THE FIVE GREATEST APPLICATIONS OF MARKOV CHAINS

PHILIPP VON HILGERS* AND AMY N. LANGVILLE†

Abstract. One hundred years removed from A. A. Markov's development of his chains, we take stock of the field he generated and the mathematical impression he left. As a tribute to Markov, we present what we consider to be the five greatest applications of Markov chains.

Key words. Markov chains, Markov applications, stationary vector, PageRank, HIdden Markov models, performance evaluation, Eugene Onegin, information theory

AMS subject classifications. 60J010, 60J20, 60J22, 60J27, 65C40

1. Introduction. The five applications that we have selected are presented in the boring but traditional chronological order. Of course, we agree that this ordering scheme is a bit unfair and problematic. Because it appears first chronologically, is A. A. Markov's application of his chains to the poem Eugeny Onegin the most important or the least important of our five applications? Is A. L. Scherr's application of Markov chains (1965) to the performance of computer systems inferior to Brin and Page's application to web search (1998)? For the moment, we postpone such difficult questions. Our five applications, presented in Sections 2-6, appear in the admittedly unjust chronological order. In Section 7, we right the wrongs of this imposed ordering and unveil our proper ordering, ranking the applications from least important to most important. We conclude with an explanation of this ordering. We hope you enjoy this work, and further, hope this work is discussed, debated, and contested.

2. A. A. Markov's Application to Eugeny Onegin. Any list claiming to contain the five greatest applications of Markov chains must begin with Andrei A. Markov's own application of his chains to Alexander S. Pushkin's poem "Eugeny Onegin." In 1913, for the 200th anniversary of Jakob Bernoulli's publication [4], Markov had the third edition of his textbook [19] published. This edition included his 1907 paper, [20], supplemented by the materials from his 1913 paper [21]. In that edition he writes, "Let us finish the article and the whole book with a good example of dependent trials, which approximately can be regarded as a simple chain." In what has now become the famous first application of Markov chains, A. A. Markov, studied the sequence of 20,000 letters in A. S. Pushkin's poem "Eugeny Onegin," discovering that the stationary vowel probability is $p = 0.432$, that the probability of a vowel following a vowel is $p_1 = 0.128$, and that the probability of a vowel following a consonant is $p_2 = 0.663$. In the same article, Markov also gave the results of his other tests; he studied the sequence of 100,000 letters in S. T. Aksakov's novel "The Childhood of Bagrov, the Grandson." For that novel, the probabilities were $p = 0.449$, $p_1 = 0.552$, and $p_2 = 0.365$.

At first glance, Markov's results seem to be very specific, but at the same time his application was a novelty of great ingenuity in very general sense. Until that time, the theory of probability ignored temporal aspects related to random events. Mathematically speaking, no difference was made between the following two events: a die

*Max Planck Institute for History of Science, Berlin, Germany (philgers@mpiwg-berlin.mpg.de)
†Department of Mathematics, College of Charleston, Charleston, SC 29424, (langvillea@cofc.edu)

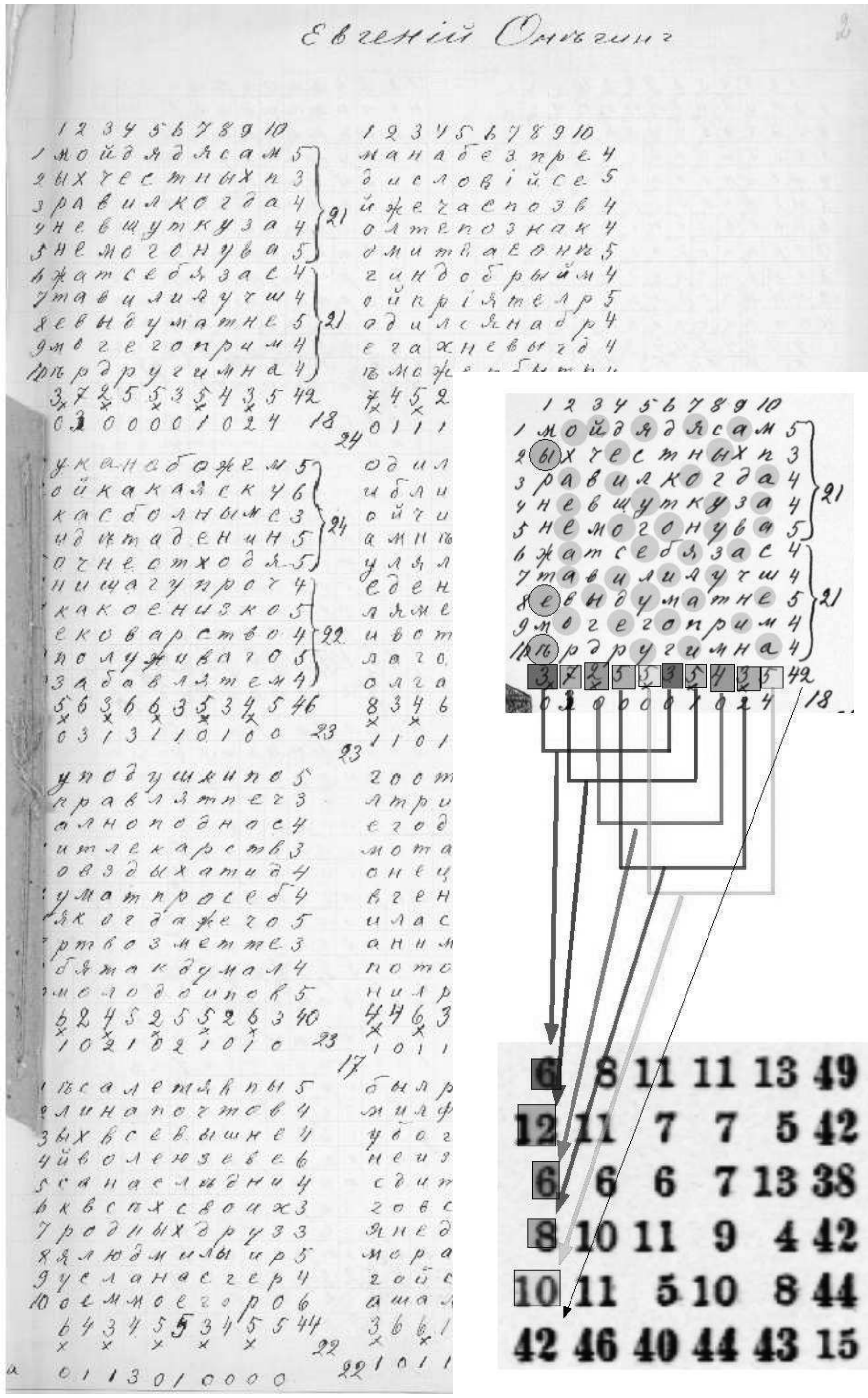

FIG. 2.1. *Left background: The first 800 letters of 20,000 total letters compiled by Markov and taken from the first one and a half chapters of Pushkin's poem "Eugeny Onegin." Markov omitted spaces and punctuation characters as he compiled the cyrillic letters from the poem. Right foreground: Markov's count of vowels in the first matrix of 40 total matrices of* 10×10 *letters. The last row of the* 6×6 *matrix of numbers can be used to show the fraction of vowels appearing in a sequence of 500 letters. Each column of the matrix gives more information. Specifically, it shows how the sums of counted vowels are composed by smaller units of counted vowels. Markov argued that if the vowels are counted in this way, then their number proved to be stochastically independent.*

thrown a thousand times versus a thousand dice thrown once each. Even dependent random events do not necessarily imply a temporal aspect. In contrast, a temporal aspect is fundamental in Markov's chains. Markov's novelty was the notion that a random event can depend only on the most recent past. When Markov applied his model to Pushkin's poem, he compared the probability of different distributions of letters taken from the book with probabilities of sequences of vowels and consonants in term of his chains. The latter models a stochastic process of reading or writing while the former is simply a calculation of the statistical properties of a distribution of letters. Figure 2.1 shows Markov's original notes in computing the probabilities needed for his Pushkin chain. In doing so, Markov demonstrated to other scholars a method of accounting for time dependencies. This method was later applied to the diffusion of gas molecules, Mendel's genetic experiments, and the random walk behavior of certain particles.

The first response to Markov's early application was issued by a colleague at the Academy of Sciences in St. Petersburg, the philologist and historian Nikolai A. Morozov. Morozov enthusiastically credited Markov's method as a "new weapon for the analysis of ancient scripts" [24]. To demonstrate his claim Morozov himself provided some statistics that could help identify the style of some authors. In his typical demanding, exacting, and critical style [3], Markov found few of Morozov's experiments to be convincing. Markov, however, did mention that a more advanced model and an extended set of data might be more successful at identifying an author solely by mathematical analysis of his writings [22].

3. C. E. Shannon's Application to Information Theory. When Claude E. Shannon introduced "A Mathematical Theory of Communication" [30] in 1948, his intention was to present a general framework for communication based on the principles of the new digital media. Shannon's information theory gives mathematically formulated answers to questions such as how analog signals could be transformed into digital ones, how digital signals then could be coded in such way that noise and interference would not do harm to the original message represented by such signals, and how an optimal utilization of a given bandwidth of a communication channel could be ensured. A famous *entropy* formula associated with Shannon's information theory is $H = -(p_1 \log_2 p_1 + p_2 \log_2 p_2 + \ldots + p_n \log_2 p_n)$, where H is the amount of information and p_i is probability of occurrence of the states in question. This formula is the entropy of a source of discrete events. In Shannon's words this formula "gives values ranging from zero—when one of the two events is certain to occur (i.e., probability 1) and all others are certain not to occur (i.e., probability 0)—to a maximum value of $\log_2 N$ when all events are equally probable (i.e., probability $\frac{1}{N}$). These situations correspond intuitively to the minimum information produced by a particular event (when it is already certain what will occur) and the greatest information or the greatest prior uncertainty of the event" [31].

It is evident that if something is known about a message beforehand, then the receiver in a communication system should somehow be able to take advantage of this fact. Shannon suggested that any source transmitting data is a Markov process. This assumption leads to the idea of determining a priori the transition probabilities of communication symbols, i.e., the probability of a symbol following another symbol or group of symbols. If, for example, the information source consists of words of the English language and excludes acronyms, then the transition probability of the letter "u" following the letter "q" is 1.

Shannon applied Markov's mathematical model in a manner similar to Markov's own first application of "chains" to the vowels and consonants in Alexander Pushkin's poem. It is interesting to pause here to follow the mathematical trail from Markov to Shannon. For further details about the theory of Markov chains, Shannon referred to a 1938 book by Maurice Fréchet [7]. While Fréchet only mentions Markov's own application very briefly, he details an application of Markov chains to genetics. Beyond Fréchet's work, within the mathematical community Markov chains had become a prominent subject in their own right since the early 1920s, especially in Russia. Most likely Fréchet was introduced to Markov chains by the great Russian mathematician Andrei Kolmogorov. In 1931 Kolmogorov met Fréchet in France. At this time Kolmogorov came to France after he had visited Göttingen, where he had achieved fame particularly due to his axiomatization of probability theory [14].

Kolmogorov speculated that if physics in Russia had reached the same high level as it had in some Western European countries, where advanced theories of probability tackled distributions of gas and fluids, Markov would not have picked Pushkin's book from the shelf for his own experiments [33]. Markov's famous linguistic experiment might have been a physics experiment instead. It is significant that Kolmogorov himself contributed an extension to the theory of Markov chains by showing that "it is a matter of indifference which of the two following assumptions is made: either the time variable t runs through all real values, or only through the integers" [15].[1] In other words, Kolmogorov made the theory suitable not only for discrete cases, but also for all kinds of physical applications that includes continuous cases. In fact, he explicitly mentioned Erwin Schrödinger's wave mechanics as an application for Markov chains.

With this pattern of prior applications of Markov theory to physical problems, it is somewhat ironic that Shannon turned away from physics and made great use of Markov chains in the new domain of communication systems, processing "symbol by symbol" [30] as Markov was the first to do. However, Shannon went beyond Markov's work with his information theory application. Shannon used Markov chains not solely as a method for analyzing stochastic events but also to generate such events.

Shannon demonstrated that a sequence of letters, which are generated by an increasing order of overlapping groups of letters, is able to reflect properties of a natural language. For example, as a first-order approximation to an English text, an arbitrary letter is added to its predecessor, with the condition that this specific letter be generated according to its relative frequency in the written English language in general. The probability distribution of each specific letter can, of course, be acquired only by empirical means. A second-order approximation considers not only the distribution of a single letter but a bigram of two letters. Each subsequent letter becomes the second part of the next bigram. Again, the transition probabilities of these overlapping bigrams are chosen to match a probability distribution that has been observed empirically. By applying such a procedure Shannon generated a sequence of order 3 by using trigrams. Shannon's first attempt at using Mar! kov chains to produce English sentences resulted in "IN NO IST LAT WHEY CRATICT FOURE BIRS GROCID" [30]. Although this line makes no sense, nevertheless, some similarities with written English are evident. With this start, Shannon felt that communication systems could be viewed as Markov processes, and their messages analyzed by means of Markov

[1] Kolmogorov saw the potential of Markov theory in the physical domain. And later, he was among the first to promote information theory and further develop it into an algorithmic information theory. In addition, due to translation issues, Kolmogorov's contribution to the mathematization of poetry is scarcely known outside of Russia.

theory. In studying artificial languages Shannon distinguished between ergodic and non-ergodic processes. Shannon used the phrase "ergodic process" to refer to the property that all processes originating from the same source have the same statistical properties.

While Shannon introduced his generative model for technical reasons, other scholars became convinced that Markov models could even play a far more general role in the sciences and arts. In fact, Shannon, himself helped popularize information theory in other diverse fields. For instance, Shannon and his colleague David W. Hagelbarger created a device that could play the game of "Even and Odd." It is reported that Shannon's so-called "Mind-Reading Machine" won most of the games against visitors at the Bell Labs who dared to challenge the machine to a match. Despite the whimsical nature of Shannon's machine, it deserves credit as being the first computer to have implemented a Markov model [32]. At this same time, the ever well-informed French psychoanalyst Jacques Lacan introduced Markov chains as the underlying mechanism for explaining the process by which unconscious choices are made. Lacan hints that Shannon's machine was the model for his theory [16, 17].

Later researchers continued Shannon's trend of using computers and Markov chains to generate objects from text to music to pictures. Artists such as musician Iannis Xenakis developed "Free Stochastic Music" based on Markov chains, and early media artists and computer experts such as Frieder Nake plotted pictures generated by Markov models. In fact, a group of influential linguists claimed that the modus operandi of language is a Markov process [12]. Such a general assumption provoked a controversial debate between Roman Jakobson and Noam Chomsky. Chomsky argued that language models based on Markov chains do not capture some nested structures of sentences, which are quite common in many languages such as English [10]. We now recognize that Chomsky's account of the limitations of Markovian language models was too harsh. The early 1980s saw a resurgence of the success of Markovian language models in speech recognition. Section 5 treats this era briefly.

4. A. L. Scherr's Application to Computer Performance Evaluation. In 1965 Allan L. Scherr completed his thesis, "An Analysis of Time-Shared Computer Systems," and received his Ph.D. in electrical engineering from M.I.T. At the time, the Compatible Time-Sharing System was new to the M.I.T. campus and allowed 300 users to interactively access the computer and its software. The goal of Scherr's thesis was to characterize the system's usage. He conducted simulation studies to predict the system's usage and wanted to compare this with real user data from similar systems. He found that no such data existed, so he conducted his own comprehensive measurements of system performance. Scherr declared his analysis of time-shared systems complete after he obtained his own real data and compared this with his simulation results.

At this point in the story, many of us can thank one of Scherr's thesis advisors for our current careers and research field. This particular advisor complained that Scherr's thesis was not yet complete as it wasn't quite academic enough. Scherr recalls that "there weren't enough mathematical formulas in it" [8]. So in response to this complaint, Scherr hobbled together a "very quick and dirty" mathematical analysis by applying a method from a recent operations research course he had taken. He used a continuous-time Markov chain to model M.I.T's Compatible Time-Sharing System. The chain not only added enough mathematics, it also led Scherr to a surprising result. Scherr's quick and dirty measure gave a very good approximation to system performance. According to Scherr, this was surprising because "this very simple, analytic model that ignored 99 percent of the details of the system was just as accurate in predicting performance and capacity as the very e! laborate, very accurate simulation model" [8].

Scherr's single-server, single-queue Markov chain captured the dominating aspects of the system. The tridiagonal transition rate matrix $\mathbf{A}$ appearing in Scherr's thesis is below. The states of the chain are the number of users interacting with the system, and thus, are labeled $\{0, 1, \ldots, n\}$.

$$\mathbf{A} = \begin{pmatrix} -\frac{n}{T} & \frac{n}{T} & & & & & \\ \frac{1}{P} & -(\frac{1}{P}+\frac{n-1}{T}) & \frac{n-1}{T} & & & & \\ & \frac{1}{P} & -(\frac{1}{P}+\frac{n-2}{T}) & \frac{n-2}{T} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \frac{1}{P} & -(\frac{1}{P}+\frac{2}{T}) & \frac{2}{T} & \\ & & & & \frac{1}{P} & -(\frac{1}{P}+\frac{1}{T}) & \frac{1}{T} \\ & & & & & \frac{1}{P} & -\frac{1}{P} \end{pmatrix}.$$

Scherr defined n as the number of users in the system. When a user interacted with the time-shared computer system, Scherr defined T as the mean time for the console portion of the interaction and P as the mean processor time per interaction. Scherr then solved the system $\boldsymbol{\pi}^T\mathbf{A} = \mathbf{0}$ for the stationary probability vector $\boldsymbol{\pi}$. Using standard queuing theory techniques, Scherr reported π_0, the long-run probability that the M.I.T. shared processor was idle, $\bar{Q} = \sum_{i=1}^{n} i\pi_i$, the average number of users waiting for service, and W, the mean response time of the system. Scherr concluded that "the think times of the users and the service times of the equipment were really all you needed to know to do a pretty accurate characterization of the response time and capacity of the system" [8].

In the years after his dissertation, researchers argued over the validity of Scherr's simple Markov analysis. Were Scherr's results—that a simple Markov model came to within 4-5 percent accuracy of the very detailed simulation—merely one atypical study or was simple analysis preferable to detailed simulation? A few decades after his dissertation, Scherr, then at IBM, had an answer to that question, an answer arrived at through practice. Scherr discovered that system updates and upgrades are the primary reason for the superiority of the simplified analytical model over detailed simulation. The simple analytical models give a better return on investment. By the time a detailed simulation is finally complete, the system is usually updated, and a new simulation required. Also, Scherr found that the simple Markov models force designers to get to the heart of the system, because only the essential information can be modeled; all extraneous information must be identified and omitted.

By the way, ten years later the performance evaluation community, which evolved as a result of Scherr's 1965 thesis [29], noted Scherr's ground-breaking contribution. In 1975, his thesis was awarded the Grace Murray Hopper Award by the Association for Computing Machinery. Each year this award is presented to a young computer professional, aged 35 or less, who has made a "single significant technical or service contribution" to the computer field. At present, there are dozens of journals and conferences devoted to computer performance evaluation. Papers from these journals and proceedings demonstrate the advancement achieved since Scherr's development of this new field. These papers describe Markov models for computer performance evaluation that have expanded far beyond Scherr's famous single-server, single queue model.

5. L. E. Baum's Application to Hidden Markov Models. When Lawrence R. Rabiner introduced Hidden Markov Models (HMMs) in the widely read proceedings of the IEEE in 1989, he stated that even though most of their principles had been known for twenty years, engineers had failed to read mathematical journals to grasp their potential and mathematicians had failed to "provide sufficient tutorial material for most readers to understand the theory and to be able to apply it to their own research" [28]. The purpose of Rabiner's paper was to rectify this under-appreciation of an important tool.

Rabiner might have overstated the case a bit as HMMs were receiving some appreciation prior to 1989. In 1980 a symposium "On the application of Hidden Markov Models to Text and Speech" was held in Princeton for the reason that it "was thought fitting to gather researchers from across the country" for one day [9]. It was at this gathering, that Lee P. Neuwirth coined the phrase "Hidden Markov Model," instead of calling it the bit unwieldy alternative of "probabilistic functions of Markov chains" [25, 26]. It is worth mentioning that Lee P. Neuwirth was the longtime director of the Communications Research Division within the Institute for Defense Analysis (IDA-CRD) in Princeton, New Jersey. Some called this institute "the most secret of the major think tanks" of the day [1].

Today, now that speech recognition software is available off the shelf, it is no longer a secret that, thanks to HMMs, "word spotting" in stream of spoken language is done by a computer—even in the presence of noise. In addition, speech cannot only be recognized algorithmically with great accuracy, but the individual voice of the speaker can be identified as well. It stands to reason that such features are of great interest, especially for intelligence services, in order to scan communication channels for key words. Moreover, HMMs enable the extraction of significant information from the acoustical pattern of a spoken language without requiring any semantic knowledge of the language in question.

John D. Ferguson, also a mathematician at the IDA, specified the theoretical impact of HMMs by solving three fundamental problems in the proceedings of that 1980 IDA symposium.

1. Compute the probability of an observed sequence based on a given model.
2. Maximize the probability of an observed sequence by adjusting the parameters using, for instance, the "Baum-Welsh" algorithm.
3. Select the most likely hidden states, given the observation, and the model.

By listing these problems Ferguson provided a building block for later tutorials on HMMs [9, 26, 28]. The meaning of these rather abstract instructions might become clearer with an example taken from Alan B. Poritz's instructive paper "Hidden Markov Models: A Guided Tour" [26]. This example is illustrated in Figure 5.1. Imagine three mugs, each containing its own mixture of stones. The stones are marked either "state 1" or "state 2." Somebody randomly draws a stone from mug "0." If the stone is marked "state 1," as shown in the illustrated example, a ball is drawn from urn "1." Otherwise, a ball is drawn from the other urn. The two urns are filled with a different mixture of black and white balls. After every draw, the ball will be replaced in the urn from which it was selected. An observer only gets to know the outcome of the drawing in terms of a sequence of symbols: "BWBWWB", where"B" stands for a black ball and "W" for a white ball.[2] Notice that the particular urn from which a ball

[2]It is interesting that Ferguson and Neuwirth were not the first scientists to consider urn problems in the Markov context. Markov and his colleague A. A. Chuprov had a heated debate over the notion of "double probability," the phrase Markov gave to a random selection between two or more urns for

is drawn is not available to the observer, only the color is. Hence the label "hidden" in the HMM acronym. An HMM is valuable because it helps uncover this hidden information, or at least gives a reasonable approximation to it.

Let's start with most difficult problem, recovering a mixture model from the observed tokens, thereby solving the second problem. In the given example, the mixture model corresponds to guessing each urn's mixture. In the late 1960s and particularly the early 1970s, Leonard E. Baum, supported by the work of other scholars, demonstrated that the underlying model can be recovered from a sufficiently long observation sequence by an iterative procedure, which maximizes the probability of the observation sequence. This procedure is called Baum-Welsh algorithm. It is based on the auxiliary Q-function, which is denoted $Q(\lambda, \bar{\lambda}) = \sum_{s \in S} P_\lambda(O, s) \log P_{\bar{\lambda}}(O, s)$, where s is a element out of set of states S and $\bar{\lambda}$ is an exstimated model of λ. In an iterative procdure $\bar{\lambda}$ becomes λ until there is no further improvement in terms of a measurement of increased probabiltiy. Baum and his collegues proved that maximization of $Q(\lambda, \bar{\lambda})$ increased likelihood of the estimated $\bar{\lambda}$ and a hidden true model, i.e. $max_{\bar{\lambda}}[Q(\lambda, \bar{\lambda})] \Rightarrow P(Q|\bar{\lambda} \geq P(Q|\lambda)$ [2, 26, 28]. Second, from this model the hidden state sequence can then be estimated, thereby solving the third problem. A technique for doing this is called the Viterbi Algorithm. This algorithm estimates the best state sequence, $Q = (q_1 q_2 \cdots q_T)$, for the given observation sequence $O = (O_1 O_2 \cdots O_T)$ by defining the quantity $\delta(i) = max_{q_1, q_2, \cdots, q_{t-1}} P[q_1 q_2 \cdots q_t = i, O_1 O_2 \cdots O_t|\lambda]$. In our example, the outcome of the unobserved draws of stones from the mug generated the hidden state sequence. Eventually, once the parameters of the hidden model denoted by the vector $\lambda = (A, B, \pi)$ are known, the probability of the observed sequence can be computed, thereby solving the first problem. Instead of calculating the probability of the observed sequence by finding each possible sequence of the hidden states and summing these probabilities, a short cut exists. It is called the forward algorithm and minimizes computational efforts from exponential growth to linear growth by calculating partial probabilities at every time step in a recursive manner [26].

To summarize, hidden Markov models require the solution of some fundamental, interrelated problems of modeling and analyzing stochastic events. Several mathematicians (only a few are mentioned in this brief survey) defined the main problems and developed different techniques to solve them. Once tutorial material spread among scholars from other fields, HMMs quickly demonstrated their potential through real-life applications. For instance, in 1989 Gary A. Churchill used HMMs to separate genomic sequences into segments [11]. Since then, HMMs have become an off-the-shelf technology in biological sequence analysis.

We close this section by returning once again to Andrei A. Markov. Even though Markov died in St. Petersburg in 1922, his work on his famous chains played a prominent role in 1980 in Princeton's secret think tank. This prominence is apparent in the bibliographies of the proceedings of the IDA symposium that year. In fact, the proceedings open with an article by Lee P. Neuwirth and Robert L. Cave that states explicitly their intention to examine the relation of HMMs to the English language "in the spirit of Markov's application"[25], referring to Eugeny Onegin. The first result of their experiment showed that the most simple model of English text provides only two states: vowel and consonant. The HMM's ability to distinguish vowels from consonants was their main result. In this experiment, Neuwirth and Cave analyzed the letters of an English text without letting the HMM know which were vowels and

---

the drawing of lots. Markov was more skeptical than Chuprov that such a scheme could lead to new results [27].

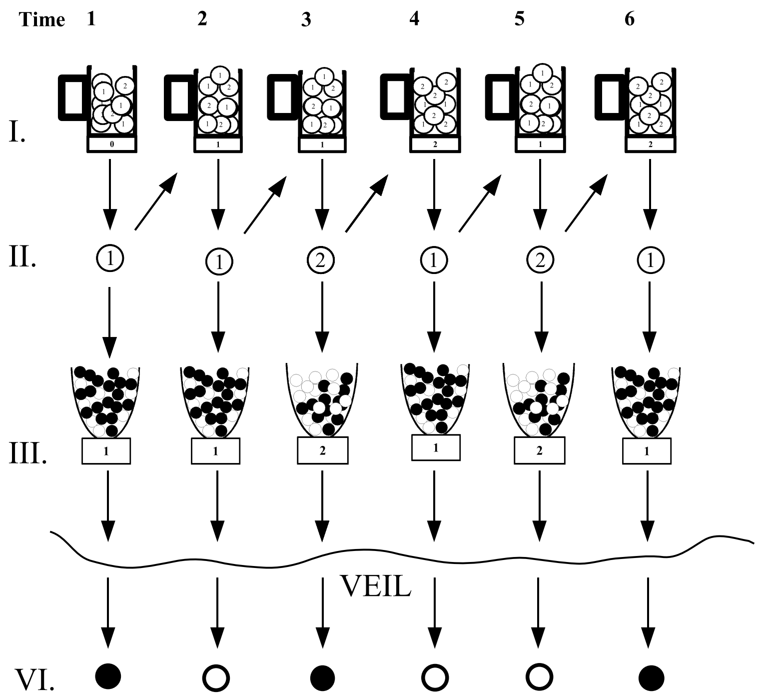

FIG. 5.1. *Sampling of the Urns veils the sampling of a sequence of Mugs.*

I. Let a_{01} be the fraction of stones with the state 1 marking in the first mug, and a_{02} the fraction of stones with the state 2 marking so that $a_{01}+a_{02}=1$. Similarly, the fraction of stones in the other two mugs are denoted a_{21}, a_{22} in one case and a_{31}, a_{32} in the other.

II. At each time $t = 1, 2, \ldots, T$, a stone is selected. The number of selections in the example above is T= 6, while the outcome of these unobserved random selections determines the sequence of states s=(1,1,2,1,2,1).

III. Say $b_1(B)$ is the fration of white balls, $b_1(W)$ the fraction of black balls in one of the two urns and analog $b_2(B)$,$b_2(W)$ in case of the other urn so that for each state s, the vector $b = (b_s(1), b_s(2), \cdots b_s(K))$ is the output probability vector of a HMM with finite output alphabet with K elements. In this example the number of elements in the alphabet is $K = 2$.

VI. The T-long observation sequence is O=(B,W,B,W,W), while $|B| = 3$ is the number of drawn balls that are black and $|W| = 3$ is the number of drawn balls that are white.

The parameter vector which describes the HMM is
$\lambda = (a_{01}, a_{02}, a_{21}, a_{22}, a_{31}, a_{32}, b_1(B), b_1(W), b_2(B), b_2(W))$. P_λ stands for probability density associated with the model. It also has become common to denote the parameter model in general by the vector $\lambda = (A, B, \pi)$, where π is the vector of the initial state probabilities, and thus, is equivalent to a_{01}, a_{02}.

which were consonants. In hindsight, Markov has to be admired for his intuition to focus on the vowel-consonant distinction, which proved to be very significant four score later. Markov was forced to stop his letter-counting experiments, when he had nearly completely lost his sight due to glaucoma. Even if Markov had had more time and better eyesight to carry his experiments further, such extensions would have been very difficult to complete, given the precomputer era he lived in, when computational efforts had to be paid in man-years.

6. S. Brin and L. Page's Application to Web Search. In the late 1990s, Sergey Brin and Larry Page, then graduate students at Stanford University, were working on their PageRank project to organize the World Wide Web's information. Their 1998 paper, "PageRank: Bringing Order to the Web" [6], contributed to the order Google brought to web search. Brin and Page saw the potential of their PageRank idea, took a leave of absence from Stanford, and formed Google in 1998.

By Brin and Page's own admission, PageRank, which is the stationary vector of an enormous Markov chain, is the driving force behind Google's success in ranking webpages. In the PageRank context, the web and its hyperlinks create an enormous directed graph. Brin and Page's vision of a web surfer taking a random walk on this graph led to the formulation of the world's largest Markov chain [23]. The Google Markov chain $\mathbf{G}$ is defined as a convex combination of two other chains: $\mathbf{H}$, a reducible Markov matrix defined by the web's hyperlinks, and $\mathbf{E}$, a completely dense rank-one Markov matrix. Specifically,

$$\mathbf{G} = \alpha\mathbf{H} + (1-\alpha)\mathbf{E},$$

where $0 < \alpha < 1$ is a parameter that Brin and Page originally set to .85, $H_{ij} = 1/O_i$ if page i links to page j, and 0, otherwise, and O_i is the number of outlinks from page i. $\mathbf{E} = \mathbf{e}\mathbf{v}^T$, where $\mathbf{v}^T > \mathbf{0}$ is the so-called personalization vector. The idea is that for $\alpha = .85$, 85% of the time surfers follow the hyperlink structure of the web and 15% of the time they jump to a new page according to the distribution in the personalization vector $\mathbf{v}^T$. For the 6-node web example in Figure 6.1, using $\alpha = .85$ and $\mathbf{v}^T = 1/6\,\mathbf{e}^T$, the $\mathbf{H}$ and $\mathbf{G}$ matrices are below.

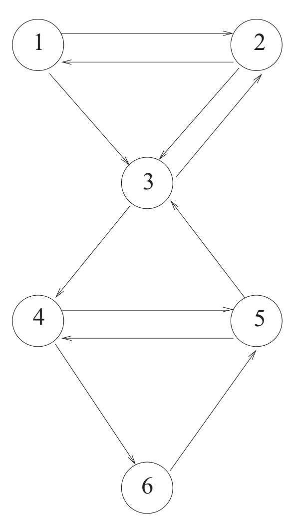

FIG. 6.1. *6-node web graph*

$$\mathbf{H} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \quad \text{and}$$

$$\mathbf{G} = \begin{pmatrix} 0.025 & 0.450 & 0.450 & 0.025 & 0.025 & 0.025 \\ 0.450 & 0.025 & 0.450 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.450 & 0.025 & 0.450 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.450 & 0.450 \\ 0.025 & 0.025 & 0.450 & 0.450 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.875 & 0.025 \end{pmatrix}.$$

$\mathbf{G}$ has been carefully formed so that it is an aperiodic, irreducible Markov chain, which means that its stationary vector $\boldsymbol{\pi}^T$ exists and is unique. Element π_i represents the long-run proportion of the time that the random surfer spends in page i. A page with a large stationary probability must be important as it means other important pages point to it, causing the random surer to return there often. The webpages returned in response to a user query are ranked in large part by their π_i values [5]. In the above example, the pages would be ordered from most important to least important as $\{3, 5, 4, 2, 6, 1\}$ since $\boldsymbol{\pi}^T = (\, 0.092 \quad 0.158 \quad 0.220 \quad 0.208 \quad 0.209 \quad 0.113 \,)$. This ranking idea is so successful that nearly all search engines use some hyperlink analysis to augment their retrieval systems.

$\boldsymbol{\pi}^T$ was originally computed using the simple power method [6], but since then a variety of methods for accelerating this computation have been proposed [18]. Judging by the hundreds of papers produced in the last few years on PageRank computation, without doubt PageRank presents a fun computational challenge for our community.

7. Ranked List. It's now time to present these five great Markov applications in their proper order. Employing a common trick among late night television hosts, we present the list in reverse order of importance, as this serves to build the anticipation and excitement inherent in all top k lists.

5. Scherr's application to Computer Performance Evaluation
4. Brin and Page's application to PageRank and Web Search
3. Baum's application to Hidden Markov Models[3]
2. Shannon's application to Information Theory
1. Markov's application to Eugeny Onegin

How did we arrive at this ordering? Good question. We considered several popular schemes for ordering elements in a set. We've already discussed the drawbacks of the chronological system. For similar reasons, the alphabetical system seemed equally unfair and unjustified. For instance, should we alphabetize by author of the application or name of the application? And what if an application was produced by two authors, as in the PageRank case? Do we order by Brin or by Page?

Having ruled out such elementary systems, we moved on to more advanced systems, such as the "Markov number" system and the "impact factor" system. The Markov number system is an analogy to the Erdos number system. In this system,

[3]Actually, there was a tie between Shannon and Baum for second place. See the next page for details.

A. A. Markov has a Markov number of 0. Markov applications produced by Markov's coauthors would receive a Markov number of 1, coauthors of coauthors, a 2, and so on. This means that Markov's own Eugeny Onegin application would be the most important of our five chosen applications, because Markov numbers are like golf scores, the lower, the better. This ranking system seemed reasonable, but it wasn't without its faults. For instance, could we find a coauthor of Scherr's who coauthored a paper with Markov? Maybe, but we admit that we didn't try very hard. This left us with a Markov number for only Markov and no Markov numbers for Shannon, Scherr, Baum, Brin, or Page.

So we proceeded to the "impact factor" system of ranking our five applications. The idea here is to rank each application according to its perceived impact on society. A laudable goal, but practically difficult. Should applications that created entire new fields, such as Shannon's information theory and Scherr's computer performance evaluation receive the highest impact factor? If so, where does that leave PageRank? While the PageRank application hasn't created a new field (yet), it does affect billions of people each day as they access Google to surf the web.

Then inspiration came to us. The proper ranking system should have been self-evident. PageRank is the most famous system for ranking items, so why not use PageRank to rank our five applications? However, this idea gave cause for caution. Could PageRank be used to rank items, one of which is PageRank itself? The idea sounded dangerous, and worth avoiding. So we sidestepped the problem by employing a PageRank-like system, called HITS [13], to rank our five applications. In order to apply the HITS system for ranking items we needed to create a graph that represented the relationships between our five Markov applications. We produced the graph in Figure 7.1. We created a link between applications if we perceived some connection between the two topics. For instance, there is a link between Information Theory and Eugeny Onegin because Information Theory uses Markov chains to capture the stochastic relationship between letters and words and Eugeny Onegin uses chains to capture the stochastic relationship between vowels and consonants. Similarly, the link from PageRank to Performance Evaluation represents the fact that Scherr's models can be used to evaluate the massive parallel machines used to compute PageRank.

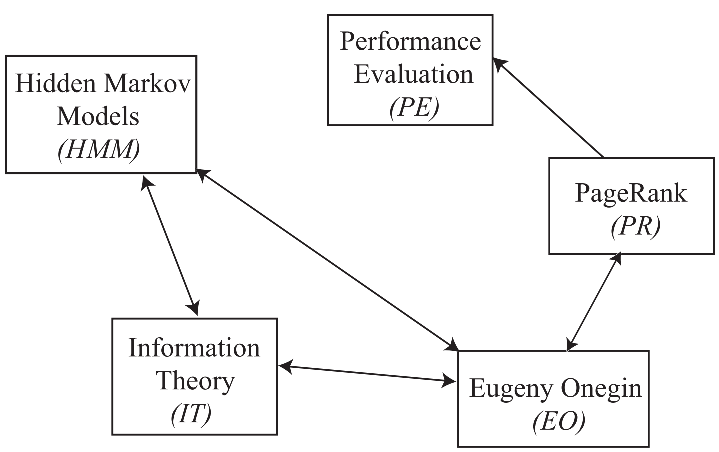

Fig. 7.1. *Graph of relationships between five Markov applications*

The adjacency matrix $\mathbf{L}$ associated with Figure 7.1 is

$$\mathbf{L} = \begin{array}{c} \\ EO \\ IT \\ PE \\ HMM \\ PR \end{array} \begin{array}{c} \begin{array}{ccccc} EO & IT & PE & HMM & PR \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array},$$

and the HITS authority vector $\mathbf{a}$ (the dominant eigenvector of $\mathbf{L}^T\mathbf{L}$) is

$$\mathbf{a} = \begin{array}{c} \begin{array}{ccccc} EO & IT & PE & HMM & PR \end{array} \\ \begin{pmatrix} .64 & .50 & .17 & .50 & .26 \end{pmatrix} \end{array},$$

meaning that the most authoritative of our five applications is Markov's application to Eugeny Onegin and the least authoritative is Scherr's application to computer performance evaluation.

Now that the mystery behind our ranking of the five greatest applications of Markov chains has been revealed, do you agree with us? Since you are all fellow Markov enthusiasts, we expect little argument about A. A. Markov's claim to the first position. It is, after all, his 150th birthday that we are celebrating. But what about the other four positions. We await your suggestions and complaints by email. Regardless, Happy 150th Markov!

REFERENCES

[1] James Bamford. The Puzzle Palace. A Report on America's Most Secret Agency. Houghton Mifflin Company, Boston, 1982.

[2] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, 41:164-171, 1970.

[3] Gely P. Basharin, Amy N. Langville, and Valeriy A. Naumov. The Life and Work of A. A. Markov. *Linear Algebra and its Applications*, 386: 3-26, 2004.

[4] Jakob Bernoulli. *Ars Conjectandi*, Opus Posthumum, Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis, Basileae, 1713 (Ch. 1-4 translated into English by B. Sung, *Ars Conjectandi*, Technical Report No. 2, Dept. of Statistics, Harvard University, 1966).

[5] Nancy Blachman, Eric Fredricksen, Fritz Schneider. *How to Do Everything with Google*. McGraw-Hill, 2003.

[6] Sergey Brin, Lawrence Page, R. Motwami, Terry Winograd. The PageRank citation ranking: Bringing order to the Web. Technical Report 1999-0120, Computer Science Department, Stanford University, 1999.

[7] Maurice Fréchet. Méthode des fonctions arbitraires. Théorie des énénements en chaine dans les cas d'un nombre fini d'états possibles. Gauthier-Villars, Paris, 1938.

[8] Karen A. Frenkel. Allan L. Scherr: Big Blue's Time-Sharing Pioneer. *Communications of the ACM*, 30(10): 824-828, 1987.

[9] John D. Ferguson. Hidden Markov Models for Language. *Symposium on the Application of Hiddden Markov Models to Text and Speech*, IDA-CRD, Princeton, 1980, 1-7.

[10] Noam Chomsky. Three models for the description of language. *IRE Transaction of Information Theory*, 2(3):113-124, 1956.

[11] Gary A. Churchill. Stochastic models for heterogeneous DNA sequences. *Bull Math Biol.* 51, 1989, 79-94.

[12] Colin Cherry, Morris Halle, and Roman Jakobson. Toward the logicaldescription of languages in their phonemic aspect. *Language*, 29 (1): 34-46, 1953.

[13] Jon Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*. 46, 1999.

[14] Andrey N. Kolmogorov. Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin, 1933.
[15] Andrey N. Kolmogorov. Zur Theorie der Markoffschen Ketten. *Math. Ann.* 112, 1936, 155-160. (Translated into English by G. Lindquist, *Selected Works of A.N. Kolmogorov*, Vol. 2, Probability Theory and Mathematical Statistics. Kluwer, Dordrecht, Boston, London,1986, 182-192.
[16] Jacques Lacan. Écrits. Le champ freudien. (Ch. 1 *Le séminare sur "la Lettre volée"*) Éditions du Seuil, Paris, 1966.
[17] Jacques Lacan. Le Séminare. Livre II. Le moi dans la théorie de Freud et dans la technique de la psychanalyse. (Ch. 15 *Pair ou impair? Au-delà de l'intersubjectivitivé.*) Éditions du Seuil, Paris, 1978.
[18] Amy N. Langville and Carl D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, Princeton, 2006.
[19] Andrei A. Markov. *Ischislenie veroyatnostej*, SPb, 1900; 2-e izd., SPb, 1908, Translated into German, *Wahrscheinlichkeitsrechnung*, Teubner, Leipzig-Berlin, 1912; 3-e izd., SPb, 1913; 4-e izd., Moskva, 1924.
[20] Andrei A. Markov. Issledovanie zamechatel'nogo sluchaya zavisimyh ispytanij, *Izvestiya Akademii Nauk*, SPb, VI seriya, tom 1, 9 3, 1907, 61-80 (Translated into French, Recherches sur un cas remarquable d'epreuves dependantes, *Acta mathematica*, Stockholm, 33, 1910, 87-104).
[21] Andrei A. Markov. Primer statisticheskogo issledovaniya nad tekstom "Evgeniya Onegina", illyustriruyuschij svyaz' ispytanij v cep', *Izvestiya Akademii Nauk*, SPb, VI seriya, tom 7, 9 3, 1913, 153-162.
[22] Andrei A. Markov. Ob odnom primenenii statisticheskogo metoda (On some application of statistical method).*Izvestija Imp. Akademii nauk*, serija VI,4:239-242, 1916.
[23] Cleve Moler. The world's largest matrix computation. *Matlab News and Notes*. Oct. (10): 12-13, 2002.
[24] N. A. Morozov, N. A. Lingvisticheskie spektry (Linguistic spectra).*Izvestija Akademii Nauk*, Section of Russian Language,20:1-4, 1915.
[25] Robert L. Cave, Lee P. Neuwirth. Hidden Markov Models for English. *Symposium on the Application of Hiddden Markov Models to Text and Speech*, IDA-CRD, Princeton, 1980, 1-7.
[26] Alan B. Poritz. Hidden Markov Models: A Guided Tour. *Proc. from ICASSP – International Conference on Acoustics, Speech, and Signal Processing*, 1988, 7-13.
[27] Kh. O. Ondar (Ed.). The Correspondence Between A.A. Markov and A.A. Chuprov on the Theory of Probability and Mathematical Statistics. Springer, New York, Heidelberg, Berlin, 1982.
[28] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. of the IEEE*, Vol. 77, 2, 1989, 257-286.
[29] Allan Lee Scherr. An Analysis of Time-Shared Computer Systems. Ph.D. thesis, Massachusetts Institute of Technology, 1962.
[30] Claude E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Jorurnal*, Vol. 27, 1948, July, 379-423, October, 623-656. Also included in *Collected Papers*, ed. by N.J. Sloane and Aaron D. Wyner. IEEE Press, Piscataway, 1993, 5-83.
[31] Claude E. Shannon. Information Theory. *Encyclopædia Britiannica*, Vol 12, 1953, 350-353.
[32] Claude E. Shannon. A Mind-Readinng(?)Machine. Bell Laboratories Memorandum, March 18, 1953. 4 pp. Also included in *Collected Papers*, ed. by N.J. Sloane and Aaron D. Wyner. IEEE Press, Piscataway, 1993, 688-690
[33] Oscar B. Sheynin. A.A. Markov's Work on Probability. *Archive for History of Exact Sciences* 39:337-375, 1988/89.