Renewal theory in analysis of tries and strings

Svante Janson

AofA'10, Vienna, June 2010

Renewal theory in analysis of tries and strings

Svante Janson

AofA'10, Vienna, June 2010

I thank Allan Gut and Wojciech Szpankowski for inspiration and helpful discussions.

Model

$\Xi^{(1)}, \Xi^{(2)}, \dots$ is a sequence of i.i.d. random infinite strings $\Xi = \xi_1\xi_2\cdots$.

For simplicity, suppose that the alphabet $\mathcal{A} = \{0, 1\}$, and that the individual letters ξ_i are i.i.d. with $\xi_i \sim \mathrm{Be}(p)$, i.e., $\mathbb{P}(\xi_i = 1) = p$ and $\mathbb{P}(\xi_i = 0) = q := 1 - p$.

Notation

Given a finite string $\alpha_1 \cdots \alpha_n \in \mathcal{A}^n$, let $P(\alpha_1 \cdots \alpha_n)$ be the probability that the random string Ξ begins with $\alpha_1 \cdots \alpha_n$:

$$P(\alpha_1 \cdots \alpha_n) = \prod_{i=1}^{n} P(\alpha_i) = \prod_{i=1}^{n} p^{\alpha_i} q^{1-\alpha_i}.$$

Given a random string $\xi_1\xi_2\cdots$, we define

$$X_i := -\ln P(\xi_i) = -\ln\bigl(p^{\xi_i}q^{1-\xi_i}\bigr) = \begin{cases} -\ln q, & \xi_i = 0,\\ -\ln p, & \xi_i = 1. \end{cases} \tag{1}$$

Then $X_1, X_2, \dots$ are i.i.d. with

$$\mathbb{E}\, X_i = H := -p\ln p - q\ln q, \tag{2}$$

the usual *entropy* of each letter ξ_i, and

$$\mathbb{E}\, X_i^2 = H_2 := p\ln^2 p + q\ln^2 q, \tag{3}$$

$$\operatorname{Var} X_i = H_2 - H^2 = pq(\ln p - \ln q)^2 = pq\ln^2(p/q). \tag{4}$$

Note that the case $p = q = 1/2$ is special; in this case $X_i = \ln 2$ is deterministic and $\operatorname{Var} X_i = 0$.

$Z_n \sim \mathrm{AsN}(\mu_n, \sigma_n^2)$ means that $(Z_n - \mu_n)/\sigma_n \xrightarrow{\mathrm{d}} N(0,1)$.
Here Z_n is a sequence of random variables and μ_n and σ_n^2 are sequences of real numbers with $\sigma_n^2 > 0$.

We denote the fractional part of a real number x by
$\{x\} := x - \lfloor x \rfloor$.

Renewal theory

Let $S_n := \sum_{i=1}^{n} X_i$, the partial sums of X_i. Thus

$$P(\xi_1 \cdots \xi_n) = \prod_{i=1}^{n} P(\xi_i) = \prod_{i=1}^{n} e^{-X_i} = e^{-S_n}. \tag{5}$$

This is a random variable, since it depends on the random string $\xi_1 \cdots \xi_n$; it can be interpreted as the probability that another random string $\Xi^{(j)}$ begins with the same n letters as observed.

Let, for $t \geq 0$ and $n \geq 1$,

$$\nu(t) := \min\{n : S_n > t\}. \tag{6}$$

We may also start with an initial random variable X_0, which is independent of $X_1, X_2, \dots$, but may have an arbitrary distribution. We then define

$$\widehat{S}_n := \sum_{i=0}^{n} X_i = X_0 + \sum_{i=1}^{n} X_i, \tag{7}$$

$$\widehat{\nu}(t) := \min\{n : \widehat{S}_n > t\}. \tag{8}$$

lattice and non-lattice

In both renewal theory and in the analysis of tries, there often are two cases: the *arithmetic* or *lattice* case when the support is a subset of $d\mathbb{Z}$ for some $d > 0$, and the *non-arithmetic* or *non-lattice* case when it is not.

arithmetic The ratio $\ln p/\ln q$ is rational. More precisely, X_i then is d-arithmetic, where d equals $\gcd(\ln p, \ln q)$. If $\ln p/\ln q = a/b$, where a and b are relatively prime positive integers, then

$$d = \gcd(\ln p, \ln q) = \frac{|\ln p|}{a} = \frac{|\ln q|}{b}. \tag{9}$$

non-arithmetic The ratio $\ln p/\ln q$ is irrational.

Tries

A *trie* is a binary tree structure designed to store a set of strings. The trie is a finite subtree of the complete infinite binary tree $\mathcal{T}_\infty$, where the nodes can be labelled by finite strings $\boldsymbol{\alpha} = \alpha_1 \cdots \alpha_k \in \mathcal{A}^* := \bigcup_{k=0}^{\infty} \mathcal{A}^k$ (the root is the empty string). A string Ξ is stored at the node labelled by $\boldsymbol{\alpha}$ if $\boldsymbol{\alpha}$ is the shortest prefix of Ξ that is not a prefix of any other string in the set.

Let D_n be the depth (= path length) of the node containing the first string in the trie constructed from n random strings $\Xi^{(1)}, \dots, \Xi^{(n)}$. Denoting the chosen string by $\Xi = \xi_1\xi_2\cdots$, the depth D_n is thus at most k if and only if no other of the strings begins with $\xi_1 \cdots \xi_k$.
Conditioning on the string Ξ, each of the other strings has this beginning with probability $P(\xi_1 \cdots \xi_k)$, and thus by independence,

$$\mathbb{P}(D_n \le k \mid \Xi) = \bigl(1 - P(\xi_1 \cdots \xi_k)\bigr)^{n-1} = \bigl(1 - e^{-S_k}\bigr)^{n-1}. \quad (10)$$

Let $X_0 = X_0^{(n)}$ be a random variable, independent of Ξ, with the distribution

$$\mathbb{P}(X_0^{(n)} > x) = \bigl(1-e^x/n\bigr)_+^{n-1} = \bigl(1-e^{x-\ln n}\bigr)_+^{n-1}, \qquad x \in (-\infty,\infty). \tag{11}$$

Then, for any $k \geq 1$,

$$\mathbb{P}(D_n \leq k) = \mathbb{P}\bigl(X_0 > \ln n - S_k\bigr) = \mathbb{P}\bigl(\widehat{S}_k > \ln n\bigr) = \mathbb{P}\bigl(\widehat{\nu}(\ln n) \leq k\bigr) \tag{12}$$

and thus

$$D_n \stackrel{\mathrm{d}}{=} \widehat{\nu}(\ln n). \tag{13}$$

Standard renewal theory theorems immediately yields the following.

Theorem

For every $p \in (0,1)$,

$$\frac{D_n}{\ln n} \xrightarrow{\mathrm{p}} \frac{1}{H}, \tag{14}$$

with H *the entropy. Moreover, the convergence holds in every* L^r, $r < \infty$, *too. Hence, all moments converge above and*

$$\mathbb{E}\, D_n^r \sim H^{-r} (\ln n)^r, \qquad 0 < r < \infty. \tag{15}$$

Theorem

More precisely:

(i) *If* $\ln p/\ln q$ *is irrational, then, as* $n \to \infty$,

$$\mathbb{E}\, D_n = \frac{\ln n}{H} + \frac{H_2}{2H^2} + \frac{\gamma}{H} + o(1). \tag{16}$$

(ii) *If* $\ln p/\ln q$ *is rational, then, as* $n \to \infty$,

$$\mathbb{E}\, D_n = \frac{\ln n}{H} + \frac{H_2}{2H^2} + \frac{\gamma}{H} + \psi_1(\ln n) + o(1), \tag{17}$$

where $\psi_1(t)$ *is a small continuous function, with period* $d = \gcd(\ln p, \ln q)$ *in* t, *given by*

$$\psi_1(t) := -\frac{1}{H} \sum_{k \neq 0} \Gamma(-2\pi \mathrm{i} k/d) e^{2\pi \mathrm{i} k t/d}. \tag{18}$$

Theorem

Suppose that $p \in (0,1)$. *Then, as* $n \to \infty$,

$$\frac{D_n - H^{-1} \ln n}{\sqrt{\ln n}} \xrightarrow{\mathrm{d}} N\Big(0, \frac{\sigma^2}{H^3}\Big),$$

with $\sigma^2 = H_2 - H^2 = pq(\ln p - \ln q)^2$. *If* $p \neq 1/2$, *then* $\sigma^2 > 0$ *and this can be written as*

$$D_n \sim \mathrm{AsN}\big(H^{-1} \ln n, H^{-3}\sigma^2 \ln n\big).$$

Moreover,

$$\operatorname{Var} D_n = \frac{\sigma^2}{H^3} \ln n + o(\ln n).$$

Imbalance in tries

Let Δ_n be the imbalance of a string in a trie, defined as the number of steps to the right minus the number of steps to the left in the path from the root to the leaf where the string is stored.
We define

$$Y_i := 2\xi_i - 1 = \begin{cases} -1, & \xi_1 = 0, \\ +1, & \xi_1 = 1, \end{cases}$$

and denote the corresponding partial sums by $V_k := \sum_{i=1}^{k} Y_i$.
Thus $\Delta_n = V_{D_n}$, with D_n as above. We have

$$(D_n, \Delta_n) = (D_n, V_{D_n}) \stackrel{\mathrm{d}}{=} \bigl(\widehat{\nu}(\ln n), V_{\widehat{\nu}(\ln n)}\bigr).$$

In particular,

$$\Delta_n \stackrel{\mathrm{d}}{=} V_{\widehat{\nu}(\ln n)}.$$

A general renewal theory theorem applies and yields

Theorem (Mahmoud)

As $n \to \infty$,

$$\Delta_n \sim \mathrm{AsN}\left(\frac{p-q}{H}\ln n,\ \frac{pq\ln^2(pq)}{H^3}\ln n\right).$$

Random networks

A *random network* is a network where nodes or edges or both are created by some random procedure.
First example: (classical random graphs studied by Erdős and Rényi and many others from 1959 and until today)

Fix two (large) numbers n (number of nodes) and m (number of edges). Number the nodes $1, \dots, n$. Draw two nodes at random and join them by an edge. Repeat m times. Denoted $G(n, m)$.

Random networks

A *random network* is a network where nodes or edges or both are created by some random procedure.
First example: (classical random graphs studied by Erdős and Rényi and many others from 1959 and until today)

Fix two (large) numbers n (number of nodes) and m (number of edges). Number the nodes $1, \dots, n$. Draw two nodes at random and join them by an edge. Repeat m times. Denoted $G(n, m)$.

A variant: Fix n (number of nodes) and a probability p. For each pair of nodes, make a random choice and connect the nodes by an edge with probability p. (Toss a biased coin, throw dice, get a random number, or use some other random procedure.)
Denoted $G(n, p)$.

Random networks

A *random network* is a network where nodes or edges or both are created by some random procedure.
First example: (classical random graphs studied by Erdős and Rényi and many others from 1959 and until today)

Fix two (large) numbers n (number of nodes) and m (number of edges). Number the nodes $1, \dots, n$. Draw two nodes at random and join them by an edge. Repeat m times. Denoted $G(n, m)$.

A variant: Fix n (number of nodes) and a probability p. For each pair of nodes, make a random choice and connect the nodes by an edge with probability p. (Toss a biased coin, throw dice, get a random number, or use some other random procedure.)
Denoted $G(n, p)$.

Examples of other, newer, random networks will follow.

Node degrees

The *degree* of a node is the number of links connecting the node to other nodes.
For Internet, as well as for many other graphs in various applications, it is evident that there is a large dispersion of the degrees for different nodes.

Examples (incoming links according to Google):

www.google.com: 649000
www.bath.ac.uk: 2180
www.math.uu.se/∼svante/papers: 2

Power laws

The classical random graphs have node degrees that are random, but with a rather small random dispersion and very small probability of having a degree that is much larger than the average. More precisely, the degree distribution is Hypergeometric or Binomial and asymptotically Poisson. Hence the distribution has exponential tails.

Many graphs from "reality" seem to have node degrees that are distributed according to a *power law*, i.e., there are constants γ and C_1 such that

$$\text{number of nodes with degree } k \approx C_1 k^{-\gamma-1}$$

or, which is roughly equivalent, with another constant C_2,

$$\text{number of nodes with degree at least } k \approx C_2 k^{-\gamma}.$$

Since the graphs are finite (although large), this can of course hold only in some (large) range and not for all k.
Graphs with a power law are often called *scale-free*.

Many graphs from "reality" seem to have node degrees that are distributed according to a *power law*, i.e., there are constants γ and C_1 such that

$$\text{number of nodes with degree } k \approx C_1 k^{-\gamma-1}$$

or, which is roughly equivalent, with another constant C_2,

$$\text{number of nodes with degree at least } k \approx C_2 k^{-\gamma}.$$

Since the graphs are finite (although large), this can of course hold only in some (large) range and not for all k.
Graphs with a power law are often called *scale-free*.

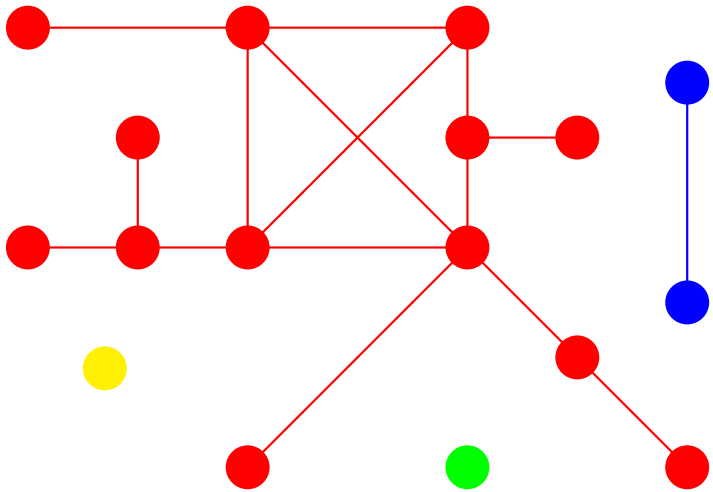
It has during the last decade been popular to study large graphs "in real life" and find such power laws for them, often with a value of γ between 1 and 2. An important example is the Internet.

Some other examples:

- references between scientific papers in the data base ISI (783339 articles with 6716198 links) and in *Physical Review D* (24296 papers)
- sexual contacts (2810 Swedes)
- collaborations (joint publications) between scientists (1520251 i Medline, and others)
- metabolic reactions in *E. coli* and other organisms (778 substances)
- interactions between proteins in yeast (*S. cerevisiae*) (1870 proteins)
- telephone calls in a given day (47000000 telephones)

Components

A *component* in a network is a connected part of the network.
Example:



This network has 4 components.

A typical case is that there is a giant component containing a large part of all nodes, together with many very small components with only one or a few nodes each.

Another case, typical for very sparse graphs, is that there are many small components but no really big one.

It is also interesting to study what happens if some of the nodes or edges in a graph are deleted. (Percolation.) What components are there in the graph that is left, and how large are they?

For example, this is important when studying the vulnerability of the Internet to technical failures or terrorist attacks: Will there still be a giant component (= only local disturbances) or are there only small components left (= the system collapses)?
Another, related, example is the spread of an infectious disease like Swine flu in a human population. In this case, the objective is the opposite: one wants to limit the spread and wants graphs with small components only.

Theorem (Erdős and Rényi)

A classical random graph $G(n,m)$ with n nodes and m edges has a giant component if $m > n/2$ but not otherwise.

More formally: If $n \to \infty$ and $m \sim cn$ for some constant c, and C_1 is the largest component of the random graph, then

$$\frac{|C_1|}{n} \xrightarrow{\mathrm{p}} \begin{cases} 0 & \text{if } c \le 1/2, \\ \rho(2c) > 0 & \text{if } c > 1/2. \end{cases}$$

If $c < 1/2$, then $|C_1| = O_p(\log n)$.
The same holds for $G(n,p)$ with $p \sim c'/n$, with $c' = 2c$ so the threshold is $c' = 1$, i.e. $p = 1/n$.

$$\rho(\lambda) = 1 - e^{-\lambda\rho(\lambda)}.$$

Susceptibility

The *susceptibility* or *mean cluster size* $\chi(G)$ is the expected size of the component containing a random node. Equivalently, it is n times the probability that two random nodes lie in the same component (and thus may be connected by a path in the graph). If the components are $C_1, C_2, \dots$, then

$$\chi(G) = \frac{\sum_i |C_i|^2}{n}.$$

Susceptibility

The *susceptibility* or *mean cluster size* $\chi(G)$ is the expected size of the component containing a random node. Equivalently, it is n times the probability that two random nodes lie in the same component (and thus may be connected by a path in the graph). If the components are $C_1, C_2, \dots$, then

$$\chi(G) = \frac{\sum_i |C_i|^2}{n}.$$

Theorem

For $G(n,p)$, as $n \to \infty$:

$$\chi(G(n,p)) \sim_p \begin{cases} \frac{1}{1-np}, & 1-np \gg n^{-1/3} \\ n\rho(np)^2, & np-1 \gg n^{-1/3}. \end{cases}$$

Distances and diameter

Given that two nodes are in the same component, we may ask for the *distance* between them, i.e., the shortest path between them in the graph. The maximum distance is the *diameter*. The *average distance* between two random nodes is often at least as interesting.

Distances and diameter

Given that two nodes are in the same component, we may ask for the *distance* between them, i.e., the shortest path between them in the graph. The maximum distance is the *diameter*. The *average distance* between two random nodes is often at least as interesting.

In many graphs, the diameter and average distance are of the order $\log n$, and thus quite small even when the number n of nodes is large. This phenomenon is often called *Small Worlds*.