Network Science

R. Lambiotte

In this section, we introduce mathematical tools used in the course. They are a collection of elementary probability theory and linear algebra, and somewhat advanced materials on stochastic processes. Some of these topics will covered during the first lesson. Some computational problems are proposed.

I. MATHEMATICAL TOOLBOX

In this section, we introduce mathematical tools used in the subsequent chapters. They are a collection of elementary probability theory and linear algebra, and somewhat advanced materials on stochastic processes.

A. Probability

1. Discrete variables

A random variable is a variable that takes its value stochastically. A discrete random variable X is defined on a set S of possible values x such that $p(x) \geq 0$ for any $x \in S$ and $\sum_{x\in S} p(x) = 1$, where $p(x)$ is the probability that X takes value x; we use p to denote the probability throughout these notes. A common example is a fair dice for which $S = \{1, 2, 3, 4, 5, 6\}$ and $p(x) = 1/6$ for each $x \in S$. If one throws the dice many times, the fraction of times with which one observes 1 tends to $1/6$.

A probabilistic event is specified by a certain subset of possible values in X. In the previous example, the event that a dice produces an odd number is represented as the event $X \in \{1, 3, 5\}$. When two events X_1 and X_2 are mutually exclusive, i.e., no value x belongs to both sets, we obtain

$$p(X_1 \text{ or } X_2) = p(X_1) + p(X_2). \tag{1}$$

Information about one event may inform the probability of another event. For instance, knowing that the dice produces an odd number increases the probability of $X = 1$ and decreases the probability of $X = 2$ to zero. Such information is quantified by the conditional probability. The conditional probability of X given Y is

$$p(X|Y) = \frac{p(X \text{ and } Y)}{p(Y)}. \tag{2}$$

By swapping X and Y in Eq. (2), we obtain $p(Y|X) = p(X \text{ and } Y)/p(X)$. By combining this equation with Eq. (2), we obtain the Bayes rule for conditional probabilities:

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)}. \tag{3}$$

Two events X and Y are said to be independent if the probability that X occurs is not affected by whether Y has occurred and vice versa. In other words,

$$p(X|Y) = p(X|\text{not } Y) = p(X). \tag{4}$$

When two events are independent, the probability that both events occur is the product of the probabilities that each event occurs, i.e.,

$$p(X \text{ and } Y) = p(X)p(Y). \tag{5}$$

In terms of the values of X and Y, we obtain

$$p(x, y) = p(x)p(y), \tag{6}$$

where $p(x, y)$ is the joint probability that $X = x$ and $Y = y$. The marginal distribution, i.e., the probability that $X = x$ regardless of the value of Y, is obtained by

$$p(x) = \sum_y p(x, y). \tag{7}$$

In principle, the random variable X can be either numerical (e.g., $1, 2, 3$) or non-numerical (e.g., white, red, black). In the former case, mostly relevant in these notes, there exist different types of tools to characterise its properties. For instance, the expected value, or average, is defined as

$$\langle x \rangle = \sum_x xp(x). \tag{8}$$

We use $\langle\cdot\rangle$ to denote the mean throughout these notes. The nth moment of X is defined by

$$\langle x^n \rangle = \sum_x x^n p(x), \tag{9}$$

where n is typically a positive integer, generalising Eq. (8). The second moment $\langle x^2 \rangle$ is related to the variance as follows:

$$\sigma^2 = \langle (x - \langle x \rangle)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2, \tag{10}$$

where σ is the standard deviation. Moments can be generalised to the case of multiple random variables, often to evaluate correlations between them. A familiar measure of linear dependence between two variables is the Pearson correlation coefficient defined by

$$\rho_{X,Y} = \frac{\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle}{\sigma_X \sigma_Y}, \tag{11}$$

where σ_X and σ_Y are the standard deviations of X and Y, respectively.

Here is a short list of frequently used discrete distributions:

1. The Bernoulli distribution takes only two possible values, 0 or 1, i.e., failure or success, with probabilities $1-p$ and p respectively. The mean $\langle x\rangle = p$ and the variance $\sigma^2 = p(1-p)$.

2. The binomial distribution describes the outcome of n independent and identically distributed random variables generated by the Bernoulli distribution with parameter p. The probability that exactly m successes are observed is given by

$$p(m) = \binom{n}{m} p^m (1-p)^{n-m}, \tag{12}$$

where $0 \leq m \leq n$. Note that $p^m(1-p)^{n-m}$ is the probability that a particular sequence containing exactly m successes is realised, and

$$\binom{n}{m} = \frac{n!}{m!(n-m)!} \tag{13}$$

is the number of sequences of length n that possess exactly m successes. We obtain $\langle m\rangle = np$ and $\sigma^2 = np(1-p)$.

3. The geometric distribution is defined via the waiting time before a success is observed, in a sequence of independent and identically distributed random variables obeying the Bernoulli distribution. The geometric distribution is defined as

$$p(m) = (1-p)^m p, \tag{14}$$

where $m = 0, 1, \ldots$. The factor $(1-p)^m$ corresponds to m consecutive failures, and p to the success on the $(m+1)$th trial. We obtain $\langle m\rangle = (1-p)/p$ and $\sigma^2 = (1-p)/p^2$.

4. The Poisson distribution is given as the limit of the binomial distribution as $n \to \infty$ while the mean np tends to a constant λ (therefore $p \to 0$). The Poisson distribution is given by

$$p(m) = \frac{m^\lambda e^{-\lambda}}{m!}, \tag{15}$$

where $m = 0, 1, \ldots$. We obtain $\langle m\rangle = \sigma^2 = \lambda$.

Ex.III.1 : Using the computer language of your choice, calculate the mean and variance of a Bernouilli process, as a function of p.

2. Continuous variables

Continuous random variables describe variables that take any value in a continuum of values, typically any real values or non-negative real values. Continuous random variables are set by their probability density function, $f(x)(\geq 0)$, defined such that the probability of observing any value between a and b is equal to

$$p(a \leq X \leq b) = \int_a^b f(x)\mathrm{d}x, \tag{16}$$

where $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$.

Most operations for discrete random variables are easily transferred to continuous random variables via replacement of sums by integrals. For instance, the joint probability density function $f(x,y)$ for continuous random variables satisfies

$$p(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x,y)\mathrm{d}x\mathrm{d}y. \tag{17}$$

The moments of the distribution are given by

$$\langle x^n\rangle = \int_{-\infty}^{\infty} x^n f(x)\mathrm{d}x. \tag{18}$$

If two random variables are independent, their joint distribution factorises into the product of their marginals:

$$f(x,y) = f(x)f(y). \tag{19}$$

Finally, it is often practical to focus on the cumulative probability $F(x)$, defined as the probability that the variable takes a value smaller than x:

$$F(x) = \int_{-\infty}^{x} f(x')\mathrm{d}x'. \tag{20}$$

By definition, $F(-\infty) = 0$ and $F(\infty) = 1$. We also often use the complementary cumulative probability, also called the survival probability or survival function, given by

$$\tilde{F}(x) = \int_x^{\infty} f(x')\mathrm{d}x' = 1 - F(x). \tag{21}$$

Classical distributions for continuous random variables include the following ones:

1. The uniform distribution takes a constant probability on interval $[a,b]$, i.e.,

$$f(x) = \frac{1}{b-a} \quad (a \leq x \leq b). \tag{22}$$

We obtain $\langle x\rangle = (b-a)/2$ and $\sigma^2 = (b-a)^2/12$.

2. The exponential distribution is defined by

$$f(x) = \lambda e^{-\lambda x} \quad (x \geq 0). \tag{23}$$

Its cumulative distribution is given by

$$F(x) = 1 - e^{-\lambda x} \quad (x \geq 0). \tag{24}$$

We obtain $\langle x\rangle = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$.

3. The Gaussian or normal distribution is defined by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (-\infty < x < \infty), \tag{25}$$

where μ is the average and σ^2 is the variance. The Gaussian distribution exhibits a bell shape. The Gaussian distribution can been seen as a continuous limit of the binomial distribution. The binomial distribution with n trials, each with probability p, converges to the Gaussian distribution with mean np and variance $np(1-p)$ owing to the central limit theorem. In particular for this reason, the Gaussian distribution is frequently observed in empirical data.

Ex..2 : Given an arbitrary continuous distribution determined by the cumulative probability $F(x)$, design a computational method that generates the corresponding random variables.

B. Renewal processes

Let us consider a system where events take place in a discrete and apparently random fashion. Those events may be emails arriving in a mail box, or atoms colliding in a gas. Such systems are often modelled, as a first order approximation, by a Poisson process, also called the homogeneous Poisson process. The Poisson process assumes that the events are independent of each other, that the rate at which the events take place is constant over time and that time is continuous. These assumptions are often violated in empirical data. For instance, in the case of emails, their reception certainly depends on the time of the day and on the day of the week. In addition, emails are often not independent processes; an email may trigger a discussion thread between two users, causing a cascade of emails. Yet, the Poisson processes are advantageous in their simplicity, which allows us to exactly calculate their properties and make them serve as a baseline model.

The Poisson process is defined as follows. Consider a time window of duration Δt and the probability q that an event takes place within time Δt. By definition, the event rate is given by $\lambda = q/\Delta t$. A Poisson process is specified by the rate λ for infinitesimally small Δt. For λ to be well-defined, $q \to 0$ must be satisfied as $\Delta t \to 0$. Consistent with this requirement, we do not allow multiple events to occur in a time window when Δt is sufficiently small. An event sequence generated by a Poisson process is shown in Fig. 1(a).

Let us derive two key properties of Poisson processes:

(1) The distribution of inter-event times, i.e., time between consecutive events: Let $p(n,t)$ be the probability of observing n events in time window $[0,t]$. By definition,

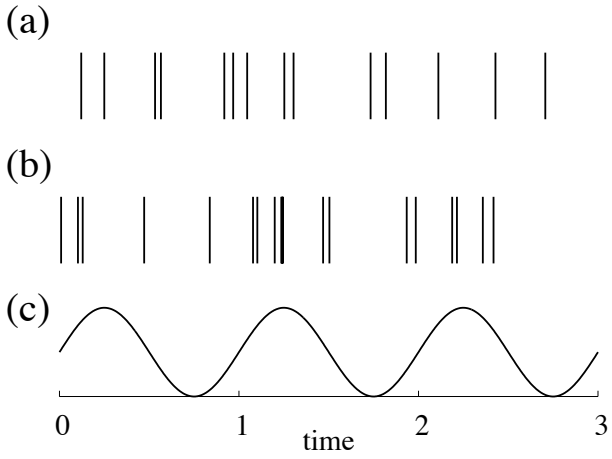

FIG. 1. Homogeneous and non-homogeneous Poisson processes. (a) An event sequence generated by the (homogeneous) Poisson process with $\lambda = 5$. (b) An event sequence generated by the non-homogeneous Poisson process with the sinusoidal rate shown in (c), i.e., $\lambda(t) = 5(1 + \sin 2\pi t)$.

we obtain

$$\begin{aligned} q =& p(1,\Delta t) = \lambda \Delta t, \\ 1-q =& p(0,\Delta t) = 1 - \lambda \Delta t, \end{aligned} \tag{26}$$

when Δt (and hence q) is small. For any $n \geq 1$, we obtain

$$\begin{aligned} p(n,t+\Delta t) =& p(n,t)p(0,\Delta t) + p(n-1,t)p(1,\Delta t) \\ =& p(n,t)(1-\lambda\Delta t) + p(n-1,t)\lambda\Delta t. \end{aligned} \tag{27}$$

Equation (27) holds true because, if there are n events in time window $[0, t+\Delta t]$, either there are n events in $[0,t]$ and no event in $[t, t+\Delta t]$, or there are $n-1$ events in $[0,t]$ and one event in $[t, t+\Delta t]$. This equation relates the probability of a system at a certain time to that at a previous time, and hence is an example of master equation, which we will encounter many times in the following.

In the limit $\Delta t \to 0$, Eq. (27) is reduced to

$$\frac{\mathrm{d}p(n,t)}{\mathrm{d}t} = \lambda p(n-1,t) - \lambda p(n,t). \tag{28}$$

For $n = 0$, we obtain

$$\frac{\mathrm{d}p(0,t)}{\mathrm{d}t} = -\lambda p(0,t), \tag{29}$$

which results in

$$p(0,t) = e^{-\lambda t}. \tag{30}$$

To derive Eq. (30), we have used the initial condition $p(0,0) = 1$, i.e., no event has occurred at $t = 0$. Because $p(0,t)$ is equal to the probability that no event occurs in $[0,t]$, the probability that the first event occurs in $[t, t+\Delta t]$ is given by $p(0,t) - p(0,t+\Delta t)$. Equation (30) implies

that the inter-event time between two consecutive events, denoted by τ, is distributed according to

$$\psi(\tau) = -\frac{\mathrm{d}p(0,\tau)}{\mathrm{d}\tau} = \lambda e^{-\lambda\tau}. \tag{31}$$

The inter-event time of a Poisson process is distributed according to the exponential distribution. The mean inter-event time is given by

$$\langle\tau\rangle = \int_0^\infty \tau\psi(\tau)\mathrm{d}\tau = \frac{1}{\lambda}. \tag{32}$$

In Poisson processes, different inter-event times τ are independent of each other because event times before the last event time t do not affect the time τ to the next event since t. This property is called the renewal property of a Poisson process. Poisson processes satisfy a stronger property, i.e., having no memory in the sense that

$$p(\tau > t_1 + t_2 | \tau > t_2) = p(\tau > t_1). \tag{33}$$

Equation (33) indicates that the length of time, t_2, for which we have waited, actually without an event, does not affect the time of the next event. The time to the next event starting from $t = t_2$, i.e., t_1, is independent of t_2 and obeys $\psi(t_1)$.

(2) The distribution of the number of events observed within a given time window: Using Eq. (28) recursively, we obtain

$$p(n,t) = \frac{(\lambda t)^n}{n!}e^{-\lambda t} \tag{34}$$

for any $n \geq 0$. Therefore, the probability of observing n events in $[0,t]$ obeys the Poisson distribution with mean and variance equal to λt. As discussed in Section I A 1, the Poisson distribution is a limiting case of the binomial distribution when the number of trials is very large and the expected number of successes remains fixed. This interpretation is consistent with the discrete-time formulation of the Poisson process because in $[0,t]$, there are $t/\Delta t$ trials in each of which an event occurs with small probability q. Therefore, the number of events in $[0,t]$ is distributed according to the binomial distribution whose mean is equal to $(t/\Delta t) \times q = \lambda t$.

A method to generate an event sequence obeying a Poisson process is to generate events one by one by independently drawing the inter-event time τ according to Eq. (31). An alternative method when the final time $t_{\max}$ is specified is to first draw the number of events n in $[0, t_{\max}]$ according to the Poisson distribution with parameter $\lambda t_{\max}$. Then, distribute each of the n events independently and uniformly on $[0, t_{\max}]$. The second method exploits the memoryless property of Poisson processes.

Let us introduce two extensions of Poisson processes. The first is non-homogeneous (also called inhomogeneous) Poisson processes, in which the event rate $\lambda(t)$ is time-dependent. In other words, an event occurs in $[t, t+\Delta t]$ with probability $\lambda(t)\Delta t$. This model is motivated by the fact that event rates seem to vary over time in a majority of empirical data. An event sequence generated by a non-homogeneous Poisson process is shown in Fig. 1(b). In this example, the rate is modulated sinusoidally as shown in Fig. 1(c). For a non-homogeneous Poisson process, Eq. (34) is extended as

$$p(n,t) = \frac{\Lambda(t)^n}{n!}e^{-\Lambda(t)}, \tag{35}$$

where

$$\Lambda(t) = \int_0^t \lambda(t')\mathrm{d}t'. \tag{36}$$

The distribution of inter-event times, conditioned by the last event at $t = 0$, is given by

$$\psi(\tau) = \lambda(\tau)e^{-\Lambda(\tau)}, \tag{37}$$

which extends Eq. (31). It should be noted that Eq. (37) is properly normalised, i.e., $\int_0^\infty \psi(\tau)\mathrm{d}\tau = 1$.

The second extension of Poisson processes, called renewal processes, considers a general distribution of inter-event times, $\psi(\tau)$. The renewal property dictates that different inter-event times are independent of each other and drawn from the same distribution $\psi(\tau)$. When $\psi(\tau) = \lambda e^{-\lambda\tau}$, we recover a Poisson process. When $\psi(\tau) = \delta(\tau - 1)$, events periodically happen at all integer times. To obtain the time of the nth event or the number of events in a given time period, we need to sum independent random variables generated according to $\psi(\tau)$. In that case, it is convenient to study the problem in a frequency domain and to consider the Laplace transform, related to the Fourier transform defined below.

Ex.III.3 : Take at random 10^4 numbers in the interval $]0,1[$, and plot the histogram of the $10^4 - 1$ lengths of the resulting intervals.

C. Random walks and diffusion

The Poisson processes provide a basic model for modelling temporal events, i.e., when random events take place. Random walk processes are its counterpart for modelling trajectories in space, i.e., when and where random events take place. Random walk processes are a standard tool to emulate diffusion on networks and also to extract information from the structure of networks, as we will show later. In this section, we derive some basic properties of random walk processes in their simplest setting, when they take place on a one-dimensional space (i.e., line) in discrete time.

In each discrete time step, a walker performs a jump whose length and direction are random variables. The probability density of transition is denoted by $f(r)$. In other words, the probability that the walker located at

x arrives in the interval $[x+r, x+r+\Delta r]$ in one jump is equal to $f(r)\Delta r$. The normalisation condition is given by $\int_{-\infty}^{\infty} f(r)\mathrm{d}r = 1$.

Our aim is to derive the density of the probability density that the walker is located at x after t steps, denoted by $p(x;t)$. Under the assumption that jumps are independent events, we obtain the following master equation:

$$p(x;t) = \int_{-\infty}^{\infty} f(x-x')p(x';t-1)\mathrm{d}x' \tag{38}$$

because the probability of visiting x at time t is the probability of having visited x' at time $t-1$ and performing a jump of displacement $x-x'$.

Equation (38) for the entire range of x is more easily solved in the Fourier domain. The Fourier transform is defined by

$$\hat{p}(k;t) \equiv \int_{-\infty}^{\infty} p(x;t)e^{-ikx}\mathrm{d}x, \tag{39}$$

The original function is recovered through the inverse Fourier transform given by

$$p(x;t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{p}(k;t)e^{ikx}\mathrm{d}k. \tag{40}$$

Probability $p(x;t)$ is thus a combination of the oscillatory functions e^{ikx}, which form a base in the space of functions. The Fourier mode $\hat{p}(k;t)$ is the projection of $p(x;t)$ onto this base. The Fourier transform of $f(x)$, $\hat{f}(k)$, is called the structure function of the random walk. The Taylor expansion around $k=0$ yields

$$\begin{aligned}\hat{p}(k;t) =& \langle e^{-ikx}\rangle \\ =& 1 - ik\langle x\rangle - \frac{1}{2}k^2\langle x^2\rangle + O(k^3).\end{aligned} \tag{41}$$

Equation (41) implies that the moments of $p(x;t)$ are obtained from the derivatives of $\hat{p}(k;t)$ at $k=0$.

The Fourier transform transfers a convolution, such as Eq. (38), to a product. For this reason, working in the Fourier domain is often recommended when dealing with problems involving summations of random variables. Equation (38) is equivalent to

$$\hat{p}(k;t) = \hat{f}(k)\hat{p}(k;t-1). \tag{42}$$

If the walker is initially located at $x=0$, such that $p(x;0)=\delta(x)$, which translates to $\hat{p}(k;0)=1$, we obtain

$$\hat{p}(k;t) = \left[\hat{f}(k)\right]^t. \tag{43}$$

Using the inverse Fourier transform (Eq. (40)), the formal solution of the random walk in the time domain is given by

$$p(x;t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left[\hat{f}(k)\right]^t e^{ikx}\mathrm{d}k. \tag{44}$$

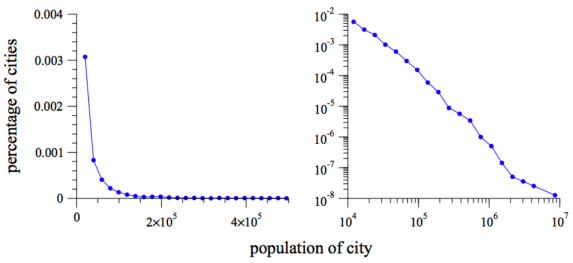

FIG. 2. Left: histogram of the populations of all US cities with population of 10 000 or more. Right: another histogram of the same data, but plotted on logarithmic scales. The approximate straight-line form of the histogram in the right panel implies that the distribution follows a power law. Data from the 2000 US Census.. Taken from Adamic, Lada A. "Zipf, power-laws, and pareto-a ranking tutorial." Xerox Palo Alto Research Center, Palo Alto, CA, http://ginger. hpl. hp. com/shl/papers/ranking/ranking. html (2000).

This solution generally depends on the details of the structure function, $\hat{f}(k)$. However, the asymptotic behaviour of the random walk as t grows only depends on some of its properties. When the first two moments of the structure function are finite, the solution converges to the Gaussian profile

$$p(x;t) = \frac{1}{(2\pi Dt)^{1/2}}e^{-\frac{(x-vt)^2}{4Dt}} \tag{45}$$

with a variance growing linearly with time.

Ex.III.4 : Take a RW on a discrete one-dimensional line. Assuming that the walker has, at each step, a probability 1/2 to go to the left and 1/2 to go to the right. Explore by numerical simulations how the probability distribution evolves over time, and verify the accuracy of Eq.(45). Provide a simple metric to test the "Gaussianity" of the distribution.

D. Power-law distributions

We have seen the emergence of two types of distributions in stochastic processes, the exponential distribution in the case of Poisson processes, and the Gaussian distribution in the case of random walk processes. Another type of distribution, i.e., power-law distribution, plays a central role in network science and in the theory of complex systems in general. In this section we overview properties of power-law distributions and raise some flags in order to properly use them when modelling complex systems.

We explain a power-law distribution for continuous variables, keeping in mind that most of the observations generalise to the case of discrete variables. Consider the

Pareto distribution given by

$$p(x) = Cx^{-\alpha} \quad (x \geq x_{\min}), \tag{46}$$

where α is the power-law exponent of the distribution, $x_{\min}(> 0)$ is the minimum value taken by the random variable and $C = (\alpha - 1)x_{\min}^{\alpha-1}$ is the normalisation constant respecting

$$\int_{x_{\min}}^{\infty} Cx^{-\alpha} \mathrm{d}x = 1. \tag{47}$$

Other power-law distributions are by definition asymptotically (i.e., for large x) the same as Eq. (46) up to a normalisation constant.

Power-law distributions mainly differ from the exponential and Gaussian distributions by the significant mass of probability carried by their tail, i.e., large values of x. The exponential and Gaussian distributions have a characteristic scale such that the probability of observing instances many times larger than this scale is negligible. In contrast, under a power-law distribution, a vast majority of instances exhibits small values while few but non-negligible instances produce very large values. Power-law distributions are associated with a broad heterogeneity in the system and are said to have a fat or long tail, because the tail of the distribution is much more populated than in exponential-like distributions. Power-law distributions are typically found in the wealth of individuals, populations of cities, the frequency of words in text, sales of books and music, citations that a scientific paper receives and so forth. Since the advent of the Pareto distribution and the associated Zipf's law, power-law distributions have been studied over a century. We stress that fat tails are also present in distributions without a power-law tail. Examples include stretched exponential distributions and log-normal distributions.

The moments of power-law distributions are given by

$$\langle x^{\beta} \rangle = \int_{x_{\min}}^{\infty} x^{\beta} p(x) \mathrm{d}x = \frac{\alpha - 1}{\alpha - 1 - \beta} x_{\min}^{\beta} \quad (\beta < \alpha - 1). \tag{48}$$

The moments for $\beta \geq \alpha - 1$ are divergent. In particular, the mean $\langle x \rangle$ does not exist for $1 < \alpha \leq 2$, and the variance does not exist for $2 < \alpha \leq 3$. These features impact various structural and dynamical properties of complex systems including networks, as we will see throughout these notes. When $\alpha \leq 1$, the distribution is ill-defined because $\int_{x_{\min}}^{\infty} p(x) \mathrm{d}x$ is divergent such that $p(x)$ cannot be normalised. When a moment, $\langle x^{\beta} \rangle$, diverges, its empirical measurement diverges as the number of samples increases and $\langle x^{\beta} \rangle$ with β only slightly smaller than $\alpha - 1$ converges very slowly. Both the divergence and slow convergence of moments are due to the appearance of extreme values. For example, the sample mean for the power-law distribution with $\alpha = 2$ diverges as we accumulate samples.

In a majority of empirical data, the distribution can be close to Eq. (46) only in a certain range of the variable. However, key observations such as the divergent moments hold true as long as a distribution behaves the same as Eq. (46) when $x \to \infty$ up to a normalisation constant. For example, the Cauchy distribution given by $p(x) = 1/\left[\pi(1 + x^2)\right]$ is qualitatively the same as Eq. (46) with $\alpha = 2$ as $x \to \infty$. It should also be noted that the tail of an empirical distribution ceases to be a power-law beyond a certain scale because of the finiteness of the system. The finite size effect typically leads to exponential cutoffs. Therefore, the power-law regime, if present, usually dominates for values that are neither too small nor too large.

The heterogeneity of power-law distributions is often associated with the presence of inequalities in the system. What fraction w of the total wealth is held by a certain fraction of the richest people when the wealth distribution is given by Eq. (46)? To answer this question, let us first calculate the fraction of the people whose wealth is at least x_0:

$$p(x \geq x_0) = \int_{x_0}^{\infty} Cx^{-\alpha} \mathrm{d}x = \left(\frac{x_0}{x_{\min}} \right)^{-\alpha+1}. \tag{49}$$

The fraction of wealth held by these richest people is given by

$$w(x_0) = \frac{\int_{x_0}^{\infty} x \cdot Cx^{-\alpha} \mathrm{d}x}{\int_{x_{\min}}^{\infty} x \cdot Cx^{-\alpha} \mathrm{d}x} = \left(\frac{x_0}{x_{\min}} \right)^{-\alpha+2} = \left[p(x \geq x_0) \right]^{\frac{\alpha-2}{\alpha-1}}, \tag{50}$$

where we have assumed that $\alpha > 2$ so that the average wealth is finite. Equation (50) neither depends on x_0 nor $x_{\min}$ explicitly, and it provides a direct relation between $w(x_0)$ and $p(x \geq x_0)$. This relation is often called the "80-20 rule", anecdotally meaning that 80% of the wealth is in the hands of the richest 20%. More precisely, setting $p(x \geq x_0) = 0.2$, $w(x_0) = 0.2^{(\alpha-2)/(\alpha-1)}$ can take any value between 0.2 and 1 depending on the value of α. In the limit $\alpha \to \infty$, the system does not exhibit a power-law tail, and we obtain $w(x_0) = 0.2$. In this case, the system is egalitarian. As α decreases, the tail of the distribution becomes fat and inequality grows. In the extreme situation with $\alpha \to 2$, the total wealth belongs to an infinitesimally small fraction of the richest people. In the econometrics literature, the measurement of this effect in empirical data can be done with the Gini coefficient.

Other properties of power-law distributions include the following:

- Power-law distributions are scale-invariant because they satisfy

  $$p(c_1 x) = c_2 p(x) \tag{51}$$

  for large x, where c_1 and c_2 are constants. Equation (51) implies that multiplying the variable, or equivalently, changing the unit in which it is measured, does not affect properties of the system.

- Power-law distributions conveniently take the form of a straight line in a log-log plot because Eq. (46) is equivalent to

$$\log p(x) = \log C - \alpha x. \tag{52}$$

When testing if empirical data are power-law distributed, it is instructive (but not conclusive) to plot their distribution on the log-log scale.

As a side note, tauberian and Abelian types of theorems help us understand power-law tails of probability distributions and functions in general. They are particularly useful for analytically understanding the long-term behaviour of stochastic dynamics when power-law statistics come into play. In short, Tauberian and Abelian theorems are the inverse of each other. The Tauberian theorem for the Laplace transform, i.e. related to the Fourier transform, is stated as follows :

Consider a function $f(t)$ whose asymptotic behaviour is given by $f(t) \approx t^{\rho-1}$ $(\rho > 0)$ for large t. The Laplace transform of $f(t)$ near $s = 0$ is given by $\hat{f}(s) \approx \Gamma(\rho)s^{-\rho}$, where $\Gamma(\rho)$ is the gamma function.

Ex.III.4 : Take an electronic version of a large book (e.g. the Bible), measure the number of occurrences of each word and then plot the distribution of these numbers. Observe the behaviour of the distribution for large values. Plot the Zipf plot of the data, that it is the relation between the rank and the number of occurrences of the words. Any connection between the Zipf plot and the distribution?

E. Maximum likelihood

The previous sections provide mechanisms by which certain families of distributions emerge. When we are confronted with empirical data, a crucial step is to find the parameter values that best reproduce the data, given a model. There exist different approaches to parameter fitting. The most popular one is probably the maximum likelihood method.

Consider a sequence of observations $\{x_i\}$ $(i = 1, 2, \ldots)$. We are trying to fit a certain model whose parameter set is denoted by θ and is assumed to have a finite support for simplicity. Maximum likelihood dictates that the parameter values are chosen to maximise the probability with which the model generates the observed data. To this end, we calculate $p(\theta|\{x_i\})$, which is related to $p(\{x_i\}|\theta)$ by Bayes' law

$$p(\theta|\{x_i\}) = p(\{x_i\}|\theta)\frac{p(\theta)}{p(\{x_i\})}. \tag{53}$$

By definition, the probability of observing certain data, $p(\{x_i\})$, is fixed, and it does affect the optimisation of θ. Moreover, in the absence of other information, it is convenient to assume that any values of θ are equally likely such that the prior distribution $p(\theta)$ is a constant. Then, $p(\theta|\{x_i\})$ and $p(\{x_i\}|\theta)$ are proportional to each other, and the locations of their maximum coincide. Therefore, it suffices to maximise $p(\theta|\{x_i\})$ in terms of θ.

As an example, consider the model in which each x_i independently obeys the same exponential distribution. The likelihood of the data is given by

$$\mathcal{L}(\{x_i\}|\lambda) = \prod_{i=1}^{n} p(x_i|\lambda), \tag{54}$$

where $p(x|\lambda) = \lambda e^{-\lambda x}$ and n is the number of observations. To find the value of λ that maximises the likelihood, we conventionally maximise the logarithm of $\mathcal{L}$. The maximum of

$$\log \mathcal{L}(\{x_i\}|\lambda) = \log \prod_{i=1}^{n} p(x_i|\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i \tag{55}$$

is obtained via

$$\frac{\partial}{\partial \lambda} \log \mathcal{L}(\{x_i\}|\lambda) = 0, \tag{56}$$

which leads to

$$\hat{\lambda} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i} = \frac{1}{\langle x \rangle}. \tag{57}$$

The maximum likelihood estimation is easy if the log likelihood takes an analytical form and its maximum is explicitly computed. Otherwise, we resort to numerical methods such as the expectation-maximisation algorithm.

There are also other situations in which likelihood maximisation needs to be done carefully. For example, suppose that we are fitting the power-law distribution, Eq. (46), to data. Usually, a power-law distribution provides a good fit of empirical data in a regime excluding small x values. Therefore, we regard $x_{\min}$ as the point where the power-law regime starts, which we are interested in estimating in addition to the power-law exponent α. The log likelihood of the data under the power-law distribution is given by

$$\log \mathcal{L}(\{x_i\}|\alpha, x_{\min}) = n \log\left(\frac{\alpha - 1}{x_{\min}}\right) - \alpha \sum_{i=1}^{n} \log\left(\frac{x_i}{x_{\min}}\right). \tag{58}$$

Setting $\partial \log \mathcal{L}/\partial \alpha = 0$ yields the maximum likelihood estimator given by

$$\hat{\alpha} = 1 + \frac{n}{\sum_{i=1}^{n} \log\left(\frac{x_i}{x_{\min}}\right)}. \tag{59}$$

However, finding the optimal $x_{\min}$ value is not a straightforward exercise because changing values of $x_{\min}$ also changes the number of observations, n, falling within the assumed power-law regime, i.e., $x \geq x_{\min}$. The likelihood monotonically decreases with increasing n because

the probability of observing an additional data point is always smaller than unity. Therefore, the maximum likelihood in terms of $x_{\min}$ is obtained by a trivial solution $\hat{x}_{\min} = \max_i x_i$, yielding $n = 0$. Other techniques must be used to estimate $\hat{x}_{\min}$. The minimisation of goodness-of-fit statistics, such as the Kolmogorov-Smirnov test, measuring the distance between the cumulative distribution of the empirical data and that of the model, is one such possibility.

Ex.III.5 : Write a code that takes as an input a sequence of real value numbers and returns the best exponential distribution for the intervals. Verify the accuracy of the prediction (including its absence of bias).

F. Entropy, information and similarity measures

The entropy of a random variable, denoted by H, is a measure of its uncertainty and quantifies how much we know about a variable before observing it. After the observation, we get rid of the uncertainty and thus gain information H about the system. For a discrete random variable X, entropy is defined as

$$H(X) = -\sum_x p(x) \log p(x). \tag{60}$$

If X can take one of n states, we obtain $0 \leq H(X) \leq \log n$. The maximum value $H(X) = \log n$ is realised when $p(x)$ is the uniform density, i.e., when $p(x) = 1/n$ for all x. The minimum value $H(X) = 0$ is realised when X is deterministic, i.e., $p(x) = \delta_{x,x_0}$ for a specific x_0, where δ is Kronecker delta. In the latter case, we know the value of X before observing it, hence the lack of uncertainty.

The joint entropy $H(X, Y)$ of a pair of discrete random variables with joint distribution $p(x, y)$ is defined as

$$H(X, Y) = -\sum_x \sum_y p(x, y) \log p(x, y). \tag{61}$$

The conditional entropy $H(Y|X)$ is defined as

$$\begin{aligned} H(Y|X) &= -\sum_x \sum_y p(x, y) \log p(y|x) \\ &= -\sum_x p(x) H(Y|X = x), \end{aligned} \tag{62}$$

and refers to the entropy of Y conditioned on the value of X and averaged over all possible values of X. The joint entropy and conditional entropy are related by the chain rule:

$$H(X, Y) = H(X) + H(Y|X). \tag{63}$$

Equation (63) states that the total uncertainty about X and Y is simply the uncertainty about X, plus the average uncertainty about Y once X is known.

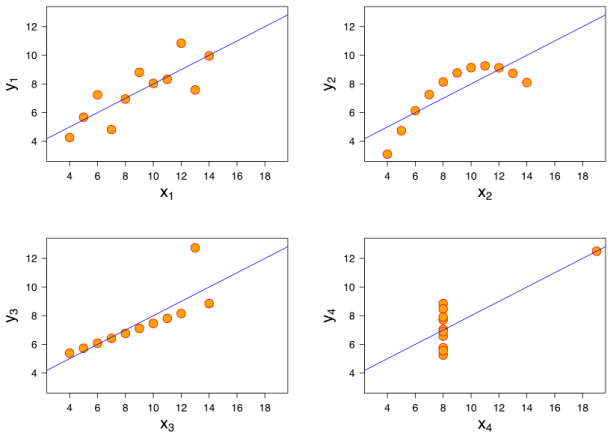

FIG. 3. Anscombe's quartet: all four sets are identical when examined using simple summary statistics, including their Pearson coefficient, but vary considerably when graphed

What does the knowledge of one variable tell us about another one? The conditional entropy $H(Y|X)$ addresses this question. More precisely, mutual information $I(X, Y)$ is defined as the amount of information gained on X by knowing the value of Y as follows:

$$I(X, Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y). \tag{64}$$

If Y is perfectly informative in the sense that it tells us everything about X, mutual information reduces to the entropy of X because $I(X, Y) = H(X) - H(X|Y) = H(X)$. Mutual information is rewritten as

$$I(X, Y) = \sum_x \sum_y p(x, y) \log \frac{p(y, x)}{p(x)p(y)}. \tag{65}$$

Equations (64) and (65) show that mutual information is symmetric, i.e., $I(X, Y) = I(Y, X)$. Mutual information measures the cost of assuming that two variables are independent when they are in fact not. Mutual information captures non-linear correlations between random variables, in contrast to linear quantities such as the Pearson correlation coefficient (see Figure 3).

There exist many situations when we have to compare two networks defined on the same set of nodes, for instance in the case of temporal networks. Mutual information can serve to this end by specifying a distribution $p(x)$ that summarises a network. Other commonly used similarity measures include the Pearson correlation coefficient and the Jaccard index. The Pearson correlation for random variables is given by Eq. (11) and adapted for a list of pairwise observations $\{(x_i, y_i); 1 \leq i \leq n\}$ as follows:

$$\frac{\sum_{i=1}^n (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sqrt{\sum_{i=1}^n (x_i - \langle x \rangle)^2 \sum_{i=1}^n (y_i - \langle y \rangle)^2}}, \tag{66}$$

where $\langle x \rangle = \sum_{i=1}^{n} x_i/n$ and $\langle y \rangle = \sum_{i=1}^{n} y_i/n$. The Jaccard index for two sets S_1 and S_2 is defined by

$$\frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}, \tag{67}$$

where $|\cdot|$ denotes the number of elements in the set. The Jaccard index takes the largest value, 1, when $S_1 = S_2$. It takes the smallest value, 0, when S_1 and S_2 do not have any common element.

Ex.III.6 : After reading Information *Theory and Statistical Mechanics*, by ET Jaynes, calculate the maximum-entropy prediction of the following system:
A traditional die (with 6-faces) is biased. The average value is 5, instead of $3,5$. Calculate the probability to observe a value 6.

G. Matrix algebra

Matrices are a standard representation of networks. Properties of matrices are crucial in order to describe linear dynamical systems and at the core of several algorithms to extract structural information from networks. In this section, we provide a short, practical summary of results from linear algebra, emphasising what will be used in later chapters.

Consider an $N \times N$ matrix A. A vector and scalar value λ satisfying

$$Au = \lambda u \tag{68}$$

are called the eigenvector and eigenvalue, respectively. There are at most N eigenvalues and associated eigenvectors. If A is a symmetric matrix, i.e., $A_{ij} = A_{ji}$ $(1 \leq i, j \leq N)$, all the eigenvalues λ_i $(1 \leq i \leq N)$ are real. In addition, the eigenvectors u_i associated with different eigenvalues λ_i are orthogonal, i.e., $\langle u_i, u_j \rangle = 0$ if $i \neq j$, where $\langle , \rangle$ is the inner product. Matrix A may have duplicated eigenvalues. Even in this case, we can select the set of N eigenvectors such that the orthogonality is respected.

Matrix A is decomposed as

$$A = \sum_{\ell=1}^{N} \lambda_\ell u_\ell u_\ell^\top, \tag{69}$$

where $\top$ represents the transposition. The validity of Eq. (69) is verified by multiplying an arbitrary eigenvector u_i to both sides of Eq. (69). Due to the orthogonality of the eigenvector, we obtain $Au_i = \lambda_i u_i$, assuming that the eigenvectors are properly normalised such that

$$\langle u_\ell, u_{\ell'} \rangle = \delta_{\ell\ell'}. \tag{70}$$

By combining Eqs. (69) and (70), we obtain

$$A^n = \sum_{\ell=1}^{N} \lambda_\ell^n u_\ell u_\ell^\top. \tag{71}$$

We are often interested in the extremal eigenvalue such as the largest eigenvalue of a symmetric matrix A, i.e., $\lambda_{\max}$. The Perron-Frobenius theorem guarantees that when all elements of A are strictly positive, $\lambda_{\max}$ is the isolated (i.e., not duplicated) largest eigenvalue. In addition, all elements of the corresponding eigenvector $u_{\max}$, called the Perron-Frobenius vector, have the same sign. Any other eigenvector u_ℓ does not show this property because, due to the orthogonality $\langle u_\ell, u_{\max} \rangle = 0$, some of the elements in u_ℓ must have the opposite signs. The Perron-Frobenius theorem also holds true for asymmetric matrices. In the asymmetric case, the statement that the largest eigenvalue is isolated is replaced by that of the modulus, or the absolute value of the eigenvalue. Matrices appearing in network analysis are often sparse, with a majority of elements being zero. The Perron-Frobenius theorem is also applicable in this situation if matrix A is primitive, i.e., if all elements of A are non-negative and all elements of A^n are positive for some integer $n > 0$. If an undirected network of interest is connected as a single component, which is usually the case in theoretical studies, matrices representing the network are usually primitive (with the exception of so-called bipartite graphs), such that the Perron-Frobenius theorem can be used.

The power method is a computationally efficient method to calculate $\lambda_{\max}$ and $u_{\max}$ of a given matrix. To do this, we start with an (almost) arbitrary initial vector x and repeat multiplying A. By multiplying x to both sides of Eq. (69), we obtain

$$x(1) \equiv Ax = \sum_{\ell=1}^{N} \lambda_\ell u_\ell \langle u_\ell^\top, x \rangle \tag{72}$$

By repeating the multiplication of A on both sides of Eq. (72), we obtain

$$x(n) \equiv A^n x = A^n x(n-1) = \sum_{\ell=1}^{N} \lambda_\ell^n u_\ell \langle u_\ell^\top, x \rangle. \tag{73}$$

If $\lambda_{\max}$ is the isolated eigenvalue, as in the case of the primitive matrix, $\lambda_{\max}^n \gg \lambda_\ell^n$ for any other eigenvalue λ_ℓ for large n. Then, in Eq. (73), all but the one term corresponding to $\lambda_{\max}$ is negligible on the right-hand side as $n \to \infty$. After many iterations, we can obtain the largest eigenvalue $\lambda_{\max}$ by looking at how much each element of $x(n)$ grows by one iterate and the corresponding eigenvector $u_{\max}$ from $x(n)$. In practice, we normalise $x(n)$ in each iterate to avoid the elements of $x(n)$ to become very large or small.

Ex.III.7 : Implement the power-method and test it on some examples.
Ex.III.8 : Calculate numerically the distribution of eigenvalues of a random symmetric matrix A of size 1000, where each entry is an independent Bernouilli random variable, subject to the constraint that $A_{ij} = A_{ji}$ to ensure symmetry.

H. Markov chains

Markov chains are stochastic dynamics on N states in discrete time. A state may be the position in a network having N nodes such that the process represents a random walk on the network. Alternatively, a state may be the number of infected people, between 0 and $N-1$, in a structureless population of $N-1$ individuals. In both cases, we number the states as 1, 2, ..., N. The state at time t ($t = 0, 1, \ldots$), which is a random variable, is denoted by X_t.

In a stochastic process on N states in general, state X_{t+1} may depend on all preceding states (i.e., full history) of the dynamics, i.e., X_0, X_1, ..., X_t. Under the Markov assumption, the conditional probability to observe a state at time $t+1$ only depends on the state at time t. In other words, a discrete-time stochastic process verifying

$$\begin{aligned} p(X_{t+1} = i_{t+1} | X_t = i_t, \ldots, X_1 = i_1, X_0 = i_0) \\ = p(X_{t+1} = i_{t+1} | X_t = i_t), \end{aligned} \tag{74}$$

is called the Markov chain. Among the class of Markov chains, we are often interested in the stationary ones, in which the conditional state-transition probability does not depend on t:

$$p(X_{t+1} = j | X_t = i) \equiv T_{ij}. \tag{75}$$

Processes verifying both properties, Markovianity and stationarity, are called stationary Markov chains. Because a realisation of the process visiting state i must go somewhere including itself in the next time step, we obtain

$$\sum_{j=1}^{N} T_{ij} = 1. \tag{76}$$

A stationary Markov chain is fully described by an initial state and an $N \times N$ transition matrix $T = (T_{ij})$. The probability that state i is visited at time t, denoted by $p_i(t)$, evolves according to

$$p_j(t+1) = \sum_{i=1}^{N} p_i(t) T_{ij} \quad (1 \le j \le N). \tag{77}$$

It should be noted that $\sum_{i=1}^{N} p_i(t) = 1$ for any t, if the initial condition is properly normalised. Equation (77) is compactly rewritten as

$$p(t+1) = p(t) T, \tag{78}$$

where $p(t) = (p_1(t) \ \cdots \ p_N(t))$. Equation (78) yields

$$p(t) = p(0) T^t. \tag{79}$$

A Markov chain is composed of different types of states. By definition, the process does not escape from an absorbing state once it has been reached. State i is absorbing if and only if $T_{ii} = 1$, which implies that $T_{ij} = 0$ for any $j \neq i$. A group of states forms an ergodic set if it is possible to go from i to j for any states i and j in the set and if the process does not leave the set once the process has reached it. An absorbing state is thus an ergodic set composed of a single state. Finally, a state is called a transient state if it is not a member of an ergodic set.

We denote the stationary density by $p^* = (p_1^*, \ldots, p_N^*)$, where $p_i^* = \lim_{t\to\infty} p_i(t)$ ($1 \le i \le N$) and hence $\sum_{i=1}^{N} p_i^* = 1$. Substitution of $p_i(t) = p_i(t+1) = p_i^*$ ($1 \le i \le N$), which holds true in the limit $t \to \infty$, in Eq. (77) yields

$$p^* = p^* T. \tag{80}$$

Therefore, the stationary density is the left eigenvector of T with eigenvalue unity. Because

$$T \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \tag{81}$$

which is a consequence of Eq. (76), T is guaranteed to have an eigenvalue of unity. If the entire set of the N states is ergodic, one can go from i to j for any i and j. In this case, p^* is unique, and iterates of Eq. (79) starting from an almost arbitrary initial condition converge p^* except in special cases.

Then, the eigenvalue of unity is in fact the largest eigenvalue of T in terms of the modulus (i.e., absolute value). Therefore, p^* is the Perron-Frobenius vector. This observation is consistent with the fact that all elements of the Perron-Frobenius vector are positive (Section I G). In addition, Eq. (71) adapted to the case of asymmetric matrices dictates that the discrepancy of $p(t)$ from p^* decays exponentially as $\propto |\lambda_{2\mathrm{nd}}|^t$, where $\lambda_{2\mathrm{nd}}$ is the second largest eigenvalue of T in terms of the modulus. In words, the second largest eigenvalue governs the relaxation time of the iterate. More generally, the speed of convergence is determined by the difference or ratio between $\lambda_{2\mathrm{nd}}$ and $\lambda_{\max}$, with the latter being equal to unity in the current case. Therefore, we often call $1 - \lambda_{2\mathrm{nd}}$ the spectral gap. A Markov chain with a large spectral gap converges rapidly.

Markov chain theory also allows us to answer other types of questions. For example, how long on average do the dynamics need to reach a certain state? What is the probability of ending in a certain absorbing state, depending on the initial condition?

Ex.III.9 : Construct the 3×3 transition matrix of a Markov chain (or your choice) and describe its properties.

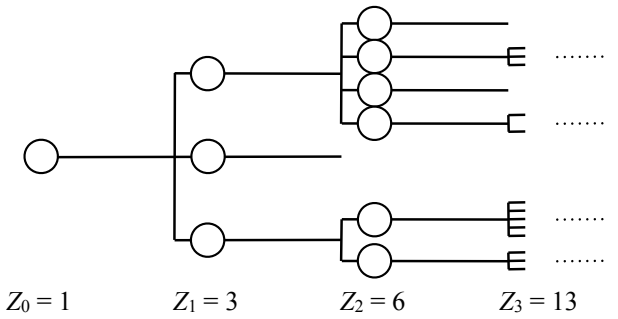

FIG. 4. Schematic of the Galton-Watson branching process.

I. Branching processes

A branching process is a Markov process in which each individual produces some (possibly zero) individuals and then dies, each of the new individuals undergoes reproduction, and so forth (Fig. 4). In network theory, branching processes are a useful tool for understanding network generation and epidemic processes on networks. In network generation, we start from a given node and explore its neighbours, neighbours of neighbours, and so on to expand the network under investigation. In epidemic processes, an initially infected node typically propagates infection to a certain number of neighbouring nodes, each of which then infects some others, and so on. In both cases, the number of nodes that a node newly recruits usually depends on the node and hence can be considered as a random number, as assumed in branching processes.

The Galton-Watson process is a prototypical branching process model defined as follows. Fix the distribution of the number of offspring, $\{p(n)\}$, where $p(n)$ is the probability that an individual reproduces n individuals. The number of individuals in the tth generation is denoted by Z_t (Fig. 4). First, there is initially one individual, i.e., $Z_0 = 1$. Second, this individual generates offspring whose number Z_1 is drawn from $\{p(n)\}$. Third, each of the Z_1 individuals in the first generation produces offspring whose number independently obeys distribution $\{p(n)\}$. The individuals born in this stage, which total Z_2, define the second generation. We repeat this procedure to define further generations until the process gets extinguished. The extinction may not occur, in which case the number of individuals grows indefinitely.

The extinction requires $p(0) > 0$. In other words, an individual does not produce any offspring with a positive probability. If $p(n)$ for large n values is large, the population would grow rather than shrink. In fact, the mean number of offspring, i.e., $\langle n\rangle \equiv \sum_{n=0}^{\infty} np(n)$ is the main determinant of a branching process. If $\langle n\rangle \leq 1$, a realisation of the process will always die out for sufficiently large t, except in the deterministic case $n = \langle n\rangle = 1$ such that each individual always yields exactly one offspring. In particular, $E[Z_t] = \langle n\rangle^t \to 0$ as $t \to \infty$. If $\langle n\rangle > 1$, $E[Z_t]$ exponentially grows and individual realisations of the process may exponentially grow as well.

We denote by q the probability that the process starting from one individual eventually dies out and, as we now show, $q = 1$ when $\langle n\rangle \leq 1$. If an individual produces n individuals, then the process will die out with probability q^n because of the independence of the sub-processes starting from n individuals. Therefore, we obtain the recursive relationship

$$q = \sum_{n=0}^{\infty} p(n)q^n. \tag{82}$$

Equation (82) always has $q = 1$ as a solution. It has a solution with $q < 1$ if and only if $\langle n\rangle > 1$. To show this, we use the fact that the solution is the intersection of

$$y = f_1(q) \equiv q$$

and

$$y = f_2(q) \equiv \sum_{n=0}^{\infty} p(n)q^n$$

. Because $\langle n\rangle > 1$, it suffices to consider the case $p(0) + p(1) < 1$. If $p(0) = 0$, $q = 0$ is a solution because

$$f_2(0) = p(0) = 0 = f_1(0).$$

If $p(0) > 0$, we obtain $0 < f_2(0) < 1$. Because $f_1(1) = f_2(1) = 1$, and

$$\mathrm{d}f_2(q)/\mathrm{d}q = \sum_{n=1}^{\infty} np(n)q^{n-1} > 0$$

and

$$\mathrm{d}^2 f_2(q)/\mathrm{d}q^2 = \sum_{n=2}^{\infty} n(n-1)p(n)q^{n-2} > 0$$

when $0 < q \leq 1$, $y = f_1(q)$ and $y = f_2(q)$ cross in $0 < q \leq 1$ if and only if $\mathrm{d}f_2(q)/\mathrm{d}q > 1$ at $q = 1$. This condition is equivalent to $\langle n\rangle > 1$. In this case, the process grows exponentially with probability $1 - q$.

Ex.III.10 : Consider a Galton-Watson process with $p(0) = a$, $p(2) = 1 - a$ and $p(i) = 0$ otherwise. Estimate numerically the evolution of $E[Z_t]$ as a function of a.

Application : Branching processes are often used to model cascades in social media. For instance, cascades of retweets on Twitter can be represented by trees and modelled by a Galton-Watson, or variations around it. From a practical point of view, the initial structure of a cascade can be fed into a machine learning framework to predict their future success, or the cascade can be used to calibrate the parameters of a branching process for such a prediction. See for instance:
Cheng, Justin, et al. "Can cascades be predicted?." Proceedings of the 23rd international conference on World wide web. ACM, 2014.
Kobayashi, Ryota, and Renaud Lambiotte. "TiDeH: Time-Dependent Hawkes Process for Predicting Retweet Dynamics." ICWSM. 2016.