Combinatorics, Probability and Computing (1997) 6, 165–178
Copyright © 1997 Cambridge University Press

Markov Processes Involving q-Stirling Numbers

D. CRIPPA, K. SIMON and P. TRUNZ

Institute for Theoretical Computer Science, ETH–Zurich,
CH–8092 Zurich, Switzerland
(e-mail: simon@inf.ethz.ch)

Received 3 January 1995; revised 5 June 1996

In this paper we consider the Markov process defined by

$$\mathrm{P}_{1,1} = 1, \qquad \mathrm{P}_{n,\ell} = (1-\lambda_{n,\ell}) \cdot \mathrm{P}_{n-1,\ell} + \lambda_{n,\ell-1} \cdot \mathrm{P}_{n-1,\ell-1}$$

for transition probabilities $\lambda_{n,\ell} = q^{\ell}$ and $\lambda_{n,\ell} = q^{n-1}$. We give closed forms for the distributions and the moments of the underlying random variables. Thereby we observe that the distributions can be easily described in terms of q-Stirling numbers of the second kind. Their occurrence in a purely time dependent Markov process allows a natural approximation for these numbers through the normal distribution. We also show that these Markov processes describe some parameters related to the study of random graphs as well as to the analysis of algorithms.

1. Introduction

In this paper we shall study the Markov process defined by

$$\begin{aligned} \mathrm{P}_{1,1} &= 1 \\ \mathrm{P}_{n,\ell} &= 0 \text{ for } \ell \notin \{1,\ldots,n\} \text{ and} \\ \mathrm{P}_{n,\ell} &= (1-\lambda_{n,\ell})\,\mathrm{P}_{n-1,\ell} + \lambda_{n,\ell-1}\,\mathrm{P}_{n-1,\ell-1} \text{ otherwise,} \end{aligned} \tag{1}$$

where $\mathrm{P}_{n,\ell}$ is the probability distribution of the underlying random variable at discrete time n and $\lambda_{n,\ell}$ is the transition probability from state ℓ to state $\ell+1$ at time n. In particular, we shall consider some special choices of transition probabilities for which the random variable described by the process assumes a concrete interpretation.

The motivation behind this work arises from the average case analysis of combinatorial algorithms, in particular the Simon algorithm [12, 1] for computing the transitive closure on acyclic digraphs and the so-called 'Approximate Counting Algorithm' (see Flajolet [3] and Morris [8]), a method for 'keeping approximate counts of large numbers in small

https://doi.org/10.1017/S0963548397002964 Published online by Cambridge University Press

registers'. In Simon [12], two variants of the above process have been analysed, namely:

$$\lambda_{n,\ell} = 1 - q^{\ell} \quad \text{and} \quad \lambda_{n,\ell} = q^{\ell}.$$

For the first transition probability, the underlying random variable models the size of the transitive closure of a node in the random acyclic digraph $G_{n,p}$. Closed forms for its distribution and moments involving q-hypergeometric series are given in Simon, Crippa and Collenberg [11]. The second one describes the width of a greedy constructed decomposition of $G_{n,p}$ into node disjoint paths [12], as well as the value of the counting register in the 'Approximate Counting Algorithm'. Flajolet [3] showed that the distribution of this random variable is given by

$$\mathrm{P}_{n,\ell} = \sum_{j=0}^{\ell-1} (-1)^j \frac{q^{\binom{j}{2}} (1-q^{\ell-j})^{n-1}}{(q)_j \, (q)_{\ell-j-1}}.$$

After having introduced some definitions in Section 2, in Section 3.1 we shall simplify this expression to

$$\mathrm{P}_{n,\ell} = (1-q)^{n-\ell} \, q^{\binom{\ell}{2}} \, S_2(\ell, n-\ell),$$

where $S_2(n,h)$ denotes the *q-Stirling number of the second kind* (see Gould [5]), and we shall present closed forms for the first two moments, results which are missing in [12] and [3]. As a byproduct we shall observe that the same process also models the size of a greedy clique and the size of a greedy stable set.

In Section 3.2, then, we shall deal with another distribution describing the random variable counting number of sources (or sinks) in a random acyclic digraph. In this case the transition probability shall be given by

$$\lambda_{n,\ell} = q^{n-1},$$

which shall lead to

$$\mathrm{P}_{n,\ell} = (1-q)^{n-k} \, q^{\binom{n}{2}} \, S_2(-n, n-k). \tag{2}$$

Note that in this case the transition probability is dependent only on the time n. This leads to a simple approximation of the discrete time process, leading to the well-known backward equation of Kolmogorov (see Feller [2, Volume II, Diffusion Processes] and Gnedenko [4]):

$$\frac{\partial f(n,\ell)}{\partial n} = -a(n,\ell) \frac{\partial f(n,\ell)}{\partial \ell} + \frac{1}{2} b(n,\ell) \frac{\partial^2 f(n,\ell)}{\partial \ell^2}. \tag{3}$$

In our case we have

$$a(n,\ell) = a(n) \quad \text{and} \quad b(n,\ell) = b(n),$$

condition which allows us to reduce (3) to

$$\frac{\partial f(n,\ell)}{\partial n} = \frac{1}{2} \frac{\partial^2 f(n,l)}{\partial \ell^2},$$

whose solution is given by the normal distribution. This approximation, together with (2), results in an approximation for the q-Stirling numbers of the second kind.

https://doi.org/10.1017/S0963548397002964 Published online by Cambridge University Press

2. Some definitions

In the following let $G = (V, E)$ be a graph with a set of vertices V and a set of edges $E \subseteq V \times V$. Without loss of generality, we shall assume that the graph does not contain loops (i.e. edges with two equal endpoints). In later sections we shall present some interpretations of the Markov process (1) related to random graphs (for more details on random graphs see Palmer [9]). Let us therefore introduce the following model:

Definition. The probability space of graphs $G_{n,p}$ is defined as the set of graphs with vertex set $V = \{1, \ldots, n\}$ and edge set

$E = \{ \{i, j\} \mid i, j \in V, i \neq j \}$ for a **undirected**,
$E = \{ (i, j) \mid i, j \in V, i \neq j \}$ for a **directed** and
$E = \{ (i, j) \mid i, j \in V, i < j \}$ for a **directed, acyclic graph**,

where every edge occurs independently with probability $p \in (0, 1)$. In the following we shall use the convention of writing q instead of $1 - p$.

Throughout this paper we shall deal with the so-called *q-notation*, a pure number theoretical notation which seems to be the appropriate tool to deal with our kind of problems (for more details see Gasper and Rahman [6], Hardy and Wright [7], and Gould [5]). Let us therefore introduce the following definitions, whereby we assume that q is different from 1.

- The *q-shifted factorial* is defined for any integer n by

$$(A; q)_0 \equiv (A)_0 = 1, \tag{4}$$

$$(A; q)_n \equiv (A)_n = \prod_{j=0}^{n-1} (1 - A\, q^j). \tag{5}$$

Whenever no misunderstanding can arise, in the q-formulae we shall drop the q.

- The *q-natural numbers* are defined for any integer n by

$$[n]_q \equiv [n] = \frac{1 - q^n}{1 - q}. \tag{6}$$

- The *q-Stirling numbers of the first kind* are defined for integers n, k as the sum of the $\binom{n}{k}$ products with k different factors which may be formed from the first n q-natural numbers $[1], [2], \ldots, [n]$ (for $k = 0$ we shall define them as 1). We shall follow the notation of Gould [5] and denote them by $S_1(n, k; q) \equiv S_1(n, k)$. As an illustration, let us compute a few examples:

$$\begin{aligned} S_1(3, 0) &= [1] = 1 \\ S_1(3, 1) &= [1] + [2] + [3] = 3 + 2\, q + q^2 \\ S_1(3, 2) &= [1]\,[2] + [1]\,[3] + [2]\,[3] = 3 + 4\, q + 3\, q^2 + q^3 \end{aligned}$$

- Similarly, the *q-Stirling numbers of the second kind*, denoted by $S_2(n, k; q) \equiv S_2(n, k)$, are defined for integers n, k as the sum of the $\binom{n+k-1}{k}$ products, each with k factors

https://doi.org/10.1017/S0963548397002964 Published online by Cambridge University Press

(repeated factors allowed), which may be formed from the first n q-natural numbers $[1], [2], \ldots, [n]$ (for $k = 0$ they shall also be defined as 1). Again let us compute a few examples:

$$\begin{aligned} S_2(2,0) &= [1] = 1 \\ S_2(2,1) &= [1] + [2] = 2 + q \\ S_2(2,2) &= [1]\,[1] + [1]\,[2] + [2]\,[2] = 3 + 3q + q^2 \\ S_2(2,3) &= [1]\,[1]\,[1] + [1]\,[1]\,[2] + [1]\,[2]\,[2] + [2]\,[2]\,[2] = 4 + 6q + 4q^2 + q^4 \end{aligned}$$

While these definitions are quite simple, they soon become useless because of the large number of terms involved. Let us therefore present recurrences which are more practical in many circumstances:

$$S_1(n,k) = S_1(n-1,k) + [n]\,S_1(n-1,k-1) \tag{7}$$

$$S_2(n,k) = S_2(n-1,k) + [n]\,S_2(n,k-1) \tag{8}$$

The relationship between q-Stirling numbers of the first kind and q-Stirling numbers of the second kind is given by

$$S_1(-n-1,k;1/q) = q^k\,S_2(n,k;q) \tag{9}$$

$$S_2(-n-1,k;1/q) = q^k\,S_1(n,k;q) \tag{10}$$

• Finally, we shall define the *broken Lambert series* as

$$L_n(q) \stackrel{\text{def}}{=} \sum_{\ell=1}^{n} \frac{q^\ell}{1-q^\ell}; \tag{11}$$

notice that the usual *Lambert Series*, $L_\infty(q)$, corresponds to the generating function of the divisor-function.

3. The Markov process

In this section we shall consider the Markov process (1) for two particular transition probabilities, namely

$$\lambda_{n,\ell} = q^\ell \quad \text{and} \quad \lambda_{n,\ell} = q^{n-1}.$$

The process with the former one defines a random variable whose distribution was provided in closed form by Flajolet [3]. In Section 3.1 we shall simplify this result to an expression involving q-Stirling numbers, and provide closed forms for the first two moments. In Section 3.2 we shall then consider the Markov process with the second transition probability. A graph theoretical interpretation for the underlying random variable shall be given together with closed forms for its distribution and moments.

3.1. The first transition probability

A possible graph theoretical interpretation for the random variable defined by the process (1) with transition probability $\lambda_{n,\ell} = q^\ell$ is given by the so-called *chain decomposition* of a graph.

https://doi.org/10.1017/S0963548397002964 Published online by Cambridge University Press

Definition. For an acyclic digraph G we inductively define the chain decomposition as a partition $\mathscr{Z} = Z_1 + \cdots + Z_k$ of V satisfying:

(i) For $n = 1$ the partition $\mathscr{Z}$ contains only the path $Z_1 = 1$.

(ii) Let $n \geq 2$ and $\mathscr{Z} = Z_1 + \cdots + Z_\ell$ be the partition of G not containing the node n. Then either one of the paths $Z_i, 1 \leq i \leq \ell$, can be extended to n or there is no Z_i for which there is an edge between the last vertex of Z_i and n. In the first case we extend the Z_i with smallest suffix to n and in the second one a new path $Z_{\ell+1}$ containing only the node n is added to $\mathscr{Z}$.

We say then that G has a decomposition of *width* k. We shall denote by $\mathscr{K}_n$ the width of the chain decomposition of a graph belonging to the $G_{n,p}$ class; further we shall write $\mathrm{K}_{n,\ell}$ for the probability of having $\mathscr{K}_n = \ell$.

It is now easy to show that $\mathrm{K}_{n,\ell}$ satisfies the following recursion:

$$\mathrm{K}_{1,1} = 1 \quad \text{and} \quad \mathrm{K}_{n,\ell} = (1 - q^\ell)\,\mathrm{K}_{n-1,\ell} + q^{\ell-1}\,\mathrm{K}_{n-1,\ell-1}. \tag{12}$$

Consider the process in which we build a $G_{n,p}$ graph G_n from a graph G_{n-1} through addition of the node n; then in G_n we shall have ℓ chains if in G_{n-1} either there are already ℓ chains and the node n is connected to at least one of the chain ends, or there are only $\ell - 1$ chains and n is not connected to any of their ends. The corresponding probabilities are

$$\mathrm{K}_{n-1,\ell}\,(1 - q^\ell) \quad \text{and} \quad \mathrm{K}_{n-1,\ell-1}\,q^{\ell-1},$$

and thus $\mathrm{K}_{n,\ell}$ satisfies the recursion (12).

In Flajolet [3] the following closed form for the distribution $\mathrm{K}_{n,\ell}$ was found by analytical techniques:

$$\mathrm{K}_{n,\ell} = \sum_{j=0}^{\ell-1} (-1)^j \frac{q^{\binom{j}{2}}\,(1 - q^{\ell-j})^{n-1}}{(q)_j\,(q)_{\ell-j-1}}. \tag{13}$$

In our first result we give a closed form for this expression.

Theorem 3.1. *The probability* $\mathrm{K}_{n,\ell}$ *that in* $G_{n,p}$ *the chain decomposition has width* ℓ *is given by*

$$\mathrm{K}_{n,\ell} = (1 - q)^{n-\ell}\,q^{\binom{\ell}{2}}\,S_2(\ell, n - \ell). \tag{14}$$

Proof. We shall prove (14) by induction on n. Considering first that $\mathrm{K}_{1,1} = 1$, the case $n = 1$ is established. With (12) we find further for $n \geq 2$:

$$\begin{aligned}
\mathrm{K}_{n,\ell} &= (1 - q^\ell)\,\mathrm{K}_{n-1,\ell} + q^{\ell-1}\,\mathrm{K}_{n-1,\ell-1} \\
&\overset{\text{I.H.}}{=} (1 - q^\ell)\,(1 - q)^{n-\ell-1}\,q^{\binom{\ell}{2}}\,S_2(l, n - \ell - 1) \\
&\quad + q^{\ell-1}\,(1 - q)^{n-\ell}\,q^{\binom{\ell-1}{2}}\,S_2(\ell - 1, n - \ell) \\
&= (1 - q)^{n-\ell}\,q^{\binom{\ell}{2}}\,\underbrace{\left([\ell]\,S_2(\ell, n - \ell - 1) + S_2(\ell - 1, n - \ell)\right)}_{\overset{(8)}{=} S_2(\ell, n-\ell)}.
\end{aligned}$$

□

https://doi.org/10.1017/S0963548397002964 Published online by Cambridge University Press

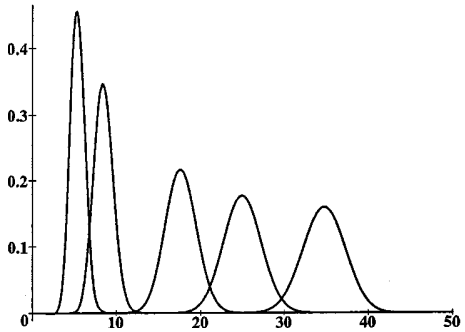


Figure 1 The evolving distribution of the number of chains in a random acyclic digraph with 50 vertices. The values for q are 0.5, 0.7, 0.9, 0.95, 0.98 (from left to right).

In Figure 1 we have plotted the evolving distribution for different values of q. We clearly see that expectation and variance both rise with increasing values of q.

Next we shall find closed forms for the first two moments of $\mathscr{K}_n$. Let us start by considering the probability generating function of $\mathscr{K}_n$.

Lemma 3.1. *The probability generating function*

$$g_n(z) \stackrel{\text{def}}{=} \sum_{\ell=1}^{n} z^{\ell}\, \mathrm{K}_{n,\ell}$$

of the random variable $\mathscr{K}_n$ has the following closed form:

$$g_n(z) = \sum_{\ell=0}^{n-1} \binom{n-1}{\ell} (-q)^{\ell}\, z\,(z)_{\ell}. \tag{15}$$

Proof. Let us define

$$G(t,y) = \sum_{\substack{n\geq 1\\ \ell\geq 0}} \mathrm{K}_{n,\ell+1}\, t^n\, y^{\ell}. \tag{16}$$

Then we have

$$\begin{aligned} G(t,y) &= t + \sum_{\substack{n\geq 2\\ \ell\geq 0}} \mathrm{K}_{n,\ell+1}\, t^n\, y^{\ell} \\ &= t + \sum_{\substack{n\geq 1\\ \ell\geq 0}} \mathrm{K}_{n+1,\ell+1}\, t^{n+1}\, y^{\ell} \\ &\stackrel{(12)}{=} t + \sum_{\substack{n\geq 1\\ \ell\geq 0}} \left((1-q^{\ell+1})\,\mathrm{K}_{n,\ell+1} + q^{\ell}\,\mathrm{K}_{n,\ell}\right) t^{n+1}\, y^{\ell} \\ &= t + t\,G(t,y) - t\,q\,G(t,y\,q) + y\,q\,t\,G(t,y\,q), \end{aligned}$$

so that

$$G(t,y) = \frac{t}{1-t} - \frac{t\,q\,(1-q)}{1-t}\,G(t,y\,q).$$

https://doi.org/10.1017/S0963548397002964 Published online by Cambridge University Press

Iteration finally yields

$$G(t,y)=\sum_{n\geq 0}\frac{(-1)^n\,t^{n+1}\,q^n\,(y)_n}{(1-t)^{n+1}}. \tag{17}$$

Accordingly we get

$$\begin{aligned}
g_n(z) &= \sum_{\ell\geq 1} z^\ell\,\mathrm{K}_{n,\ell}\\
&= z\sum_{\ell\geq 1} z^\ell\,\mathrm{K}_{n,\ell+1}\\
&= z\,G(t,z)\Big|_{t^n}\\
&= z\sum_{\ell\geq 0}\frac{(-1)^\ell\,t^{\ell+1}\,q^\ell\,(z)_\ell}{(1-t)^{\ell+1}}\Bigg|_{t^n}\\
&= z\sum_{\ell\geq 0}\sum_{h\geq 0}(-1)^\ell\,t^{\ell+h+1}\binom{\ell+h}{h}\,q^\ell\,(z)_\ell\Bigg|_{t^n}\\
&= z\sum_{\ell\geq 0}(-1)^\ell\binom{n-1}{\ell}\,q^\ell\,(z)_\ell,
\end{aligned}$$

where $P(t)|_{t^n}$ denotes the coefficient of t^n in the polynomial $P(t)$. □

From (15) we can easily derive the expectation and the second moment of $\mathscr{K}_n$.

Theorem 3.2. *The expected value of $\mathscr{K}_n$ is given by*

$$\mathbf{E}\left[\mathscr{K}_n\right]=1+\sum_{\ell=1}^{n-1}\binom{n-1}{\ell}(-1)^{\ell-1}\,q^\ell\,(q)_{\ell-1} \tag{18}$$

and the second moment by

$$\mathbf{E}\left[(\mathscr{K}_n)^2\right]=\mathbf{E}\left[\mathscr{K}_n\right]+2\sum_{\ell=1}^{n-1}(-1)^{\ell+1}\,q^\ell\,(q)_{\ell-1}\binom{n-1}{\ell}\left(1-L_{\ell-1}(q)\right). \tag{19}$$

Proof. To prove this theorem we notice that

$$\left[\frac{d}{dz}\,g_n(z)\right]_{z=1}=\left[\sum_{\ell=1}^{n}\ell\,z^{\ell-1}\mathrm{K}_{n,\ell}\right]_{z=1}=\mathbf{E}\left[\mathscr{K}_n\right]$$

and

$$\left[\frac{d^2}{dz^2}\,g_n(z)\right]_{z=1}=\left[\sum_{\ell=1}^{n}\ell\,(\ell-1)\,z^{\ell-2}\mathrm{K}_{n,\ell}\right]_{z=1}=\mathbf{E}\left[(\mathscr{K}_n)^2\right]-\mathbf{E}\left[\mathscr{K}_n\right].$$

In $g_n(z)$ the only term containing z is $z\,(z)_\ell$ so that we only need to determine the first two derivatives of this term. For $\ell=0$ we have

$$\left[\frac{d}{dz}\,z\,(z)_0\right]_{z=1}=\left[\frac{d}{dz}\,z\right]_{z=1}=1$$

https://doi.org/10.1017/S0963548397002964 Published online by Cambridge University Press

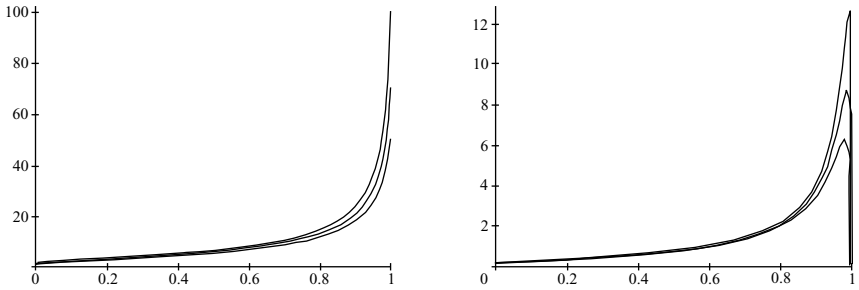


Figure 2 The evolving expectation (left) and variance for the number of chains.

and for $\ell \geq 1$ we find

$$\left[\frac{d}{dz}\, z\,(z)_\ell\right]_{z=1} = \left[\frac{d}{dz}\, z\,(1-z)(zq)_{\ell-1}\right]_{z=1} = (q)_{\ell-1}.$$

On the other hand we observe for the second derivative

$$\begin{aligned}
\left[\frac{d^2}{dz^2}\, z\,(z)_\ell\right]_{z=1} &= \left[\frac{d^2}{dz^2}\, z\,(1-z)\,(zq)_{\ell-1}\right]_{z=1} \\
&= \left[\frac{d}{dz}\left((1-2z)\,(zq)_{\ell-1} + z\,(1-z)\,\frac{d}{dz}\,(zq)_{\ell-1}\right)\right]_{z=1} \\
&= -2\,(q)_{\ell-1} - 2\underbrace{\left[z\,\frac{d}{dz}\,(zq)_{\ell-1}\right]_{z=1}}_{-(q)_{\ell-1}\sum_{j=1}^{\ell-1}\frac{q^j}{1-q^j}} \\
&= -(q)_{\ell-1}\,(2-2\,L_{\ell-1}(q))\,.
\end{aligned}$$

Expressions (18) and (19) follow directly therefrom. □

Note that we can model the number of vertices in a greedy constructed stable set using the same probability process. Instead of dealing with the end vertices of a chain we look at the vertices in the stable set. Further, if we substitute q with p we can also model the number of vertices in a greedy constructed maximal clique using a dual argument.

In Figure 2 we have plotted the evolution of the expectation and the variance of $\mathscr{K}_n$ for a random graph with 50 (lowest curves), 70 and 100 nodes.

3.2. The second transition probability

In this section we shall consider the Markov process (1) with the transition probability $\lambda_{n,\ell} = q^{n-1}$. It turns out that the random variable modeled by this process has another graph theoretical interpretation, namely the number of *sources* in a random acyclic digraph.

Definition. A node v in a digraph G is called a *source* if it does not have any predecessor. We shall denote by $\mathscr{S}_n$ the number of sources in a random acyclic digraph $G_{n,p}$.

https://doi.org/10.1017/S0963548397002964 Published online by Cambridge University Press

Furthermore, we denote by $\mathrm{S}_{n,\ell}$ the probability of having exactly ℓ sources in such a graph.

Again it is not difficult to show that $\mathrm{S}_{n,\ell}$ satisfies the Markov process (1) with transition probability

$$\lambda_{n,\ell} = q^{n-1}. \tag{20}$$

Consider in fact the process in which we build a graph G_n from G_{n-1} through addition of the node n. The graph G_n can have exactly ℓ sources if in G_{n-1} either there exist only $\ell - 1$ sources and no node of G_{n-1} is connected to n, or there are already ℓ sources and there is at least a node in G_{n-1} connected to n. The corresponding probabilities are

$$\mathrm{S}_{n-1,\ell-1}\, q^{n-1} \quad \text{and} \quad \mathrm{S}_{n-1,\ell}\,(1 - q^{n-1}),$$

proving therefore what we have asserted.

At this point we can show the following result.

Theorem 3.3. *Let $G = G_{n,p}$ be a random acyclic digraph. Then the random variable $\mathscr{S}_n$ counting the number of sources in G has the following distribution, expectation and variance:*

$$\mathrm{S}_{n,\ell} = (1-q)^{n-\ell}\, q^{\binom{n}{2}}\, S_2(-n, n-\ell), \tag{21}$$

$$\mathbf{E}\left[\mathscr{S}_n\right] = [n] \tag{22}$$

and

$$\mathbf{Var}\left[\mathscr{S}_n\right] = [n]_q - [n]_{q^2}. \tag{23}$$

Proof. We first prove (21) by induction on n. For $n = 1$ the equation is satisfied, so let us assume $n \geq 2$. Using (1) and (20) we get

$$\begin{aligned}
\mathrm{S}_{n,\ell} &= (1-q^{n-1})\,\mathrm{S}_{n-1,\ell} + q^{n-1}\,\mathrm{S}_{n-1,\ell-1} \\
&\overset{\text{I.H.}}{=} (1-q^{n-1})\,(1-q)^{n-\ell-1}\, q^{\binom{n-1}{2}}\, S_2(-n+1, n-\ell-1) \\
&\quad + q^{n-1}\,(1-q)^{n-\ell}\, q^{\binom{n-1}{2}}\, S_2(-n+1, n-\ell) \\
&= (1-q)^{n-\ell}\, q^{\binom{n}{2}} \left(\frac{[n-1]}{q^{n-1}}\, S_2(-n+1, n-\ell-1) + S_2(-n+1, n-\ell) \right) \\
&\overset{(8)}{=} (1-q)^{n-\ell}\, q^{\binom{n}{2}} \left(\frac{[n-1]}{q^{n-1}}\, S_2(-n+1, n-\ell-1) \right. \\
&\quad \left. + S_2(-n, n-\ell) + [-n+1]\, S_2(-n+1, n-\ell-1) \right) \\
&= (1-q)^{n-\ell}\, q^{\binom{n}{2}} \left(S_2(-n, n-\ell) \right. \\
&\quad \left. + \underbrace{\left(\frac{[n-1]}{q^{n-1}} + [-n+1] \right)}_{0}\, S_2(-n+1, n-\ell-1) \right).
\end{aligned}$$

https://doi.org/10.1017/S0963548397002964 Published online by Cambridge University Press

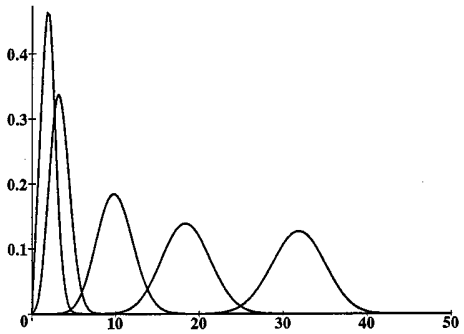


Figure 3 The evolving distribution of the number of sources in a random acyclic digraph with 50 vertices. The values for q are 0.5, 0.7, 0.9, 0.95, 0.98 (from left to right).

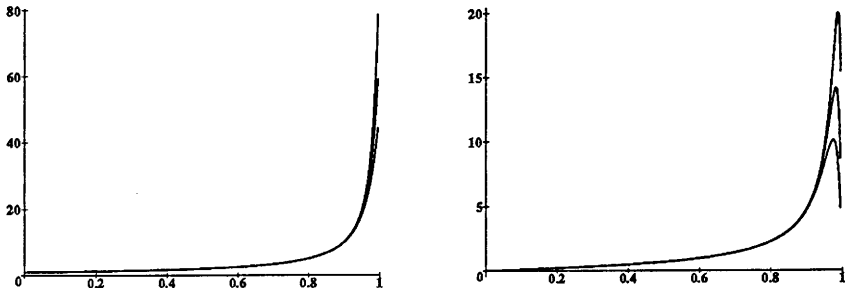


Figure 4 The evolving expectation (left) and variance for the number of sources.

Let us now define the random variables $\mathscr{V}_i, 1 \le i \le n$, taking value 1 if node i is a source and 0 otherwise. $\mathscr{V}_i$ therefore has a Bernoulli distribution with probability q^{i-1}, namely the probability that node i is a source in G. Accordingly, we get

$$\mathbf{E}\left[\mathscr{V}_i\right] = q^{i-1} \quad \text{and} \quad \mathbf{Var}\left[\mathscr{V}_i\right] = q^{i-1}(1-q^{i-1}).$$

This leads immediately to

$$\mathbf{E}\left[\mathscr{S}_n\right] = \mathbf{E}\left[\sum_{i=1}^{n} \mathscr{V}_i\right] = \sum_{i=1}^{n} \mathbf{E}\left[\mathscr{V}_i\right] = \sum_{i=1}^{n} q^{i-1} = \frac{1-q^n}{1-q} = [n].$$

On the other hand, the $\mathscr{V}_i$'s are pairwise independent random variables, and hence we get

$$\mathbf{Var}\left[\mathscr{S}_n\right] = \sum_{i=1}^{n} \mathbf{Var}\left[\mathscr{V}_i\right] = \sum_{i=1}^{n} q^{i-1}(1-q^{i-1}) = \frac{1-q^n}{1-q} - \frac{1-(q^2)^n}{1-q^2} = [n]_q - [n]_{q^2}. \quad \square$$

We have plotted in Figure 3 the evolution of the distribution representing the number of sources in a random acyclic digraph with 50 vertices and q going from 0.5 to 0.98.

In Figure 4 we have plotted the evolution of the expectation and the variance of $\mathscr{S}_n$ for a random graph with 50 (lowest curves), 70 and 100 nodes.

https://doi.org/10.1017/S0963548397002964 Published online by Cambridge University Press

In this section we have presented a distribution which describes the number of sources in a random directed acyclic graph. Note that for symmetry reasons the same process can also be used to describe the number of sinks in a random acyclic digraph. This holds even though the process which describes the number of sinks is completely different from the process describing the number of sources.

4. Approximations

In this section we shall consider the approximation of the discrete distribution $\mathrm{S}_{n,\ell}$ by a function $f(n,\ell)$, continuous in both variables. The classical example for this kind of approach is the binomial distribution $BIN(n,p)$,

$$\mathrm{B}_{n,\ell} = \binom{n}{\ell} p^{\ell} (1-p)^{n-\ell}, \tag{24}$$

which can be approximated by the normal distribution $N(\mu,\sigma^2)$

$$\mathrm{N}_{n,\ell} = \frac{1}{\sqrt{2\,\pi\,\sigma^2}} \exp\left(\frac{\ell-\mu}{\sigma^2}\right) \tag{25}$$

with expectation $\mu = np$ and variance $\sigma^2 = p\,(1-p)\,n$. A simple approach leading to these results comes from the interpretation of the normal distribution as a diffusion process. One dimensional diffusion processes can by described by the partial differential equation

$$\frac{\partial f(x,t)}{\partial t} = -a(x,t)\,\frac{\partial f(x,t)}{\partial x} + \frac{1}{2}\,b(x,t)\,\frac{\partial^2 f(x,t)}{\partial x^2}, \tag{26}$$

known as the Kolmogorov backward equation (see Feller [2]).

Let $\mathscr{Y}(t)$ be a random variable describing the position of a particle at time t and let $f(x,t)$ be the probability function of $\mathscr{Y}(t)$. Further let the random variable $\mathscr{Y}_{x,t}(\tau)$ describe the position of the particle at time $\tau \geq t$ on the condition that $\mathscr{Y}(t) = x$. Then $f(x,t)$ satisfies equation (26) and the coefficients $a(x,t)$ and $b(x,t)$ are given by

$$a(x,t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t}\,\mathbf{E}\left[\,\mathscr{Y}_{x,t}(t+\Delta t) - \mathscr{Y}_{x,t}(t)\,\right] \tag{27}$$

$$b(x,t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t}\,\mathbf{E}\left[\,(\mathscr{Y}_{x,t}(t+\Delta t) - \mathscr{Y}_{x,t}(t))^2\,\right]. \tag{28}$$

From a physical point of view $a(x,t)$ represents the mean speed value of $\mathscr{Y}(t)$ at position x and $b(x,t)$ is proportional to the mean kinetic energy of the system.

The simplest nontrivial case in which we can solve equation (26) is given by

$$a(x,t) = 0 \qquad \text{and} \qquad b(x,t) = 1,$$

reducing the Kolmogorov equation to

$$\frac{\partial f(x,t)}{\partial t} = -\frac{1}{2}\,\frac{\partial^2 f(x,t)}{\partial x^2}. \tag{29}$$

This process describes the Brownian motion and in this case the solution of (29) is the normal distribution $N(0,1)$.

The connection between a Markov process of type (1) and a diffusion process arises from

https://doi.org/10.1017/S0963548397002964 Published online by Cambridge University Press

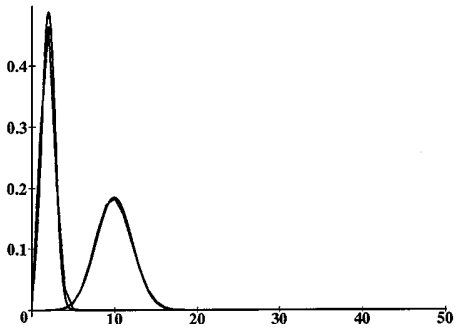


Figure 5 Source distribution and its approximation in a graph with 50 nodes for $q = 0.5$ (left curves) and $q = 0.9$.

the interpretation of n as time and ℓ as position. Then the Markov process corresponds to a diffusion process (see Risken [10]) with

$$a(n,\ell) \approx \lambda_{n,\ell} \quad \text{and} \quad b(n,\ell) \approx \lambda_{n,\ell}.$$

Unfortunately, equation (26) is not easy to solve in the general case. However, if a and b are dependent only on the time and not on the position, i.e.

$$a(t,x) = a(t) \quad \text{and} \quad b(t,x) = b(t),$$

then we can apply a variable substitution to (26) (see Gnedenko [4, §53, §54]) and reduce it to (29), whose solution will be given by $N(\mu, \sigma^2)$.

This is exactly the case for the random variable $\mathscr{S}_n$. As we already have closed forms for its expectation and variance it seems natural to use them in the normal distribution. Accordingly, we get

$$\begin{aligned} \mathrm{S}_{n,\ell} &\approx \frac{1}{\sqrt{2\pi\,\mathbf{Var}\left[\mathscr{S}_n\right]}} \exp\left(-\frac{1}{2}\frac{(\ell - \mathbf{E}\left[\mathscr{S}_n\right])^2}{\mathbf{Var}\left[\mathscr{S}_n\right]}\right) \\ &= \frac{1}{\sqrt{2\pi\,\left([n]_q - [n]_{q^2}\right)}} \exp\left(-\frac{1}{2}\frac{(\ell - [n])^2}{[n]_q - [n]_{q^2}}\right). \end{aligned}$$

This together with the exact formula for the distribution

$$\mathrm{S}_{n,\ell} = (1-q)^{n-\ell}\, q^{\binom{n}{2}}\, S_2(-n, n-\ell)$$

results in the following approximation for the q-Stirling numbers of the second kind:

$$S_2(-n,\ell) \approx (1-q)^{-\ell}\, q^{-\binom{n}{2}} \frac{1}{\sqrt{2\pi\,\left([n]_q - [n]_{q^2}\right)}} \exp\left(-\frac{1}{2}\frac{(\ell - [n])^2}{[n]_q - [n]_{q^2}}\right). \tag{30}$$

This approximation for $\mathrm{S}_{n,\ell}$ seems to be pretty good as we can see in Figure 5. However, there are some remarks to be made. While the distribution is very well approximated for values of ℓ such that $\mathrm{S}_{n,\ell}$ is significantly above 0, the approximation becomes too small for values of ℓ such that $\mathrm{S}_{n,\ell}$ is close to 0. In the latter case the approximation

https://doi.org/10.1017/S0963548397002964 Published online by Cambridge University Press

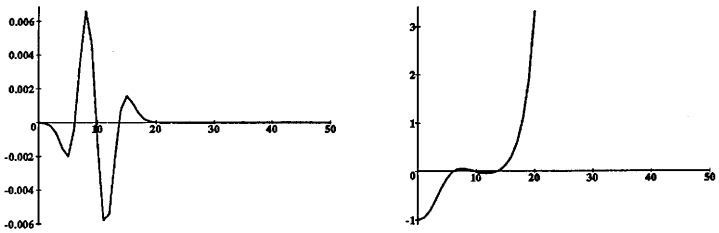


Figure 6 Absolute (left) and relative (right) error between the source distribution and its approximation for $n = 50$, $q = 0.9$.

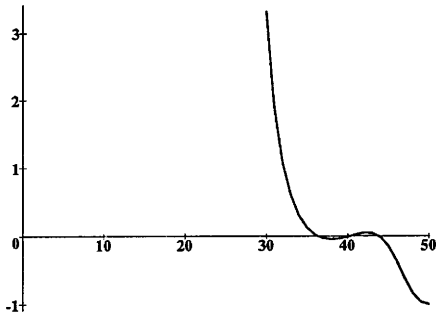


Figure 7 Relative difference between q-Stirling numbers of the second kind and their approximation for $n = 50$ and $q = 0.9$.

could deviate from the correct value by several orders of magnitude. This is still a very small difference in absolute terms, but if we let $\tilde{S}_{n,\ell}$ be the value computed with the approximation formula, then the relative error, computed as

$$\frac{S_{n,\ell} - \tilde{S}_{n,\ell}}{\tilde{S}_{n,\ell}},$$

may be a very large factor. In Figure 6 we have plotted the absolute and relative error between the distribution and the approximation for $n = 50$ and $q = 0.9$ (see Figure 5).

In Figure 7 we have plotted relative difference between the q-Stirling numbers of the second kind and their approximation.

As we can see, the situation for the q-Stirling numbers is similar. The approximation is pretty accurate around the expected value of the source distribution, but the accuracy is rather poor in the regions where the source distribution has values close to 0.

If we look at Figure 4 it becomes clear that expectation and variance become larger for values of q near 1. As a large variance implies a flat curve, for values of q close to 1 there will be a good number of values ℓ for which $S_{n,\ell}$ is sufficiently large. On the other hand, for values of q significantly deviating from 1, the approximation becomes inaccurate for most of the ℓ-values. Nonetheless, for graph theory purposes the values of q close to 1 are the most interesting, since we usually deal with sparse graphs.

https://doi.org/10.1017/S0963548397002964 Published online by Cambridge University Press

5. Concluding remarks

In this paper we have studied the random variable described by a general Markov process for two particular transition probabilities. It turns out that the distributions of these two random variables can be easily described in terms of q-Stirling numbers of the second kind. As far as we know this is the first time that these number theoretical functions arise in conjunction with a probability distribution.

One of the two random variables, the number of sources in a random acyclic digraph, has a transition probability which is only time dependent. This has allowed us to determine an approximation for its distribution through the normal distribution, leading finally to an approximation for some q-Stirling numbers of the second kind. By the transform (9) we get then an approximation for some q-Stirling numbers of the first kind.

Although these approximations are not satisfactory over all range of parameters, they definitely represent a first approach to the approximation of q-Stirling numbers, a result that would be very meaningful to the theory of q-hypergeometric series.

References

[1] Crippa, D. and Simon, K. (1994) q-Distributions in Random Graphs: Transitive Closure and Reduction. *Technical report*, ETH Zurich, Department of Computer Science.
[2] Feller, W. (1971) *An Introduction to Probability Theory and Its Applications*. Vol. 1. Wiley.
[3] Flajolet, P. (1985) Approximate Counting: A Detailed Analysis. *BIT* **25** 113–134.
[4] Gnedenko, B. (1980) *Lehrbuch der Wahrscheinlichkeitsrechnung*. Vol. 274. Verlag Harri Deutsch.
[5] Gould, H. W. (1961) The q-Stirling Numbers of First and Second Kinds. *Duke Mathematical J.* **28** 281–289.
[6] Gasper, G. and Rahman, M. (1990) *Basic Hypergeometric Series*. Cambridge University Press.
[7] Hardy, G. and Wright, E. (1989) *An Introduction to the Theory of Numbers*. Oxford Science Publications.
[8] Morris, R. (1978) Counting Large Numbers of Events in Small Registers. *Comm. ACM*. **21** 840–842.
[9] Palmer, E. M. (1985) *Graphical Evolution*. Wiley-Interscience.
[10] Risken, H. (1984) *The Fokker–Planck Equation*. Springer–Verlag.
[11] Simon, K., Crippa, D. and Collenberg, F. (1993) On the Distribution of the Transitive Closure in Random Acyclic Digraphs. *Lecture Notes in Computer Science*. **726** 345–356.
[12] Simon, K. (1988) Improved Algorithm for Transitive Closure on Acyclic Digraphs. *Theoretical Computer Science*. **58** 325–346.

https://doi.org/10.1017/S0963548397002964 Published online by Cambridge University Press