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 08
Percolation Phase Transition

December 7, 2022

Notes:

- **Lecture L08:** Percolation Phase Transition 07.12.2022
- **Educational objective:** We use generating functions to derive a critical point for the emergence of a giant connected component in random networks with arbitrary degree distributions.
  - Emergence of a giant connected component
  - Molloy-Reed criterion in random networks
  - Robustness against random node failures

- **Exercise 06:** k-regular random graphs and epidemic spreading due 14.12.2022

Motivation

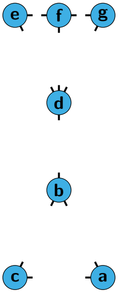
- **Molloy-Reed model**: ensemble of random networks with fixed degree sequence or distribution

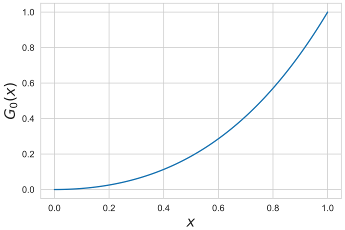
- we can use **generating functions** to encode **distribution of degrees and excess degrees**

$$G_0(x) := \sum_{k=0}^{\infty} P(k) x^k$$

$$G_1(x) := \frac{G_0'(x)}{G_0'(1)}$$

- we explained **friendship paradox** in random networks based on **non-zero variance of degree distribution**



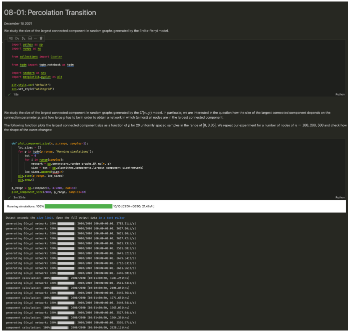


Notes:

- In the last lecture, we introduced generating functions and we applied them to study the expected neighbour degree of networks generated by the **Molloy-Reed or configuration model**, i.e. the statistical ensemble of microstates with a fixed degree sequence of degree distribution
- We particularly found that we can explain the friendship paradox, which we observed both for empirical networks as well as for random microstates of the Molloy-Reed ensemble, based on the variance of the degree distribution. We found that any non-zero variance in the degree distribution necessarily implies that the friendship paradox holds on average, i.e. the mean neighbour degree is larger than the mean degree.
- Today, we will show how we can use **generating functions** to make statements about the connectivity and robustness of random networks with given degree distribution. In the last lecture we have also mentioned that we can consider the Molloy-Reed ensemble as a generalization of the simpler random graph models, if we fix the degree distribution to a Binomial (Poisson/Normal) distribution. If we learn how to analytically derive expected properties for this more general ensemble, we can also apply those results to random networks.

Practice Session

- how large is the **largest connected component** of a random network?
- how does the size of the largest connected component in $G(n, p)$ model depend on p?



practice session

see notebook 08-01 in `gitlab` repository at

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

Notes:

- In the first practice session, we motivate the problem that we will address today. It is based on a very simple question: How large is the largest connected component in a random graph generated by the $G(n, p)$ model. Specifically, we will study how the size of the largest connected component changes as we increase or decrease the probability p to generate a link.
- What do you expect? How does the size of a largest connected component change if we increase p by, say, 10 %?

Percolation phase transition in networks

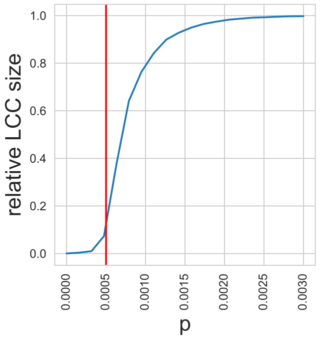
- as we increase p in the $G(n, p)$ model, we observe an **abrupt transition** from a disconnected to a connected phase

percolation phase transition

in the $G(n, p)$ model a giant connected component **emerges abruptly** as we increase p beyond a **critical point** that depends on the network size

- **percolation in physics**: can a liquid pass through a porous material?

- we can use generating functions to **predict the critical point** at which giant connected component emerges in random networks



relative size of largest connected component in networks generated by $G(n, p)$ model for $n = 2000$

Notes:

- Different from what you might have expected, we find that – as we increase the link probability p – a giant connected component (i.e. a largest connected component that contains almost all nodes) **emerges abruptly** beyond a certain **critical point** for the link probability (see figure above for a network with $n = 2000$ nodes). The transition between the disconnected phase (left) and the connected phase (right) actually becomes increasingly sharp as we increase the size of the network.
- This is an interesting example for a sudden phase transition, a class of phenomena that are common in physics. In particular, the transition between a disconnected and connected phase is an example for a percolation phase transition. In physics and material science, percolation theory addresses the question whether a liquid can pass through a porous material or not. This problem can be modelled as a connectivity problem in a lattice graph, i.e. whether a path connects two sides of the material.
- We will use generating functions to analytically understand this critical point, which is important for multiple reasons:
    1. it is a paradigmatic example of how we can use using statistical ensembles to explain an "emergent", i.e. collective property of complex networks → L01
    2. it is an example for non-linear behaviour in complex systems, where tiny causes (like the addition/removal of few links) can have major effects (like the emergence of destruction of connectedness)
    3. we will derive a statement about arbitrary classes of random networks with given degree distribution, which we can generalize to study network robustness
    4. finally, it is related to the epidemic threshold, which can be defined based on the basic reproduction number R_0 that you may have heard in the news.

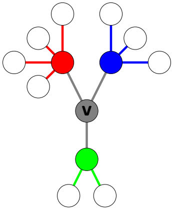
Reminder: Excess degree distribution

- distribution of the **degrees of neighbours** w without (v, w)

$$G_1(x) := \frac{\sum_{k=0}^{\infty} \frac{k}{\langle k \rangle} P(k) x^k}{x}$$

$$= \frac{\sum_{k=0}^{\infty} k P(k) x^{k-1}}{\langle k \rangle} = \frac{G_0'(x)}{G_0'(1)}$$

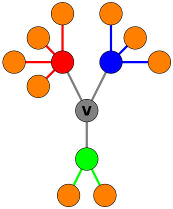
(often called **excess degree distribution**)



$\{4, 3, 2\}$

- distribution of **number of second-nearest neighbours** of v

$$\sum_{k=0}^{\infty} P(k)[G_1(x)]^k = G_0(G_1(x))$$



$4 + 3 + 2 = 9$

Notes:

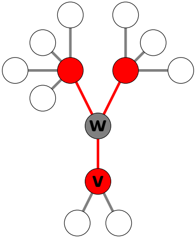
- Consider a node v chosen uniformly at random in a Molloy-Reed microstates. What other quantities can we calculate? For the distribution of the **degrees of neighbours w of a randomly chosen v**, the following holds:
  - The probability that a randomly chosen node w has degree k is $P(k)$. However, we must additionally account for the fact that w is also a neighbour of v (i.e. w was **not** chosen uniformly at random).
  - A node with degree k has k chances to be randomly chosen as neighbour of v, so the probability that w has degree k is proportional to $kP(k)$
  - We need to normalize this to obtain a probability, i.e. we divide each probability $P(k)$ by $\sum_k kP(k) = \langle k \rangle$.
  - In addition, we must discount for link (v, w), which decrements the resulting degree by one. We can achieve by dividing the generating function by x.
- We obtain a **new generating function $G_1(x)$**, which generates the probability that a random neighbour of a randomly chosen node v has degree k (without (v, w)). This is often called the **excess degree distribution** (cf. EX 04). Using this function we can calculate the **distribution of the number of nodes at distance two** to a randomly chosen node v as follows:
  - We sum all degrees of neighbors w of v (without considering (v, w)).
  - We thus sum k realizations of neighbour degrees (which are generated by G_1) where k is generated by G_0. Hence the distribution that we are looking for is generated by the composition of G_0 and G_1!
- This holds if there is **zero clustering**, i.e. if we can ignore the case that a neighbour of w is also a neighbour of v (which would lead to a closed triad).

Following a randomly chosen link

- consider a **random neighbor** *w* of randomly chosen node *v*
- distribution of **degrees** of node *w* is generated by

$$\sum_k \frac{k}{\langle k \rangle} P(k) x^k = x G_1(x)$$

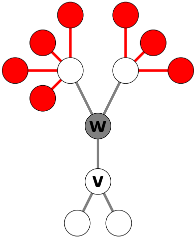
(**excess degree distribution** plus one)



$d_w = 3$

- distribution of **second-order neighbours** of *w* (in "forward direction")

$$G_1(G_1(x))$$



7

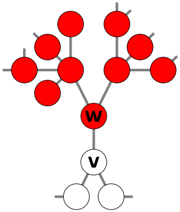
Notes:

- In the following, we use generating functions to study the **expected size of connected components**. For this, we first reconsider the generating function $xG_1(x)$, which generates the degrees of a random neighbor w of a random node v. This is the function that we used to explain the friendship paradox.
- We now literally go one step further and calculate a generating function that generates the **distribution of the number of neighbors at distance two to node v**, limiting ourselves to the direction of link (v, w) (i.e. we start at node v, move to w and count all nodes at distance two through the link (v, w)). We first make some observations:
  - G_1 generates the distribution of degrees of a node arrived at by following an edge, while discounting for the edge we arrived through.
  - We have to sum this quantity for a number of nodes whose distribution is again generated by G_1, i.e. we can get a generating function for the distribution of that sum by composing G_1 with itself!
  - WNote that this is equivalent to $G_0(G_0(x))$ for a sparse random network with a Poisson degree distribution. And this is the reason why we could simply take $\langle k \rangle$ to the power of l when we studied the diameter of random graphs → L05
- Remember: “Generating functions can give stunningly quick derivations of various probabilistic aspects of the problem that is represented by your unknown sequence” → L07

Component size: recursive definition

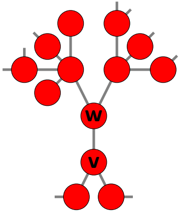
- **distribution of component sizes** of neighbour w of random node v (in "forward direction" including w)

$$H_1(x) := xG_1(H_1(x))$$



- **distribution of component sizes** for randomly chosen node v (including v)

$$H_0(x) := xG_0(H_1(x))$$



Notes:

- We now have almost everything to define a generating function for the distribution of component sizes in a random network with arbitrary degree distribution.
  - Let us assume that an (unknown) generating function H_1 generates the **distribution of component sizes of node w in *forward direction***, i.e. we do not count nodes in that part of the component that we arrived from (we only count red nodes above).
  - Using this assumption, we can derive a *self-consistency condition*: starting from node w that we arrive at, we sum up all component sizes in forward direction for each of the neighbors of w (without considering the node from which we arrive at w)
  - Each of these neighbors we found by following a link, so the distribution of component sizes for the neighbors of w are again generated by H_1.
  - We further know that the number of neighbors of w – again not considering the link that we arrived through – is generated by G_1.
  - The total component size is thus generated by the composition $G_1(H_1(x))$. We finally additionally account for node w, i.e. we add one by multiplying the resulting function with x.
  - By definition, the resulting generating function must again be H_1, i.e. **we found a self-consistency condition that the (unknown) function H_1 must satisfy**
- Using H_1 we can now write down a generating function for the **distribution of component sizes of a randomly chosen node v**. Here we just sum the forward component sizes for all neighbors (whose number is generated by G_0) and again add one for node v by multiplying the function with x.

Expected component size

- **expected component size** of randomly chosen node v is $H_0'(1)$
- with $H_0(x) := xG_0(H_1(x))$ we have

$$H_0'(x) = G_0(H_1(x)) + xG_0'(H_1(x))H_1'(x)$$

$$\langle s \rangle = H_0'(1) = G_0(\underbrace{H_1(1)}_{1}) + G_0'(\underbrace{H_1(1)}_{1})H_1'(1)$$

- with $H_1(x) := xG_1(H_1(x))$ we have

$$H_1'(x) = G_1(H_1(x)) + xG_1'(H_1(x))H_1'(x)$$
$$H_1'(1) = 1 + G_1'(1)H_1'(1)$$

- and thus

$$H_1'(1) = \frac{1}{1 - G_1'(1)}$$

Notes:

- Due to the recursive definition of H_1 this result may not appear to be very helpful, but it turns out that it is enough to solve our original question. For this we consider that the **expected component size of a random node** is the first derivative at $x = 1$ of our function $H_0(x)$. Since we actually do not know the function H_1 (we only know a self-consistency condition), we cannot write down a closed form expression. But maybe this is not needed?
- Let us calculate the expected component size by substituting H_0 with its definition. We apply the product rule $(f \cdot g)' = f' \cdot g + f \cdot g'$ as well as the chain rule $(f \circ g)' = (f' \circ g) \cdot g'$ for derivatives.
- We then substitute $x = 1$ and recall that for a probability generating function f we have $f(1) = 1$. This yields $H_0'(1) = 1 + G_0'(1)H_1'(1)$.
- We find that, to calculate the expected component size, we do not need a closed-form expresion of the generating function H_1, it is enough to know $H_1'(1)$ since we are only interested in the first raw moment.
- What we can say for $H_1'(x)$? Using the self-consistency condition $H_1(x) = xG_1(H_1(x))$, we can apply the product and chain rule and get

$$H_1'(1) = 1 + G_1'(1)H_1'(1)$$

- Here we have used:

$$x = 1 + yx \Rightarrow 1 = x - yx \Rightarrow \frac{1}{x} = 1 - y \Rightarrow x = \frac{1}{1-y}$$

Expected component size

- we get

$$\langle s \rangle = H_0'(1) = \underbrace{G_0(1)}_{1} + G_0'(1) H_1'(1)$$

- with $H_1'(1) = \frac{1}{1 - G_1'(1)}$ we get an expression for the **expected component size** solely based on the known functions G_0 and G_1

$$\langle s \rangle = 1 + \frac{G_0'(1)}{1 - G_1'(1)}$$

- remember: for randomly chosen node v
  - $G_0'(1)$ is expected degree $\langle k \rangle$
  - $G_1'(1)$ is expected neighbour degree $\langle k_n \rangle$ minus one

Notes:

- Substituting H_1' we find the above expression for the expected component size of a randomly chosen node. Remarkably, this expression only depends on the two generating functions G_0 and G_1, for which we know the closed-form expression as long as we know the degree distribution of the network!
- We note that we have the expected degree in the counter while $G'(1)$ in the denominator is the mean neighbor degree minus one.
- How does this help us to answer the question whether we have a **giant connected component**?
- Let n be the number of nodes in the network. The presence of a giant connected component implies that the mean component size diverges for $n \to \infty$. Otherwise, we necessarily have (a possible infinite number of) components that all contain only a finite number of nodes.

Diverging expected component size

- consider size $s(n)$ of largest connected component in a random microstate with n nodes
- $\lim_{n\to\infty} s(n) < \infty \Rightarrow$ no giant connected component
- existence of **giant connected component** implies $\langle s \rangle \to \infty$ for $n \to \infty$

$$\langle s \rangle = 1 + \frac{G_0'(1)}{1 - G_1'(1)}$$

- when does $\langle s \rangle$ diverge?
- we observe two cases:
  1. $G_0'(1) = \langle k \rangle \to \infty$ i.e. **network is not sparse**
  2. $G_1'(1) \to 1$ for **sparse networks** (i.e. $\langle k \rangle$ finite)

Notes:

- Hence, to study whether a giant connected component exists, we can take $n \to \infty$ and study whether the expected component size $\langle s \rangle$ diverges. We find that the convergence behavior of $\langle s \rangle$ is uniquely determined by $G_0'(1)$ (the expected degree) and $G_1'(1)$ (the expected neighbor degree).
- Under which conditions does the expected component size diverge for $n \to \infty$?
- There are two cases where this is the case:
    1. $\langle k \rangle \to \infty$, i.e. the network becomes denser as n grows. Trivially, in this case we have a giant connected component for $n \to \infty$, since the network converges to a fully connected network.
    2. If $\langle k \rangle$ is finite, we find that the mean neighbor degree minus one must necessarily converge to one, i.e. $G'(1) \to 1$ for $n \to \infty$.

Emergence of giant connected component

- when is $G_1'(1) \to 1$ for $n \to \infty$?

$$G_1'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\frac{G_0'(x)}{G_0'(1)} = \frac{\mathrm{d}}{\mathrm{d}x}\frac{\sum_k kP(k)x^{k-1}}{\sum_k kP(k)} = \frac{\sum_k k(k-1)P(k)x^{k-2}}{\sum_k kP(k)}$$

$$G_1'(1) = \frac{\sum_k k(k-1)P(k)}{\sum_k kP(k)} = \frac{\sum_k k^2 P(k)}{\sum_k kP(k)} - 1 = \frac{\langle k^2 \rangle}{\langle k \rangle} - 1$$

- for **sparse** random networks with arbitrary degree distribution:

$$G_1'(1) \to 1 \Leftrightarrow \frac{\langle k^2 \rangle}{\langle k \rangle} \to 2$$

- we find a **critical point** at which a giant connected component emerges in sparse random networks

Notes:

- Let us study the non-trivial case, i.e. where the divergence in a sparse network is due to $G_1'(1) \to 1$.
- We first use the definition of $G_1(x)$ to calculate the derivative at point $x = 1$. $G_1'(1)$ is the ratio between the second and the first raw moment minus one. A giant connected component thus emerges at a critical point that is given by the solution $\frac{\langle k^2 \rangle}{\langle k \rangle} - 1 = 1$.
- We just derived an important critical point for general random graphs with arbitrary degree distributions that is defined by **the ratio between the second and first raw moment of the distribution**.
- This critical point was first derived by means of a different, graph-theoretic analysis in $\longrightarrow$ M Molloy and B Reed, 1995 . The same result has been derived by means of generating functions in $\longrightarrow$ MEJ Newmann, SH Strogatz, DJ Watts, 2001

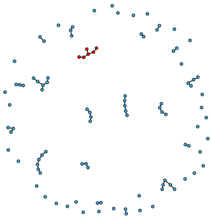
Application to $G(n, p)$ model

sparse random graphs with **Poisson degree distribution** $(n \to \infty)$

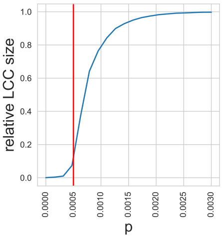
$$P(k) = \frac{n^k p^k e^{-np}}{k!} \Rightarrow G_0(x) = e^{np(x-1)}$$

$$\langle k \rangle = np, \qquad \langle k^2 \rangle = np + n^2 p^2$$

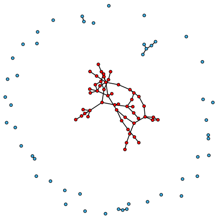
$$\frac{\langle k^2 \rangle}{\langle k \rangle} = \frac{np + n^2 p^2}{np} = 2 \iff 1+np = 2 \iff np = 1 \iff p = \frac{1}{n}$$



$np = 0.95 < 1$
$(n = 100)$



largest component size in random graphs $(n = 2000)$



$np = 1.05 > 1$
$(n = 100)$

Notes:

- We now apply this to explain our result for Erdős-Rényi networks, i.e. we consider an ensemble of networks with a fixed Poisson degree distribution. We use the generating function for the Poisson degree distribution introduced in the exercise (and earlier in this lecture).
- We actually do not even need the generating function since the critical point only depends on the first and the second raw moment of the distribution. For a Poisson degree distribution, both mean and variance are $\lambda = np$ (i.e. mean and variance are the same). In general, the variance $Var(X) = \sigma^2$ is related to the first and the second raw moment of a distribution as follows:

$$Var(X) = \langle k^2 \rangle - \langle k \rangle^2$$

- we can thus calculate the second raw moment as:

$$\langle k^2 \rangle = Var(X) + \langle k \rangle^2$$

- For the Poisson distribution we get $\langle k^2 \rangle = np + n^2 p^2$. From this we calculate the critical point as $p = \frac{1}{n}$ or equivalently $np = 1$.
- This finding has an intuitive interpretation: for a giant connected component to emerge in a random Erdős-Rényi network, each node needs to be connected – on average – to minimally one other node. Note how this is related to the critical threshold of one for the basic reproduction number in the spreading of an epidemic

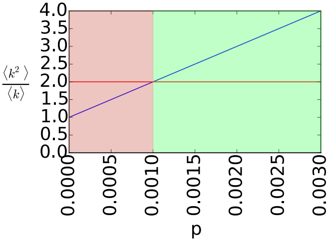
$\longrightarrow$ Exercise Sheet 06

Molloy-Reed criterion

- for $G(n, p)$ microstates **we expect a giant connected component** to exist iff

$$np > 1 \Leftrightarrow 1 + np > 2 \Leftrightarrow \frac{\langle k^2 \rangle}{\langle k \rangle} > 2$$

- increase of p necessarily increases connected components, i.e. **expected component size is monotonous**



Molloy-Reed criterion

for random microstates from the Molloy-Reed ensemble we expect a giant connected component to exist iff

$$\frac{\langle \mathbf{k}^2 \rangle}{\langle \mathbf{k} \rangle} > 2$$

$\longrightarrow$ M Molloy, B Reed, 1998

example: $G(n, p)$ model

plot of $\frac{\langle k^2 \rangle}{\langle k \rangle} = 1 + np$ (blue line) for $G(n, p)$ model with $n = 1000$

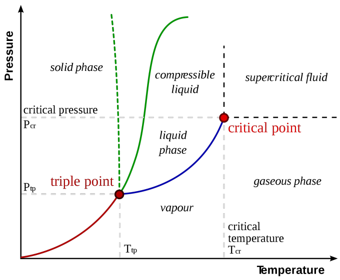
we expect a giant connected component for $p > 0.001$ but not for $p < 0.001$

Notes:

- So far we only argued about the critical point at which the giant connected component emerges but what about regions where the fraction $\frac{\langle k^2 \rangle}{\langle k \rangle}$ is smaller or larger than 2? So what happens if we increase the mean degree $\langle k \rangle = np$ in the $G(n, p)$ model?
- If we increase $\langle k \rangle$ we necessarily increase $\langle k^2 \rangle$ even more, so $\frac{\langle k^2 \rangle}{\langle k \rangle}$ increases. In other words: there is a monotonous relationship between the expected degree and the fraction $\frac{\langle k^2 \rangle}{\langle k \rangle}$. If there is (asymptotically) a giant connected component at the point $\frac{\langle k^2 \rangle}{\langle k \rangle} = 2$, there must also be a giant connected component for $\frac{\langle k^2 \rangle}{\langle k \rangle} > 2$.
- Hence, we have found a critical point for the **emergence of the giant connected component**.
- The resulting condition is called the Molloy-Reed criterion. It was first derived in the form above in → M Molloy and B Reed, 1998 .

Critical phenomena in complex systems

- **sudden emergence** of a giant connected component constitutes a **phase transition** in networks
- paradigmatic example for a class of **critical phenomena** in complex systems
- small change in a **control parameter** leads to **abrupt** **change in collective properties**
- critical/non-linear phenomena can make it hard to **predict the behavior of complex systems**



liquid-vapour phase diagram of water

exemplary critical phenomena

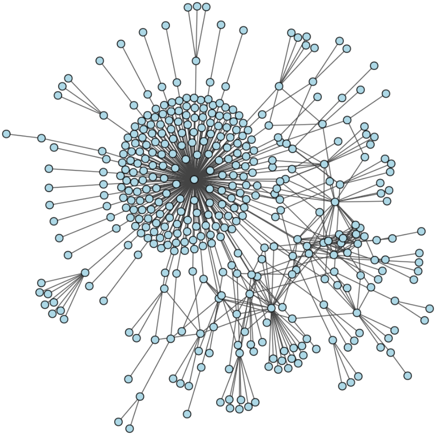
- percolation in porous media
- ferromagnetism
- phase transitions in thermodynamics
- epidemic threshold (e.g. $R_0 = 1$)
- avalanche / earthquake dynamics

Notes:

- This critical point in random graphs (with arbitrary degree distributions) is a paradigmatic example for **critical phenomena in complex systems**. We found a phase transition, i.e. a point in the parameter space of an ensemble at which the collective, *emergent properties* of microstates change abruptly.
- In the case of networks, the emergent property is the connectedness of the system. It changes abruptly as we vary the parameter p in the $G(n, p)$ model slightly. Similar problems are studied in different areas of statistical physics and material science.
- As an example consider the *percolation* of liquids through porous media. We can model this by a random lattice, where occupied cells block a “flow”. At which point can a liquid flow from top to bottom? A path of “open” cells from top to bottom is called a “percolating cluster”. Similar like the giant connected component, percolating cluster emerge abruptly (and the behavior of the cluster size near the critical point has been shown to depend on the dimensionality of the lattice).
- Another prominent example for phase transition phenomena can be found in thermodynamic systems: a small change in aggregate statistical quantities like, e.g. temperature or pressure, can lead to abrupt changes in the bulk properties of a material. A discussion of these and other similarities can be found in → JE Cohen, 1988

Robustness of complex networks

- **robustness** is an important systemic property
- is this network robust?
- depends on notion of robustness, i.e. “robust against what?”



network robustness

we consider how the removal of nodes and/or links affects the **connectedness** of the remaining network

- we do **not consider** failure cascades, i.e. **failures do not propagate** across links

example

collaboration network of developers and users in the Open Source community GENTOO

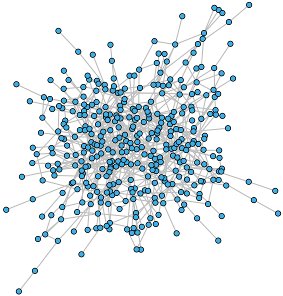
Notes:

- From a certain point of view, we have studied when the giant connected component emerges as we add more and more random links in a network that is initially empty. However, we can also ask when the giant connected component ogf an initially connected network disappears as we keep removing links. This is naturally related to the question how robust a network or system is.
- The question how "robust" a system is naturally leads us to ask: "robust against what?". Depending on the context, there are different notions of robustness: How robust is a **power grid** against **cascading failures**? How robust is a **social network** against the **spreading of diseases/fake news**? How robust is a **communication network** against **inadvertent synchronization** (e.g. for systems like router networks, in which synchronous exchanges of protocol messages are detrimental)? How robust is a **power grid** against a **loss of synchronization** (e.g. for systems like a power grid, where synchronization of power frequency is crucial for the system's function)?
- All of the above notions of robustness make sense and they have been studied for networked systems. However, in the following we limit ourselves to the simplest possible notion of **topological robustness**: the robustness of a system's connectedness against the loss of nodes and/or links.
- The benefit of this simple approach is that we do not need to model a dynamical process (like e.g. cascading processes, disease spreading). We can instead study this problem analytically by adapting the generating functions framework which we already introduced.

Systemic risk in complex networks

stochastic node failure model

- nodes fail uniformly at random
- uniform failure probability $P(X = v) \equiv q$
- we remove failed nodes as well as incident links



- how do such **random node failures** affect the network?
- how large are "surviving" connected components?
- can we calculate the **expected effect** in an ensemble of **random microstates**?

example

random microstate generated using Erdős-Rényi model with $n = 400$ nodes and $p = 0.0075$

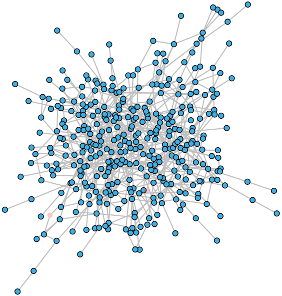
Notes:

- In general, we do not know *which* elements of a system are going to fail. We thus use a simple **stochastic failure model** for nodes that assumes that all nodes fail with a given probability q uniformly at random. This means that we consider a situation where (i) all nodes have the same failure characteristics, and (ii) node failures are independent.
- Both assumptions are probably unrealistic in real scenarios: the failure of one node can affect the remaining nodes, thus increasing their likelihood to fail. Moreover, in many systems some nodes are more or less likely to fail than others (because certain nodes may be particularly exposed to or protected against failures). But we need to start somewhere, so let us forget these complications and see what we can already learn from this simple model.
- To study the robustness of a network we will remove failing nodes from the network and study the relative size of the largest surviving connected component. We are eventually interested in the expected effect of random node failures in an ensemble of random microstates.
- To get an intuition for the effect of random failures, we study a random microstate generated by the Erdős-Rényi model (shown on the right).

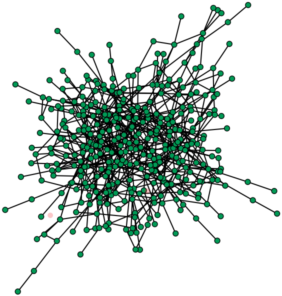
Example: random $G(n, p)$ microstate

example

random microstate generated by $G(n, p)$ model
with $n = 400$ and $p = 0.0075$
node failure probability $q = 0.01$



failed nodes (in red)



largest "surviving" connected component (in green)

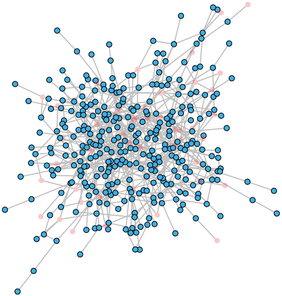
Notes:

- In a first experiment we let each node fail with probability $q = 0.01$, i.e. we expect 1 % of nodes to be removed. The failed nodes that are going to be removed are shown in red in the figure on the left.
- In the figure on the right we removed those (red) nodes as well as all of their links to other nodes. We color nodes in the largest connected component of the remaining network in green. We observe that the random failure of 1% of the nodes does not affect the connectivity of the surviving network at all, i.e. the surviving 396 nodes in the network still form a single connected component.

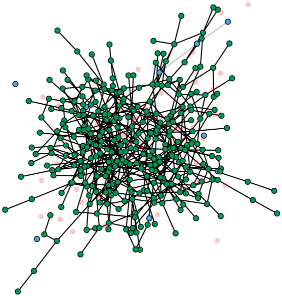
Example: random $G(n, p)$ microstate

example

random microstate generated by $G(n, p)$ model
with $n = 400$ and $p = 0.0075$
node failure probability $q = 0.1$



failed nodes (in red)



largest "surviving" connected component (in green)

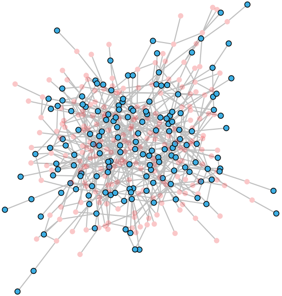
Notes:

- We repeat this experiment by increasing the failure probability of nodes by a factor of ten, i.e. we let 10% of the nodes fail uniformly at random. In the figure on the left, those 10% of failing nodes are again marked in red.
- The figure on the right shows the surviving largest connected component in the network, where I have removed all red nodes as well as their links. We find that, for this network, the failure of 10% of the nodes does not really affect the connectivity of the surviving network. In the surviving network, the vast majority of nodes still form a single connected component.

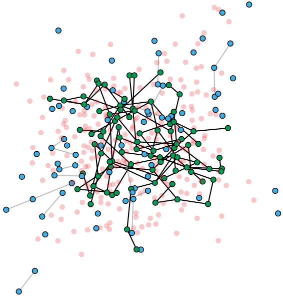
Example: random $G(n, p)$ microstate

example

random microstate generated by $G(n, p)$ model
with $n = 400$ and $p = 0.0075$
node failure probability $q = 0.6$



failed nodes (in red)



largest "surviving" connected component (in green)

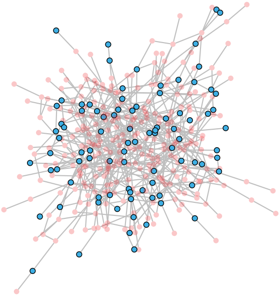
Notes:

- We further increase the failure probability to $q = 0.6$, i.e. we let more than half of the nodes (marked in red in the network on the left) fail.
- We now observe that a number of nodes in the surviving network shown on the right become disconnected from the largest connected component. What is remarkable is that the surviving network still has a largest connected component that spans the majority of nodes in the network. From a connectivity point of view, the system is not heavily affected even though more than half of the nodes were removed.

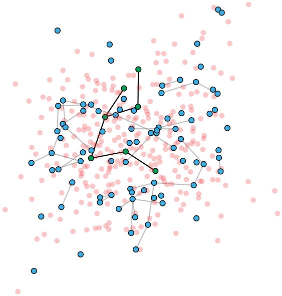
Example: random $G(n, p)$ microstate

example

random microstate generated by $G(n, p)$ model
with $n = 400$ and $p = 0.0075$
node failure probability $q = 0.75$



failed nodes (in red)



largest "surviving" connected component (in green)

Notes:

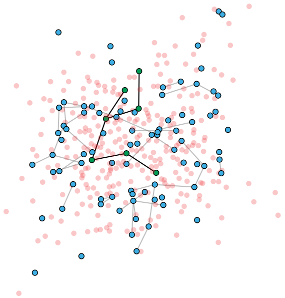
- We now further increase the failure rate to 75% of the nodes (again marked in red in the network shown on the left). This is a mere 15 % percent increase of the failure rate above our previous experiment.
- For this particular microstate, something interesting happens between $q = 0.6$ and $q = 0.75$: the giant connected component is destroyed and the surviving network is disrupted into many small connected components.
- Considering this non-intuitive effect, a natural question arises: Can we explain (and predict) this “breaking point” of the network analytically? How does this breaking point depend on the parameters of the Erdős-Rényi model? And can we make general statements for random networks with arbitrary degree distribution?

Generating function analysis of robustness

- generating functions help us to predict **emergence** of giant connected component
- how can we predict its **destruction?**
- we are interested in **surviving connected component sizes**

idea for analysis

1. consider random failures in **random microstate** from Molloy-Reed ensemble
2. assume that **degree distribution is generated by** G_0
3. mark **failed nodes** in red and **surviving nodes** in blue
4. calculate expected size of connected components for blue nodes



largest "surviving" connected component (in green)

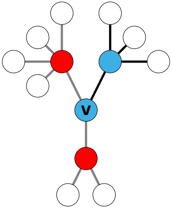
Notes:

- Considering that we used generating functions to predict the **emergence** of the giant connected component, it seems natural to adopt the same approach to predict the point at which the giant connected component is destroyed. We will thus modify our analysis from the last lecture.
- The idea behind our analysis is simple: Let us consider a random network with a fixed degree distribution, whose probability mass function is generated by a generating function G_0. We then apply our stochastic failure model, i.e. we assume that nodes fail uniformly (and independently) with a probability q.
- Rather than actually removing failed nodes, we simply mark failed nodes in red (assuming that the "surviving" nodes are marked in blue)
- With this, we can now reformulate the question about the size of the largest *surviving* connected component: what is the largest connected component that is only formed by blue nodes (i.e. we disregard all red nodes as well as their links)?
- If we can write down generating functions equivalent to G_0, G_1, H_0, and H_1 from last week's lecture, we can simply repeat the analysis that yielded the Molloy-Reed criterion.

Generating functions for surviving network

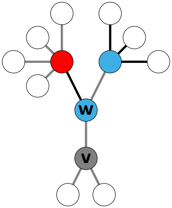
- for random node failure probability q, probability that **random node v has degree k and survives** is generated by

$$F_0(x) := \sum_{k=0}^{\infty} (1-q)\, P(k) x^k = (1-q)\, G_0(x)$$



- probability that **neighbour w of random node v survives and has k links** without (v, w) is generated by

$$F_1(x) := \sum_{k=0}^{\infty} \frac{\frac{k}{\langle k \rangle}(1-q) P(k) x^k}{x} = (1-q)\, G_1(x)$$



Notes:

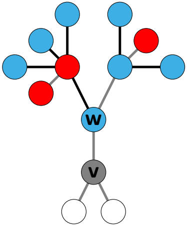
- Consider a random network with degree distribution $P(k)$ in which some nodes have failed (coloured in red). Note that the original degree distribution $P(k)$ includes links to surviving and failed nodes (i.e. to blue and red nodes).
- What is the probability that a node has survived **and** has degree k (in the original network)?
  - since these events are independent, we can multiply the probability $(1-q)$ that a randomly chosen node v survives with the probability $P(k)$ that v has degree k.
  - Replacing $P(k)$ in G_0 (defined in L07), we obtain a new generating function F_0 for which the k-th derivative generates $(1-q)P(k)$.
  - Note that for $q=0$ (no failures) we have $F_0 = G_0$. For $q=1$ (all nodes fail) we have $F_0 \equiv 0$. The latter case translates to the fact that the probability that nodes with *any* degree survive is zero.
  - Also note that we have relaxed the constraint that the coefficients of x^k are a properly normalized probability mass function, i.e. the coefficients do not sum to one. This means that we have $F_0(1) = (1-q) \leq 1$, which is the fraction of surviving nodes for failure probability q. We only consider the surviving nodes in the network, i.e. F_0 generates the part of the degree distribution that is due to surviving nodes.
- Using G_1 from L07, we define a function that generates the probability that a surviving node w, which we arrive at by following a random link, has k links (not counting the link we arrived through).

Surviving component sizes

- let distribution of **surviving component size** s **for random neigbour** w **of random node** v (in "forward" direction) be generated by H_1

  for $s = 0 \Rightarrow w$ has failed, i.e.

$$P(s=0) = \left[\frac{1}{0!}\frac{d^0}{dx^0}H_1(x)\right]_{x=0} = H_1(0) = q$$

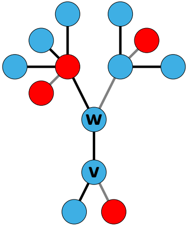


for $s \geq 0 \Rightarrow$ **recursive** definition $\rightarrow$ slide 6

$$H_1(x) := q + xF_1(H_1(x))$$

- distribution of surviving component size s for **randomly chosen node** v

$$H_0(x) := q + xF_0(H_1(x))$$



Notes:

- Using F_0 and F_1 we can define a new generating function H_1 that generates the distribution of **surviving** connected component sizes along the lines explained above. We again consider the connected component size in forward direction for a node w that we arrived at by following a random link.
- We observe two cases, which correspond to different derivatives of H_1 at the point $x = 0$:
    1. The first case is that the surviving component size is zero, which implies that node w itself must have failed. This happens with probability q and we must define H_1 such that $P(s=0) = \left[\frac{1}{0!}\frac{d^0}{dx^0}H_1(x)\right]_{x=0} = H_1(0) = q$ (remember that $0! := 1$). We observe that any reasonable candidate for our function $H_1(x)$ must have a constant term q that does not depend on x.
    2. The second case is that $s > 0$, which implies that node w has survived. This happens with probability $1-q$. This case corresponds to the higher derivatives of H_1 at $x = 0$. In this case we sum up the component sizes of neighbouring nodes (adding one for node w) just like we did in the definition of H_1 before. In other words, with probability $1-q$ (a factor already included in the definition of F_1), we sum up the forward component sizes of all neighbors of w by composing F_1 and H_1.
    3. Combining these two idea yields a new recursive definition for H_1.
- We finally use the same approach as before to define $H_0(x)$ based on $H_1(x)$.

Surviving component sizes 2/2

- with $F_1(x) = (1-q)G_1(x)$ we can calculate

$$H_1(x) = q + xF_1(H_1(x)) = q + (1-q)xG_1(H_1(x))$$

- with $F_0(x) = (1-q)G_0(x)$ we obtain

$$H_0(x) = q + xF_0(H_1(x)) = q + (1-q)xG_0(H_1(x))$$

- **expected size of surviving connected component** $\langle s \rangle$ is given by $H_0'(1)$

Notes:

- Similar to our analysis of the percolation phase transition, we now have everything to calculate the mean surviving connected component size under a random failure model in a random network with arbitrary degree distribution generated by G_0.
- We first consider our updated self-consistency condition for H_1. We substitute F_1 in the definition of H_1 by $(1-q)G_1(x)$ and obtain the expression above. We then substitute F_0 in the definition of H_0 by $(1-q)G_0(x)$ and obtain the expression for H_0 above.
- Just like before, we can calculate the mean component size based on the first derivative of H_0 at the point $x=1$. This analysis was first introduced in $\rightarrow$ DS Callaway, 2000

Expected surviving component size

- by applying the chain and product rule, we find

$$H_0'(1) = (1-q) + (1-q)G_0'(1)H_1'(1)$$

- we can further calculate

$$H_1'(1) = (1-q) + (1-q)G_1'(1)H_1'(1)$$

- solving by $H_1'(1)$ we then find the expression

$$H_1'(1) = \frac{1-q}{1-(1-q)G_1'(1)}$$

- for the **expected surviving component size**, we finally get

$$\langle s \rangle = H_0'(1) = (1-q)\left[1 + \frac{(1-q)\langle k \rangle}{1-(1-q)G_1'(1)}\right]$$

Notes:

- To calculate $H_0'(1)$ we need to apply the product rule $(f \cdot g)' = f'g + fg'$ and the chain rule $(f \circ g)' = f \circ g \cdot g'$
- We obtain the following:
  - With $H_0(x) = q + (1-q)xG_0(H_1(x))$ we first have $H_0'(x) = (1-q)G_0(H_1(x)) + x(1-q)G_0'(H_1(x))H_1'(x)$.
  - For $x = 1$ we get $H_0'(1) = (1-q) + (1-q)G_0'(1)H_1'(1)$.
- For $H_1'(1)$ we then have:
  - With $H_1(x) = q + (1-q)xG_1(H_1(x))$
  - We first have $H_1'(x) = (1-q)G_1(H_1(x)) + x(1-q)G_1'(H_1(x))H_1'(x)$
  - For $x = 1$ we get $H_1'(1) = (1-q) + (1-q)G_1'(1)H_1'(1)$
  - Solving the equation for $H_1'(1)$ we obtain $H_1'(1) = \frac{1-q}{1-(1-q)G_1'(1)}$
- Substituting $H_1'(1)$ in the expression $H_0'(1)$ above, we obtain an expression for the mean surviving component size that only depends on the (known) generating function G_1, the mean degree $\langle k \rangle$ given by $G_0'(1)$, and the node failure probability q

Survival of giant connected component

- analogous to our previous analysis, for $n \to \infty$ **survival of giant connected component** implies $\langle s \rangle \to \infty$, i.e.

$$\langle s \rangle = (1-q)\left[1 + \frac{(1-q)\langle k \rangle}{1-(1-q)G_1'(1)}\right] \to \infty$$

- **case 1:** $\langle k \rangle \to \infty$, i.e. networks are not sparse
- **case 2:** $(1-q)G_1'(1) \to 1$ for sparse networks
- for $G_1'(1)$ we found

$$G_1(x) = \frac{\sum_k^\infty kP(k)x^{k-1}}{\langle k \rangle} \Rightarrow G_1'(1) = \frac{\langle k^2 \rangle}{\langle k \rangle} - 1$$

- we expect **giant surviving connected component** for failure probabilities q below **critical failure rate** q_c

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

Notes:

- We use the same argument as before, i.e. for $n \to \infty$ the presence of a giant surviving connected component implies that the mean surviving component size diverges.
- In other words: we are interested in the critical point at which the mean size of the surviving connected component switches from infinite to finite (i.e. the critical point below which the giant connected component is destroyed).
- Considering the expression for the mean size of the surviving connected component, we again identify two cases in which the mean size is infinite:
    1. The first case corresponds to infinite mean degree, i.e. the network is not sparse. This is clear since, in a fully connected network, the surviving nodes always form a giant connected component for any failure probability q. So trivially such a network is maximally robust in the sense that no failure probability can ever destroy the surviving giant connected component.
    2. For sparse networks, we have $(1-q)G_1'(1) \to 1$, from which we can derive a critical failure rate q_c at which $\langle s \rangle$ changes from infinite to finite.
- This critical failure rate q_c bounds the failure probability from above, i.e. larger failure probabilities naturally result in smaller (finite) surviving connected components.

Robustness of Erdős-Rényi networks

- for **Erdős-Rényi networks** with n nodes and mean degree np

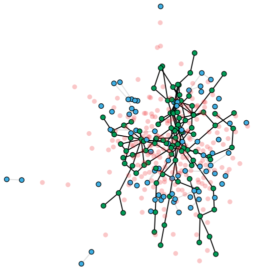
$$\frac{\langle k^2 \rangle}{\langle k \rangle} = \frac{np + n^2 p^2}{np} = 1 + np$$

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

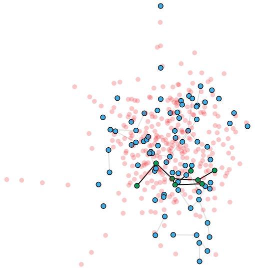
example

random microstate from $G(n, p)$ ensemble with $n = 400$, $p = 0.0075$

critical failure probability $q_c = 1 - \frac{1}{np} = \frac{1}{3} \approx 0.67$



$q = 0.6 < q_c$



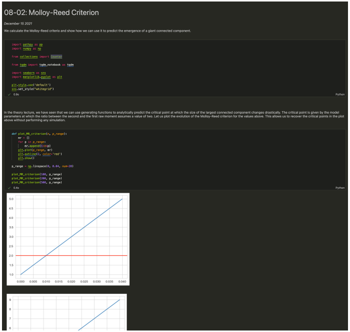
$q = 0.75 > q_c$

Notes:

- Let us apply this to our example of Erdős-Rényi networks. Using the first and the second raw moment of the Poisson distribution, we can derive a critical failure probability based on the parameters of the $G(n, p)$ model.
- Our motivating example was generated from a $G(n, p)$ model with $n = 400$ nodes and $p = 0.0075$ and thus $np = 3$.
- We analytically calculate a critical failure probability $q_c = 1 - \frac{1}{3} = \frac{2}{3}$, which explains our finding that the giant connected component is destroyed above a failure probability of 0.6 (cf. the figures on the right).
- This explains why even comparably sparse random Erdős-Rényi networks are remarkably robust against random node failures. We can use the expression above to calculate the critical point at which we expect the giant connected component to disappear for different values of the mean degree:
  - $np = 1 \Rightarrow q_c = 0$
  - $np = 2 \Rightarrow q_c = 0.5$
  - $np = 4 \Rightarrow q_c = 0.75$
  - $np = 10 \Rightarrow q_c = 0.9$
- Can you provide an intuitive interpretation for the critical point $1 - \frac{1}{np}$ based on the Molloy-Reed criterion?

Practice Session

- we use the Molloy-Reed model to calculate the critical point p_c of the $G(n, p)$ model and validate the Molloy-Reed criterion
- we implement a random node failure model and apply it to random networks
- we validate our results about the robustness of random networks for the Erdős-Rényi model



practice session

see notebooks 08-02 and 08-03 in `gitlab` repository at

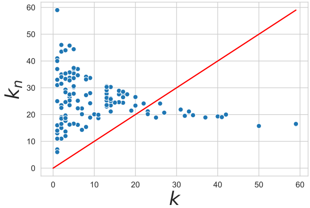
→ https://gitlab.informatik.uni-wuerzburg.de/ml4nets_notebooks/2022_wise_sna_notebooks

Notes:

- In the final two notebooks of this week's practice session, we empirically explore the Molloy-Reed criterion and implement a simple random node failure model.
- We further validate our results about the destruction of the giant connected components in random networks.

In summary

- we have applied **generating functions** to study statistical ensembles of networks with fixed degree distribution

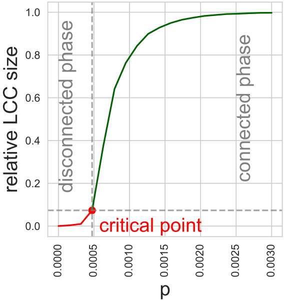


giant connected component

- sudden emergence at **critical point**
- critical point given by **Molloy-Reed criterion**
- example for **phase transitions in networks**

robustness against random node failures

- we can use generating functions to calculate expected size of **largest suriving connected component**
- giant connected component **suddenly disappears** as failure probability increases
- in **Erdös-Renyi networks**, tolerance against random node failures only depends on mean degree



Notes:

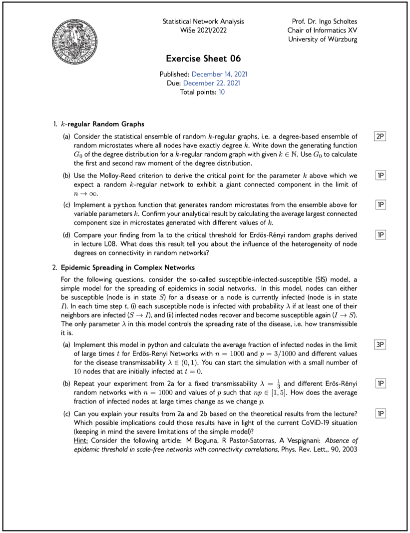
- In summary, we used generating functions to analytically study two important phenomena in networks based on statistical ensembles.
- We first studied the sudden emergence of the giant connected component in random networks with arbitrary degree distribution. We related the sudden emergence to the convergence behaviour of the expected component size, which can be calculated based on generating functions. We further derived an expression based on the first two raw moments of a degree distribution that we can use determine whether we expect a giant connected component.
- We further considered a simple random node failure model and studied its impact on random networks. We were able to update our generating function formalism such that we can express the expected size of the surviving connected component and we used this to derive a critical point for the node failure probability above which the giant connected component disappears.
- Applying generating functions to networks with scale-fee distributions, next week we will close the chapter on statistical ensembles of random networks.

Exercise sheet 06

- **sixth exercise sheet** available on WueCampus
  - apply generating functions to k-regular random graphs
  - calculate the connectivity phase transition in networks with zero variance
  - understand how the degree distribution of networks influences epidemic spreading

- solutions are due **December 14th** (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 06

Published: December 14, 2021
Due: December 22, 2021
Total points: 10

1. **k-regular Random Graphs**
   (a) Consider the statistical ensemble of random k-regular graphs, i.e. a degree-based ensemble of random microstates where all nodes have exactly degree k. Write down the generating function G_0 of the degree distribution for a k-regular random graph with given $k \in \mathbb{N}$. Use G_0 to calculate the first and second raw moment of the degree distribution. 2P
   (b) Use the Molloy-Reed criterion to derive the critical point for the parameter k above which we expect a random k-regular network to exhibit a giant connected component in the limit of $n \to \infty$. 1P
   (c) Implement a `python` function that generates random microstates from the ensemble above for variable parameters k. Confirm your analytical result by calculating the average largest connected component size in microstates generated with different values of k. 1P
   (d) Compare your finding from 1a to the critical threshold for Erdös-Rényi random graphs derived in lecture L08. What does this result tell you about the influence of the heterogeneity of node degrees on connectivity in random networks? 1P

2. **Epidemic Spreading in Complex Networks**
   For the following questions, consider the so-called susceptible-infected-susceptible (SIS) model, a simple model for the spreading of epidemics in social networks. In this model, nodes can either be susceptible (node is in state S) for a disease or a node is currently infected (node is in state I). In each time step t, (i) each susceptible node is infected with probability λ if at least one of their neighbors are infected ($S \to I$), and (ii) infected nodes recover and become susceptible again ($I \to S$). The only parameter λ in this model controls the spreading rate of the disease, i.e. how transmissible it is.
   (a) Implement this model in python and calculate the average fraction of infected nodes in the limit of large times t for Erdös-Renyi Networks with $n = 1000$ and $p = 3/1000$ and different values for the disease transmissability $\lambda \in (0, 1)$. You can start the simulation with a small number of 10 nodes that are initially infected at $t = 0$. 3P
   (b) Repeat your experiment from 2a for a fixed transmissability $\lambda = \frac{1}{3}$ and different Erös-Rényi random networks with $n = 1000$ and values of p such that $np \in [1, 5]$. How does the average fraction of infected nodes at large times change as we change p. 1P
   (c) Can you explain your results from 2a and 2b based on the theoretical results from the lecture? Which possible implications could those results have in light of the current CoViD-19 situation (keeping in mind the severe limitations of the simple model)?
   Hint: Consider the following article: M Boguna, R Pastor-Satorras, A Vespignani: *Absence of epidemic threshold in scale-free networks with connectivity correlations*, Phys. Rev. Lett., 90, 2003 1P

Notes:

Self-study questions

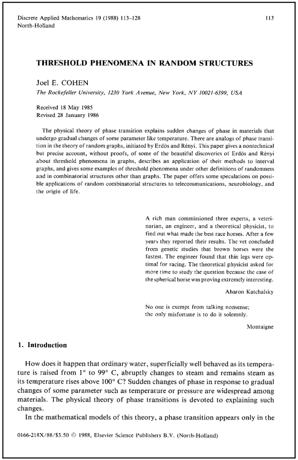
1. Under which conditions can we use generating functions to study connected components?
2. What is the condition for the existence of a giant connected component in the $G(n, p)$ and the $G(n, m)$ model?
3. What is the Molloy-Reed criterion?
4. In an Erdös-Rényi network, what does the Molloy-Reed criterion imply for $G_0'(1) = \langle k \rangle$ and $G_1'(1) = \langle k_n \rangle - 1$?
5. What is the condition for the existence of a giant connected component in the $G(n, p)$ and the $G(n, m)$ model?
6. In an Erdös-Rényi network, what does the Molloy-Reed criterion imply for $G_0'(1) = \langle k \rangle$ and $G_1'(1) = \langle k_n \rangle - 1$?
7. Use the Mollo-Reed criterion to calculate the critical point for the degree k of k-regular random graphs.
8. What is the critical failure probability q_c (uniform failure probabilities) for random networks with arbitrary degree distribution?
9. What is the critical failure probability q_c (uniform failure probabilities) for a $G(n, p)$ network?

Notes:

References

reading list

- P Erdős and A Rényi: **On the evolution of random graphs**, Publ. Math. Inst. Hung. Acad. Sci. 5, 1960
- JE Cohen: **Threshold phenomena in random structures**, Discrete Applied Mathematics, 1988
- M Molloy, B Reed: **A critical point for random graphs with a given degree sequence**, Random Structures & Algorithms, 1995
- M Molloy, B Reed: **The size of the giant component of a random graph with a given degree sequence**, Combinatorics, Probability and Computing, 1998
- MEJ Newman, SH Strogatz, DJ Watts: **Random graphs with arbitrary degree distributions and their applications**, Physical Review E, 2001
- SN Dorogovtsev, AV Goltsev, JFF Mendes: **Critical phenomena in complex networks**, Reviews of Modern Physics, 2008
- SN Dorogovsev, JFF Mendes: **Evolution of Networks**, Oxford University Press, 2003 ⟶ Chapter 6



Discrete Applied Mathematics 19 (1988) 113–128 113
North-Holland

THRESHOLD PHENOMENA IN RANDOM STRUCTURES

Joel E. COHEN
The Rockefeller University, 1230 York Avenue, New York, NY 10021-6399, USA

Received 18 May 1985
Revised 28 January 1986

The physical theory of phase transition explains sudden changes of phase in materials that undergo gradual changes of some parameter like temperature. There are analogs of phase transition in the theory of random graphs, initiated by Erdös and Rényi. This paper gives a nontechnical but precise account, without proofs, of some of the beautiful discoveries of Erdös and Rényi about threshold phenomena in graphs, describes an application of their methods to interval graphs, and gives some examples of threshold phenomena under other definitions of randomness and in combinatorial structures other than graphs. The paper offers some speculations on possible applications of random combinatorial structures to telecommunications, neurobiology, and the origin of life.

A rich man commissioned three experts, a veterinarian, an engineer, and a theoretical physicist, to find out what made the best race horses. After a few years they reported their results. The vet concluded from genetic studies that brown horses were the fastest. The engineer found that thin legs were optimal for racing. The theoretical physicist asked for more time to study the question because the case of the spherical horse was proving extremely interesting.

Aharon Katchalsky

No one is exempt from talking nonsense;
the only misfortune is to do it solemnly.

Montaigne

1. Introduction

How does it happen that ordinary water, superficially well behaved as its temperature is raised from 1° to 99° C, abruptly changes to steam and remains steam as its temperature rises above 100° C? Sudden changes of phase in response to gradual changes of some parameter such as temperature or pressure are widespread among materials. The physical theory of phase transitions is devoted to explaining such changes.

In the mathematical models of this theory, a phase transition appears only in the

Notes: