Claude E. Shannon: A Retrospective on His Life, Work, and Impact

Robert G. Gallager, *Life Fellow, IEEE*

Invited Paper

***Abstract*—Claude E. Shannon invented information theory and provided the concepts, insights, and mathematical formulations that now form the basis for modern communication technology. In a surprisingly large number of ways, he enabled the information age. A major part of this influence comes from his two-part monumental 1948 paper, "A Mathematical Theory of Communication." We attempt here to provide some clues as to how a single person could have such a major impact. We first describe Shannon's life and then study his publications in the communication area. We next consider his research style in the context of these publications. Finally, we consider the process under which the impact of his work evolved from the creation of a beautiful and challenging theory to the establishment of the central principles guiding digital communication technology. We end with some reflections on the research environment that stimulates such work both then and now.**

***Index Terms*—Coding theorems, digital communication, information theory, Shannon.**

I. Claude Shannon's Life

A NATIVE of the small town of Gaylord, MI, Claude Elwood Shannon was born on April 30, 1916. His mother was a language teacher and principal of the local Gaylord High School, and his father was a businessman and a Judge of Probate.

Claude went through the public school system, graduating from Gaylord High School at the age of 16. The young Claude led a normal happy childhood with little indication of his budding genius. As in later life, he was not outgoing, but was friendly when approached. He was interested in such things as erector sets and model planes and was curious about how various devices worked.

After high school, Shannon enrolled in the University of Michigan, Ann Arbor, where, in 1936, he received bachelor's degrees in both electrical engineering and mathematics. His dual interest in these fields continued through his professional career. It was at Michigan also that his lifelong interest in Boolean algebra began.

While trying to decide what to do next, he saw a notice on a bulletin board advertising for someone to operate Vannevar Bush's differential analyzer (an early analog computer) at the Massachusetts Institute of Technology. Claude applied for the job and was accepted as a research assistant and graduate student in the MIT Electrical Engineering Department.

After arriving at MIT, Claude became interested both in the analog aspects of the computer and in the complex switching circuit controlling it. Along with his academic subjects, he started to explore the possibility that Boolean algebra could be used to understand such switching circuits.

After his first academic year, Claude spent the summer of 1937 at Bell Telephone Laboratories working on the relationship between Boolean algebra and switching. Back at MIT in the fall, he fleshed out these ideas and showed how to use Boolean algebra both for the analysis and synthesis of relay circuits. This was used both for his MIT Master's thesis and for his first published paper [3].

The importance of this work was quickly recognized as providing a scientific approach for the rapidly growing field of switching. Switching circuits were of great importance in the telephone industry, and subsequently in the development of computers. The paper won the 1940 Alfred Noble prize for the best paper in engineering published by an author under 30. It is widely recognized today as the foundation of the switching field and as one of the most important Master's theses ever written.

Partly on the advice of Vannevar Bush, Shannon started to look for a Ph.D. topic in the area of genetics. He switched from Electrical Engineering to Mathematics and aimed to establish a mathematical basis for genetics. His Ph.D. dissertation, "An Algebra for Theoretical Genetics," was completed in 1940. This thesis was never published and remained largely unknown until recently. Its results were important, but have been mostly rediscovered independently over the intervening years.

Claude was never interested in getting recognition for his work, and his mind was always full of new ideas, so many of his results were never published. While he was doing his Ph.D. research, he was also becoming interested in the fundamental problems of communication, starting to nibble around the edges of what would later become his monumental "A Mathematical Theory of Communication." He also continued to work on switching theory. Thus, it is not surprising that he focused on these areas after completing his thesis rather than on publication of the thesis.

Manuscript received August 3, 2001.

The author is with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139 USA (e-mail: gallager@lids.mit.edu).

Communicated by P. M. Siegel, Editor-in-Chief.
Publisher Item Identifier S 0018-9448(01)08945-3.

The summer of 1940 was spent at Bell Labs exploring further topics in switching. Claude then accepted a National Research Fellowship at the Institute for Advanced Study at Princeton. It was here, during the academic year 1940–1941, that he started to work seriously on his nascent mathematical theory of communication.

By the summer of 1941, war was imminent, and Shannon joined an elite group at Bell Labs working on fire control for anti-aircraft batteries. In his spare time, Claude continued to work on switching and on his rapidly developing theory of communication. He also published two papers, [28], [29], on the theory of differential analyzers. These were outgrowths of his earlier work on the differential analyzer at MIT. Along with developing a theory for these analog computers, they also contributed to an understanding of how digital computers could accomplish similar computational tasks.

During the war, Shannon also became interested in cryptography. He realized that the fundamental issues in cryptography were closely related to the ideas he was developing about communication theory. He was not cleared for the major cryptographic projects at Bell Labs, so he could explain his ideas to the relevant cryptographers, but they could not talk about their applications. It appears, however, that his results were important in the speech scrambling device used by Roosevelt and Churchill during the war.

Shannon wrote up his cryptography results in the classified paper, "A Mathematical Theory of Cryptography" in 1945; this became available in the open literature in 1949 as "Communication Theory of Secrecy Systems" [4]. This paper established a mathematical theory for secrecy systems, and has had an enormous effect on cryptography. Shannon's cryptography work can be viewed as changing cryptography from an art to a science.

Some of the notions of entropy that Shannon had worked out for his evolving theory of communication appeared in [4]. Since he reported these ideas first in his classified cryptography paper, some people supposed that he first developed them there. In fact, he worked them out first in the communication context, but he was not yet ready to write up his mathematical theory of communication.

By 1948, all the pieces of "A Mathematical Theory of Communication" [1], [2] had come together in Shannon's head. He had been working on this project, on and off, for eight years. There were no drafts or partial manuscripts—remarkably, he was able to keep the entire creation in his head. In a sense, this was necessary, because his theory was about the entire process of telecommunication, from source to data compression to channel coding to modulation to channel noise to demodulation to detection to error correction. The theory concerned the performance of the very best system possible and how to approach that performance (without explicitly designing the system). An understanding of each piece of the system was necessary to achieve this objective.

The publication of this monumental work caused a great stir both in the technological world and in the broader intellectual world. Shannon employed the provocative term "information" for what was being communicated. Moreover, he was able to quantify "information" for both sources and channels. This new notion of information reopened many age-old debates about the difference between knowledge, information, data, and so forth. Furthermore, the idea that something called information could be quantified stimulated much excitement and speculation throughout the intellectual community.

Whether Shannon's quantifiable definition of information will someday have a major impact on larger questions of either human or artificial intelligence is still an open question. It is certainly true, however, that [1], [2] totally changed both the understanding and the design of telecommunication systems, as we shall show below.

Claude remained in the mathematical research group at Bell Labs until 1956 and created a constant stream of new and stimulating results. There was a remarkable group of brilliant people to interact with, and he tended to quickly absorb what they were working on and suggest totally new approaches. His style was not that of the expert who knows all the relevant literature in a field and suggests appropriate references. Rather, he would strip away all the complexity from the problem and then suggest some extremely simple and fundamental new insight.

Claude tended to work alone for the most part. He would work on whatever problem fascinated him most at the time, regardless of whether it was of practical or conceptual importance or not. He felt no obligation to work on topics of value to the Bell System, and the laboratory administration was happy for him to work on whatever he chose. The Bell Labs administration was well known for supporting basic research in mathematics and science, but we must admire them for also encouraging Claude's research on topics that appeared slightly frivolous at the time.

In the years immediately after the publication of [1], [2], Claude had an amazingly diverse output of papers on switching, computing, artificial intelligence, and games. It is almost as if all these topics were on the back burner until all the conceptual issues in his theory of communication had been worked out. In retrospect, many of these papers have been important for Bell Labs.

One of the wonderful aspects of Claude is how his work and play came together. For example, the problem of programming a computer to play chess fascinated him [30], [31]. Chess is an interesting game from an artificial intelligence perspective, because there is no randomness in the game, but also there is no hope for a computer to tabulate all possible moves. The chess playing programs devised since then, which now can beat human chess champions, follow in a direct line from Shannon's pioneering work.

A similar semiserious project was Theseus. Theseus was a mechanical mouse, designed to solve mazes. Once it had solved the maze, it would remember the solution. If the walls of the maze were changed, or the position of the cheese changed, the mouse would recognize the change and find the new solution. Along with being amusing, this was an early and instructive example of machine learning. A short, but very popular, film was made of Shannon and Theseus.

A more tongue-in-cheek project of the period was the Throbac Computer, which calculated using Roman numerals. Another project was a penny matching machine that searched for patterns in the adversary's play.

Shannon had been interested in questions of computability and Turing machines since before the war, and had a number of interesting discussions with Alan Turing during the war. In [32], he showed how a universal Turing machine could be constructed with only two internal states. Along with its importance, this is a beautifully written paper, which provides an excellent tutorial introduction to Turing machine theory.

In other fundamental research, Claude worked with Edward Moore on computing with unreliable components [33]. Von Neumann had looked at this problem earlier, but had obtained weaker results. Moore and Shannon assumed that the computing elements were error-prone relays, with independently occurring errors. They showed how to achieve arbitrary reliability by using enough redundancy. Although this is a theoretically important result, it does not seem to have impacted the actual design of reliable computers.

Claude met his wife, Mary Elizabeth (Betty) Moore, at Bell Labs, where she worked as a numerical analyst. They shared a good natured intellectual sense of humor and a no-nonsense but easy-going style of life. They brought up three children, and although Claude was always thinking about some currently fascinating idea, he was also always available for his family. The family shared a love of toys, many of which Claude built himself. They had collections of unicycles, looms, chess sets, erector sets, musical instruments, as well as a gasoline powered pogo stick and the mechanical mouse Theseus. Claude was well known for riding a unicycle through the halls of Bell Labs while juggling.

Betty often helped Claude in his work, sometimes checking his numerical calculations, and sometimes writing his papers as he dictated them. It seems astonishing that anyone could dictate a paper and have it come out right without many editing revisions, but Claude disliked writing, and thus kept thinking about a subject until everything was clear.

In 1956, Claude spent a year visiting MIT, and then the next year visiting the Center for the Study of Behavioral Sciences in Palo Alto, CA. In 1958, he accepted a permanent appointment at MIT as Donner Professor of Science, with an appointment both in Electrical Engineering and in Mathematics. The Shannons bought a large gracious home in Winchester, MA, overlooking Mystic Lake, where there was plenty of room for all their toys and gadgets, and where they occasionally hosted parties for MIT students and faculty.

There was a very active group of graduate students and young faculty studying information theory at MIT around 1958. For them, Claude Shannon was an idol. Many of these students are now leaders in the digital communication field, and have made their mark both in research and practice.

Shannon's role as a faculty member at MIT was atypical. He did not teach regular courses, and did not really like to talk about the same subject again and again. His mind was always focused on new topics he was trying to understand. He was happy to talk about these new topics, especially when he obtained some new insights about them. Thus, he gave relatively frequent seminars. He once gave an entire seminar course with new research results at each lecture.

It was relatively rare for him to be the actual supervisor of a student's thesis, but yet he had an enormous influence on the students' lives. As in his earlier life, he was not outgoing, but he was very friendly and helpful when contacted. Many students summoned up the courage to approach him at some point, and he would usually find an interesting and novel way for them to look at their problems. These interactions were important in two ways. First, they helped the students directly in their research, and second, the students started to understand how to formulate and approach problems in a more fundamental way. Students learned to look at carefully constructed toy problems before getting lost in technical detail.

In his research at MIT, Shannon turned back to information theory and extended the theory in a number of ways as will be discussed later. He also continued to work or play with his many mechanical gadgets. He developed an elaborate strategy for winning at roulette by taking advantage of small imbalances in the roulette wheel. However, he tired of this before becoming successful, as he really was not interested in making money with the scheme, but only in whether it could be done.

He and Betty also became interested in the stock market. He developed some theories about investment growth that were never published; however, he gave a seminar on investment theory at MIT that attracted hundreds of eager listeners. On a more practical level, Claude and Betty invested very successfully, both in the general market and, more particularly, in several companies started by talented friends.

By the 1980s, it was increasingly clear that Claude was having memory problems, and he was later diagnosed with Alzheimer's disease. He spent the final years of his life in a private hospital, but was good-natured as usual and enjoyed Betty's daily visits. Finally, everything in his body started to fail at once, and he died on February 24, 2001.

II. A Mathematical Theory of Communication [1], [2]

This is Shannon's deepest and most influential work. It established a conceptual basis for both the individual parts and the whole of modern communication systems. It was an architectural view in the sense that it explained how all the pieces fit into the overall space. It also devised the information measures to describe that space.

Before 1948, there was only the fuzziest idea of what a message was. There was some rudimentary understanding of how to transmit a waveform and process a received waveform, but there was essentially no understanding of how to turn a *message* into a transmitted *waveform*. There was some rudimentary understanding of various modulation techniques, such as amplitude modulation, frequency modulation, and pulse code modulation (PCM), but little basis on which to compare them.

Most readers of this paper are familiar with Shannon's theory, and many have read [1], [2] in detail. However, we want to briefly retrace this work, in order to illustrate its remarkable simplicity and unity, its mathematical precision, and its interplay between models and reality.

Shannon started by explaining that messages should be thought of as choices between alternatives. In [1], this set of alternatives is discrete, whereas in [2] it is arbitrary.

The discrete theory draws on Hartley's work [5], which showed that (for many examples) the number of possible alternatives from a message source over an interval of duration T grows exponentially with T, thus suggesting a definition of information as the logarithm of this growth. Shannon extended this idea by attaching a probability measure to the set of alternatives, and by making a clean separation between source and channel. He pointed out that it is the *choice* between a set of alternatives which is important, not the representation (integer, letter, hieroglyph, binary code, etc.) of the choice. The representation of interest to the user may be mapped into any convenient representation for transmission (for example, mapping letters of the alphabet into bytes). That mapping is established ahead of time at both transmitter and receiver, and then an arbitrarily long sequence of choices can be communicated.

The major example used for illustrating the assignment of probabilities to alternatives is that of English text (of course, the particular language is not important). Shannon pointed out that some letters of the alphabet have higher relative frequency than others—e.g., "e" is much more likely than "q." Also, the letters are not used independently (e.g., "u" typically follows "q"), only letters that form English words can be used between spaces, and only sequences of words that obey the rules of English can be used.

Shannon then proposed studying artificial mathematical languages that model some, but not all, of these statistical constraints. For example, the simplest such model assumes independence between successive letters and uses experimentally derived relative frequencies as letter probabilities. A Markov source is a more complex model in which the state represents some known history, such as the previous letter or several letters. The transition to the next state is then labeled by the next letter, using experimentally derived conditional relative frequencies as letter probabilities.

The use of simple toy models to study real situations appears not to have been common in engineering and science before Shannon's work. Earlier authors in various sciences used simple examples to develop useful mathematical techniques, but then focused on an assumed "correct" model of reality. In contrast, Shannon was careful to point out that even a Markov source with a very large state space would not necessarily be a faithful model of English text (or of any other data). The purpose of a model is to provide intuition and insight. Analysis of the model gives precise answers about the behavior of the model, but can give only approximate answers about reality.

In summary, data sources are modeled as discrete stochastic processes in [1], and primarily as finite-state ergodic Markov sources. Shannon showed in a number of ways, including the growth rate of alternatives and the number of binary digits per unit time needed for any representation, that such source models are characterized by a certain information rate.

In 1948, and even today, to view a message source as a random process is a rather strange idea, in that we do not usually think of the messages we create as random. However, this point of view is appropriate for a communication engineer who is building a device to communicate unknown messages. Thus, the interpretation of information in Shannon's theory had nothing to do with the "meaning" of the message, but was simply a measure of the inherent requirements involved in communicating that message as one of a set of possible messages.

Shannon next considered channels in [1]. In his picture, a channel accepts a sequence of letters at its input and produces a noise-corrupted version of those letters at its output. He introduced the concept of encoding, which had hardly been considered previously. The channel encoder converts the source output sequence to an appropriate input sequence for the channel. A corresponding decoder tries to convert the channel output sequence back to the original source sequence.

Shannon then proved his most dramatic and unexpected result, the channel coding theorem. He shows that a channel is characterized by a single number, its *capacity*. If the information rate of a source model is less than the channel capacity, then it can be transmitted virtually error-free over the channel by appropriate processing. Conversely, if the source information rate exceeds the channel capacity, then significant distortion must result no matter what processing is employed.

In [2], these results were extended to analog sources and to analog channels with waveform inputs and outputs. For analog sources, the notion of information rate was extended to that of information rate relative to a fidelity (or distortion) criterion. Shannon showed that there is a concept of capacity for analog channels that is essentially the same as for discrete channels, although the mathematical details are considerably more complex.

Other researchers, such as Kolmogorov and Wiener, were independently starting to model transmitted waveforms as stochastic processes at this time. However, they were more interested in questions of estimation and filtering of a given waveform in the presence of noise. They had no sense of the transmitted waveform as an arbitrarily processed function of the source output, and thus had no sense of alternative choices or of information. Their work nowhere suggests the notions of capacity or of information rate.

A. The Source Coding Theorem

Let X be a discrete chance variable[1] with finitely many outcomes denoted by $1, 2, \ldots$. Let P_i be the probability of outcome i. Shannon defined the *entropy* of X as

$$H(X) = -\sum_i P_i \log_2 P_i. \tag{1}$$

The entropy is a function only of the probabilities, and not of the labels attached to the possible outcomes. As a simple extension, the conditional entropy of a chance variable Y conditioned on another chance variable X is

$$H(Y|X) = -\sum_{i,j} P_i p_{ij} \log_2 p_{ij} \tag{2}$$

where $P_i = \Pr(X = i)$ and $p_{ij} = \Pr(Y = j | X = i)$. Viewing the pair XY as a chance variable in its own right, the entropy $H(XY)$ is given by (1) as

$$H(XY) = -\sum_{i,j} P_i p_{ij} \log_2 P_i p_{ij}.$$

[1] A chance variable is a mapping from a probability space to a given set. If the set is the set of real or complex numbers, then the chance variable is called a random variable.

It is then easy to see that $H(XY) = H(X) + H(Y|X)$. Shannon gave many additional extensions, interpretations, equalities, and inequalities between entropy expressions which have been repeated in all texts on Information Theory and need no repetition here.

For an ergodic Markov chain in which $p_{ss'}$ denotes the conditional probability of a transition to state s' from state s and π_s denotes the steady-state probability of state s, it follows from (2) that the entropy per transition of the Markov chain is

$$H(S'|S) = -\sum_{s,\, s'} \pi_s p_{ss'} \log_2 p_{ss'}.$$

Now consider a source modeled by an ergodic finite-state Markov chain. Assume throughout that each alphabet letter appears on at most one outgoing transition from each state, Then, given an initial state, an output letter sequence corresponds to a unique state sequence, so the entropy of the source is equal to that of the Markov chain. If we relabel $p_{ss'}$ as p_{si}, where i denotes the source letter associated with the transition from s to s', then the entropy per transition of the Markov source is

$$H(X|S) = -\sum_{s,\, i} \pi_s p_{si} \log_2 p_{si}. \tag{3}$$

The main justification for these definitions of entropy is the source coding theorem [1, Theorems 3 and 4], which relate entropy to the probability of typical long source sequences and to the number of binary digits required to represent those sequences. These theorems have been generalized and reproven in many ways since 1948. We prefer Shannon's original proof, which is very short and eloquent. We give it here (making a few details more explicit) to demonstrate both its mathematical precision and its central role later in proving the noisy channel coding theorem.

We start with the simplest case in which the source output is a sequence of independent and identically distributed (i.i.d.) source letters from a finite alphabet, say $1, 2, \ldots$. Letting $P_i > 0$ denote the probability[2] of letter i, the probability of a sample sequence $\boldsymbol{x} = (i_1, i_2, \ldots, i_n)$ is $\prod_{k=1}^{n} P_{i_k}$. Letting n_i denote the number of appearances of letter i in $\boldsymbol{x}$, this may be rewritten as

$$\Pr(\boldsymbol{x}) = \prod_i P_i^{n_i}. \tag{4}$$

For the i.i.d. case, define a sequence $\boldsymbol{x}$ of length n to be *δ-typical*[3] if

$$nP_i(1-\delta) \le n_i \le nP_i(1+\delta), \qquad \text{for all } i.$$

For brevity, we express this condition as $n_i = nP_i(1 \pm \delta)$. Taking the logarithm of (4) and dividing by n, a δ-typical sequence has the property that

$$\begin{aligned} \frac{\log_2 \Pr(\boldsymbol{x})}{n} &= \sum_i \frac{n_i}{n} \log_2 P_i \\ &= \sum_i P_i(1 \pm \delta) \log_2 P_i \\ &= -H(X)(1 \pm \delta). \end{aligned} \tag{5}$$

The δ-typical sequences are simply those for which the relative frequency of each letter in the sequence is approximately equal to the probability of that letter. We see from (5) that

$$\Pr(\boldsymbol{x}) = 2^{-nH(X)(1 \pm \delta)}. \tag{6}$$

By the law of large numbers,[4] for any $\epsilon > 0$ and the given $\delta > 0$, there is an N_0 such that for all sequence lengths $n \ge N_0$, the set T_δ of δ-typical sequences has probability

$$\Pr(T_\delta) \ge 1 - \epsilon. \tag{7}$$

Equations (6) and (7) comprise the essence of Shannon's Theorem 3 for this simple case. They say that, for sufficiently large n, the set of δ-typical sequences is overwhelmingly probable, and that each typical sequence has approximately the same probability in the sense of (6). This is an unusual sense of approximation, since one typical sequence can have a probability many times that of another, but it is sufficient for many of the needs of information theory.

In Shannon's Appendix 3, this argument is generalized to finite-state ergodic Markov sources. Again, for each state s, let π_s be the steady-state probability of state s and let $p_{si} > 0$ be the transition probability of the letter i conditional on state s. For any given sample sequence $\boldsymbol{x}$ of length n, and for any given initial state s_0, let n_{si} be the number of transitions from state s using letter i. Then, as in (4), the probability of $\boldsymbol{x}$ given s_0 is

$$\Pr(\boldsymbol{x}|s_0) = \prod_{s,\, i} p_{si}^{n_{si}}.$$

We say that a sample sequence $\boldsymbol{x}$ with starting state s_0 is *δ-typical* if, for each (s, i), the number n_{si} of (s, i) transitions in $\boldsymbol{x}$ is $n\pi_s p_{si}(1 \pm \delta)$.

As in (4), the probability of any given δ-typical sequence of length n is

$$\Pr(\boldsymbol{x}|s_0) = \prod_{s,\, i} p_{si}^{n_{si}} = \prod_{si} p_{si}^{n\pi_s p_{si}(1 \pm \delta)}. \tag{8}$$

Taking the logarithm and dividing by n

$$\begin{aligned} \frac{\log_2 \Pr(\boldsymbol{x}|s_0)}{n} &= \sum_{s,\, i} \pi_s p_{si}(1 \pm \delta) \log_2 p_{si} \\ &= -H(X|S)(1 \pm \delta). \end{aligned} \tag{9}$$

As in the i.i.d. case, the typical sequences for a given s_0 are those for which the relative frequency of each state transition is close to the average. The weak law of large numbers for finite-state

[2] Discrete events of zero probability can often be ignored by leaving them out of the sample space; here, if $P_i = 0$, it can be removed from the alphabet. For the nondiscrete case, more care is required.

[3] Shannon referred to the set of these sequences as the high probability set. Today this is called a strongly typical sequence, although the detailed use of δ is nonstandard. Our use here is especially simple, although it is restricted to finite alphabets.

[4] Shannon quoted the strong law of large numbers here, but the weak law is also sufficient.

ergodic Markov chains says that, for any $\epsilon > 0$ and $\delta > 0$, there is an N_0 such that, for all $n \geq N_0$ and all starting states s_0, the set of δ-typical sequences $T_\delta(s_0)$ for starting state s_0 has probability

$$\Pr(T_\delta(s_0)) \geq 1 - \epsilon.$$

This proves [1, Theorem 3], which says (slightly paraphrased) the following.

Theorem: For every ϵ, $\delta > 0$, there exists an N_0 such that for all $n \geq N_0$, the set $T_\delta(s_0)$ of δ-typical n-sequences for each starting state s_0 satisfies $\Pr(T_\delta(s_0)) \geq 1 - \epsilon$, and for each $\boldsymbol{x} \in T\delta(s_0)$

$$\begin{aligned} H(X|S)(1+\delta) &\geq \frac{-\log_2[\Pr(\boldsymbol{x}|s_0)]}{n} \\ &\geq H(X|S)(1-\delta). \end{aligned} \tag{10}$$

An important use of this theorem is to estimate the number $m(T_\delta(s_0))$ of δ-typical sequences. Since

$$\Pr(\boldsymbol{x}|s_0) \leq 2^{-nH(X|S)(1-\delta)}$$

for each δ-typical sequence, and since

$$\Pr(T_\delta(s_0)) \geq 1 - \epsilon$$

we must have

$$m(T_\delta(s_0)) \geq (1-\epsilon)2^{nH(X|S)(1-\delta)}. \tag{11}$$

Similarly, using the opposite limits on $\Pr(\boldsymbol{x}|s_0)$ and $\Pr(T_\delta(s_0))$

$$m(T_\delta(s_0)) \leq 2^{nH(X|S)(1+\delta)}. \tag{12}$$

Equations (11) and (12) comprise a slightly weakened version of Shannon's Theorem 4. Equation (12) shows that, for any $\delta > 0$, it is possible to map all δ-typical sequences in $T_\delta(s_0)$ into binary sequences of length at most $nH(X|S)(1+\delta)$.

Note that in (10), the *bounds* on $\Pr(\boldsymbol{x}|s_0)$ for δ-typical sequences are independent of the starting state s_0. The *set* of δ-typical sequences is a function of s_0, however. Often $\Pr(\boldsymbol{x})$ is of interest rather than $\Pr(\boldsymbol{x}|s_0)$. Define $\boldsymbol{x}$ to be δ-typical if it is δ-typical for at least one starting state. Then, for large enough n, (10) is valid for $\Pr(\boldsymbol{x})$ and (11) and (12) are valid for the enlarged set T_δ of δ-typical sequences if δ is replaced by 2δ.

The results above ignore source n-sequences outside of the typical set, which is sometimes undesirable. Other results in [1] use variable-length coding techniques to show that, for large enough n, the set of *all* n-sequences from an ergodic Markov source with entropy $H(X|S)$ per letter can be compressed into an *expected* length arbitrarily close to $H(X|S)$ binary digits (bits) per letter, but never less than $H(X|S)$ bits per letter. This result is more important in practice, but [1, Theorems 3 and 4] give the most fundamental insight into the meaning of entropy.

It is important to remember that these results apply to models of sources rather than to the sources themselves. It can be shown that if a sequence is δ-typical in a refinement of a given model, then it is also δ-typical in the given model. However, since these models are all stationary, and real sources are never truly stationary (both machines and humans have finite lives), more elaborate models must be treated with care, and detailed results about their convergence as $n \to \infty$ may have limited engineering significance. The real escape from this modeling dilemma came later with the introduction of universal source codes (e.g., [6], [7]), which exploit whatever redundancy exists in the source without the need for a probabilistic model.

B. The Noisy Channel Coding Theorem

The noisy channel coding theorem is certainly the crowning jewel of [1], [2].

In [1], the input and output are regarded as discrete sequences, so the channel is viewed as including the modulation and demodulation. In [2], the input and output are regarded as waveforms, so modulation and demodulation are viewed as part of the input and output processing. In each case, Shannon defined various simplified models of real communication channels. For each such model, the capacity C_t in bits per second is defined.

The noisy channel coding theorem [1, Theorem 11] states that for any source whose entropy per second H_t is less than C_t, it is possible to process (encode) that source at the channel input, and to process (decode) the received signal at the output, in such a way that the error rate (in source symbol errors per second) is as small as desired. Furthermore, if H_t is greater than C_t, arbitrarily small equivocation is impossible.

Achieving a small error probability with a given source and channel when $H_t < C_t$ usually requires large delay and high complexity. Even so, this result was very surprising in 1948 since most communication engineers thought that small error probability could only be achieved by decreasing H_t. Perhaps the only reason this result is less surprising today is that we have heard it so often.

It is now common, even outside the engineering community, to refer to sources in terms of data rate in bits per second and to channels in terms of transmitted bits per second.

The typical sequence arguments of the last section help to understand part of this result. For the models considered, there are about $2^{nH(X|S)}$ approximately equiprobable typical source n-sequences when n is large. If the source emits a symbol each τ seconds, then the channel can successfully transmit the source output if and only if the channel can transmit a choice from approximately $2^{nH(X|S)}$ equiprobable alternatives per interval $n\tau$. Put more simply, the source output can be sent if and only if the channel is capable of transmitting $nH(X|S)$ binary digits reliably per interval $n\tau$ for n large.

We will make this argument more precise later, but for now the reader should appreciate the remarkable insight that Shannon gave us simply from the properties of typical source sequences. Since a long source output sequence is highly likely to be one of a set of more or less equiprobable alternatives, there is no essential loss of generality in mapping these sequences into binary digits, and then transmitting the binary digits over the channel.

The above discussion of sources does not explain why it is possible to transmit binary digits reliably at a rate arbitrarily close to C_t bits per second. It is surprising that Shannon's proof of this result is as simple as it is. The proof applies the typical sequence arguments of the last section to the channel input and output, and then adds one more ingenious twist. Shannon's proof is a little sketchy here, but all of the ideas are clearly presented. We will present his proof for the special case of a discrete memoryless channel (DMC), adding the requisite details.

The input to a discrete channel is a sequence $\boldsymbol{X} = (X_1, X_2, \ldots)$ of letters from some finite alphabet, denoted $\{1, 2, \ldots\}$ and the output is a corresponding sequence $\boldsymbol{Y} = (Y_1, Y_2, \ldots)$ from a possibly different finite alphabet, denoted $\{1, 2, \ldots\}$. The channel is noisy in the sense that the outputs are not determined by the inputs, but rather have only a stochastic dependence on the input. Thus, given any input sequence, the output sequence is a stochastic process with a known distribution conditional on the input. However, the channel input sequence is arbitrary. Choosing the encoding relationship between the source output and the channel input is the most important degree of freedom that we have in designing a system for reliable communication. The channel input and output may be described as a joint stochastic process once we know how this source/channel input processing is done.

We consider a particularly simple type of noisy channel known as a discrete memoryless channel[5] (DMC). Here, each output Y_k in the output sequence $\boldsymbol{Y} = (Y_1, Y_2, \ldots)$ is statistically dependent only on the corresponding input X_k, and at each time k there is a given conditional probability $p_{ij} > 0$ of output j given input i, i.e., $\Pr(Y_k = j | X_k = i) = p_{ij}$ independent of k. Thus, for any sample input sequence $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and output sequence $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, the conditional probability $\Pr(\boldsymbol{y}|\boldsymbol{x})$ is given by

$$\Pr(\boldsymbol{y}|\boldsymbol{x}) = \prod_{k=1}^{n} p_{x_k y_k}. \tag{13}$$

Shannon began his analysis of a noisy channel by representing the channel input and output by chance variables X and Y. These can be viewed as individual letters or sequences of letters. They involve not only the channel representation, but also a stochastic input representation. He defined the transmission rate[6] I for this input choice as

$$I = H(X) - H(X|Y). \tag{14}$$

Shannon interpreted the transmission rate as the *a priori* uncertainty $H(X)$ about the input less the conditional uncertainty, or equivocation, $H(X|Y)$ about the input after the output is observed. By manipulating entropy expressions, (14) can also be represented as

$$I = H(Y) - H(Y|X). \tag{15}$$

Now view X and Y in these expressions as n-sequences $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. From (13)

$$H(\boldsymbol{Y}|\boldsymbol{X}) = \sum_k H(Y_k|X_k).$$

From (15), then

$$I = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X}) \le \sum_{k=1}^{n} H(Y_k) - H(Y_k|X_k) \tag{16}$$

with equality if the inputs are statistically independent.

Shannon's definition of the channel capacity C in bits per symbol for an arbitrary discrete channel is essentially the supremum of $\frac{1}{n}[H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{Y})]$ over both the input distribution and the sequence length n. He did not spell out the specific conditions on the channel for his subsequent results to be valid. However, for a DMC, (16) shows that this supremum is the same for all n and is achieved by i.i.d. inputs, in which the input i is chosen with the probability P_i that achieves the maximization

$$C = \max_{P_1, \ldots, P_n} \left[\sum_{i,j} P_i p_{ij} \log_2 \frac{p_{ij}}{\sum\limits_{l} P_l p_{lj}} \right]. \tag{17}$$

Shannon gave an explicit formula for C when the maximizing distribution satisfies $P_i > 0$ for all i.

We can now outline Shannon's proof that an arbitrarily small error probability can be achieved on a DMC when $H_t < C_t$. We start with an artificial source that produces i.i.d. inputs with the optimizing input probabilities P_i in (17). With this input distribution, the input/output pairs

$$((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$$

are i.i.d.. We then define the δ-typical set of these input/output pairs. The next step is to choose 2^{nR} codewords randomly, for a given $R < C$. We then define a decoding rule for each such randomly chosen code. Finally, we evaluate an upper bound on error probability averaged over this ensemble of randomly chosen codes. We show that this bound approaches 0 as $n \to \infty$. Obviously, some code must be as good as the average for each n.

To give the details, let $(X_1, X_2, \ldots, X_n)$ be an i.i.d. input sequence with $\Pr(X_k{=}i) = P_i$ for $1 \le k \le n$. The channel output sequence is then i.i.d. with probabilities

$$\Pr(Y_k{=}j) = \sum_i P_i p_{ij}.$$

The input/output pairs $X_1Y_1, X_2Y_2, \ldots, X_nY_n$ are i.i.d. with probabilities $\Pr(X_kY_k{=}ij) = P_i p_{ij}$. For an input/output sequence $\boldsymbol{xy} = x_1y_1, x_2y_2, \ldots, x_ny_n$, let n_{ij} be the number of input/output pairs x_ky_k that take the value ij. As in (4), we have

$$\Pr(\boldsymbol{xy}) = \prod_{i,j} (P_i p_{ij})^{n_{ij}}. \tag{18}$$

From the general definition of δ-typicality for i.i.d. chance variables, $\boldsymbol{xy}$ is δ-typical if $n_{ij} = nP_i p_{ij}(1 \pm \delta)$ for each i, j. As in (5), for any δ-typical $\boldsymbol{xy}$

$$\begin{aligned} -\frac{\log_2 \Pr(\boldsymbol{xy})}{n} &= \sum_{i,j} P_i p_{ij}(1 \pm \delta) \log_2 P_i p_{ij} \\ &= H(XY)(1 \pm \delta). \end{aligned} \tag{19}$$

[5] Shannon [1] considered a more general channel called a finite-state channel in which the noise and next state are probabilistic functions of the input and previous state. Shannon's proof works in essence for this more general case, but a number of subtle additional conditions are necessary (see, for example, [8, Sec. 4.6]).

[6] Shannon used the variable R for transmission rate, but this rate is now usually called mutual information.

If $\boldsymbol{xy}$ is δ-typical, then $\boldsymbol{x}$ is also δ-typical. To see this, let n_i be the number of inputs that take the value i. Then

$$n_i = \sum_j n_{ij} = \sum_j P_i p_{ij}(1 \pm \delta) = P_i(1 \pm \delta).$$

Similarly, if $\boldsymbol{xy}$ is δ-typical, then $\boldsymbol{y}$ is δ-typical. For a δ-typical $\boldsymbol{xy}$ pair, (6) then gives us the following relations:

$$\Pr(\boldsymbol{xy}) = 2^{-nH(XY)(1\pm\delta)} \tag{20}$$

$$\Pr(\boldsymbol{x}) = 2^{-nH(X)(1\pm\delta)} \tag{21}$$

$$\Pr(\boldsymbol{y}) = 2^{-nH(Y)(1\pm\delta)}. \tag{22}$$

Finally, for each δ-typical output $\boldsymbol{y}$, define the *fan* $F_{\boldsymbol{y}}$ to be the set of input sequences $\boldsymbol{x}$ such that the pair $\boldsymbol{xy}$ is δ-typical. If $\boldsymbol{y}$ is not δ-typical, then $F_{\boldsymbol{y}}$ is defined to be empty. For a typical $\boldsymbol{xy}$ pair, $\Pr(\boldsymbol{y}) \leq 2^{-nH(Y)(1-\delta)}$ and $\Pr(\boldsymbol{xy}) \geq 2^{-nH(XY)(1+\delta)}$. Thus, the number of elements in $F_{\boldsymbol{y}}$ must satisfy

$$\begin{aligned} |F_{\boldsymbol{y}}| &\leq 2^{nH(XY)(1+\delta) - nH(Y)(1-\delta)} \\ &= 2^{n(H(X|Y)+\delta(H(XY)+H(Y)))}. \end{aligned} \tag{23}$$

We next choose a code consisting of M input sequences $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$, each of length n. We choose each letter of each sequence independently at random, using letter i with the capacity-achieving probability P_i. We will then average the error probability over this ensemble of randomly chosen codewords.

The decoding rule proposed by Shannon is a "typical-set" rather than a maximum-likelihood decoding rule. Given a received sequence $\boldsymbol{y}$, the rule is to choose the unique message m' such that the codeword $\boldsymbol{x}_{m'}$ is in the fan $F_{\boldsymbol{y}}$. If $F_{\boldsymbol{y}}$ contains either no codewords or more than one codeword, then the decoder refuses to choose, and a decoding error is said to occur.

If the input to the encoder is message m, then a decoding error will occur only if $\boldsymbol{x}_m\boldsymbol{y}$ is not δ-typical (i.e., if $\boldsymbol{y}$ is not δ-typical or if $\boldsymbol{x}_m \notin F_{\boldsymbol{y}}$), or if $\boldsymbol{x}_{m'} \in F_{\boldsymbol{y}}$ for any other codeword $\boldsymbol{x}_{m'}$. Letting T_δ be the set of δ-typical $\boldsymbol{xy}$ sequences, the union bound then upper-bounds the probability of error as

$$\Pr(E) \leq (1 - \Pr(T_\delta)) + (M-1)\Pr(\boldsymbol{X}_{m'} \in F_{\boldsymbol{Y}}). \tag{24}$$

This equation makes use of the fact that, over this random ensemble of codes, the error probability does not depend on the message m. It also makes use of the fact that the input/output sequence $\boldsymbol{X}_m\boldsymbol{Y}$ is a set of n independent input/output pairs each with the probabilities $P_i p_{ij}$ for which the above definition of δ-typicality applies.

Each codeword $\boldsymbol{X}_{m'}$ other than the transmitted word is independent of the received sequence $\boldsymbol{Y}$. Each δ-typical choice for $\boldsymbol{X}_{m'}$ has a probability at most $2^{-nH(X)(1-\delta)}$. Thus, using the bound on $|F_{\boldsymbol{y}}|$ in (23), which is valid for all $\boldsymbol{y}$, we see that

$$\begin{aligned} \Pr(\boldsymbol{X}_{m'} \in F_{\boldsymbol{Y}}) &\leq 2^{-nH(X)(1-\delta)} 2^{n\{H(X|Y)+\delta[H(XY)+H(Y)]\}} \\ &= 2^{-n\{C+\delta[H(X)+H(XY)+H(Y)]\}}. \end{aligned} \tag{25}$$

The rate of the code in bits per channel letter is $R = \frac{1}{n}\log_2 M$. If $R < C$, then $\eta = C - R > 0$. Upper-bounding $M-1$ by M, it follows from (24) and (25) that

$$\Pr(E) \leq (1 - \Pr(T_\delta)) + 2^{-n\{\eta+\delta[H(X)+H(XY)+H(Y)]\}}. \tag{26}$$

To complete the proof of the coding theorem (i.e., that $\Pr(E)$ can be made arbitrarily small for $R < C$), we choose

$$\delta = \eta/\{2(H(X)+H(XY)+H(Y))\}$$

Equation (26) then becomes

$$\Pr(E) \leq (1 - \Pr(T_\delta)) + 2^{-n\eta/2}.$$

For any $\epsilon > 0$, we then choose n large enough that both $\Pr(T_\delta) \geq 1 - \epsilon/2$ and $2^{-n\eta/2} \leq \epsilon/2$. Thus, for sufficiently large n, $\Pr(E) \leq \epsilon$. Since this is true for the average over this ensemble of codes, it is also true for at least one code.[7]

The error probability here is the probability that a block of input data is decoded incorrectly. The probability of error per binary input (averaged over the block) is at least as small, and the error probability per transmitted source symbol is arbitrarily small, provided that errors in the state are prevented from causing a propagation of source letter errors.

Shannon went one step further in his Theorem 11, where he states that arbitrarily small error probability can also be achieved in the case for which $R = C$. He does not indicate how this extension is made, but it is quite simple if we interpret error probability appropriately. For $\eta > 0$ arbitrarily small, let $R' = R(1-\eta)$. We have seen that an arbitrarily small block error probability ϵ is achievable at rate R' with some block length n. Encode the source sequence into a sequence of these codewords, but send only a fraction $1-\eta$ of that sequence, and accept errors in the remaining fraction η of unsent codewords. As an average in time over the input data, the error probability is then at most $\epsilon + \eta$, which can be made arbitrarily small. This does not assert that we can achieve a rate $R = C$ within a single block with small block error probability, but it does assert that reliable communication is possible in the time average sense above.

Finally, Shannon gave a converse to the coding theorem when $R > C$ in terms of the equivocation $H(X|Y)$. In 1948, the Fano inequality [22] (which lower-bounds the error probability in terms of equivocation) did not exist, so Shannon simply showed that $H(X|Y) \geq R - C$. The Fano inequality (and later, the strong converse) were certainly important, but the fundamental insights were all there in [1].

In summary, Shannon really did prove the noisy channel coding theorem, except for spelling out a few of the details. Perhaps more importantly, he provided fundamental insights that were very simple, beautiful, and useful in later developments.

C. Analog Channels and Sources

The second part of "A Mathematical Theory of Communication" [2] extends the results of [1] to analog channels and analog sources.

Shannon began this extension by presenting the sampling theorem as a method of representing waveforms limited to frequencies at most W by a sequence of time samples with a sample interval of $1/(2W)$ seconds. The sampling theorem had been known earlier to mathematicians, but it had not been used previously by communication engineers.

From the sampling theorem, Shannon argued that waveforms limited to a bandwidth W and a time T have about $2TW$ degrees of freedom (in an asymptotic sense, when TW is large).

[7] The error probability for such a good code is an average over the M codewords. A good code can be made uniformly good for each codeword by deleting the half of the codewords that have highest error probability. Shannon did not discuss uniformly good codes, but Feinstein's later proof [34] achieved this uniform property.

The notion of $2TW$ degrees of freedom was known at the time, largely through the work of Nyquist [10]. Evidently, however, the signal space concepts of such importance today were in their infancy then.

The sampling theorem provides a convenient way to represent analog sources and channels in terms of sequences of chance variables, thus providing a link between [1] and [2]. In [1], these sequences consist of discrete chance variables, whereas in [2], they mainly consist of continuous-valued random variables. Shannon was certainly aware of the issues of intersymbol interference that had been so eloquently treated by Nyquist, but he was primarily interested in simple models that would permit the development of his theory with the fewest distractions; the sampling theorem offered such a model.

The entropy of a continuous valued random variable X with a probability density $p(x)$ was then defined as

$$H(X) = -\int p(x) \log p(x)\, dx.$$

Many of the relations between entropies, joint entropies, conditional entropies, and so forth are valid for this new type of entropy. Shannon pointed out, however, that this form of entropy is measured relative to the coordinate system, and is therefore less fundamental than discrete entropy. Fortunately, however, the difference of entropies, such as $H(X) - H(X|Y)$, is essentially independent of the coordinate system.

One particularly important result here is that if X is a Gaussian random variable with mean zero and variance N, then $H(X) = \frac{1}{2} \log_2 2\pi eN$. Moreover, for any random variable X with second moment $\overline{X^2} \le N$, $H(X) \le \frac{1}{2} \log_2 2\pi eN$. Thus, a Gaussian random variable has the maximum entropy for a given second moment.

Shannon next developed one of the best known and important results in communication theory, the capacity of the ideal band-limited additive white Gaussian noise (AWGN) channel. He considered a channel in which the input is limited to the band $(-W, W)$ and the noise is white Gaussian noise. The noise outside of the signal band is irrelevant, so both input and output can be represented by sequences of random variables at a rate of $2W$ samples per second. Each output variable Y_k can be represented as the input X_k plus the noise Z_k, where the noise variables are i.i.d. Gaussian with mean zero and variance N, and are independent of the input.

As with a DMC, the transmission rate $I = H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{Y})$ for the input/output sequence is upper-bounded by

$$\sum_k [H(X_k) - H(X_k|Y_k)]$$

with equality when the input variables are independent.

It is easily seen that the transmission rate between input and output is unbounded unless some constraint is put on the input variables. The most convenient and practical constraint is on the input power, either a mean-square constraint $\overline{X_k^2} \le P$ on each input, or a mean-square constraint $\overline{\boldsymbol{X}^2} \le 2WTP$ on a sequence of $2WT$ inputs. Shannon showed that a Gaussian input distribution with mean zero and variance P maximizes the transmission rate for each constraint above, yielding a transmission rate per letter equal to $\frac{1}{2} \log_2 (P+N)/N$. The capacity in bits per second is then given by the famous equation

$$C_t = W \log_2 \frac{P+N}{N}. \tag{27}$$

Shannon went on to outline how bit sequences with rates less than or equal to C_t can be transmitted with arbitrarily small error probability. The Gaussian input and output variables are approximated by finite, finely quantized sets. This yields a DMC whose capacity can be made arbitrarily close to that of the continuous channel by arbitrarily fine quantization. Thus, [1, Theorem 11] can be reused, with appropriate continuity conditions, on the channel transition probabilities.

The capacity result of (27) is also extended somewhat for the case of non-Gaussian additive noise variables, again assuming statistical independence between input and noise. Shannon defined the entropy power $\overline{N}$ of a random variable Z as the variance of a Gaussian random variable having the same entropy as Z. Clearly, $\overline{N} \le N = \overline{Z^2}$. He then showed that

$$W \log_2 \frac{P+\overline{N}}{\overline{N}} \le C \le W \log_2 \frac{P+N}{\overline{N}}. \tag{28}$$

The final topic in [2] is a brief outline of rate-distortion theory. (Shannon would return several years later [13] to develop this topic much more completely.)

In [1], Shannon had shown how many bits per symbol are required to represent a discrete source. For a continuous source, it generally requires an infinite number of binary digits to represent a sample value of a single variable exactly. Thus, to represent waveform sources such as voice, it is necessary to accept some distortion in digital representations, and it is natural to expect a tradeoff between rate (in bits per symbol) and the level of distortion.

Rate-distortion theory permits distortion to be defined in a variety of ways, such as mean square, maximum, weighted mean square, etc. The problem is then to determine the minimum average number $R(D)$ of bits per second that are required to represent the source within a given mean distortion level D.

Shannon's solution to this problem is entirely in the spirit of the rest of Information Theory. Let $\boldsymbol{X}$ represent the source output sequence. To be specific, we may take $\boldsymbol{X}$ to be $2WT$ time samples from a band-limited source.[8] Let $\boldsymbol{Y}$ represent the corresponding channel output sequence. As usual, the channel is defined by the conditional probability density of $\boldsymbol{Y}$ given $\boldsymbol{X}$. The rate (in bits per second) of a source relative to a mean distortion D is then defined as

$$R(D) = \inf \frac{H(\boldsymbol{X}) - \boldsymbol{H}(\boldsymbol{X}|\boldsymbol{Y})}{T} \tag{29}$$

where the infimum is taken over T and over probability distributions on $\boldsymbol{X}$ such that the mean distortion between sequences $\boldsymbol{X}$ and $\boldsymbol{Y}$ of length $2WT$ is at most DT. There are a number of mathematical issues here involving measurability, but Shannon was clearly interested in the general idea rather than in producing a careful theorem.

Shannon restricted his attention to sources and distortion measures for which the infimum above can be approximated arbitrarily closely by a channel with finite alphabets. He then gave a random coding argument very similar to that of his

[8] Shannon did not restrict himself in this way and uses a generic form of input.

Theorem 11. The major difference is that the code consists of approximately $2^{2WTR(D)(1+\delta)}$ output sequences rather than a set of input codewords. The argument is again based on jointly typical input/output sequences. The distortion between the input sequence and output sequence of each such jointly typical pair is approximately DT. For large enough T, each input sequence is with high probability in the fan $F(\boldsymbol{y})$ of one of these codewords $\boldsymbol{y}$.

The above argument says roughly that a source can be compressed into about $R(D)$ bits per second in such a way that the corresponding binary sequence represents the source with an average distortion per second of D.

However, the expression (29) leads to a much stronger claim. If a source sequence $\boldsymbol{X}$ is processed in a completely arbitrary way and then passed through a channel of capacity C_t with output $\boldsymbol{Y}$, then the combination of processor and channel may be regarded as simply another channel with capacity at most C_t. It follows that if $C_t < R(D)$, then from (29) the average distortion between $\boldsymbol{X}$ and $\boldsymbol{Y}$ must be greater than D.

In other words, whether we insist on mapping $\boldsymbol{X}$ into a binary stream with average distortion at most $R(D)$ before transmission over a channel, or we allow $\boldsymbol{X}$ to be processed in any way at all, a channel of capacity $C_t \geq R(D)$ is required to achieve average distortion D. This is the essence of what is now called the source/channel separation theorem, or, more succinctly, the binary interface theorem. If a discrete or analog source with a distortion constraint can be transmitted by any method at all through a given channel, then it can alternatively be transmitted by the following two-stage process: first, encode the source into a binary stream that represents the source within the distortion constraint; second, using channel coding, send the binary stream over the channel essentially without errors.

Shannon never quite stated this binary interface property explicitly, although it is clear that he understood it. This result, which was essentially established in 1948, forms the principal conceptual basis for digital communication. Notice that when we say "digital communication," we do not imply that the physical channel is digital, only that the input to the modulator is discrete, and we do not imply that the source is discrete, but only that it is to be represented by a discrete sequence. Thus, "digital communication" implies only that there is a discrete interface between the source and channel, which without loss of generality can be taken to be binary. This establishes the architectural principle that all interfaces may be standardized to be binary interfaces without any essential loss in performance. This means that source coding and channel coding can be treated as independent subjects, a fact that has been implicitly (but not explicitly) recognized since 1948.

Information theory has sometimes been criticized for ignoring transmission delay and decoding complexity. However, if Shannon had been required to take these additional considerations into account, information theory would probably never have been invented. The simple and powerful results of information theory come from looking at long time intervals and using the laws of large numbers. No doubt Shannon saw that it was necessary to exclude considerations of delay and complexity in order to achieve a simple and unified theory. For example, he never even mentions the delay problems involved in using the sampling theorem as a bridge between discrete sequences and continuous waveforms. Later work has extended information theory to address delay and complexity in various ways.

III. Shannon's Other Major Works in Information Theory

The seeds for the modern age of digital communication were all present in [1] and [2]. In subsequent years, Shannon continued to play a critical role both in generalizing his theory and in making it more precise. The original papers were in some sense an extended outline, presenting all the major results and tools, but not including many later refinements that improved the theory conceptually and tailored it for applications.

We discuss these subsequent papers briefly, starting with two important papers that were almost concurrent with [1] and [2].

A. PCM and Noise

The first subsequent paper was "The Philosophy of PCM" [11], whose coauthors were B. R. Oliver and J. R. Pierce. This is a very simple paper compared to [1], [2], but it had a tremendous impact by clarifying a major advantage of digital communication.

In typical large communication systems, a message must travel through many links before reaching its destination. If the message is analog, then a little noise is added on each link, so the message continually degrades. In a digital system, however, "regenerative repeaters" at the end of each link can make decisions on the discrete transmitted signals and forward a noise-free reconstructed version, subject to a small probability of error. The end-to-end probability of error grows approximately linearly with the number of links, but, with Gaussian noise, a negligible increase in signal-to-noise ratio compensates for this. The only distortion is that introduced in the initial sampling and quantization.

Uncoded PCM also requires bandwidth expansion, converting one source sample into multiple bits. This paper conceded this bandwidth expansion, and did not emphasize the message of [1], [2] that digital transmission with efficient source and channel coding is ultimately at least as bandwidth-efficient as analog transmission. It was many years before this message became widely accepted.

The enduring message of this paper is that digital transmission has a major advantage over analog transmission in faithful reproduction of the source when communication is over multiple-link paths. Today, we look back and say that this is completely obvious, but in those days engineers were not used to making even mildly conceptual arguments of this type. Since the argument was very strong, and there were many tens of decibels to be gained, PCM and other digital systems started to become the norm.

It is probable that this paper had a greater impact on actual communication practice at the time than [1], [2]. However, [1], [2] has certainly had a greater impact in the long run. Also, the advantages of PCM would have certainly been explained by someone other than Shannon, whereas it is difficult to conceive of someone else discovering the results of [1], [2].

The second major paper written at about the same time as [1], [2] is "Communication in the Presence of Noise" [12]. This is a more tutorial amplification of the AWGN channel results of [2]. This paper reiterates the sampling theorem, now in a geometric signal-space perspective. The coding theorem for AWGN channels is proven in detail, using the geometry of orthogonal signals and the spherical symmetry of the noise. Finally, the theory is extended to colored Gaussian noise, and the famous power allocation result now known as "water-pouring" is derived.

This was the paper that introduced many communication researchers to the ideas of information theory. The notions of discrete sources and channels were not very familiar at that time, and this paper was more accessible to people accustomed to analog communication.

B. Shannon's Later Communication Work

After these 1948–1949 papers, Shannon turned his attention away from information theory for several years while he made some major contributions to switching, artificial intelligence, and games. During this interval, he wrote a few short tutorial papers on information theory, and published "Prediction and Entropy of Printed English," [14], which greatly expanded on the early results on this topic in [1]. However, his next major contributions to information theory came in the mid-1950s.

The first of these papers is "The Zero-Error Capacity of a Noisy Channel" [15], a delightful puzzle-type paper whose nature is primarily combinatoric. When no errors at all are permitted, the probabilistic aspects of channel coding disappear, and only graph-theoretic aspects remain. Surprisingly, the zero-error capacity seems to be harder to determine than the ordinary capacity. Also, it was shown that feedback from receiver to transmitter can increase the zero-error capacity of memoryless channels, which, surprisingly, is not true for the ordinary capacity.

The second is "Certain Results in Coding Theory for Noisy Channels" [16], presented at a conference in 1955 and published in 1957. The main thrust of this paper was to show that the probability of error could be made to decrease exponentially with code block length at rates less than capacity.

The coding theorem of [1] was originally presented as an asymptotic result, with a proof that suggested that very long constraint lengths would be required to achieve low error probability at rates close to capacity. In 1955, coding theory was still in its infancy, and no one had much sense of whether the coding theorem was simply a mathematical curiosity, or would someday transform communications practice. Coding theorists had attempted to find the best codes as a function of block length, but without success except in a few very special cases. Information theorists therefore began to seek upper and lower bounds on error probability as exponential functions of block length.

The first three such results, [35], [17], [16], appeared in 1955. The first, by Feinstein [35], showed that error probability decreases exponentially with block length for $R < C$, but was not explicit about the exponent. Elias [17] then developed the random coding upper bound and the sphere-packing lower bound for the binary symmetric channel and showed that the exponent in these two bounds agree between a certain critical rate and capacity. He also showed that this random coding bound applies to linear codes, encouraging continued linear code research. Finally, he invented convolutional codes and showed that they also could achieve the same asymptotic performance.

Shannon's paper [16], presented at the same conference as [17], used Chernoff bounds to develop an exponential random coding bound for the general DMC and some finite-state channels. Shannon's bounds were not as tight as later results, but his techniques and insights led to those later results.

The third of Shannon's later major papers on information theory is "Probability of Error for Optimal Codes in a Gaussian Channel" [18]. This paper was concerned with the exponential dependence of error probability on block length for the AWGN channel. This paper was unusual for Shannon, in that the ideas were carried through with a high level of detail, with careful attention not only to exponents but also to numerical coefficients.

This paper was the first to introduce an expurgated form of the random coding bound for transmission rates close to zero. The sphere-packing bound was also improved for rates near zero. These were some of the major new ideas needed for later work on error probability bounds. In addition, Shannon considered codes with three different constraints on the set of codewords, first, equal-energy, then peak-energy, and finally average-energy. The results were substantially the same in all cases, and one might argue this set the stage for later constant-composition results.

The fourth is "Coding Theorems for a Discrete Source with a Fidelity Criterion" [13]. This is an expansion of the results at the end of [2]. Shannon began here with a simple discrete source with i.i.d. letters and a single-letter distortion measure, and gave a simple and detailed proof of the rate-distortion theorem. He then generalized to more general sources and distortion measures, finally including analog sources.

The fifth paper in this sequence is "Two-Way Communication Channels" [19]. This applies information theory to channels connecting two points A and B for which communication is desired in both directions, but where the two directions interfere with each other. This was the first of a long string of papers on what is now called multiuser or network information theory.

The most striking thing about this paper is how surprisingly hard the problem is. The most basic information-theoretic problem here is to find the capacity region for the channel, i.e., the maximum rate at which B can transmit to A as a function of the rate from A to B. Shannon showed that the region is convex, and established inner and outer bounds to the region; however, in many very simple cases, the region is still unknown.

Fortunately, nicer results were later developed by others for multiple-access and broadcast channels. It is interesting to note, though, that Shannon stated, at the end of [19], that he would write another paper discussing a complete and simple solution to the capacity region of multiple-access channels. Unfortunately, that later paper never appeared.

The final paper in this set is "Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels" [20], [21], coauthored with the present author and E. R. Berlekamp. This was Shannon's final effort to establish tight upper and lower bounds on error probability for the DMC. Earlier, Robert Fano [22] had discovered, but not completely proved, the sphere-packing lower bound on error probability. In [20], [21], the sphere-packing bound was proven rigorously, and another lower bound on error probability was established which was stronger at low data rates. The proof of the sphere-packing bound given here was quite complicated; it was later proven in [23] in a simpler way.

IV. Shannon's Research Style

The great mathematician Kolmogorov summed up Claude Shannon's brilliance as a researcher very well. He wrote: "In our age, when human knowledge is becoming more and more specialized, Claude Shannon is an exceptional example of a scientist who combines deep abstract mathematical thought with a broad and at the same time very concrete understanding of vital problems of technology. He can be considered equally well as one of the greatest mathematicians and as one of the greatest engineers of the last few decades."

While recognizing his genius, however, many mathematicians of the day were frustrated by his style in [1], [2] of omitting precise conditions on his theorems and omitting details in his proofs. In Section II, we repeated some of his major proofs, partly to show that the omitted details are quite simple when the theorems are specialized to simple cases such as the DMC.

It appears that Shannon's engineering side took the dominant role in his theorem/proof style here. It was clear that DMCs are not sufficiently general to model interesting phenomena on many interesting real channels. It was also clear that finite-state channels are sufficiently general to model those phenomena. Finally, given Shannon's almost infallible instincts, it was clear that the coding theorem was valid for those finite-state channels appropriate for modeling the real channels of interest.

Was this theorem/proof style, with occasionally imprecise conditions, another stroke of genius or a failing? Bear in mind that this paper contained the blueprint of communication systems for at least the subsequent 50 years. It also explained clearly why all of these major results are true, under at least a broad range of conditions. Finally, the ideas form a beautiful symphony, with repetition of themes and growing power that still form an inspiration to all of us. This *is* mathematics at its very best, as recognized by Kolmogorov. If these theorems had been stated and proven under the broadest possible conditions, the paper would have been delayed and would probably have been impenetrable to the engineers who most needed its unifying ideas.

What was it that made Shannon's research so great? Was he simply such a towering genius that everything he touched turned to gold?

In fact, Shannon's discoveries were not bolts from the blue. He worked on and off on his fundamental theory of communication [1], [2] from 1940 until 1948, and he returned in the 1950s and 1960s to make improvements on it [13], [16], [18], [20], [21]. This suggests that part of his genius lay in understanding when he had a good problem, and in staying with such a problem until understanding it and writing it up.

This is not to suggest that Shannon listed various problems of interest, ordered them in terms of importance or interest, and then worked on them in that order. Rather, he worked on whatever problem most fascinated him at the moment. This might mean working on a curious aspect of some game, extending his theory of communication, thinking about artificial intelligence, or whatever. A look at his bibliography makes clear how many complementary interests he had. Working on whatever is currently fascinating might seem a little frivolous and undisciplined, but fortunately Shannon was guided by superb instincts.

Claude Shannon tended to be fascinated by puzzles and toy problems that exemplified more generic problems. He was fascinated not by problems that required intricate tools for solution, but rather by simple new problems where the appropriate approach and formulation were initially unclear. He would often consider many problems, in various stages of understanding, in his mind at once. He would jump from one to the other as new clues jumped into his mind. In the case of [1], [2], where many totally new ideas had to be fitted together, this gestation process required eight years. In other simpler cases, such as the seminar course he gave at MIT, a new idea was developed and presented twice a week.

Shannon was also fascinated by developing mathematical theories for subjects (e.g., switching, communication, cryptography, the stock market). This was closely related to his fascination with puzzles, since in both cases the end point was understanding the right way to look at a topic. He would approach this with toy models, sometimes conceptual, sometimes physical. The toy models would then lead to generalizations and new toy models.

Shannon's research style combined the very best of engineering and mathematics. The problems that fascinated him were engineering problems (in retrospect, even chess is a toy version of an important engineering problem). Abstraction and generalization, focusing on both simplicity and good approximate models, are the essence of both mathematics and engineering. Turning them into an elegant mathematical theory is, of course, great mathematics.

Shannon did not like to write, but he wrote very well, with remarkable clarity and ability to convey his sense of delight in problems and their solutions. He was not interested in the academic game of accruing credit for individual research accomplishments, but rather with a responsibility for sharing his ideas. He would state results as theorems, but was clearly more interested in presenting the idea than in the precise statement or proof.

A. Can Shannon's Research Style Be Cloned?

Many information theory researchers seem to have absorbed some aspects of Claude Shannon's research style. The combination of engineering and mathematics, the delight in elegant ideas, and the effort to unify ideas and tools are relatively common traits that are also highly admired by others.

The more controversial trait that we focus on here is Shannon's habit of working on whatever problem fascinated him most at the time. A more colorful expression is that he followed his nose. More specifically, he followed his nose in uncharted areas where the biggest problem was to understand how to look at the problem. We call this Shannon-style research. We should not confuse this with ivory-tower research, since Shannon's research remained very close to engineering topics.

Should we encourage ourselves, and encourage others, to try to do Shannon-style research? An easy, but I think dangerous, answer is that Shannon earned the right to follow his nose from his early research successes. In this view, people who have not earned that right should be expected to do goal-oriented research, i.e., to solve well-posed problems.

The difficulty with this view is that goal-oriented research (unless the goal is quite broad and the time scale long) provides little guidance in how to follow one's nose successfully. Engineering education also provides little or no guidance. Engineering students are trained to solve restricted classes of problems by learning algorithms that lead them through long calculations with little real thought or insight.

In graduate school, doctoral students write a detailed proposal saying what research they plan to do. They are then expected to spend a year or more carrying out that research. This is a reasonable approach to experimental research, which requires considerable investment in buying and assembling the experimental apparatus. It is a much less reasonable approach to Shannon-style research, since writing sensibly about uncharted problem areas is quite difficult until the area becomes somewhat organized, and at that time the hardest part of the research is finished.

My belief is that we should encourage both ourselves and others to acquire and improve the ability to do Shannon-style research. This is the kind of research that turns an area from an art into a science. Many areas of telecommunication technology are still primarily arts, and much of the network field is an art. Shannon-style research is relatively rare and desperately needed in these areas.

Shannon rarely wrote about his research goals. In learning to do Shannon-style research, however, writing about goals in poorly understood areas is very healthy. Such writing helps in sharing possible approaches to a new area with others. It also helps in acquiring the good instincts needed to do Shannon-style research. The development of good instincts is undoubtedly more valuable for a researcher than acquiring more facts and techniques.

Fortunately, the information theory field has a sizable number of senior and highly respected researchers who understand both the nature and the value of Shannon-style research. Effort is always needed, of course, in educating research administrators in the distinct character and long-term value of this style.

In summary, Shannon is a notable exemplar of an instinct-driven style of research which has had remarkable results. It is important to encourage this style of research in a variety of engineering fields.

V. Shannon's Impact on Telecommunication

For the first quarter century after the publication of "A Mathematical Theory of Communication," information theory was viewed by most informed people as an elegant and deep mathematical theory, but a theory that had relatively little to do with communication *practice*, then or future. At the same time, it was quickly recognized as the right theoretical way to view communication systems, as opposed to various mathematical theories such as probability, filtering, optimization, etc., that dealt only with isolated aspects of communication (as well as with aspects of many other fields). In fact, when [1], [2] were republished in book form [24][9] the following year, "A Mathematical Theory" had been replaced by "The Mathematical Theory."

In more recent years, the recognition has been steadily growing that information theory provides the guiding set of principles behind the practice of modern digital communication. On the other hand, it is difficult to separate the roles of economics, politics, entrepreneurship, engineering, and research in the growth of new technologies. Thus, we cannot be definitive about Shannon's impact, but can only suggest possibilities.

In what follows, we first discuss Shannon's impact on information theory itself, then his impact on coding theory, and, finally, the impact of information theory and coding theory on communication technology.

A. The Evolution of Information Theory

The birth of information theory in 1948 led to intense intellectual excitement in the early 1950s. The Institute of Radio Engineers (a precursor to the IEEE) started to publish occasional issues of the Transactions on Information Theory, which became regular in 1956. There were also a number of symposia devoted to the subject. The people working in this nascent field were quite interdisciplinary, probably more so than today. There were mathematicians trying to give rigorous proofs to precise statements of the major theorems, there were physicists trying to interpret the theory from the entropy concepts of statistical mechanics, there were engineers curious about applicability, and there were people from many fields entranced by the word "information."

In order to understand the mathematical issues, we need to understand that probability theory had been put on a rigorous measure-theoretic foundation by Kolmogorov only in 1933. Despite Kolmogorov's genius and insight, mathematical probability theory remained quite a formal and unintuitive subject until Feller's 1950 book [26] showed how to approach many simpler problems with simple but correct tools. Before 1950, much of the nonmathematical literature on probability was vague and confused. Thus, it is not surprising that mathematicians felt the need to generalize and reprove Shannon's basic theorems in formal measure-theoretic terms.

McMillan [36] generalized the source coding theorem from ergodic Markov sources to general ergodic sources in 1953. Similarly, Feinstein [34] gave a rigorous, measure-theoretic proof of the noisy channel coding theorem for memoryless channels in 1954. One of the reasons that we repeated Shannon's original proof of these two major theorems (with some added details) was to clarify the simple elegance of his proof from both a mathematical and engineering perspective. Ultimately, it was the simplicity of Shannon's ideas which led to engineering understanding.

[9] Warren Weaver was a coauthor on the book, but his only contribution was to write a short introduction.

One of the difficulties that arose from these and subsequent mathematical attempts was an increasing emphasis on limit theorems. As we noted before, sources are not ergodic in reality, and neither are channels. It is only the models that have these properties, and the models must provide insight about the reality. Differences between different types of convergence and small differences between the generality of a class of models do not always provide such insight.

Fortunately, in the more recent past, pure mathematicians and engineers have usually worked in harmony in the information theory field. Pure mathematicians now often pay attention to modeling issues, engineers often pay attention to mathematical precision, and the two talk to each other about both models and precision. Even more important, there are more and more researchers in the field who, like Shannon, are equally comfortable with both engineering and mathematics. It appears that Shannon is largely responsible for this harmony, since he understood both mathematics and engineering so well and combined them so well in his work.

The efforts of physicists to link information theory more closely to statistical mechanics were less successful. It is true that there are mathematical similarities, and it is true that cross pollination has occurred over the years. However, the problem areas being modeled by these theories are very different, so it is likely that the coupling will remain limited.

In the early years after 1948, many people, particularly those in the softer sciences, were entranced by the hope of using information theory to bring some mathematical structure into their own fields. In many cases, these people did not realize the extent to which the definition of information was designed to help the communication engineer send messages rather than to help people understand the meaning of messages. In some cases, extreme claims were made about the applicability of information theory, thus embarrassing serious workers in the field.

Claude Shannon was a very gentle person who believed in each person's right to follow his or her own path. If someone said something particularly foolish in a conversation, Shannon had a talent for making a reasonable reply without making the person appear foolish. Even Shannon, however, was moved to write an editorial called the "Bandwagon" in the TRANSACTIONS ON INFORMATION THEORY [27] urging people, in a very gentle way, to become more careful and scientific.

In later years, applications of information theory to other fields, and *vice versa*, has been much more successful. Many examples of such interdisciplinary results are given in [9].

B. Coding Theory

It is surprising that Shannon never took great interest in coding techniques to achieve the results promised by his theory. At the same time, however, his results provided much of the motivation for coding research and pointed the direction for many of the major achievements of coding. At a fundamental level, the coding theorems and the promise of digital communication provided a direct motivation for discovering both source and channel codes that could achieve the promise of information theory.

In source coding, for example, Huffman coding is simple and beautiful, but clearly depends on Shannon's early example of source coding. Universal source coding has been an active and important research field for many years, but it depends heavily on both the source modeling issues and the typical sequence arguments in [1]. There is less evidence that modern voice compression depends heavily on rate-distortion theory.

Error-correction coding can be divided roughly into two parts, algebraic techniques and probabilistic techniques. Both depend on [17], which depends on [1], for the assurance that linear codes and convolutional codes are substantially optimum. Other than this, however, algebraic coding does not depend heavily on [1].

Probabilistic coding techniques, on the other hand, depend heavily on Shannon's work. Both Viterbi decoding and sequential decoding are based on the premise that most convolutional codes are good, which comes from [17] and [1]. One can argue that two important insights in the development of turbo codes are that most codes of a given constraint length are relatively good, and that error probability goes down rapidly with constraint length. These insights come from [1], [17], and [16]. Low-density parity-check codes are very directly dependent on [1], [17], and [16].

C. The Evolution of Practical Applications

For many years after 1948, both information theory and coding theory continued to advance. There were a few high-end applications, but integrated-circuit technology was not sufficiently advanced for economic large-scale commercial coding applications. Indeed, the theory also matured very slowly.

There was a certain impatience in the 1960s and 1970s with the length of time that it was taking for the theory to become practical. When I received my doctorate in 1960, several people suggested that information theory was dying. In 1970, many attendees at a major workshop on communication theory seriously considered the proposition that the field was dead. Curiously, this was just about the time that digital hardware prices were falling to the point that coding applications were becoming practical (as Irwin Jacobs pointed out at that workshop).

Today we are constantly amazed by how quickly new technologies and applications are developed, and then replaced by yet newer technologies and applications. Many leaders of industry and academy seem to accept without question the message that engineers must become increasingly quick and nimble, and that *research must become equally quick.*

I believe that this latter view is absolutely wrong, and shows an alarming lack of appreciation for Shannon-style research. The history of digital communications (and many other fields) illustrates why this view is wrong. Today, digital communications has matured to the point that there are many established methods of accomplishing the various functions required in new systems. Given this toolbox, new developments and applications can be truly rapid. However, a modern cellular system, for example, is based on various concepts and algorithms that have been developed over decades, which in turn depend on research going back to 1948.

Shannon developed his communication theory from 1940 until 1948. During the 50 years since then, communication theory has developed to its present stage, with seamless connections from abstract mathematics to applications. Imagine a research administrator in 1945 saying, “Hurry up, Claude. We can’t wait forever for your theory—we have real systems to build.”

Shannon-style research, namely, basic research that creates insight into how to view a complex, messy system problem, moves slowly. It requires patience and reflection. It can easily be destroyed in an atmosphere of frantic activity.

We are all aware of the rapid product cycles in modern engineering companies. New product generations are being designed before the previous generation is even released. Basic research must be carried on outside the critical paths of these product cycles. The results of basic research can subsequently be inserted into the product cycle after it is sufficiently digested. This requires that researchers be sufficiently close to product development.

Shannon frequently said that he was not particularly interested in applications, but rather was interested in good or interesting problems. It is too late to ask him what he meant by this, but I suspect he was saying that he was not interested in getting involved in the product cycle. He was clearly interested in the generic issues that needed to be understood, but realized that these issues proceeded by their own clock rather than that of the product cycle.

It is clear that engineering, now as in the past, requires a range of talents, including product engineers, generalists who recognize when ideas are ripe for products, mathematicians who refine and generalize theories, and Shannon-style researchers who provide the conceptual structure on which all the above is based. The danger today is that the Shannon-style researcher may not be appreciated in either industrial or academic environments.

Any engineering professor in a university today who proclaimed a lack of interest in applications would become an instant pariah. However, it appears that our field is an excellent example of a field where long-term research on “good problems” has paid off in a major way. We should probably use this example to help educate academic, government, and industrial leaders about the nature and benefit of Shannon-style research. And we should be very thankful to Claude Shannon for giving us such a good example.

Acknowledgment

The author wishes to thank D. Forney for countless suggestions for sharpening the meaning, rephrasing the awkward expressions, and changing the overall organization of this paper. J. Massey has also contributed greatly. Other very helpful reviewers include I. Csiszar, S. Savari, B. Schein, E. Telatar, and D. Tse.

References

[1] C. E. Shannon, “A mathematical theory of communication (Part 1),” *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.

[2] ——, “A mathematical theory of communication (Part 2),” *Bell Syst. Tech. J.*, vol. 27, pp. 623–656, 1948.

[3] ——, “A symbolic analysis of relay and switching circuits,” *Trans. AIEE*, vol. 57, pp. 713–723, 1938.

[4] ——, “Communication theory of secrecy systems,” *Bell Syst. Tech. J.*, vol. 28, pp. 656–715, 1949.

[5] R. V. L. Hartley, “Transmission of information,” *Bell Syst. Tech. J.*, vol. 7, p. 535, 1924.

[6] A. Lempel and J. Ziv, “On the complexity of finite sequences,” *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 75–81, Jan. 1976.

[7] J. Ziv and A. Lempel, “Compression of individual sequences by variable-rate coding,” *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530–536, Sept. 1978.

[8] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[10] H. Nyquist, “Certain topics in telegraph transmission theory,” *Trans. AIEE*, vol. 47, p. 617, 1928.

[11] B. M. Oliver, J. R. Pierce, and C. E. Shannon, “The philosophy of PCM,” *Proc. IRE*, vol. 36, pp. 1324–1331, 1948.

[12] C. E. Shannon, “Communication in the presence of noise,” *Proc. IRE*, vol. 37, pp. 10–21, Jan. 1949.

[13] ——, “Coding theorems for a discrete source with a fidelity criterion,” in *IRE Nat. Conv. Rec.*, 1959, pp. 142–163.

[14] ——, “Prediction and entropy of printed English,” *Bell Syst. Tech. J.*, vol. 30, pp. 50–64, 1951.

[15] ——, “The zero-error capacity of a noisy channel,” *IRE Trans. Inform. Theory*, vol. IT-2, pp. S8–S19, Sept. 1956.

[16] ——, “Certain results in coding theory for noisy channels,” *Inform. Contr.*, vol. 1, pp. 6–25, 1957.

[17] P. Elias, “Coding for noisy channels,” in *IRE Conv. Rec.*, 1955, pp. 37–46.

[18] C. E. Shannon, “Probability of error for optimal codes in a Gaussian channel,” *Bell Syst. Tech. J.*, vol. 38, pp. 611–656, 1959.

[19] ——, “Two-way communication channels,” in *Proc. 4th Berkeley Symp. Probability and Statistics, June 20–July 30, 1960*, J. Neyman, Ed. Berkeley, CA: Univ. Cal. Press, 1961, vol. 1, pp. 611–644.

[20] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, “Lower bounds to error probability for coding on discrete memoryless channels I,” *Inform. Contr.*, vol. 10, pp. 65–103, 1967.

[21] ——, “Lower bounds to error probability for coding on discrete memoryless channels II,” *Inform. Contr.*, vol. 10, pp. 522–552, 1967.

[22] R. M. Fano, *Transmission of Information*. Cambridge, MA: MIT Press/Wiley, 1961.

[23] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. London, U.K.: Academic, 1981.

[24] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL: Univ. Illinois Press, 1949.

[25] A. N. Kolmogorov, *Foundations of the Theory of Probability*. New York: Chelsea, 1950, 2nd. ed. 1956.

[26] W. Feller, *An Introduction to Probability Theory and its Applications*. New York: Wiley, 1950, vol. I.

[27] C. E. Shannon, “The bandwagon: Editorial,” *IRE Trans. Inform. Theory*, vol. IT-2, p. 3, Mar. 1956.

[28] ——, “Mathematical theory of the differential analyzer,” *J. Math. Phys.*, vol. 20, pp. 337–54, 1941.

[29] ——, “The theory and design of linear differential machines,” Report to the Services 20, Div 7-311-M2, Bell Labs, Jan. 1942.

[30] ——, “Programming a computer for playing chess,” *Philos. Mag.*, ser. 7, vol. 41, pp. 256–275, Mar. 1950.

[31] ——, “A chess-playing machine,” *Scientific Amer.*, vol. 182, pp. 48–51, Feb. 1950.

[32] ——, “A universal Turing machine with two states,” Memo. 54-114-38, Bell Labs., 1954.

[33] E. F. Moore and C. E. Shannon, “Reliable circuits using crummy relays,” Memo. 54-114-42, Bell Labs., 1954. (Republished in the *J Franklin Inst.*, Sept.–Oct. 1956).

[34] A. Feinstein, “A new basic theorem of information theory,” *IRE Trans. Inform. Theory*, vol. PGIT-4, pp. 2–22, Sept. 1954b.

[35] ——, “Error bounds in noisy channels without memory,” *IRE Trans. Inform. Theory*, vol. IT-1, pp. 13–14, Sept. 1955.

[36] B. McMillan, “The basic theorems of information theory,” *Ann. Math. Statist.*, vol. 24, pp. 196–219, 1953.