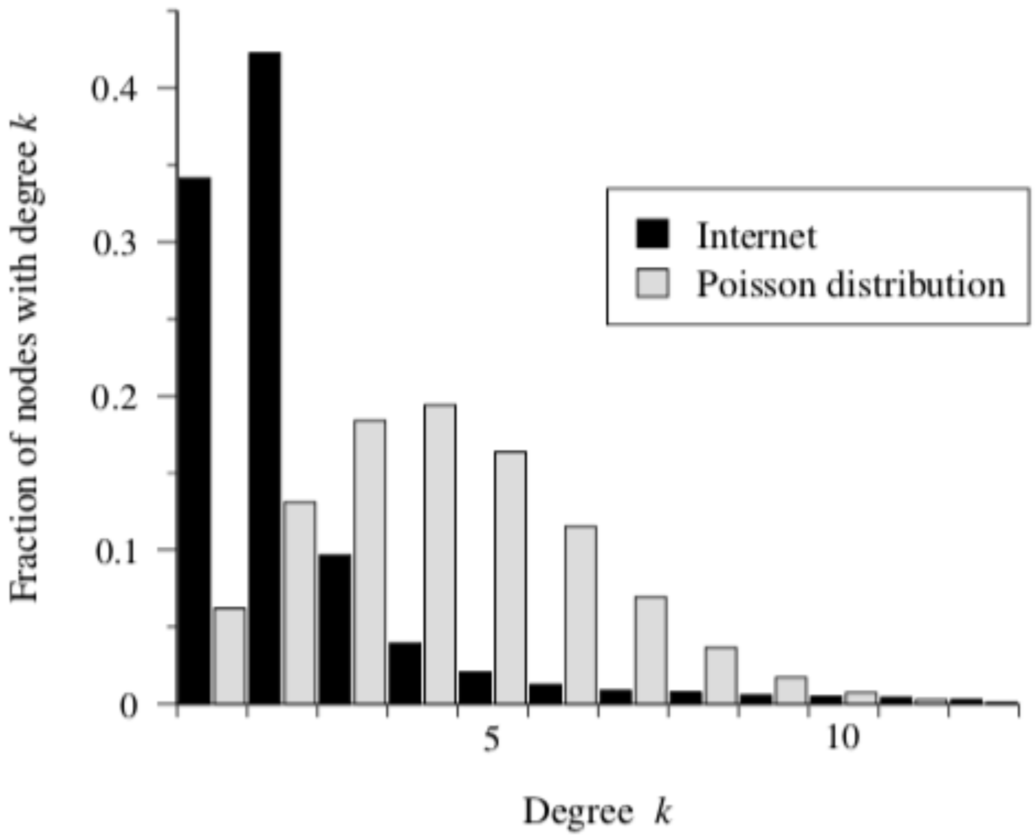
Every pair of nodes i,j is connected with probability p*. Total of n* nodes

- Binomial degree distribution, $c = p(n\text{-}1)$
- Low C $= c/n$
- Low mean path length l ~ log(n)

Random graph process
Start with n vertices with 0 edges. Each step add a missing edge.
(Video)

Erdös-Rényi Random graph

- Degree distribution ….



Erdös-Rényi Random graph

- Not a realistic model but good toy model

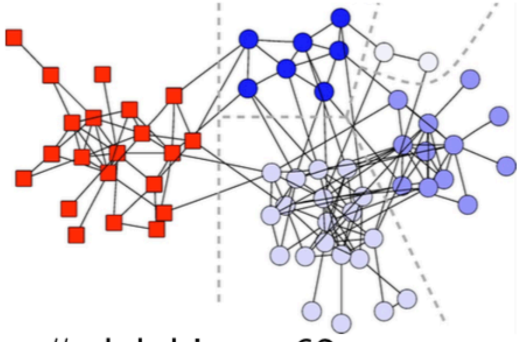
- Serves as a null model

A differentiation between graphs which are truly modular and those which are not can ... only be made if we gain an understanding of the intrinsic modularity of random graphs.

-- Reichardt and Bornholdt

Erdös-Rényi Random graph

- Serves as a null model

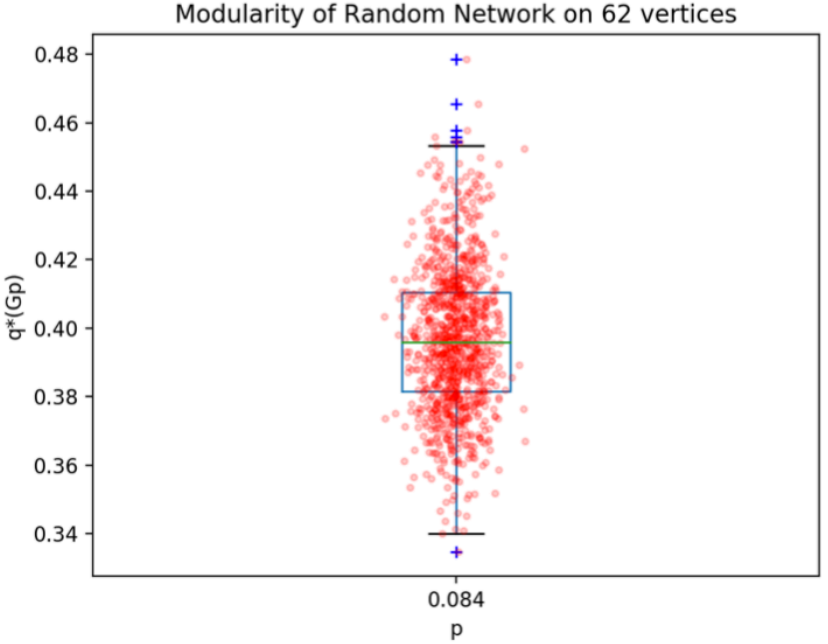


$\#$ dolphins $=62$
$\#$ edges $=159$

$q^* = 0.52$

8.4% of possible edges

$q^*(\text{dolphins}) > q^*(\text{random network})??$



Configuration Model

Start with degree sequence d_1, … d_n
Place d_i half edges on each node
Choose a random matching of half edges

Serves as a null model.

Can choose degree sequence.
Low clustering coefficient (-> 0 as network size increases)

Preferential Attachment Model

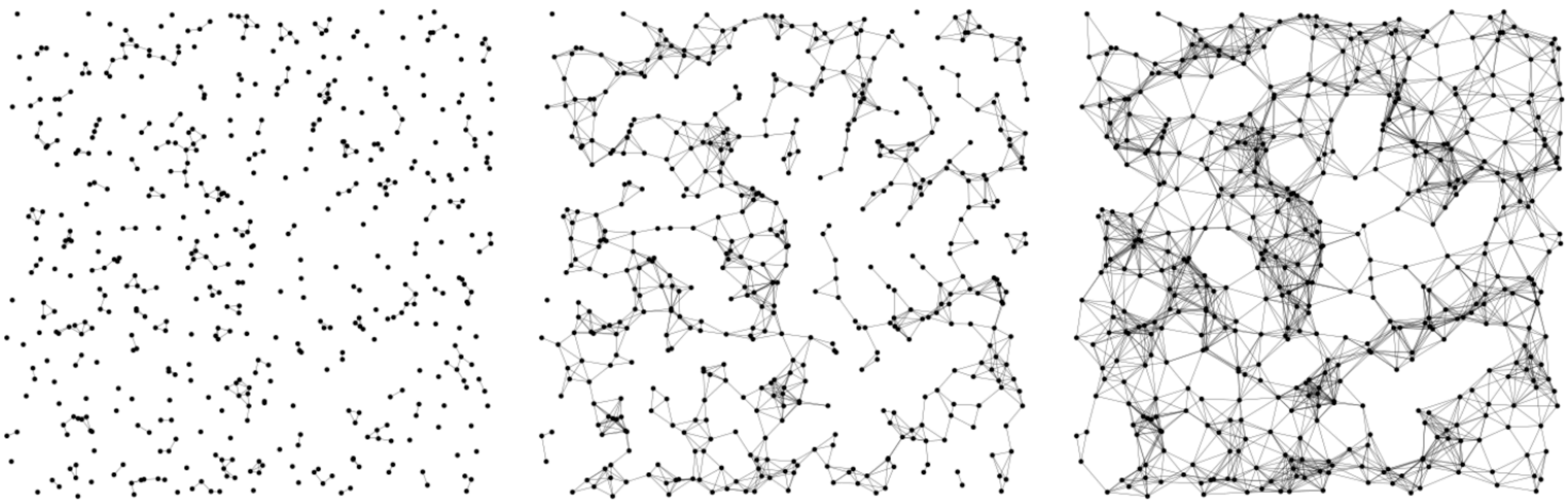
- Animation - https://www.youtube.com/watch?v=4GDqJVtPEGg
- Start with a single edge, or a node with a ‘half-edge’.
- Step i,
  - add vertex v_i
  - pick a previously present vertex v_j with probability proportional to deg(v_j).
  - Add edge v_i ~ v_j

Modifications: add v_i to ‘m’ vertices each step, make probability proportional to deg(v_j)^c, for some constant c.

- Varying c: https://www.youtube.com/channel/UC-P96HKdvFs0Sy4Lp76THlA

Random Geometric Graph

- Place n points uniformly. Join any two vertices with distance less than r.



500 points. r=0.03, r=0.06, r=0.09

Tobias Muller

- # KPKVB model - random hyperbolic graph

  - Hyperbolic plane curvature -alpha^2

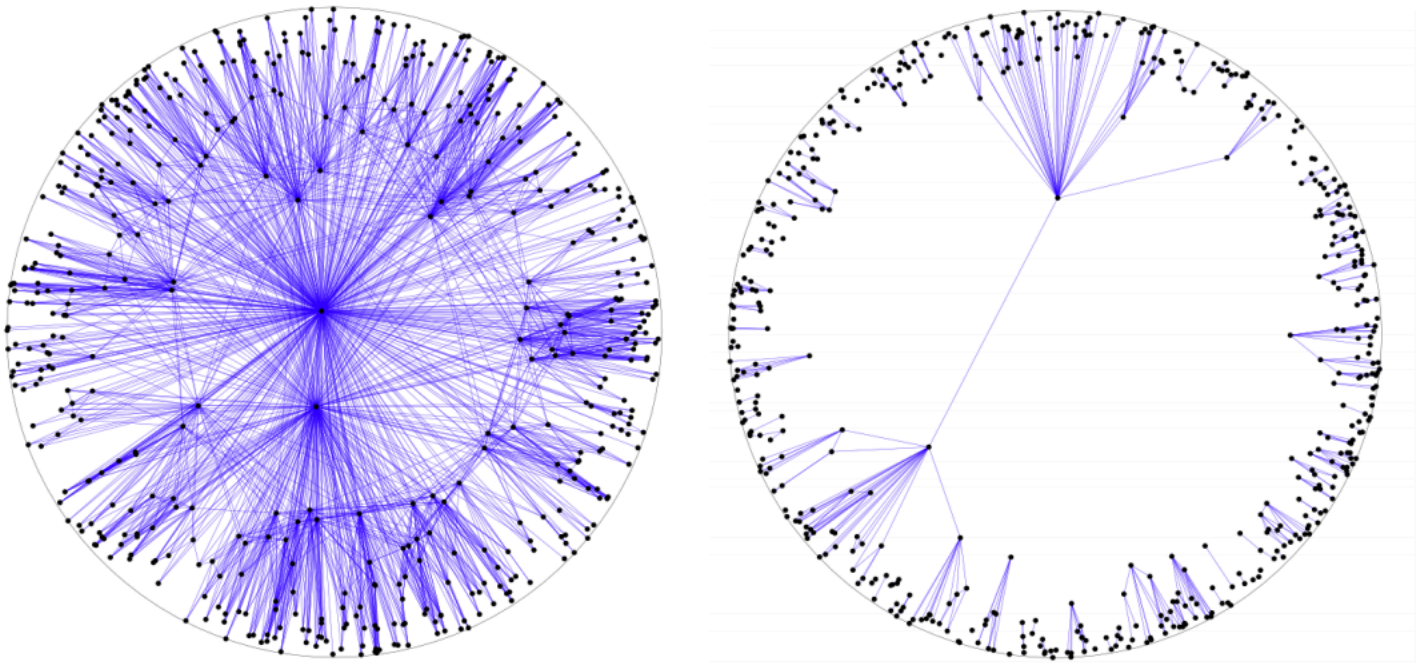


Figure 1: The random graph $G(N;\alpha,\nu)$ with $N = 500$ vertices, $\nu = 2$ and $\alpha = 0.7$ and $3/2$.

Müller and Fountoulakis, Law of large numbers for the largest component in a hyperbolic model of complex networks, 2018

Real networks

TABLE I. The general characteristics of several real networks. For each network we have indicated the number of nodes, the average degree $\langle k \rangle$, the average path length ℓ, and the clustering coefficient C. For a comparison we have included the average path length ℓ_{rand} and clustering coefficient C_{rand} of a random graph of the same size and average degree. The numbers in the last column are keyed to the symbols in Figs. 8 and 9.

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

- KPKVB model - random hyperbolic graph
  - Krioukov-Papadopoulos-Kitsak-Vahdat-Boguñá
  - Power law degree distribution
  - Clustering coefficient
  - Hard to prove results in this model

Müller and Fountoulakis, Law of large numbers for the largest component in a hyperbolic model of complex networks, 2018

Networks II

Modelling Complex Systems

Some of this lecture is adapted from:
Albert and Barabasi, Reviews of Modern Physics 74 (2002)
M. Barthelemy, Physics Reports 499 (2011)
Newman, Networks (2018) - ebook available Uppsala University Library
-previous slides of David Sumpter.

Modelling Networks with (random) graphs

- Lattice graphs
- Erdos-Renyi random graph/Binomial random graph
- Chung-Lu random graph
- Configuration model
- Preferential attachment model
- Geometric random graph
- Random hyperbolic graph/KPKVB model
- Small world network*

How well does the behaviour of each model replicate that in real networks?

Recap-
Five (of many) network measures

- Average degree
- Degree distribution
- Mean path length
- Clustering coefficient *
- Maximum modularity/
  Community partitions

What values do these take in real networks?

Real networks

TABLE I. The general characteristics of several real networks. For each network we have indicated the number of nodes, the average degree $\langle k \rangle$, the average path length ℓ, and the clustering coefficient C. For a comparison we have included the average path length ℓ_{rand} and clustering coefficient C_{rand} of a random graph of the same size and average degree. The numbers in the last column are keyed to the symbols in Figs. 8 and 9.

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

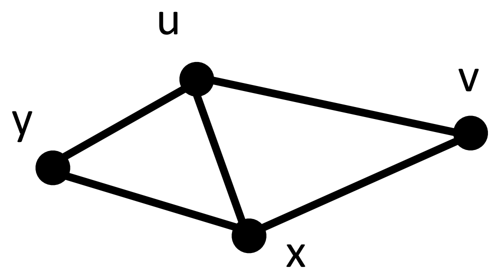
(Global) Clustering Coefficient

Measures
- probability that a randomly chosen two path forms a triangle
- high in social networks: you are friends with your friends friends.

$$c(G) = \frac{\sum_{v \in V} N_3(v)}{\sum_{v \in V} N_2(v)}$$

$N_3(v) =$ #unlabelled triangles having vertex v

$N_2(v) =$ #unlabelled 2-stars having central vertex v



Graph has 2 triangles -> numerator is 6
2-stars: $N_2(u) = N_2(x) = 3$, $N_2(y) = N_2(v) = 1$

$\therefore \quad c = 6/8 = 3/4$

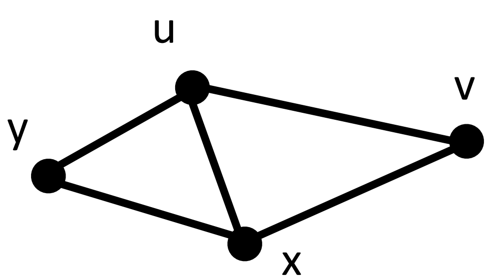
(Global vs. Local) Clustering Coefficient

Local clustering coefficient, $c_L(G)$, also studied, compare:

$$c(G) = \frac{\sum_{v\in V} N_3(v)}{\sum_{v\in V} N_2(v)} \qquad c_L(G) = \frac{1}{|V|}\sum_{v\in V}\frac{N_3(v)}{N_2(v)}$$

$N_3(v) =$ #unlabelled triangles having vertex v

$N_2(v) =$ #unlabelled 2-stars having central vertex v



Graph has 2 triangles
-> $N_3(u) = N_3(x) = 2,\ N_3(y) = N_3(v) = 1$

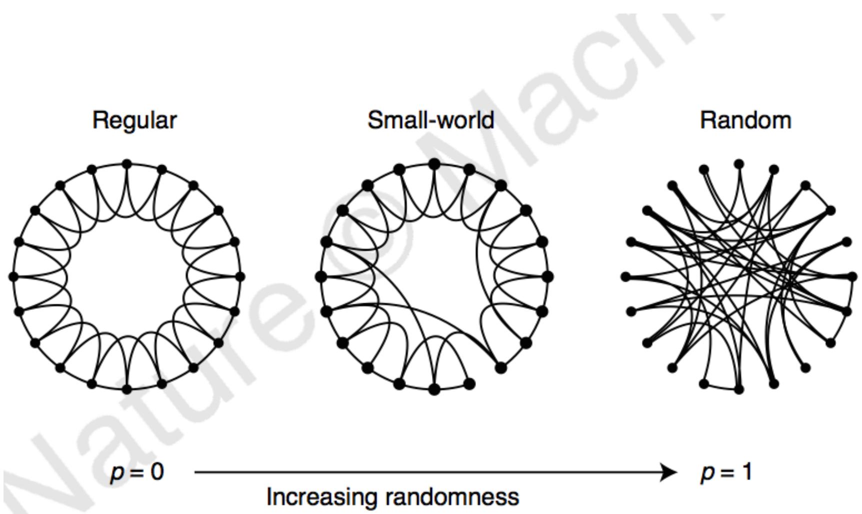
2-stars: $N_2(u) = N_2(x) = 3,\ N_2(y) = N_2(v) = 1$

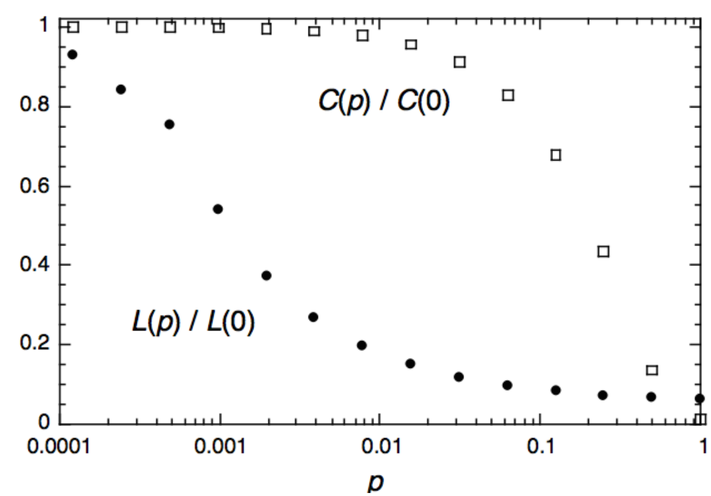
$\therefore\ c_L = 1/4(2/3 + 2/3 + 1 + 1) = 5/6$

Small world network

- Watts & Strogatz model interpolates between a structured and random network

- Low diameter + high clustering = small world



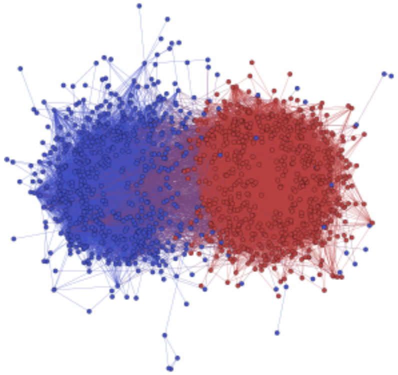


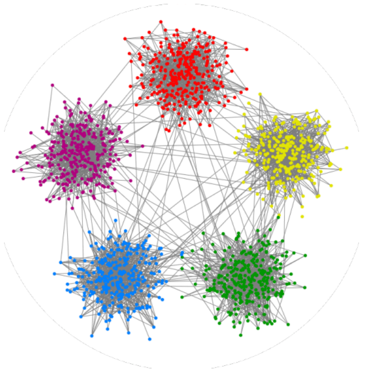
Watts and Strogatz, *Nature* 393 (1998)

Stochastic Block model

- Generates random graphs with "planted communities". Also called planted partition model

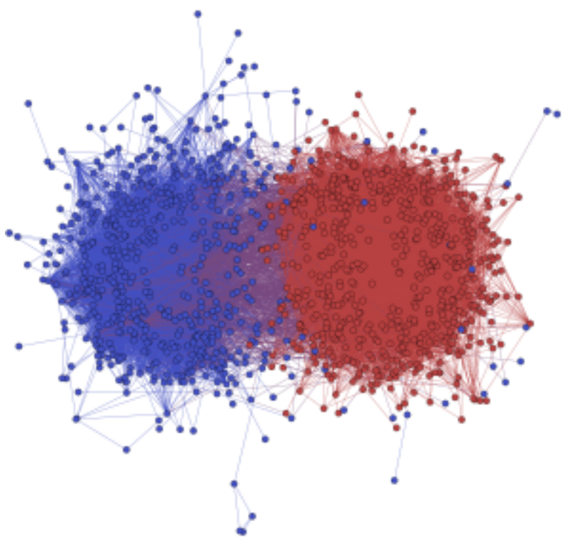
- (For two communities):
- Parameters n, k=2, p, q. (p>q)
  Start with n nodes.
  For each node colour **red** prob. 1/2, otherwise **blue**
- For each pair of vertices uv: if monochromatic join with probability p, otherwise with probability q.



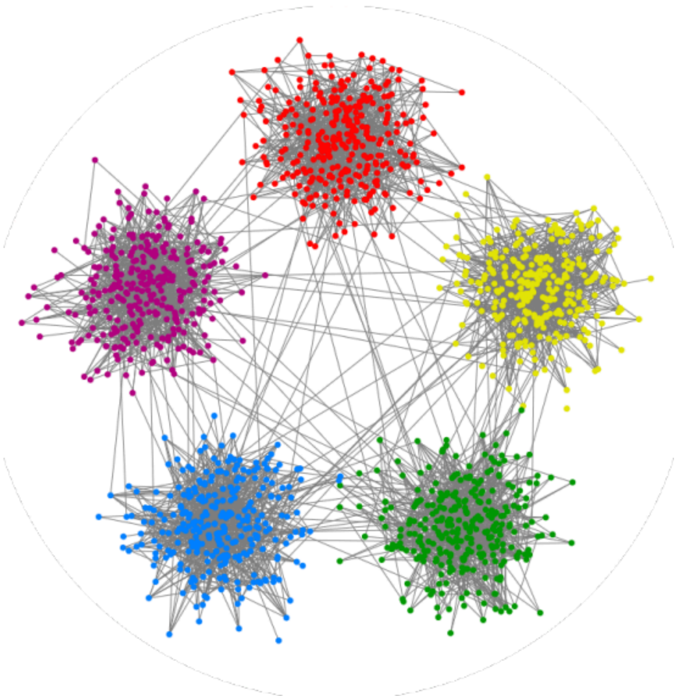


Stochastic Block model

- Generates random graphs with “planted communities”. Also called planted partition model
-



Political blog, US 2004 election
~Adam Glance 2006



SBM (n, k, p, q):
(1000, 5, 0.02, 0.001)

Aside: Distinguishing graph models

You are told you have a random graph from model A or model B, each probability 1/2. Can you say which model (with good likelihood).

(e.g. Erdos-Renyi/binomial random graph distribution, or Stochastic block model)

Active area of research!

Questions

- for what parameter values can you distinguish?
- what test statistics on networks distinguish?
- what algorithms can distinguish? (Fast?)

Monte-Carlo tests

Q: Is the test statistic on our network, t^*, expected if network is drawn from the null distribution.

e.g. modularity of network - compare network to configuration model same degree

- useful when we don’t know the distribution
- need to be able to sample from the null distribution
- discrete data has ties (break randomly and method still valid).

Monte-Carlo tests

Q: Is the test statistic on our network, t^*, expected if network is drawn from the null distribution.

e.g. modularity of network - compare network to configuration model same degree

Method for $\alpha = m/(n+1)$

- sample n from the null distribution and calculate test statistic $t_1, \ldots t_n,$

e.g. sample n configuration models and calculate modularity score of each,

- order $t^*, t_1, \ldots, t_n$
- if t^* among top m values reject null hypothesis for distribution of network
- rule of thumb, take m at least 5.

Monte-Carlo tests

If null hypothesis true all orderings of data are equally likely, the probability that the one you observe is among the top m is $m/(n+1)$

Also called 'parametric bootstrapping', 'conditional uniform graph test'

Method for $\alpha = m/(n+1)$

- sample n from the null distribution and calculate test statistic $t_1, \dots t_n,$

e.g. sample n configuration models and calculate modularity score of each,

- order $t^*, t_1, \dots, t_n$
- if t^* among top m values reject null hypothesis for distribution of network

- rule of thumb, take m at least 5.

Monte-Carlo tests

If null hypothesis true all orderings of data are equally likely, the probability that the one you observe is among the top m is $m/(n+1)$

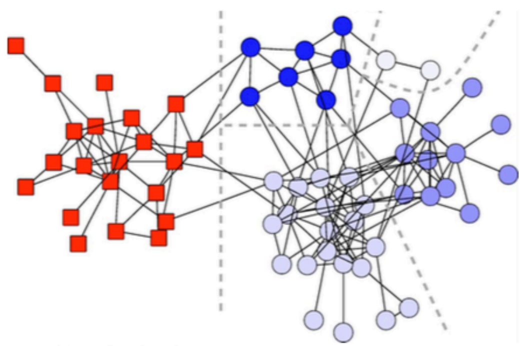
Also called ‘parametric bootstrapping’, ‘conditional uniform graph test’

Watch

- https://www.youtube.com/watch?v=QT2xj9k00q0
- 1:07-1:11 discusses Monte-Carlo test

Monte Carlo example

- with Erdos-Renyi random graph as null model.

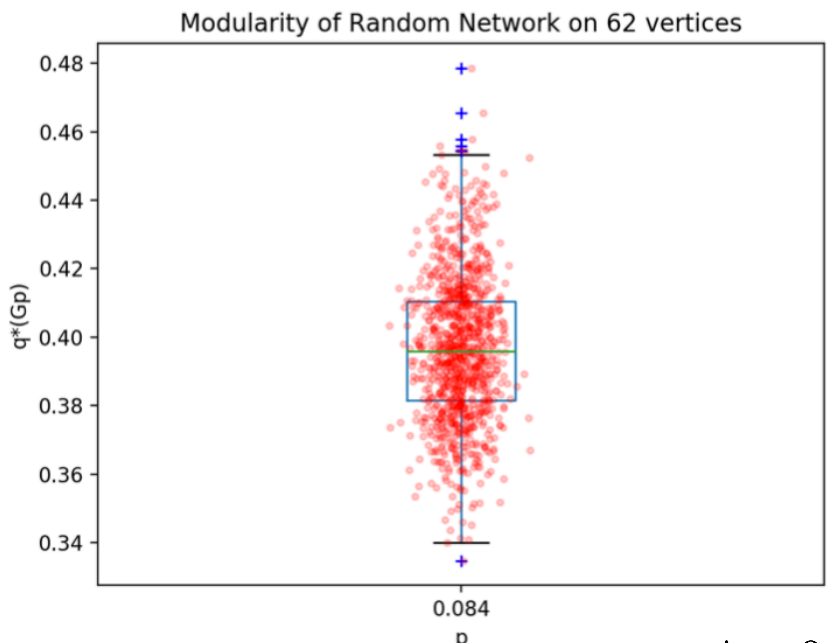


$\#$ dolphins $= 62$
$\#$ edges $= 159$

$q^* = 0.52$

8.4% of possible edges

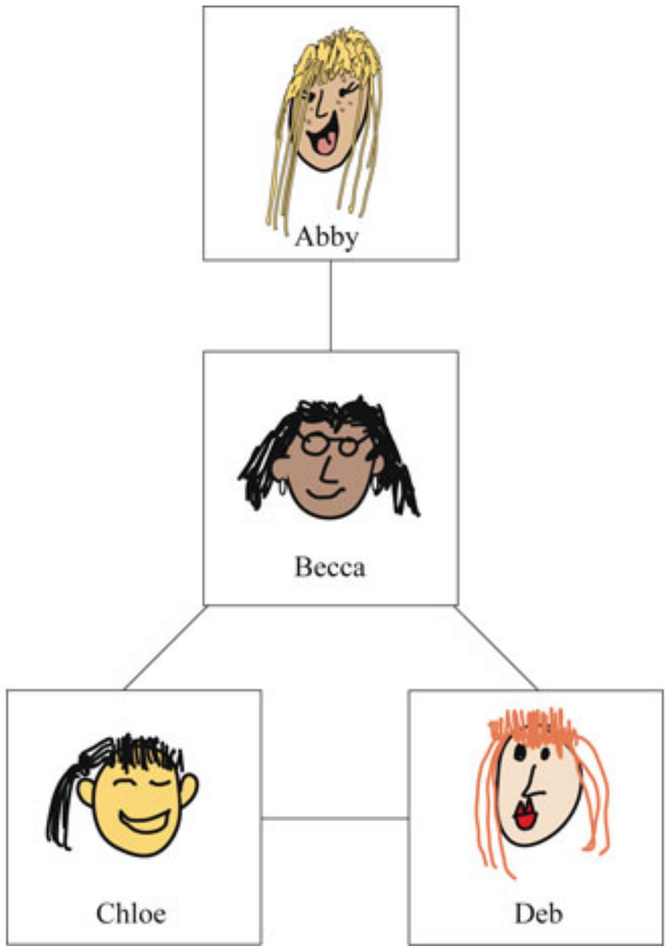
$q^*(\text{dolphins}) > q^*(\text{random network})??$



- The test statistic on our network is $t^* = 0.52$
- Red dots on graph above give test statistic on graphs sampled from the null model. Note t^* is greater than value of test statistic on any generated graph.

- Calculate p so that null model has same expected degree: $p = \#\text{edges}/\binom{\#\text{dolphins}}{2}$

Friendship Paradox



https://opinionator.blogs.nytimes.com/2012/09/17/friends-you-can-count-on/

Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media

Friendship Paradox Redux: Your Friends Are More Interesting Than You

Nathan O. Hodas
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
nhodas@isi.edu

Farshad Kooti
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
kooti@usc.edu

Kristina Lerman
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
lerman@isi.edu

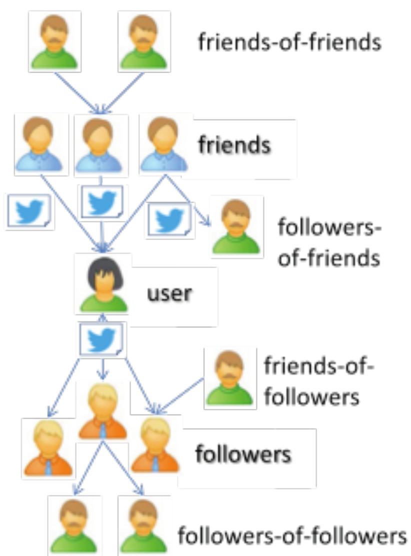
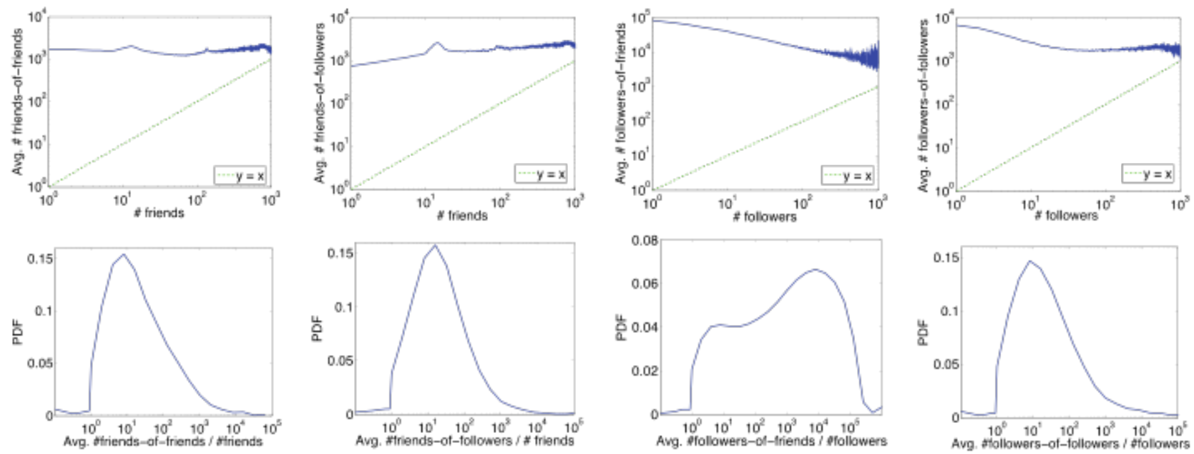


Figure 1: *An example of a directed network of a social media site with information flow links. Users receive information from their friends and broadcast information to their followers.*



(i) (ii) (iii) (iv)