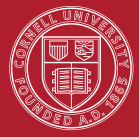


Cornell University

Lecture 10- Random Graph Models and Real Networks

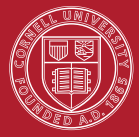
Graph-Based Data Science For Networked Systems
ECE 5260 – ORIE 5735

Anna Scaglione

Cornell Tech

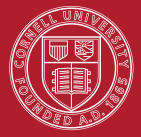
February 25, 2026

Content



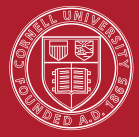
Cornell University

Outline



Cornell University

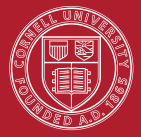
Introduction



Cornell University

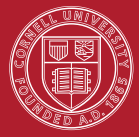
- In the last lecture we extended random graphs models to match some properties of real networks
- We also discussed the Small World effect, the empirical observation that networks tend to have average distance which is $O(\log N)$
- Today with talk about few more characterisitcs that are present in real networks, and their connectivity

Outline



Cornell University

Outline

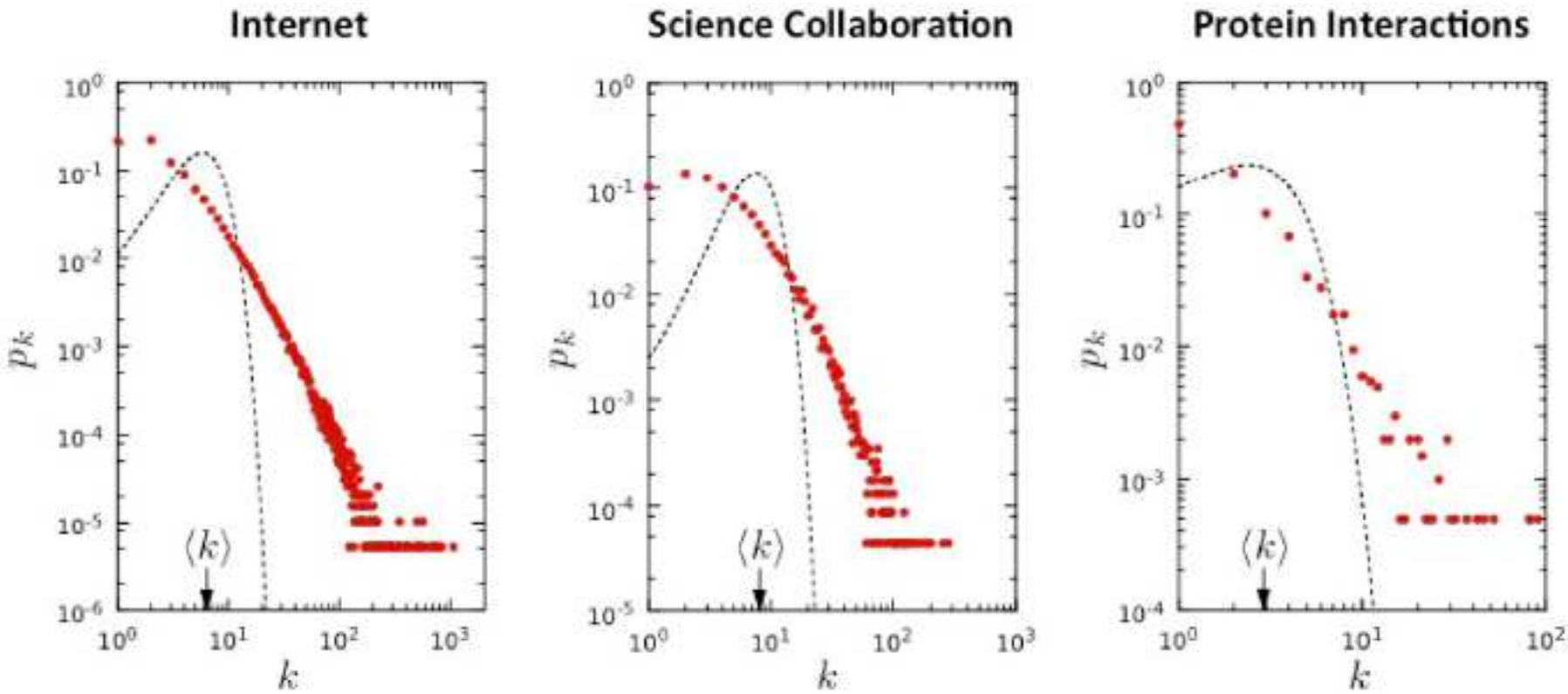


Cornell University

Power law in the degree distribution

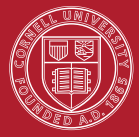
Cornell University

- Empirically several networks diplay a power law trend in their degree distribution



In *scale free networks* a substantial percentage of nodes that have a degree orders of magnitude larger than the average degree

The name "scale free"



- Any function $f(x)$ that has the property:

$$f(ax) \propto f(x)$$

  is called "scale free"
- Power law functions have this property:

$$f(x) = x^{-\alpha} \quad f(ax) = a^{-\alpha}x^{-\alpha} \propto f(x)$$

Power Law Degree Distribution

Let p_k denote the degree distribution, i.e. $p_k := P(d_u = k)$.

Scale-free networks

In scale free networks the degree distribution is such that (C is a constant to normalize the distribution)

$$\ln p_k = -\alpha \ln k + const. \quad \Rightarrow p_k = Ck^{-\alpha}$$

- $-\alpha$ is the negative slope of the trend in the previous slides figures. Typical values are in the range $2 < \alpha < 3$
- When $\alpha \leq 1$ the series $k^{-\alpha}$ does not converge.

Power Law Degree Distribution

Let p_k denote the degree distribution, i.e. $p_k := P(d_u = k)$.

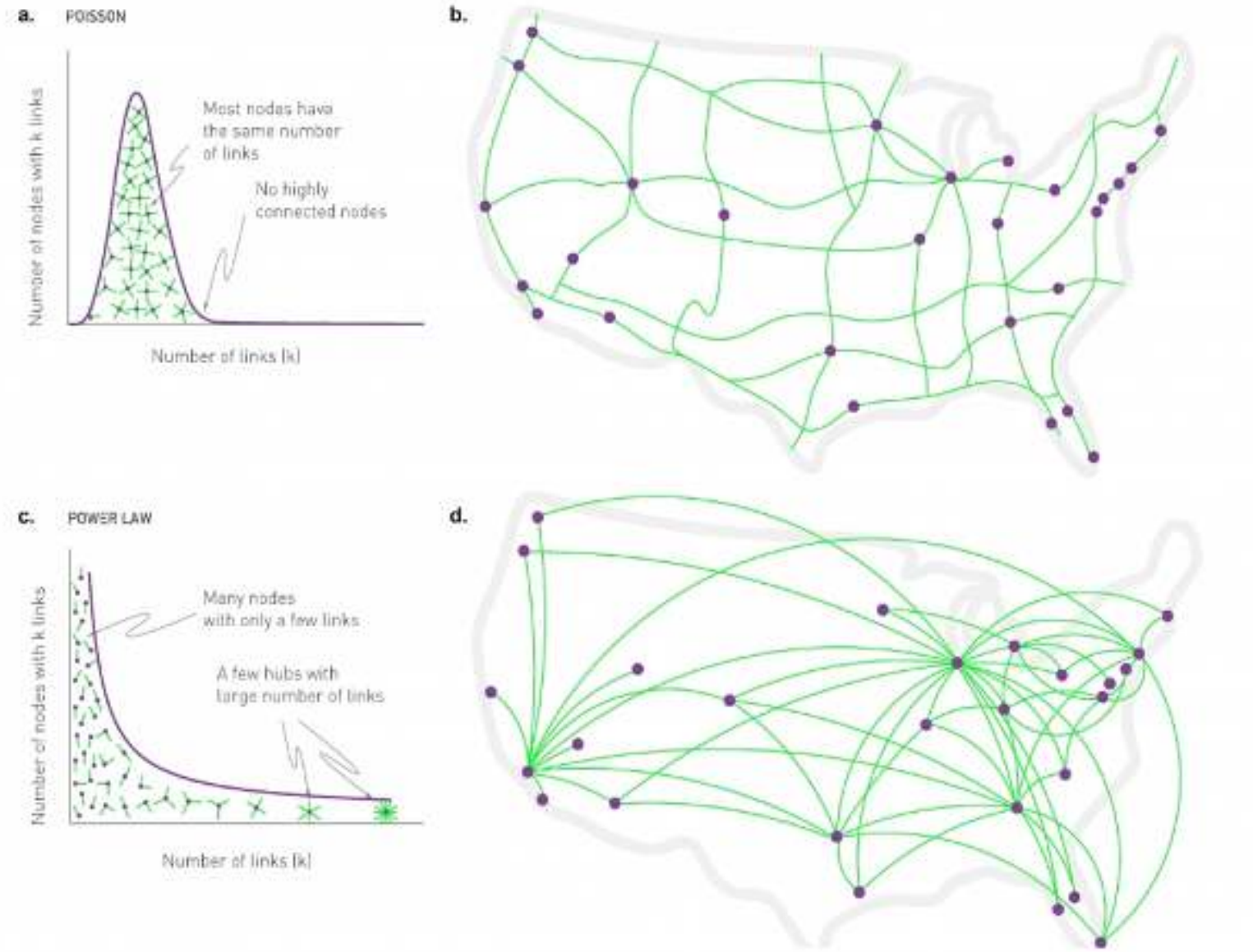
Scale-free networks

In scale free networks the degree distribution is such that (C is a constant to normalize the distribution)

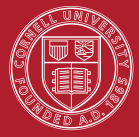
$$\ln p_k = -\alpha \ln k + const. \quad \Rightarrow p_k = Ck^{-\alpha}$$

- $-\alpha$ is the negative slope of the trend in the previous slides figures. Typical values are in the range $2 < \alpha < 3$
- When $\alpha \leq 1$ the series $k^{-\alpha}$ does not converge.

What does scale-free degree imply?

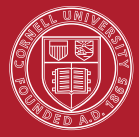


Networks that exhibit a power law trend



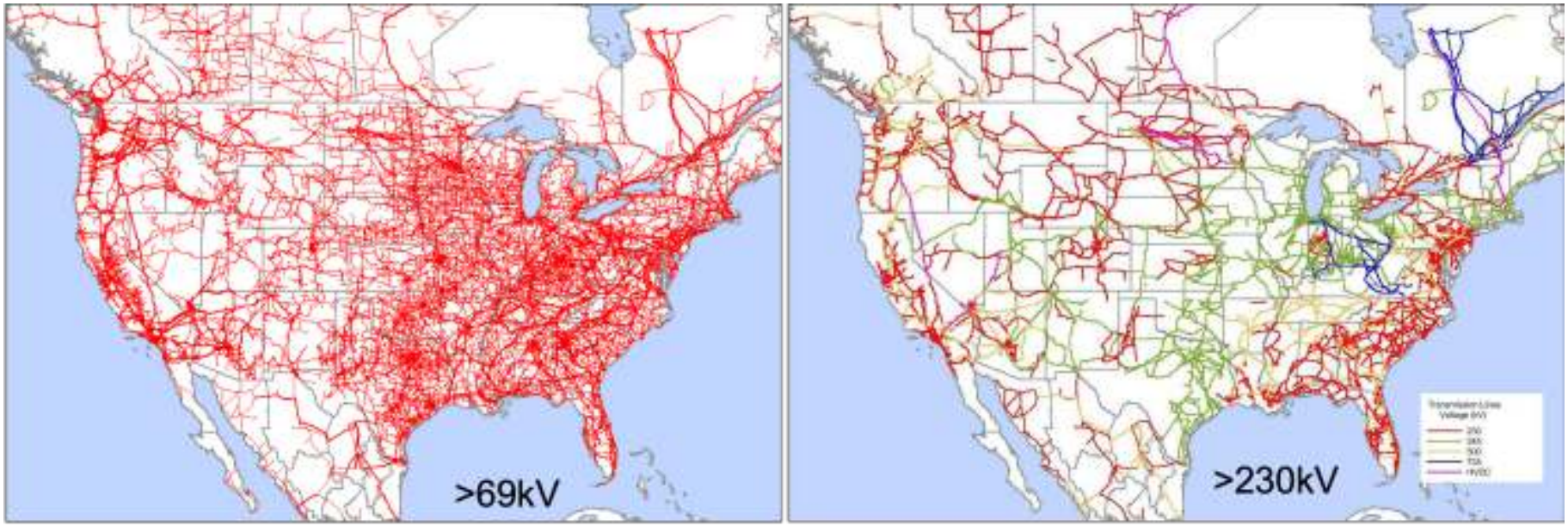
- Among the networks that showcase this trend are:
  1. Some Social networks and collaboration networks.
  2. Computer networks
  3. The webgraph of the World Wide Web
  4. Semantic networks
  5. Protein-protein interaction networks
  6. Airline networks
  7. ....

Not all networks are scale-free

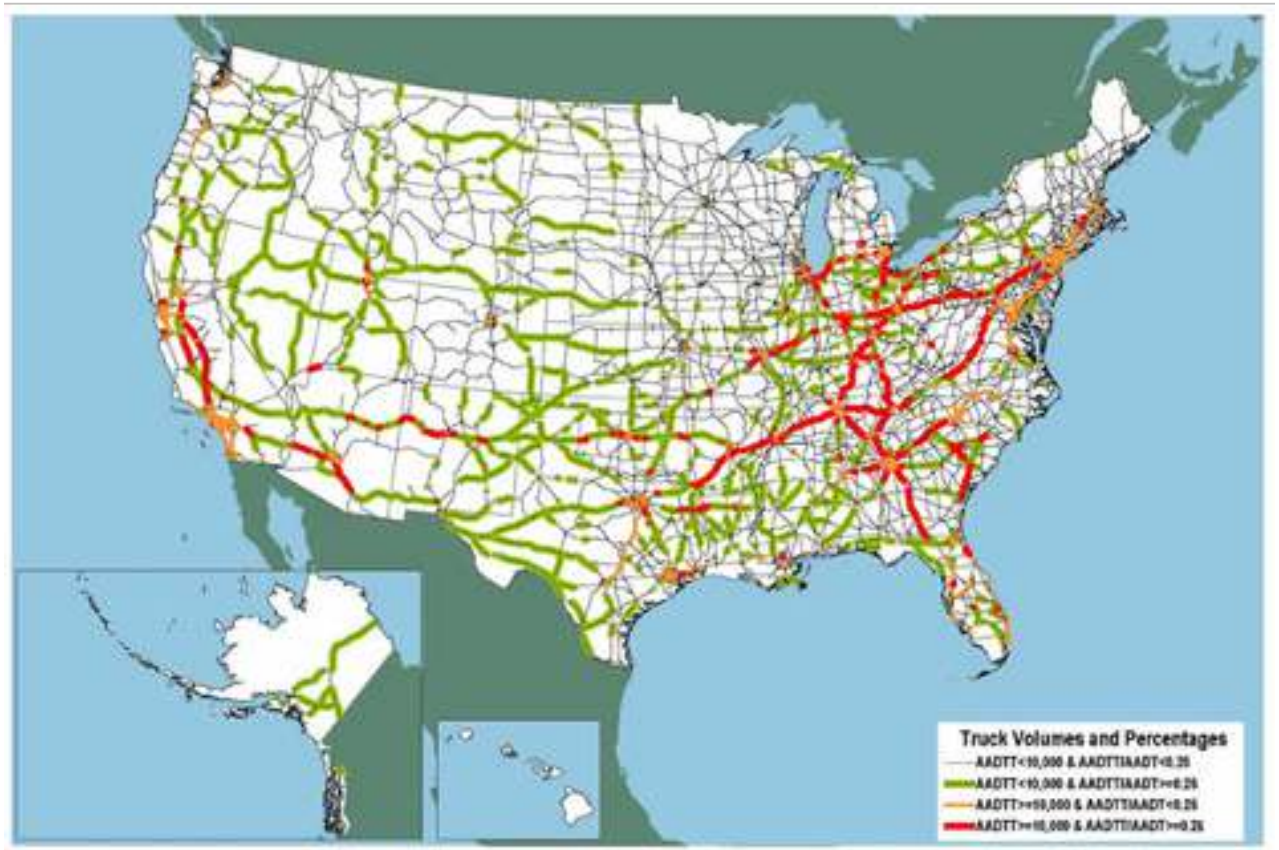


- Some infrastructure networks, like the Internet, can have nodes that manage a much larger amount of traffic, but many physical infrastructure have limits on what edge and node degrees
- Practically speaking networks that support flows and that are bound by physical constraints may not follow this model
- These networks have an exponential tail in the degree distribution, not a power law

A non scale-free network: power systems

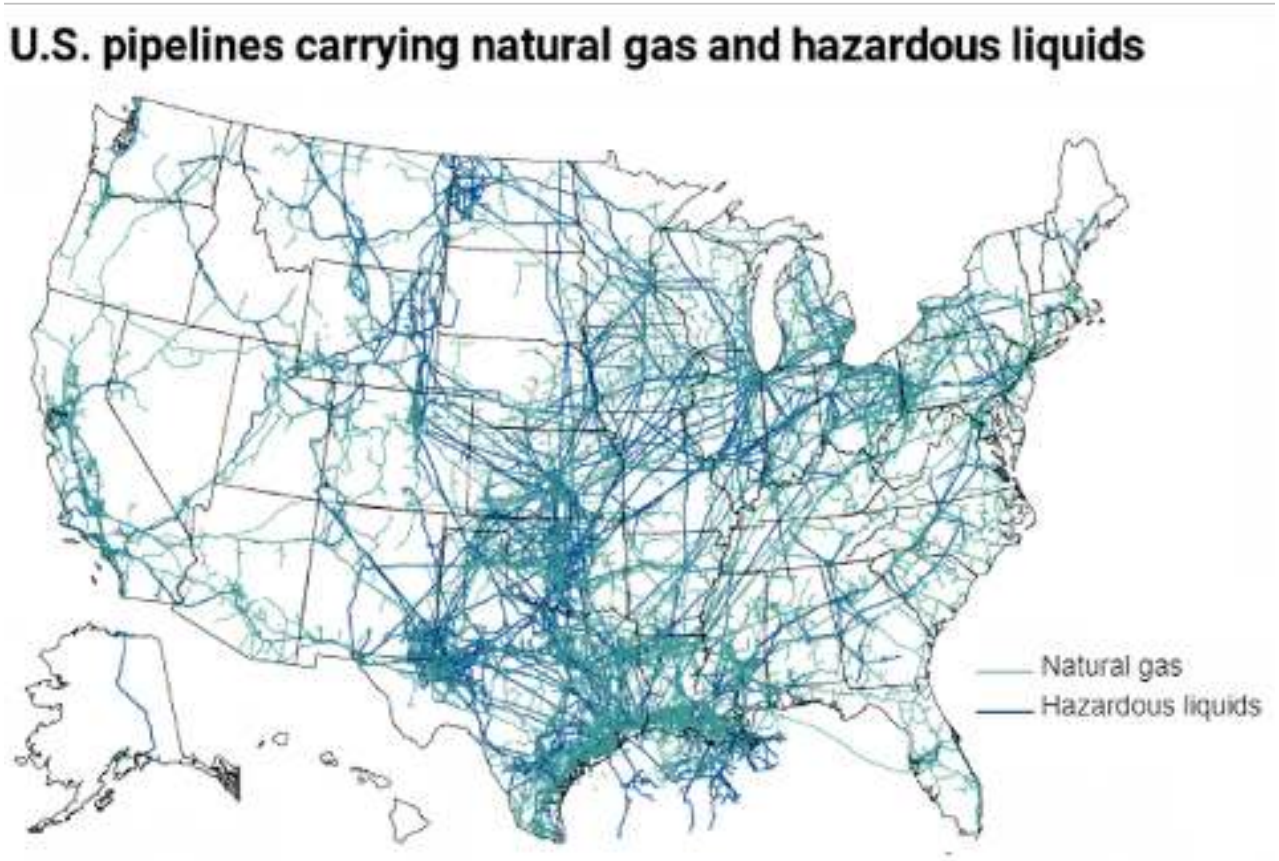


Other infrastructures: transportation

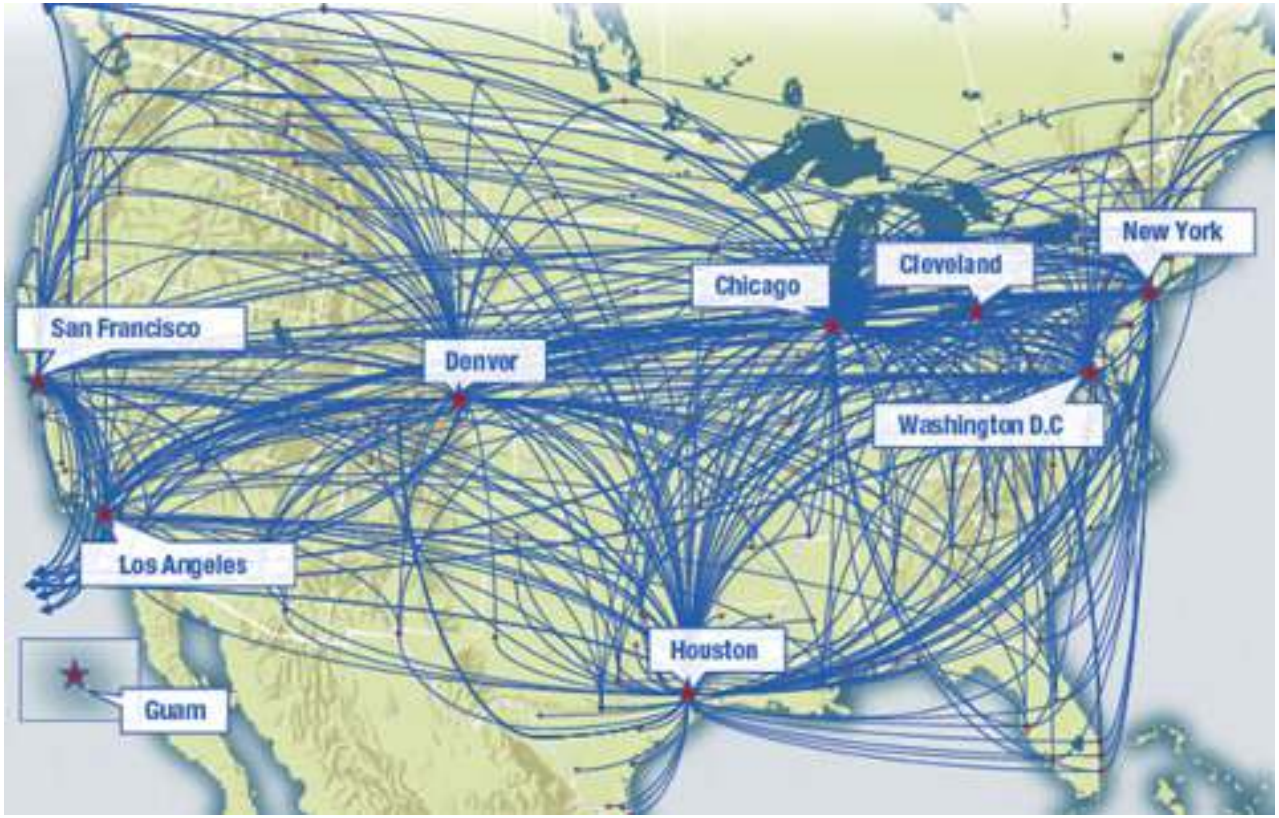


This image shows the highways system and the truckload

Other infrastructures: gas, oil

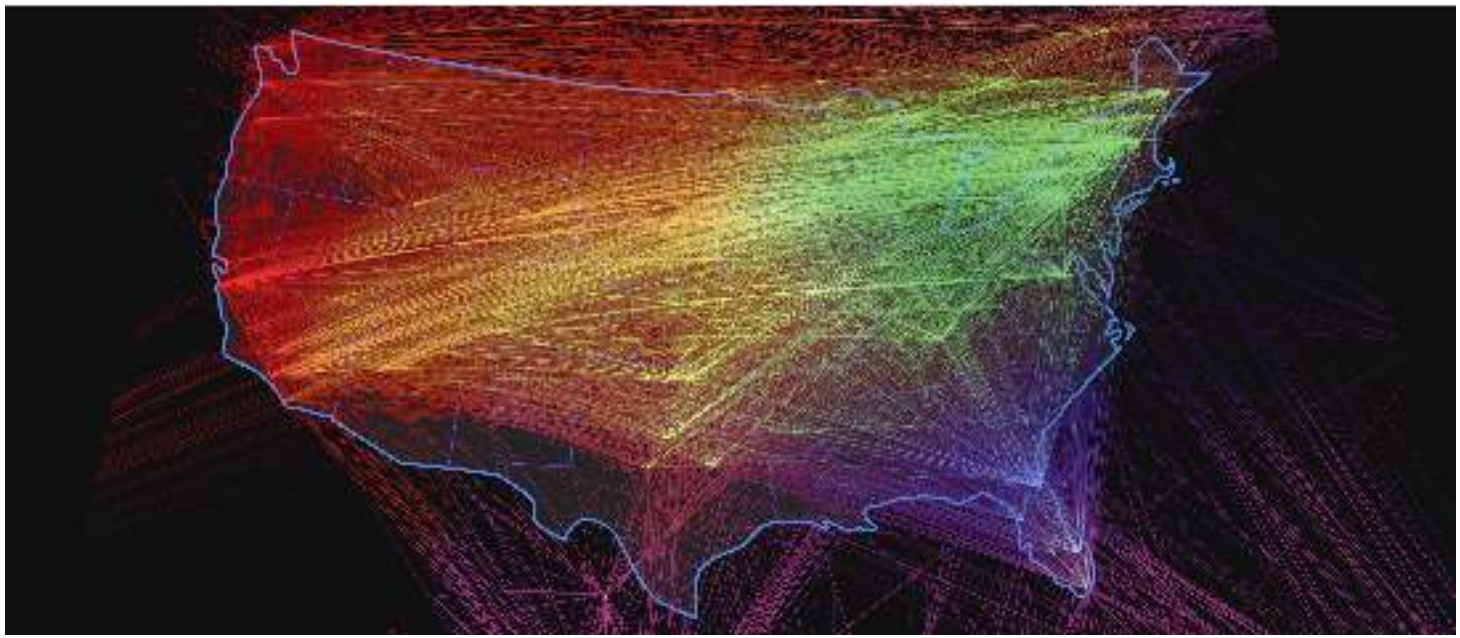


Other infrastructures: air transportation



When geography, and capacity of the line, is not a limitation, the trend of having hubs dominates.

The Internet



The privatization of the backbone of the Internet has lead to a tremendous growth of connectivity and the creation of hubs

Complementary Cumulative Distribution Function

- Histograms can be *noisy* because some representative data are missing, so if the bins contain only consecutive integer values, there can be gaps
- The Complementary Cumulative Distribution Function (CCDF) $:= P(d_u > k)$ fills the gaps because its estimate adds up the frequency of all $d_u > k$
- The power law trend will emerge naturally in log scale, but with a negative slope $\alpha - 1$. In fact

$$P(d_u > k) = \sum_{x=k+1}^{+\infty} p_k = C \overbrace{\sum_{x=k+1}^{+\infty} x^{-\alpha}}^{\approx \int_k^{+\infty} x^{-\alpha} dx = \frac{k^{-(\alpha-1)}}{(\alpha-1)}} \approx \frac{Ck^{-(\alpha-1)}}{(\alpha-1)}$$

Complementary Cumulative Distribution Function

- Histograms can be *noisy* because some representative data are missing, so if the bins contain only consecutive integer values, there can be gaps
- The Complementary Cumulative Distribution Function (CCDF) $:= P(d_u > k)$ fills the gaps because its estimate adds up the frequency of all $d_u > k$
- The power law trend will emerge naturally in log scale, but with a negative slope $\alpha - 1$. In fact

$$P(d_u > k) = \sum_{x=k+1}^{+\infty} p_k = C \overbrace{\sum_{x=k+1}^{+\infty} x^{-\alpha}}^{\approx \int_k^{+\infty} x^{-\alpha} dx = \frac{k^{-(\alpha-1)}}{(\alpha-1)}} \approx \frac{Ck^{-(\alpha-1)}}{(\alpha-1)}$$

Complementary Cumulative Distribution Function

- Histograms can be *noisy* because some representative data are missing, so if the bins contain only consecutive integer values, there can be gaps
- The Complementary Cumulative Distribution Function (CCDF) $:= P(d_u > k)$ fills the gaps because its estimate adds up the frequency of all $d_u > k$
- The power law trend will emerge naturally in log scale, but with a negative slope $\alpha - 1$. In fact

$$P(d_u > k) = \sum_{x=k+1}^{+\infty} p_k = C \overbrace{\sum_{x=k+1}^{+\infty} x^{-\alpha}}^{\approx \int_k^{+\infty} x^{-\alpha} dx = \frac{k^{-(\alpha-1)}}{(\alpha-1)}} \approx \frac{Ck^{-(\alpha-1)}}{(\alpha-1)}$$

Comments on α

- Once the trend is discerned rather than a linear fit a better heuristic estimate $\hat{\alpha}$ of the slope is to use ($d_{\min} := \min(d_v)$):

$$\hat{\alpha} = 1 + N \left(\sum_{v=1}^{N} \ln \frac{d_v}{d_{\min} - 1} \right)^{-1}$$

- It is interesting to visualize the impact of the scale free property on the so called **Lorenz curves**

Lorenz curve

For every k, $P(d_v > k)$ is the fraction of nodes that has degree higher than k. There is a certain probability W that an edge is connected to this group. This can be seen as the fraction of edges that is connected to this group. Lorenz curve shows the trend of W as function of the value $P = P(d_v > k)$, $0 \leq P \leq 1$.

Comments on α

- Once the trend is discerned rather than a linear fit a better heuristic estimate $\hat{\alpha}$ of the slope is to use ($d_{\min} := \min(d_v)$):

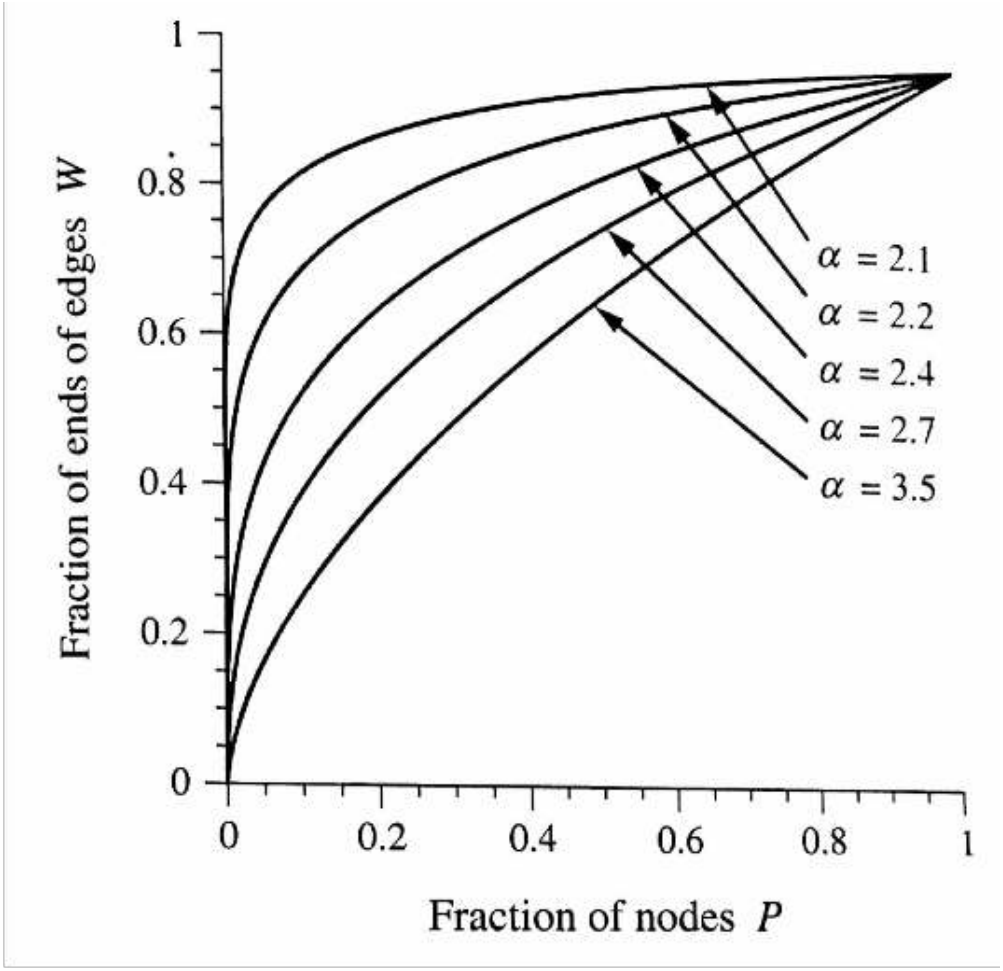
$$\hat{\alpha} = 1 + N \left(\sum_{v=1}^{N} \ln \frac{d_v}{d_{\min} - 1} \right)^{-1}$$

- It is interesting to visualize the impact of the scale free property on the so called **Lorenz curves**

Lorenz curve

For every k, $P(d_v > k)$ is the fraction of nodes that has degree higher than k. There is a certain probability W that an edge is connected to this group. This can be seen as the fraction of edges that is connected to this group. Lorenz curve shows the trend of W as function of the value $P = P(d_v > k)$, $0 \leq P \leq 1$.

Comments on α

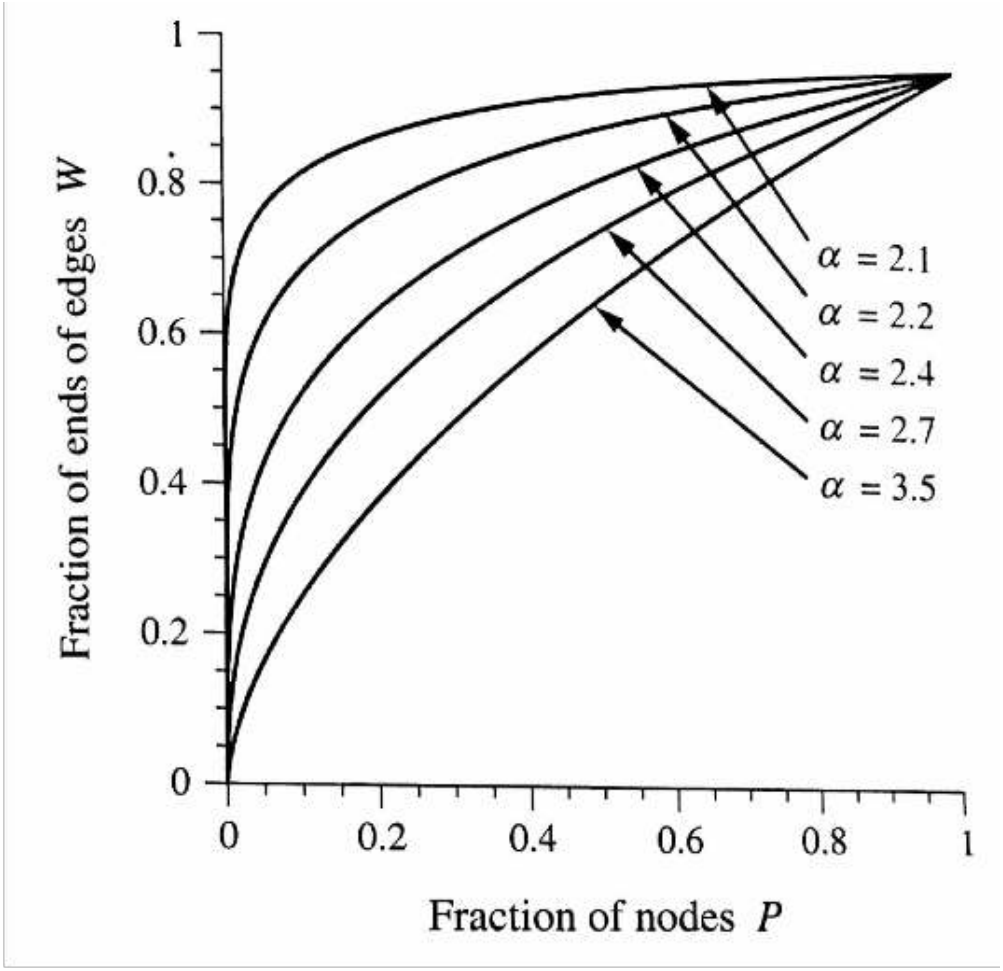


- For a pure power-law degree ($N \to +\infty$), Lorenz curve is:

$$W = P^{\frac{\alpha-2}{\alpha-1}}$$

- The smaller is α the slower the tail decays, the greater the percentage of nodes that is attached to a relatively small number of nodes.
- These nodes are **hubs**.

Comments on α



- For a pure power-law degree ($N \to +\infty$), Lorenz curve is:

$$W = P^{\frac{\alpha-2}{\alpha-1}}$$

- The smaller is α the slower the tail decays, the greater the percentage of nodes that is attached to a relatively small number of nodes.
- These nodes are **hubs**.

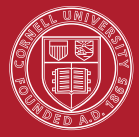
Comments on α

- The main reason scale-free network emerge is that they feature path distances even smaller than the distances observed in an equivalent random network
- In fact, the average distance tends to depend on α as follows

$$\text{average distance} = \begin{cases} const. & \alpha = 2 \\ \ln \ln N & 2 < \alpha < 3 \\ \frac{\ln N}{\ln \ln N} & \alpha = 3 \\ \ln N & \alpha > 3 \end{cases}$$

- For networks that are scale free, most often $2 < \alpha < 3$.
- The anomaly happens for $\alpha = 2$, because the degree variance does not converge for $N \to +\infty$

Comments on α

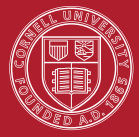


- The main reason scale-free network emerge is that they feature path distances even smaller than the distances observed in an equivalent random network
- In fact, the average distance tends to depend on α as follows

$$\text{average distance} = \begin{cases} const. & \alpha = 2 \\ \ln \ln N & 2 < \alpha < 3 \\ \frac{\ln N}{\ln \ln N} & \alpha = 3 \\ \ln N & \alpha > 3 \end{cases}$$

- For networks that are scale free, most often $2 < \alpha < 3$.
- The anomaly happens for $\alpha = 2$, because the degree variance does not converge for $N \to +\infty$

Comments on α



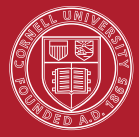
Cornell University

- The main reason scale-free network emerge is that they feature path distances even smaller than the distances observed in an equivalent random network
- In fact, the average distance tends to depend on α as follows

$$\text{average distance} = \begin{cases} const. & \alpha = 2 \\ \ln \ln N & 2 < \alpha < 3 \\ \frac{\ln N}{\ln \ln N} & \alpha = 3 \\ \ln N & \alpha > 3 \end{cases}$$

- For networks that are scale free, most often $2 < \alpha < 3$.
- The anomaly happens for $\alpha = 2$, because the degree variance does not converge for $N \to +\infty$

Comments on α



- The main reason scale-free network emerge is that they feature path distances even smaller than the distances observed in an equivalent random network
- In fact, the average distance tends to depend on α as follows

$$\text{average distance} = \begin{cases} const. & \alpha = 2 \\ \ln \ln N & 2 < \alpha < 3 \\ \frac{\ln N}{\ln \ln N} & \alpha = 3 \\ \ln N & \alpha > 3 \end{cases}$$

- For networks that are scale free, most often $2 < \alpha < 3$.
- The anomaly happens for $\alpha = 2$, because the degree variance does not converge for $N \to +\infty$

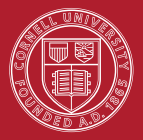
Comments on α

- To understand why the variance does not converge for $N \to +\infty$ is sufficient to look at the divergence of the mean square

$$\mathbb{E}[d_v^2] = \sum_{k=1}^{+\infty} k^2 p_k = \sum_{k=1}^{+\infty} k^2 C k^{-\alpha} = \begin{cases} \sum_{k=1}^{+\infty} C & \text{if } \alpha = 2 \\ \sum_{k=1}^{+\infty} C k^{-1} & \text{if } \alpha = 3 \end{cases}$$

  which diverges in both cases to $+\infty$ (for $\alpha = 3$ we have the harmonic series which is divergent)

Minimum and maximum degree

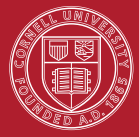


- So far we have done calculations assuming that $p_1 > 0$ so that the minimum possible degree $d_{\min} = 1$
- In general, the normalization costant C for $p_k = Ck^{-\alpha}$, will depend on $d_{\min}$ approximately as follows:

$$1 = \sum_{k=d_{\min}}^{+\infty} Ck^{-\alpha} \approx \int_{d_{\min}}^{+\infty} x^{-\alpha} dx = \frac{d_{\min}^{-(\alpha-1)}}{(\alpha-1)}$$

$$\Rightarrow \quad C \approx (\alpha-1) d_{\min}^{(\alpha-1)}$$

Minimum and maximum degree



- So far we have done calculations assuming that $p_1 > 0$ so that the minimum possible degree $d_{\min} = 1$
- In general, the normalization costant C for $p_k = Ck^{-\alpha}$, will depend on $d_{\min}$ approximately as follows:

$$1 = \sum_{k=d_{\min}}^{+\infty} Ck^{-\alpha} \approx \int_{d_{\min}}^{+\infty} x^{-\alpha} dx = \frac{d_{\min}^{-(\alpha-1)}}{(\alpha-1)}$$

$$\Rightarrow \quad C \approx (\alpha-1) d_{\min}^{(\alpha-1)}$$

Minimum and maximum degree

- Now if we assume that for large N there is only one node with the maximum degree, the CCDF evaluated at $d_{\max}$ should tend to be $1/N$ (i.e. the frequency of the degree $d_{\max}$)
- This implies

$$\frac{1}{N} = \sum_{k=d_{\max}}^{+\infty} Ck^{-\alpha} \approx C\frac{d_{\max}^{-(\alpha-1)}}{\alpha-1} \approx (\alpha-1)d_{\min}^{(\alpha-1)}\frac{d_{\max}^{-(\alpha-1)}}{\alpha-1}$$

$$\Rightarrow \frac{d_{\max}}{d_{\min}} \approx N^{\frac{1}{\alpha-1}} \Rightarrow d_{\max} \approx d_{\min}N^{\frac{1}{\alpha-1}}$$

- Similar calculations for an exponential degree distribution (which I will skip), i.e. $p_k = Ce^{-\lambda k}$, yield a very different result: $d_{\max} \approx d_{\min} + \frac{N}{\lambda}$

Minimum and maximum degree

- Now if we assume that for large N there is only one node with the maximum degree, the CCDF evaluated at $d_{\max}$ should tend to be $1/N$ (i.e. the frequency of the degree $d_{\max}$)
- This implies

$$\frac{1}{N} = \sum_{k=d_{\max}}^{+\infty} Ck^{-\alpha} \approx C\frac{d_{\max}^{-(\alpha-1)}}{\alpha-1} \approx (\alpha-1)d_{\min}^{(\alpha-1)}\frac{d_{\max}^{-(\alpha-1)}}{\alpha-1}$$

$$\Rightarrow \quad \frac{d_{\max}}{d_{\min}} \approx N^{\frac{1}{\alpha-1}} \quad \Rightarrow \quad d_{\max} \approx d_{\min} N^{\frac{1}{\alpha-1}}$$

- Similar calculations for an exponential degree distribution (which I will skip), i.e. $p_k = Ce^{-\lambda k}$, yield a very different result: $d_{\max} \approx d_{\min} + \frac{N}{\lambda}$

Minimum and maximum degree

- Now if we assume that for large N there is only one node with the maximum degree, the CCDF evaluated at $d_{\max}$ should tend to be $1/N$ (i.e. the frequency of the degree $d_{\max}$)
- This implies

$$\frac{1}{N} = \sum_{k=d_{\max}}^{+\infty} Ck^{-\alpha} \approx C\frac{d_{\max}^{-(\alpha-1)}}{\alpha-1} \approx (\alpha-1)d_{\min}^{(\alpha-1)}\frac{d_{\max}^{-(\alpha-1)}}{\alpha-1}$$

$$\Rightarrow \quad \frac{d_{\max}}{d_{\min}} \approx N^{\frac{1}{\alpha-1}} \quad \Rightarrow \quad d_{\max} \approx d_{\min} N^{\frac{1}{\alpha-1}}$$

- Similar calculations for an exponential degree distribution (which I will skip), i.e. $p_k = Ce^{-\lambda k}$, yield a very different result: $d_{\max} \approx d_{\min} + \frac{N}{\lambda}$

Minimum and maximum degree

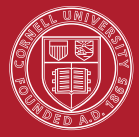
- Now if we assume that for large N there is only one node with the maximum degree, the CCDF evaluated at $d_{\max}$ should tend to be $1/N$ (i.e. the frequency of the degree $d_{\max}$)
- This implies

$$\frac{1}{N} = \sum_{k=d_{\max}}^{+\infty} Ck^{-\alpha} \approx C\frac{d_{\max}^{-(\alpha-1)}}{\alpha-1} \approx (\alpha-1)d_{\min}^{(\alpha-1)}\frac{d_{\max}^{-(\alpha-1)}}{\alpha-1}$$

$$\Rightarrow \quad \frac{d_{\max}}{d_{\min}} \approx N^{\frac{1}{\alpha-1}} \quad \Rightarrow \quad d_{\max} \approx d_{\min}N^{\frac{1}{\alpha-1}}$$

- Similar calculations for an exponential degree distribution (which I will skip), i.e. $p_k = Ce^{-\lambda k}$, yield a very different result: $d_{\max} \approx d_{\min} + \frac{N}{\lambda}$

The case $\alpha < 2$

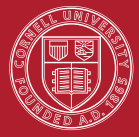


- Note what happens for $\alpha < 2$: $d_{\max}$

$$d_{\max} \approx d_{\min} N^{\frac{1}{2+\epsilon-1}} = d_{\min} N^{\frac{1}{1+\epsilon}} > d_{\min} N$$

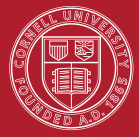
- This case is highly problematic, because the maximum degree would have to grow faster than N!
- This indicates that there may be some problem in maintaining the trend beyond a certain point, and the distributions are essentially truncated.

Outline



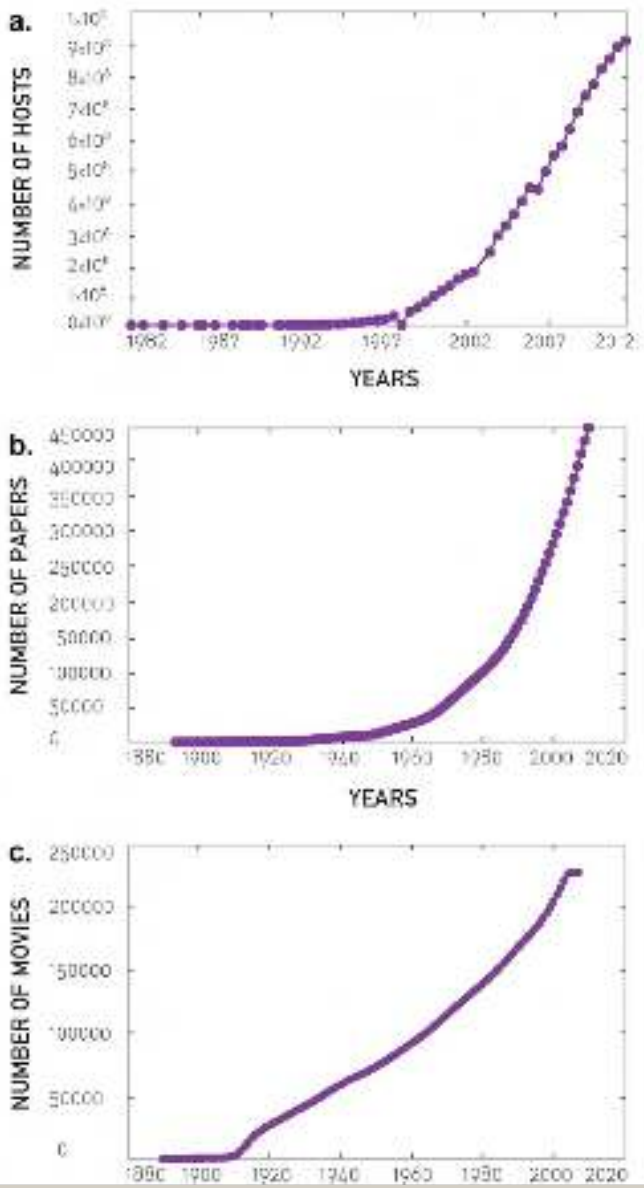
Cornell University

The generation of scale-free networks



- Why do so different systems as the WWW that link web-pages or cellular networks that are interacting through chemical reactions converge to a similar scale-free architecture?
- Is there a random mechanism that could be responsible for the emergence of the scale-free property?
- One way to look at this is by seeing how, as the network expands, nodes select to attach themselves to other nodes: this is the so called **Barabási Albert**

The generation of scale-free networks

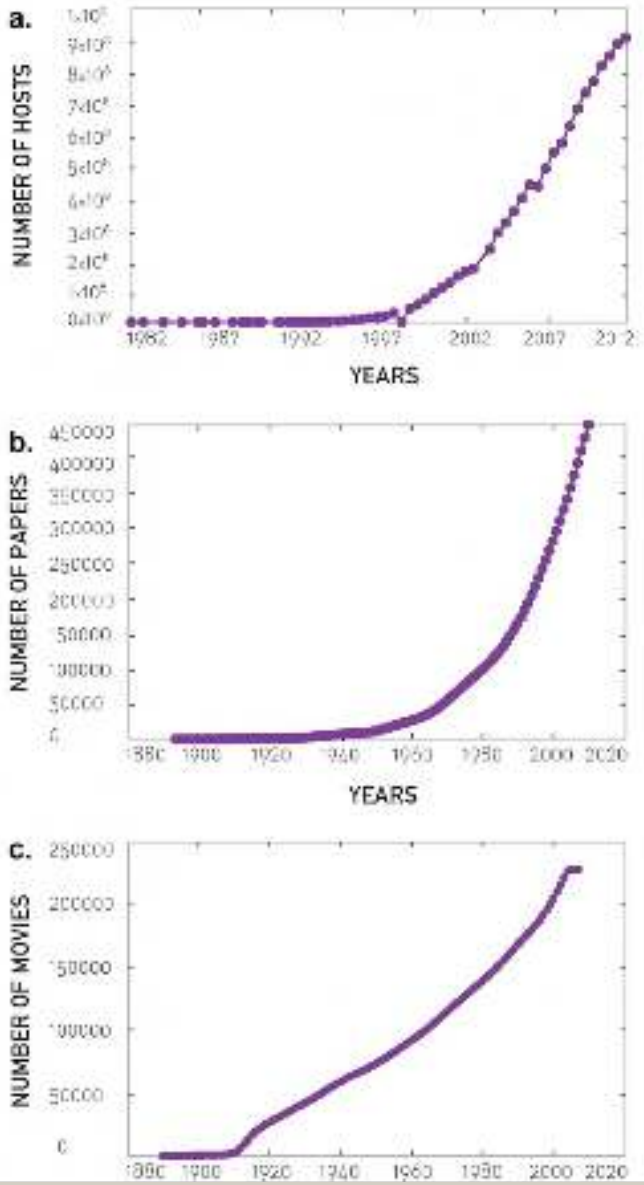


- In real networks new nodes tend to link to the more connected nodes.
- The **preferential attachment model** is based on adding a new node at each $t = 1, \ldots, N$.
- Node v is attached at random to the existing ones $u = 1, \ldots, t-1$ with one of its m edges with probability:

$$p_{tu} = \frac{d_u}{\sum_{k=1}^{t-1} d_k} = \frac{d_u}{2m(t-1)},$$

where $2m(t-1)$ are the edges added so far (handshaking theorem)

The generation of scale-free networks

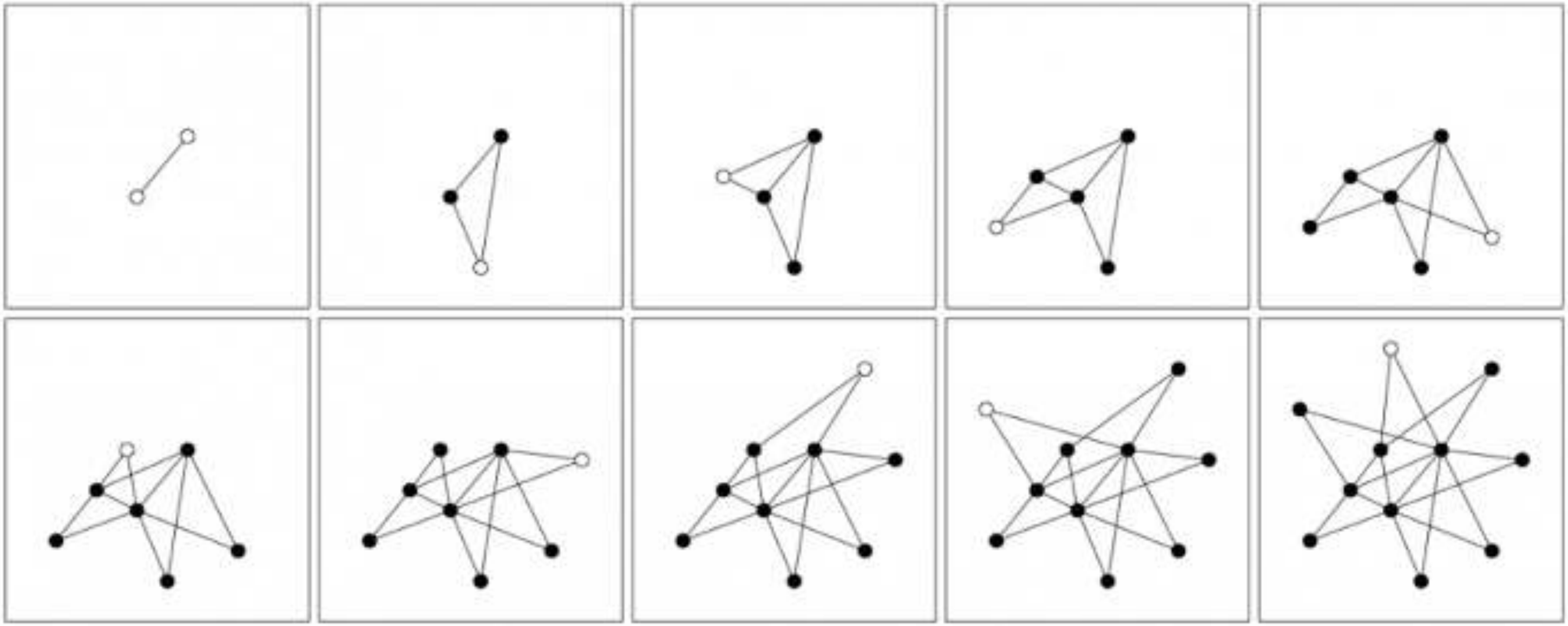


- In real networks new nodes tend to link to the more connected nodes.
- The **preferential attachment model** is based on adding a new node at each $t = 1, \ldots, N$.
- Node v is attached at random to the existing ones $u = 1, \ldots, t-1$ with one of its m edges with probability:

$$p_{tu} = \frac{d_u}{\sum_{k=1}^{t-1} d_k} = \frac{d_u}{2m(t-1)},$$

where $2m(t-1)$ are the edges added so far (handshaking theorem)

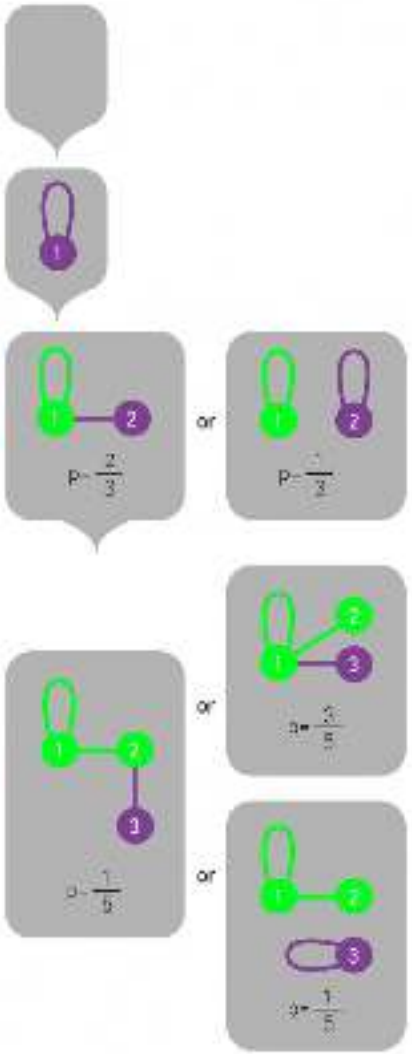
The generation of scale-free networks



- As it turns out this random generation model creates a scale free degree distribution p_k

The generation of scale-free networks

Cornell University

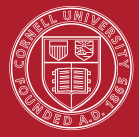


- This is because the nodes that are already in the network tend to be in greater demand
- The derivation of p_k is done with a trick: looking at the growth of the average degree $\mathbb{E}[d_u] = \kappa_u(t)$ as a differential equation

$$\frac{d\kappa_u(t)}{dt} = m\frac{\kappa_u(t)}{2mt} = \frac{\kappa(t)}{2t}$$

where we ignored the -1 for $t \gg 1$

Solution



- The solution of this differential equation is:

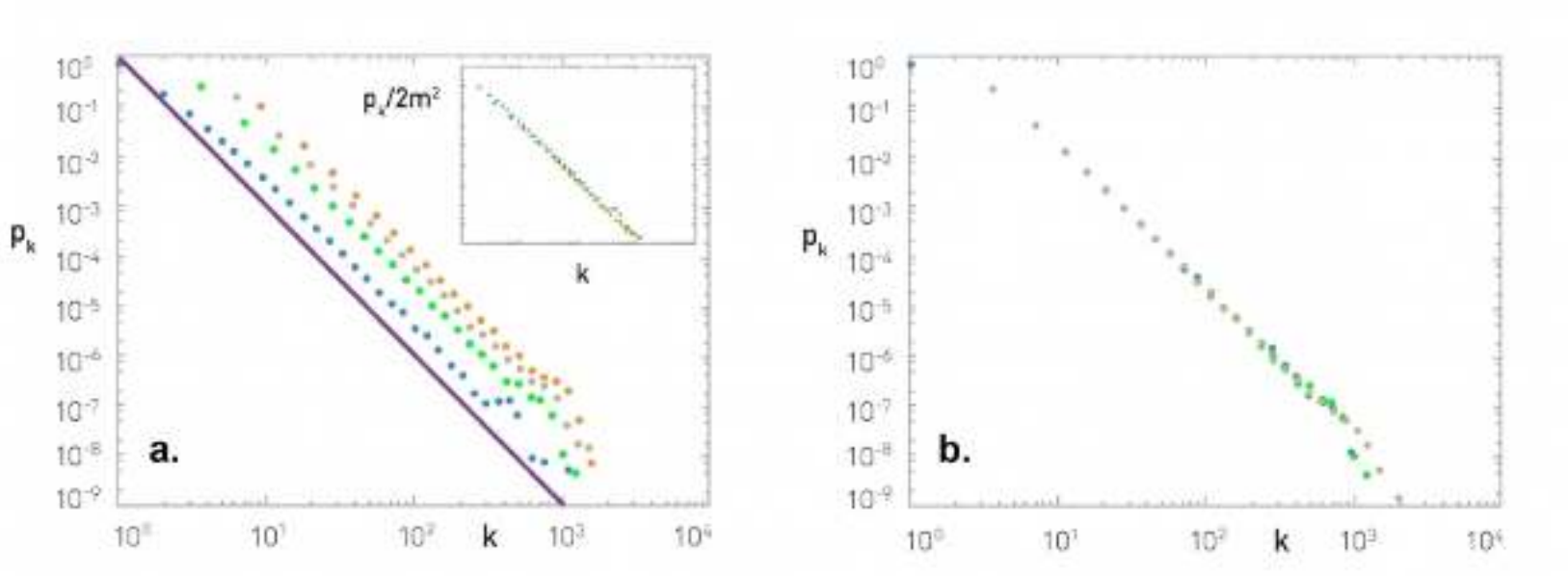
$$\kappa_u(t) = m\sqrt{\frac{t}{t_u}}$$

  where t_u is the time when node u was added (in our notation is actually $t_u = u$)
- The observation is obviously that the smaller is t_u the larger is its expected degree.
- Also the trend is a power law with a **specific exponent**, the average degree grows as $O(t^{1/2}) = O(|\mathcal{V}|^{1/2})$
- What is the degree distribution?

Degree distribution of Barabási Albert model

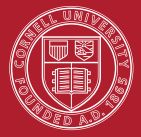
- I will omit the calculations, but in addition to the average degree, one can obtain analytically the degree distribution is

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)} \approx 2m^2 k^{-\alpha}, \quad \alpha = 3$$



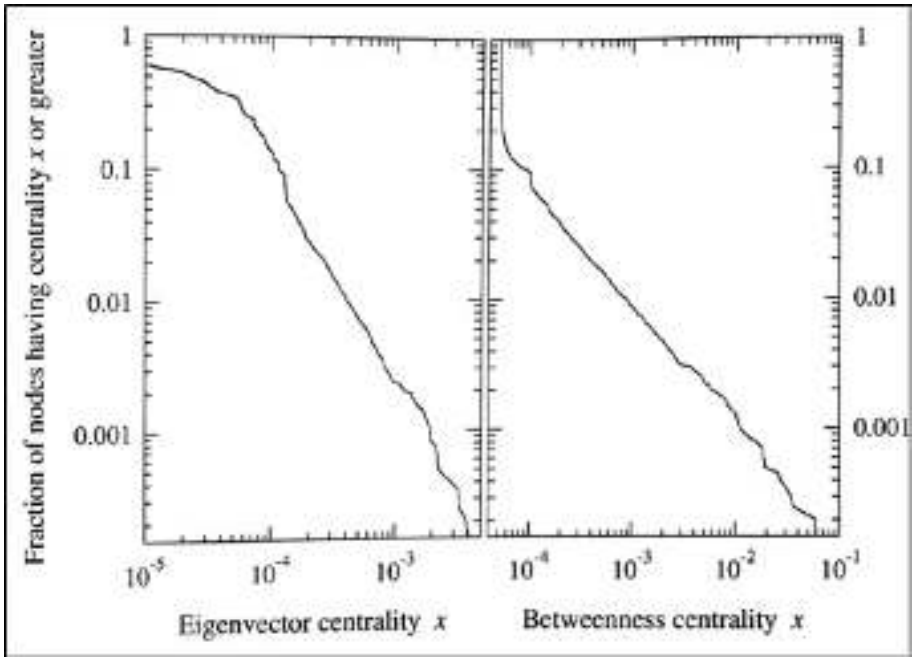
BA networks N=100,000 and m=1 (blue), 3 (green), 5 (grey), and 7 (orange). The fact that the curves are parallel to each other indicates that α is independent of m/ The slope of the purple line is -3, corresponding to the predicted degree exponent α=3.
Since the formula predicts $p_k \sim 2m^2$, $p_k/2m^2$ should be independent of m. Indeed, by plotting $p_k/2m^2$ vs. k, the data points shown in the main plot collapse into a single curve.

Outline



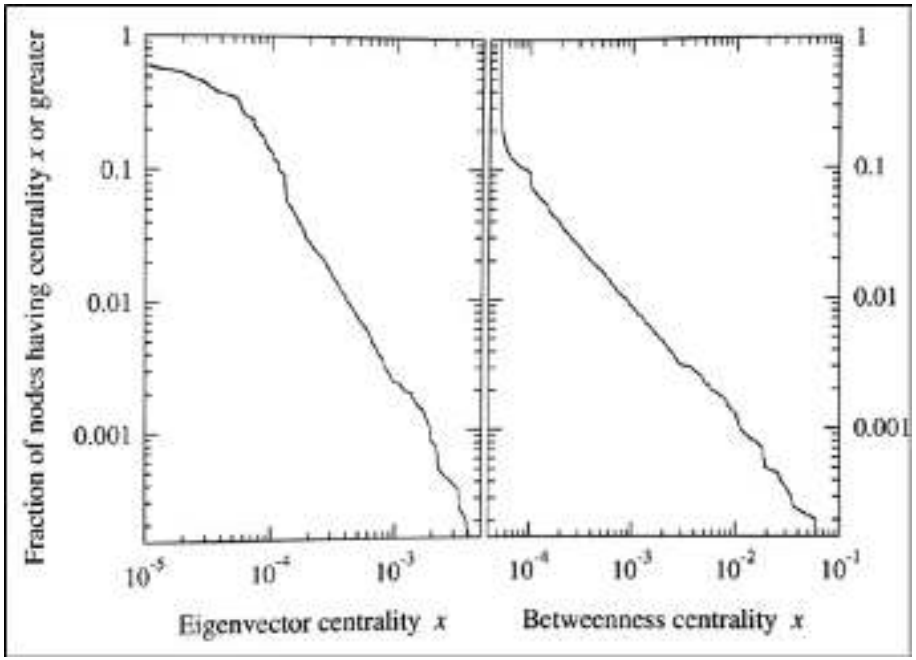
Cornell University

Power law in other features



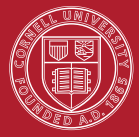
- For scale free networks the **eigenvector centrality** also has a power-law trend. Clearly, it is closely related to the degree centrality
- Interestingly some real networks also display a **betweenes centrality** with a power law trend, even if the network is not scale-free.

Power law in other features



- For scale free networks the **eigenvector centrality** also has a power-law trend. Clearly, it is closely related to the degree centrality
- Interestingly some real networks also display a **betweenes centrality** with a power law trend, even if the network is not scale-free.

Clustering coefficient

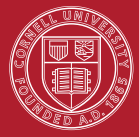


- Recall the definition of clustering coefficient:

$$C = \frac{(\text{number of closed paths of length two})}{(\text{number of paths of length two})}$$

careful, this is not the same C we used before for $p_k = Ck^{-\alpha}$, this is the clustering coefficient!

Clustering coefficient

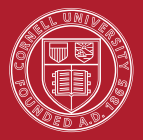


- Recall the definition of clustering coefficient:

$$C = \frac{(\text{number of closed paths of length two})}{(\text{number of paths of length two})}$$

careful, this is not the same C we used before for $p_k = Ck^{-\alpha}$, this is the clustering coefficient!

Table of large real networks parameters



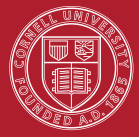
Cornell University

	Network	Type	n	m	c	S	ℓ	α	C	C_{WS}	r	Ref(s).
Social	Film actors	Undirected	449 913	25 516 482	113.43	0.980	3.48	2.3	0.20	0.78	0.208	20,466
	Company directors	Undirected	7 673	55 392	14.44	0.876	4.60	–	0.59	0.88	0.276	131,369
	Math coauthorship	Undirected	253 339	496 489	3.92	0.822	7.57	–	0.15	0.34	0.120	133,219
	Physics coauthorship	Undirected	52 909	245 300	9.27	0.838	6.19	–	0.45	0.56	0.363	347,349
	Biology coauthorship	Undirected	1 520 251	11 803 064	15.53	0.918	4.92	–	0.088	0.60	0.127	347,349
	Telephone call graph	Undirected	47 000 000	80 000 000	3.16			2.1				10,11
	Email messages	Directed	59 812	86 300	1.44	0.952	4.95	1.5/2.0		0.16		156
	Email address books	Directed	16 881	57 029	3.38	0.590	5.22	–	0.17	0.13	0.092	364
	Student dating	Undirected	573	477	1.66	0.503	16.01	–	0.005	0.001	–0.029	52
	Sexual contacts	Undirected	2 810					3.2				304, 305
Information	WWW nd.edu	Directed	269 504	1 497 135	5.55	1.000	11.27	2.1/2.4	0.11	0.29	–0.067	16,41
	WWW AltaVista	Directed	203 549 046	1 466 000 000	7.20	0.914	16.18	2.1/2.7				84
	Citation network	Directed	783 339	6 716 198	8.57			3.0/–				404
	Roget's Thesaurus	Directed	1 022	5 103	4.99	0.977	4.87	–	0.13	0.15	0.157	272
	Word co-occurrence	Undirected	460 902	16 100 000	66.96	1.000		2.7		0.44		146, 175
Technological	Internet	Undirected	10 697	31 992	5.98	1.000	3.31	2.5	0.035	0.39	–0.189	102, 168
	Power grid	Undirected	4 941	6 594	2.67	1.000	18.99	–	0.10	0.080	0.003	466
	Train routes	Undirected	587	19 603	66.79	1.000	2.16	–		0.69	–0.033	425
	Software packages	Directed	1 439	1 723	1.20	0.998	2.42	1.6/1.4	0.070	0.082	–0.016	352
	Software classes	Directed	1 376	2 213	1.61	1.000	5.40	–	0.033	0.012	–0.119	453
	Electronic circuits	Undirected	24 097	53 248	4.34	1.000	11.05	3.0	0.010	0.030	–0.154	174
	Peer-to-peer network	Undirected	880	1 296	1.47	0.805	4.28	2.1	0.012	0.011	–0.366	6,409
Biological	Metabolic network	Undirected	765	3 686	9.64	0.996	2.56	2.2	0.090	0.67	–0.240	252
	Protein interactions	Undirected	2 115	2 240	2.12	0.689	6.80	2.4	0.072	0.071	–0.156	250
	Marine food web	Directed	134	598	4.46	1.000	2.05	–	0.16	0.23	–0.263	245
	Freshwater food web	Directed	92	997	10.84	1.000	1.90	–	0.20	0.087	–0.326	321
	Neural network	Directed	307	2 359	7.68	0.967	3.97	–	0.18	0.28	–0.226	466,470

Table 10.1: Basic statistics for a number of networks. The properties measured are: type of network, directed or undirected; total number of nodes n; total number of edges m; mean degree c; fraction of nodes in the largest component S (or the largest weakly connected component in the case of a directed network); mean distance between connected node pairs ℓ; exponent α of the degree distribution if the distribution follows a power law (or "–" if not; in/out-degree exponents are given for directed networks); clustering coefficient C from Eq. (7.28); clustering coefficient C_{WS} from the alternative definition of Eq. (7.31); and the degree correlation coefficient r from Eq. (7.64). The last column gives the citation(s) for each network in the References. Blank entries indicate unavailable data.

From the book "Networks" by Mark Newman

Clustering coefficient

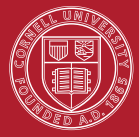


- If you look at social networks values for the average degree you cannot match the formula
- This can be only explained by recognizing that in social networks these links (with collaborators, friends...) are not born at random
- What is interesting is that technological networks do not have a high clustering coefficient

Assortativity of real networks

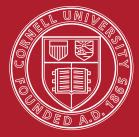
- Another interesting difference between technological, biological and social networks is in the degree assortativity they display
- Social networks tend to be assortative. The way we befriend/collaborate is similar to that of our peers:
  - Highly connected people tend to be connected with other highly connected people, and viceversa people with fewer connections tend to stick with people that have fewer connections
- Technological and biological networks tend to be disassortative or in general with small r. This may be attributed to the presence of hubs.

Assortativity of real networks



- Another interesting difference between technological, biological and social networks is in the degree assortativity they display
- Social networks tend to be assortative. The way we befriend/collaborate is similar to that of our peers:
  - Highly connected people tend to be connected with other highly connected people, and viceversa people with fewer connections tend to stick with people that have fewer connections
- Technological and biological networks tend to be disassortative or in general with small r. This may be attributed to the presence of hubs.

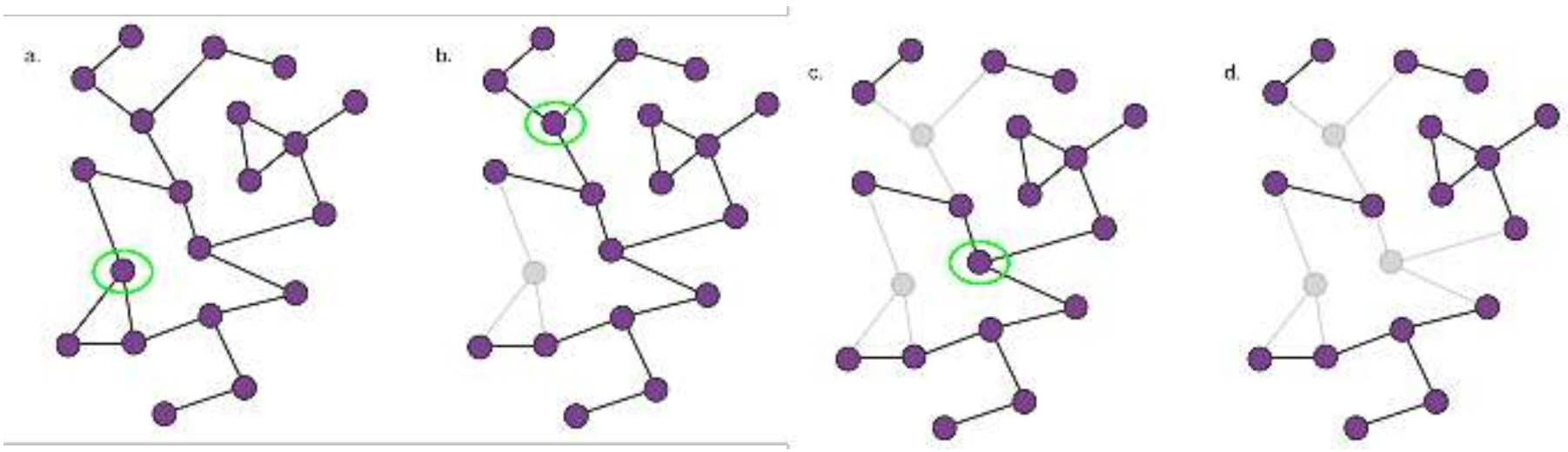
Outline



Cornell University

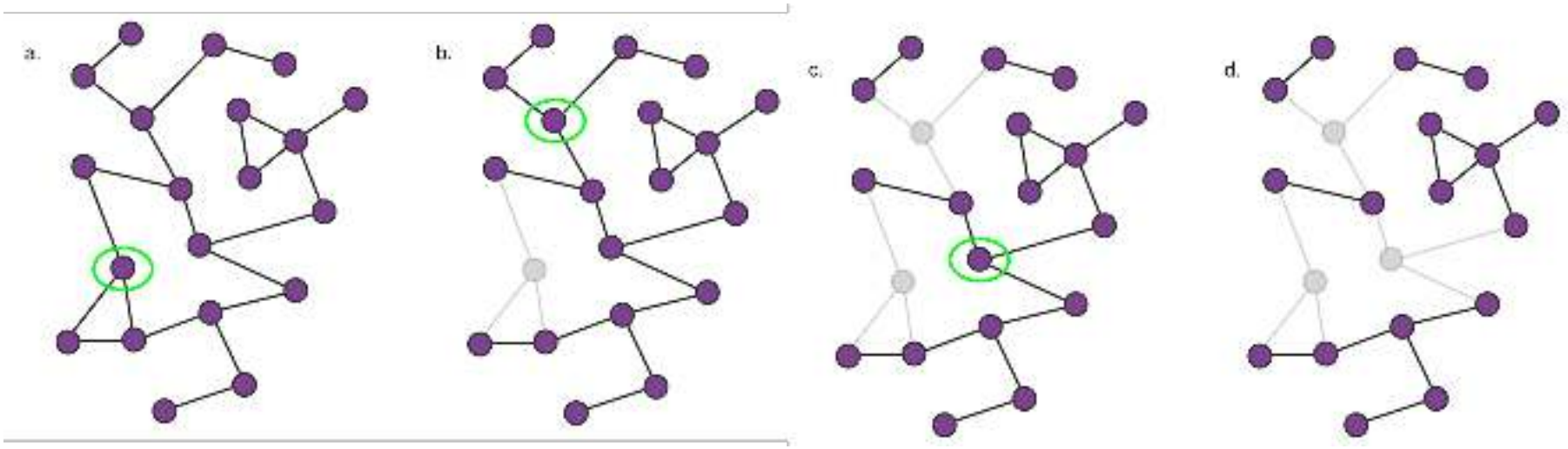
Robustness: Percolation theory

- A single node removal has only limited impact on a network's integrity
- The removal of several nodes, however, can break a network into several isolated components

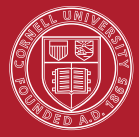


Robustness: Percolation theory

- A single node removal has only limited impact on a network's integrity
- The removal of several nodes, however, can break a network into several isolated components

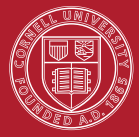


Percolation theory



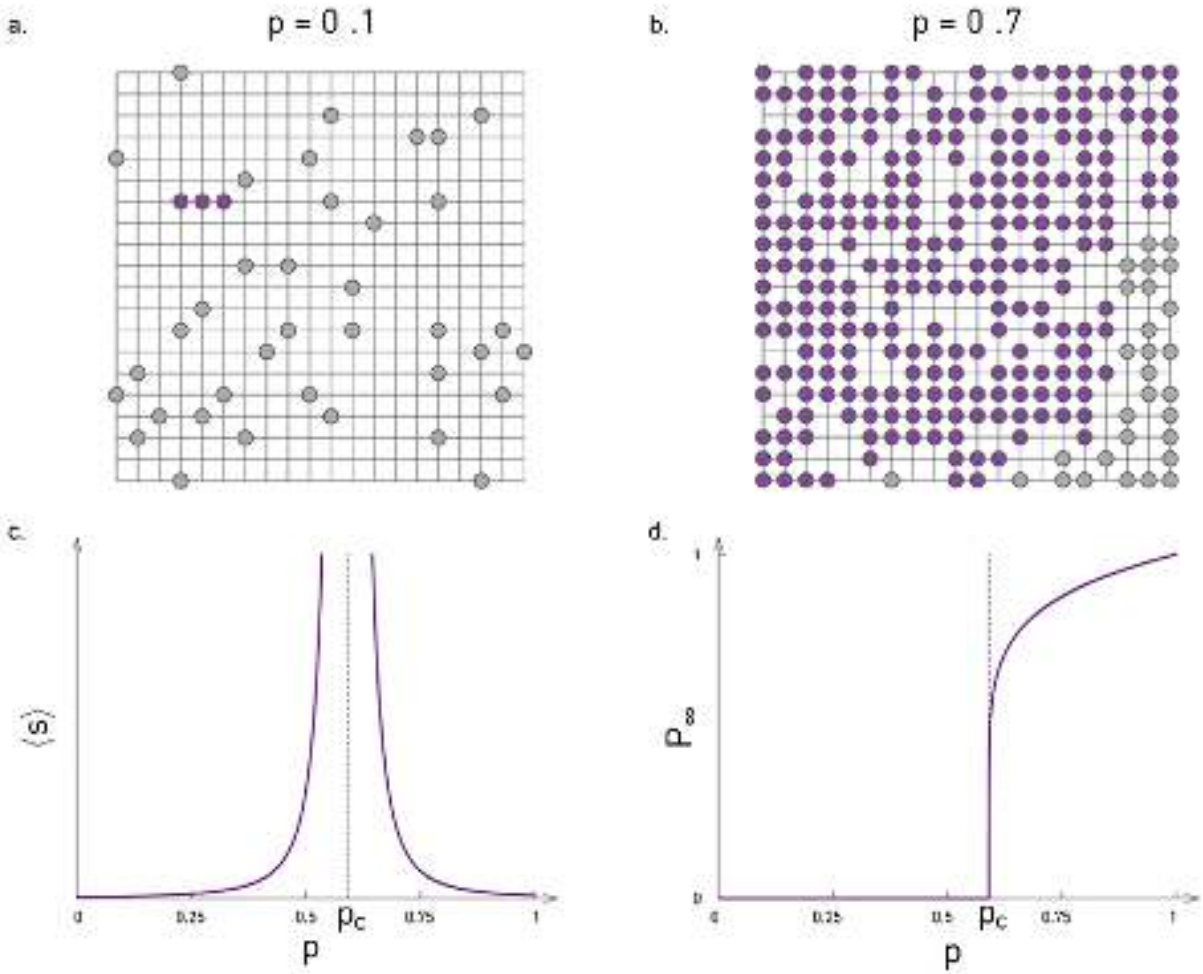
- The intuition on the trend is provided looking at percolation theory
- In it, nodes are added at random with probability p over a n-simensional grid (a lattice)
- The theory shows that as $p > p_c$ a critical value, becomes connected. We can calculate the following:
  1. The expected size of the clusters that are connected is
  $$\mathbb{E}[S] \propto |p - p_c|^{-\gamma_p}$$
  2. The probablity p^∞ that a node chosen at random belongs to the largest cluster is:
  $$p^\infty \propto |p - p_c|^{-\beta_p}$$
  3. The mean distance between nodes is:
  $$\mathbb{E}[\text{distance}(u, v)] \propto |p - p_c|^{-\nu}$$

Percolation theory



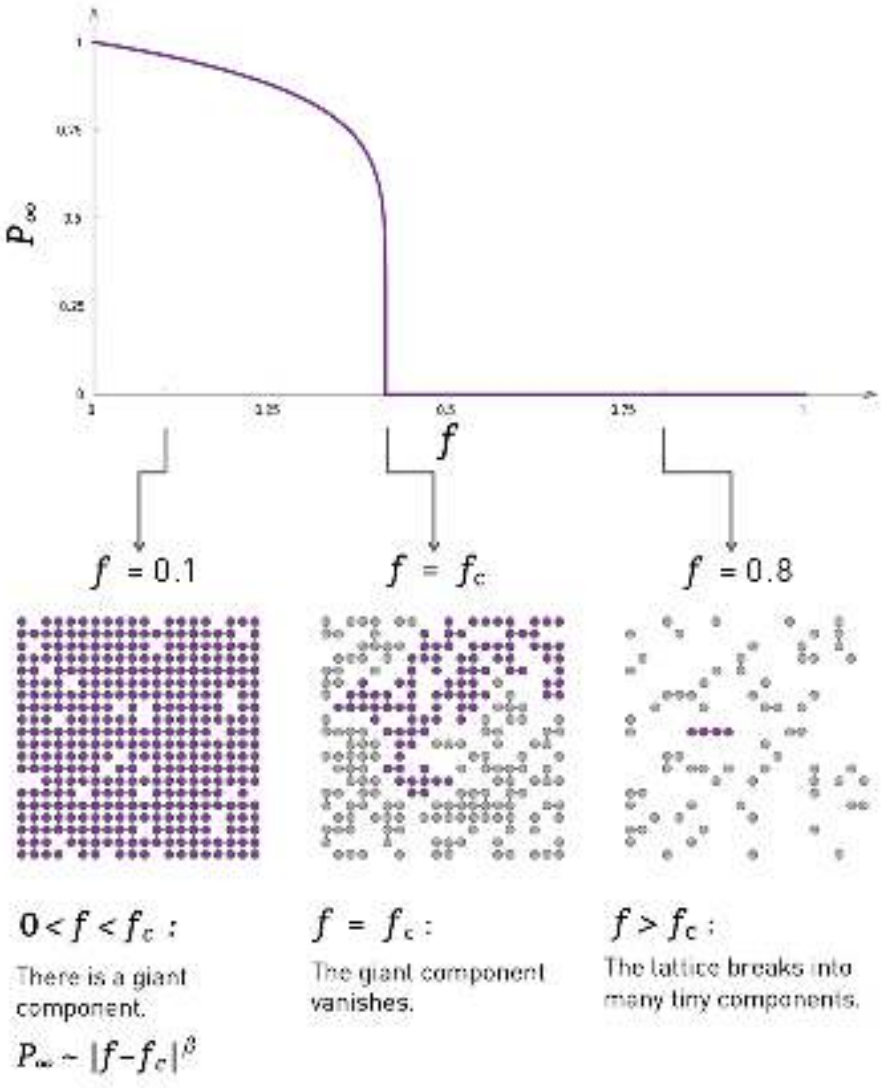
- The intuition on the trend is provided looking at percolation theory
- In it, nodes are added at random with probability p over a n-simensional grid (a lattice)
- The theory shows that as $p > p_c$ a critical value, becomes connected. We can calculate the following:
  1. The expected size of the clusters that are connected is
  $$\mathbb{E}[S] \propto |p - p_c|^{-\gamma_p}$$
  2. The probablity p^∞ that a node chosen at random belongs to the largest cluster is:
  $$p^\infty \propto |p - p_c|^{-\beta_p}$$
  3. The mean distance between nodes is:
  $$\mathbb{E}[\text{distance}(u, v)] \propto |p - p_c|^{-\nu}$$

Percolation theory



The exponents (γ_p, β_p, ν) in the formulas just shown depend on the number of dimensions of the lattice, while p_c depends on its geometry (e.g. a square two dimensional lattice has $p_c \approx 0.6$ while a triangular lattice has $p_c = 0.5$).

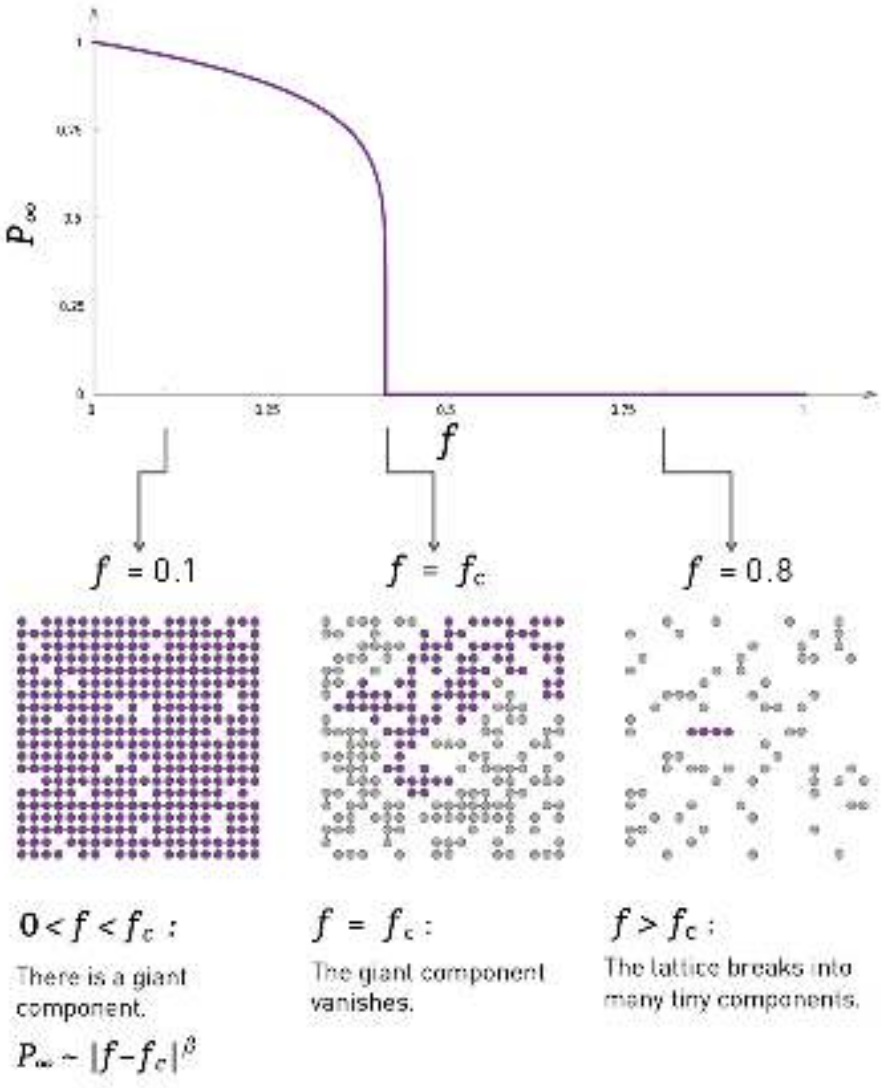
Inverse percolation



- Removal of nodes is the reverse process: we start removing them from the grid
- If we call f the probability of removal, it is clear that the trends should go exactly as the previous one with $p = 1 - f$, and therefore $f_c = 1 - p_c$

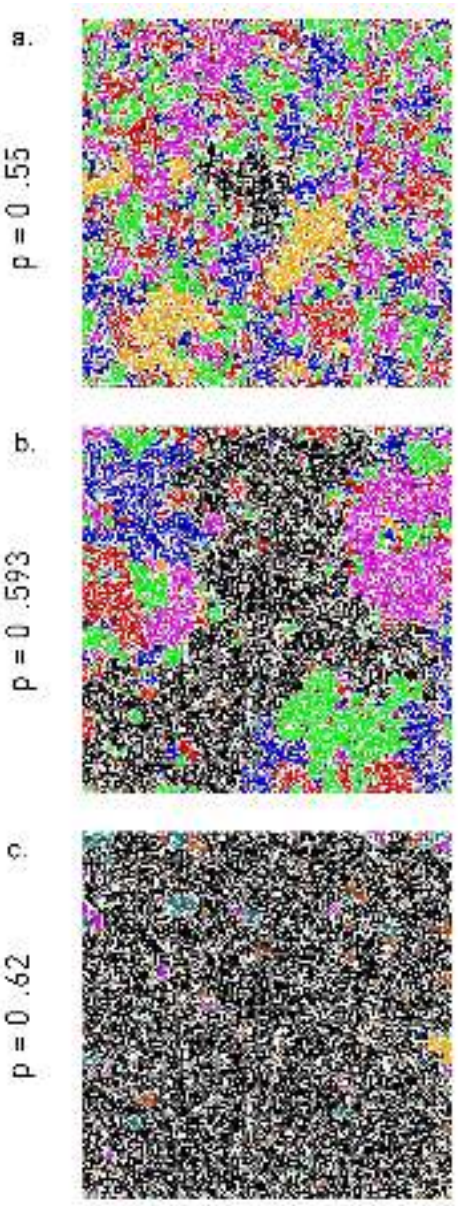
Inverse percolation

Cornell University



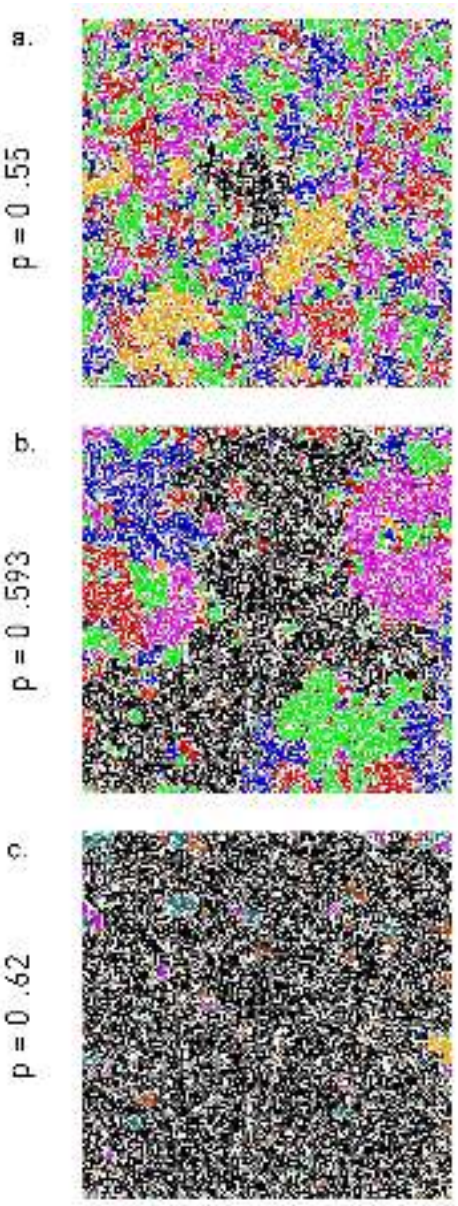
- Removal of nodes is the reverse process: we start removing them from the grid
- If we call f the probability of removal, it is clear that the trends should go exactly as the previous one with $p = 1 - f$, and therefore $f_c = 1 - p_c$

Example: Spread of fire in a forest



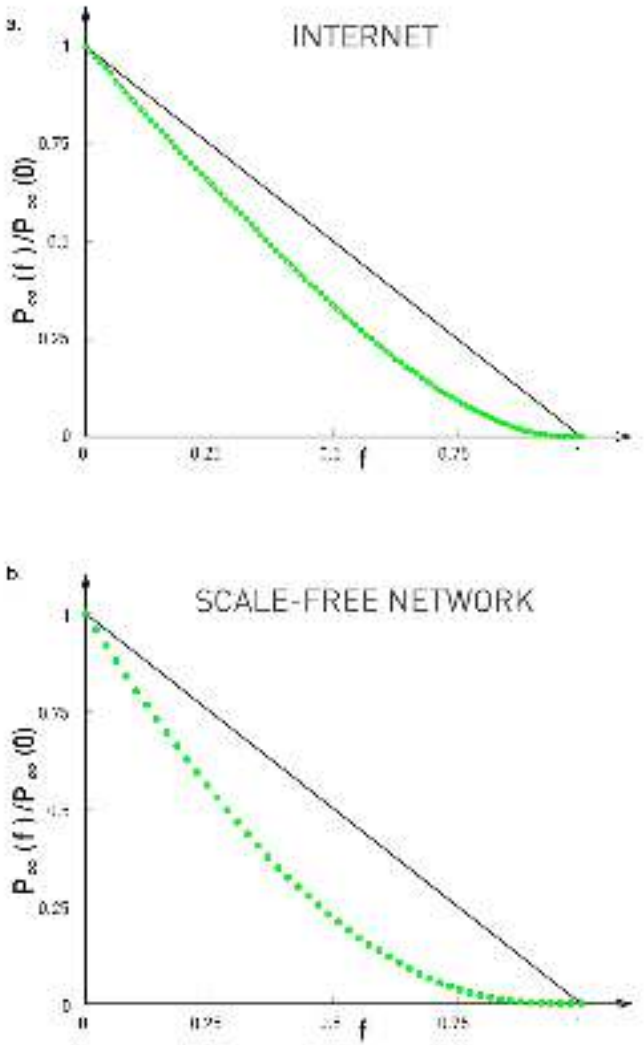
- The spread of a fire in a forest to is a good example of the use of percolation theory
- The "pebble" is a tree and starts, chosen at random, to burn and burning its neighbors
- If the forrest is not very dense p is small $\rightarrow$ the fraction that burns down is small.
- If p is large the largest component tends to burn down.

Example: Spread of fire in a forest



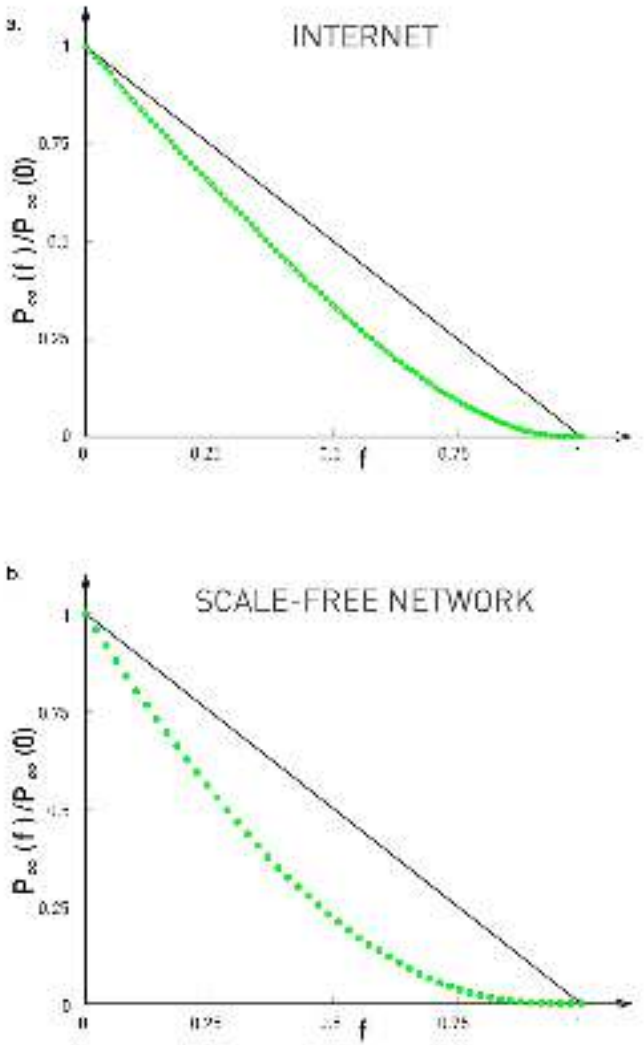
- The spread of a fire in a forest to is a good example of the use of percolation theory
- The "pebble" is a tree and starts, chosen at random, to burn and burning its neighbors
- If the forrest is not very dense p is small $\rightarrow$ the fraction that burns down is small.
- If p is large the largest component tends to burn down.

Robustness of scale-free networks



- We said that regular lattices are very different from scale-free networks
- Empirically, if one starts removing at random switches from the Internet with failure rate f the trend observed is that only when $f \approx 1$ the network shatters
- This trend is also true for general scale free networks
- The random scale free network tested on the left figure has the same $\alpha \approx 2.5$ of the Internet

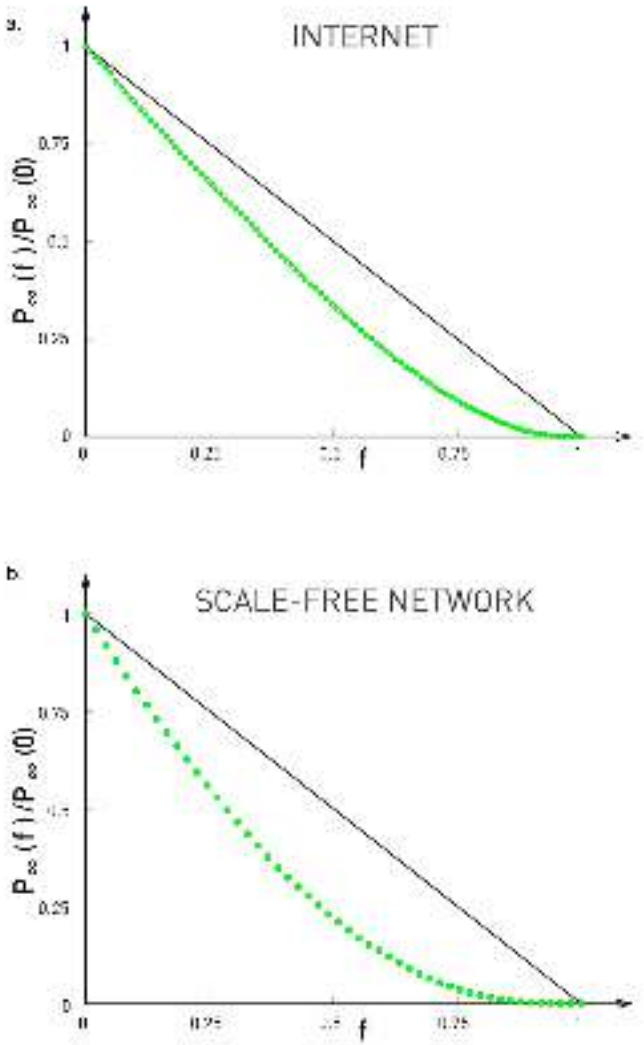
Robustness of scale-free networks



- We said that regular lattices are very different from scale-free networks
- Empirically, if one starts removing at random switches from the Internet with failure rate f the trend observed is that only when $f \approx 1$ the network shatters
- This trend is also true for general scale free networks
- The random scale free network tested on the left figure has the same $\alpha \approx 2.5$ of the Internet

Robustness of scale-free networks

Cornell University



- We said that regular lattices are very different from scale-free networks
- Empirically, if one starts removing at random switches from the Internet with failure rate f the trend observed is that only when $f \approx 1$ the network shatters
- This trend is also true for general scale free networks
- The random scale free network tested on the left figure has the same $\alpha \approx 2.5$ of the Internet

Molloy-Reed criterion

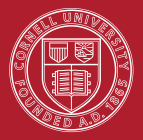
- **Simple observation**: For a network to have a giant component, most nodes that belong to it must be connected to at least two other nodes
- Based on this idea Molloy and Reed showed that a randomly wired network has a giant component if:

$$\frac{\mathbb{E}[d_v^2]}{\mathbb{E}[d_v]} > 2$$

- Using this criterion and an involved mathematical proof, one reaches the following critical failure threshold:

$$f_c = 1 - \frac{1}{\frac{\mathbb{E}[d_v^2]}{\mathbb{E}[d_v]} - 1}$$

Poisson vs Scale Free networks



- In general $\mathbb{E}[d_v^2] = \sigma_{d_v}^2 + (\mathbb{E}[d_v])^2$
- For a Poisson random network, d_v is a Poisson random variable, which has the same mean and variance $pN = \mathbb{E}[d_v]$. Poisson random variable the mean and variance are equal, and therefore $\mathbb{E}[d_v^2] = \mathbb{E}[d_v] + (\mathbb{E}[d_v])^2 = \mathbb{E}[d_v](1 + \mathbb{E}[d_v])$, which implies the critical failure rate is:

$$f_c^{ER} = 1 - \frac{1}{\mathbb{E}[d_v]}$$

- Note that in the Poisson regime $p = c/N$, hence $\mathbb{E}[d_v] = c$ and $f_c^{ER} = 1 - c^{-1}$, i.e. the lower is c the easier is to attack the network effectively, as one would expect.

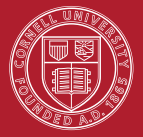
Poisson vs Scale Free networks

- In general $\mathbb{E}[d_v^2] = \sigma_{d_v}^2 + (\mathbb{E}[d_v])^2$
- For a Poisson random network, d_v is a Poisson random variable, which has the same mean and variance $pN = \mathbb{E}[d_v]$. Poisson random variable the mean and variance are equal, and therefore $\mathbb{E}[d_v^2] = \mathbb{E}[d_v] + (\mathbb{E}[d_v])^2 = \mathbb{E}[d_v](1 + \mathbb{E}[d_v])$, which implies the critical failure rate is:

$$f_c^{ER} = 1 - \frac{1}{\mathbb{E}[d_v]}$$

- Note that in the Poisson regime $p = c/N$, hence $\mathbb{E}[d_v] = c$ and $f_c^{ER} = 1 - c^{-1}$, i.e. the lower is c the easier is to attack the network effectively, as one would expect.

Poisson vs Scale Free networks



Cornell University

- For a scale free network instead:

$$f_c = 1 - \frac{1}{\left|\frac{2-\alpha}{3-\alpha}\right| A - 1}$$

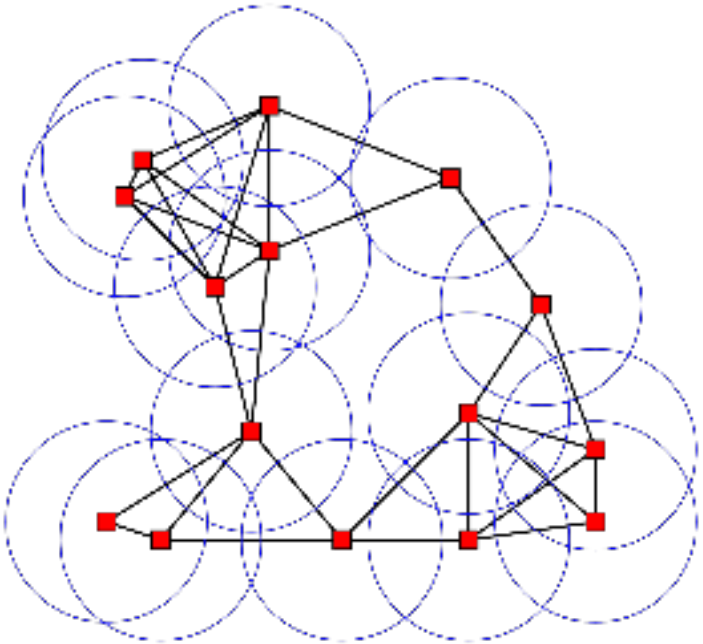
where the A depends on $d_{\min}$ $d_{\max}$ and α and N as follows:

$$A = \begin{cases} d_{\min} & \alpha > 3 \\ d_{\min}^{\alpha-2} d_{\max}^{3-\alpha} & 2 < \alpha < 3 \\ d_{\max}, & 1 < \alpha < 2 \end{cases}$$

where we can also use the relationship $d_{\max} \approx d_{\min} N^{\frac{1}{\alpha-1}}$ to understand the trend with N

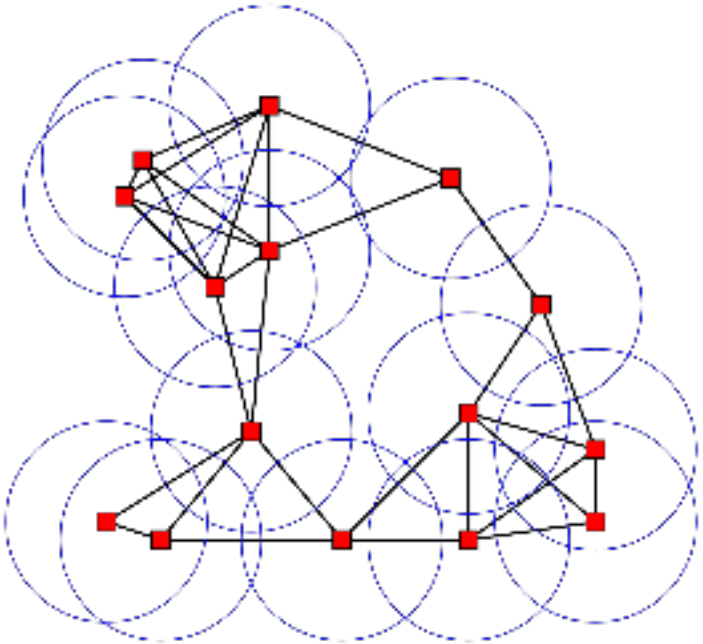
- The important point is that for $\alpha < 3$ the value of $\mathbb{E}[d_v^2]/\mathbb{E}[d_v] \to +\infty$ as $N \to +\infty$ which means that $f_c \to 1$, as seen in real networks with a scale free degree distribution. It is hard to break them!

Random Geometric graphs



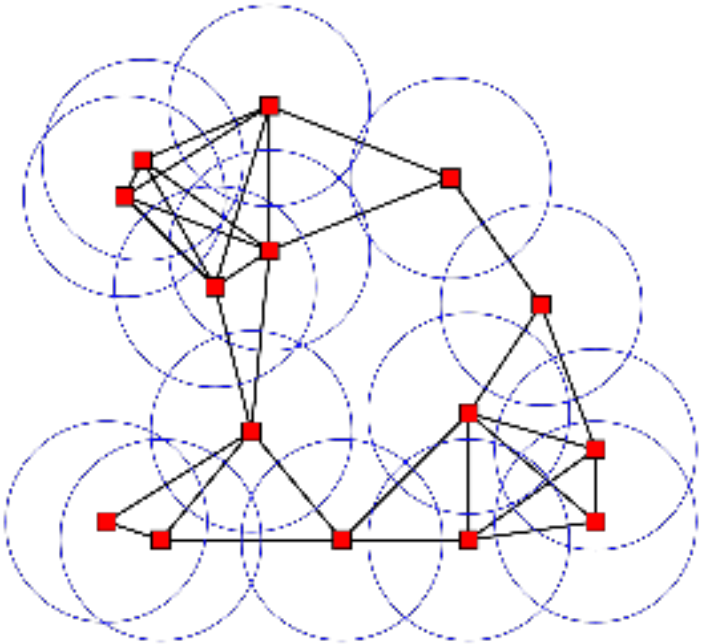
- This is another interesting case studied to understand coverage in a network embedded in space
- The nodes are placed at random according to a 2 dimensional Poisson point process. If the area is one, the density α nodes per unit area.
- Each node has a disc of radius r around them, and an edge joins them if the centers are separated by $< R = 2r$: does a giant component emerge?
- The answer is yes, for an appropriate α and/or r

Random Geometric graphs



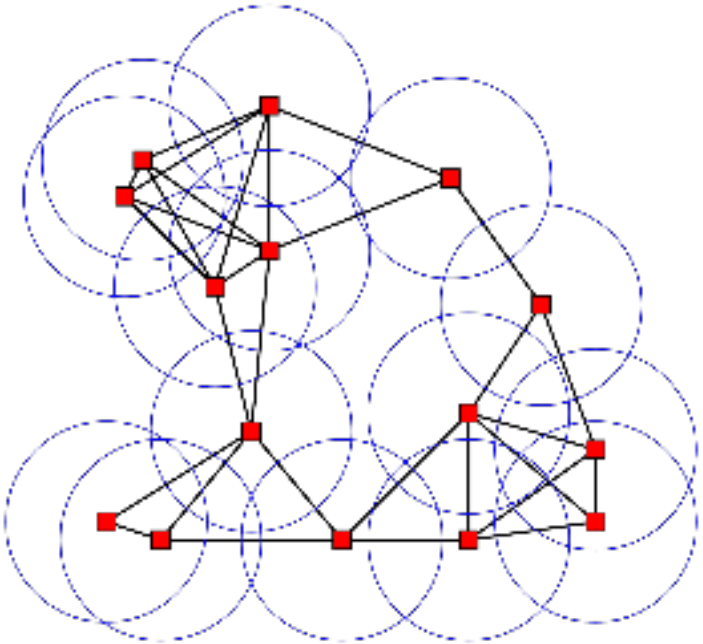
- This is another interesting case studied to understand coverage in a network embedded in space
- The nodes are placed at random according to a 2 dimensional Poisson point process. If the area is one, the density α nodes per unit area.
- Each node has a disc of radius r around them, and an edge joins them if the centers are separated by $< R = 2r$: does a giant component emerge?
- The answer is yes, for an appropriate α and/or r

Random Geometric graphs



- This is another interesting case studied to understand coverage in a network embedded in space
- The nodes are placed at random according to a 2 dimensional Poisson point process. If the area is one, the density α nodes per unit area.
- Each node has a disc of radius r around them, and an edge joins them if the centers are separated by $< R = 2r$: does a giant component emerge?
- The answer is yes, for an appropriate α and/or r

Random Geometric graphs



- This is another interesting case studied to understand coverage in a network embedded in space
- The nodes are placed at random according to a 2 dimensional Poisson point process. If the area is one, the density α nodes per unit area.
- Each node has a disc of radius r around them, and an edge joins them if the centers are separated by $< R = 2r$: does a giant component emerge?
- The answer is yes, for an appropriate α and/or r

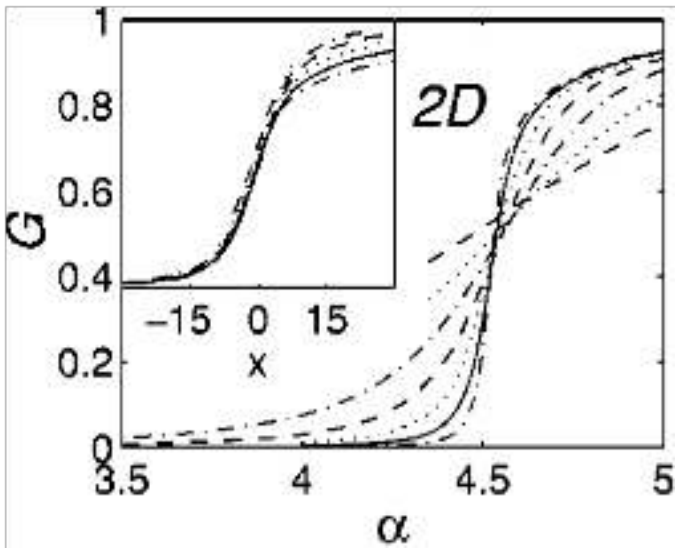
Phase transition for RGG

Cornell University

- In 2D a disk area[1] is $V = \pi r^2$. The area where two points have to fall to connect is $V_e = \pi R^2$, which is , called exclusion area.
- Since the total volume of the box we throw points in is 1, falling in an area $A \leq 1$ has probability A. Hence, the probability that a point's coordinate falls in the exclusion area of another node is equal V_e.

In the Poisson point process α is the average number of points per unit area. The phase transition occurs when the average number of points per unit area is equal to the expected number of points in the exclusion area NV_e:

$$\alpha = NV_e = N\pi R^2 \quad \Rightarrow \quad R = \sqrt{\frac{\alpha}{N\pi}}$$



Fraction of nodes in the giant component.

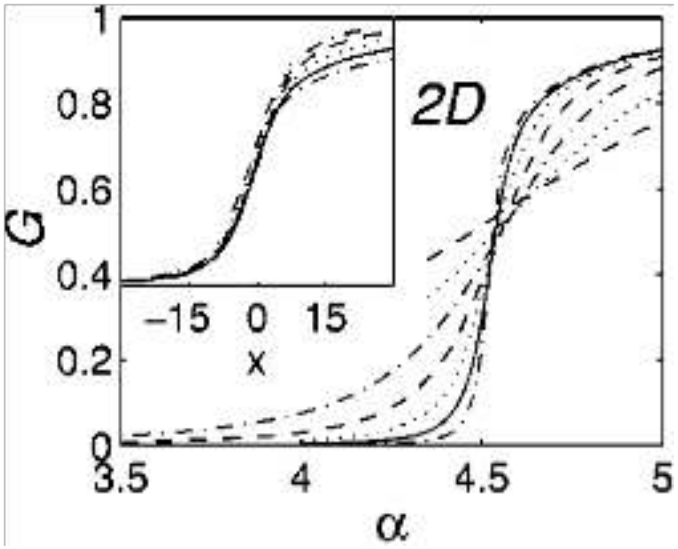
[1] In general for a n dimensions space the volume is $V = \pi^{n/2}/\Gamma(\frac{n+2}{2})$

Phase transition for RGG

- In 2D a disk area[1] is $V = \pi r^2$. The area where two points have to fall to connect is $V_e = \pi R^2$, which is , called exclusion area.
- Since the total volume of the box we throw points in is 1, falling in an area $A \leq 1$ has probability A. Hence, the probability that a point's coordinate falls in the exclusion area of another node is equal V_e.

In the Poisson point process α is the average number of points per unit area. The phase transition occurs when the average number of points per unit area is equal to the expected number of points in the exclusion area NV_e:

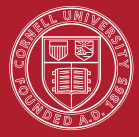
$$\alpha = NV_e = N\pi R^2 \quad \Rightarrow \quad R = \sqrt{\frac{\alpha}{N\pi}}$$



Fraction of nodes in the giant component.

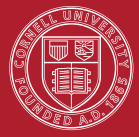
[1] In general for a n dimensions space the volume is $V = \pi^{n/2}/\Gamma(\frac{n+2}{2})$

Outline



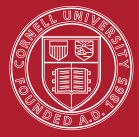
Cornell University

Conclusions



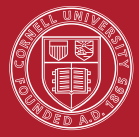
- In this lecture we discussed scale-free networks and their unique features
- We introduced the so called *preferential attachment model*, a random generative model introduced by Barabási and Albert
- We also introduced random geometric graphs and the critical parameters threshold that makes them connected.
- We introduced the notion of network robustness
- This lecture concludes the first module of the course where we focused on better understanding the graphs themselves.

Conclusions



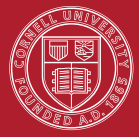
- In this lecture we discussed scale-free networks and their unique features
- We introduced the so called *preferential attachment model*, a random generative model introduced by Barabási and Albert
- We also introduced random geometric graphs and the critical parameters threshold that makes them connected.
- We introduced the notion of network robustness
- This lecture concludes the first module of the course where we focused on better understanding the graphs themselves.

Conclusions



- In this lecture we discussed scale-free networks and their unique features
- We introduced the so called *preferential attachment model*, a random generative model introduced by Barabási and Albert
- We also introduced random geometric graphs and the critical parameters threshold that makes them connected.
- We introduced the notion of network robustness
- This lecture concludes the first module of the course where we focused on better understanding the graphs themselves.

Conclusions



- In this lecture we discussed scale-free networks and their unique features
- We introduced the so called *preferential attachment model*, a random generative model introduced by Barabási and Albert
- We also introduced random geometric graphs and the critical parameters threshold that makes them connected.
- We introduced the notion of network robustness
- This lecture concludes the first module of the course where we focused on better understanding the graphs themselves.