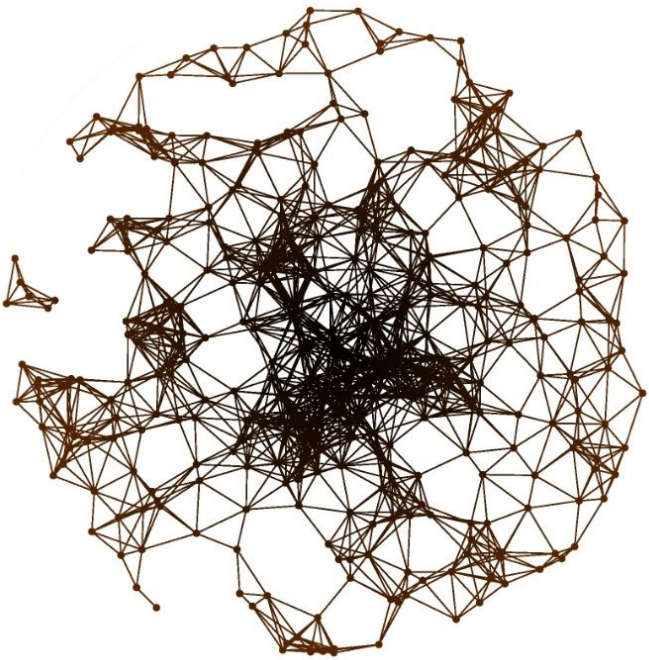
platform wiskunde nederland

Willekeur en structuur in netwerken

Syllabus Vakantiecursus 2022

Amsterdam, 26 en 27 augustus 2022

Eindhoven, 2 en 3 september 2022



platform wiskunde nederland

Willekeur en structuur in netwerken

Syllabus Vakantiecursus 2022

Amsterdam, 26 en 27 augustus 2022

Eindhoven, 2 en 3 september 2022

Programmacommissie

prof. dr. Wil Schilders (PWN, TU/e) (voorzitter)

dr. Jeroen Spandaw (TUD)

drs. Kees Temme (Gymnasium Hilversum, UVA)

dr. Benne de Weger (TU/e) (eindredactie syllabus)

drs. Peter Ypma (Goudse Waarden, Betapartners)

e-mail: vakantiecursus@platformwiskunde.nl

Platform Wiskunde Nederland

Science Park 123, 1098 XG Amsterdam

Telefoon: 020-592 4006

Website: http://www.platformwiskunde.nl

Vakantiecursus 2022

De Vakantiecursus Wiskunde voor leraren in de exacte vakken in HAVO, VWO, HBO en andere belangstellenden is een initiatief van de Nederlandse Vereniging van Wiskundeleraren, en wordt georganiseerd door het Platform Wiskunde Nederland. De cursus wordt sinds 1946 jaarlijks gegeven op het Centrum Wiskunde en Informatica te Amsterdam, en later ook aan de Technische Universiteit Eindhoven.

Deze cursus wordt mede mogelijk gemaakt door een subsidie van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), en een bijdrage van 4TU.AMI, het toegepaste wiskunde-instituut van de 4 Nederlandse technische universiteiten. Organisatie vindt plaats in nauwe samenwerking met het Centrum voor Wiskunde en Informatica (CWI) en de Technische Universiteit Eindhoven (TU/e).

De presentaties van de sprekers zullen zo veel mogelijk beschikbaar komen op de PWN-website: `https://www.platformwiskunde.nl`.

Met dank aan

Ondersteuning PWN: Sjoukje Talsma.

Historie

De eerste vakantiecursus wordt in het jaarverslag 1946 van het Mathematisch Centrum als volgt vermeld:

> *Op 29 en 31 Oct. '46 werd onder auspiciën van het M.C. een druk bezochte en uitstekend geslaagde vacantiecursus gehouden voor wiskundeleeraren in Nederland. Op 29 October stond de wiskunde, op 31 October de didactiek van de wiskunde op de voorgrond. De sprekers waren: Prof.Dr. O. Bottema, "De prismoide", Dr. A. Heyting, "Punten in het oneindige", Mr. J. v. IJzeren, "Abstracte Meetkunde en haar betekenis voor de Schoolmeetkunde.", Dr. H.D. Kloosterman, "Ontbinding in factoren", Dr. G. Wielenga, "Is wiskunde-onderwijs voor alpha's noodzakelijk?", Dr. J. de Groot, "Het scheppend vermogen van den wiskundige" en Dr. N.L.H. Bunt, "Moeilijkheden van leerlingen bij het beginnend onderwijs in de meetkunde".*
>
> *Aan het einde van de vacantiecursus werden diverse zaken besproken die het wiskunde-onderwijs in Nederland betroffen. Een Commissie werd ingesteld, die het M.C. over de verder te organiseren vakantiecursussen van advies zou dienen. Hierin namen zitting een vertegenwoordiger van de Inspecteurs van het V.H. en M.O. benevens vertegenwoordigers van de lerarenverenigingen Wimecos en Liwenagel.*
>
> *Ook werd naar aanleiding van "wenschen" die tijdens de cursus naar voren gekomen waren ingesteld: "een colloquium over moderne Algebra, een dispuut over de didactiek van de wiskunde, beiden hoofdzakelijk bedoeld voor de leeraren uit Amsterdam en omgeving, terwijl tevens vanwege het M.C. een cursus over Getallenleer werd toegezegd te geven door de heeren v.d. Corput en Koksma. (Colloquium, dispuut en cursus zijn in 1947 gestart en verheugen zich in blijvende belangstelling).*

Docenten

prof. dr. Nelly Litvak (hoofddocent)
Faculty of Electrical Engineering, Mathematics and Computer Science,
Universiteit Twente,
Department of Mathematics and Computer Science,
Technische Universiteit Eindhoven
web: `https://people.utwente.nl/n.litvak`
e-mail: `n.litvak@utwente.nl`

dr. Pim van der Hoorn
Department of Mathematics and Computer Science,
Technische Universiteit Eindhoven
web: `https://www.tue.nl/en/research/researchers/pim-van-der-hoorn/`
e-mail: `w.l.f.v.d.hoorn@tue.nl`

dr. Clara Stegehuis
Faculty of Electrical Engineering, Mathematics and Computer Science,
Universiteit Twente
web: `https://www.clarastegehuis.nl/`
e-mail: `c.stegehuis@utwente.nl`

Programma

Vrijdag 26 augustus 2022 / 2 september 2022

15.00–15.30		*Ontvangst, koffie*
15.30–15.35		*Openingswoord*
15.35–16.20	Nelly Litvak	Modelleren van schaarse netwerken met de Erdős-Rényi stochastische graaf
16.20–16.45		*Pauze*
16.45–17:30	Nelly Litvak	Kleine patronen tellen in stochastische grafen
17.30–18.30		*Diner*
18.30–19.15	Nelly Litvak	Vrijwel zekere garantie dat een stochastische graaf is verbonden
19.15–19.45		*Pauze*
19.45–20.30	Nelly Litvak	Modelleren van schaalvrije netwerken

Zaterdag 27 augustus 2022 / 3 september 2022

09.00–10.00		*Ontvangst, koffie*
10.00–10.45	Nelly Litvak	Opkomst van power laws in het preferential attachment model
10.45–11.15		*Pauze*
11.15–12.00	Nelly Litvak	Geometrie voor het modelleren van driehoeken
12.00–13.00		*Lunch*
13.00–13.45	Pim van der Hoorn	Wie is het belangrijkst in een netwerk?
13.45–14.30	Clara Stegehuis	Hoe maak je een netwerk efficiënt?
14.30		*Afsluiting*

1 Preliminaries, content and goals of this course

Nelly Litvak

1.1 Complex networks, modeled as random graphs

Many real-life systems are *networks*. A network is a set of objects connected by some relationship. For example, a railroad is a collection of stations connected by rails. In a social network, people are connected by friendships. Internet is a network of routers connected by wires. In our brain, neurons are connected if they fire together.

A *graph* is a natural mathematical model for a network of any nature. In a graph, each object is represented as a *vertex*, and if there is a relationship between two vertices, then there is an *edge* between them. *Undirected edges* represent symmetric relations, and we draw them as lines. For example, communications between two Internet routers usually go in both directions. If the relation is not symmetric, we model this using *directed edges* and use arrows to draw them. For example, if somebody follows you on Twitter, you might not follow back.

Self-test: Look at the networks in Figure 1.1a–1.1d. What are the vertices and what are the edges? Are the edges directed or undirected? The answer will be given in Section 1.4.

In this course we will learn to model large real-life networks, such as social networks or the World Wide Web, using so-called *random graphs*.

Make the next step yourself: What do you think exactly is *random* about a random graph? There is no wrong answer, just think logically yourself, and in the next line we will explain what 'random' means in this course.

Usually in research, and always in this course, we will assume that in a random graph the vertices are fixed, but the edges are placed at random. This makes sense because relationships between objects often emerge at random, like friendships in a social network. Also, even if the network is not random, such as the Internet, its structure is so complicated that it is often useful to describe it using statistical summaries and model as a random object. A *random graph model* is in fact a set of rules, according to which the random edges are chosen. Different rules result in different models with different mathematical properties.

For example, in one model, edges between all vertex pairs can be equally likely, while other models assign higher probabilities to some vertex pairs.

Self-test: Assume you want to construct a random graph of n vertices. What is the easiest way to place edges at random?

- Can you write a formal mathematical description of this random graph model? If so, write it down. If not, what is on your way?
- Can you write down an algorithm that generates such random graph on a computer? If so, write it down. If not, what is on your way?

When we reach Chapter 3, look back at this exercise: is the rule that you came up the same as the Erdős-Rényi random graph model?

In this course we will study several by now well understood, even classical, random graph models. More specifically, we will study how particular rules of placing random edges result in graphs that share some of the fundamental empirical properties of real-life networks. In Section 1.4 we will list the properties of real-life networks that we will learn to model in this course. The rest of the chapters are about mathematical formalization of these properties and how they emerge from the edge-placing rules defined by the random graph models.

1.2 Position of this course with respect to related domains

In this section we will briefly explain how this course aligns with related domains such as 'graph theory', 'network science', and 'data science'.

The branch of mathematics that studies graphs is called *graph theory*. It has a long history dating back to Leonhard Euler in 1736[1]. The *theory of random graphs* is much younger. It has spurred from the graph theory in the pioneering work by Paul Erdős and Alfréd Rényi in the 1950s[2].

It is interesting that, initially, random graphs were invented and used to solve difficult graph-theoretic problems. This line of research has by now matured into the so-called *probabilistic method* in graph theory[3]. This is of course a very different purpose than modeling, say, the World Wide Web, which did not even exist when random graphs were invented! The fact that the E-R model turned out to be useful for understanding networks is another brilliant demonstration of

[1] The seven bridges of Königsberg, in: The world of mathematics, Simon and Schuster, New York, 1956, pp. 573–580.

[2] On random graphs I, Publicationes Mathematicae Debrecen, Vol. 6, 1959, pp. 290–297.

[3] Alon, Noga and Spencer, Joel H., The probabilistic method, John Wiley & Son, 2016.

how mathematics can be useful for the purposes that did not even exist when this mathematics was created. There are numerous such examples in science, most notable perhaps is the relatively recent use of the number theory in cryptography.

Today the theory of random graphs has largely diverted from the traditional scope and methodology of graph theory. It is now a well developed and quickly growing area of modern mathematics that provides mathematical models and tools for studying real-life networks. As such the theory of random graphs is contributing into a broad interdisciplinary domain of *network science*. Methodologically, the theory of random graphs often builds on *statistical mechanics*, *theoretical probability*, and *statistics*.

The emergence of the theory of random graphs in its current form is motivated by the overwhelming interest in networks that took off around the end of 1990s. Of course, the networks were studied long before that. In 1965, Derek de Solla Price analyzed the network of scientific citations. In such networks, the vertices are scientific papers, and a (directed) edge means that one paper cites another[4]. Social networks were studied, too. Milgram conducted his famous small-world experiment in 1960s. The work by Granovetter on the 'strength of weak ties' dates back to 1977.

So why this sudden massive interest? This has everything to do with computer technology and data. The Internet is a giant network itself, and it gave rise to the World Wide Web, Facebook, Twitter, Wikipedia, and many other online networks that greatly influence our life. Moreover, we now have the technology to stream, store, analyze and share large amounts of data, including the network data. Data became a crucial game changer in studies of networks. This way, the theory of random graphs, together with the modern network science are integrated into a broader scope of *data science*.

Self-test. Can you draw a diagram to show the relation of the domains 'graph theory', 'theory of random graphs', 'networks science', and 'data science'? There is no one right answer, just check your own understanding.

1.3 The goal of this course

In this course we will describe and analyze mathematically several fundamental properties of complex networks using the theory of random graphs. For that, we will study a number of basic random graph models and discuss their pros and cons for modeling real-life networks.

At the end of the course the students will be able to:

- describe mathematically the empirical properties of real-life complex net-

[4]Derek J. De Solal Price, Networks of scientific papers, Science, 1965, pp. 510–515.

works;

- choose and explain a random graph model that adequately represents these properties;
- provide a mathematical argument why the property of interest is represented adequately in the chosen random graph model.

The course is designed for MSc-level students and professionals with a large variety of technical backgrounds:

- The students with greater interest in mathematical theory, may dive deeper into mathematical derivations and proofs.
- The students with interest in numerical studies and applications may instead choose to investigate properties of random graphs empirically using simulations.

Self-test.

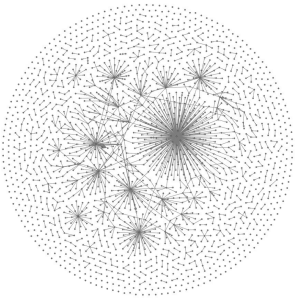
- What do you want to learn in this course? Can you write it down or say it out loud specifically?
- Do you prefer to prove theorems or to code and run numerical experiments? What are the advantages and the disadvantages of each of these research methods?

1.4 Properties of real-life networks addressed in this course

The connections in complex networks, such as hyperlinks in the World Wide Web, or online friendships, appear in rather unpredictable ways. Yet, surprisingly, many networks of a completely different nature share common properties. This is why we talk about a particular *structure* of a network. The presence of a predictable structure does not contradict the fact that individual network connections are random, because when random connections occur on a massive scale, they form clear patterns. These patterns are exactly what we mean by the 'structure of a network'. This structure can be captured in a mathematical model, and in this course we will learn how to do this for large real-life networks.

In this course we will address four such patterns, or structural properties, of complex networks.

Sparse Consider a social network. Even if the network is very large, one can maintain only so many friendships. We say that social networks are often *sparse*, meaning that the number of connections per person is limited, and

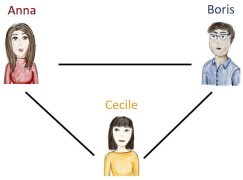


(a) Network of retweets about the ProjectX in Haren, Neltherlands 21-09-2012 7:00. The average number of retweets per user is small compared to the network size. This network is *sparse*. Image: Marijn ten Thij.

(b) Any railway station can be reached by train from any other. The railway network is *connected*. Image: Wikimedia Commons

# other webpages, from which a webpage can be reached by following a hyperlink	EU Web crawl 2015
average	85.743
largest	20 252 239

(c) The number of other webpages, from which a webpage can be reached varies greatly per webpage. The Webgraph is *scale-free*. Source: P. Boldi et al. BUbiNG: Massive crawling for the masses. In *WWW 2014*, `https://law.di.unimi.it/datasets.php`.



(d) Since Anna is a friend of Boris and Cecile, it is likely that Boris and Cecile are friends, too. Social networks have many *triangles*. Drawing: Natalia Litvak.

Figure 1.1

does not increase very much with the network size. Figure 1.1a shows an interesting example: the network of retweets about Project X in Haren, the Netherlands, in 2012. A birthday invitation of a 16-year-old girl went viral in social media and ended up in a destructive riot. Dots are Twitter users and each tiny arrow (a directed edge) represents a retweet from one user to another. The figure shows this network in the morning before the riot. We see that the network is sparse, on average there are only 1.5 retweets per user. Over the night of the riot the network increased in size more than 10 times, but the average number of retweets remained small, it went only a little bit above two.

Self-test. Can you give an example of a sparse real-life network? Why

do you think it is sparse? Can you explain the reason why in this network the average degree does not grow with the network size?

Connected Consider the network of railway stations connected by railroads, as the NS network in Figure 1.1b. The vertices are the stations, and the (undirected) edges are the railway connections between them. A passenger can travel by train from any station to any other. The railroad network is *connected*. The Internet is another powerful example of a connected network. Internet is extremely complex and completely decentralized, yet, the data can be transferred across the planet from any Internet router to any other!

Self-test. Can you give an example of a connected network? Why do you think it is connected? Do you think the World Wide Web is connected?

Scale-free In the Web graph, vertices are the webpages, and (directed) edges are the hyperlinks. By clicking on a hyperlink, we can go from one webpage to another. From how many other pages a typical webpage can be reached? In Figure 1.1c we show the average and the maximum of this number in the `.eu` domain of the Web graph in 2015. We see that on average, a webpage can be reached from 85.7 other webpages. However, this number differs from one webpage to another, and the maximum is over 200 000 times larger than average! We say that such network is *scale-free*. This unusual term means that there is no such thing as a 'typical webpage'. The number of hyperlinks pointing to a page can have very different scales – from a few, to hundreds, thousands, and millions.

Self-test: Can you give an example of a scale-free network? Why do you think it is scale-free? Can you give an example of some other quantities that are scale-free in the sense that the maximum is by many orders of magnitude larger than average?

Answer: Some other examples of scale-free quantities are: incomes of people, city sizes, and sizes of files sent over the Internet.

Triangles How do people in social networks usually meet? Often I know friends of my friends, they can be my friends, too. Groups of friends create clusters with many *triangles*, such as in Figure 1.1d: if Anna is a friend of Boris and Cecile, then it is not surprising that Boris and Cecile know each other as well. Having many triangles is another typical structural property of complex networks.

Self-test: Can you give an examples of networks where you may expect many triangles? What other connection patterns among small groups of people might be typical for a social network?

Surprisingly, many real-life networks of a completely different nature (social net-

works, the Internet, networks of neurons in the brain, networks of bank transactions, protein-protein interactions, etc.) share these properties: they are *sparse*, *connected*, *scale-free*, and have many *triangles*. In this course we will learn how to describe these structural properties of random connections in mathematical terms, and how to capture them in random graph models.

1.5 Communities, small world, and other properties of real-life networks

There are many other very interesting and common structural properties of real-life networks that are not part of this course. For example, we often see *communities*. In social networks, communities can be defined by interests, language, or geography.

Another famous property of real-life networks is the *'small world phenomenon'*: most pairs of vertices are connected by a short path of edges. In social networks, this phenomenon is also known as 'six degrees of separation' stating that *'everybody on this planet is separated only by six other people'* (John Guare).

The communities and the small world phenomenon, of course, too, have been studied using random graphs. The research on random graphs and complex networks is happening right now, and the authors of these notes are a part of this collective scientific effort.

We hope that this course will equip you with mathematical tools for thinking about complex networks, and leave you with exciting feeling of exploration and endless opportunities of this quickly developing branch of modern mathematics.

1.6 Work in progress

As a teacher I believe that it is not very important what I write or say, it is mostly important what students do.

What should students do in a course on random graphs? One of the best ways of learning is self-testing, this is why I want this book to be full of self-tests. At each step, I want to give a very coherent and small piece of theory and let the students test right away whether they've got it right. I also want to explain the answer right away so that the students get feedback immediately.

Writing such a syllabus is not easy, and it so happened that I had an impossible dead-line before the Vakanticursus 2022. I chose to stick to my idea of what a good syllabus should be, which means that I could not complete all chapters. I have completed this introductory Chapter 1, Chapter 2 that introduces main

techniques, and Chapters 3 and 4 for classes 1 and 2. Unfortunately, for classes 3–6, I had to resort to only a short summary.

I know this is not ideal but I feel I could not do this differently at this point. If you are not satisfied with the syllabus in its current form, I am really sorry. I also apologize for the multiple typo's that I – no doubt – have made and haven't found. But I am also very curious about feedback at the Vakanticursus 2022 on the style of the chapters that I managed to complete!

Thank you for joining this course. We will do our best to make it interesting and useful for you. I will of course continue working on this syllabus. I hope to share a more complete and better version with you soon enough.

Kind regards, Nelly Litvak.

Enschede, 16-08-2022

2 Network as a graph

Nelly Litvak

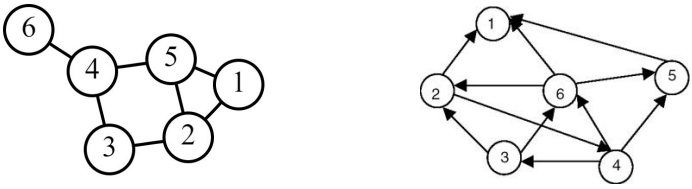
2.1 Mathematical representation of vertices and edges

A network is modeled as a graph $G = (V, E)$.

The capital letter V denotes the set of vertices. We will associate vertices with numbers: $V = \{1, 2, \ldots, n\}$, and write this as $V = [n]$. Usuallly we will denote vertices by small letters $i, j, k \in V$.

The capital letter E denotes the set of edges. Mathematical description of the edges is our first step to the abstract mathematical representation of a network.

In Figure 2.1a we see an undirected graph, where edges are drawn as lines. In Figure 2.1b we see a directed graph, where edges are drawn as arrows. This visual representation is clear, but it is not suitable for operating with graphs in a mathematical derivation or a computer program. For these purposes, we need a formal mathematical definition of an edge.



(a) Undirected graph

(b) Directed graph

Figure 2.1

Make the next step yourself. If $i, j \in V$ are two vertices, how would you denote an edge between them? For example, in Figure 2.1a on the left, there is an undirected edge between $i = 1$ and $j = 2$. How would you write down this edge as a mathematical object? And how would you write down the directed edge from $j = 2$ to $i = 1$ in Figure 2.1b? What will be the difference between the notation for a directed and an undirected case?

Clearly, an edge is defined by a pair of vertices, so the notation for an edge between i and j will involve both i and j. For example, we could write ij. But then, how do we know whether this edge is directed or undirected?

Mathematical representation of undirected edges. In the graph theory, and in the theory of random graphs, we mathematically represent an *undirected* edge as an *unordered set* of two vertices. So, the edge between i and j is denoted by $\{i, j\}$. The curly brackets are used because it is a standard notation for an unordered set. *'Unordered'* means that the order of i and j is irrelevant. The formal way to write it, is:

$$\{i, j\} = \{j, i\} \text{ for all } i, j \in V.$$

This equality makes sense because in an undirected graph the edge from i to j is the same as the edge from j to i (think, for example, of edge $\{1, 2\}$ in Figure 2.1a). When there are several edges between i and j we may index them as in $\{i, j\}_1$, $\{i, j\}_2$, etc.

In this course we will mainly deal with undirected graphs.

Mathematical representation of directed edges. In the graph theory, and in the theory of random graphs, we mathematically represent a *directed* edge as an *ordered set*, or, a *vector*, of two elements. So, the edge from i to j is denoted by (i, j). The round brackets are used because it is a standard notation for a vector. In a vector, the order of its elements is improtant. The formal way to write it is:

$$(i, j) \neq (j, i) \text{ if } i \neq j.$$

This inequality makes sense because in an directed graph the edge from i to j is not the same as the edge from j to i. For exmaple, in Figure 2.1b, we see an edge $(2, 1)$, but edge $(1, 2)$ does not exist in this directed graph.

2.2 Key technique: A sum of indicators

It is very important that you understand this section completely because this technique will be used throughout the course

Pre-requisites. The following preliminary knowledge is required for this section:

- Probability distribution of a discrete random variable (required);
- Expectation of a discrete random variable (required);
- Law of large numbers (desirable);
- Linearity of expectations (required).

In what follows we assume that you are able to *explain* and *apply* these notions. If you feel that this is not the case, please revise your knowledge before proceeding further.

As we discussed before, in a random graph, an edge may exist or not. More precisely, existence of an edge in a random graph is a *random event*. This kind of randomness with a binary outcome – yes or no – is very common among random phenomena. In probability theory, such random binary outcome is fomally described using so-called *Bernoulli* random variables. A Bernoulli random variable is 1 for 'yes', and 0 for 'no'.

So, in a random graph, there is a Bernoulli random variable corresponding to each edge. It is common to write down these Bernoulli random variables as so-called *indicators*. In the theory of random graphs, indicators are very useful because we can write many relevant quantities, such as the number of edges or triangles, as sums of indicators. This is very good news because from probability theory we know a lot of properties of sums of indicators. Therefore, once we have written a numerical quantity as a sum of indicators, we can derive analytically many properties of this quantity, and eventually derive many useful results for a random graph.

We will use indicators and their sums throughout this course. The goal of this section is to introduce this technique. In Section 2.2.1, we will explain what the indicators are, and in Section 2.2.2, as an example, we will use indicators to derive the formula for the expected number of edges.

2.2.1 An indicator of an edge

Assume that $G = ([n], E)$ is an undirected random graph. Each edge $\{i, j\}$ may exist or not. When edge $\{i, j\}$ exists, it is a part of the edge set E. Therefore, we can formally write the event $[\textit{edge } \{i, j\} \textit{ exists}]$ as $[\{i, j\} \in E]$ (we used the square brackets to separate the description of an event from the rest of the text.)

Denote by p_{ij} the probability that edge $\{i, j\}$ exists. We can formally write this as

$$\mathbb{P}(\{i, j\} \in E) = p_{ij}, \quad i, j \in [n].$$

Self-test. Why do we add $i, j \in [n]$ at the end of the formula above? Will the meaning of the formula change, if we do not add it? First attempt answer yourself, then read the answer. It is very useful to give an answer, even (especially!) a wrong one. Research shows that learning happens when we make mistakes.

Answer. We add $i, j \in [n]$ to state that the formula holds for all possible i and j in $[n]$. We *must* add this because the formula is complete only if all

symbols are defined. If we do not add this, then the formula is incomplete and carries no information because we do not know for which values of i and j the formula is true.

In *any* formula we *always* must specify the *range* of variables. This can be done in the formula itself, as we did above, or in the text.

Now we will give the main definition of this section. We introduce the *indicator* of edge $\{i, j\}$, and we denote this indicator by I_{ij}.

Definition 1. *For each $i, j \in [n]$, the indicator of edge $\{i, j\}$, denoted by I_{ij}, is a Bernoulli random variable that is 1 if $\{i, j\} \in E$ (edge $\{i, j\}$ exists), and 0 if $\{i, j\} \notin E$ (edge $\{i, j\}$ does not exist).*

The *probability distribution* of I_{ij} is given by the following table, where the values are written in the first line, and the corresponding probabilities are written in the second line:

I_{ij}	0	1
probability	$1 - p_{ij}$	p_{ij}.

, $\quad i, j \in [n]$.

Self-test. Write down the probability distribution of I_{ij} yourself, without looking at the table above. Do you have any slightest difficulty with this? What is the source of this difficulty? Is the difficulty in understanding the definition of the indicator? Then make a small example, e.g. take $n = 4$, draw a graph of 4 vertices, and go through this section again, translating each step into your example. Is the difficulty in understanding the table? Then revise the topic 'Probability distribution of a discrete random variable'.

We can now easily obtain the *expectation* of I_{ij} as follows:

$$\mathbb{E}(I_{ij}) = 0 \cdot (1 - p_{ij}) + 1 \cdot p_{ij} = p_{ij}, \quad i, j \in [n]. \tag{2.1}$$

In words, the expectation of indicator I_{ij} is the probability that edge $\{i, j\}$ exists. We can explain this intuitively as follows. If we construct the random graph infinitely many times, then, in the limit, by the law of large numbers, p_{ij} will be the fraction of times when our graph will contain edge $\{i, j\}$.

Self-test. Write down the derivation of $\mathbb{E}(I_{ij})$ yourself, without looking at the derivation in (2.1). Do you have any slightest difficulty with this? What is the source of this difficulty? Is the difficulty in understanding the definition of the indicator? Then make a small example, e.g. take $n = 4$, draw a graph of 4 vertices, and go through this section again, translating each

step into your example. Is the difficulty in understanding and computing the expectation? Then revise the topic 'Expectation of a discrete random variable'.

In the next section we will use the sum of indicators for the first time, for computing the expected number of edges.

2.2.2 Example: Expected number of edges

Recall that we work with a random graph $G = ([n], E)$, where E is a set of edges placed at random. We assume that the random graph is undirected. We also assume that the random graph is *simple*. The statement 'graph is simple' means that the graph does not have double edges (no double edges means that there cannot be more than one edge between two vertices) and does not have self-loops (a self-loop is an edge $\{i, i\}$, from vertex $i \in [n]$ to itself).

We are now interested in the total number of these random edges. The number of edges is the size of set E. We will denote the size of E using the standard notation $|E|$.

The quantity $|E|$ is a random variable because the existence of each edge is a random event. For example, in a graph of 3 vertices, $|E|$ can be 0 when no edge exists, and 3 when all edges exist.

Our goal now is to compute $\mathbb{E}(|E|)$, the expectation of the random variable $|E|$. How can we do this? The notation $|E|$ is compact and clear, but we cannot do any computation with it. Hence, we need to write down $|E|$ in a different way, that yields itself for analysis.

We will now proceed writing $|E|$ as a sum of indicators. In this example we will do this in great detail. In the rest of the course, even in more complicated examples, we will assume that the student understands this technique, and we will use it without additional explanation. Therefore, please make sure that you are able to explain each step in full sentences. If you can do so before reading our explanations, you may skip some of the explanations.

Make the next step yourself: Can you write $|E|$ as a sum of indicators I_{ij}'s? As you will see below, there are several ways to do it, and we will discuss all of them, but right now try to produce at least one. You may use the hint below.

Hint. Notice that each edge contributes exactly 1 to $|E|$ when it exists, adn 0 when it does nto exist.

Don't know how to proceed? If you cannot write down the sum, please write down what exactly holds you back. After that, you may proceed with reading, but when you finish the section, please come back to the

question what exactly was holding you back, and write down which part of the explanation helped you to overcome this difficulty.

Below we will first write four possible expressions, and after that we will explain each of them in detail:

$$|E| = \sum_{\{i,j\}\in E} 1 = \sum_{\{i,j\}\in E} I_{ij} \overset{\text{crucial step!}}{=} \sum_{\{i,j\}\subset[n]} I_{ij} = \sum_{i=1}^{n}\sum_{j=i+1}^{n} I_{ij}. \tag{2.2}$$

We will now explain all four expressions one by one.

The first sum $\sum_{\{i,j\}\in E} 1$ is simply a definition of $|E|$. We merely add 1 for each existing edge. The range of summation (written under the summation sign), is the set of edges E. In other words, this sum goes through all existing edges and adds 1 for each of them. The *summands* in this formula are *deterministic* (all equal 1), but this summation has a *random range* E. Why is this a problem? Because, if we want to compute, e.g., $\mathbb{E}(|E|)$ from this formula, we will have to compute a weighted sum over all possible *values* of E. Recall that E is not a number, it is a *set* of edges, so its values are sets as well. For example, in a graph of $n = 3$ vertices, E can be $\{\{1,2\},\{1,3\},\{2,3\}\}$ if all vertices are connected, it can be $\{\{1,3\}\}$ if $\{1,3\}$ is the only edge, and it can be $\emptyset$ if no edge exists. Altogether, for $n = 3$, there are $2^3 = 8$ possible values of E. This is quite many for such a small graph! What will happen for the general n? Using S as a running index of all possible values of E, we arrive at

$$\mathbb{E}(|E|) = \sum_{S\subseteq\left\{\{i,j\}:i,j\in[n]\right\}} |S|\cdot\mathbb{P}(E=S).$$

This sum contains $2^{\frac{n(n-1)}{2}}$ terms. (Do you know why $2^{\frac{n(n-1)}{2}}$? If not, no problem, you may skip this for now, we will come back to this in Chapter 4.) This is a huge number, it has 14 digits already for $n = 10$. The formula maybe simplifies in a simple models, but when we consider a somewhat realistic model, e.g. with inhomogeneity and/or dependence between edges, going through all possible sets of edges is simply impossible!

This is exactly the reason why we want to use indicators instead. Rather than summing *deterministic numbers* over a *random set*, we prefer to sum *random variables* over a *deterministic set*. Advantages: 1) Avoid the enumeration of sets. 2) Probability theory has many strong and useful results on sums of indicators.

The second sum $\sum_{\{i,j\}\in E} I_{ij}$ introduces indicators. It is identical to the first sum because we sum over all edges $\{i,j\}\in E$, and for these edges we have that

$I_{ij} = 1$. In other words, in the second sum, we simply replaced the 1's by I_{ij}'s that are equal to one.

Self-test. In the second sum, is the range of the summation deterministic or random? Are the summands deterministic or random? The answer is in the next paragraph.

The range of summation in the second sum did not change, it is still a random set E. The summands I_{ij} are de facto deteministic because we sum only over edges $\{i, j\}$ that exist in the graph. The usefulness of the second sum is in injecting indicators into the formula.

We will now make a crucial step from the second sum to the *third* sum, $\sum_{\{i,j\} \subset [n]} I_{ij}$.

Make the next step yourself. What is the difference between the second and the third sum? Can you explain why the third sum is equal to the second sum? In the third sum, is the range of summation deterministic or random? In the third sum, are the summands deterministic or random?

We will answer the above questions one by one.

Most importantly, in the third sum, we have extended the range of summation from the set of *existing* edges E to the set of *all possible undirected pairs* $\{i, j\} \subset [n]$. (The notation $\{i, j\} \subset [n]$ means that $\{i, j\}$ is a subset of set $[n] = \{1, 2, ..., n\}$. Written under the sum, it means that we sum over all such subsets of two vertices.) Extending the range of summation effectively means adding more summands. In our case, these extra summands are the indicators of edges outside of E.

Why the sum does not change? Because all edges outside of E do not exist, so their corresponding indicators equal zero. In short, we make the step from the second sum to the third sum by adding a bunch of zeros. This changes the number of summands but does not change the sum.

In the third sum, the range of summation is *deterministic* because we sum over all possible pairs of vertices.

In the third sum, the summands are *random* because each indicator I_{ij} can be 0 or 1 depending on the random event of whether edge $\{i, j\}$ exists or not.

Finally, in the fourth sum we change summation over edges to the equivalent summation over vertices. We make this step because summation over pairs of vertices is inconvenient for computation.

Make the next step yourself: Assume that you have to write a computer program that outputs the third sum. How will you do this? What

about the fourth sum?

If you think about the question above, you will realize that it is not easy to explain to a computer how to enumerate all vertex pairs (unless there is already a command for it). Eventually, you will have to translate this into two loops over vertices, and this is exactly what the fourth sum does. The double summation corresponds to the two loops. The external summation runs over all possible i, the first element of $\{i, j\}$. Importantly, the summation over j runs from $i+1$ and not from 1.

Make the next step yourself: Why does the summation over j run from $i+1$?

The summation over j runs from $i+1$ because in an undirected graph edge $\{i, j\}$ equals to edge $\{j, i\}$, and therefore $I_{ij} = I_{ji}$. We want to count each edge only once. For example, if we have added I_{13} then we do not need to add I_{31} anymore. When we sum over j from $i+1$, we make sure that $j > i$ and therefore we count each undirected edge exactly once.

Now we are ready to compute $\mathbb{E}(|E|)$, using the last expression (the fourth sum) in (2.1). For this, we use the very powerful result from probability theory – the linearity of expectations. The linearity of expectations says that the expectation of a sum *always* equals to the sum of expectations. This holds true for *any* sum of random variables, even if the random variables are dependent. This is a very convenient property because we have written $|E|$ as a sum of *indicators*, and expectation of an indicator is just the probability of 1.

Altogether, here is how we obtain the expected number of edges:

$$\mathbb{E}(|E|) \overset{(2.1)}{=} \mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=i+1}^{n} I_{ij}\right) \overset{\text{linearity of expectations}}{=} \sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathbb{E}(I_{ij}) = \sum_{i=1}^{n}\sum_{j=i+1}^{n} p_{ij}. \tag{2.3}$$

In words, the average number of edges is the sum of probabilities of all edges. This sounds logical because indeed, when edges become more likely, the average number of edges increases. Yet, this result is very elegant! Recall that the rules of how edges appear can be quite complicated. The edges may be dependent on many factors, and on each other. Nevertheless, the average number of edges is simply the sum of probabilities, and this is true for *any* undirected simple random graph!

One could guess the result in (2.3) intuitively. We chose to present a detailed formal derivation because this way we can demonstrate how to use indicators. Later we will use this technique in exactly the same way in more complicated and less intuitive cases.

2.2.3 Indicators not only of the edges

During the course, we will use indicators not only for the edges. We will use indicators I of other objects, such as wedges and triangles. We will also use indicators of *events*, for example, an 'event' could be that the number of edges is not greater than, say, $100\times$ the number of vertices. It is important to realize that an indicator is *always* a Bernoulli random variable that has a binary outcome 0 or 1, and its expectation *always* equals to the probability of 1.

2.3 The degree of a vertex: definition

The *degree* of vertex $i \in [n]$ is the number of vertices connected to i by an edge.

How to write this down as a formula?

Make the next step yourself. Can you write down the mathematical expression for the set of all vertices connected to i by an edge?

We formally write the set of all vertices connected to $i \in [n]$ by an edge as

$$\{j \in [n] : \{i, j\} \in E\}.$$

In words, this set includes vertex $j \in [n]$ if and only if $\{i, j\} \in E$.

The degree of i is the size of this set. This is exactly a mathematical definition of the degree:

Definition 2. *Consider graph $G = ([n], E)$. The degree of vertex $i \in [n]$, denoted by d_i, is the number of vertices connected to i by an edge. Formally:*

$$d_i = |\{j \in [n] : \{i, j\} \in E\}| \text{ for any } i \in [n]. \tag{2.4}$$

Self-test. Write down the degrees of all vertices in the graph in Figure 2.2.

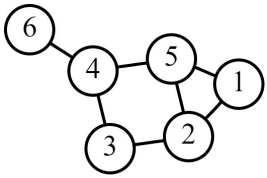


Figure 2.2: An undirected graph $G = ([6], E)$.

Answer: $d_1 = 2$, $d_2 = 3$, $d_3 = 2$, $d_4 = 3$, $d_5 = 3$, $d_6 = 1$.

2.4 The degree of a vertex as a sum of indicators

We will now proceed to express the degree as a sum of indicators.

Notice the similarity between $|E|$ and d_i. Both are the number of edges, only $|E|$ is the total number of edges in the graph, while d_i is the number of edges attached to vertex i. This means that we can express d_i using indicators of edges by following exactly the same steps as in Section 2.2.2 (even slightly easier because we do not need the double summation). When we do this, we get:

$$d_i = \sum_{j=1}^{n} I_{ij}, \tag{2.5}$$

and

$$\mathbb{E}(d_i) = \sum_{j=1}^{n} p_{ij}. \tag{2.6}$$

These formulas hold for any simple undirected random graph.

Self-test: Derive (2.5) and (2.6) by repeating the steps that we used to obtain (2.1) and (2.3).

2.5 The sum of all degrees

In this course we will often talk about the *average degree* of $G = ([n], E)$. The average degree of a graph with n vertices is given by the natural formula

$$\mu_n = \frac{d_1 + d_2 + \cdots + d_n}{n}. \tag{2.7}$$

In the numerator is the sum of all degrees, or the total degree. In this section we will derive a very basic relation between the sum of all degrees and the number of edges.

Make the next step yourself: As before, assume that G is a simple undirected graph. Complete the next formula by writing on the right-hand side a (very simple) expression that depends only on $|E|$:

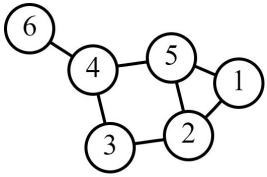
$$\sum_{i=1}^{n} d_i = \dots.$$

Answer:

$$\sum_{i=1}^{n} d_i = 2|E|. \tag{2.8}$$

We will now obtain (2.8), first using an intuitive logical argument, and then algebraically using indicators.

For the intuitive argument, let us look again at the example in Figure 2.2:



Imagine that we go through all vertices, one by one, adding their degree. We start with vertex 1 and add degree $d_1 = 2$. This is equivalent to saying that we have added the two edges of vertex 1: edge $\{1, 2\}$ and edge $\{1, 5\}$. So, in fact, we simply count edges. Next, we move to vertex 2, and add 3 to the sum, corresponding to the 3 edges of vertex 2. By doing so, we add edge $\{1, 2\}$ again. If we proceed this way, each edge $\{i, j\} \in E$ will be counted exactly twice: as part of d_i and as part of d_j. Therefore, the sum of all degrees simply equals twice the number of edges, which is exactly what is written in (2.8).

The intuitive argument in the previous paragraph is quite easy, but we need to be very careful that we do not make a logical error, and it will become much more complicated when we want to derive more intricate formulas. Indicators allow to obtain the same result algebraically, using only formulas. Such derivation is easier to check for errors, and it generalizes to many other cases.

Make the next step yourself: Can you derive (2.8) using indicators?

Hint: In $\sum_{i=1}^{n} d_i$, replace d_i by the right-hand side of (2.5).

The calculations are very easy:

$$\sum_{i=1}^{n} d_i \overset{(2.5)}{=} \sum_{i=1}^{n} \sum_{j=1}^{n} I_{ij} \overset{\text{compare to (2.2)}}{=} 2|E|.$$

Factor 2 in the last expression appears because the summation over j is from 1 to n and not from $i + 1$ to n, as it was in the fourth sum in (2.2).

Self-test. Can you explain why $\sum_{i=1}^{n} \sum_{j=1}^{n} I_{ij} = 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} I_{ij}$?

If you have difficulty with this self-test, please check this on a small example, and/or go back to the explanation of the *fourth* sum in Section 2.2.2.

2.6 Random variable D_n

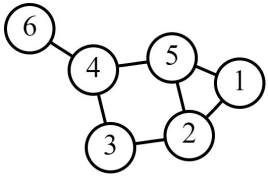
Degrees of vertices, and variation thereof from one vertex to another, and most basic characteristics of real-life networks. In this course we will often talk about *degree distribution*. In this section we will define what exactly we mean by that.

Let $G = ([n], E)$ be a simple undirected graph, and suppose we pick one vertex at random out of the n vertices of G. Formally, let U be a discrete random variable that takes values in $[n]$ with equal probability $1/n$. Then U is the number of a randomly chosen vertex. This vertex U will have some degree, d_U. Regardless whether G is deterministic or random graph, the degree of a random vertex U is a random variable, because of the randomness of the vertex number, U.

Notation d_U is not very convenient, for example, because it does not include n. In this course, dependence on n is important in this course because we often want to take $n \to \infty$. Therefore, we introduce a different notation: we will denote the degree of a randomly chosen vertex by D_n. We emphasize that $D_n \equiv d_U$, and we will use d_U sometimes in derivations.

Definition 3. *Random variable D_n is the degree of a vertex chosen uniformly at random from $[n]$.*

Self-test: Write down the degree distribution of D_6 in our earlier example:



Answer:

D_6	1	2	3
probability	1/6	2/6	3/6

If you have difficulty with this self-test, please revise the topic 'Probability distributions of a discrete random variable'

The probability distribution that you have just produced in the self-test, is exactly what we call the *degree distribution* of the graph. This was a small example, now we will give the general definition:

Definition 4. *The degree distribution of graph $G = ([n], E)$ is the probability distribution of random variable D_n.*

This definition holds for both deterministic and random graph G, and both situations are important for us: the *data* of real-life networks are deterministic because we already know which edges exist, but the *models* for real-life networks

are random graphs because these networks emerged as a result of some random process.

2.7 The average degree

If D_n is a random variable, then what is its expectation?

Make the next step yourself. Write down $\mathbb{E}(D_n)$.

Since D_n is the degree of a randomly chosen vertex, then, with probability $1/n$, vertex i will be chosen, resulting in $D_n = d_i$. When we substitute this into the definition of $\mathbb{E}(D_n)$, we get:

$$\mathbb{E}(D_n) = \mathbb{E}(d_U) \overset{\text{condition on } U}{=} \sum_{i=1}^{n} P(U=i)\,\mathbb{E}(d_U|U=i) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(d_i) = \mathbb{E}(\mu_n). \tag{2.9}$$

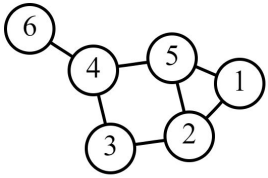
Self-test. Recall the definition of μ_n in (2.7). Can you explain why in the derivation above the last equality holds?

If you have difficulty with this self-test please revise the topic 'Linearity of expectations'.

The difference between $\mathbb{E}(D_n)$ and the average degree μ_n can be confusing, so we want to make this clear in the bullets below:

- We will use notation $\mu_n = \frac{1}{n}\sum_{i=1}^{n} d_i$ for the *empirical average degree* as given in (2.7). We emphasize that if degrees are random, then μ_n is a random variable, just as $|E|$.
- Random variable D_n is a degree of a specific, randomly chosen, vertex. Factor $1/n$ appears in $\mathbb{E}(D_n)$ due to probability $1/n$ of each vertex, and in μ_n due to averaging over n vertices.
- It always holds that $\mathbb{E}(D_n) = \mathbb{E}(\mu_n)$.
- If degrees are deterministic, then it holds that $\mathbb{E}(D_n) = \mu_n$. If degrees are random, then μ_n is a random variable while $\mathbb{E}(D_n)$ is still a fixed real number, therefore, being mathematical objects of a different type, they cannot be equal.

Self-test. Compute $\mathbb{E}(D_n)$, in our earlier example:



Is $\mathbb{E}(D_n)$ the same as μ_n in this example?

Answer: 2.667. Here $\mathbb{E}(D_n)$ and μ_n are the same because the degreres are deterministic.

If you have difficulty with this self-test, please revise the topic 'Expectation of a discrete random variable'

2.8 Apply, revise, repeat

This completes our story about the mathematical formalization of a network as a graph. The notions and techniques presented here are key to the course. In what follows we will often apply these notions and techniques, assuming that the students can explain and use them.

It is normal if later in the course the students have some struggles with this material in the context of specific random graph models, for example, sums of indicators or the source of randomness in D_n. We want to emphasize that this is a natural learning process, and we strongly encourage you to revise the corresponding sections of this chapter in light of the model/problem at hand.

3 Modeling sparse networks with the Erdős-Rényi random graph

Lezing 1, Nelly Litvak

3.1 Outline of this chapter

The title of this chapter contains two terms: *sparse networks* and *Erdős-Rényi random graph*. We will start with the latter, and this will be our very first random graph model in this course. We will define the model in Section 3.2 and derive the formulas for number of edges and degrees in Section 3.3.

Then we continue with formal mathematical definition of sparse networks in Section 3.4.

Our next question is: is the Erdős-Rényi random graph sparse? Turns out, we can make it sparse using a powerful technique called *parametrization*. We will explain this in Section 3.5.

Finally, in Section 3.6, we will state the convergence of the degree distribution to the Poisson distribution in a sparse sequence of E-R random graphs.

3.2 Definition of the Erdős-Rényi random graph model

How can we model a network as a random graph? Let us think about the simplest possible way. It is time to come back to the self-test in Chapter 1:

Self-test. Assume you want to construct a random graph of n vertices. What is the easiest way to place edges at random? Can you write a formal mathematical description of this random graph model?

Let us now define the Erdős-Rényi random graph model.

Definition 5. *The Erdős-Rényi (E-R) random graph is a simple undirected random graph $G = ([n], E)$, where for each $i, j \in [n]$, $i \neq j$, edge $\{i, j\}$ exists with the same probability p, independently of anything else. We denote this random graph by $ER_n(p)$.*

Self-test. Compare the definition of $ER_n(p)$ to your own ideas of the simplest possible model. Was it the same?

In general, in a random graph, each edge $\{i, j\}$ is placed with probability p_{ij}, so it is very natural to simplify the model by letting all these probabilities be the same. Formally, $p_{ij} = p$ for all $i, j \in [n]$. We show this schematically in Figure 3.1, where solid lines are existing connections and dashed lines are possible but not realized connections.

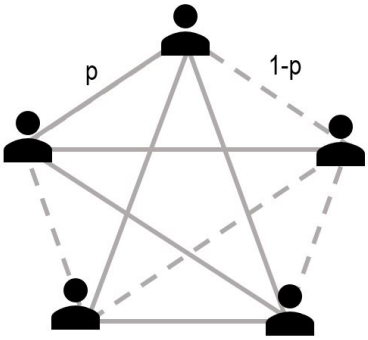


Figure 3.1: Social network is modeled as an Erdős-Rényi random graph: each friendship exists with probability p.

Oftentimes the description offered by students ends there, but this is not enough because edges can be *dependent* on each other. Think about the triangles in Figure 1.1d. If Anna has an edge to Boris and Cecile, then the edge between Boris and Cecile is very likely to exist. The E-R model simplifies things further by assuming that all edges are *independent*. If a social network is created by the E-R model, then the edge between Boris and Cecile will not depend on whether they both know Anna, it will simply randomly exist with probability p. The independence of edges is a crucial property of the E-R model, as we will see already in Section 3.3.

Self-test. Go back to your own ideas of the simplest possible model. Did you explicitly include the independence of edges?

This is of course not how social networks form in reality, and we will come back to it. Nevertheless, the E-R model is very useful exactly due to its simplicity. As it often happens in science, deep understanding of the simplest possible mathematical model is an important step towards explaining real-world phenomena. In our case the phenomena in question will be sparseness of real-life networks (this chapter), and presence of motifs (Chapter 4).

It is useful to see what the E-R graph looks like for different n and different p, so please look at the *Network pages* website:

`https://www.networkpages.nl/CustomMedia/`

When you scroll down, you will find a simulator of the E-R random graph.

Self-test. Using the simulator, look at several realizations of the E-R random graph. (Ignore the λ for the moment, we will explain it in Section 3.5.)

First, alter n (Nr of vertices) for fixed p (Edge probability), and then alter p for fixed n.

What looks exactly as you expected? Explain what exactly you expected correctly, why you expected this, and what exactly in the figures confirms that that your idea was correct.

What looks differently than you expected? If your initial intuition was not correct, explain what you expected to be different, how you see that it was not correct, and where was the mistake in your original logic.

3.3 Probability distribution of the number of edges, and degrees in the E-R random graph

Pre-requisites. The following preliminary knowledge is required for this section:

- Binomial distribution (required).

In what follows we assume that you are able to *define, recognize* and *apply* the binomial distribution. If you feel that this is not the case, please revise your knowledge before proceeding further.

Consider a random graph $G = ([n], E)$ and assume that this graph is $ER_n(p)$.

What can we say about the number of edges $|E|$? Here the indicators are really useful. Recall from (2.2) that $|E|$ is a sum of indicators:

$$|E| = \sum_{i=1}^{n} \sum_{j=i+1}^{n} I_{ij}.$$

In $ER_n(p)$, edges are independent and exist with the same probability. In other words, I_{ij} are independent and identically distributed Bernoulli random variables. This is very informative for us because we know that the *sum of independent and identically distributed Bernoulli random variables has a Binomial distribution.* The 'success probability' in this case is p, and the 'number of Bernoulli experiments' is the maximal possible number of edges $\frac{n(n-1)}{2}$.

Self-test. Can you explain why the maximum of $|E|$ is $\frac{n(n-1)}{2}$?

This question is important because when we use a sum of indicators, we will often need to compute how many indicators are there in the sum. We ask you to pause here until you can formulate the answer in full sentences. If you are not confident at the moment, the hint below may help you to proceed.

Hint. The number of edges $|E|$ is maximal when each pair $\{i, j\}$ is in E. Now, it is useful to realize that each edge $\{i, j\}$ is simply a subset of two elements of $[n]$. So, the maximal possible $|E|$ is the same as the number of possible subsets of size 2 out of n elements. This number goes by the name *'n-choose-2'* and has notation

$$\binom{n}{2}.$$

If you are not sure why $\binom{n}{2} = \frac{n(n-1)}{2}$, please look it up on the Internet.

We use the symbol $A \sim F$ to state that random variable A has distribution F. Then we write:

$$|E| \sim Binomial\left(\frac{n(n-1)}{2}, p\right).$$

The expectation and variance of the Binomial distribution in our case become:

$$\mathbb{E}(|E|) = \frac{n(n-1)}{2}\, p, \quad \mathrm{Var}(|E|) = \frac{n(n-1)}{2}\, p(1-p).$$

What about the degree? Consider the degree of a fixed vertex $i \in [n]$.

Make the next step yourself. What is the probability distribution of d_i?

Hint. Look at formula (2.5).

Since the graph is simple, so edge $\{i, i\}$ does not exist, we rewrite (2.5) as

$$d_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} I_{ij}, \quad \text{for any } i \in [n].$$

We now see that d_i is the sum of $n-1$ independent identically distributed indicators, therefore

$$\begin{aligned} &d_i \sim Binomial(n-1, p), \\ &\mathbb{E}(d_i) = (n-1)p, \\ &\mathrm{Var}(d_i) = (n-1)p(1-p), \quad \text{for any } i \in [n]. \end{aligned}$$

Finally, D_n is distributed as d_i with probability $1/n$, and all d_i are identically distributed, so we have

$$D_n \sim Binomial(n-1, p).$$

3.4 Mathematical definition of a sparse network

What is a sparse network and how can we express this in a formula? We start with an informal definition we had before:

A network is sparse if the degrees of vertices are small compared to the network size.

Now we will replace the words by mathematical expressions, wherever we can, based on the previous material of this reader.

The *network size* is clearly n.

The next question is: which mathematical notion describes *'the degrees of vertices'*? There are different ways of doing this. The easiest choice is just the average degree μ_n.

Then we arrive at the following:

A network is sparse if μ_n is small compared to n.

It remains to define what it means that μ_n is *'small'* compared to n, and this *is* a somewhat tricky question.

Let us look at some examples.

- Suppose $n = 1000$ and $\mu_n = 2$. We will probably all agree that the network is sparse.
- Now suppose $n = 1000$ and $\mu_n = 500$. We will probably all agree that the network is quite dense, and definitely is not sparse.
- What about $n = 1000$ and $\mu_n = 60$? This is quite uncertain because 60 is not such a small number, but it is also quite small compared to 1000.

This example illustrates that when n is *fixed*, we cannot really define what 'sparse' means. Any decision whether the network is sparse when $n = 1000$ and $\mu_n = 60$, will be quite arbitrary. This is why, as we often do in mathematics, we

dismiss the notion of sparseness for *fixed* n altogether. Instead, sparseness is an *asymptotic notion.*

(The words 'asymptotic notion' may sound scary but you are definitely well familiar with common asymptotic notions in mathematics such as $O(1)$ or $o(x)$ as $x \to \infty$. For example, $f(x) = O(1)$ if $f(x)$ is bounded by a constant for all $x \in \mathbb{R}$. When you look at Definition 6 below, you will immediately see the similarity.)

'Asymptotic notion' means that 'sparse' networks are defined in terms of a limit of μ_n when $n \to \infty$. This is why we do not talk about a 'sparse graph' but about a 'sparse *sequence* of graphs'. It is a sequence because we consider different $n = 1, 2, 3, \ldots$, and assume there is a graph $G_n = ([n], E_n)$ for each n. Once we have a graph sequence, we can view μ_n as a *function* of n, and define sparseness in terms of the limit of this function as $n \to \infty$. More specifically, we will say that:

A graph sequence is sparse if μ_n remains bounded as $n \to \infty$.

We are almost there, except for the fact that in the random graph model the degrees can be random, and then μ_n is random as well. The limit of a random variable is a subtle notion, and there are different ways to deal with this. At this point, to keep things simple, we define a sparse sequence of random graphs using $\mathbb{E}(\mu_n) = \mathbb{E}(D_n)$ instead of μ_n.

The formal definition of a sparse graph sequence is given below.

Definition 6. *A graph sequence $G_n = ([n], E_n)$ is sparse when there exists constant $M > 0$ such that*

$$\mu_n \leq M, \quad \text{for all } n=1,2,\ldots \quad \text{if } d_1, d_2, \ldots, d_n \text{ are deterministic,}$$

or

$$\mathbb{E}(\mu_n) = \mathbb{E}(D_n) \leq M, \text{ if } d_1, d_2, \ldots, d_n \text{ are random.}$$

Remark 7. *Many sparse random graph sequences satisfy a stronger condition, namely that μ_n converges to a limit. If the degrees are random, so μ_n is random as well, it is usually assumed that μ_n converges to its limit 'in probability'. In that spirit, a neater definition of sparseness is as follows: a network is sparse when $\mu = O_{\mathbb{P}}(1)$, that is, for any $\varepsilon, \delta > 0$, holds: $\lim_{n\to\infty} \mathbb{P}(\mu_n/n^{\delta} > \varepsilon) = 0$. This is 'neater' than using $\mathbb{E}(\mu_n)$, as we did in Definition 6, because sparseness is defined through the empirically observed degrees rather than through their theoretical expectation.*

We emphasize that boundedness or convergence of $\mathbb{E}(\mu_n)$ is technically not equivalent to the boundedness or convergence of μ_n when degrees are random, even though most of the time both will hold together in this course.

At this point we chose to work with $\mathbb{E}(\mu_n)$ to define sparseness because it essentially captures the same phenomenon without diving into the subtleties of convergence of random variables.

In the light of Definition 6, let us come back to the question whether the network is sparse when $n = 1000$ and $\mu_n = 60$. Strictly according to the definition, we cannot answer this question for the single $n = 1000$, we need to know how μ_n changes with n in the entire *sequence* G_n. If, say, $\mu_n \leq 70$ for all $n = 1, 2, \ldots$, then this *graph sequence* is sparse. If $\mu_n = 0.06n$, then the *graph sequence* is not sparse.

Self-test. Can we say that a network is sparse when $n = 1000$ and $\mu_n = 2$? Try to answer yourself, then look at the answer provided in the next paragraph.

Formally, the question in the last self-test is ill-posed. We cannot say about a single network instance whether it is sparse or not, we must look at the entire sequence. For example, if $\mu_n = 0.002n$ then $\mu_n \to \infty$ as $n \to \infty$, and the graph sequence is not sparse.

That said, in practice, when we see a real-life network with 1000 nodes and on average 2 connections per node, it is reasonable to view it as an instance of a sparse sequence, and usually we will choose a sparse sequence of random graphs to model networks like that. We need asymptotic notions for the correct and useful mathematics, but inspiration for these notions still lies in what we observe in real-life networks of fixed size.

3.5 Parametrization for creating a sparse sequence of Erdős-Rényi random graphs

Let us now come back to $ER_n(p)$, $n = 1, 2, \ldots$. Is this sequence sparse? According to Definition 6 we have to look at

$$\mathbb{E}(\mu_n) = \mathbb{E}(D_n) = (n-1)p$$

as a function of n.

At the first sight it looks like μ_n is linear in n. However, it does not need to be! We can change this by using a powerful technique called *parametrization*. The idea is to not view p as a given constant in $(0, 1)$ but make it a function of n, so $p = p(n)$. Why is it useful? Because, for instance, different $p(n)$ will result in different μ_n as a function of n.

Make the next step yourself. Do you see how to choose $p = p(n)$ so that the sequence $ER_n(p(n))$ becomes sparse?

For the sparse sequence we need that μ_n is bounded. We can achieve this by setting

$$p := p(n) = \frac{\lambda}{n}, \quad \lambda > 0.$$

This is exactly the same λ that you saw in the simulator in Section 3.2.

Make the next step yourself. Write down the distribution, the expectation and the variance of $|E|$, d_i, D_n, $i \in [n]$ in $ER_n(\lambda/n)$.

For $|E|$ in $ER_n(\lambda/n)$ we obtain:

$$|E_n| \sim Binomial\left(\frac{n(n-1)}{2}, \frac{\lambda}{n}\right), \quad \mathbb{E}(|E|) = \frac{(n-1)\lambda}{2},$$

so the mean number of edges grows linearly with n.

For d_i in $ER_n(\lambda/n)$ we obtain:

$$d_i \sim Binomial\left(n-1, \frac{\lambda}{n}\right), \quad \mathbb{E}(d_i) = \frac{n-1}{n}\lambda = \lambda\left(1 - \frac{1}{n}\right),$$

and

$$\lim_{n\to\infty} \mathbb{E}(d_i) = \lambda.$$

Let us now verify that Definition 6 of a sparse network holds. The degrees d_i are random, so μ_n is random as well, and we take the expectation:

$$\mathbb{E}(\mu_n) = \mathbb{E}(D_n) = \frac{(n-1)\lambda}{n} = \lambda\left(1 - \frac{1}{n}\right) < \lambda,$$

and hence now the graph sequence is sparse. Moreover,

$$\lim_{n\to\infty} \mathbb{E}(\mu_n) = \lambda,$$

so we have convergence of $\mathbb{E}(\mu_n)$ to λ, which is stronger than boundedness.

Having $p = \frac{\lambda}{n}$ means that the probability that vertex i connects to vertex j decreases when the network grows, and the average number of connections of vertices remains approximately constant. This is in fact quite a natural modeling assumption. For example, in a social network an individual typically will not get many more friends when the network becomes larger. Rather, the probability of connection of one individual to another will decrease because there will be too many total strangers in terms of location, interests, etc.

3.6 Convergence of the degrees to the Poisson distribution in a sparse sequence of E-R random graphs

Since $d_i \sim Binomial\left(n-1, \frac{\lambda}{n}\right)$, and

$$\lim_{n\to\infty} \mathbb{E}(d_i) = \lim_{n\to\infty} (n-1)\frac{\lambda}{n} = \lambda,$$

the distribution of d_i converges to the $Poisson(\lambda)$ distribution:

$$\lim_{n\to\infty} \mathbb{P}(d_i = x) = \frac{\lambda^k}{k!}\, e^{-\lambda}, \quad x = 0, 1, \ldots. \tag{3.1}$$

Self-test. Can you explain what (3.1) says and why it holds? If not, it is useful to look up convergence of the binomial distribution to the $Poisson(\lambda)$ distribution. You can also derive (3.1) yourself, it is not at all difficult. You will need to use the formula for the binomial distribution:

$$\mathbb{P}(d_i = x) = \binom{n-1}{x} p^x (1-p)^{n-1-x}, \quad x = 0, 1, \ldots, n-1,$$

and the standard limit

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{\lambda}.$$

Convergence to the Poisson distribution is a very good news because this distribution is very well understood and has many nice properties. We will often see convergence to the Poisson distribution in the course when a sum of indicators (such as the degree) has a finite expectation in the limit when $n \to \infty$.

4 Counting motifs in random graphs
Lezing 2, Nelly Litvak

In real-life networks we often observe so-called *motifs* such as *triangles* in Figure 1.1d, but also *wedges*, *cycles*, and *cliques*, see Figure 4.1. (A clique is a set of vertices that are all connected to each other, like a group of close friends in a school class.) Notice that a triangle is a cycle, but also a clique of three.

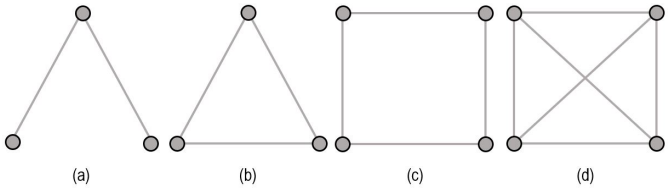


Figure 4.1: (a) wedge; (b) triangle; (c) 4-cycle; (d) 4-clique: all four vertices are connected to each other by an edge.

Motifs like this are defined mathematically as *connected induced subgraphs*, and in network science they are also often called *graphlets*. The word *'induced'* means that if an edge is not in a graphlet, then this edge is also not in E. The word *'connected'* means that in a graphlet, there is always a path of edges between any pair of vertices.

Such graphlets are very informative. For example, as we discussed before, social networks have many triangles. On the other hand, planned networks such as railways will usually not have a triangle between three neighboring stations! Rather, there will be longer 'cycles' of stations that offer detours, for example, in case of disruption.

In this chapter we will learn how to count motifs in random graphs. The Erdős-Rényi (E-R) random graph is an ideal model to start with due to its simplicity. In particular, we will discover how many triangles and cycles one may expect to see in a sparse E-R random graph.

For a motif of any form, the analysis of its count follows exactly the same technique as counting edges, using indicators. In this course we will show only the first basic steps of this analysis, and we expect all students to understand them completely. In Section 4.1 we will demonstrate how to compute the mean number of

wedges. In Section 4.2 we will extend this analysis to any motif. In Section 4.3 we go back to the sparse graph and discuss the number of motifs when $n \to \infty$.

4.1 The number of wedges

We again work with a simple undirected graph $G = ([n], E)$. We will derive the mean number of wedges using indicators.

The first question is, what *is* the number of wedges in formal mathematical terms?

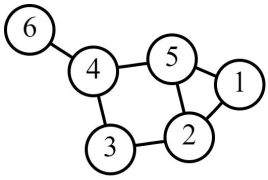
Similarly to $|E|$ and d_i, the number of wedges is the size of a set that contains all wedges. We denote the set of all wedges by $\mathcal{W}$:

$$\mathcal{W} = \{\text{all wedges in } G\}.$$

If we want to mathematically define the set of wedges, we have to start with defining what a wedge is in mathematical terms.

4.1.1 Mathematical definition of a wedge

Any wedge consists of three vertices and two edges. Consider the small graph in our earlier example:



Here, for instance, $6-4-5$ is a wedge. How can we write this down?

Make the next step yourself. Denote the wedge $6-4-5$ in the example by w_{645} and write its formal definition.

A wedge $6-4-5$ is an ordered set of three vertices $w_{645} = (6, 4, 5)$. We use the round brackets to denote the ordered set. The order is important because this is the only way to show the special position of vertex 4. When we change the position of this vertex, we get a different wedge, for example, wedge $(4, 5, 6)$ is not the same as $(6, 4, 5)$. Indeed, as you can see, wedge $(4, 5, 6)$ does not exist in this graph.

Of course, not every ordered set of vertices is a wedge. We have wedge w_{546} only when $\{4, 5\}, \{4, 6\} \in E$ and $\{5, 6\} \notin E$.

We arrive at the following formal definition of a wedge:

Definition 8. *A wedge is an ordered set of three vertices, such that the second vertex is connected by an edge to the first and the third vertex, but the first and the third vertex are not connected. Formally, for all* $i, j, k \in [n]$*,*

$$w_{ijk} = (i, j, k) \text{ such that } \{i, j\} \in E,\ \{i, j\} \in E,\ \{i, k\} \notin E.$$

4.1.2 Indicator of a wedge

Denote by I^w_{ijk} the indicator of a wedge. Indicator I^w_{ijk} is 1 if (i, j, k) is a wedge, and 0 otherwise. Whether (i, j, k) is a wedge or not depends only on the edges between these vertices, so I^w_{ijk} is completely defined by the edges $\{i, j\}$, $\{j, k\}$, $\{i, k\}$. In fact, we can express I^w_{ijk} through the indicators of the edges.

Make the next step yourself. Write down I^w_{ijk} through indicators of edges I_{ij}, I_{jk}, I_{ik}. The answer is given below. Use the hint.

Hint: Use a product.

We have:

$$I^w_{ijk} = I_{ij} I_{jk} (1 - I_{ik}), \quad i, j, k \in [n]. \tag{4.1}$$

Self-test. Verify that the formula (4.1) is correct.

Hint: First verity (4.1) when (i, j, k) is a wedge. Then see what changes when (i, j, k) is not a wedge.

The answer is found by a simple enumeration of all combinations of I_{ij}, I_{jk}, I_{ik}. Even if you experience difficulty with this question, do not give up, you can do it!

4.1.3 Formal definition of the set of wedges

We will now proceed writing down the set of all wedges $\mathcal{W}$. Since a wedge is an ordered set of three vertices, set $\mathcal{W}$ consists of such ordered triples (i, j, k), but it must include only those triples that form a wedge, and it must include each wedge exactly once.

Self-test. Will any ordered set (i, j, k) correspond to a different wedge? If not, which wedges will be equal? Write down mathematically your statement about equality or inequality of wedges. We give the answer in the next paragraph.

Clearly, swapping the positions of the first and the third veetex in a wedge does not change anything, it is still the same wedge. In our small example,

$$w_{645} = w_{546}.$$

To make sure that we list each wedge exactly once we will impose that between vertices 5 and 6 the vertex with a smaller number goes first, so in $\mathcal{W}$ we will include w_{546} but not w_{645}.

We are now ready to write down the mathematical definition of $\mathcal{W}$.

Make the next step yourself. Write down the formal definition of the set of all wedges $\mathcal{W}$. If you cannot do this, then look at the answer in (4.2) below, and complete the self-test after (4.2), in a written form, without skipping any bullet.

Based on the discussion above, we have:

$$\mathcal{W} = \{(i,j,k) : i,j,k \in [n], i < k, \{i,j\} \in E, \{j,k\} \in E, \{i,k\} \notin E\}. \qquad (4.2)$$

Self-test. Write down the answers to the following questions:

1. Why do we write the external (red) curly brackets?
2. Why is (i,j,k) in round brackets?
3. Why do we write $i,j,k \in [n]$?
4. Why do we write $i < k$?
5. Why do we write $\{i,j\} \in E, \{j,k\} \in E, \{i,k\} \notin E$?

Answer:

1. Because $\mathcal{W}$ is a set.
2. Because $\mathcal{W}$ is a set of wedges, and each wedge is an ordered set of three vertices. The round brackets are used to denote an ordered set.
3. Because any triplet of vertices can form a wedge.
4. Because $w_{ijk} = w_{kji}$. By imposing $i < k$, we count each wedge exactly once.
5. Because (i,j,k) is a wedge if and only if the graph has edges $\{i,j\}$ and $\{j,k\}$ and does not have edge $\{i,k\}$.

4.1.4 The number of wedges

We are interested in the number of wedges, which is the size of the set of all wedges, $|\mathcal{W}|$. We will write this using indicators.

Denote by I^w_{ijk} the indicator of w_{ijk}. Then we have

$$|\mathcal{W}| = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=i+1}^{n} I^w_{ijk}. \tag{4.3}$$

We hope you appreciate that writing the range of the three summations is easy because we did a thorough preliminary work of defining $\mathcal{W}$. Let us take a closer look. The range of summation over vertices is determined by the statement $i, j, k \in [n], i < k$ in (4.2), while the indicator is 1 when (i, j, k) is a wedge.

From (4.3) we obtain the mean number of wedges, using the linearity of expectations:

$$\mathbb{E}|\mathcal{W}| = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=i+1}^{n} \mathbb{E}(I^w_{ijk}). \tag{4.4}$$

We can go one step further, using (4.1):

$$\mathbb{E}|\mathcal{W}| = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=i+1}^{n} \mathbb{E}(I^w_{ijk}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=i+1}^{n} \mathbb{E}\left(I_{ij}I_{jk}(1-I_{ik})\right). \tag{4.5}$$

Here $\mathbb{E}\left(I_{ij}I_{jk}(1-I_{ik})\right)$ is the probability that (i, j, k) is a wedge, so there is a very strong similarity to the formula for the mean number of edges (2.3). Unfortunately, in general, in (4.5), we have to stop here, and we cannot right away rewrite $\mathbb{E}\left(I_{ij}I_{jk}(1-I_{ik})\right)$ through the edge probabilities.

Self-test. Why can't we replace $\mathbb{E}\left(I_{ij}I_{jk}(1-I_{ik})\right)$ by $p_{ij}p_{jk}(1-p_{ik})$? The answer is in the next paragraph.

The reason is that, in general, $\mathbb{E}\left(I_{ij}I_{jk}(1-I_{ik})\right)$ is *not* equal to $\mathbb{E}(I_{ij})\mathbb{E}(I_{jk})\mathbb{E}(1-I_{ik})$. The exception is the case when the edges are *independent*, then the equality does hold. This is why it is very easy to find $\mathbb{E}(|\mathcal{W}|)$ in the E-R random graph, as we shall see in the next section.

4.1.5 The mean number of wedges in the E-R random graph

In the E-R random graph, the edges are *independent*. This pays off greatly when we derive the formula for the mean number of motifs. In particular, for the wedges we get:

$$\mathbb{E}(|\mathcal{W}|) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=i+1}^{n} \mathbb{E}\left(I_{ij}I_{jk}(1-I_{ik})\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=i+1}^{n} p^2(1-p). \tag{4.6}$$

In the last summation, all terms are the same, so the final answer is $p^2(1-p)$, multiplied by the number of terms.

Self-test. How many terms are there in the last summation in (4.6)? The answer is in the next paragraph.

There are $n(n-1)(n-2)$ ordered triples of vertices, and we need to divide this number by 2 because $i < k$. The number $\frac{n(n-1)(n-2)}{2}$ is the number of possible wedges in the graph. So, we get

$$E(|\mathcal{W}|) = \frac{n(n-1)(n-2)}{2}\, p^2(1-p).$$

This formula is very simple, and it follows directly from the fact that $|\mathcal{W}|$ is a sum of indicators. This generalizes to the number of graphlets of any form in the E-R random graph.

4.2 The mean number of graphlets in the E-R random graph

Similarly as we did with the wedges, we can write down the number of graphlets of any form as a sum of indicators. In the most general terms, let g be a graphlet of r vertices and m edges. Denote by $\mathcal{G}$ the set of such graphlets, and let $I^g_{i_1 i_2 \dots i_r}$ be the indicator of graphlet $(i_1, i_2, \dots, i_r)$. Then

$$|\mathcal{G}| = \sum_{\substack{\text{all } (i_1,i_2,\dots,i_r) \in [n]^r \\ \text{that can form distinct graphlets}}} I^g_{i_1 i_2 \dots i_r} \tag{4.7}$$

and

$$\begin{aligned} \mathbb{E}(|\mathcal{G}|) &= \sum_{\substack{\text{all } (i_1,i_2,\dots,i_r) \in [n]^r \\ \text{that can form distinct graphlets}}} \mathbb{E}(I^g_{i_1 i_2 \dots i_r}) \\ &= \sum_{\substack{\text{all } (i_1,i_2,\dots,i_r) \in [n]^r \\ \text{that can form distinct graphlets}}} \mathbb{P}(I^g_{i_1 i_2 \dots i_r} = 1). \end{aligned} \tag{4.8}$$

Make the next step yourself. Write down the expression for $\mathbb{P}(I^g_{i_1 i_2 \dots i_r} = 1)$ in the E-R random graph when r has m edges.

In the E-R random graph, $\mathbb{P}(I^g_{i_1 i_2 \dots i_r} = 1)$ is the product that has multiple p for each existing edge in g, and multiple $(1-p)$ for each edge not in g. The number of non-existing edges in g is $\frac{r(r-1)}{2} - m$. (Why?). Then (4.8) becomes:

$$\mathbb{E}(|\mathcal{G}|) = [\text{number of possible distinct graphlets}] \times p^m \, (1-p)^{\frac{v(v-1)}{2} - m}. \quad (4.9)$$

For example, denote by Δ the set of triangles in the E-R random graph. Then

$$\mathbb{E}(|\Delta|) = \frac{n(n-1)(n-2)}{6} p^3. \quad (4.10)$$

Self-test.

1. Explain (4.10).
2. Compute the mean number of induced squares in the E-R random graph, as in Figure 4.1(c).
3. Compute the mean number of 4-cliques in the E-R random graph, as in Figure 4.1(d).

Answer.

1. First we count the number of ordered triples (i, j, k) that can form *distinct* triangles. Since the order of vertices in a triangle plays no role, and there are 6 possible ways to order a triplet, we have that the total number of possible triangles, is $\frac{n(n-1)(n-2)}{6}$. Next, all three edges are present, so we must multiply $\frac{n(n-1)(n-2)}{6}$ by p^3.
2. $\frac{n(n-1)(n-2)(n-3)}{8} p^4 (1-p)^2$.
3. $\frac{n(n-1)(n-2)(n-3)}{4!} p^6$.

4.3 The number of graphlets when $n \to \infty$ in a sequence of sparse E-R random graphs

Recall that a sparse E-R random graph $ER_n(\lambda/n)$ models the phenomenon that the average degree remains bounded when the network grows. Many real-life networks are sparse, but they have graphlets as well, for example, social networks will have many triangles. In this section we will see what happens to the number of graphlets in $ER_n(\lambda/n)$ when $n \to \infty$.

It turns out that we can say a lot about the *distribution* of the number of graphlets as $n \to \infty$, by looking only at the *mean* number of graphlets. This is because the number of graphlets is *always* the sum of indicators, as we wrote explicitly in (4.7).

Moreover, in an E-R random graph, the probability of a graphlet does not depend on the vertices involved, all graphlets of the same form have the same probability. Hence, in the E-R random graph all indicators in (4.7) are *identically distributed*.

If indicators were also *independent*, as it was the case for the indicators of edges I_{ij}, then their sum would have a binomial distribution.

Make the next step yourself. Look back at the formula for the number of wedges (4.3). Does $|\mathcal{W}|$ follow a Binomial distribution?

Answer: No. The wedges are not independent. For example, wedges $(1,2,3)$ and $(1,2,4)$ share edge $\{1,2\}$, and so they are dependent through this edge. If wedge $(1,2,3)$ exists, it means that edge $\{1,2\}$ exists, and this makes wedge $(1,2,4)$ more likely. Formally, $\mathbb{P}(I^w_{124}=1) = p^2(1-p)$, but $\mathbb{P}(I^w_{124}=1 | I^w_{123}=1) = p(1-p)$.

(Do you have troubles explaining the last formula? Then draw the four vertices and make $(1,2,3)$ a wedge. Given that $(1,2,3)$ is already a wedge, what is now the probability that $(1,2,4)$ is a wedge as well?)

In (4.7), indicators are not independent, they influence each other through shared edges. Nevertheless, loosely speaking, they become *'almost'* independent as $n \to \infty$. For example, I^w_{123} in (4.3) is dependent only on the indicators that involve either $\{1,2\}$, or $\{2,3\}$, or $\{1,3\}$. There are only $O(n)$ of such indicators. But there are in total $O(n^3)$ indicators of wedges, so I^w_{123} is independent of most of the summands in (4.3).

Same reasoning applies to any finite graphlet in an E-R random graph. The idicators are not independent, but each indicator is independent of the great majority of the others. Intuitively, this is why, as $n \to \infty$, $|\mathcal{G}|$ has the same limiting properties as a binomial random variable.

In the remainder of this section we will present most basic limiting properties of $|\mathcal{G}|$. If you are comfortable with limiting properties of the Binomial distribution, then these results will sound very familiar to you. Proving these results however might be technical because of the dependencies between the graphlets. The book by Remco van der Hofstad[1] contains many useful proof techniques, e.g. Theorems 2.4, 2.5.

[1] Random graphs and complex networks, Cambridge University Press, 2016

4.3.1 The limit of $\mathbb{E}(|\mathcal{G}|)$

In $ER_n(\lambda/n)$, formula (4.9) becomes

$$\mathbb{E}(|\mathcal{G}|) = [\text{number of possible distinct graphlets}] \times \left(\frac{\lambda}{n}\right)^e \left(1-\frac{\lambda}{n}\right)^{\frac{r(r-1)}{2}-m}. \tag{4.11}$$

Here is what we can say about this formula:

- The first term, the *number of possible distinct graphlets*, is of the order n^r because choice of every vertex has the order of n options.
- The second therm, $\left(\frac{\lambda}{n}\right)^m$, is of order n^{-m}.
- For the third term, it holds that

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{\frac{r(r-1)}{2}-m} = 1,$$

  so this term will not affect the limit.

Altogether, the limit of $\mathbb{E}(|\mathcal{G}|)$ will be determined by the main term of order n^{r-m}. Then there are only three options:

1. $r < m$, and $\lim_{n\to\infty}\mathbb{E}(|\mathcal{G}|) = 0$,
2. $r = m$, and $\lim_{n\to\infty}\mathbb{E}(|\mathcal{G}|) = \mu_g > 0$,
3. $r > m$ is possible only if $r = m+1$ since a graphlet is connected. (Why is this statement true?) Then $\mathbb{E}(|\mathcal{G}|)$ is of order n.

In the next subsections we will consider these cases one by one.

4.3.2 Vanishing graphlets: $\lim_{n\to\infty}\mathbb{E}(|\mathcal{G}|) = 0$

As an example, consider the number of 4-cliques as in Figure 4.1(d). Denote the set of 4-cliques in a graph by $\mathcal{C}_4$. Then we have

$$\lim_{n\to\infty}\mathbb{E}(|\mathcal{C}_4|) = \lim_{n\to\infty}\frac{n(n-1)(n-2)(n-3)}{4!}\frac{\lambda^6}{n^6} = 0.$$

So, on average, in the limit, the number of 4-cliques in a sparse E-R random graph is zero.

This is on average, but when we generate an instance of such a random graph, we do not see an average, we see only one specific realization. Can we conclude that typically in such realization there will be no 4-clique? The answer is yes, but we need a little bit of work to prove it mathematically.

First of all, what do we want to prove? We are interested whether a random graph $ER_n(\lambda/n)$ contains a 4-clique.

Make the next step yourself. Express the event [*a random graph contains a 4-clique*] mathematically using random variable $|\mathcal{C}_4|$.

Saying *'a graph contains a 4-clique'* is equivalent to saying *'the number of 4-cliques is more than zero'*, or *there is at least one 4-clique*. Mathematically we can write is as

$$|\mathcal{C}_4| \geq 1.$$

If a graph typically will not contain 4-cliques then the probability of this event should be very small. So, we are interested in

$$\mathbb{P}(|\mathcal{C}_4| \geq 1),$$

and we want to show that this probability converges to zero as $n \to \infty$.

Since we know $\mathbb{E}(|\mathcal{C}_4|)$, the appropriate proof technique uses the *Markov inequality*. This is a very well-known inequality in the probability theory, given below.

Theorem 9 (Markov inequality)**.** *Let X be a non-negative random variable, $\mathbb{E}(X) < \infty$. Then for any $a > 0$ it holds that*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

In our case, we get

$$\mathbb{P}(|\mathcal{C}_4| \geq 1) \leq \frac{\mathbb{E}(|\mathcal{C}_4|)}{1} = \mathbb{E}(|\mathcal{C}_4|),$$

so

$$\lim_{n\to\infty} \mathbb{P}(|\mathcal{C}_4| \geq 1) \leq \lim_{n\to\infty} \mathbb{E}(|\mathcal{C}_4|) = 0.$$

The conclusion is that in a sparse E-R random graph, when n grows, we will almost never see a 4-clique.

In $ER_n(\lambda/n)$, exactly the same result will hold for any graphlet that contains more edges than vertices.

Self-test. Can you intuitively explain why the probability to see a 4-clique reduces as n grows?

> **Possible answer.** The number of possible cliques will grow with n, but the probability to get all 6 edges in the clique will reduce faster.

This is already not very good news in terms of modeling e.g. social networks. Clearly, in a real-life social network, when the network grows, the number of completely connected small groups should grow as well.

4.3.3 Poisson number of graphlets: $\lim_{n\to\infty} \mathbb{E}(|\mathcal{G}|) = \mu_g$

As an example, consider the number of triangles. In $ER_n(\lambda/n)$, we have

$$\lim_{n\to\infty} \mathbb{E}(|\Delta|) = \lim_{n\to\infty} \frac{n(n-1)(n-2)}{6} \frac{\lambda^3}{n^3} = \frac{\lambda^3}{6}.$$

So, in the limit, the mean number of triangles is a constant.

Self-test. Do you think that the number of triangles will be a constant in any realization of the random graph? The answer is given in the next paragraph.

The number of triangles itself is of course not a constant, it is a random variable, and it will be different in each realization. We know that its mean converges to a constant. But what about its probability distribution?

We know that the $Binomial\left(\frac{n(n-1)(n-2)}{6}, \frac{\lambda^3}{n^3}\right)$ distribution converges to the $Poisson\left(\frac{\lambda^3}{6}\right)$ distribution. We also know that the distribution of $|\Delta|$ is *not* Binomial because some triangles are dependent through their shared edges. Nevertheless, it turns out that this moderate dependency does not destroy the convergence to the Poisson distribution, only the proof of this convergence becomes (much) harder. At this point we will state the convergence result without proof.

Theorem 10. *In $ER_n(\lambda/n)$, as $n \to \infty$, the distribution of the number of triangles, $|\Delta|$, converges to the $Poisson\left(\frac{\lambda^3}{6}\right)$ distribution. Specifically,*

$$\lim_{n\to\infty} \mathbb{P}(|\Delta| = x) = \frac{1}{x!} \left(\frac{\lambda^3}{6}\right)^x e^{-\frac{\lambda^3}{6}}, \quad x = 0, 1, 2, \ldots.$$

Convergence to the Poisson distribution holds for $|\mathcal{G}|$ when g has the same number of vertices and edges, with the Poisson parameter equal to $\lim_{n\to\infty} \mathbb{E}(|\mathcal{G}|)$.

The implication of Theorem 10 is that in every realization of $ER_n(\lambda/n)$ we will see some random number of triangles that will approximately follow a Poisson distribution with fixed mean. In particular, the number of triangles will not grow with n. In other words, compared to the network size, triangles will become rare. In the context of modeling real-life networks, especially social networks, this is

not very realistic because such networks typically contain many triangles. This is one of the drawback of the sparse E-R random graph in terms of modeling real-life networks.

4.3.4 Law of large numbers: $\mathbb{E}(|\mathcal{G}|) = O(n)$

Pre-requisite knowledge:

The variance of a linear function of random variables (required to understand why the result in Theorem 11 holds).

As an example, let us again look at the number of wedges. In $ER_n(\lambda/n)$, we have

$$\mathbb{E}(|\mathcal{W}|) = \frac{n(n-1)(n-2)}{2}\frac{\lambda^2}{n^2}\left(1-\frac{\lambda}{n}\right).$$

This number goes to infinity linearly with n. Then what number of wedges would we typically see in $ER_n(\lambda/n)$? Clearly, $|\mathcal{W}|$ is random, but it turns out that $|\mathcal{W}|$ is of the same order of magnitude as its average.

To see this, it is more informative to divide $\mathbb{E}(|\mathcal{W}|)$ by n so that the resulting expression converges to a finite limit:

$$\lim_{n\to\infty}\frac{E(|\mathcal{W}|)}{n} = \lim_{n\to\infty}\frac{n(n-1)(n-2)}{2n}\frac{\lambda^2}{n^2} = \frac{\lambda^2}{2}.$$

It turns out that $\frac{|\mathcal{W}|}{n}$ converges to the same limit $\frac{\lambda^2}{2}$ as its average. Since $|\mathcal{W}|$ is a random variable, there are several ways to define what convergence means. Here we use so-called *convergence in probability*. We will first state the theorem, and then will explain its meaning and why it holds.

Theorem 11. *In $ER_n(\lambda/n)$, as $n\to\infty$, the scaled number of wedges, $\frac{|\mathcal{W}|}{n}$, converges in probability to $\frac{\lambda^2}{2}$. This convergence in probability means that for any $\varepsilon > 0$,*

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\frac{|\mathcal{W}|}{n}-\frac{\lambda^2}{2}\right|>\varepsilon\right)=0.$$

This result is called the Law of Large Numbers for $|\mathcal{W}|$.

Theorem 11 states that when n grows, the deviation of $\frac{|\mathcal{W}|}{n}$ from $\frac{\lambda^2}{2}$ by more than ε becomes unlikely: the probability of this event goes to zero. Informally, $\frac{|\mathcal{W}|}{n}$ comes closer and closer to $\frac{\lambda^2}{2}$ when $n\to\infty$.

Does this mean that $|\mathcal{W}| \approx \frac{\lambda^2}{2}n$? No. Random variable $|\mathcal{W}|$ will vary a lot around $\frac{\lambda^2}{2}n$ from one realization to another. However, these deviations from

the main, linear in n, term $\frac{\lambda^2}{2}n$, will be of the order smaller than n (to be precise, the strandard deviation of $|\mathcal{W}|$ is of the order $\sqrt{n}$, as we will establish later in this section).

Theorem 11 is called 'the law of large numbers' for $|\mathcal{W}|$ because the classical law of large numbers states that a sum of independent identically distributed random variables, divided by the number of terms, converges in probability to its mean. Theorem 11 is similar in spirit because $|\mathcal{W}|$ is a sum of indicators, and n is the order of its main term.

Theorem 11 is proved using another well-knwon result in probability theory, the *Chebychev inequality*. We state this inequality in the next Theorem.

Theorem 12 (Chebychev inequality)**.** *Let $X \in \mathbb{R}$ be a random variable such that $\mathbb{E}(X) < \infty$ and $Var(X) < \infty$. Then for any $\varepsilon > 0$ it holds that*

$$\mathbb{P}(|X - \mathbb{E}(X)| > \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}.$$

Theorem 11 is proved by applying Chebychev's inequality, and taking the limit:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{|\mathcal{W}|}{n} - \frac{\lambda^2}{2}\right| > \varepsilon\right) \leq \lim_{n\to\infty} \frac{\mathrm{Var}\left(\frac{|\mathcal{W}|}{n}\right)}{\varepsilon^2}.$$

Since ε is a constant, the result is proved after we prove that $\lim_{n\to\infty} \mathrm{Var}\left(\frac{|\mathcal{W}|}{n}\right) = 0$.

From the properties of the variance, we know that

$$\mathrm{Var}\left(\frac{|\mathcal{W}|}{n}\right) = \frac{1}{n^2}\mathrm{Var}(|\mathcal{W}|),$$

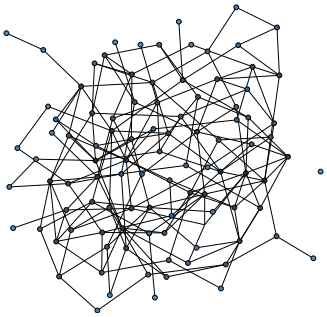
so the main task in the proof is computing $\mathrm{Var}(|\mathcal{W}|)$.

We know that the *Binomial* $\left(\frac{n(n-1)(n-2)}{2}, \frac{\lambda^2}{n^2}\right)$ distribution has variance of the same order of magnitude n as the mean. The number $|\mathcal{W}|$ does not follow the Binomial distribution because some wedges are dependent. Computing $\mathrm{Var}(|\mathcal{W}|)$ is the tricky technical part of the proof. Nevertheless, it turns out that, since most wedges are independent, the moderate dependence does not change the order of magnitude of $\mathrm{Var}(|\mathcal{W}|)$. So, $\mathrm{Var}(|\mathcal{W}|)$ is of the order n (hence, the standard deviation is of the order $\sqrt{n}$). It follows that $\lim_{n\to\infty} \frac{\mathrm{Var}(|\mathcal{W}|)}{n^2} = 0$.

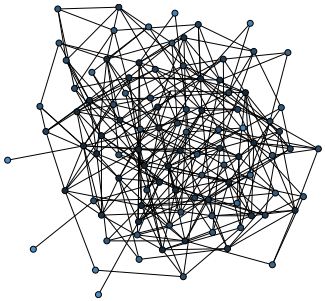
Similar laws of large numbers hold for any graphlet in $ER_n(\lambda/n)$ that has more vertices than edges. Any such graphlet has $r = m + 1$, and the fraction $\frac{|\mathcal{G}|}{n}$ converges in probability to the limit of its mean.

5 Vrijwel zekere garantie dat een stochastische graaf is verbonden
Lezing 3, Nelly Litvak

Veel belangrijke levensechte netwerken zijn verbonden. Op het internet kan informatie van elke router naar elke andere router worden verzonden. Maar als we een netwerk modelleren als een stochastische graaf en dan willekeurig verbindingen maken, zal deze graaf dan nog steeds verbonden zijn? In deze sessie zullen we deze vraag beantwoorden voor de Erdős-Rényi stochastische graaf. De eenvoud van dit E-R model zal ons helpen om de eerste stappen te beredeneren: in de E-R stochastische graaf hebben we n knooppunten, en elk paar knooppunten is met elkaar verbonden met kans p. Laten we nu eens kijken naar de extreme gevallen, iets wat wiskundigen vaak doen wanneer ze tegen een nieuw probleem aanlopen. Wanneer $p = 0$, dan zullen er geen verbindingen zijn, en zal de graaf niet verbonden zijn. Als $p = 1$, dan zal elk paar knooppunten met elkaar verbonden zijn, en is de graaf dus vanzelfsprekend verbonden. Dus, als p groter wordt, dan wordt de kans dat de graaf verbonden is ook groter. Als we dus p groot genoeg kiezen, kunnen we er vrijwel zeker van zijn dat de graaf verbonden is. Maar wat is groot genoeg? Gebaseerd op Figuren 5.1a en 5.1b kunnen we misschien aannemen dat $p = 0.05$ zal voldoen? Of zal het afhangen van de waarde van n? We zullen samen deze prachtige wiskundige puzzel op gaan lossen. En er is meer! Het antwoord zal leiden tot een fascinerend verhaal over wat stochastische grafen gemeen hebben met water dat in ijs verandert.



(a) E-R random graph with $n = 100$, $p = 0.04$. The graph is disconnected.



(b) E-R random graph with $n = 100$, $p = 0.05$. The graph is connected.

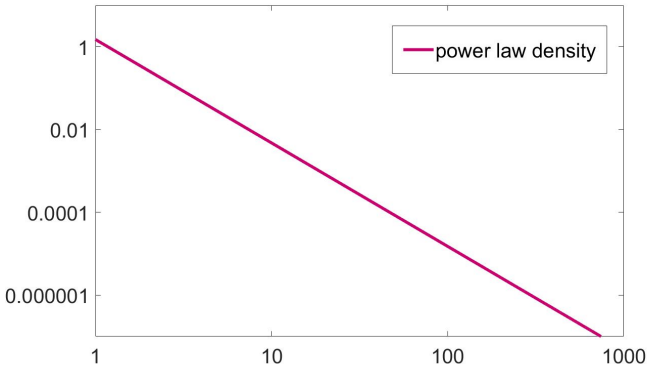
Figuur 5.1

6 Modelleren van schaalvrije netwerken
Lezing 4, Nelly Litvak

Hoeveel verbindingen heeft een knoop in een netwerk? Een van de meest verbazingwekkende eigenschappen van levensechte netwerken is dat het aantal verbindingen van een knoop, dit noemen we de graad van een knoop, heel erg kan variëren tussen verschillende knopen. Sommige knopen hebben maar een paar verbindingen, terwijl anderen er wel miljoenen kunnen hebben. In de introductie hebben we al gezien dat dit het geval is voor de Webgraaf (zie Figuur 1.1c). Maar ook het retweet netwerk in Figuur 1.1a heeft deze eigenschap: de meeste twitteraars hebben geen retweets, maar sommige 'sterren' in het twitteruniversum worden heel vaak geretweet! Op dezelfde manier is het internet in Figuur 1.1b ook een schaalvrij netwerk. In de wiskunde kunnen we dit modelleren met de zogenoemde *power law* verdeling. Deze power law is geformuleerd als volgt:

$$\frac{\#\text{ knopen met } k \text{ verbindingen}}{\text{totaal } \#\text{ knopen}} \approx \text{constante} \cdot k^{-\tau}, \quad \tau > 1.$$

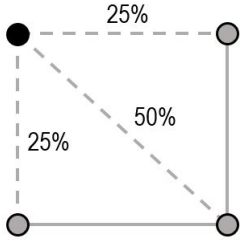
We noemen deze relatie een power law door de negatieve exponent (power) van k. Je kan de power law herkennen door het te plotten op een zogenoemde log-log schaal, zoals we gedaan hebben in Figuur 6.1: op de horizontale as hebben we nu 1, 10, 100, ... in plaats van 1, 2, 3, ..., en op de verticale as hebben we nu 1, 0.1, 0.01, De power law kan worden herkend aan deze kenmerkende rechte lijn in de log-log plot. Er is echter een verhitte wetenschappelijke discussie gaande of de graad van knopen in real-life netwerken echt kan worden beschouwd als een power law. Er is zelfs een artikel geschreven over deze discussie in het NRC! ('Hoe machtig is het superknooppunt?' door Alex van den Brandhof, NRC, 20-12-2019.) We zullen verder niet ingaan op deze discussie. Voor deze cursus is het het belangrijkst om te weten dat de power law een algemeen aanvaard wiskundig begrip is dat het schaalvrije verschijnsel van netwerken goed weergeeft. In deze sessie zullen we op twee verschillende manieren bewijzen dat power laws een valide model zijn voor het schaalvrije verschijnsel. Daarna gaan we terug naar de stochastische grafen, en zullen we zien dat de Erdős-Rényi (E-R) stochastische graaf niet deze power-law-gradenverdeling heeft. Kunnen we dan het E-R model zo manipuleren dat de power laws wel kunnen worden gebruikt? Het antwoord is ja: deze manipulatie heet een gegeneraliseerde stochastische graaf, en zorgt ervoor dat de power laws weer terug komen, zoals we kunnen zien in de rechte lijn in Figuur 6.1.



Figuur 6.1: Probability density function of a power law, in the log-log scale, $\tau = 2.5$.

7 Opkomst van power laws in het preferential attachment model
Lezing 5, Nelly Litvak

We kunnen dit schaalvrije verschijnsel observeren, meten en zelfs modelleren, maar toch is er nog een vraag onbeantwoord: waarom zijn netwerken eigenlijk schaalvrij? Een poging om deze vraag te beantwoorden is door middel van het preferential attachment model: een dynamisch wiskundig model van de groei van een netwerk. Dit model formaliseert het mechanisme dat ook bekend is als 'rijken worden rijker', of 'bekende mensen worden nog bekender'. Dit werkt globaal als volgt: we beginnen met een netwerk met drie knopen, zoals de drie grijze punten in Figuur 7.1. Wanneer de volgende knoop verschijnt (het zwarte punt in de figuur), kan deze knoop een verbinding maken met èèn van de drie knopen (stippellijnen). Om dat te doen gebruikt de nieuwe knoop het mechanisme van 'rijken worden rijker': de kans om te verbinden met een grijze knoop is evenredig met het aantal verbindingen dat de knoop op dit moment heeft. In de figuur heeft een van de grijze knopen twee verbindingen, waardoor deze een hogere kans heeft om er nog eentje te krijgen. Dus, hoe meer verbindingen een knoop krijgt, hoe makkelijker het is om nog meer verbindingen te krijgen. Precies het 'rijken worden rijker'-mechanisme! Het is heel natuurlijk dat met zulke mechanismen sommige knopen erg goed verbonden zullen zijn met de rest van de graaf. In deze sessie zullen we gaan kijken naar de geschiedenis van deze preferential attachment modellen en zullen we wiskundig aantonen dat deze modellen zorgen voor power laws.

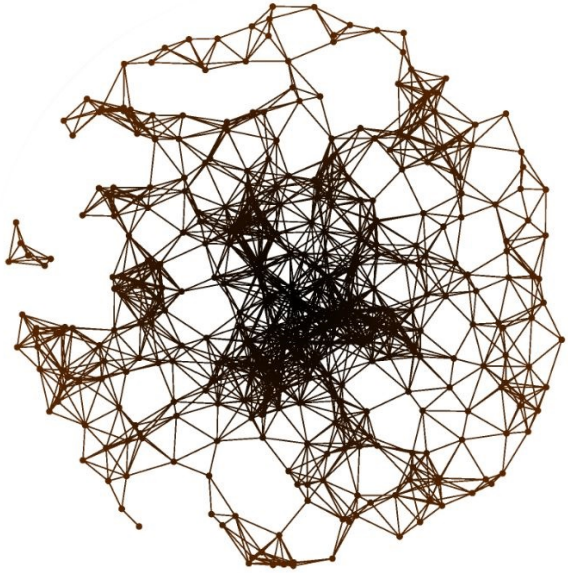


Figuur 7.1: A new (black) vertex arrives in the network, and connects to existing vertices with probabilities proportional to their degrees.

8 Geometrie voor het modelleren van driehoeken

Lezing 6, Nelly Litvak

We hebben al gezien dat een schaarse Erdős-Rényi stochastische graaf maar weinig driehoeken heeft. Dus nu is de vraag: is er een geschikt model voor een netwerk met veel driehoeken? Een fundamentale manier om deze vraag te beantwoorden is door het introduceren van geometrie in een netwerk. Dit kunnen we doen door de knopen in een multi-dimensionale ruimte te plaatsen. Soms is zo'n geometrie al aanwezig in een netwerk, bijvoorbeeld in een netwerk van vliegvelden die verbonden zijn door directe vluchten: elk vliegveld heeft een locatie. Om het iets abstracter maken: we kunnen de 'locatie' van een knoop defini eren aan de hand van de eigenschappen van deze knoop. Een voorbeeld: we kunnenmensen 'plaatsen' in een multi-dimensionale ruimte aan de hand van hun leeftijd en hobby's. In een geometrische stochastische graaf heeft een paar knopen een verbinding als ze dichtbij elkaar zijn. Neem bijvoorbeeld een sociaal netwerk: mensen met dezelfde eigenschappen hebben een grotere kans om vrienden met elkaar te zijn. Dit is een heel natuurlijk en aantrekkelijk idee! Veel netwerkwetenschappers geloven daarom ook dat geometrische stochastische grafen de enige manier zijn om realistische modellen te krijgen van complexe netwerken. In deze sessie zullen we wiskundig bewijzen dat er inderdaad veel driehoeken zijn in geometrische stochastische grafen, zoals te zien is in Figuur 8.1. Maar we zullen ook naar andere eigenschappen kijken: hoe kunnen we een geometrische stochastische graaf schaars maken? Kan het ook schaalvrij zijn?



Figuur 8.1: A geometric random graph. Image: Pim van der Hoorn.

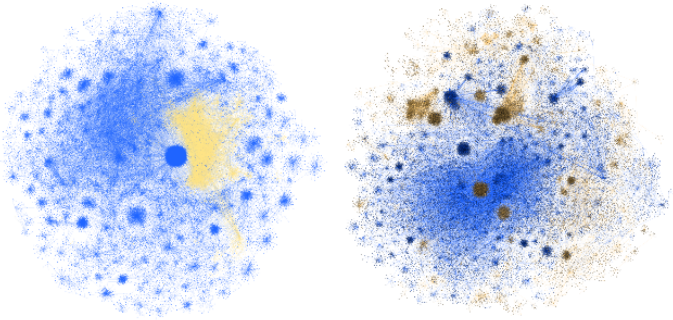
9 Wie is het belangrijkst in een netwerk?
Lezing 7, Pim van der Hoorn

Net zoals post op sociale-media, is niet elke knoop in een netwerk even relevant. Dus, welke knopen zijn het belangrijkst? Deze simpele vraag blijkt niet zo simpel te beantwoorden. De voornaamste reden is dat "belangrijk" verschillende dingen kan betekenen, afhankelijk van het netwerk of simpelweg de vraag die je hiermee wilt beantwoorden. Desalniettemin hebben onderzoekers verschillende manieren ontwikkeld om het belang van knopen te meten en ze met elkaar te kunnen vergelijken. Deze methodes vallen onder de noemer centraliteitsmaten. Zij rangschikken de knopen in een netwerk gebaseerd op wat belangrijk betekent. In het eerste deel van deze sessie zullen we kijken naar verschillende centraliteitsmaten die te maken hebben met paden en navigatie in netwerken. We zullen leren wat de intuïtie achter hun definitie is en wat ze ons kunnen vertellen over het belang van de knopen in een netwerk. Hierna gaan we onze mouwen opstropen en de kersverse kennis toepassen op misschien wel het meest notoire transportnetwerk in Nederland: het spoorweg-netwerk van de NS (Figuur 9.1). We zullen zien hoe verschillende maten de stations op een andere manier rangschikken en bespreken wat dit ons kan vertellen over hoe belangrijk de stations echt zijn. Uiteindelijk kunnen we misschien een station aanwijzen dat het belangrijkst is en ook uitleggen waarom. En nee, het antwoord hoeft niet altijd Utrecht Centraal te zijn.

Figuur 9.1: Wat is het belangrijkste station in Nederland's meest notoire transport netwerk? Afbeelding: Wikimedia Commons

10 Hoe maak je een netwerk efficiënt?
Lezing 8, Clara Stegehuis

Hoe kunnen berichten en fake news zo snel viral gaan op social media? En wat is de rol van netwerken? In deze lezing gaan we interactief op zoek naar de belangrijkste netwerkeigenschappen die voor snelle verspreiding zorgen. We tekenen efficiënte en minder efficënte netwerken voor verspreidingen en we onderzoeken met welke wiskundige netwerkeigenschappen dit samenhangt.



Figuur 10.1: Fake news (geel/bruin) dat zich verspreidt over Twitter. Links: Fake news over vliegtuigsporen in de lucht mengt zich met gewone berichten over de lucht. Rechts: antivaxberichten mengen zich met berichten over de griep.

Voor wie is PWN interessant?

Beroepswiskundigen
Wiskundeleraren
Bedrijven
Leerlingen en studenten
Breed publiek

Platform Wiskunde Nederland is hét landelijke loket voor alles wat met wiskunde te maken heeft.

PWN behartigt de belangen van, en fungeert als spreekbuis voor, de gehele Nederlandse wiskunde.

Platform Wiskunde Nederland | Science Park 123 | kamer L013 | 1098 XG Amsterdam | 020 592 40 06

Ga voor meer informatie naar:
www.platformwiskunde.nl

platform
wiskunde nederland