International Journal of Mathematics and
Computer Science, **7**(2012), no. 2, 113–128

An Existence Theorem in Information Theory

Michelle Foster[1], **Peter Johnson**[2]

[1]Department of Mathematics and Computer Science
Alabama State University
Montgomery, AL 36104, USA

email: mjfoster74@gmail.com

[2]Department of Mathematics and Statistics
Auburn University
Auburn, AL 36849, USA

email: johnspd@auburn.edu

(Received October 4, 2012, Accepted October 27, 2012)

Abstract

A *probabilistic finite state source automaton* (pfssa) is a kind of machine for generating source text over a given alphabet. The theorem referred to in the title gives necessary and sufficient conditions on a k-dimensional array of probabilities, purporting to be the relative k-gram frequencies of some statistically stable source text, for the existence of a pfssa that will generate source text exhibiting the proposed relative k-gram frequencies.

1 Introduction

A *probabilistic finite state source automaton* (pfssa) with source alphabet $S = \{s_1, \ldots, s_m\}$ is a strongly connected directed graph (digraph) $\mathcal{D}$ whose arcs are labelled with pairs (q, s), $q \in (0, 1]$, $s \in S$, satisfying:

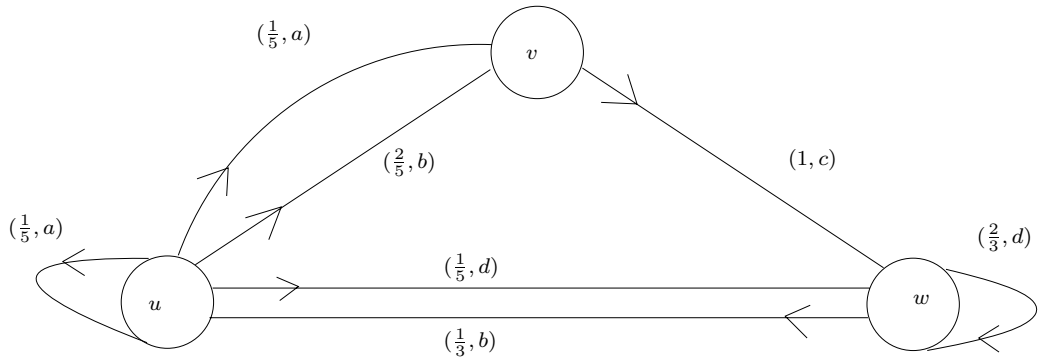
(i) each $s \in S$ appears on some arc;

Key words and phrases: Probabilistic finite state automata, strongly connected directed graph, relative k-gram frequencies.
AMS (MOS) Subject Classifications: 94A05, 60G52.

(ii) for each node, the sum of the q labels on the arcs leaving the node is 1 (one); and

(iii) for any ordered pair (u, v) of nodes of $\mathcal{D}$, and any $s \in S$, s appears as a label on at most one arc going from u to v.

Example 1. *A pfssa with* $S = \{a, b, c, d\}$*, and* 3 *states (nodes).*



The nodes in a pfssa are also called *states*. A "gremlin" moves from state to state in the pfssa, moving along arcs in the directions of the arcs. In any state, the gremlin chooses the arc along which to leave that state according to the probabilities indicated by the q labels on the arcs. The source letter label on the chosen arc is written into the source text. [We will think of the writing as being from left to right.] In the preceding example, for instance, when the gremlin finds itself in state v, it will, with certainty, head for state w along the only arc leaving v and the letter c will be added to the sequence of source letters making up the source text. The possibilities in the other two states are more interesting. We leave the statistical analysis of the source text produced by this pfssa as a recreation; we will look into how to carry out that analysis shortly.

Given how a pfssa works to produce source text, requirements (i) and (ii) are obviously natural. Requirement (iii) is not absolutely essential, but is simplifying: if $\mathcal{D}$ satisfies (i) and (ii) but not necessarily (iii), for each pair u, v of states of $\mathcal{D}$ and $s \in S$, if there are arcs in $\mathcal{D}$ from u to v with s as the letter label, combine them into one arc from u to v with label (q, s), q being the sum of the probability labels on the arcs being combined. The resulting directed graph with labelled arcs will satisfy (i), (ii), and (iii), and will be indistinguishable from $\mathcal{D}$, statistically.

The requirement that a pfssa be strongly connected—meaning that for any ordered pair of different states in the digraph, there is a proper walk in the digraph from the state in the first coordinate to the state in the second—also may not be absolutely essential. If a proposed pfssa were not strongly connected, then no matter where the system's gremlin is first spotted, with probability 1 it would eventually wind up confined to a strong component of the original digraph—i.e., a subdigraph which is maximal, among subdigraphs of the given digraph, with respect to the property of being strongly connected—and this strong component may as well have been the pfssa all along, if we take the long view of the source text. So, we could justify our restriction to strongly connected digraphs on "practical" grounds.

We do not think the restriction requires any justification, at all. From the comments of one of our referees, to whom we are grateful for making plain the possibility of confusion on this point, we see the need to make clear that this is not a paper about pfssa's as yet another species in the vast menagerie of probabilistic finite-state automata (pfsa's: among these, the pfssa is a special case of a stochastic sequential machine, treated in [4]; and more recently, our referee informs us, of a "transducer"). Rather, the role of the pfssa here is as a statistically stable source, an emitter of statistically stable source text (next section).

Here is an open question: given a statistically stable source (or rather, given the text from that source), is there a pfssa, or any sort of concretely realizable machine, that will produce text that is statistically indistinguishable from that of the given source? One corollary of our main result (Theorem 1, section 4) is that for every statistically stable source and every non-negative integer k, there is a pfssa that produces text which has exactly the same k^{th} order statistical properties (definitions to come) as the given text.

This is far from the answer to that open question, but maybe it is a start; and because statistically stable sources are fundamental entities in information theory, perhaps the result justifies interest in pfssa's in information theory. In fact, we claim that pfssa's were introduced informally, and not by any particular name, at the creation of information theory, in [5]. The particular pfssa's casually proposed by Shannon for simulating real human-language source text were of the special type introduced in section 3, which we call *simulants*.

2 Relative k-gram frequencies

Suppose that $S = \{s_1, \ldots, s_m\}$ is an alphabet. (It will be our unspecified source alphabet throughout.) For any positive integer k, a *k-gram* over S is a word of length k made of letters from S; in other words, a k-gram over S is an element of S^k.

A source with alphabet S is *statistically stable* if there is a function $f : \bigcup_{k=1}^{\infty} S^k \to [0,1]$ such that for each k and each word $s_{i_1} \ldots s_{i_k} \in S^k$, $f(s_{i_1} \ldots s_{i_k})$ is the true probability that k consecutive letters selected at random from the text produced by the source will be the word $s_{i_1} \ldots s_{i_k}$. We will refer to $f(s_{i_1} \cdots s_{i_k})$ as the *relative frequency* of $s_{i_1} \cdots s_{i_k}$ in the text produced by the source.

The statistical stability of a given source of which we know only a finite portion of the text it produces, and of which we have no knowledge of the inner workings, is conjectural, impossible to verify with certainty, although there are clearly statistical tests that can be run on the known source text to test the hypothesis of stability.

But a pfssa, considered as a source of text, is a statistically stable source, and the relative k-gram frequencies in its text can be computed directly. Here is a brief account verifying these assertions. A more detailed account is given in [1], but we hope that the following will serve.

Suppose that $\mathcal{D}$ is a pfssa with state set $\mathcal{V} = \{v_1, \ldots, v_r\}$. For each i, $j \in \{1, \ldots, r\}$, let q_{ij} be the sum of the probability labels on the arcs of $\mathcal{D}$ that start in v_i and go to v_j. The $r \times r$ matrix $Q = [q_{ij}]$ is the matrix of *state transition probabilities* for $\mathcal{D}$ with respect to the ordering $v_1, \ldots, v_r$ of $\mathcal{V}$.

Braving the fact that we do not know when in the past the pfssa started generating text, nor in which state the resident gremlin dwelt at the start, if, indeed, there was a start, we make the assumption that there are positive probabilities $p_1, \ldots, p_r$ such that, at a randomly selected pulse of time when the gremlin is in a particular state, just before it makes a move which generates a letter of text, the probability is p_i that the state of temporary residence is v_i, $i = 1, \ldots, r$. (The assumption of the existence of these probabilities greatly simplifies discussion of $\mathcal{D}$ and the text it emits but is this assumption valid? Can the existence of $p_1, \ldots, p_r$ be satisfactorily proven? There is a proof from the foundations of ergodic theory (see [6])—whether or not this proof is satisfactory we leave to philosophers to debate. However, although we are not well-versed in the theory of probabilistic finite-state automata, we get the impression that the existence of these state probabilities in a strongly connected pfsa is widely accepted, in a folkloric way.)

If the p_i exist then, by elementary probability and the assumed structure and workings of $\mathcal{D}$, the p_i must satisfy

$$p_j = \sum_{i=1}^{r} p_i q_{ij}, \text{ or, letting } \bar{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_r \end{pmatrix},$$

$Q^T\bar{p} = \bar{p}$. At this point classical linear algebra provides a result that supports the assumption of the existence of the p_i: from the Perron-Frobenius theorem [3] it can be deduced that if Q is the matrix of state transition probabilities of a strongly connected pfsa, then the equation $Q^T\bar{p} = \bar{p}$ has exactly one solution $\bar{p}$ among the probability vectors, and the entries of the unique probability vector satisfying the equation are all positive.

Once the existence of the state probabilities is granted, and those probabilities are found, the existence of the relative k-gram frequencies in the text generated by the pfssa $\mathcal{D}$, and their calculation, is straightforward, by elementary probability theory. To calculate the relative frequency of $s_{i_1} \cdots s_{i_k}$ in the output of $\mathcal{D}$, one adds the probabilities of the various proper walks of length k in $\mathcal{D}$ such that the sequence of letter labels encountered on the arcs traversed during the walk is $s_{i_1}, \ldots, s_{i_k}$. The probability of any such walk is the probability p_s of the gremlin being in the state v_s from which the walk begins times the product of the probability labels on the arcs traversed.

For example, if $\mathcal{D}$ is the 3-state pfssa in the preceding section, and the states are ordered u, v, w, then

$$Q = \begin{bmatrix} 1/5 & 3/5 & 1/5 \\ 0 & 0 & 1 \\ 1/3 & 0 & 2/3 \end{bmatrix},$$

and the unique probability vector in the null space of

$$Q^T - I_3 \quad \text{is} \quad \begin{pmatrix} p_u \\ p_w \\ p_w \end{pmatrix} = \begin{pmatrix} 1/4 \\ 3/20 \\ 3/5 \end{pmatrix};$$

we calculate

$$f(dba) = \frac{1}{4}(\frac{1}{5} \cdot \frac{1}{3} \cdot \frac{1}{5}) + \frac{1}{4}(\frac{1}{5} \cdot \frac{1}{3} \cdot \frac{1}{5}) + \frac{3}{5}(\frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{5}) + \frac{3}{5}(\frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{5}),$$

while $f(abd) = 0$, simply because there is no proper walk of length 3 in $\mathcal{D}$ with letter labels a, b, d, in that order, on the arcs along the walk.

Now suppose we have a statistically stable source of unknown structure, with alphabet $S = \{s_1, \ldots, s_m\}$ and relative frequency function $f : \bigcup_{k=1}^{\infty} S^k \to [0,1]$. By elementary arguments we have the *consistency condition*:

For all $k \geq 2$ *and* $i_1, \ldots, i_{k-1} \in \{1, \ldots, m\}$, $\sum_{j=1}^{m} f(s_{i_1} \ldots s_{i_{k-1}} s_j) = \sum_{j=1}^{m} f(s_j s_{i_1} \ldots s_{i_{k-1}}) = f(s_{i_1} \cdots s_{i_{k-1}})$.

A corollary: the values of f on S^k determine the values of f on $\bigcup\limits_{1 \leq t < k} S^t$.

It is worth noting that if $k \geq 3$ and $[a(i_1, \ldots, i_k)]$, $1 \leq i_1, \ldots, i_k \leq m$, is a k-dimensional array of numbers (or, more generally, of elements of an abelian group) satisfying $\sum_{j=1}^{m} a(i_1, \ldots, i_{k-1}, j) = \sum_{j=1}^{m} a(j, i_1, \ldots, i_{k-1})$ for all $i_1, \ldots, i_{k-1} \in \{1, \ldots, m\}$, and if we define b on $\{1, \ldots, m\}^{k-1}$ by $b(i_1, \ldots, i_{k-1}) =$ the common value of $\sum_{j=1}^{m} a(i_1, \ldots, i_{k-1}, j)$ and $\sum_{j=1}^{m} a(j, i_1, \ldots, i_{k-1})$, then b satisfies $\sum_{j=1}^{m} b(i_1, \ldots, i_{k-2}, j) = \sum_{j=1}^{m} b(j, i_1, \ldots, i_{k-2})$ for all $i_1, \ldots, i_{k-2} \in \{1, \ldots, m\}$:

$$\begin{aligned} & \textstyle\sum_{j=1}^{m} b(i_1, \ldots, i_{k-2}, j) = \sum_{j=1}^{m} \sum_{t=1}^{m} a(t, i_1, \ldots, i_{k-2}, j) \\ = \; & \textstyle\sum_{t=1}^{m} \sum_{j=1}^{m} a(t, i_1, \ldots, i_{k-2}, j) = \sum_{t=1}^{m} b(t, i_1, \ldots, i_{k-2}). \end{aligned}$$

3 Simulants

Suppose that $\mathcal{S}$ is a statistically stable source with alphabet $S = \{s_1, \ldots, s_m\}$ and relative frequency function f. The *simulant* of $\mathcal{S}$ of order 0, denoted $\mathcal{S}^{(0)}$, is the pfssa with a single state U and with a loop at U for each letter $s_j \in S$, with label $(f(s_j), s_j)$.

If $r > 0$, the simulant of $\mathcal{S}$ of order r, denoted $\mathcal{S}^{(r)}$, is an edge-labelled digraph with node set $\{s_{i_1} \cdots s_{i_r} \mid i_1, \ldots, i_r \in \{1, \ldots, m\}$ and $f(s_{i_1} \cdots s_{i_r}) > 0\}$, with an arc from $s_{i_1} \cdots s_{i_r}$ to $s_{j_1} \cdots s_{j_r}$ if and only if $j_t = i_{t+1}$, $t = 1, \ldots, r-1$, and $f(s_{i_1} s_{j_1} \cdots s_{j_r}) > 0$, and with (probability, letter) label $\left(\frac{f(s_{i_1} s_{j_1} \cdots s_{j_r})}{f(s_{i_1} \cdots s_{i_r})}, s_{j_r}\right)$ on such an arc.

That is to say, for each $(r+1)$-gram $s_{i_1} \cdots s_{i_{r+1}}$ with positive relative frequency, in $\mathcal{S}^{(r)}$ there is an arc from node $s_{i_1} \cdots s_{i_r}$ to node $s_{i_2} \cdots s_{i_{r+1}}$ with label $\left(\frac{f(s_{i_1} \cdots s_{i_{r+1}})}{f(s_i \cdots s_{i_r})}, s_{i_{r+1}}\right)$. The idea here is one of simulation: if you are at the node, or in the state, $s_{i_1} \cdots s_{i_r}$, it means that the last r letters read in the source text were $s_{i_1}, \ldots, s_{i_r}$, in that order; for each possible next letter s_j, the probability that s_j will indeed be the next letter, given that the previous r letters have been $s_{i_1} \cdots s_{i_r}$, is the conditional probability $\frac{f(s_{i_1} \cdots s_{i_r} s_j)}{f(s_{i_1} \cdots s_{i_r})}$, and if s_j is, indeed, the next letter, then the new state to which the gremlin

will have to move is $s_{i_2} \cdots s_{i_r} s_j$. So, the r^{th} order simulant of the source simulates the behavior of the source after each r-gram in the source text, and the simulation "rolls" naturally from each r-gram to the next, one letter at a time.

The proof of the following is given in [1], but we give a proof here, for future reference.

Proposition 1. *If $\mathcal{S}$ is a statistically stable source then, for each $r \geq 0$, $\mathcal{S}^{(r)}$ is strongly connected, and is therefore a pfssa. Further, if S is the alphabet of $\mathcal{S}$, and f is its relative frequency function, and $\tilde{f}$ is the relative frequency function of $\mathcal{S}^{(r)}$, then $f = \tilde{f}$ on $\bigcup_{k=1}^{r+1} S^k$. When $r > 0$ the state probability of each state $s_{i_1} \cdots s_{i_r}$ in $S^{(r)}$ is $f(s_{i_1} \cdots s_{i_r})$.*

Proof. Clearly all claims of the Proposition hold for $r = 0$. Suppose $r > 0$ and $f(s_{i_1} \cdots s_{i_r})$, $f(s_{j_1} \cdots s_{j_r}) > 0$. Then for each chosen point in the source text, with probability 1 each of $s_{i_1} \cdots s_{i_r}$ and $s_{j_1} \cdots s_{j_r}$ must occur infinitely often as r-grams in the source text beyond that point. We may assume that $s_{i_1} \cdots s_{i_r} \neq s_{j_1} \cdots s_{j_r}$. Consider an occurrence of $s_{i_1} \cdots s_{i_r}$ in the source text, and an occurrence of $s_{j_1} \cdots s_{j_r}$, further on. As you scan from one to the other, every $(r+1)$-gram encountered has positive relative frequency, just because it has actually occurred. Thus there is a proper walk in $\mathcal{S}^{(r)}$ from the node $s_{i_1} \cdots s_{i_r}$ to the node $s_{j_1} \cdots s_{j_r}$.

Each $s \in S$ appears in the source text, and therefore appears as a letter label on some arc of $\mathcal{S}^{(r)}$. Therefore $\mathcal{S}^{(r)}$ satisfies (i), in the definition of a pfssa.

If $i_1, \ldots, i_{r+1} \in \{1, \ldots, m\}$, and $f(s_{i_1} \cdots s_{i_{r+1}}) > 0$, the transition probability from state $s_{i_1} \cdots s_{i_r}$ to state $s_{i_2} \cdots s_{i_{r+1}}$ is $f(s_{i_1} \cdots s_{i_{r+1}})/f(s_{i_1} \cdots s_{i_r})$, the probability label on the sole arc from the one state to the other. It follows that the sum of the probability labels on arcs leaving state $s_{i_1} \cdots s_{i_r}$ is $\sum_{j=1}^{m} f(s_{i_1} \cdots s_{i_r} s_j))/f(s_{i_1} \cdots s_{i_r}) = 1$, by part of the consistency condition satisfied by f. Therefore (ii), in the definition of a pfssa, is satisfied by $\mathcal{S}^{(r)}$.

Requirement (iii) is obviously satisfied by $\mathcal{S}^{(r)}$, since for any ordered pair of states in $\mathcal{S}^{(r)}$, there is at most one arc in $\mathcal{S}^{(r)}$ from the first to the second.

To show that $f = \tilde{f}$ on $\bigcup_{k-1}^{r+1} S^k$, it suffices to show that $f = \tilde{f}$ on S^{r+1}, by the corollary of the consistency condition, since both $\mathcal{S}$ and $\mathcal{S}^{(r)}$ are statistically stable sources. First, we will see that the state probability $p(s_{i_1} \cdots s_{i_r})$ of a state $s_{i_1} \cdots s_{i_r}$ (such that $f(s_{i_1} \cdots s_{i_r}) > 0$) in $\mathcal{S}^{(r)}$ is $f(s_{i_1} \cdots s_{i_r})$. To see this, it suffices to verify that the system of equations $\bar{p} = Q^T \bar{p}$, the satisfying of which is necessary and sufficient for a probability

vector $\bar{p} = [p(s_{i_1} \cdots s_{i_r}); f(s_{i_1} \cdots s_{i_r}) > 0]$ to be the vector of state probabilities in $\mathcal{S}^{(r)}$, consists of the equations

$$p(s_{i_1} \cdots s_{i_r}) = \sum_{\{j \in \{1,\ldots,m\} | f(s_j s_{i_1} \cdots s_{i_r}) > 0\}} p(s_j s_{i_1} \cdots s_{i_{r-1}}) \frac{f(s_j s_{i_1} \cdots s_{i_r})}{f(s_j s_{i_1} \cdots s_{i_{r-1}})}$$

Plugging $p(s_{i_1} \cdots s_{i_r}) = f(s_{i_1} \cdots s_{i_r})$ for all $s_{i_1} \cdots s_{i_r} \in S^r$ gives the equations

$$f(s_{i_1} \cdots s_{i_r}) = \sum_{j=1}^{m} f(s_j s_{i_1} \cdots s_{i_r}),$$

which are satisfied as part of the consistency condition for f.

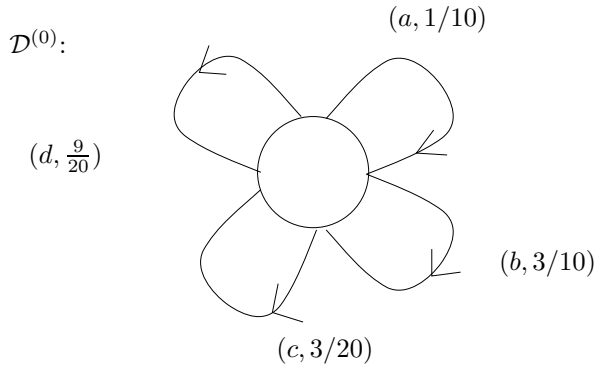
Now suppose that $s_{i_1} \cdots s_{i_{r+1}} \in S^{r+1}$. The walks in $\mathcal{S}^{(r)}$ that would generate $s_{i_1} \cdots s_{i_{r+1}}$ start from some state $s_{j_1} \cdots s_{j_r}$, then go to $s_{j_2} \cdots s_{j_r} s_{i_1}$, then to $s_{j_3} \cdots s_{j_r} s_{i_1} s_{i_2}$ (if $r \geq 3$), etc., and necessarily finishing with a move from state $s_{i_1} \cdots s_{i_r}$ to $s_{i_2} \cdots s_{i_r} s_{i_r+1}$. Let $B \subseteq S^r$ be the set of states in $\mathcal{S}^{(r)}$ from which such a walk can start.

If $f(s_{i_1} \cdots s_{i_{r+1}}) = 0$, then either one of $s_{i_1} \cdots s_{i_r}$, $s_{i_2} \cdots s_{i_{r+1}}$ is not a state in $\mathcal{S}^{(r)}$ (because f assigns value 0 to one of them, considered as an r-gram in the source text), or they are both states of $\mathcal{S}^{(r)}$, but there is no arc from the former to the latter. In either case, $B = \emptyset$ and $\tilde{f}(s_{i_1} \cdots s_{i_{r+1}}) = 0$.

If $f(s_{i_1} \cdots s_{i_{r+1}}) > 0$ then $s_{i_1} \cdots s_{i_{r+1}}$ does occur in the source text, and, therefore, the $(2r+1)$-gram $s_{j_1} \cdots s_{j_r} s_{i_1} \cdots s_{i_{r+1}}$ occurs for at least one $s_{j_1} \cdots s_{j_r} \in B$. The relative frequency of $s_{i_1} \cdots s_{i_{r+1}}$ in the text produced by $\mathcal{S}^{(r)}$ is the sum, over such $s_{j_1} \cdots_{j_r} \in B$, of products

$$f(s_{j_1} \cdots s_{j_r}) \frac{f(s_{j_1} \cdots s_{j_r} s_{i_1})}{f(s_{j_1} \cdots s_{j_r})} \cdots \frac{f(s_{i_1} \cdots s_{i_{r+1}})}{f(s_{i_1} \cdots s_{i_r})},$$

because of the way probabilities are assigned to the arcs of $\mathcal{S}^{(r)}$, and because the state probability of $s_{j_1} \cdots s_{j_r}$ in $\mathcal{S}^{(r)}$ is $f(s_{j_1} \cdots s_{j_r})$. By elementary probability, such a product is the relative frequency of $s_{j_1} \cdots s_{j_r} s_{i_1} \cdots s_{i_{r+1}}$ among $(2r+1)$-grams in the source text produced by $\mathcal{S}$, i.e. $f(s_{j_1} \cdots s_{j_r} s_{i_1} \cdots s_{i_{r+1}})$, and, again by elementary probability, the sum of these relative frequencies over r-grams $s_{j_1} \cdots s_{j_r}$ that do occur as the first r letters of a $(2r+1)$-gram whose last $r+1$ letters are $s_{i_1} \cdots s_{i_{r+1}}$, i.e., over $s_{j_1} \cdots s_{j_r} \in B$, will be $f(s_{i_1} \cdots s_{i_{r+1}})$, the relative frequency of $s_{i_1} \cdots s_{i_{r+1}}$ among $(r+1)$-grams in the original source text. Thus $f(s_{i_1} \cdots s_{i_{r+1}}) = \tilde{f}(s_{i_1} \cdots s_{i_{r+1}})$, whether $f(s_{i_1} \cdots s_{i_{r+1}}) = 0$ or not. Thus $f = \tilde{f}$ on S^{r+1}, and, therefore, on $\bigcup_{k=1}^{r+1} S^k$. □



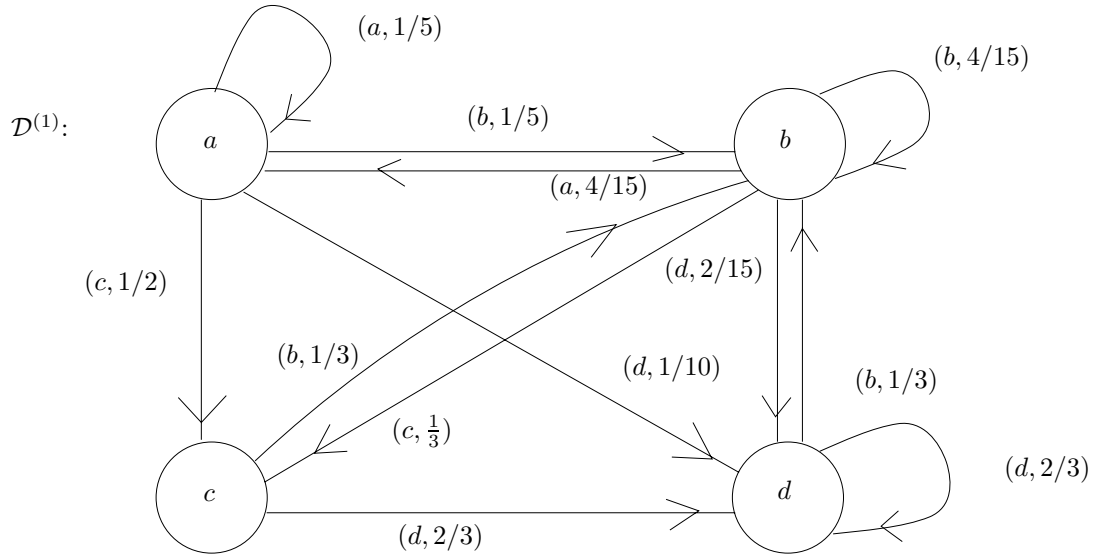


Figure 1: The zeroth and first order simulants of the pfssa $\mathcal{D}$ in the Introduction. $\mathcal{D}^{(2)}$ has 12 nodes.

4 The main result

Throughout this section, m, $k \geq 2$ will be integers, and $A : \{1, \ldots, m\}^k \to [0, \infty)$ will be a k-dimensional array of non-negative numbers. The questions to be considered are: under what circumstances is there a statistically stable source with an m-letter alphabet $S = \{s_1, \ldots, s_m\}$ and relative frequency function f such that $A(i_1, \ldots, i_k) = f(s_{i_1} \cdots s_{i_k})$ for all $i_1, \ldots, i_k \in \{1, \ldots, m\}$? And when will there be a "solution" source real-

izable as a pfssa? There are 5 conditions of interest.

(1) $\sum_{1\leq i_1,\ldots,i_k\leq m} A(i_1,\ldots,i_k) = 1$.

(2) For all $i_1,\ldots,i_{k-1} \in \{1,\ldots,m\}$, $\sum_{j=1}^{m} A(i_1,\ldots,i_{k-1},j) = \sum_{j=1}^{m} A(j,i_1,\ldots,i_{k-1})$.

(3) For all $i_1,\ldots,i_{k-1},j_1,\ldots,j_{k-1} \in \{1,\ldots,m\}$, with $(i_1,\ldots,i_{k-1}) \neq (j_1,\ldots,j_{k-1})$, such that

$$\sum_{r=1}^{m} A(i_1,\ldots,i_{k-1},r), \sum_{r=1}^{m} A(r,j_1,\ldots,j_{k-1}) > 0,$$

there is a sequence $t_1,\ldots,t_p \in \{1,\ldots,m\}$, $p \geq k$, such that $(t_1,\ldots,t_{k-1}) = (i_1,\ldots,i_{k-1})$, $(t_{p-k+2}\cdots,t_p) = (j_1,\ldots,j_{k-1})$, and, for all $z \in \{1,\ldots,p-k+1\}$,
$A(t_z,\ldots,t_{z+k-1}) > 0$.

(4) Each $j \in \{1,\ldots,m\}$ is an entry in some sequence $(i_1,\ldots,i_k) \in \{1,\ldots,m\}^k$ such that $A(i_1,\ldots,i_k) > 0$.

(5) There does not exist a proper subset I of $\{1,\ldots,m\}$ such that, letting $J = \{1,\ldots,m\}\backslash I$, $A \equiv 0$ on $\{1,\ldots,m\}^k\backslash(I^k \cup J^k)$.

Lemma 1. *Condition [5] implies (4), and (3) and (4) imply (5).*

Proof. If (4) fails then some $i \in \{1,\ldots,n\}$ is in no k-tuple $(i_1,\ldots,i_k) \in \{1,\ldots,m\}^k$ such that $A(i_1,\ldots,i_k) > 0$. But then, taking $I = \{i\}$, we see that (5) fails.

Now assume that (3) and (4) hold. Suppose there is a proper subset I of $\{1,\ldots,m\}$ such that $A = 0$ on $\{1,\ldots,m\}^k\backslash(I^k \cup J^k)$, where J denotes $\{1,\ldots,m\}\backslash I$. Take any $i \in I$, $j \in J$; by (4), there exist $(i_1,\ldots,i_k),(j_1,\ldots,j_k) \in \{1,\ldots,m\}^k$ such that $A(i_1,\ldots,i_k)$, $A(j_1,\ldots,j_k) > 0$ and $i \in \{i_1,\ldots,i_k\}$, $j \in \{j_1,\ldots,j_k\}$. Then $\{i_1,\ldots,i_k\} \subseteq I$ and $\{j_1,\ldots,j_k\} \subseteq J$. Since

$$0 < A(i_1,\ldots,i_k) \leq \sum_{r=1}^{m} A(i_1,\ldots,i_{k-1},r)$$

and

$$0 < A(j_1,\ldots,j_r) \leq \sum_{r=1}^{m} A(r,j_2,\ldots,j_k),$$

by (3) there exist $t_1, \ldots, t_p \in \{1, \ldots, m\}$ such that $(t_1, \ldots, t_{k-1}) = (i_1, \ldots, i_{k-1})$, $(t_{p-k+2}, \ldots, t_p) = (j_2, \ldots, j_k)$, and A assigns a positive value to every block of k consecutive entries in the sequence $(t_1, \ldots, t_p)$. But then, since the first $k-1$ entries are in I, and the last $k-1$ entries are in J, and $k \geq 2$, there would have to be a block of k consecutive entries of $(t_1, \ldots, t_p)$ containing entries from both I and J. This would contradict the supposition that A assigns the value 0 to every such element of $\{1, \ldots, m\}^k$. □

Lemma 2. *Suppose that (2) holds,* $1 \leq t \leq k-1$ *and that* $j_1, \ldots, j_t \in \{1, \ldots, m\}$ *occupy, in that order, positions* $s, \ldots, s+t-1$ *in some integer vector* $\bar{i} = (i_1, \ldots, i_k) \in \{1, \ldots, m\}^k$ *such that* $A(i_1, \ldots, i_k) > 0$. *If* $s+t-1 < k$ *then* $j_1, \ldots, j_t$ *also occupy, in that order, positions* $s+1, \ldots, s+t$ *of another vector* $\bar{i}' \in \{1, \ldots, m\}^k$ *such that* $A(\bar{i}') > 0$*; and if* $s > 1$ *then* $j_1, \ldots, j_t$ *occupy positions* $s-1, \ldots, s+t-2$ *of a vector* $\bar{i}$" $\in \{1, \ldots, m\}^k$ *such that* $A(\bar{i}'') > 0$.

Proof. $0 < A(\bar{i}) \leq \sum_{j=1}^m A(i_1, \ldots, i_{k-1}, j) = \sum_{j=1}^m A(j, i_1, \ldots, i_{k-1})$, from which the existence of $\bar{i}'$ when $s+t-1 < k$ is easily seen; the other conclusion is seen similarly. □

Theorem 1. *The following are equivalent.*

(a) There is a statistically stable source with an m-letter alphabet $S = \{s_1, \ldots, s_m\}$ *whose relative frequency function* f *satisfies* $f(s_{i_1} \cdots s_{i_k}) = A(i_1, \ldots, i_k)$ *for all* $i_1, \ldots, i_k \in \{1, \ldots, m\}$.

(b) There is a pfssa with an m-letter alphabet S *whose relative frequency function* f *satisfies* $f(s_{i_1} \cdots s_{i_k}) = A(i_1, \ldots, i_k)$ *for all* $i_1, \ldots, i_k \in \{1, \ldots, m\}$.

(c) A *satisfies conditions (1), (2), (3), and (4).*

Proof. Clearly (b) implies (a). If (a) holds then by Proposition 1 the simulant of order $k-1$ of that source is a pfssa satisfying the requirements in (b). Thus (a) and (b) are equivalent.

If a source satisfying (a) exists, let $\mathcal{D}$ be its simulant of order $k-1$, with states $s_{i_1} \cdots s_{i_{k-1}}$, for $i_1, \ldots, i_{k-1} \in \{1, \ldots, m\}$ such that $f(s_{i_1} \cdots s_{i_{k-1}}) > 0$. Since $f(s_{i_1} \cdots s_{i_k}) = A(i_1, \ldots, i_k)$ for all $i_1, \ldots, i_k \in \{1, \ldots, m\}$, clearly A must satisfy conditions (1), (2), and (4). Suppose that $i_1, \ldots, i_{k-1}, j_1, \ldots, j_{k-1} \in \{1, \ldots, m\}$, $(i_1, \ldots, i_{k-1}) \neq (j_1, \ldots, j_{k-1})$, and $\sum_{r=1}^m A(i_1, \ldots, i_{k-1}, r)$, $\sum_{r=1}^m A(r, j_1, \ldots, j_{k-1}) > 0$. Then $f(s_{i_1} \cdots s_{i_{k-1}})$, $f(s_{j_1} \cdots s_{j_{k-1}}) > 0$, so $s_{i_1} \ldots s_{i_{k-1}}$ and $s_{j_1} \cdots s_{j_{k-1}}$ are states in $\mathcal{D}$. Because $\mathcal{D}$ is strongly connected (Proposition 1), there is a proper walk in $\mathcal{D}$ from state $s_{i_1} \cdots s_{i_{k-1}}$ to state

$s_{j_1} \cdots s_{j_{k-1}}$. By the way simulants are defined, and by the assumption that $f(s_{t_1} \cdots s_{t_k}) = A(t_1, \ldots, t_k)$ for all $t_1, \ldots, t_k \in \{1, \ldots, m\}$, the existence of that walk implies the existence of a sequence $t_1, \ldots, t_p \in \{1, \ldots m\}$ satisfying the requirements of (3) with respect to A and $i_1, \ldots, i_{k-1}$,
$j_1, \ldots, j_{k-1}$. Thus (a) implies that A satisfies condition (3), and so (a) implies (c).

It remains to be seen that (c) implies (a), or (b). Assuming (c), we define a pfssa by forming the simulant $\mathcal{D}$ of order $k-1$ of a source that would satisfy (a), if there were such a source. To define $\mathcal{D}$, let $S = \{s_1, \ldots, s_m\}$ be an alphabet; let $B : \{1, \ldots, m\}^{k-1} \to [0, \infty)$ be defined by $B(i_1, \ldots, i_{k-1}) = \sum_{j=1}^{m} A(i_1, \ldots, i_{k-1}, j) = \sum_{j=1}^{m} A(j, i_1, \ldots, i_{k-1})$ (since A satisfies (1), the codomain of B is actually $[0, 1]$); let the states of $\mathcal{D}$ be the words $s_{i_1} \cdots s_{i_{k-1}} \in S^{k-1}$ such that $B(i_1, \ldots, i_{k-1}) > 0$, and let these be connected by labelled arcs as they would be in a simulant of order $k-1$: There can be an arc from $s_{i_1} \cdots s_{i_{k-1}}$ only to states $s_{i_2} \cdots s_{i_{k-1}} s_j$ such that $A(i_1, \ldots, i_{k-1}, j) > 0$, and such an arc will have probability label $A(i_1, \ldots, i_{k-1}, j)/B(i_1, \ldots, i_{k-1})$ and letter label s_j. It remains to be shown that the digraph with labelled arcs thus defined is a pfssa with alphabet S and that, if f is its relative frequency function, then $f(s_{i_1} \cdots s_{i_k}) = A(i_1, \ldots, i_k)$ for all $i_1, \ldots, i_k \in \{1, \ldots, m\}$.

Requirement (iii) for $\mathcal{D}$ in the definition of a pfssa, follows, as in the case of the simulants, from the fact that there is at most one arc going from u to v, for each pair u, v of nodes of $\mathcal{D}$. Requirement (ii) follows from the definition of B.

To see that $\mathcal{D}$ satisfies (i), we apply Lemma 2 with $t = 1$ to conclude that if $j \in \{1, \ldots, m\}$ then, because (2) holds, there must exist $i_1, \ldots, i_{k-1} \in \{1, \ldots, m\}$ such that $A(i_1, \ldots, i_{k-1}, j) > 0$. Then s_j is the letter label on the arc from node $s_{i_1} \cdots s_{i_{k-1}}$ to node $s_{i_2} \cdots s_j$.

It is straightforward to see that condition (3) implies that $\mathcal{D}$ is strongly connected. So $\mathcal{D}$ is a pfssa. Now it remains to be seen that, if f is the relative frequency of $\mathcal{D}$, then $f(s_{i_1} \cdots s_{i_k}) = A(i_1, \ldots, i_k)$ for all $i_1, \ldots, i_k \in \{1, \ldots, m\}$.

By condition (1),

$$\sum_{1 \le i_1, \ldots, i_{k-1} \le m} B(i_1, \ldots, i_{k-1}) = \sum_{1 \le i_1, \ldots, i_{k-1}, j \le m} A(i_1, \ldots, i_{k-1}, j) = 1,$$

so $\{B(i_1, \ldots, i_{k-1}) \mid s_{i_1} \cdots s_{i_{k-1}}$ is a node of $\mathcal{D}\}$ is a positive probability vector indexed by the states (nodes) of $\mathcal{D}$. As in the proof of Proposition 1, it is straightforward—in this case by (2) and the definition of B—to see

that the equations necessary and sufficient for $B(i_1, \ldots, i_{k-1})$ to be the state probability of a node $s_{i_1} \cdots s_{i_{k-1}}$ of $\mathcal{D}$ are satisfied:

$$\begin{aligned} B(i_1, \ldots, i_{k-1}) &= \textstyle\sum_{j=1}^{m} A(j, i_1, \ldots, i_{k-1}) \\ &= \textstyle\sum_{\{j \mid A(j,i_1,\ldots,i_{k-1})>0\}} B(j, i_1, \ldots, i_{k-2}) \frac{A(j,i_1,\ldots,i_{k-1})}{(B(j,i_1,\ldots,i_{k-2})}. \end{aligned}$$

(In case $k = 2$, $B(i_1, \ldots, i_{k-1}) = B(i_1)$ and $B(j, i_1, \ldots, i_{k-2}) = B(j)$; a similar common-sense adjustment is needed when $k = 3$.)

The proof to follow that $A(i_1, \ldots, i_k) = f(s_{i_1} \cdots s_{i_k})$ for all $i_1, \ldots, i_k \in \{1, \ldots, m\}$ could have been used for the proof of the corresponding statement in Proposition 1, but assumptions there permitted an easier proof. However, the proof here starts from the same point: for any $i_1, \ldots, i_k \in \{1, \ldots, m\}$, $f(s_{i_1} \cdots s_{i_k})$ is the sum, over $(j_1, \ldots, j_{k-1}) \in \{1, \ldots, m\}^{k-1}$ such that $B(j_1, \ldots, j_{k-1}) > 0$ and there is an $s_{i_1} \cdots s_{i_k}$-generating walk in $\mathcal{D}$ starting from the state $s_{j_1} \cdots s_{j_{k-1}}$, of the products

$$B(j_1, \ldots, j_{k-1}) \tfrac{A(j_1,\ldots,j_{k-1},i_1)}{B(j_1,\ldots,j_{k-1})} \tfrac{A(j_2,\ldots,j_{k-1},i_1,i_2)}{B(j_2,\ldots,j_{k-1},i_1))} \cdots \tfrac{A(i_1,\ldots,i_k)}{B(i_1,\ldots,i_{k-1})}$$

$$= A(j_1, \ldots, j_{k-1}, i_1) \tfrac{A(j_2,\ldots,j_{k-1},i_1,i_2)}{B(j_2,\ldots,j_{k-1},i_1)} \cdots \tfrac{A(i_1,\ldots,i_k)}{B(i_1,\ldots,i_{k-1})}$$

Call this product $P(j_1, \ldots, j_{k-1}, i_1, \ldots, i_k)$, for short. Then $f(s_{i_1} \cdots s_{i_k}) = \sum_{j_1,\ldots,j_{k-1}} P(j_1, \ldots, j_{k-1}, i_1, \ldots, i_k)$, where the sum, as stated above, is over a certain subset of $\{1, \ldots, m\}^{k-1}$.

If $B(t_1, \ldots, t_{k-1}) = 0$, then $A(j, t_1, \ldots, t_{k-1}) = A(t_1, \ldots, t_{k-1}, j) = 0$ for all $j \in \{1, \ldots, m\}$. In such a case, let $\frac{A(t_1,\ldots,t_{k-1},j)}{B(t_1,\ldots,t_{k-1})} = 0$ for all $j \in \{1, \ldots, m\}$; this is a notational convention which is apparently trivial, but which allows us to observe that for each $i_1, \ldots, i_k \in \{1, \ldots, m\}$,

$$\begin{aligned} f(s_{i_1} \cdots s_{i_k}) &= \textstyle\sum_{1 \le j_1,\ldots,j_{k-1} \le m} P(j_1, \ldots, j_{k-1}, i_1, \ldots, i_k) \\ &= \textstyle\sum_{1 \le j_2,\ldots,j_{k-1} \le m} \left[\sum_{j_1=1}^{m} A(j_1, \ldots, j_{k-1}, i_1)\right] \frac{P(j_2,\ldots,j_{k-1},i_1,\ldots,i_k)}{B(j_2,\ldots,j_{k-1},i_1)} \\ &= \textstyle\sum_{1 \le j_2,\ldots,j_{k-1} \le m} B(j_2, \ldots, j_{k-1}, i_1) \frac{P(j_2,\ldots,j_2-1,i_1,\ldots,i_k)}{B(j_2,\ldots,j_{k-2},i_1)} \\ &= \textstyle\sum_{1 \le j_2,\ldots,j_{k-1} \le m} P(j_2, \ldots, j_{k-1}, i_1, \ldots, i_k) \\ &= \cdots = \textstyle\sum_{j=1}^{m} P(j, i_1, \ldots, i_k) \\ &= (\textstyle\sum_{j=1}^{m} A(j, i_1, \ldots, i_{k-1})) \frac{A(i_1,\ldots,i_k)}{B(i_1,\ldots,i_{k-1})} = A(i_1, \ldots, i_k). \end{aligned}$$

□

Theorem 2. *If, in Theorem 1, $k = 2$ and (c) is replaced by:*
(c') A satisfies (1), (2), and (5),
then the resulting statement is true.

Proof. As before, (a) and (b) are equivalent, and imply that A satisfies (1), (2), (3), and (4). By Lemma 1, A also satisfies (5).

Now suppose that A satisfies (1), (2), and (5). By Lemma 1 A therefore satisfies (4), so, if $\mathcal{D}$ is formed as in the proof of Theorem 1, $\mathcal{D}$ has nodes (states) $s_1, \ldots, s_m$, with an arc from s_i to s_j if and only if $A(i,j) > 0$. It is straightforward that (ii) and (iii) hold. By (4), (2), and Lemma 2, for every $j \in \{1, \ldots, m\}$ there exists $i \in \{1, \ldots, m\}$ such that $A(i,j) > 0$ so (i) holds for $\mathcal{D}$.

It remains only to show that $\mathcal{D}$ is strongly connected; if it is, then $f(s_i s_j) = A(i,j)$ for all $i, j \in \{1, \ldots, m\}$ will follow as in the proof of Theorem 1.

Consider a strong component $\mathcal{U}$ of $\mathcal{D}$ which is a sink in the a cyclic digraph of strong components of $\mathcal{D}$. This means that there are no arcs leaving $\mathcal{U}$, i.e. going from nodes in $\mathcal{U}$ to nodes outside of $\mathcal{U}$. Let $V(\ )$ and $E(\ )$ stand for "vertex set of" and "arc set of", respectively; we have

$$\sum_{\{(i,j)|(s_i,s_j)\in E(\mathcal{U})\}} A(i,j) =$$
$$\sum_{\{i|s_i\in V(\mathcal{U})\}} \sum_{\{j|(s_i,s_j)\in E(\mathcal{U})\}} A(i,j) =$$

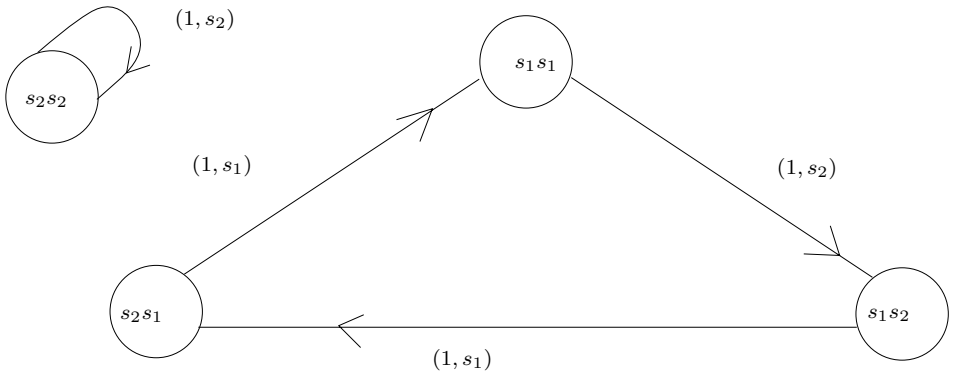
$$\sum_{\{i|s_i\in V(\mathcal{U})\}} \sum_{\{j|(s_i,s_j)\in E(\mathcal{D})\}} A(i,j) = \sum_{\{i|s_i\in V(\mathcal{U})\}} B(i)$$

$$= \sum_{\{j|s_j\in V(\mathcal{U})\}} \sum_{\{i|(s_i,s_j)\in E(\mathcal{D})\}} A(i,j)$$

$$= \sum_{\{(i,j)|(s_i,s_j)\in E(\mathcal{U})\}} A(i,j) + \sum_{\substack{\{(i,j)|s_i\notin V(\mathcal{U}), s_j\in V(\mathcal{U}) \\ \text{and } (s_i,s_j)\in E(\mathcal{D})\}}} A(i,j)$$

The second equality holds because $\mathcal{U}$ is a sink. To simplify: because A satisfies (2), the sum of the values of $A(i,j)$ over arcs $(s_i, s_j) \in E(\mathcal{U})$ is equal to the same sum plus the sum of $A(i,j)$ over arcs (s_i, s_j) of $\mathcal{D}$ coming into $\mathcal{U}$ from outside. Since $A(i,j) > 0$ for every arc $(s_i, s_j) \in E(\mathcal{D})$, it follows that there are no arcs coming into $\mathcal{U}$ from the outside, as well as no arcs leaving $\mathcal{U}$, by the choice of $\mathcal{U}$ as a sink. If $\mathcal{U}$ is not all of $\mathcal{D}$, then, taking $I = \{i \mid s_i \in V(\mathcal{U})\}$, we see that (5) is violated. Therefore, $\mathcal{U} = \mathcal{D}$ and so $\mathcal{D}$ is strongly connected. □

Example 2. If $k \geq 3$, conditions (1), (2), and (5) do not suffice for A to give the relative k-gram frequencies of a statistically stable source. Here is a small example, with $k = 3$ and $m = 2 : A(1,1,2) = A(1,2,1) = A(2,1,1) = A(2,2,2) = 1/4$ and $A(i_1, i_2, i_3) = 0$ for the other 4 elements of $\{1,2\}^3$. It is straightforward to see that A satisfies (1), (2), and (5), but not (3). The digraph $\mathcal{D}$ with labelled arcs associated with A as in the proof of Theorem 1 is:



Part of the proof of Theorem 2 can be adapted to the cases $k \geq 3$ to show that no counter-example to the conclusion of Theorem 2 for any $k \geq 3$ can be constructed in which the digraph $\mathcal{D}$ is connected, but not strongly connected. To be more precise, there is another theorem lurking here, which we state informally: If A satisfies (1), (2), and (4), and if the edge-labelled digraph $\mathcal{D}$ associated with A as in the proof of Theorem 1 is connected, then A gives the relative k-gram frequencies of some statistically stable source.

The big questions that we would like to tackle in the vague, distant future are: Under what conditions on $A : \bigcup_{k=1}^{\infty} \{1, \ldots, m\}^k \to [0, \infty)$ is there a statistically stable source with an m-letter alphabet with relative frequency function f such that $f(s_{i_1} \cdots s_{i_k}) = A(i_1, \ldots, i_k)$ for all k and for all $i_1, \ldots, i_k \in \{1, \ldots, m\}$? And when will there be a "solution" source among the pfssa's? The results of this paper may be of some small help in the attack on these questions. There are natural operations on pfssa's: see [1], [2], and [4]. Given an A, as above, satisfying natural necessary conditions to define a relative frequency function, the results here give an infinite sequence of pfssa's, $(\mathcal{D}_r)$, such that the restriction of the relative frequency function of $\mathcal{D}_r$ to $\bigcup_{k=1}^{r} S^k$ is given by the restriction of A to $\bigcup_{k=1}^{r} \{1, \ldots, m\}^k$. It feels as though Mother Nature is beckoning us to do something with the $\mathcal{D}_r$ to obtain, if not a pfssa, at least a statistically stable source with relative frequency function given by A.

References

[1] Michelle J. Foster, Operations on Probabilistic Finite State Source Automata, Ph.D. dissertation, Auburn University, August, 2000.

[2] Michelle Foster and Peter Johnson, Some operations on probabilistic finite state source automata, *Congressus Numerantium* 147 (2000), 53-64.

[3] Peter Lancaster and Miron Tismenetsky, *The Theory of Matrices with Applications*, Academic Press, 1985.

[4] Azaria Paz, *Introduction to Probabilistic Automata*, Academic Press, 1971.

[5] Claude Shannon, A mathematical theory of communication, *Bell System Technical Journal* 27 (July 1948), 379-423 and 623-656.

[6] Paul C. Shields, *The Ergodic Theory of Discrete Sample Paths*, Graduate Studies in Mathematics, Volume 13, American Mathematical Society, 1996.