Statistical Network Analysis

Prof. Dr. Ingo Scholtes

Chair of Machine Learning for Complex Networks
Center for Artificial Intelligence and Data Science (CAIDAS)
Julius-Maximilians-Universität Würzburg
Würzburg, Germany

ingo.scholtes@uni-wuerzburg.de

Lecture 09
Scale-Free Networks

December 14, 2022

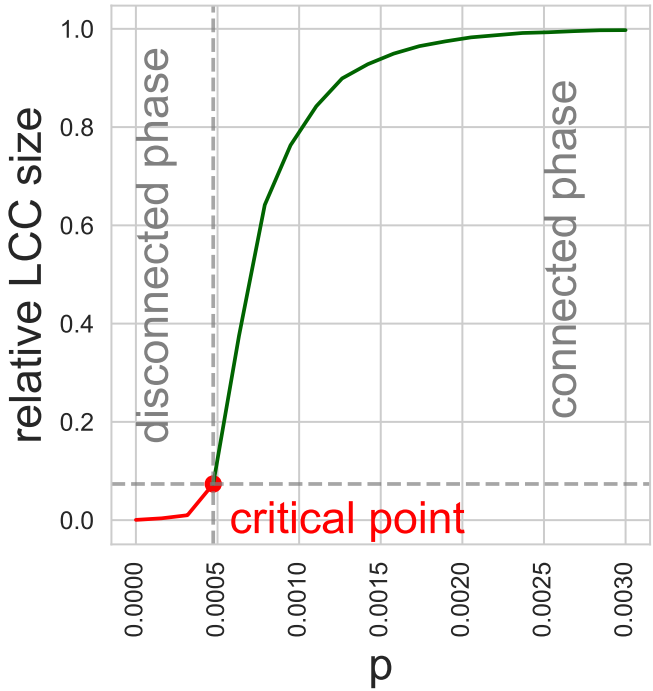
Notes:

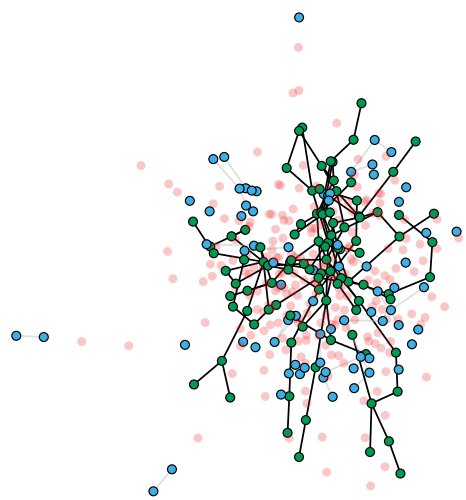
- **Lecture L09:** Scale-free Networks 14.12.2022
- **Educational objective:** We explore how the degree distribution influences the robustness of random networks. We further introduce scale-free networks and discuss fallacies in the application of ensemble-based methods to real systems.
  - Scale-free networks
  - Expected Properties of Random Scale-Free Networks
  - Limitations of Ensemble-based Approaches

- **Exercise 07:** Scale-free Networks due 21.12.2022

Motivation

- we used **probability generating functions** to study the expected connected component size in random graphs
- we discovered a **sudden transition** between a disconnected and connected phase
- we adapted our framework to study the effect of **random node failures** on the giant connected component
- giant connected component suddenly disappears if node failure probability exceeds **critical threshold**





Notes:

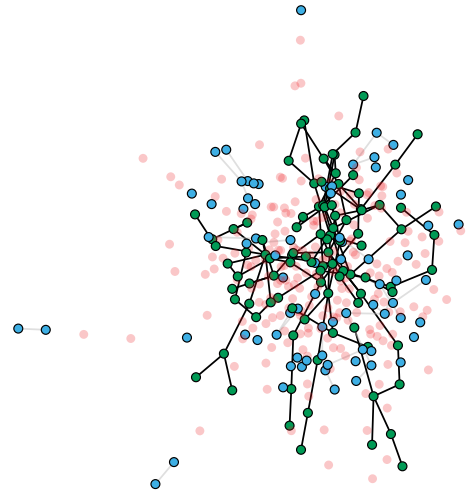
- We start with a brief reminder of the previous lecture, where we used probability generating functions to calculate the expected connected component size in random networks with arbitrary degree distributions.
- Using a limit argument, we found a critical point for the ratio between the first two moments of the degree distribution at which a giant connected component suddenly emerges. In random Erdős-Rényi networks, this critical point translates to a critical link probability p above which we expect (asymptotically, i.e. for $n \to \infty$) a giant connected component that connects the majority of nodes.
- We further applied the same framework to calculate the expected size of the surviving connected components in a network if a random fraction of nodes are removed. Not surprisingly, we again found an asymptotic critical point for the failure probability above which the giant connected component is destroyed.
- In today's lecture, we will further explore how the degree distribution influences this critical behaviour in networks, which will motivate an interesting class of networks that is often found in real-world systems.

Robustness of random Erdős-Rényi networks

- for **random Erdő-Rényi networks** we derived the critical failure probability → L08

$$\frac{\langle k^2 \rangle}{\langle k \rangle} = \frac{np + n^2 p^2}{np} = 1 + np$$

$$q_c = 1 - \left(\frac{\langle k^2 \rangle}{\langle k \rangle} - 1 \right)^{-1} = 1 - \frac{1}{np}$$



failure probability $q = 0.6 < q_c$ for $G(n, p)$ network with $np = 3$

- for a giant connected component to survive we need – on average – one “surviving” neighbour per node

Notes:

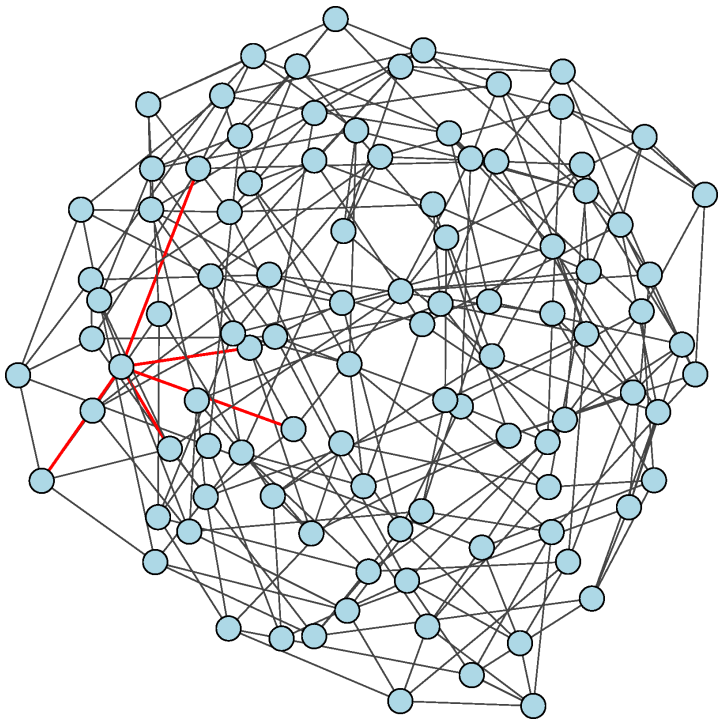
- We briefly repeat the argument for the critical failure probability in random Erdős-Rényi networks, which is again due to the ratio between the first two raw moments of the degree distribution (here, a Poisson distribution in case of sparse networks).
- We can intuitively interpret this critical point: in a random network, a giant connected component survives if – on average – each node has at least one surviving neighbor. In the example, where the expected degree is three, this means that two out of three nodes can fail randomly, which brings the remaining network to the critical threshold with a mean degree of one.
- However, it is important to note that this is a property of the Poisson distribution, i.e. the rule that we need at least one neighbour per node on average to have a surviving giant connected component does not generalize to other degree distributions.

Robustness of random k-regular networks

- for **random k-regular networks**, i.e. networks in which all nodes have the same degree $k \equiv const$ we have

$$\frac{\langle k^2 \rangle}{\langle k \rangle} = \frac{k^2}{k} = k$$

$$q_c = 1 - \frac{1}{k-1}$$



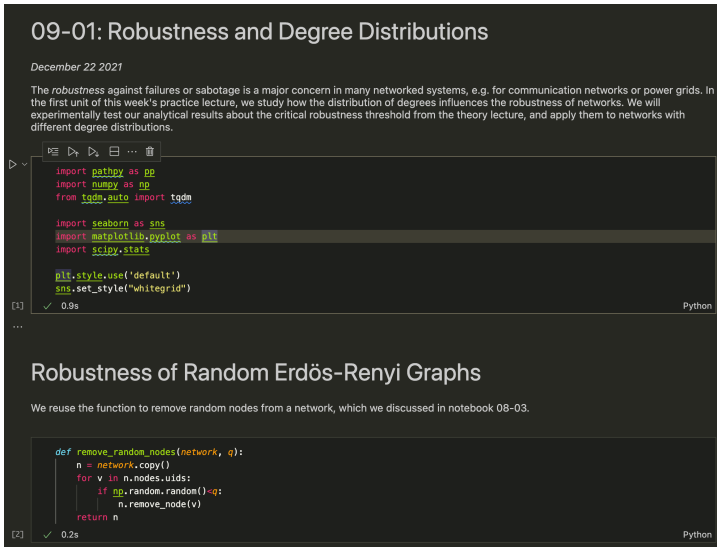
- compute the critical failure rate for $k = 2, k = 3, k = 4, \ldots$ and compare it to Erdős-Rényi networks

Notes:

- To better understand this, let us consider k-regular networks. Since the variance of the degree distribution is zero we have $\langle k^2 \rangle = \langle k \rangle^2 = k^2$ → L07
- Using our formula for the critical failure probability, we find $1 - \frac{1}{k-1}$, i.e. q_c is inversely proportional to the degree k (minus one).
- Remember that the Molloy-Reed criterion is $\frac{\langle k^2 \rangle}{\langle k \rangle} > 2$, i.e. for a k-regular network with $k = 1$ we are still below the critical threshold for the emergence of a giant connected component, while for $k = 2$ we are just at the point of the phase transition.
- For $k = 1$ (where we have no giant connected component), the critical failure rate $q_c = 1 - \frac{1}{0}$ is undefined. For $k = 2$, we find $q_c = 0$. We are just at the critical point where a giant connected component emerges, so we cannot afford any node to fail!
- For $k = 3$, we find $q_c = 0.5$, while for $k = 4$, we find $q_c = \frac{2}{3}$ (the same failure probability that we found for an Erdős-Rényi network with $np = 3$.
- This is interesting because in a k-regular network k trivially is also the mean degree of the network. For random Erdős-Rényi networks, we expect a giant connected component for $\langle k \rangle > 1$, while here we have $\langle k \rangle = k > 2$. This results from the zero variance in the degree distribution, which suggests that a larger variance makes networks more robust.

Practice Session

- we experimentally validate critical node failure probabilities for networks with different degree distributions
- we explore the effect of degree heterogeneity on robustness



observation

networks with broader degree distributions tend to be more robust against random node failures

practice session

see notebook 09-01 in `gitlab` repository at

→ `https://gitlab.informatik.uni-wuerzburg.de/ml4nets_notebooks/2022_wise_sna_notebooks`

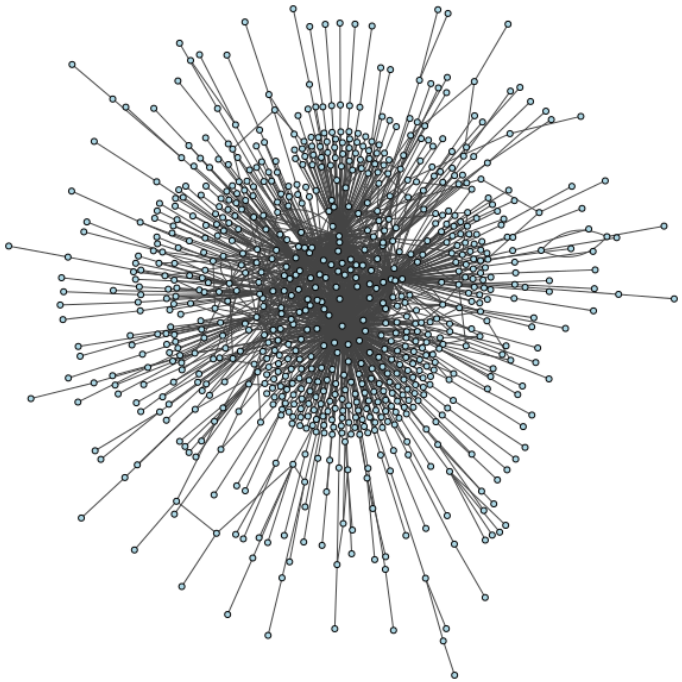
Notes:

- This observation leads us to an interesting follow-up question: what is the robustness of networks with a broad degree distribution, i.e. networks where degrees have a large variance than in random Erdős-Rényi networks?
- Intuitively, we expect that broader degree distributions lead to two competing effects:
    1. On the one hand, it is unlikely that the most connected nodes fail at random, i.e. we expect the effect of randomly failing nodes to be comparably small.
    2. On the other hand, the impact of a single failure of the most connected nodes is much higher than in a random Erdős-Rényi network.
- Our finding about the relation between the surviving connected component size and the moments of the degree distribution allows us to analytically calculate how these two competing effects affect network robustness.
- In the practice session, we experimentally explore this.

Scale-free networks

- for $k \in \{1, \ldots, N\}$ and exponent $\gamma > 1$, a **power law distribution** is characterized by the probability mass function

$$P(k) = k^{-\gamma} \left(\sum_{i=1}^{N} i^{-\gamma} \right)^{-1}$$



- power laws give rise to **scale invariance**

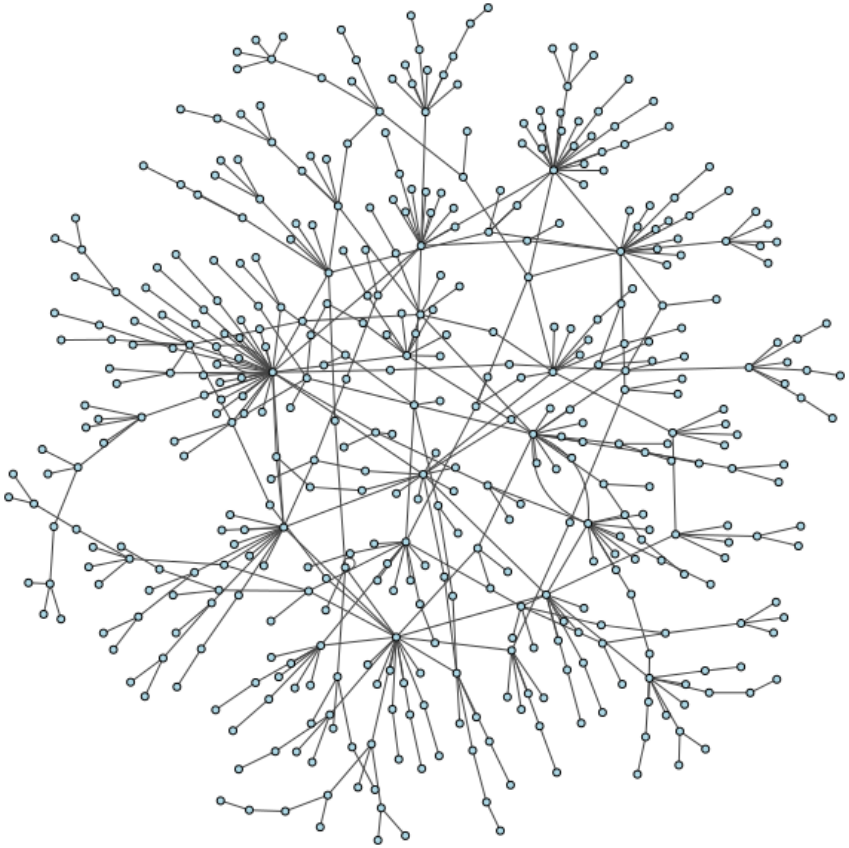
$$P(\alpha k) \propto (\alpha k)^{-\gamma} \propto \alpha^{-\gamma} P(k) \propto P(k)$$

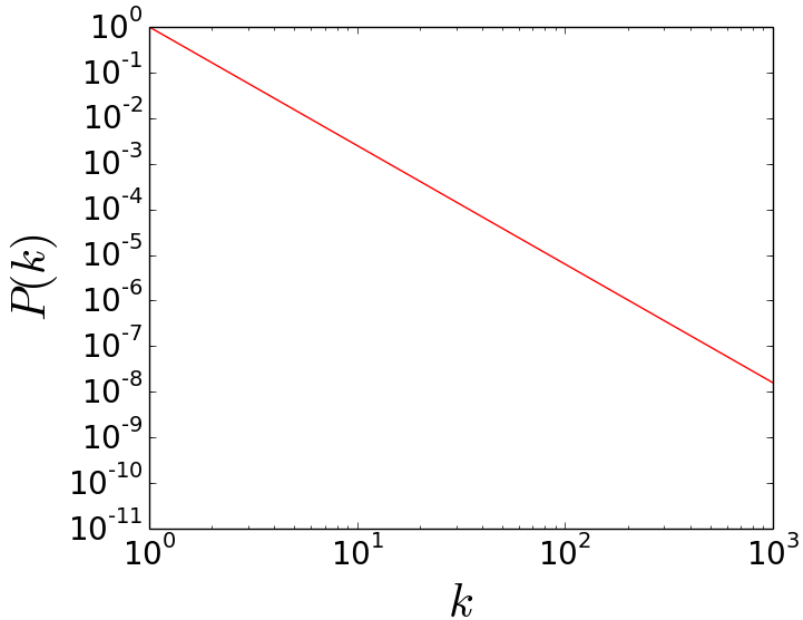
- networks with **scale-invariant** degree distribution are called **scale-free networks**

Notes:

- A frequently studied class of networks with broad degree distributions are those, where the degree distributions follow a power law like the one above. This is only one example of a “broad distributions”, others include
  - Pareto distribution
  - Zipf-Mandelbrot distribution
  - Yule-Simon distribution
  - Weibull distribution
- Note that we have again excluded $k = 0$, since $0^{-\gamma} = \frac{1}{0^{\gamma}}$ is undefined
- The power law gives rise to an effect called scale invariance, i.e. the shape of the distribution looks the same, independent of the value k. Note that this is different for other distributions (like the Poisson or the Normal distribution) which we have seen before. Here, due to the exponential term we do not have scale invariance, i.e. multiplying the degree k with a constant will overproportionally change the probability to observe a node with degree k.
- Due to the lack of a dependence of the distribution on a certain “scale of degrees” networks with a power law degree distribution are sometimes called scale-free networks.

Scale-free network with $\gamma = 2.6$





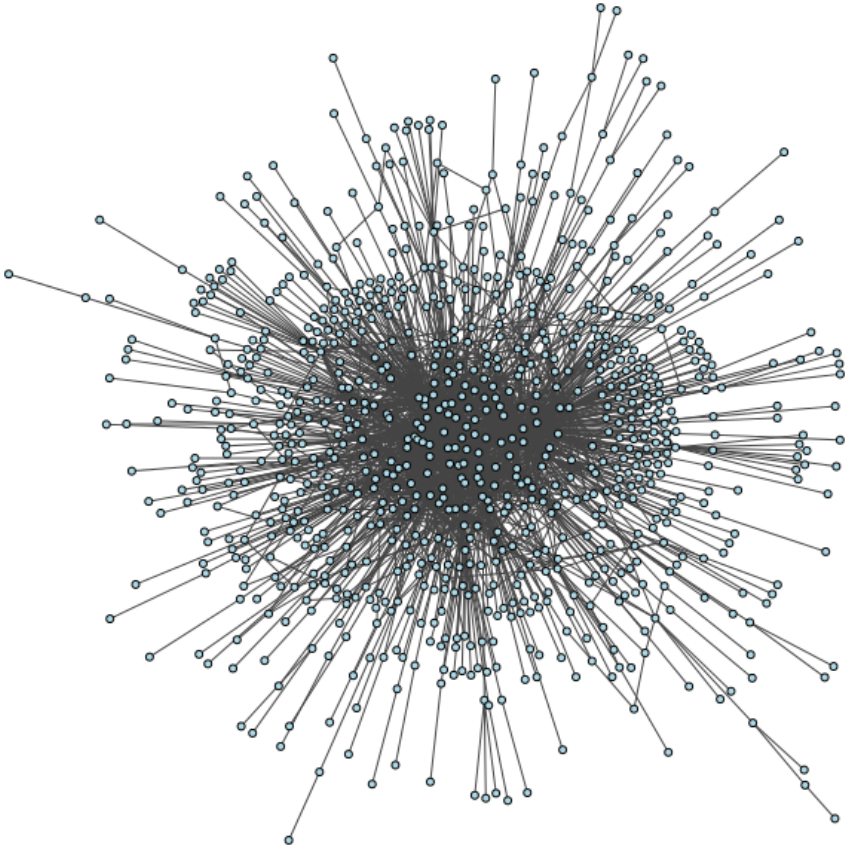
example

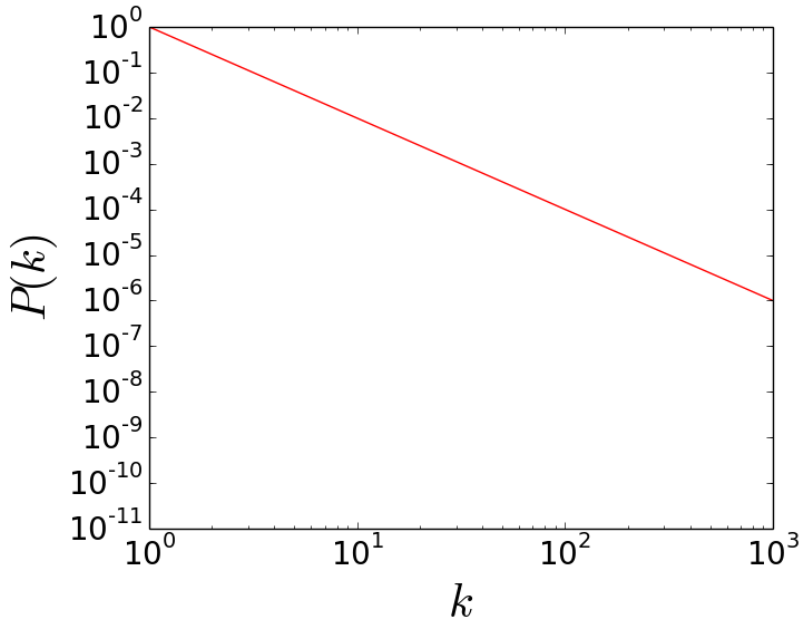
random microstate from Molloy-Reed ensemble (left) with degree distribution given by power law with $\gamma = 2.6$ (right)

Notes:

- The single parameter γ of the power law distribution has important implications for the distribution of degrees (and thus the network topologies generated by a Molloy-Reed model).
- The network on the left shows (the giant connected component of) a random realization generated by the Molloy-Reed model, using a Zipf degree distribution with $\gamma = 2.6$. The underlying probability mass function (in logarithmic scale) is shown on the right.
- For $\gamma = 2.6$ we find that the heterogeneity of degrees is rather mild, i.e. the probability of nodes with very large degree (say 100 in this example) is still very small (about 10^{-5}).

Scale-free network with $\gamma = 2$





example

random microstate from Molloy-Reed ensemble (left) with degree distribution given by power law with $\gamma = 2.0$ (right)

Notes:

- If we decrease the exponent, the distribution becomes broader, which translates to the fact that we generate nodes with very high degree with high probability.
- For $\gamma = 2$ the probability to observe a node with large degree (e.g. 100) is now almost two orders of magnitude larger than before (approx. 10^{-3}).
- The slope of the probability mass function flattens out, i.e. we have comparably large probabilities for very high degree nodes.
- In other words: a **decrease** of the exponent γ translates to an **increase** in the heterogeneity of the degree distribution, i.e. the distribution becomes broader.

Infinite limit: Zeta degree distribution

- for large networks, we approximate finite power law degree distribution by the infinite **Zeta distribution**

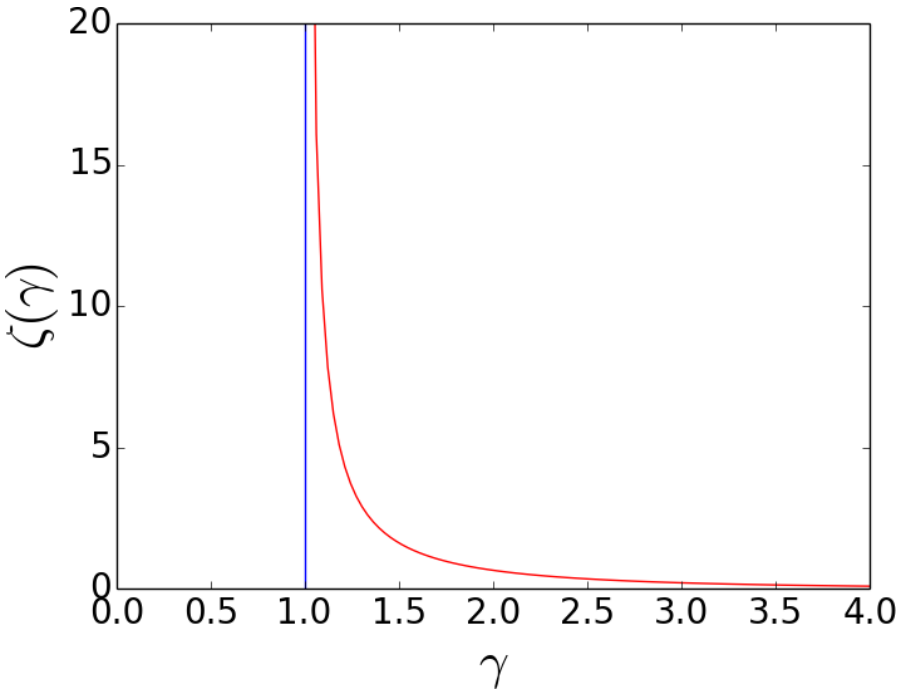
Zeta distribution

- discrete distribution on $\mathbb{N}$
- for **exponent** $\gamma > 1$ and degree $k \in \mathbb{N}$ probability mass function is given as

$$P(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}$$

- with $\zeta(\gamma)$ the **Riemann zeta function**, i.e.

$$\zeta(\gamma) = \sum_{i=1}^{\infty} i^{-\gamma}$$



Riemann zeta function

plot of Riemann zeta function $\zeta(\gamma)$ for exponents $\gamma > 1$

Notes:

- Just like our findings about the emergence of the gcc, our robustness analysis was based on a convergence argument for infinite networks. We thus need to generalise the power law distribution for networks with an infinite number of nodes n.
- For $n \to \infty$, the normalization constant $\sum_{k=1}^{N} k^{-\gamma}$ turns into the **infinite series** $\sum_{i=1}^{\infty} k^{-\gamma}$, which is the Riemann zeta function $\zeta(\gamma)$.
- The plot on the right shows the Zeta function in the range between zero and four. The blue vertical line indicates the point $x = 1$ at which $\zeta(x)$ (the red line) diverges, i.e. the normalization constant would be infinite.
- You may remember certain values of $\zeta(x)$ from you math classes:
- $\zeta(2) = \sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$
- $\zeta(4) = \sum_{i=1}^{\infty} \frac{1}{i^4} = \frac{\pi^4}{90}$
- $\zeta(6) = \sum_{i=1}^{\infty} \frac{1}{i^6} = \frac{\pi^6}{945}$
- The Riemann zeta function continues these well-known (convergent) series in continuous space.

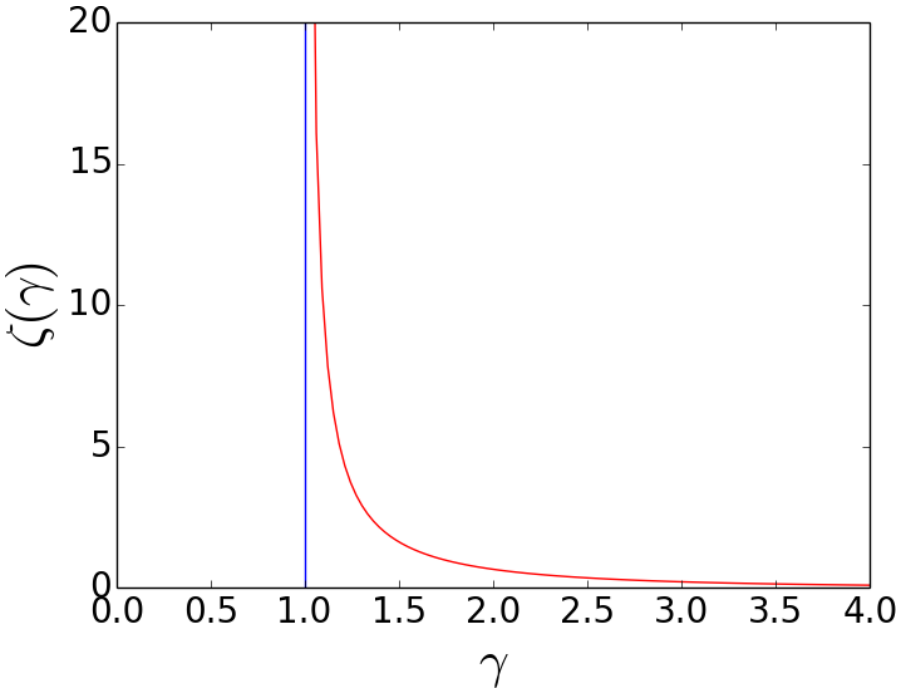
Moments of the Zeta distribution

- for the Zeta distribution the *m*-**th raw moment** is given as

$$\langle k^m \rangle = \sum_{k=1}^{\infty} k^m P(k) = \sum_{k=1}^{\infty} \frac{k^{m-\gamma}}{\zeta(\gamma)}$$

$$= \frac{\zeta(\gamma - m)}{\zeta(\gamma)}$$

- $\langle k \rangle$ finite for $\gamma \in (2, \infty)$
- $\langle k^2 \rangle$ finite for $\gamma \in (3, \infty)$



Riemann zeta function

plot of Riemann zeta function $\zeta(\gamma)$ for exponents $\gamma > 1$

scale-free networks

we say that a network with zeta degree distribution with exponent γ is **scale-free** iff $\gamma \in (2, 3)$

Notes:

- Our analysis of connectivity and robustness is based on the raw moments of a degree distribution. These are easy to calculate for the Zeta distribution because of the following relation between the m-th raw moment and the values of the Riemann Zeta function. Since $k^{m-\gamma} = k^{-(\gamma-m)}$ we find that

$$\langle k^m \rangle \sum_{k=1}^{\infty} \frac{k^{m-\gamma}}{\zeta(\gamma)} = \frac{\zeta(\gamma - m)}{\zeta(\gamma)}$$

- This implies that any random network whose degree distribution follows a Zeta degree distribution with parameter $\gamma > 2$ is sparse, i.e. the mean degree is finite for $n \to \infty$.
- Importantly, the second raw moment (and thus the variance) is only finite for $\gamma > 3$. For $\gamma \in (2, 3)$ we thus have an interesting regime where:
  - The mean degree is finite, i.e. the number of links scales as a constant with network size
  - The second raw moment (and thus the variance) is infinite, i.e. arbitrarily large degrees can occur with "reasonably large" probability
- In some works, the term scale-free networks specifically refers to networks with a power law distribution and exponent $\gamma \in (2, 3)$, i.e. power law networks with infinite variance and finite mean degree.

Robustness of scale-free networks

- for **scale-free networks** with degree distribution exponent γ

$$\frac{\langle k^2 \rangle}{\langle k \rangle} = \frac{\zeta(\gamma - 2)}{\zeta(\gamma - 1)}$$

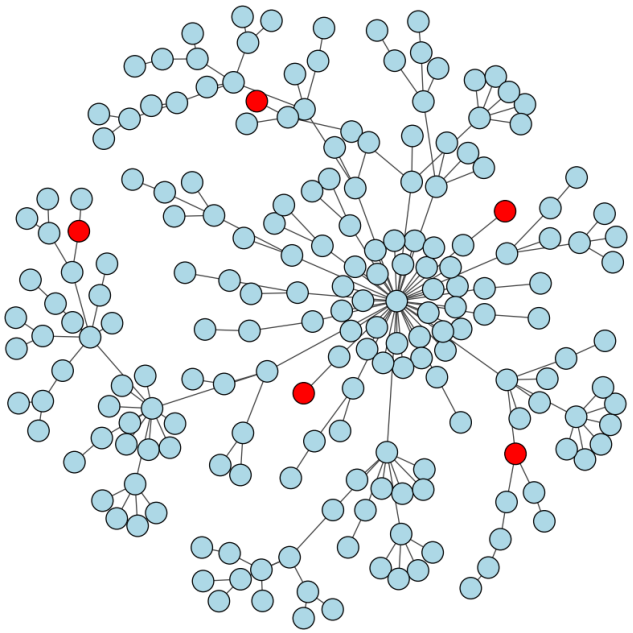
- **critical failure probability** is

$$q_c = 1 - \left(\frac{\zeta(\gamma - 2)}{\zeta(\gamma - 1)} - 1 \right)^{-1}$$

- for $\gamma \in (2, 3)$ we have $\zeta(\gamma - 2) \to \infty$ while $\zeta(\gamma - 1) \to$ const and thus

$$q_c \to 1$$

- we expect random scale-free networks to be **super-robust against random failures**



Notes:

- We can now analytically calculate the critical failure probability (again assuming random node failures with uniform probability q).
- The example on the right shows an example microstate generated using the Molloy-Reed model for $\gamma = 2.1$ and $n = 100$ nodes, where nodes have a uniform failure probability of $q = 0.05$.
- We can apply the expression for the critical failure rate q_c from slide 14:

$$q_c = 1 - \left(\frac{\langle k^2 \rangle}{\langle k \rangle} - 1 \right)^{-1}$$

- Substituting the raw moments of the Zeta distribution, we have

$$\frac{\langle k^2 \rangle}{\langle k \rangle} = \frac{\zeta(\gamma - 2)}{\zeta(\gamma - 1)}$$

- From the convergence behaviour of the Riemann zeta function, we know that $\zeta(\gamma)$ diverges for $\gamma < 1$
- This implies that for $\gamma < 3$ we have $\zeta(\gamma - 2) \to \infty$, while $\zeta(\gamma - 1)$ converges to a constant value. We thus find a special case where the Molloy-Reed criterion assumes infinite values. This implies that the critical failure rate is one, i.e. in the limit of infinite network size the giant connected component of scale-free networks cannot be destroyed unless all nodes fail!

Diameter of scale-free networks

- number of **neighbours at distance two** of randomly chosen node v

$$G_0(G_1(x))$$

- number of **neighbours at distance three** of node v

$$G_0(G_1(G_1(x)))$$

- for **diameter** D in network with n nodes one can derive

$$D = 1 + \frac{\log(n) - \log(\langle k \rangle)}{\log(\langle k^2 \rangle) - log(\langle k \rangle)}$$

- for **scale-free networks with** $\gamma \in (2, 3)$: inclusion of exponential cutoff in tail of degree distribution yields → R Cohen and S Havlin, 2003

$$D \approx \frac{\log\log(n)}{\log(\gamma - 2)}$$

Notes:

- Let us close by studying the diameter of scale-free networks. We have only addressed this for Erdős-Rényi networks, where the analysis is particularly simple due to $G_0 = G_1$ (cf. L06, slide 13/14). In the general case where $G_0 \neq G_1$, we can analyse the diameter using generating functions. We can use the repeated composition to calculate the number of neighbours at larger and larger distances. For the full analysis (which we omit here), please refer to → R Cohen and S Havlin, 2003
- Here it is sufficient to say that this approach can be used to calculate the diameter as given above (which again depends only on the first and second raw moment of the distribution). For scale-free networks with $\gamma \in (2, 3)$ we find an expression for the diameter above, stating that it scales in a double-logarithmic fashion with the network size. The second raw moment of the distribution appears in the denominator of this expression. For $\gamma \in (2, 3)$ we have $\gamma - 2 \in (0, 1)$. This implies that the smaller the exponent γ (i.e. the larger the heterogeneity) the larger the (absolute value) of $\log(\gamma - 2)$, and thus the smaller the diameter.
- This translates to the fact that (i) networks where the degree distribution has an infinite second raw moment have a very small diameter (constant for practical network sizes), and (ii) the more heterogeneous the degree distribution the smaller the diameter.

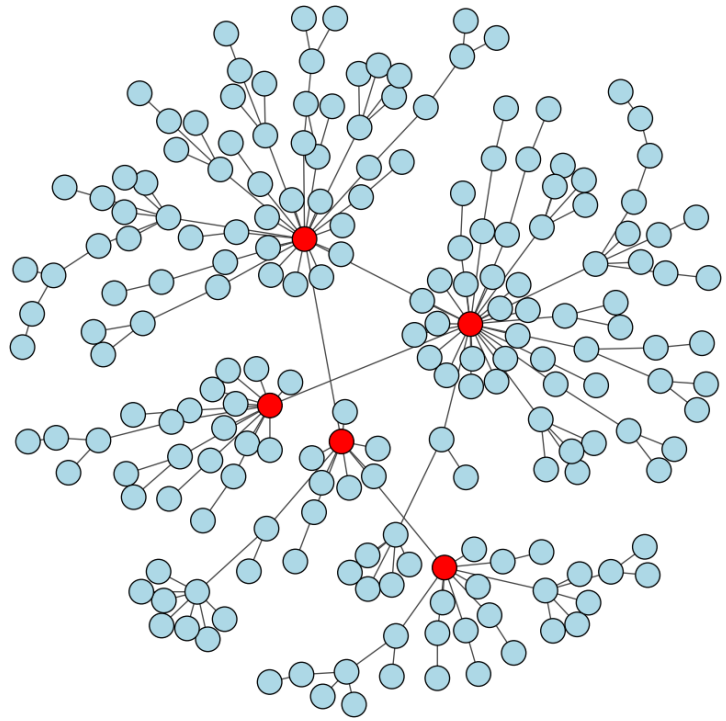
Degree-dependent failures

- nodes with higher degree may be more likely to fail $\Rightarrow$ **WHY?**
- attack strategy: **remove all nodes with degree $k > k_t$**
- we can express this with a degree-dependent failure probability $q(v) = q(d_v)$ with

$$q(k) := \begin{cases} 1 & \text{if } k \geq k_t \\ 0 & \text{if } k < k_t \end{cases}$$

- using the **Heaviside step function** Θ we have

$$q(k) = \Theta(k_t - k)$$



example

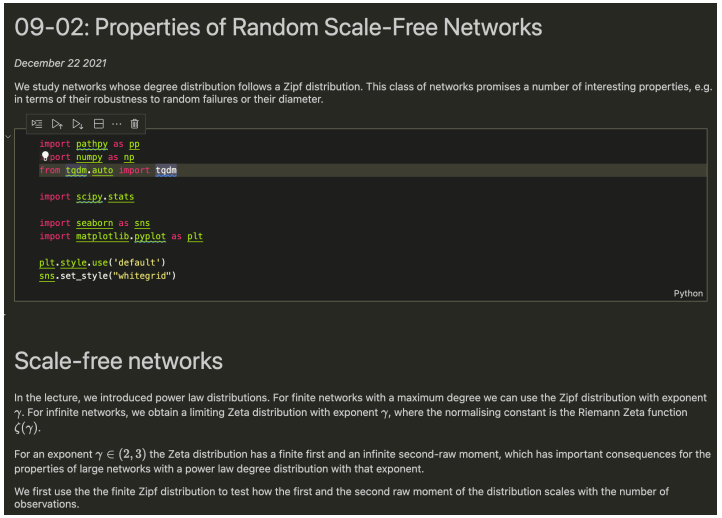
random scale-free network with $n = 200$ nodes
failure probability $q(k) = \Theta(10 - k)$

Notes:

- Let us close our discussion by critically reflecting whether a uniform failure model is a realistic model.
- Degree-dependent failure probabilities, in which nodes with higher degree fail with higher probability, have a number of natural interpretations:
  - Targeted attack or sabotage: an attacker may want to maximize damage with a minimum of effort, thus choosing the most connected nodes.
  - Load-dependent failures of nodes could lead more connected nodes to have a higher failure probability.
  - In many systems, a risk to fail results from the very fact of being connected to others (e.g. in systems where errors can propagate), which means that more connected nodes are more likely to fail.
- We can analytically express such a failure model by replacing the uniform failure probabilities by a Heaviside step function (named after Oliver Heaviside), which is zero for all $x < 0$ and one for all $x \geq 0$.
- In our example removing the five most-connected nodes already has a major impact, i.e. the connected component is destroyed.

Practice Session

- we study the robustness against random node failures and targeted attacks
- we experimentally test our theoretical predictions in random scale-free networks


09-02: Properties of Random Scale-Free Networks

December 22 2021

We study networks whose degree distribution follows a Zipf distribution. This class of networks promises a number of interesting properties, e.g. in terms of their robustness to random failures or their diameter.

```
import pathpy as pp
import numpy as np
from tqdm.auto import tqdm

import scipy.stats

import seaborn as sns
import matplotlib.pyplot as plt

plt.style.use('default')
sns.set_style("whitegrid")
```

Python

Scale-free networks

In the lecture, we introduced power law distributions. For finite networks with a maximum degree we can use the Zipf distribution with exponent γ. For infinite networks, we obtain a limiting Zeta distribution with exponent γ, where the normalising constant is the Riemann Zeta function $\zeta(\gamma)$.

For an exponent $\gamma \in (2, 3)$ the Zeta distribution has a finite first and an infinite second-raw moment, which has important consequences for the properties of large networks with a power law degree distribution with that exponent.

We first use the the finite Zipf distribution to test how the first and the second raw moment of the distribution scales with the number of observations.

observation

even for large scale-free networks, we find considerable discrepancies that are due to **finite-size effects**

practice session

see notebook 09-02 in `gitlab` repository at

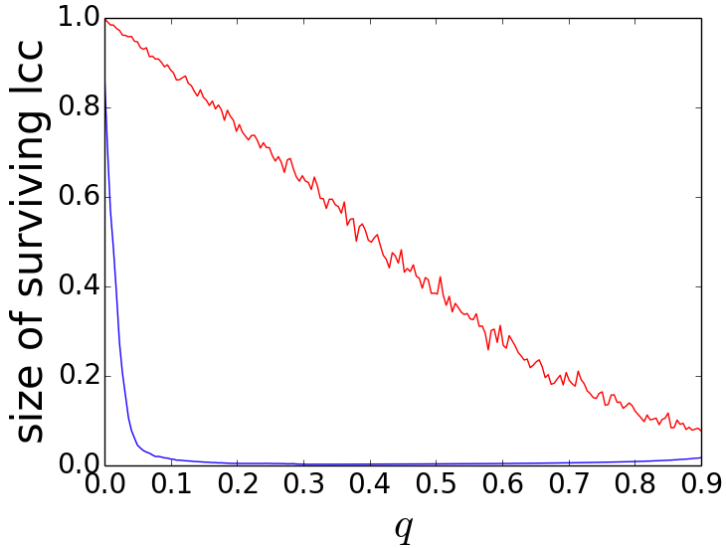
→ `https://gitlab.informatik.uni-wuerzburg.de/ml4nets_notebooks/2022_wise_sna_notebooks`

Notes:

- In the practice session, we experimentally study the relative size of the surviving connected component of scale-free networks under a random failure model. We find that the surviving connected component size decreases linearly with the failure probability q. This is at odds with our theoretical prediction. We would expect that for a scale-free network with exponent between 2 and 3, the size of the largest surviving connected component is close to one.
- The reason for this difference between theory and reality is the finite size of (real) networks. In other words: for a network with exponent $\gamma = 2.5$ the prediction (which we have obtained in the limit of infinite networks) is not a good approximation for a finite network with 1000 nodes. We have seen that this finite size was not an issue for Erdős-Rényi networks, but it is an issue for scale-free networks.

Scale-free networks: robust yet fragile

- attack scenario: expected size of largest surviving connected component?
- removal of tiny fraction q of most-connected nodes sufficient to destroy giant connected component
- **random scale-free networks are fragile**
- sometimes paraphrased as **"robust yet fragile"** nature of scale-free networks



example

average size of largest connected components for random scale-free networks with $n = 1000$ nodes and $\gamma = 2.5$

removal of fraction of q most-connected (blue) and random nodes (red)

Notes:

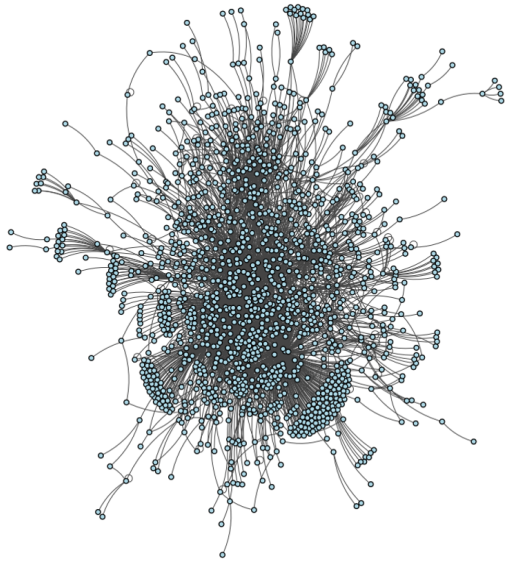
- While we omit the generating function analysis of the attack scenario, we can numerically explore it by a simulation. In the plot on the right, we compare the effect of a fraction of q random node failures (red curve) with a model where we remove the fraction of q most connected nodes (i.e. for each target fraction q we can determine a value of k_{max} for the degree-dependent failure probability given in the previous slide
- The blue line in the plot shows the mean relative size of the largest surviving connected component in a scale-free network with $\gamma = 2.5$ and $n = 1000$ nodes.
- We find that the size of the largest surviving component quickly drops to zero. This is in stark contrast to our finding of "super-robustness". It implies that random scale-free networks are robust against random failures but very fragile when considering targeted attacks on the most connected nodes.
- This theoretical finding has led to a huge interest in scale-free networks, and it has – at time – led researchers to make far-reaching statements about real systems. In the following, we scrutinise some of those statements.

AS-level Internet topology

- **argument 1:** "degree distribution of Internet topology approximately follows a power law with exponent 2.34"
- **argument 2:** "theory of scale-free networks suggests that Internet topology is fragile"

> *"The topological weaknesses of the current communication networks, rooted in their inhomogeneous connectivity distribution, seriously reduce their attack survivability. This could be exploited by those seeking to damage these systems."* ⟶ R Albert et al. 2000

- **should we be worried?**



example

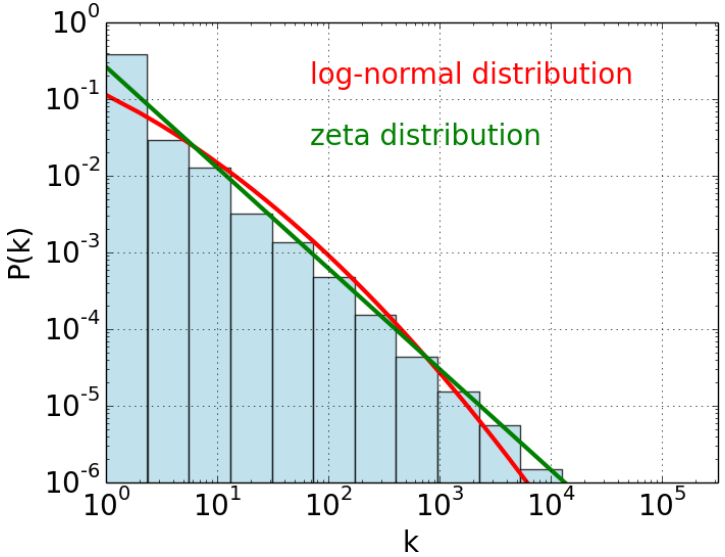
degree distribution (bottom) of AS-level Internet topology (top) as of late 1990s

Notes:

- In particular, the theoretical finding of the robust-yet-fragile nature of scale-free networks can trick us into the following chain of arguments about real-world systems.
- As an example we use the AS-level Internet topology, which has been extensively discussed in the literature. But similar claims have been made about other biological, social, and technical systems.
- Reminder: AS = autonomous systems, i.e. each node is a computer network under the administrative control of one organization.
- The first step in the chain of arguments is the fact that the degree distribution of the Internet topology approximately follows a power law with an exponent $\gamma = 2.34$ in the range between two and three. For this parameter, the theory of scale-free networks predicts an ultra-small diameter, super-robustness but also a high fragility under targeted attacks.
- The quote above is from $\longrightarrow$ R Albert, H Jeon, AL Barabasi, 2000
- It turns out that this argumentation introduces three fallacies, which we discuss in the following.

Fallacy 1: what is the degree distribution?

- probability distribution $P(k)$ of degrees needed for
  - generating function analysis
  - numerical exploration of statistical ensemble
- **analytical form of degree distribution is foundation for theoretical claims**
- visual resemblance is not enough!
- need rigorous statistical hypothesis testing
- particularly challenging for power law distributions



example

degree distribution of Molloy-Reed microstate with fixed log-normal degree distribution and two fits to the degree distribution

Notes:

- The first fallacy is related to statement 1 on the previous slide. Remember that both the analytical findings, as well as our numerical experiments was based on a statistical ensemble defined by a particular degree distribution. The shape of this distribution critically affects the results, so we need to be careful what distribution we use.
- Specifically, if we want to make statements about the expected properties of a real networked system, we need to estimate the distribution based on the observed (finite) degree sequence
- This is a non-trivial problem by itself and in early works on scale-free networks it has often been addressed (in a rather simple way) by visually fitting the slope of a line on a log-log plot showing the histogram of degrees (see figure on the right)
- This is not a good idea, because whether or not such a line is a good (or even the best possible) fit for an empirical distribution is hard if not impossible to decide visually.
- Consider the example on the right, which compares both a log-normal distribution and a zeta distribution fitted to a log-normal degree distribution. Which fit would you prefer?
- Also note that the log-normal distribution is the distribution with maximum entropy under the constraint that both the mean and variance of the logarithm of a random variable are given. That is from a model selection perspective, there are arguments that speak in favor of a log-normal distribution (more on this later in the course).

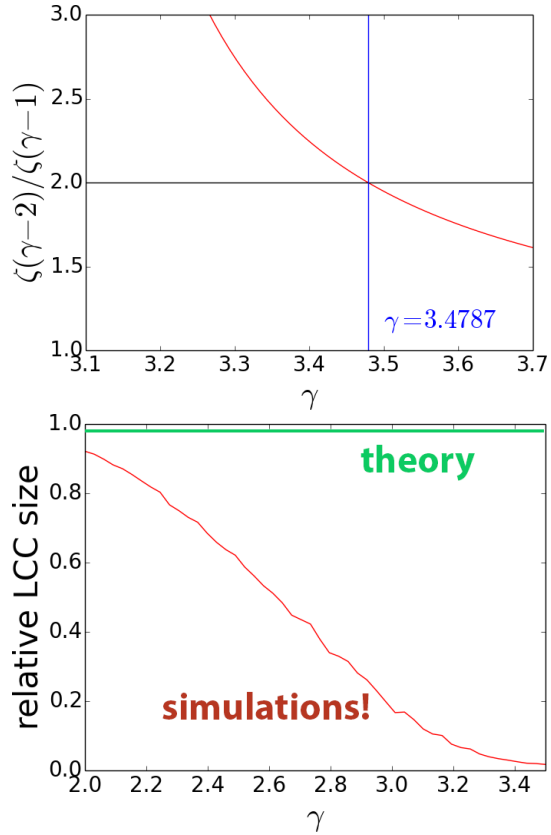
Fallacy 2: finite vs. infinite systems

- analytical statements hold in the limit of infinite systems but **real systems are finite**
- above which network size n is the infinite limit a good approximation?

example

expected size of largest connected component in random scale-free networks should be close to one for all $\gamma < 3.4787$

- for $\gamma > 3$ **theoretical predictions do not apply even to very large realisations**



Molloy-Reed criterion for random scale-free networks (top) and size of LCC in networks with 5000 nodes (bottom)

Notes:

- A second fallacy is related to the finite-size effects, which we have seen before. Our theoretical predictions hold in the limit of infinite network size, while real systems are large but necessarily finite. For random networks, we have seen that the results for infinite networks closely approximate the results for finite networks, even for small network sizes with a few hundred nodes. But for other distributions this may be different!
- As an example, consider the mean relative size of the largest connected component in random scale-free networks generated for different exponents γ with $n = 5000$ nodes.
- The Molloy-Reed criterion for a giant connected component (see analytical arguments in L06) is $\frac{\langle k^2 \rangle}{\langle k \rangle} > 2$
- For scale-free networks with exponent γ, the Molloy-Reed criterion yields the condition:

$$\frac{\zeta(\gamma - 2)}{\zeta(\gamma - 1)} > 2$$

- In the infinite limit, a giant connected component is expected to exist for all $\gamma < 3.47$, however the plot below shows that this is not the case for real networks with 5000 nodes. How closely a finite network resembles the infinite limit further depends on the exponent γ (and thus the degree distribution) → Exercise Sheet 07

Fallacy 3: microstates vs. real systems

- our findings hold for **random microstates** in ensemble of networks with fixed degree distribution
- empirical studies show that **Internet topology** is …

> *"…not only robust to worst-case deletions (here, worst cases are low connectivity core vertices) but also shows high tolerance to deleting other vertices. In particular, loss of high-degree edge routers disconnects only low-bandwidth users and has no other effect on overall connectivity."*
>
> ⟶ JC Doyle et al. 2005

John C. Doyle
born 1954

Walter Willinger
born ?

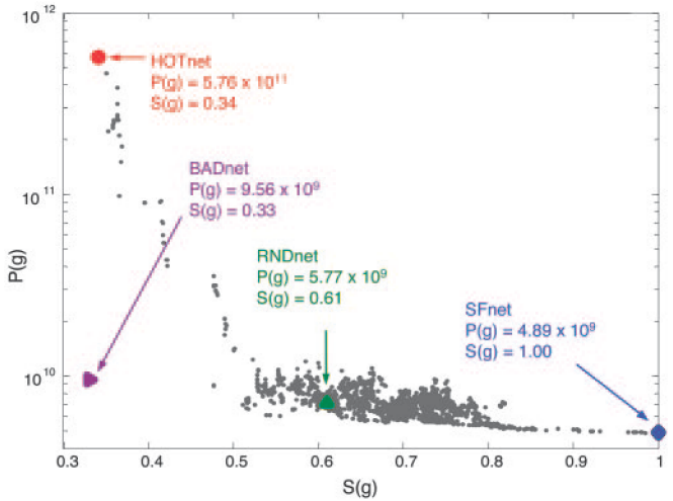


Fig. 4. The diversity of graphs having the same degree sequence *D*. Despite having the identical budget, technology constraint, degree sequence, and traffic-demand model, when computing and plotting for the four network models in Fig. 1 their *S*(*g*) (*x* axis) and network performance *P*(*g*) (*y* axis) metrics, the four models occupy completely different areas in the *S*(*g*) versus *P*(*g*) plane.

image credit: CalTech, acems.org, ⟶ JC Doyle et al., 2005 , fair use

Notes:

- Finally – and maybe most importantly – we have analysed **expected properties of random microstates** drawn from the statistical ensemble of networks with given degree distribution. This does **not** imply that expected properties hold for **all microstates**. We can have outliers that have very different properties. Depending on the ensemble, we can even have cases where the majority of microstates have properties that are quite different from the mean.
- Moreover, **real networks** are created, e.g., by sophisticated engineering processes or evolutionary principles (rather than being drawn at random). This means that a specific real network can be robust against attacks, even though its degree sequence follows a power law distribution with $\zeta \in (2, 3)$, and even though the majority of microstates with the same degree sequence are not robust!
- The figure on the right (and the explanation below) is taken from $\longrightarrow$ JC Doyle et al., 2005
  $P(g)$ is the "performance" of a network g, defined as its throughput in terms of communication flows. $S(g)$ can be interpreted as "likelihood", i.e. how likely it is to generate a graph g at random (e.g. from the Molloy-Reed model). SFNET and RNDNet: scale-free networks generated by two different random processes, BADnet: network with scale-free degree distribution, specifically designed to have poor performance, HOTnet: network representing an actual communication network (Abilene network).

Limitations of ensemble-based approaches

- statistical ensembles are **good null models**, but **bad models**

Breakdown of the Internet under intentional attack

Reuven Cohen[1] *, Keren Erez[1], Daniel ben-Avraham[2], and Shlomo Havlin[1]
[1] *Minerva Center and Department of Physics, Bar-Ilan university, Ramat-Gan, Israel*
[2] *Department of Physics, Clarkson University, Potsdam NY 13699-5820, USA*

We study the tolerance of random networks to intentional attack, whereby a fraction p of the most connected sites is removed. We focus on scale-free networks, having connectivity distribution $P(k) \sim k^{-\alpha}$ (where k is the site connectivity), and use percolation theory to study analytically and numerically the critical fraction p_c needed for the disintegration of the network, as well as the size of the largest connected cluster. We find that even networks with $\alpha \leq 3$, known to be resilient to random removal of sites, are sensitive to intentional attack. We also argue that, near criticality, the average distance between sites in the spanning (largest) cluster scales with its mass, M, as $\sqrt{M}$, rather than as $\log_k M$, as expected for random networks away from criticality. Thus, the disruptive effects of intentional attack become relevant even before the critical threshold is reached.

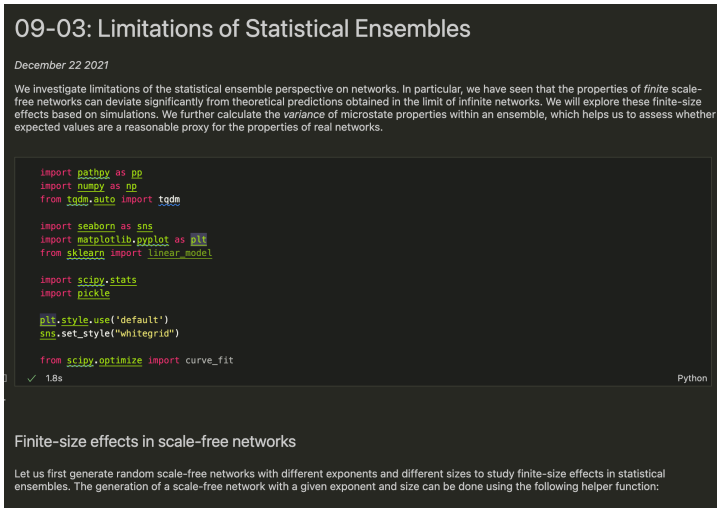
"This raises the more basic question of the applicability to highly evolved systems of unstructured, ensemble-based approaches, of which SF networks are just one example, and a largely parallel story in biology further suggests that the answer may be negative." → John C. Doyle et al. 2005

Notes:

- In summary, findings about random scale-free networks **do not** tell us much about the properties of a particular real network (even if its degree distribution exactly follows a power law distribution). They rather tell us something about the **expected properties**.
- In general, statistical ensembles are not suitable as models for a system since they significantly oversimplify the way the network topology was generated.
- However, they are good null models which can be used to see what properties we can already expect at random, and thus to distinguish patterns in networks from noise (i.e. randomness). After all, if data on a real system are available there is no point in claiming that, e.g., the robustness of this systems corresponds to that of a random network. It is much more reasonable to study the actual robustness of the real system instead. The ensemble perspective then allows us to compare this robustness to what we would expect at random, thus highlighting potential mechanism that shape the system in question (and that may make it less or more robust than expected).
- This difference between random networks and real systems is sometimes not made clear enough (see the exemplary title and abstract of a network science paper above). A nice discussion of this problem (for engineered systems) can be found in → JC Doyle et al., 2005

Practice Session

- we experimentally explore finite size effects
- we study the variance of microstates in statistical ensembles
- we demonstrate challenges that arise in the fitting of scale-free distributions


09-03: Limitations of Statistical Ensembles

December 22 2021

We investigate limitations of the statistical ensemble perspective on networks. In particular, we have seen that the properties of *finite* scale-free networks can deviate significantly from theoretical predictions obtained in the limit of infinite networks. We will explore these finite-size effects based on simulations. We further calculate the *variance* of microstate properties within an ensemble, which helps us to assess whether expected values are a reasonable proxy for the properties of real networks.

```
import pathpy as pp
import numpy as np
from tqdm.auto import tqdm

import seaborn as sns
import matplotlib.pyplot as plt
from sklearn import linear_model

import scipy.stats
import pickle

plt.style.use('default')
sns.set_style("whitegrid")

from scipy.optimize import curve_fit
```

✓ 1.8s Python

Finite-size effects in scale-free networks

Let us first generate random scale-free networks with different exponents and different sizes to study finite-size effects in statistical ensembles. The generation of a scale-free network with a given exponent and size can be done using the following helper function:

practice session

see notebook 09-03 – 09-04 in `gitlab` repository at

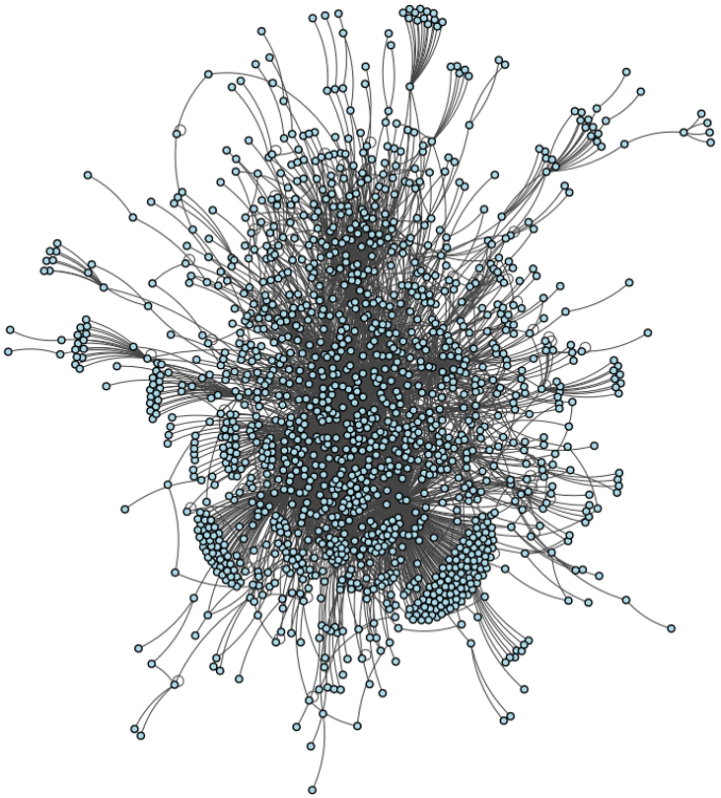
→ `https://gitlab.informatik.uni-wuerzburg.de/ml4nets_notebooks/2022_wise_sna_notebooks`

Notes:

- In the final practice session for this lecture (and this year) we experimentally study limitations of ensemble-based analysis of complex networks, namely finite-size effects, the variance of microstates within an ensemble, as well as issues arising in the fitting of scale-free degree distributions.

Benefits of ensemble-based approaches

- **what are statistical ensembles good for?**
  - they show us which features of real networks are worth a closer look
  - they are a valuable baseline for models of complex systems
- **good models need additional ingredients**
  - technological constraints
  - strategic behavior
  - design decisions
  - growth process
- good news: **domain expertise still needed**



AS-level Internet topology

Notes:

- In general ensemble-based approaches are useful to sharpen our intuition of "what is to be expected" vs. "what is surprising", i.e. which properties of a networked system deserve "further explanation" and which do not.
- To obtain good models for complex systems, most of the time we must include further ingredients like, e.g. physical, economic, technological constraints, costs and pay-offs of certain links, strategic behavior of nodes, growth process of the system, etc.
- Example: For the Internet topology, which additional ingredients may be needed?
  - Technological constraints: bandwidth of routers and links
  - Economic aspects: who has an interest to be connected to whom?
  - Geographical aspects: there are limits of how long communication links can be
  - Design choices: engineers of distributed communication systems will carefully avoid single points of failure (the Internet was designed for exactly this purpose)
  - There are different classes of networks (transit and stub networks) that have very different roles and follow different economic rules
- Think of **real social networks**: is the degree distribution really enough to characterise them? Certainly not! We already know that real systems will have a higher clustering coefficient than expected at random.

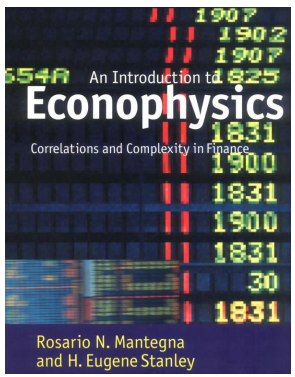
In summary

- carefully distinguish between **organised and disorganised complexity** → Warren Weaver, 1948

- random scale-free networks are **disorganised complex networks** and thus accessible to an ensemble perspective

- Internet (like most real systems) is an **organised complex network**, subject to technological, physical, and economic constraints

Warren Weaver
1894 – 1978



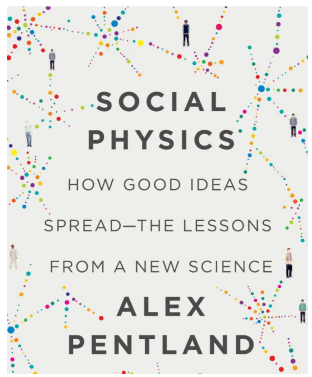


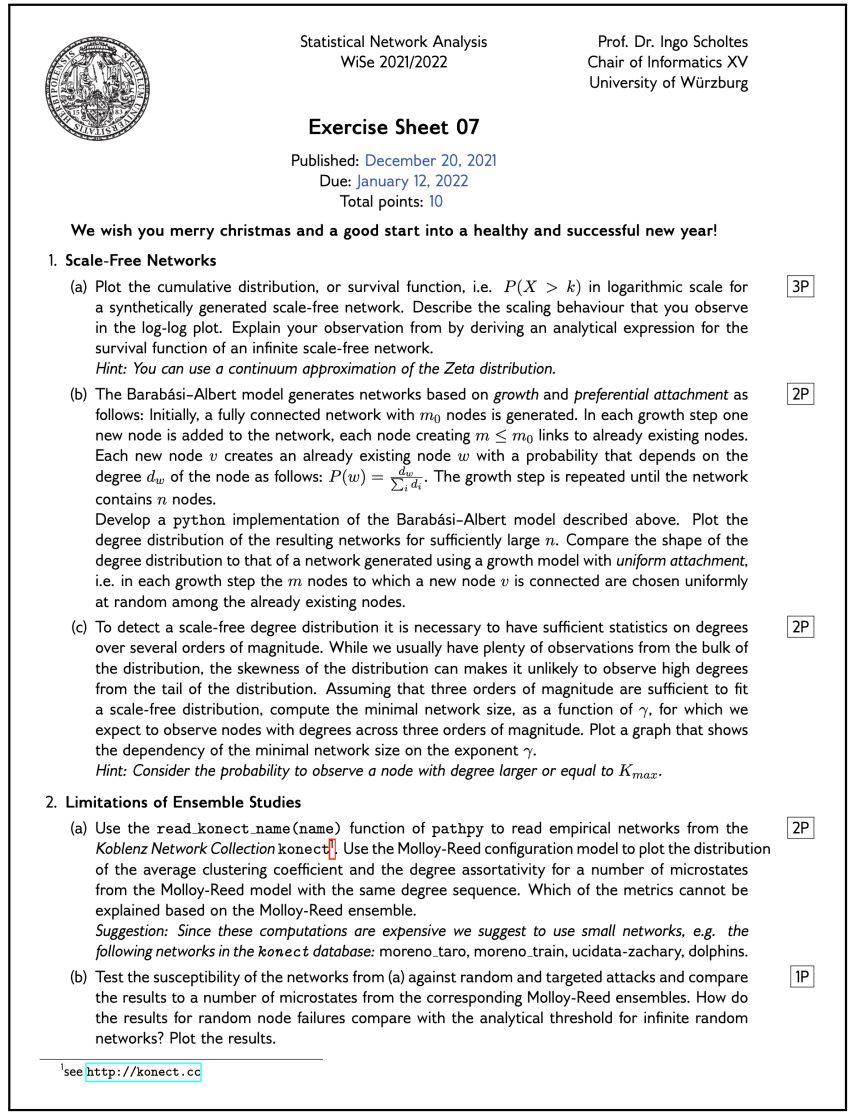
image credit: Rockefeller Foundation, public domain

Notes:

- The discussion about the difference between ensemble-based approaches and models of real (complex) systems is not new. It has just been forgotten for almost six decades.
- In his article “Science and Complexity”, published in American Scientist in 1948, Warren Weaver has warned us to carefully distinguish between what he called “organised” and “disorganised” complexity.
- He argued that ensemble methods (drawing inspiration from physics) are useful to study complex phenomena in systems where elements follow simple mechanistic, or stochastic rules. He called this “disorganized complexity” and following this terminology random scale free networks are one example for “disorganised complex networks”.
- Real systems have different characteristics, since they have been engineered (or they have evolved) for a specific purpose. Warren Weaver called this “organised complexity” and he argued that this marks the line of distinction between systems that can be studied by “physics-inspired approaches“ and those that cannot.

Exercise sheet 07

- **seventh exercise sheet** available on WueCampus
  - study properties of scale-free distributions
  - implement a growth model for scale-free networks
  - compare random microstates to real networks
- solutions are due **December 21st** (via WueCampus)
- present your solution to earn bonus points



Statistical Network Analysis
WiSe 2021/2022

Prof. Dr. Ingo Scholtes
Chair of Informatics XV
University of Würzburg

Exercise Sheet 07

Published: December 20, 2021
Due: January 12, 2022
Total points: 10

We wish you merry christmas and a good start into a healthy and successful new year!

1. **Scale-Free Networks**

(a) Plot the cumulative distribution, or survival function, i.e. $P(X > k)$ in logarithmic scale for a synthetically generated scale-free network. Describe the scaling behaviour that you observe in the log-log plot. Explain your observation from by deriving an analytical expression for the survival function of an infinite scale-free network. 3P
Hint: You can use a continuum approximation of the Zeta distribution.

(b) The Barabási–Albert model generates networks based on *growth* and *preferential attachment* as follows: Initially, a fully connected network with m_0 nodes is generated. In each growth step one new node is added to the network, each node creating $m \leq m_0$ links to already existing nodes. Each new node v creates an already existing node w with a probability that depends on the degree d_w of the node as follows: $P(w) = \frac{d_w}{\sum_i d_i}$. The growth step is repeated until the network contains n nodes. 2P
Develop a `python` implementation of the Barabási–Albert model described above. Plot the degree distribution of the resulting networks for sufficiently large n. Compare the shape of the degree distribution to that of a network generated using a growth model with *uniform attachment*, i.e. in each growth step the m nodes to which a new node v is connected are chosen uniformly at random among the already existing nodes.

(c) To detect a scale-free degree distribution it is necessary to have sufficient statistics on degrees over several orders of magnitude. While we usually have plenty of observations from the bulk of the distribution, the skewness of the distribution can makes it unlikely to observe high degrees from the tail of the distribution. Assuming that three orders of magnitude are sufficient to fit a scale-free distribution, compute the minimal network size, as a function of γ, for which we expect to observe nodes with degrees across three orders of magnitude. Plot a graph that shows the dependency of the minimal network size on the exponent γ. 2P
Hint: Consider the probability to observe a node with degree larger or equal to K_{max}.

2. **Limitations of Ensemble Studies**

(a) Use the `read_konect_name(name)` function of `pathpy` to read empirical networks from the *Koblenz Network Collection* `konect`[1]. Use the Molloy-Reed configuration model to plot the distribution of the average clustering coefficient and the degree assortativity for a number of microstates from the Molloy-Reed model with the same degree sequence. Which of the metrics cannot be explained based on the Molloy-Reed ensemble. 2P
Suggestion: Since these computations are expensive we suggest to use small networks, e.g. the following networks in the `konect` database: moreno_taro, moreno_train, ucidata-zachary, dolphins.

(b) Test the susceptibility of the networks from (a) against random and targeted attacks and compare the results to a number of microstates from the corresponding Molloy-Reed ensembles. How do the results for random node failures compare with the analytical threshold for infinite random networks? Plot the results. 1P

[1] see `http://konect.cc`

Notes:

- Let me close this lecture by wishing you merry christmas, happy holidays and a good start into a healthy and successful new year 2022!

Self-study questions

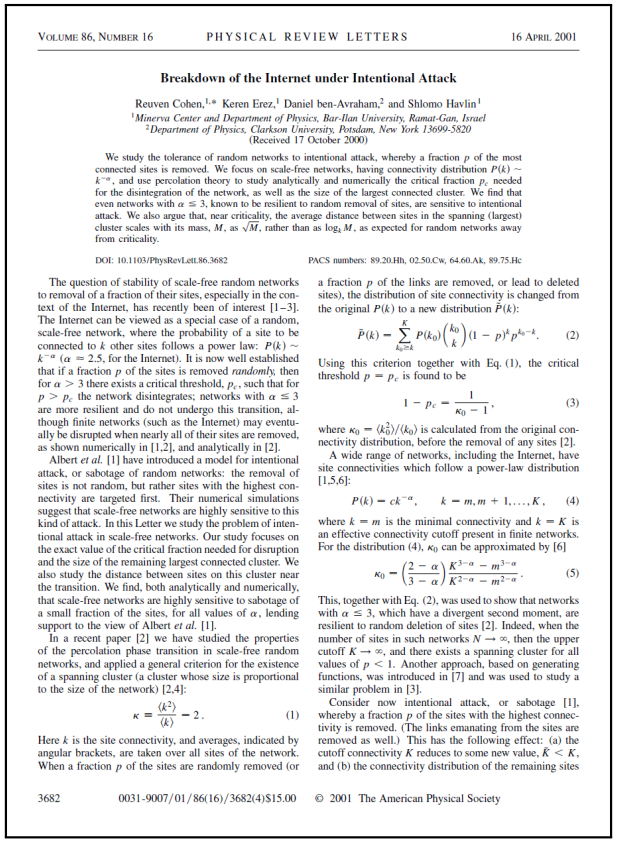
1. Explain why a k-regular random network is less robust than a random Erdős-Rényi network with the same expected degree.
2. What are scale-free networks? How can we characterize their degree distribution in the finite and infinite case?
3. What is the critical failure probability q_c (uniform failure probabilities) for a scale-free network with exponent γ in the limit of infinite size?
4. How does the diameter of scale-free networks scale with their size?
5. What is the reason for the discrepancy between the predicted and the empirical surviving lcc for random failures in scale-free networks?
6. What is the robust-yet-fragile nature of random scale-free networks?
7. Explain three possible fallacies that can occur when applying the ensemble perspective to real systems?

Notes:

References

reading list

- DS Callaway, ME Newman, SH Strogatz, DJ Watts: **Network Robustness and Fragility: Percolation on Random Graphs**, Physical Review Letters, 2000
- R Cohen, K Erez, D Ben-Avraham, S Havlin: **Breakdown of the internet under intentional attack**, Physical Review Letters, 2001
- R Cohen and S Havlin: **Scale-Free Networks Are Ultrasmall**, Physical Review Letters, 2003
- A Albert, H Jeong and AL Barabási: **Error and attack tolerance of complex networks**, Nature, 2000
- JC Doyle et al. **"The "robust yet fragile" nature of the Internet**, PNAS, 2005
- W Weaver: **Science and Complexity**, American Scientist, 1948
- A Clauset, D Kempe, C Moore: **On the Bias of Traceroute Sampling: or, Power-Law Degree Distributions in Regular Graphs**, Journal of the ACM, 009



VOLUME 86, NUMBER 16 PHYSICAL REVIEW LETTERS 16 APRIL 2001

Breakdown of the Internet under Intentional Attack

Reuven Cohen,[1,*] Keren Erez,[1] Daniel ben-Avraham,[2] and Shlomo Havlin[1]

[1]*Minerva Center and Department of Physics, Bar-Ilan University, Ramat-Gan, Israel*

[2]*Department of Physics, Clarkson University, Potsdam, New York 13699-5820*

(Received 17 October 2000)

We study the tolerance of random networks to intentional attack, whereby a fraction p of the most connected sites is removed. We focus on scale-free networks, having connectivity distribution $P(k) \sim k^{-\alpha}$, and use percolation theory to study analytically and numerically the critical fraction p_c needed for the disintegration of the network, as well as the size of the largest connected cluster. We find that even networks with $\alpha \leq 3$, known to be resilient to random removal of sites, are sensitive to intentional attack. We also argue that, near criticality, the average distance between sites in the spanning (largest) cluster scales with its mass, M, as $\sqrt{M}$, rather than as $\log_k M$, as expected for random networks away from criticality.

DOI: 10.1103/PhysRevLett.86.3682 PACS numbers: 89.20.Hh, 02.50.Cw, 64.60.Ak, 89.75.Hc

The question of stability of scale-free random networks to removal of a fraction of their sites, especially in the context of the Internet, has recently been of interest [1–3]. The Internet can be viewed as a special case of a random, scale-free network, where the probability of a site to be connected to k other sites follows a power law: $P(k) \sim k^{-\alpha}$ ($\alpha \approx 2.5$, for the Internet). It is now well established that if a fraction p of the sites is removed *randomly*, then for $\alpha > 3$ there exists a critical threshold, p_c, such that for $p > p_c$ the network disintegrates; networks with $\alpha \leq 3$ are more resilient and do not undergo this transition, although finite networks (such as the Internet) may eventually be disrupted when nearly all of their sites are removed, as shown numerically in [1,2], and analytically in [2].

Albert *et al.* [1] have introduced a model for intentional attack, or sabotage of random networks: the removal of sites is not random, but rather sites with the highest connectivity are targeted first. Their numerical simulations suggest that scale-free networks are highly sensitive to this kind of attack. In this Letter we study the problem of intentional attack in scale-free networks. Our study focuses on the exact value of the critical fraction needed for disruption and the size of the remaining largest connected cluster. We also study the distance between sites on this cluster near the transition. We find, both analytically and numerically, that scale-free networks are highly sensitive to sabotage of a small fraction of the sites, for all values of α, lending support to the view of Albert *et al.* [1].

In a recent paper [2] we have studied the properties of the percolation phase transition in scale-free random networks, and applied a general criterion for the existence of a spanning cluster (a cluster whose size is proportional to the size of the network) [2,4]:

$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle} - 2 \,. \tag{1}$$

Here k is the site connectivity, and averages, indicated by angular brackets, are taken over all sites of the network. When a fraction p of the sites are randomly removed (or a fraction p of the links are removed, or lead to deleted sites), the distribution of site connectivity is changed from the original $P(k)$ to a new distribution $\tilde{P}(k)$:

$$\tilde{P}(k) = \sum_{k_0=k}^{K} P(k_0) \binom{k_0}{k} (1-p)^k p^{k_0-k} \,. \tag{2}$$

Using this criterion together with Eq. (1), the critical threshold $p = p_c$ is found to be

$$1 - p_c = \frac{1}{\kappa_0 - 1} \,, \tag{3}$$

where $\kappa_0 = \langle k_0^2 \rangle / \langle k_0 \rangle$ is calculated from the original connectivity distribution, before the removal of any sites [2].

A wide range of networks, including the Internet, have site connectivities which follow a power-law distribution [1,5,6]:

$$P(k) = ck^{-\alpha}, \qquad k = m, m+1, \ldots, K \,, \tag{4}$$

where $k = m$ is the minimal connectivity and $k = K$ is an effective connectivity cutoff present in finite networks. For the distribution (4), κ_0 can be approximated by [6]

$$\kappa_0 = \left(\frac{2-\alpha}{3-\alpha}\right) \frac{K^{3-\alpha} - m^{3-\alpha}}{K^{2-\alpha} - m^{2-\alpha}} \,. \tag{5}$$

This, together with Eq. (2), was used to show that networks with $\alpha \leq 3$, which have a divergent second moment, are resilient to random deletion of sites [2]. Indeed, when the number of sites in such networks $N \to \infty$, then the upper cutoff $K \to \infty$, and there exists a spanning cluster for all values of $p < 1$. Another approach, based on generating functions, was introduced in [7] and was used to study a similar problem in [3].

Consider now intentional attack, or sabotage [1], whereby a fraction p of the sites with the highest connectivity is removed. (The links emanating from the sites are removed as well.) This has the following effect: (a) the cutoff connectivity K reduces to some new value, $\tilde{K} < K$, and (b) the connectivity distribution of the remaining sites

Notes: