BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 16, Number 1, January 1987

ON THE DYNAMICS AND EVOLUTION OF SOME SOCIOTECHNICAL SYSTEMS

ELLIOTT W. MONTROLL

edited by
BRUCE J. WEST

1. Introduction. With statistical mechanics having been a central theme in my research program, it gives me special pleasure to have been invited to lecture in memory of one of its great founding fathers, J. Willard Gibbs. I feel personally closer to Gibbs than to two of the other fathers, Maxwell and Boltzmann. During the late 1940s and early 1950s, I had the good fortune of having many discussions with Edwin Bidwell Wilson, collaborator on the Gibbs-Wilson vector analysis book, probably the Gibbs student closest to the master, and eighth Gibbs lecturer. Since Wilson's favorite topics for discussion were Gibbs and the National Academy of Sciences, after n of these I began to feel that Gibbs was a third grandfather, one whom I never had the joy of knowing, and that I was prematurely a member of the inner circle of the National Academy of Sciences. Incidentally, Wilson (author of a once-popular advanced calculus book, editor of the *Proceedings of the National Academy of Sciences* during the period when it evolved into an internationally prominent journal, and for a time professor of vital statistics at the Harvard School of Public Health) considered himself to be the middleman of statistical mechanics in the United States. He was Gibbs' student and a teacher of Richard Tolman, whose treatises on the subject were classics of the 1920s and 1930s. Wilson, while head of the MIT Physics Department in the early 1920s, hired J. S. Slater, who became a teacher of Jack Kirkwood, the man who directed more Ph.D. students and postdoctorals in statistical mechanics than any other American professor.

Received by the editors October 23, 1984.

1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 90B20, 93A15, 93C10, 94A15.

This paper was presented as the 55th Josiah Willard Gibbs Lecture on January 13, 1982 at the 88th annual meeting of the American Mathematical Society in Cincinnati, Ohio. Professor E. W. Montroll died on December 3, 1983 before the lecture notes had been put in final form. Various phases of editing the manuscript were done by his friends at the Institute for Physical Science and Technology, University of Maryland, especially Dr. M. F. Shlesinger and Marilyn Spell. The final editing, including the addition of the appendix based on some of Professor Montroll's earlier work, was done by his friend and former student Bruce J. West. Dr. West claims full responsibility for any errors, misrepresentations and/or omissions in the final manuscript.

©1987 American Mathematical Society
0273-0979/87 $1.00 + $.25 per page

Wilson also liked to identify himself as Calvin Coolidge's scientific advisor. Apparently, Coolidge enjoyed reading the rotogravure section of the Sunday newspapers, which frequently contained some picture news about science. Wilson was not surprised on a Monday morning when he got a phone call from Coolidge querying him about one of the science articles. Ever so seldomly Coolidge would also consult Wilson about some science-directed bill that he was asked to sign.

The broad subject from which this lecture will draw special topics, the role of mathematics in the social sciences, was one that especially interested Gibbs in his later years even though he did not make direct contributions to it himself. One of his last students was Irving Fisher who wrote, partly under Gibbs' direction, a Ph.D. thesis entitled *Mathematical investigations in the theory of value and prices*. To quote Fisher:[1]

> Professor Gibbs showed a lively interest in this youthful work and was especially interested in the fact that I had used geometric constructions and methods including his own vector notation.

I recently turned to the April 1930 *Bulletin of the American Mathematical Society* to reread the Irving Fisher seventh Gibbs Lecture,[1] *The application of mathematics to the social sciences*, delivered to the Society on the 29th of December 1929, and I realized that a bit of trauma must have entered Fisher's life between the date he received the invitation to speak and the day of the lecture. Fisher was the chief academic spokesman for the wonders of the stock market boom of the late 20s. As one of the major academic economists and an eternal optimist, he was frequently invited to lecture to business and investment groups on the future of the market and was much quoted in the newspapers and on the radio. His state of mind during the winter of 1929 may be deduced from three of the pronouncements that I extracted from John Galbraith's *The Great Crash*.[2]

October 15th:

> "Stock prices have reached what looks like a permanently high plateau. I expect to see the market a good deal higher than it is today..."

October 21st:

> "The decline represents only a shaking out of the lunatic fringe... the market has not yet reflected the beneficent effects of prohibition which has made the American worker more productive and dependable."

TUESDAY, OCTOBER 29TH WAS THE WORST DAY IN TRADING HISTORY.

November 3rd:

> "It was the Psychology of Panic. It was mob psychology and it was not primarily that the price level was unsoundly high. The fall of the market was very highly due to the psychology by which it went down because it went down."

The crash notwithstanding, Fisher's Gibbs Lecture was an excellent review of the history and the current status of attempts to apply mathematics to the social sciences; its published version may still be read with profit by those interested in the history of ideas and applied mathematics.

In addition to reflecting Gibbs' late interests, I have also selected the subject of this lecture in partial response to a plea of the late President Handler of the National Academy of Sciences in his retiring presidential address of April, 1981:[3]

> ...what I would particularly like to direct to your attention is the pressing need... for the development of sophisticated analytical approaches to large sociotechnical systems.

2. The entropy function in sociotechnical systems. A major contribution of Ludwig Boltzmann, who with Gibbs is considered a founding father of statistical mechanics, was the identification of a statistical construct with the thermodynamic entropy of a material system, the first example of such a system that he examined being a perfect gas[4] (1877). His construct gave a measure of randomness or disorder in the system and allowed him to take the view that a complex system of atoms and molecules achieved as random a state as possible consistent with constraints introduced by conservation laws, i.e., conservation of numbers of particles and of the total energy of the system.

Basically, the function that became ideal for this view was

$$H = -\sum_{i=1}^{N} p_i \log p_i \quad \text{with } p_i \geqslant 0, \tag{1}$$

a function we shall call the entropy function. Usually the p_i's are postulated to be normalized so that

$$\sum_{i=1}^{N} p_i = 1. \tag{2}$$

Frequently, the subscript i is identified with a possible state of a physical system, with the p_i representing the probability that the system achieve the ith state. The fact that H measures the degree of randomness in a system may be seen by comparing its value for two extreme cases. If all states are equally likely, $p_i = 1/N$ of all i, a most random case

$$H = \log N. \tag{3a}$$

The larger the N, the larger the entropy; i.e., the more random the system, the larger the entropy. In the most specified, least random case, the system is certainly in a single state, say the state j with probability 1 so that $p_j = 1$ with all other p_i being 0. Then

$$H = 0. \tag{3b}$$

Thus absolute certainty has the minimum entropy.

The definition of H may be extended to the case that the states are represented by a continuous variable x. If x has the range $-\infty \leqslant x \leqslant \infty$ then we define H as

$$H = -\int_{-\infty}^{\infty} p(x)\log p(x)\,dx, \tag{4a}$$

$p(x)\,dx$ being the probability that the variate lies between x and $x + dx$, with

$$\int_{-\infty}^{\infty} p(x)\,dx = 1. \tag{4b}$$

It is easy to show (and was already known to Boltzmann) that when the dispersion σ of $p(x)$, is given so that

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 p(x)\,dx \equiv \langle x^2 \rangle, \tag{5}$$

the function that maximizes the entropy is the Gauss distribution

$$p(x) = (2\pi\sigma^2)^{-1/2}\exp(-x^2/2\sigma^2) \quad \text{so that } H = \log\sigma(2\pi e)^{1/2}. \tag{6}$$

The larger σ, the larger H. When x is restricted to the positive half-line and the mean value of x is given by the value μ,

$$\mu = \int_{0}^{\infty} xp(x)\,dx \equiv \langle x \rangle, \tag{7}$$

it can be shown that the distribution $p(x)$ that maximizes the entropy is the exponential

$$p(x) = \mu^{-1}\exp(-x/\mu). \tag{8}$$

Both the Gaussian and exponential distributions are basic in statistical mechanics. In the case of a perfect gas the kinetic energy, which is conserved in molecular collisions, is quadratic in the momentum. If the average value of the sum of the squares of the momentum is constant and the entropy function is maximized under this auxiliary condition, one obtains the Maxwell distribution function for the moment or the velocity of an individual molecule. The canonical ensemble distribution of the energy of a system is the exponential.

For seventy years the domain of the entropy function was restricted to the field of statistical mechanics and the applications of that subject to the physical sciences, until in 1948 Claude Shannon[5] (36th Gibbs Lecturer), in elaborating on the pioneering work of Ralph Hartley and Harry Nyquist, identified the entropy function as an ideal measure of the information transferred in communication systems. Thus the entropy function appeared in the characterization of the output of a sociotechnical system.

Shannon observed that, if after long experience with message transfer through a communication channel incorporating a code that employs N symbols identified by $j = 1, 2, \ldots, N$, it is found that the jth symbol appeared with probability p_j, then the maximum information transfer rate becomes

$$H = -c\sum_{j=1}^{N} p_j \log p_j. \tag{9}$$

The constant c depends on the base of the logarithm and on the rate that the source device emits symbols. The information transmission rate may also be related to the continuum entropy when the message is propagated in a continuous wave form. Let W be the bandwidth of the transmitter. Then the message wave form may be Fourier analyzed as a linear combination of W harmonics. If an ensemble of continuous messages is coded so that the Fourier coefficients all have a Gauss distribution with a common dispersion, then the information transmission rate is proportional to H given in equation (6). This follows from the assertion following (5) concerning the Gauss distribution.

Noisy circuits carry less information. If P is the signal power and N is the power of interfering Gaussian noise, then, as Shannon (equation (9)) showed, the information transfer rate is proportional to $H = W \log[1 + (P/N)] \simeq WP/N$. The asymptotic form is valid when $N \gg P$, conforming to the engineers' rule of thumb that broad bandwidth circuits carry information at a higher rate and that a simple way to overcome noise is to enhance the signal.

The communication systems considered by Shannon were composed of a message input element, a transmission channel, and a message output element. Since the entropy function appeared in a natural way for the information transfer rate in such a system, we might ask if the function could also be important in other sociotechnical systems that are composed of an analogous set of three components.

A. ENTROPY FUNCTION IN A TRAFFIC STREAM.[6] An example that immediately comes to mind is a highway transportation system that has an input provision for a vehicle and a road providing the channel for travel to an exit point. About twenty-five years ago, Robert Herman, the author, and colleagues at the General Motors Research Center performed car-following experiments and made numerous observations on flow on single-lane roads (and multilane highways under high-density conditions so that weaving from one lane to another was rare).[7-9] I now show that an entropy function evolves naturally from the observed stimulus-response equation that describes the manner in which a car follows its predecessor in a platoon.

Let us consider a platoon of N cars identified as $n = 1, 2, \ldots, N$ flowing along a long, single-lane highway void of traffic signals. An equation found to describe with remarkable accuracy the response of a follower (identified by $n + 1$) to the behavior of a leader (identified by n) is[7-10]

$$d\mathrm{v}_{n+1}(t + \Delta)/dt = \lambda_0 \left\{ \frac{\mathrm{v}_n(t) - \mathrm{v}_{n+1}(t)}{x_n(t) - x_{n+1}(t)} \right\} \tag{10}$$

in which $\mathrm{v}_n(t)$ is the velocity of car n at time t, $x_n(t)$ is the location of the front end of that car at time t, and Δ is the time lag between the stimulus provided by the lead car and the response by the follower. The time lag Δ, which varies from person to person, is about 1.5 seconds. Equation (10) is a quantitative reflection of the fact that an $(n + 1)$st driver accelerates when his relative speed is too slow and decelerates when it is too fast. When the driver he follows is far ahead of him, his response is not as sensitive as when close.

Integration of the stimulus-response equation (10) yields an equation of state for traffic, a relationship between vehicular flow rate and density in single-lane traffic.[10,11] First we integrate (10) between t_1 and t_2 to obtain, for all n,

$$(11)\quad v_{n+1}(t_2 + \Delta) - \lambda_0 \log\left[x_n(t_2) - x_{n+1}(t_2)\right] = v_{n+1}(t_1 + \Delta) - \lambda_0 \log\left[x_n(t_1) - x_{n+1}(t_1)\right]$$

so that

$$(12)\quad v_{n+1}(t + \Delta) - \lambda_0 \log d_{n+1}(t) = \text{constant}$$

with

$$(13)\quad d_{n+1}(t) \equiv x_n(t) - x_{n+1}(t)$$

being the space available per car at the location between the nth and the $(n + 1)$st cars at time t. The constancy of equation (12) is a consequence of the left-hand side of (11) being a function of only t_2, and the right-hand side only of t_1.

The traffic density at the location of car n, $\rho_n(t)$, is the reciprocal of the space available per car: $\rho_n = 1/d_n$ = number of cars per unit length. In a freely moving stable stream of traffic, $v_n(t + \Delta)$ with $\Delta \approx 1.5$ sec is practically the same as $v_n(t)$ and (12) becomes (ρ_c being the bumper-to-bumper close packing density at which $v_n = 0$)

$$(14)\quad v_n(t) = -\lambda_0 \log\left[\rho_n(t)/\rho_c\right].$$

The local traffic flow rate (dropping the explicit dependence upon time) is then

$$(15)\quad q_n = \rho_n v_n = -\lambda_0 \rho_c (\rho_n/\rho_c) \log(\rho_n/\rho).$$

Notice that $0 < \rho_n/\rho_c \leqslant 1$ and that the dimensions of our variables might be cars per hour for q, cars per mile for ρ, and miles/hr for v. By averaging over N cars in a line of traffic, the mean flow rate is proportional to an entropy function in the variables (ρ_n/ρ_c),[6]

$$(16)\quad q = \lambda_0 \rho_c \left\{ -\frac{1}{N} \sum_{n=1}^{N} \frac{\rho_n}{\rho_c} \log\left(\frac{\rho_n}{\rho_c}\right) \right\}.$$

It is to be noted that the ratio ρ_n/ρ_c is not normalized since the sum over all ρ_n/ρ_c does not have to be 1 or N. It is possible to construct a set of normalized p_n by defining a mean density ρ by

$$(17a)\quad \rho = \frac{1}{N} \sum_{n=1}^{N} \rho_n.$$

Then the quantity $p_n = \rho_n/\rho N$ is positive and has the property

$$(17b)\quad \sum_{n=1}^{N} p_n = 1.$$

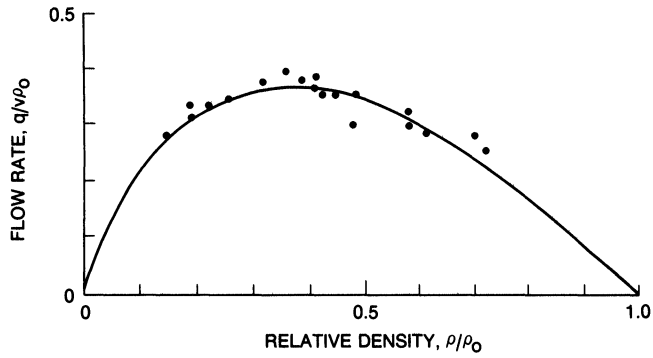

FIGURE 1. Variation of flow (cars per hour) with density. The points are taken from *Lincoln Tunnel Data* by H. Greenberg.[12] The curve is determined from equation (18).

The p_n may be introduced into (16) to yield

$$q = \rho\lambda_0 \log(\rho_c/\rho) + \rho\lambda_0\left\{\log\frac{1}{N} - \sum_{n=1}^{N} p_n \log p_n\right\}. \tag{18a}$$

The flow rate is a maximum at a given traffic density if all ρ_n are identical so that $\rho_n = \rho$ and $p_n = 1/N$. Then the term in the parenthesis in (18a) vanishes. However, if drivers behave differently from each other (as of course they do) so that some ρ_n deviate from ρ, the entropy term in (18a) does not achieve its maximum value and the term in the parenthesis is negative, yielding a reduction in the flow rate q.

If we set

$$\rho_n = \rho + \Delta\rho_n,$$

substitute it into (18a), remembering that $p_n = \rho_n/\rho N$, and assume that $\Delta\rho_n$ is small, then, since $\Sigma\Delta\rho_n = 0$,

$$q = \rho\lambda_0 \log(\rho_c/\rho) - \frac{\rho\lambda_0}{2N}\sum_n (\Delta\rho_n/\rho)^2. \tag{18b}$$

The throughput q is plotted as a function of ρ in Figure 1, omitting the negative contribution of the fluctuation term in (18b). The points on the graph were obtained from observations of traffic flow in a large sample of more than 24,000 vehicles in the Holland Tunnel in New York City.[12] The value of λ_0 that gave the best fit to the tunnel data was nineteen miles per hour.

B. ENTROPY FUNCTION IN THE CATALOGUES OF SEARS, ROEBUCK AND COMPANY. A communication system is composed of a message input element, a channel for message propagation, and a message output element. A highway transportation system has an input provision for a traveler and a road

providing the channel for travel to an exit point. We have seen that the flow-through rate in each of these systems may be related to an entropy function. Another important sociotechnical system is the merchandising system. Goods flow into a warehouse or distribution center of the retailing firm, remain temporarily as an inventory, and finally are carried out by or delivered to the customer. Hence, a company's profit depends upon the flow-through rate of goods and on the price associated with the goods. The similarity between merchandising flows and the previous two examples suggests that the entropy function may appear in an analysis of that process.

Further development of this idea requires merchandising data. Fortunately, Sears, Roebuck and Company (SR) has left us a rich legacy of information on this subject in its annual catalogues,[13] which form a magnificent data base of Americana of the past eighty-five years. Prices listed in the catalogues were generally right for their times and the items listed reflect the public taste of the time. At first, through its mail-order operation, the firm made available to the farm family products found in cities of medium to large size; then it tried to compete with city merchants for the urban trade. The catalogues may be regarded as a merchandise model of a medium-sized city, listing available goods at reasonable prices.

The preparation of the catalogues was a major concern of SR. Basically, each page was audited to produce its share of the profit. For example,[14] in 1930 the goals set ranged from \$5,000 to \$20,000 per page, depending upon the responsible merchandising department. Since the profit that year[15] was \$14,300,000 and the catalogues ran 1000 to 1500 pages, the profit per page averaged about \$10,000. Expensive goods, properly illustrated, often attracted attention to pages containing cheaper bargain items. Many pages reserved a small space for the tentative introduction of new products. If the response was favorable, the allocation increased the next year. As annual sales of an item declined, its space allocation decreased: sometimes it even disappeared completely from the catalogue. Various department heads, anxious for raises and promotions, were very competitive in the preparation of pages that listed items that were hoped to outsell those of their colleagues.

Although the SR catalogues have been woven into the lives of millions, Robert Herman and I[16] may have been the first to regard the lists of prices as a statistician's delight, to be exploited as a microcosm of the merchandising world. Motivated by reasons expressed in reference 16, we found the distribution function of prices by year listed in many of the catalogues. Since prices range from a few cents to hundreds of dollars, we "expanded" the scale of low-cost items and "contracted" that of higher-priced ones by recording the data as the logarithm of the price (to the base 2), $\log_2 P$. Of course, we were aware (as many before us dating back to D. Bernoulli) that $\log P$ is psychologically a more important variable than the price itself because one is especially sensitive to relative price changes, $(\Delta P)/P \simeq \Delta \log P$.

Examination of the price distribution from many catalogues indicates that, in a given catalogue, the distribution of $\log_2 P_i$ (P_i being the price of the ith item) is very close to the normal distribution[6]. Three examples are shown in Figure 2.[17] We also investigated the mean $\log_2 P$ and the dispersion of $\log_2 P_i$.

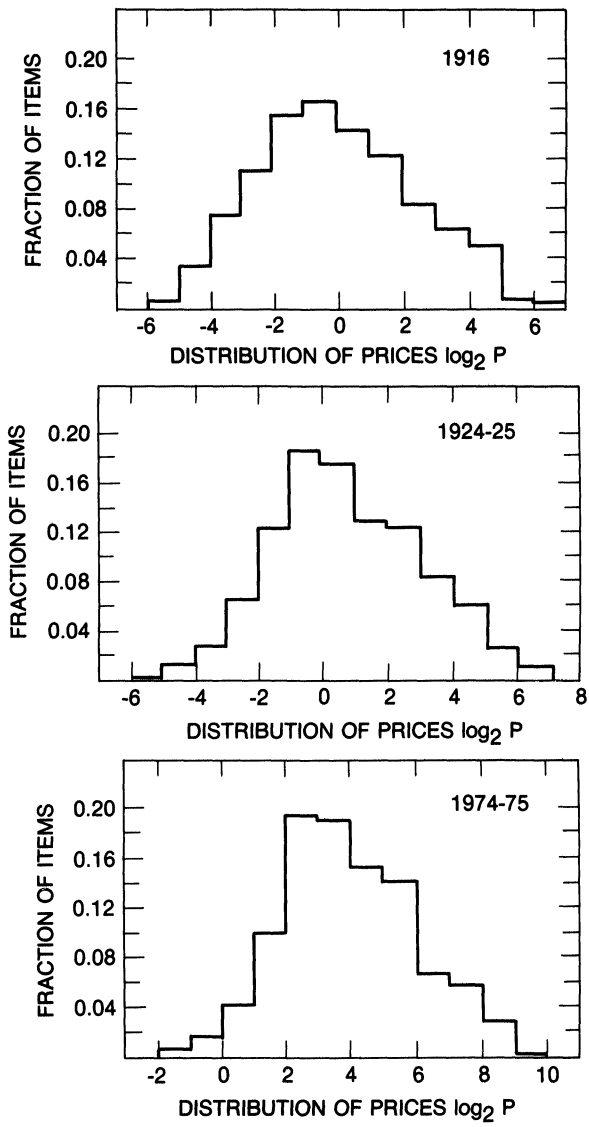

FIGURE 2. Histogram of distribution of prices in Sears Roebuck catalogues for years 1916, 1924–25, and 1974–75. The fraction of items in each price range in each catalogue is plotted as a function of $\log_2 P$, P being the price [from ref. 17].

If N is the number of prices sampled, we define

$$\log \bar{P} \equiv \langle \log_2 P \rangle \equiv \frac{1}{N} \sum_{i=1}^{N} \log_2 P_i , \tag{19}$$

$$\sigma^2_{\log_2 P} \equiv \frac{1}{N}\sum_{i=1}^{N} (\log_2 P_i - \langle \log_2 P \rangle)^2. \tag{20}$$

The findings for these quantities for eighteen years appears in Table 1.

The variation in $\langle \log_2 P \rangle$ over the years reflects changes in cost of living through the twentieth century. Catalogue prices changed in two manners: (i) by the change in price of an invariant item such as a clothespin, a 1910 specimen being indistinguishable from one of 1940; and (ii) by the change in the nature of the item listed to reflect an evolving technology and a varying public taste. The 1910 bicycle was quite different from a 1970 model. The 1910 buggy whip had disappeared from the catalogue and the CB transmitter was known only to science fiction writers in 1925. Many interesting deductions follow from changes in $\langle \log_2 P \rangle$,[16] but it is upon the third and sixth columns, $\sigma^2_{\log_2 P}$, of Table 1 that I wish to direct attention.

Year[a]	$<\log_2 P>$	σ	Year[a]	$<\log_2 P>$	σ
1900	0.150	2.43	1939-40	0.627	2.62
1902	0.212	2.34	1946-47	0.532	2.15
1908	-0.0228	2.29	1948-49	1.336	2.37
1916	-0.068	2.38	1951-52	1.785	2.34
1924-25	0.422	2.32	1962	2.403	2.24
1929-30	0.998	2.26	1972-73	3.030	2.27
1932-33	0.691	1.91	1973-74	3.322	2.05
1934-35	0.673	2.22	1974-75	3.870	2.12
1935-36	0.537	2.39	1975-76	4.060	2.03

TABLE 1. Standard deviation of $\log_2 P$ from mean $\langle \log_2 P \rangle$ for various years in the period 1900–1976. $\bar{\sigma} = 2.26$; $\langle (\sigma - \bar{\sigma}^2) \rangle^{1/2} = 0.17$.

[a]An entry identified by a single year corresponds to a spring–summer catalogue; an entry identified by a number such as "1924–25" corresponds to a winter catalogue.

As one superficially scans successive catalogues, one is impressed with the tremendous variety of articles available and the steady change from year to year. We have been as much impressed by the existence of an almost invariant statistical quality—an "economic constant of the motion"—for the marketing operation. It is remarkable that, for more than seventy-five years, the dispersion $\sigma_{\log_2 P}$ (defined by equation (20)) has hardly changed. The average value of $\sigma_{\log_2 P}$ is 2.26 with $\langle (\sigma - \bar{\sigma})^2 \rangle^{1/2} = 0.17$. Table 1 shows the largest observed deviation of $\sigma_{\log_2 P}$ from 2.26 to be 1.91, in the 1932–33 winter catalogue, at the depth of the Great Depression. That catalogue contained a statement to the effect that, because of the high cost of catalogue production and somewhat reduced demand for high-priced furniture, the furniture listing is meager. A separate furniture catalogue was available upon request. The combination of the regular 1932–33 catalogue with the furniture catalogue would lead to a $\sigma_{\log_2 P}$ value closer to 2.26.

Having observed the consistency of σ, let us construct a simple inflation model to "explain" it. Suppose that, in a given year, all prices are changed by the same factor, α. Then the transition experienced by the price of the nth catalogue item in that year would be $P_i \to \alpha P_i$ so that the transition of $\log_2 P_i$ would be $\log_2 P_i \to \log_2 P_i + \log_2 \alpha$ and the difference $[\log_2 P_i - \langle \log_2 P \rangle]$ would remain invariant because the α-dependent contributions of each term cancel. On this basis, $\sigma_{\log P}$ defined by (20) remains invariant under the constant-inflation-factor postulate.

The inflation model may be made more realistic by assuming that the ith item has its own inflation factor α_i expressed as an average inflation factor plus a small correction $\Delta\alpha_i$; $\alpha_i = \alpha + \Delta\alpha_i$ with $\langle \Delta\alpha_i \rangle = 0$. Then, to first order in $(\Delta\alpha_i/\alpha)$, $\log P_i$ in one year is transformed in the next to

$$\log \alpha_i P_i \simeq \log P_i + \log \alpha + (\Delta\alpha_i/\alpha).$$

Hence, to the first order,

$$\frac{1}{N} \cdot \sum_{i=1}^{N} \log P_i \to \frac{1}{N} \sum_i \log P_i + \log \alpha$$

and

$$(21) \quad \sigma^2_{\log P} \to \sigma^2_{\log P} + \frac{2}{N} \sum_{i=1}^{N} \left(\log_2 P_i - \log_2 \overline{P}\right)(\Delta\alpha_i/\alpha) + \frac{1}{N} \sum_{i=1}^{N} (\Delta\alpha_i/\alpha)^2.$$

In a year with a mean inflation rate of 10%, $\alpha = 1.1$. A reasonable range for $\Delta\alpha_i$ might be $-0.1 < \Delta\alpha_i < 0.1$, yielding the range $-0.09 < \Delta\alpha_i/\alpha < 0.09$, so that typically $(\Delta\alpha_i/\alpha)^2 \simeq 0.01$. When the inflation rate is independent of the price of the item, the cross term of first order in $\Delta\alpha_i/\alpha$ in (21) vanishes. However, when the inflation rate for low-priced items is generally higher than that for higher-priced ones (a common situation), the middle term in (21) becomes negative and cancels the positive last term. Without that influence, σ^2 grows each year.

The constancy of σ^2 combined with the discussion following (5) implies that the normal distribution of $\log P_i$ maximizes the entropy function associated with that variable. Hence, in their marketing wisdom, Sears, Rosenwald, their staff, and their successors, created catalogues with goods priced so that year after year the price distribution maximized the entropy function associated with $\log P_i$.

The entropy function itself, defined by (4a), for a log-normal distribution function has the form

$$H = -\int_0^\infty \left[\log\left(P/\overline{P}\right)\right]^2 p\left(\log P/\overline{P}\right) d\left[\log\left(P/\overline{P}\right)\right],$$

where $p(x)$ is the normal distribution function defined by (6). $\operatorname{Log} P/\overline{P}$ is similar to the utility function of classical economics, originally used by Bernoulli in his analysis of the St. Petersburg gambling paradox. The H is the weighted average of the square of the function resembling the utility function.

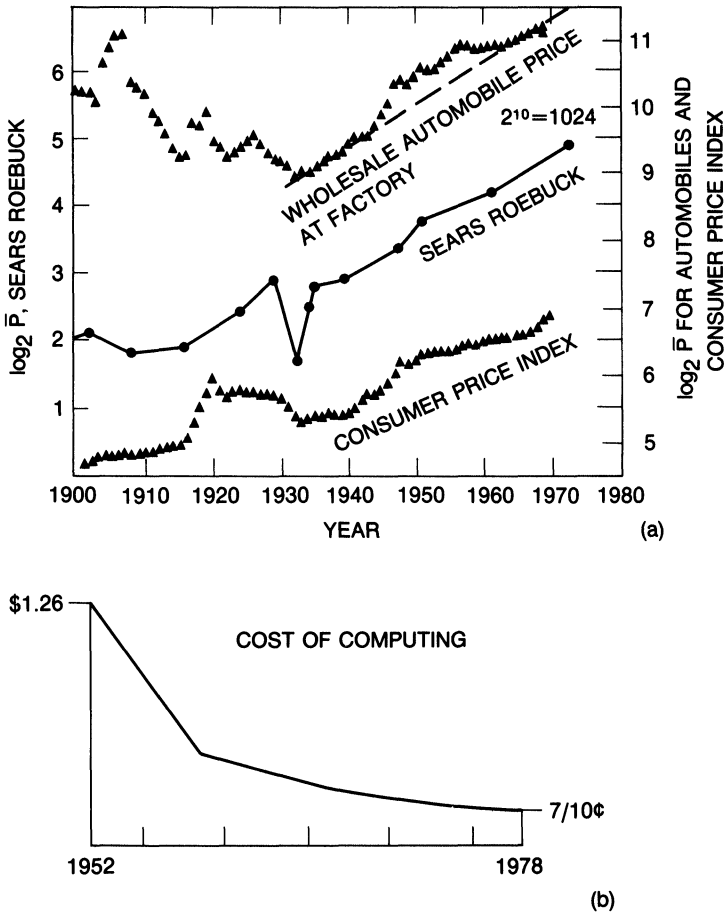

FIGURE 3. (a) Time variation of the logarithm to the base 2 of the average price in dollars of items found in general SR catalogues. Also shown are time variation of the log of average wholesale automobile prices in dollars and the log of U. S. Department of Labor consumer price index for the period from 1900 to 1975.[18] (b) The recent decline in the price of computation.[20]

Many other quantitative conclusions may be drawn from SR catalogues. One of interest to academics was especially noted by Cohn.[14] In 1905, guitars enjoyed as great a popularity as they again did in the 1960s. Among the many styles available that year was the college name group. Cohn wrote:

> ...one wonders whether the head of Sears' music department, when he priced and named his guitars, was not at the same time passing judgment upon the merits of the universities according to some secret or unconscious criteria of his own.

Note the valuations:

The Stanford	$4.25	The Princeton	$13.75
The Cambridge	$8.95	The Yale	$16.95
The Cornell	$11.35	The Harvard	$21.45

The items in a Sears Roebuck catalogue are products of an extensive and diverse technological network. Hence, the annual variation in the mean price represents an average over numerous technologies. On this basis one might determine whether a given technology is evolving well or poorly relative to other technologies by comparing price variation of its typical products with the SR mean price variation.

As an example we have plotted in Figure 3a[18] the average factory sale price to automobile dealers in the United States in the period 1900–1975 as obtained from *Automobile Facts and Figures*, 1967 and 1975 editions.[19] Notice the rapid drop in prices from 1908 until 1917, the period of dramatic evolution of the motor car from a basket on wheels to a modern vehicle. The trend in auto prices relative to the SR index continued to fall until about 1935. Since then wholesale automobile prices have paralleled the SR index. By the 1930s innovations in production line operation and in marketing pioneered by the auto industry were adopted by most other producers. Furthermore, a modern automobile, being a conglomeration of steel, glass, upholstery, electric gadgets, etc., is itself a small "Sears Roebuck Catalogue" of items.

The computer industry with its ingenious integrated circuit technology has produced superior devices at declining prices even in times of abnormally high inflation. We have, in the lower part of Figure 3b, included the variation in cost of making a specified calculation in different machines over the past twenty years as advertised by IBM in a recent issue of *Scientific American*.[20] If this data were put on the upper part of Figure 3 it would indicate that the cost reduction rate of computation parallels that of the automobile during its most dramatic period, circa 1910. When computation prices finally parallel the SR index, the most innovative period of computation development will be over. In Figure 4 the price of eggs is shown to have been almost constant for many years, while university tuition has increased more rapidly than the SR index. Poultry culture becomes ever more automated. Professors have not become more efficient operators. Furthermore, through federal pressure and their own motivation, university administrations expand more rapidly than teaching staffs. Hence, as with practically all personal service activities, costs escalate more rapidly than those of most factory-made products.

C. ON INCOME DISTRIBUTION. We have seen that money is spent on items whose prices have a log-normal distribution and that entropy has been maximized in a peculiar way in the distribution process. It would be interesting to explore the possibility that a symmetry exists between the manner in which money is made and the manner in which it is spent. We show in this section that over the first ninety-nine percentile of the U. S. population the distribution of annual incomes is log-normal and that there is an entropy principle analogous to that observed for Sears Roebuck prices. We start our discussion with some general remarks on the income distribution.

It is commonly observed that, over a large range of an independent variable, distributions might be of a standard type such as normal or log-normal but then suffer a transition in the last few percentile of a population into an inverse power law. This transition is analyzed here through a special example, the U. S. annual income distribution. That distribution is plotted in Figure 5

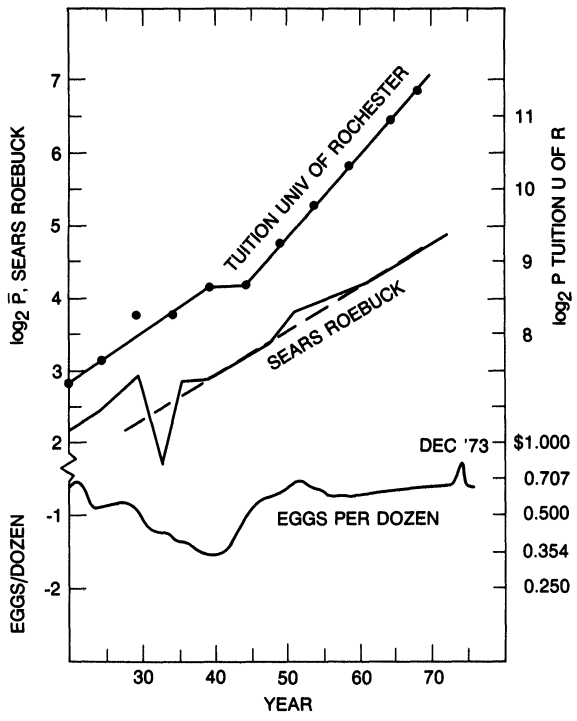

FIGURE 4. Comparison of time variation of egg prices, university tuition, and SR price index.

for the period 1935–36 on log-normal graph paper.[21] On such graph paper a cumulative log-normal distribution would be a straight line. That is the case for the first 98–99 percentile; however, afterwards a transition to a Pareto inverse power-law distribution occurs. One of the earliest observers of the log-normal distribution of incomes was R. Gibret.[22] More recent critical examination of the fitting of the log-normal distribution to data is given in references 23 and 24. Badger[21] has given a useful summary of the application of various statistical distribution functions to income data.

We now indicate how the log-normal distribution might be interpreted in terms of a maximum entropy strategy. Then we suggest a model to "describe" the transition to the Pareto form.

Through various transactions, money is transferred from individual to individual in a manner analogous to the transfer of energy from gas molecule to gas molecule through collisions. By transfer of goods or services (or welfare or charity), every family has someone with an annual income. One might argue that the many transactions cause money to become randomly distributed but, through various constraints due to training, motivation, risk-taking, inheritance, luck, intimidation, skill, etc., some people obtain larger annual incomes than others. We will still apply the entropy principle, but at first without any clear understanding of the constraint that implies the observed distribution.

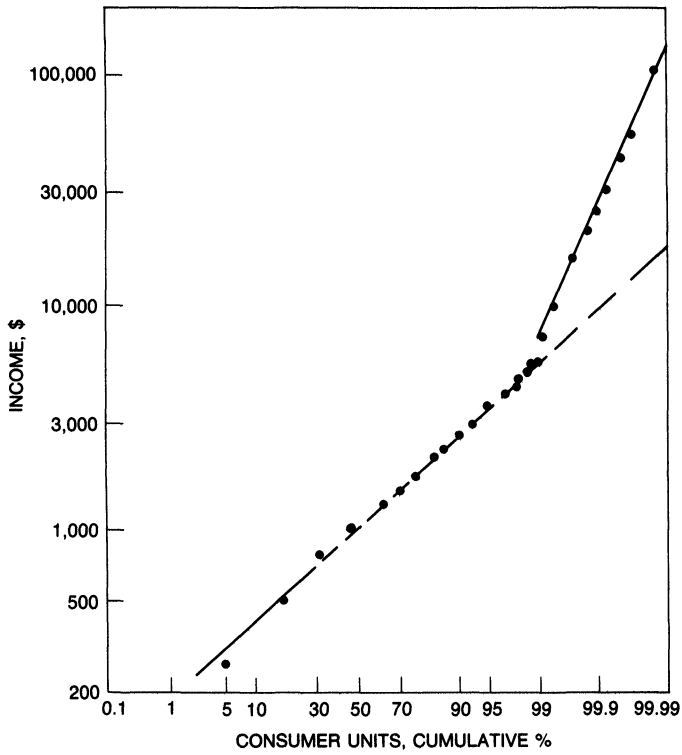

FIGURE 5. Distribution of families and single individuals by income level, 1935/1936. Data are from reference 25. Most of the data follow a log-normal distribution, while the last 1% is governed by a Pareto tail.

Let us suppose that the distribution of annual incomes is log-normal, as indicated in Figure 5. Then the probability that one's annual income lies between x and $x + dx$ is

$$(22) \qquad (2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{(\log[x/\bar{x}])^2}{2\sigma^2}\right\}\frac{dx}{x} = p(x)\,dx.$$

The factor dx/x is exactly the variation of the Bernoulli utility function $U(x)$ defined so that[26]

$$(23) \qquad dU = dx/x.$$

The classical significance of this form is that a process involving a transfer of money dx has a different meaning to persons of different levels of income. Transactions made by persons of different income levels might be more equivalent if they involved the same fraction of the income of the participants. Hence, according to D. Bernoulli, the basic function which determines one's course of action is the utility function

$$(24) \qquad U(x) = \log(x/\bar{x}).$$

Quintiles	1944	1947	1950	1951	1954	1956	1959	1962
Lowest	4.9	5.0	4.8	5.0	4.8	4.8	4.6	4.6
Second	10.9	11.0	10.9	11.3	11.1	11.3	10.9	10.9
Third	16.2	16.0	16.1	16.5	16.4	16.3	16.3	16.3
Fourth	22.2	22.0	22.1	22.3	22.5	22.3	22.6	22.7
Highest	45.8	46.0	46.1	44.9	45.2	45.3	45.6	45.5
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Top 5%	20.7	20.9	21.4	20.7	20.3	20.3	20.3	19.6

TABLE 2. Percent distribution of family personal income by quintiles and top five percent of consumer units for selected years (data from ref. 21)

Notice that with U considered to be the basic function of our process the normal distribution of U would follow from the maximization of an entropy function[21,27]

$$H = -\int p(U)\log p(U)\, dU \tag{25}$$

under the auxiliary condition $p(U)$ being normalized and

$$\langle U^2\rangle = \int U^2 p(U)\, dU = \text{constant}. \tag{26}$$

That the integral of U^2 is essentially constant over a long time interval is apparent from the data in Table 2. We may write

$$\int U^2 p(U)\, dU = \int (\log x/\bar{x})^2 p(\log x/\bar{x})(\bar{x}/x)\, d(x/\bar{x}). \tag{27}$$

From Table 2, the fraction of the national family income in a given population quintile remained almost constant over the period of eighteen years of the selected data. The mean income shifted, generally going to a higher level, but relative to the mean the distribution in a given interval remained invariant. Hence in the transition from one year to another incomes would have suffered an annual inflation factor (or deflation factor) α so that

$$x \to \alpha x, \quad \bar{x} \to \alpha\bar{x} \quad \text{and } x/\bar{x} \to \alpha x/\alpha\bar{x} = x/\bar{x}$$

but yet (27) would have remained invariant. This is a consequence of there being no basic scale in the process.

The above analysis is interesting; however, it gives us no insight into the appearance of the Pareto inverse power law tail beyond the 99 percentile in Figure 5. No one would dispute the fact that the wealthy differ from the lower 99% in the manner that they accumulate income. While most people are paid by the hour, or the number of widgets they produce, the wealthy frequently accumulate their extra wealth by some amplification process; that process

varying from case to case. At the height of the Beatles' popularity any new recording by them was purchased by millions of fans. The leverage people in the investment business have their style of amplification. During certain periods of prosperity easy money becomes available for investment, sometimes in stock, sometimes in real estate, or perhaps in silver or Rembrandts. A common characteristic of such times is that the daring may exploit the easy money to acquire some speculative commodity through a small margin payment, say ten percent, with a promise to pay the remainder later. If the commodity doubles in price a ten-percent margin payment is amplified into a ninefold profit. J. P. Morgan was given his first million by his father. He invested a considerable fraction of that in the manner described above, reinvesting the profit, and so on, to become much richer than he would have had he accepted the offer of a privatdozentship in mathematics at Göttingen University offered to him by Felix Klein. Perhaps one of the most common lower-level modes of amplification is for an individual to organize an operation with others working for him so that his income is amplified through the efforts of others (a modest-sized business, for example).

We now introduce a model to indicate how Pareto-Lévy tails may be derived from a log-normal distribution (or indeed from any one of a broad class of distributions with second moments) by accounting for the process of amplification, by the amplification of amplifications, etc.[28] Let $g(x/\bar{x})$ denote the basic distribution written in terms of the dimensionless quantity $x/\bar{x}$, $\bar{x}$ being the mean value of the observed x if the tail of the distribution is neglected. With a small probability, λ, suppose that in the new amplifier class one has the same distribution function g that is natural for the process but that $\bar{x}$ is amplified to $N\bar{x}$. In the second stage of amplification, which we postulate to occur with a probability λ^2, the mean value of $\bar{x}$ becomes $N^2\bar{x}$. The new distribution $G(y)$ (with $y \equiv x/\bar{x}$) that allows for the possibility of continuing levels of amplification is

$$G(y) = (1-\lambda)\left[g(y) + \frac{\lambda}{N}g(y/N) + \frac{\lambda^2}{N^2}g(y/N^2) + \cdots\right] \tag{28}$$

where λ is a parameter that determines the range of the initial distribution $g(y)$. The factor $(1-\lambda)$ is introduced to ensure the proper normalization of $G(y)$. It is easy to see that by replacing y by y/N in (28) that

$$G(y) = \frac{\lambda}{N}G(y/N) + (1-\lambda)g(y). \tag{29}$$

The determination of the complete solution of our inhomogeneous scaling formula (29) is rather complex but it is easy to obtain our desired asymptotic properties of $G(y)$. First suppose $\lambda \to 0$. Then there is no amplifier class in the population and $G(y)$ becomes the same as $g(y)$. If λ is small, say 0.01, and N is about 10 then $G(y)$ is still close to $g(y)$ since the first term in (29) may be neglected. However, when y becomes large $g(y) \to 0$. Let us suppose this decay is faster than that of $G(y)$. Then the asymptotic form of $G(y)$ is determined by the simpler scaling formula

$$G(y) = (\lambda/N)G(y/N).$$

If we suppose that $G(y) = Ay^{-1-\mu}$, then direct substitution yields

$$\mu = \log(1/\lambda)/\log N. \tag{30}$$

Thus the Pareto exponent appears as a fractional dimension. The evaluation of A requires a more subtle analysis since in general it may be periodic in $\log \lambda$ with period $\log N$.

The best value of μ to fit the tail of the 1935–36 data was found by Badger[21] to be 1.63. If we put the probability of being in the special amplifier class as $\lambda = 0.01$ the average amplification factor N would be about 16.8. This number is not surprising since one of the most common modes of significant income amplification is to organize a modest-sized business with the order of 15–20 employees.

3. On the dynamics of technological evolution.

A. EVOLUTION AS A SEQUENCE OF REPLACEMENTS. We have found that in the description of certain sociotechnical systems the entropy function plays an important role. In all of the systems discussed a considerable degree of sophistication has evolved over many years of experience. In this section we consider the manner in which technological systems evolve.

1860	1.30	1900	2.48	1940	6.78
1865	1.00	1905	2.47	1945	5.53
1870	1.49	1910	2.04	1950	5.96
1875	1.60	1915	2.34	1955	8.13
1880	1.78	1920	2.55	1960	10.00
1885	2.17	1925	3.41	1965	11.53
1890	2.28	1930	4.35	1970	12.62
1895	2.63	1935	4.83	1975	10.50

TABLE 3. Index of ratio of industrial daily wage to wholesale farm price index 1860–1975 (Ratio: 1865 = 1.00; data from ref. 29)

Lifestyles in modern society are ever changing. Indeed, such has been the case in most societies—the changes at some times being more rapid than others. An evolving technology has, for several hundred years, been an effective driving force for change influencing numerous components of our lifestyles, such as our diet, our ease of travel, the duration of the workday, the nature of available goods, and our life expectancy. In this section we characterize an evolving technology (and society) as a sequence of replacements. A remarkable feature of such a technology is its autocatalytic nature, each innovation catalyzing the generation of the next.

Affluence might be measured in terms of the amount of goods purchasable by a typical worker for a day's wage. As a first example of the effect of technological development on affluence in the United States in the past hundred years, we list in Table 3 the ratio of the index of the industrial daily wage to the wholesale farm price index for the period 1860–1975 (with the ratio for 1865 being equal to one). The data was obtained from reference 29

	1800	1840	1880	1900	1920	1940	1960
Wheat							
Man-hours per acre	56	35	20	15	12	7.5	3.1
Yield per acre, bu.	15	15	13.2	13.9	13.8	15.9	24.9
Man-hours per 100 bu.	373	233	152	108	87	47	12
Corn							
Man-hours per acre	86	69	46	38	32	25	7.9
Yield per acre, bu.	25	25	25.6	25.9	28.4	30.3	56.9
Man-hours per 100 bu.	344	276	180	147	113	83	14
Cotton							
Man-hours per acre	185	135	119	112	90	98	54
Yield of lint per acre in lbs.	147	147	179	191	160	245	454
Man-hours per bale	601	439	318	280	269	191	57

TABLE 4. Productivity of workers in wheat, corn, and cotton culture in the United States (data from ref. 29)

[Tables E-1 through E-41, D-574, D-590, and D-626, and in annual "Statistical Abstracts": reference 30 (since 1955)]. The farm price data corresponds to the same mix of produce for each year of data recorded. The increase in the index by an order of magnitude is the result of improved agricultural efficiency and of the increased productivity of a typical nonagricultural worker so that he can command a higher wage.

In Table 4 we have recorded improved agricultural yields in wheat, corn, and cotton since 1800. This spectacular development of agricultural technology is the result of proper choice and use of fertilizer, research in plant genetics, mechanization, improved soil management, irrigation, and other factors. Modern agricultural evolution has been a continuing replacement of one technique or plant strain by another.

A second example of the influence of technology on affluence is the increase over the years of the distance a typical worker can travel on a day's wage. Roman data (circa 300 A.D.) is available from Diocletian's wage and price control ordinance.[27,31] An unskilled workman could travel about eleven miles on a day's wage, while a carpenter or stone mason could do double that distance. There was no improvement for the next 1,500 years. In England in 1790 the stagecoach fare from Manchester to London (195 miles) was £2/5 or four miles to the shilling.[32] A laborer then made 14d per day, giving him slightly more than a four mile deluxe ride. By renting a horse he could do about as well as the Romans did. A coal miner at 7d per day did not do so well, but a foreman at 21d per day[32] could travel commercially about seven miles on a day's wages. While the coaches were the best vehicles available, they did shake the passenger a bit. In New England the tavern density was one per linear mile to ease the traveler of his pain.

The 1795 stagecoach fare from New York to Georgetown (Washington, D. C., did not yet exist) was $16 for three stages: New York to Philadelphia ($6), Philadelphia to Baltimore ($6), Baltimore to Georgetown ($4). This is to

be compared with the 1983 bus fare of $22.50 (on the basis of $45 for the round trip) for a smoother four and a half hour ride, warmer in the winter and cooler in the summer. Typical colonial stagecoach fares in nonmountainous, well-settled regions ranged from five to seven cents per mile, rising to ten cents per mile in wilder parts of the country. The government travel allowance was fourteen cents per mile in 1815; today it is twenty. It is clear that if a poor man needed to take a long trip he probably walked. Those slightly better off might own a horse, but only the rich could afford commercial vehicles. Ocean travel was also expensive. An Englishman without funds who wished to try his luck in the colonies had to indenture himself for seven years (about one-third of his remaining life expectancy at age twenty) to pay for his boat ride.

The first technological breakthrough to broaden the modern worker's travel horizon on a day's wage beyond that of the Romans was the construction of canals. When the Duke of Bridgewater's canal[32] was completed in 1765, the price of transport of £5 per ton from Liverpool to Birmingham was reduced to £1/10.

The canals were a basic component in the birth of the industrial revolution. The early steam engine provided the motor power to make possible the cheap mass production of simple objects. However, mass production is meaningless without mass markets. Before canals, transport costs were frequently greater than production costs, so that a change in production cost had only a second-order effect on the price at a distant market. The canals at low cost carried coal and raw materials for fabrication to the industrial centers and then cheaply delivered finished products to distant markets. By 1810, hardly any significant English town was more than ten miles from a canal.

The most drastic change in price, trip time, and comfort came with the railroads which replaced the canals. The railroad fare in the period 1890–1915 was about two cents per mile. Hence, on the $5 a day that Henry Ford offered his workers, they could travel 250 miles. An autoworker in 1978 made about $20,000 a year and a coal miner about $25–30,000 a year. This corresponded to a range of about $100 a day. In 1978, at regular airfare of about seven cents a mile, a trip of almost 1,500 miles was possible by air. (The lower bus fare permitted 3,000 miles.) With careful planning, taking advantage of the old Texas International's $99 flights from Los Angeles to New York, or Freddie Laker's special transatlantic rates, a skilled industrial worker could then fly 3,000 miles on a day's wage. His poorer brother who got along on the minimum wage ($3 an hour) could still travel about 750 miles by bus, a factor of 75 times better then the old Roman and three times better than Henry Ford's employees. Since we did not base our calculation on wages after taxes, the above numbers are somewhat exaggerated. On the other hand, if our typical worker wished to drive a car with mama and the two kids, more passenger miles might be possible on a day's wage.

We have asserted that technological and social evolution is a consequence of a sequence of replacements of one technique (or idea, tradition, or artifact) by another. This statement is in the Darwinian spirit of survival of the fittest, with each new mutant or species struggling to find its niche, sometimes at the

expense of displacing or replacing the older forms. Once a virile mutant or new form established itself, it would be expected to propagate, continuing to replace its competitors until it reached an equilibrium saturation level.

A simple mathematical model of growth to saturation is the logistic model, introduced by Verhulst[33] in an investigation of the population expansion of nations. Let $n = n(t)$ be the population of a given species at time t. Then, with saturation level θ, the model is characterized by the growth equation:

$$dn/dt = kn[1 - (n/\theta)], \tag{31}$$

k being a rate constant (to be determined empirically). If we let $x = n/\theta$ be the fraction of the way to saturation, and $y = 1/x$, then

$$-d(y - 1)/dt = k(y - 1) \tag{32}$$

so that

$$y - 1 = [y(0) - 1] \exp(-kt);$$

or, with $x_0 = x(0)$,

$$\log [x/(1 - x)] = \log [x_0/(1 - x_0)] + kt. \tag{33}$$

This formula suggests that to test the logistic model for a particular case one should plot the relevant data in the form $x/(1 - x)$ on semi-log graph paper as a function of the time and observe whether the points lie on the required straight line.

J. Fisher and R. Pry[34] have successfully exploited the logistic model to describe the market penetration of many new products and technologies. If a superior new product or process excites the trade sufficiently to absorb 10–15% of the market, it is highly likely that it will win an increasingly larger fraction until it completely dominates the market or until its own new competitor appears. In the logistic replacement model, $x/(1 - x)$ represents the ratio of the fraction of the market captured by the new to that remaining for the old. We have reproduced in Figure 6 the remarkable logistic fit published by Fisher and Pry for several industrial replacements. These authors have produced many other equally impressive graphs for other technologies.

C. Marchetti and N. Nakicenovic[35] have given an excellent summary of world energy usage and source substitution by employing the logistic model as shown in Figure 7. There one sees the replacement of wood by coal and coal by oil. Natural gas seems to be on the road to becoming the primary oil substitute.

R. Herman and the author[36] have shown that as basic an evolutionary process as the industrial revolution may also be modeled by logistic dynamics. As the industrial revolution evolved, the fraction of the labor force in agriculture declined while the fraction in industry grew.

Before 1840 the ratio of nonagricultural workers to agricultural workers in the labor force of the United States remained fairly constant over many decades. The ratio of the fraction of nonagricultural workers to agricultural

workers in the United States is plotted as a function of time in Figure 8 on semi-log graph paper (as is the corresponding fraction for Sweden). The U. S. data was obtained from the U. S. Bureau of the Census, Statistical Abstracts of the United States. It is remarkable how well the data fits the straight line defined by the logistic equation for a period of about one hundred years.

The rate at which agricultural workers left the farms in the early 1940s exceeded that expected on the basis of the logistic equation. The acceleration

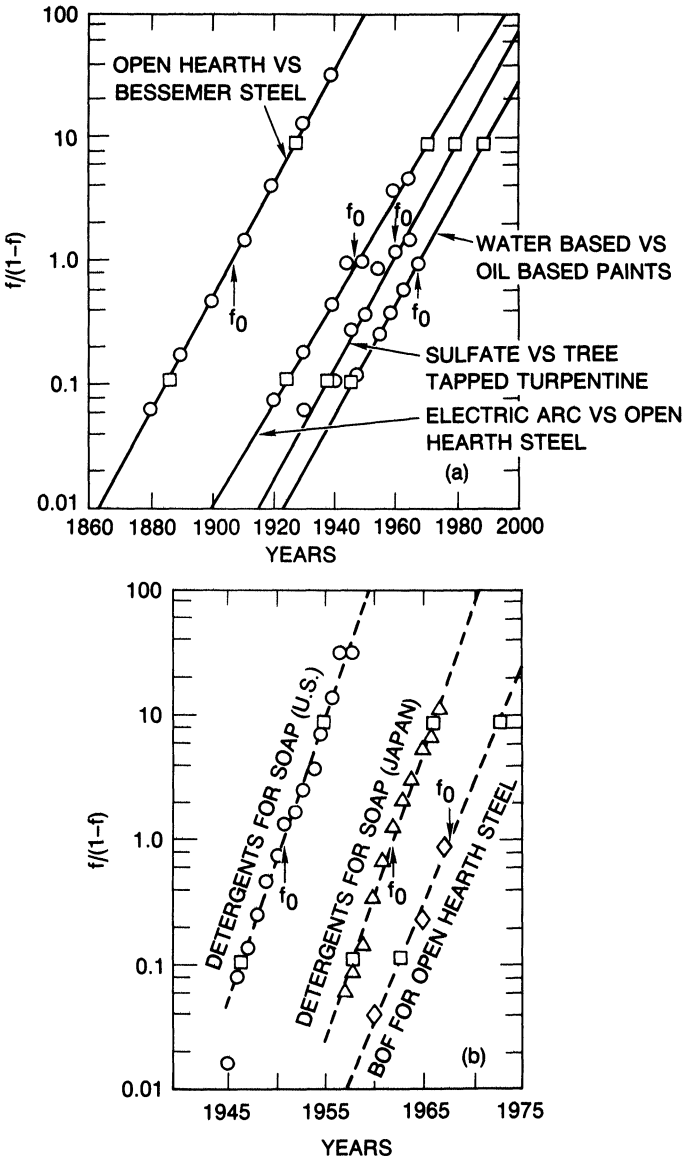

FIGURE 6. Substitution data and fit to model for a number of products and processes: all data are for the U. S. except where indicated.[34]

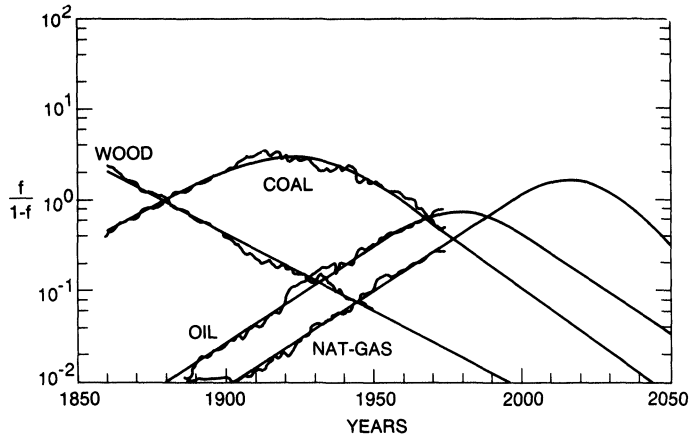

FIGURE 7. The logistic curves appear as straight lines, from which we see that it takes about 100 years to go from 1 percent to 50 percent of the market. Also we see that all perturbations are reabsorbed elastically without influencing the trend.

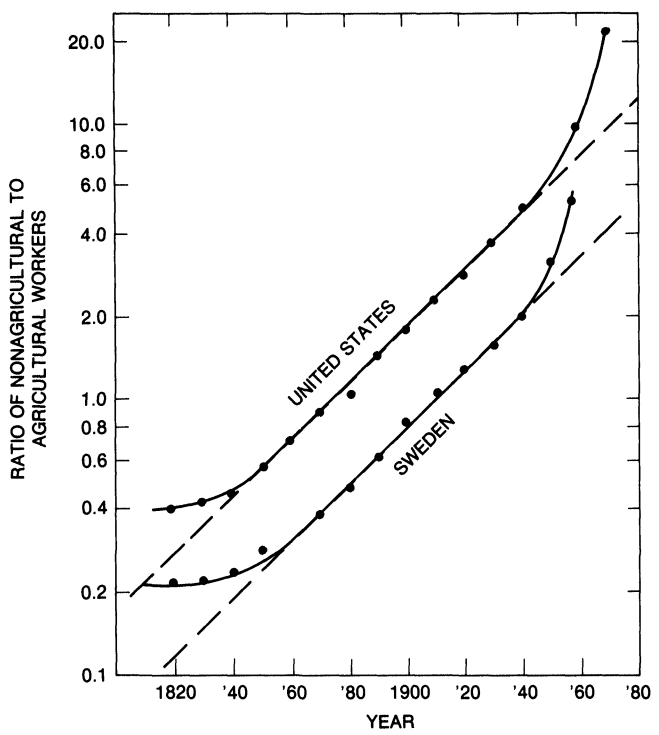

FIGURE 8. Variation of the ratio of nonagricultural workers to agricultural workers in the U. S. and Swedish labor force.

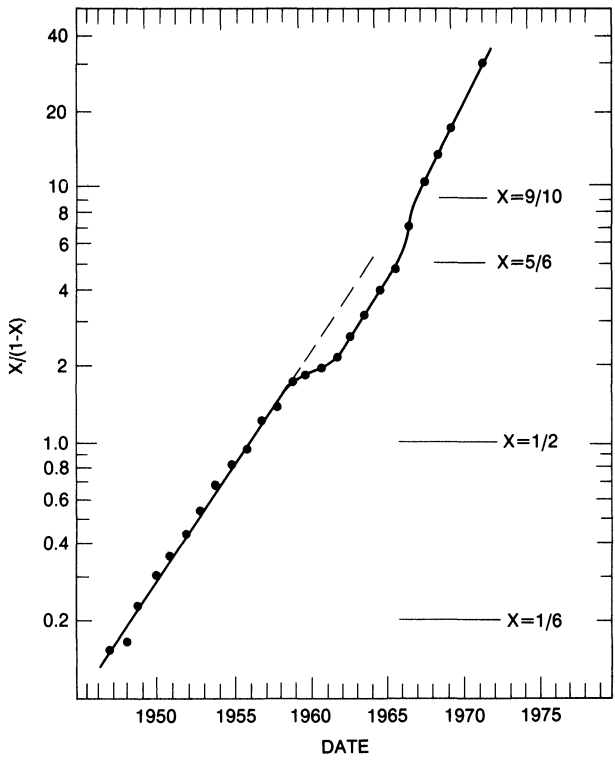

FIGURE 9. Manner in which intercity passenger travel by rail has been replaced by air travel.[37]

was a response to a force generated by telegrams from President Franklin Roosevelt that started with the word "Greetings." After World War II many of the young men who responded to the greetings did not return to the farm. The curve for $x/(1-x)$ (nonagricultural to agricultural worker ratio) for Sweden follows that of the United States. While the young Swedish farmer did not receive a greetings telegram, he was enticed to relocate by the higher wages paid by companies that were selling factory products to pleading customers around the world.

Even when the logistic form of the replacement model is violated one may still expand his intuition on replacement dynamics through the understanding of the cause of the violation. In Figure 9 we sketch the replacement pattern of rail by air in intercity passenger travel in the United States. In the figure[37]

$$x = \frac{\text{annual air passenger miles}}{\text{annual air and rail passenger miles}}. \tag{34}$$

Notice that in the period 1947–1959 the data are in accord with equation (33). The deterrence in replacement evolution during 1960–1961 seems to have been the result of several unusually long strikes by airline workers and of the public's response to a series of serious and unusual airplane accidents. By 1962, the system recovered and the replacement curve continued with its old slope until the late 1960s. Then, accelerated replacement to the end of the decade

occurred at the time the largest rail passenger carrier, the Penn-Central line, was suffering through its prebankruptcy and bankruptcy. The management of passenger rail service was reorganized in the 1970s by Amtrak.

The role of intercity buses was omitted from the above discussion, because the fraction of total passenger traffic maintained by them changed very little during the period investigated.

Year	Gross Tonnage, Tons $\times 10^{-3}$	
	Steam	Sail
1840	202	1978
1845	326	2091
1850	526	3010
1855	770	4442
1860	868	4480

TABLE 5. Gross tonnage according to type in United States merchant fleet (data from ref. 30)

The replacement of sail by steam in the U. S. Merchant Marine in the nineteenth century followed a similar pattern.[37] Practical steamship operations started in the 1820s, and during the two decades 1830–1850 the logarithm of the fraction of tonnage in steamers to that in sailing ships followed a logistic straight line, as it did again during the interval 1880–1915.

At first, steamboats appeared in the river traffic and then in coastal waterways; later they operated on transatlantic runs, but only after the Civil War could they successfully compete with the clipper ships on the longer Pacific passages and on the voyages around the Horn connecting the East Coast with San Francisco.

The first fast clipper ships appeared in the 1830s, when steamboats were becoming numerous. They were built in large numbers during the decade 1845–1855 (peaking between 1850 and 1853). Two important events of 1849 stimulated their production, perturbing the takeover by steam: discovery of gold in California and repeal of the British Navigation Acts and the breaking of the China trade monopoly long enjoyed by the British merchant marine. The expansion of trade in the West Coast, the Orient, and Australia by adventurous American skippers created an enormous demand for the speedy clippers, as indicated in Table 5.

A financial slump in 1854 essentially stopped clipper ship construction, and during the panic of 1857 practically all types of construction were terminated.

Steamboat construction was favored over sail for the increased local transport required by the Civil War. The considerably improved steamboat models dominated naval construction in the postwar reconstruction period 1865–1873 only to be abated by the panic of 1873 whose effects persisted for several years [see Figure 10]. With the return to normal, the steamboat replacement curve proceeded along its logistic course until 1915 when the shipping requirements of World War I stimulated an accelerated naval construction program. By then, no one considered new sailing vessels to be suitable for commercial shipping.

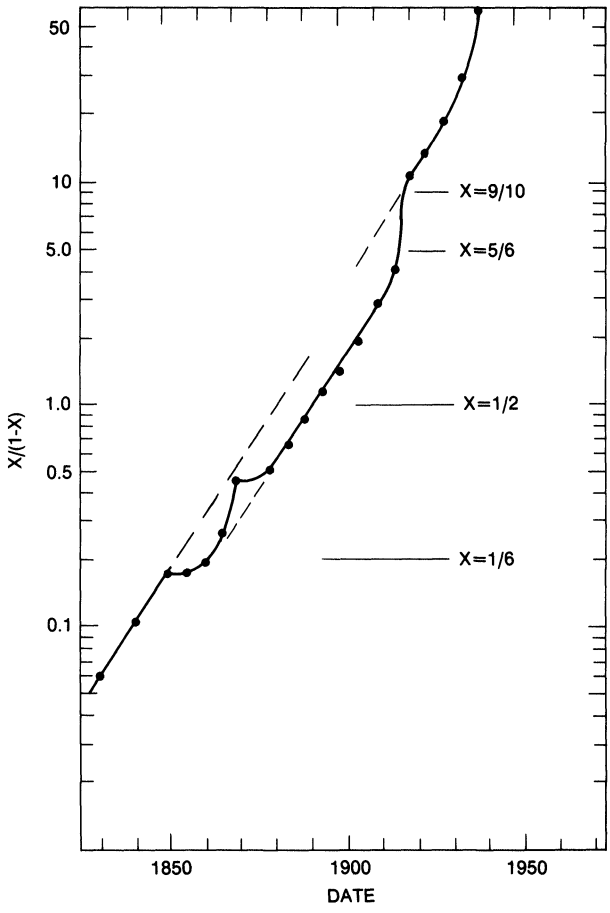

FIGURE 10. Manner of replacement of sailing ships by steamships in the U. S. merchant fleet.[37]

A remarkable feature of the last two examples is that after an intermittent deterring force has been lifted from a logistic process, the process is restored with its original rate constant.

Marchetti has found another surprising application of the logistic model, namely to the evolution of the efficiency of technical devices and processes. Some of his findings are summarized in Figure 11, where data is plotted for efficiency of the steam engine, of lamps, and finally of ammonia production. He said:

> In a sense inventors, wandering in the world of all possible machines, picked the ones that looked best, ready to throw them away for the next better ones like Alice in Wonderland with her flowers. Here only one parameter was taken as an indicator of performance, but a very important and subtle one: thermodynamic efficiency.

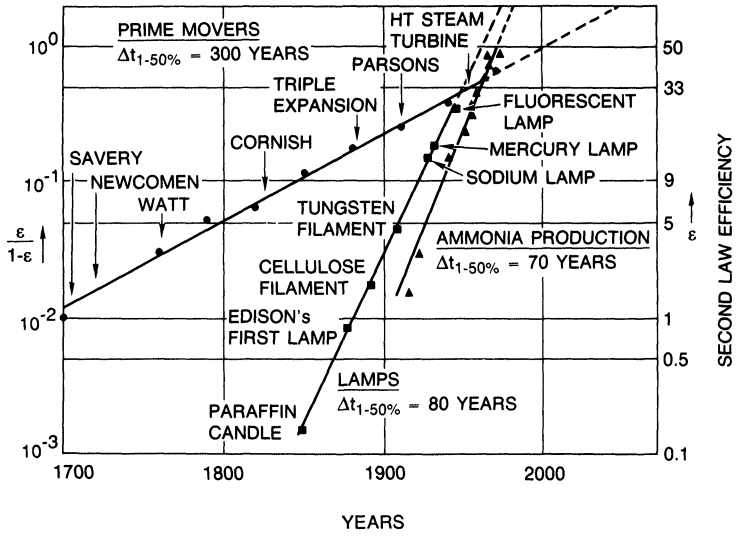

FIGURE 11. Historical trends in efficiency ($\Delta t_{1-50\%}$ is time necessary to evolve from 1% to 50% efficiency; ε is second law efficiency.)

B. CRITICAL POINTS AND "PHASE TRANSITIONS" IN TECHNOLOGICAL EVOLUTION. When a device, developed to perform an important function, enjoys a considerable and growing market penetration, its own success generates a demand pressure for its further improvement. Owners of a fast boat want a faster boat. Users of programmable pocket computers capable of performing forty-step operations cry for one to perform eighty-step operations.

As the demand for improved performance grows, some natural limitation in the basic device technology becomes apparent, causing improvements to become increasingly expensive, and thus motivating the search for alternative technologies to meet the demand. We shall see that production and operating accidents also grow, as well as normal operating and development costs, when these natural limits are approached. The main purpose of this section is to exhibit several relationships between physical laws and technological transitions (and therefore technological evolution). The first example to be considered is a ship of a given class operating on a calm sea surface. As the velocity of the ship increases, it is subjected to an increased resistance by the water, thus requiring more power. Figure 12 represents a typical operating curve of resistance per ton of displacement plotted as a function of $V/L^{1/2}$, V being the velocity and L the length of the ship. This is effectively $F^{1/2}$, with $F = V^2/Lg$ being the Froude number of the ship motion. At high velocities resistance is associated with the formation of the bow wave. All displacement ships generate a bow wave whose amplitude and wavelength increase with the velocity. The larger the amplitude the greater the fraction of the ship's power converted to raising water vertically to a bow wave rather than in the horizontal propulsion of the vessel. When the velocity (in knots) reaches

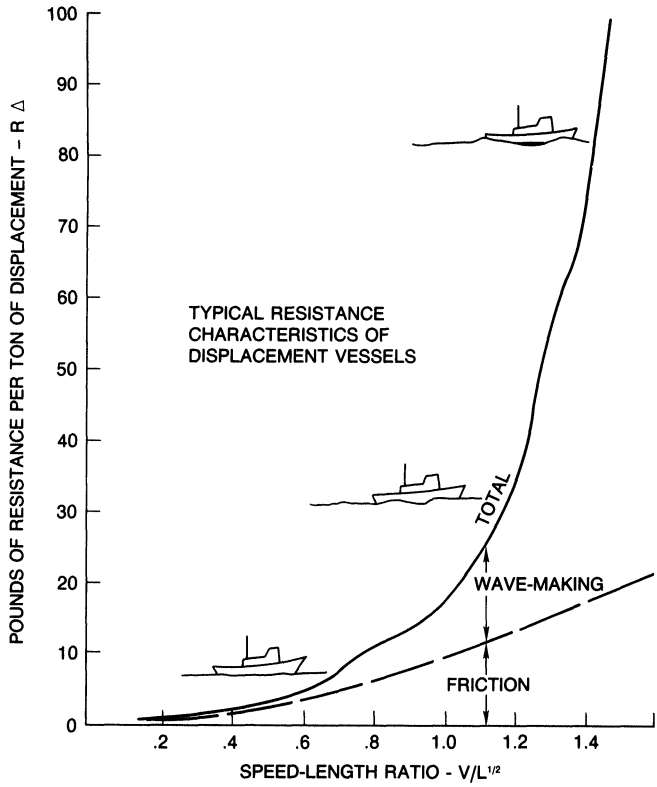

FIGURE 12. Resistance versus $V/L^{1/2}$ in a characteristic speed-power curve for a displacement ship: Note the ship-wave profiles.[38]

$0.6L^{1/2}$ (again the length in feet), there are about two and a half bow waves per hull length. Finally, at the critical speed $V_0 = 1.3L^{1/2}$ the wavelength of the bow wave becomes equal to the length of the hull.[38] Since the upper crest of the wave is aft of the bow and the lower crest fore of the stern, the ship is effectively *going uphill* upon the achievement of the critical speed. To remain in this tipped configuration a ship then requires considerable more propulsion power than one moving horizontally at low speeds. A 750-foot hull has the critical speed of 35.6 knots; for a 1,000-foot one, it is about 41 knots. Therefore it becomes very expensive to operate a large ship whose crusing speed exceeds 35 to 40 knots. One motivation for increasing the size of oil tankers is that the longer the tanker, the greater will be the critical speed.

If one wishes to overcome the bow wave limitation with a sea-going vessel, he must find a scheme to eliminate the bow wave. This indeed has been accomplished through the invention of hydrofoils and planing ships, with the ship essentially flying as it skips over the water. In this new mode of operation the resistance drops to the level indicated on the dashed curve of Figure 13. Thus, one finds a λ form of the specific power curve as he passes through the transition. The curve in Figure 13 is analogous to the heat capacity curve of a

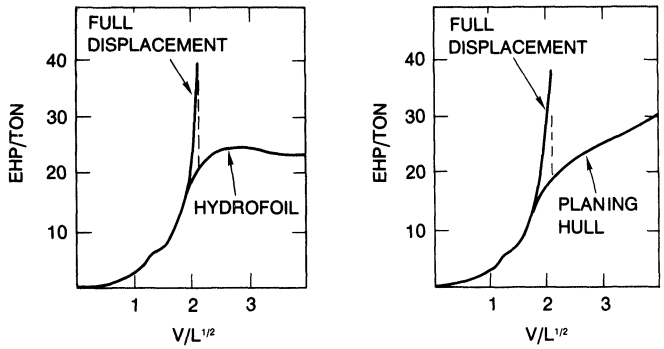

FIGURE 13. Qualitative comparison of sealed power (EHP/ton) versus ship speed ($V/L^{1/2}$) for displacement, planing, and hydrofoil hulls.

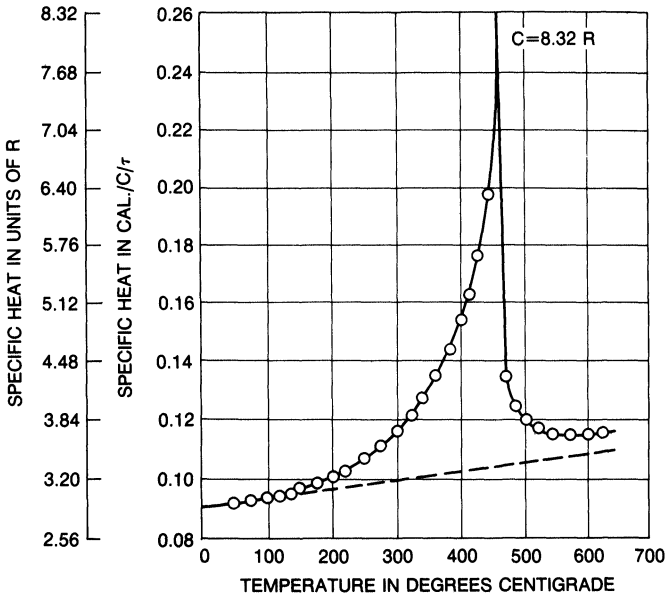

FIGURE 14. Heat capacity versus temperature of CuZn (β-brass) alloy.[39]

solid characterized by several types of internal degrees of freedom (as is the case in magnetic materials, binary alloys, liquid helium, etc., see Figure 14).[39] By definition, the heat capacity of a material is

$$C = dE/dT,$$

the amount of energy required to raise the temperature of the material one degree Kelvin. As a critical point is reached more energy is needed to raise the temperature one degree. In the case of a β-brass (50% Cu, 50% Zn) the alloy is completely ordered at low temperatures as a periodic simple cubic lattice of alternating Cu and Zn atoms. As the critical temperature is approached the

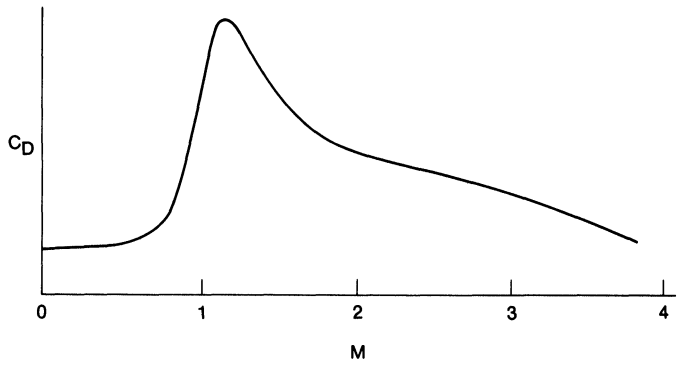

FIGURE 15. Drag coefficient C_D of a wing section at a constant angle of attack through the transonic range as a function of the Mach number M.

arrangement of Cu and Zn atoms becomes randomized with a certain amount of energy being used in randomizing the system rather than raising the temperature. This is analogous to the energy used by a ship to develop the bow wave, energy then not available for propulsion.

The operating curves in Figure 12 and 13 refer to normal operating conditions, that is, for a ship in a calm sea. When a ship operates near its natural limit in the presence of the fluctuations common to a rough sea with high waves, an extra operating cost is added in the form of ship damage. At high speed in a rough sea while the ship is tipped at the normal angle with its bow wave, interference with a natural wave of the opposite phase may leave the bow unsupported in mid-air so that it "slams." A continuing sequence of slams shakes the crew and damages the ship.

The transition from the subsonic flight regime to supersonic also follows a λ-like variation[40] in the drag coefficient, as evident in Figure 15, the peak appearing slightly above Mach 1.

Basic limitations in successive stages in the evolution of computer technology arose from demands for speed and storage capacity. The first automatic programmable digital computers were conceived of independently by Howard Aiken (Harvard) and George Stibitz (BTL) in 1937. Their electromechanical relay devices were completed in 1943. While these machines could multiply numbers ten times faster than desk calculators, improvements beyond another order of magnitude were limited by the response times of mechanical components whose motions could not significantly exceed the speed of sound.

Since the switching time of electron tubes did not suffer from this limitation, the tube technology was the basis of the second generation of programmable digital machines, pioneered by Eckert, Mauchley, and Goldstein at the University of Pennsylvania. Their product, the ENIAC, required thousands of tubes, numbers far exceeding those in other electronic devices, so that tube reliability became a basic limitation in this development of technology.

The two most advanced machines[41] put in operation in 1950 were the Whirlwind (MIT) and SEAC (NBS), the Whirlwind depending entirely on

tubes and the SEAC following the English preference for the employment of mercury sonic delay lines as storage device. One of the main champions for delay lines was M. V. Wilkes, who used them on the ESDAC (Cambridge, 1949).[42] The English view was evident in a statement by A. D. Booth[43] concerning the design of "modest-sized machines" (1953):

> ...it has long been maintained by the author that for reliable operation an electronic computer must use a minimum possible number of electronic valves. APE(R)C designed with this view has less than 450.

The English position was based upon the realization that electron tubes have a finite lifetime so that any network of a sufficiently large number of tubes would require an elaborate tube maintenance program. Such was an important feature of the Whirlwind design.

If n is the number of tubes in a network, the rate of tube failure is characterized by

$$-dn/dt = \lambda n.$$

If the tube failure rate constant λ, is small, then in a small time interval Δt the expected number of failures would be

$$-\Delta n = \lambda t \Delta t.$$

Hence, if λ is measured in inverse hours and the number of tubes in the network is $n \approx 1/\lambda$, on the average one tube would fail per hour. In a larger network with $n \gg 1/\lambda$, several tubes might fail per hour. A critical number n^* would exist so that in machines with $n > n^*$ tubes, a tube would fail in a time interval of the order of that required to search for the defective tube and replace it. The cost per calculation in a machine whose tube number approaches the critical number would become enormous.

As the problem of a vacuum tube lifetime loomed more seriously, attention to it was diverted by the development of the transistor, a smaller, cheaper, more durable, less power-hungry device, that was available to perform the same functions as the vacuum tube. By the late 1950s ferrite magnetic memory cores also appeared as ideal memory components. With these advances the technological phase transition for computers sketched in Figure 16 was experienced in the late 1950s and early 1960s.

We have plotted schematically the computer operating cost per component as a function of the number of components. The upper dashed curve represents electron tube operating costs and the lower dotted curve solid-state machine operating costs, both as a function of the number of components. The rapid rise in the dashed curve reflects the increase in down time required for tube testing and replacement. There exists a critical tube number beyond which, at any moment it is certain that a tube will "die" in a time interval of the order of that required to detect and replace it. At that critical point the cost of a computation would become enormous. Fortunately, the technology of the solid-state computer "phase" became available so that the curve followed by the industry tended to be like the solid one of Figure 16.

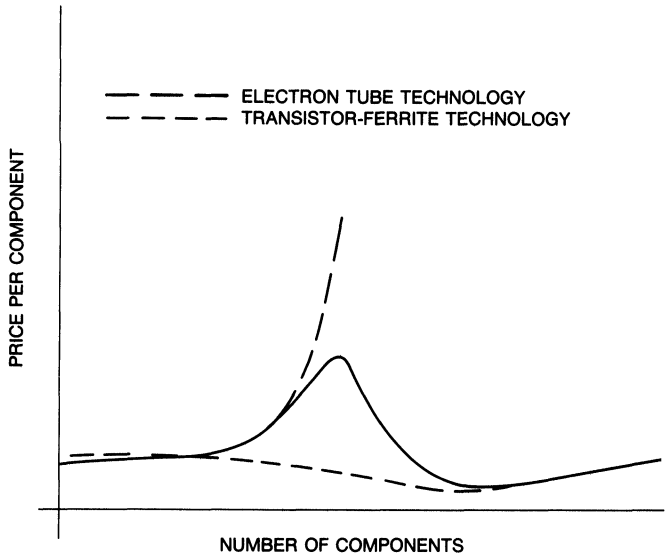

FIGURE 16. Qualitative comparison of the unit contribution to computing cost by a single electron tube or transistor element versus the number of components in a calculating machine.

With the employment of transistors and ferrite cores very large reliable memory banks could be constructed. The main limitation of the early solid-state technology became apparent as one tried to make calculations faster by decreasing the switching time required for basic binary operations. As switching times decrease to the 10^{-8}–10^{-9} second level, the time of a calculation becomes limited by the velocity of light, with the lower limit depending upon the distance traversed by a signal during the calculation. This physical limitation was overcome through the technological phase transition associated with the introduction of integrated circuits with hundreds of circuit elements per square centimeter on a substrate. As the packing density of circuit elements increases, the expected limitations on improvement might be associated[44] with (1) heat generated during the switching operations, which will increase the diffusion constant of doping impurities in the silicon chips, thus blurring the identity of individual circuit elements, and (2) diffraction-limited optical problems associated with preparation of masks for photoetching technology, and (3) effects of background radiation (from cosmic rays and similar sources) on the false switching of very small circuit elements.

The assertion that a technology advances through a series of phase transitions and substitution implies that, if one plots the output of a technology as a function of time, it might be considered a succession of logistic curves. Each curve reaches a saturation level. With each level of evolutionary innovation to a new logistic develops. Such a graph is drawn in Figure 17, which summarizes the improvement in number of feet of advance a day in hard rock drilling over 150 years.[27] The critical innovations are indicated in the figure.

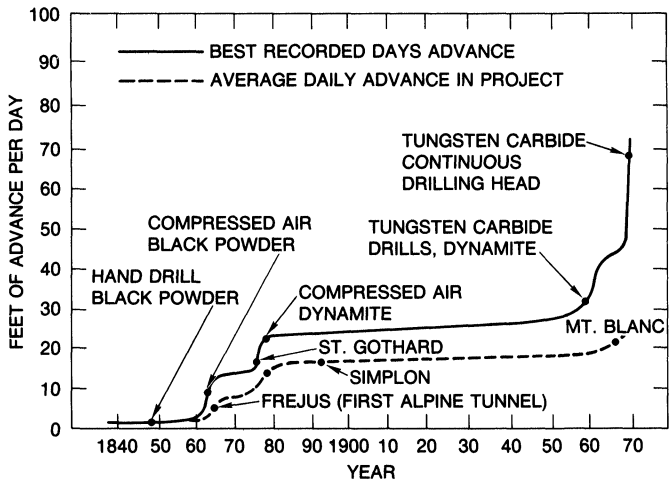

FIGURE 17. Progress in tunnel drilling in hard rock: all drilling records before 1965 were taken from Alpine tunnels so that conditions were essentially the same. While there were great days in the Mt. Blanc tunnel, drilled with the latest tools, the average rate of advance was not significantly higher than in tunnel construction at the turn of the century, because of various unexpected holdups. The 72-foot-a-day record of 1967 was made in the drilling of a water tunnel in St. Louis, Michigan with the latest continuous drilling machine.[27]

We close our discussion of phase transitions in technology with the presentation of several graphs showing the growth and decline in the number of companies concerned with some given technologies. As a new technology grows, numerous firms are established to produce and market the output of that technology. At first the number of firms and their total output seems to increase exponentially. Generally, the new firms are born independently of each other so that an entrepreneur or investor does not have sufficient data to make a good estimate of the share of the market potentially available to him. Hence, the number of firms sometimes increases more rapidly than the market would warrant, leading to its supersaturation. As long as investors are patient and bankers are lenient in calling their loans, this state of supersaturation may persist for some time, but with a business slump or money panic the weaker firms will not be able to pay their notes or bills and will be forced into bankruptcy (or may be absorbed by stronger firms). Thus, a phase transition occurs and a state in which there are many small firms is transformed into one with a small number of larger firms.

This "condensation" effect is exhibited in Figure 18, where the number of operating railroads in the United States is plotted as a function of time. The growth is evident until the panic of 1907, which started the steady decline in the number. (The figure is based on Table Q284, pp. 735–736 of reference 30.)

The number of banks in the United States follows a similar pattern with the breaking point occurring in 1921, a post-World War I depression year. (The bank data summarized in Figure 19 is taken from Table X-580, pp. 1019–1020,

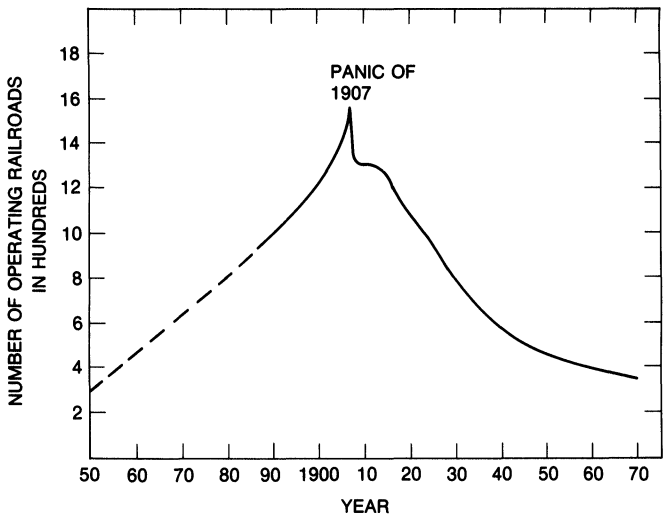

FIGURE 18. Number of operating railroads (in hundreds) in the years 1850 to 1970.[45]

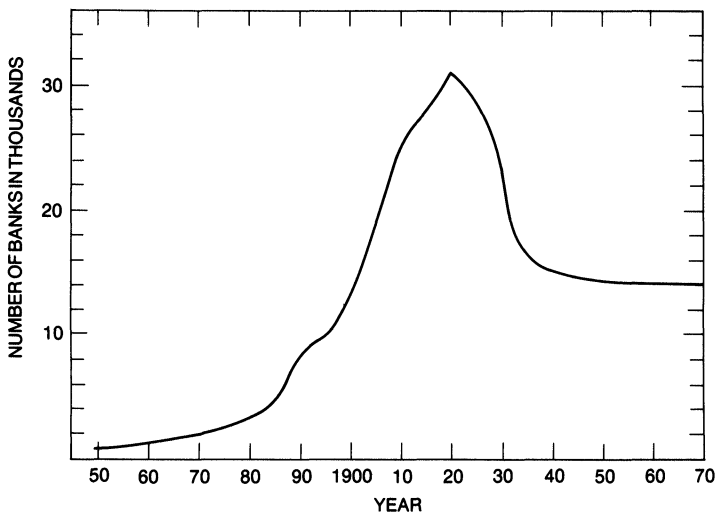

FIGURE 19. Number of banks (in thousands) in the years 1850 to 1970.[45]

of reference 30.) In the decade 1921–1930, 6,987 banks failed, with the heaviest mortality among small banks in small towns in areas of depressed staple agriculture.[45] "Hazardous practices of country bankers, encouraged by loose state banking laws and slack enforcement, needed only the collapse of crop prices and land values to render them fatal." The number of banks closed in the Depression years were: 1928, 491; 1929, 642; 1930, 1,345; 1931, 2,298 (the peak year); and 1932, 1,436. The peak month was October 1931 with 544 failures. The large city banks joined the rural procession into insolvency in 1930.

Bank holidays became common in the autumn of 1931 and continued through the winter. "When Roosevelt stood that Saturday afternoon [March 4, 1933] to take the oath as president, the economic heart of the country as symbolized by the banks, had stopped beating."[45] One of the first acts of the new administration was to declare Monday, March 6, the beginning of an extended bank holiday. Upon its termination on March 15, 5% (about 1,000) of the banks in the United States were declared insolvent. Various government actions during the holiday shored up public confidence in the surviving banks. The president's first fireside radio chat asked the public to return savings to the banks and refrain from making unnecessary withdrawals. His plea succeeded, providing more security to the banks than the accompanying legislation.

Condensation curves similar to those in Figures 18 and 19 can also be constructed for the number of automobile firms, number of telephone companies, and so on.

C. THE NAVIER - STOKES EQUATION AND THE USE OF DIMENSIONLESS CONSTANTS IN SCALING EXPERIMENTS IN HYDRODYNAMICS.The most fruitful modeling strategy for complex fluid dynamical systems is the use of dimensionless constants for design of experiments involving small-scale physical models of full-sized objects under investigation. The design of airplanes, ships, dams, harbors, canals, etc. would be impossible without scaling experiments. The physical basis and the experimental practice of this strategy will be described in terms of the Navier-Stokes equation for the flow field of an incompressible viscous fluid.

Let $\mathbf{v} \equiv \mathbf{v}(\mathbf{r}, t)$ be the velocity of a fluid element located at $\mathbf{r}$ at time t of an incompressible fluid of density ρ and with kinematic viscosity ν. The Navier-Stokes equation is[46]

$$\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} = -\nabla(p/\rho) + \nu \nabla^2 \mathbf{v} + \mathbf{F}/\rho, \tag{35a}$$

$p \equiv p(\mathbf{r}, t)$ being the pressure on the fluid element at $(\mathbf{r}, t)$ and $\mathbf{F}$ the external force acting on the fluid. The equation of continuity (a consequence of the conservation of mass) for an incompressible fluid is

$$\nabla \cdot \mathbf{v} = 0. \tag{35b}$$

Since (35a) is a vector equation, it corresponds to three scalar equations. These combined with (35b) yield four equations for the four variables, the pressure and the three components of the velocity vector. The force function will be specialized to be a gravitational force with

$$\mathbf{F}/\rho = \mathbf{g}. \tag{35c}$$

The Navier-Stokes equation is applicable to many processes involving fluids, such as the motion of ships and subsonic airplanes, and the flow of fluids through pipes. In the description of each of these processes one must introduce the boundary conditions required to specialize (35) to the process being considered. The nonlinear term $\mathbf{v} \cdot \nabla \mathbf{v}$ in (35a) is a tremendous mathematical nuisance. It makes hydrodynamics a difficult subject. On the other hand, it contributes to making it physically a rich subject.

Since the state of the art of solving nonlinear partial differential equations has not become sufficiently advanced to solve (35a) under the conditions required for the design of ships and airplanes or even for the prediction of the weather, it is important to develop experimental techniques for the investigation of the solutions of (35a) for engineering applications.

Scaling theory is based on the transformation of equation (35) to an equivalent equation for dimensionless quantities. The velocity, the pressure, and even g in (35a) are all variables with dimensions—their numerical values depend on the units chosen. To obtain a dimensionless equivalent of (35) we measure the local velocity, pressure, etc., as a multiple of some important basic dimensions of the object responsible for the flow pattern being investigated. Suppose, for example, that we wish to investigate the flow pattern of air around an airplane in flight. We let

V = average velocity of body being investigated;
L = an important unit of length of the body
(say the average width of an airplane wing);
P = average pressure in absence of body.

Then we can define a set of dimensionless quantities v', p', x', etc., by

$$v = Vv', \qquad x = Lx', \qquad p = \mathrm{P}p'. \tag{36a}$$

If we are concerned only with steady flow patterns, we can set $\partial \mathbf{v}/\partial t = 0$ in (35a). Now if we set

$$\nabla' = i\partial/\partial x' + \cdots = L\nabla, \tag{36b}$$

equation (35a) can be shown to have the form

$$\mathbf{v}' \cdot \nabla \mathbf{v}' = -P\nabla' p + (1/R)\nabla'^2 \mathbf{v}' + 1/F \tag{37}$$

with dimensionless coefficients defined by

$$R = \text{Reynolds number} = VL/\nu; \tag{38a}$$

$$P = \text{pressure number} = \mathrm{P}/\rho V^2; \tag{38b}$$

$$F = \text{Froude number} = V^2/Lg. \tag{38c}$$

Sommerfeld called the combination VL/ν the Reynolds number to honor Osborne Reynolds' pioneering studies on the onset of turbulence in flow of fluids through pipes. $F = V^2/Lg$ is named after William Froude, a junior naval architect under Isambard Brunel and Scott-Russell (of recently revived soliton fame) in the design of the *Great Eastern*. That great, underpowered, unprofitable iron ship (1858), from which the first successful Atlantic cable was laid, was a wonder of its time. Unfortunately, since its design required a giant leap from the state of the art, it was plagued by numerous engineering and management faults[47] (including poor cost estimation, a common curse of giant leaps). Froude's experiences with the *Great Eastern* motivated him to consider the possibility of estimating power requirements for ships from model tests.

There are certain flow regimes with the property that two out of the three terms on the right-hand side of (37) can be neglected relative to the remaining one. For example, suppose only the $1/F$ term need be retained. Then the flow velocity field and engineering design parameters that depend upon the flow field would be a function of only the dimensionless quantity F. Hence small-scale model experiments could be made to obtain design data for full-scale engineering of the device of interest.

It is easy to show that for a 1000-foot ship operating in the 40 ft/sec range, with pressures being measured in units of atmospheric presssure (and noting that the kinematic coefficient of viscosity of water at 15°C is $\nu = 1.23 \times 10^{-5}$ft/sec);

$$1/F = 20, \quad P = 0.69, \quad \text{and } R^{-1} = 10^{-9}.$$

Hence, the $1/F$ term is the most important one on the right-hand side of (37). Then ship modeling can, to a first approximation, be based on Froude modeling; i.e., modeling with a dimensionless constant that depends on g.

In Figure 12, we have plotted a typical operating curve of a ship as a function of $F^{1/2}$. A 10-foot ship model moving 4 ft/sec has the Froude number of a 1000-foot real ship at 40 ft/sec. Hence, by plotting the ratio of pounds of resistance per ton of displacement (a dimensionless quantity) of a 10-foot model towed at 4 ft/sec in a towing tank as a function of $1/F$, one can determine the power required to overcome the resistance expected by the full-scale ship.

In aerodynamics the first term on the right side of (37) is most important. Consider an airplane with a wing of width 10 ft, designed to operate at a speed of 800 ft/sec (about 545 miles/hr). Measuring the pressure in atmospheres (and using the kinematic coefficient of viscosity of air at 15°C, $\nu = 1.59 \times 10^{-4}$ ft^2/sec), $1/F = 5 \times 10^{-4}$, $p = 1.45$, and $1/R = 2 \times 10^{-8}$.

If, as suggested by these numbers, we need retain only the first terms on the right side of (37), the resulting equation is the Bernoulli equation of a nonviscous fluid $\nabla[\frac{1}{2}v^2 + p(\rho)] = 0$. Since the pressure difference between the bottom and top of the wing section of an airplane, as developed by circulation of air around the wing, determines the "lift" of the wing, it is not surprising that the pressure term is most important in our regime of interest. A wind tunnel[46] is the traditional device for measuring the lift and drag (and their ratio) on a model airplane in a flow stream. Since the length L does not enter into the pressure number, the lift-to-drag ratio would be the same on a small airplane model as on a full-scale object of the same shape.

4. The tyranny of many dimensionless constants. We have noted that simple questions about ships, airplanes, and flow of fluids through pipes can each, to a first approximation, be discussed in terms of a single dimensionless constant. Even without an understanding of hydrodynamics, ships sailed around the world; it was only with the advent of steam propulsion that naval architects sought to make calculations of the power required for a ship to perform in a desired manner. When sailors depended on the wind they tinkered with ship shapes and sail configurations to empirically improve performance.

It is remarkable that within ten years of the first flight of the Wright brothers, Igor Sikorsky (1913) built a four-engine giant (92-foot wing-span) capable of carrying a payload of 4.5 tons. Sikorsky's *Ilya Mourometz* (1914) stayed airborne for 6.5 hours and carried six passengers.[48] It could have served as a prototype for a commercial airliner. However, World War I intervened and soon pilots were shooting at each other in aerial dogfights, and bombing cities of their opponents. By 1924, twenty-one years after Kitty Hawk, the Imperial Airways flew Handley-Page airliners in routes from Cairo to Africa and India.[48] The 1928 model H-P 42s had a passenger capacity of 38.

On December 2, 1942, Enrico Fermi and his colleagues produced the first sustained fission chain reaction, demonstrating the feasibility of a nuclear power plant. By fifteen years later, the first commercial nuclear power plant was operating.

The highly successful space program was initiated only a few decades after primitive rocket experiments. Robert Goddard's first success in shooting a rocket over a mile vertically was achieved on May 31, 1935; twenty-six years later the Soviet cosmonaut was the first human to encircle the globe in a rocket-launched satellite.

The record of these successes has led the public, and even numerous scientists, to believe that with a little money and ingenuity, any desired scientific goal could be achieved. Unfortunately, this is not always true. Consider the magnetically confined fusion program which started as Project Sherwood in 1951. The aim of the project was to accelerate a plasma of ion-deuterons to a point where, at an energy equivalent to a temperature of about $200 \times 10^6\,^\circ\mathrm{K}$, nuclear fusion would occur with a tremendous release of energy. Since 1/6500 of the hydrogen in ocean water is composed of deuterium it is considered by optimists that, if the controlled fusion process can be successfully achieved, the energy problems of the world would be solved. Unfortunately, thirty-three years and hundreds of millions of dollars later, energy by magnetically confined fusion seems even further away than it did in 1951. What has happened? Why has this branch of physics failed to live up to expectations?

I contend that the magnetically confined fusion program has fallen victim to the tyranny of many dimensionless constants. The great engineering successes of the past have involved processes which could, to a first approximation, be characterized by a small number of dimensionless constants. Hence only a small number of model experiments were necessary to determine the feasibility of a project and to estimate the cost and difficulties to be surmounted. Even the space program was broken down into a number of subprojects, each of which could be analyzed in terms of a small number of dimensionless constants so that the results of many *independent* model tests could be used as a basis of the required full-scale engineering designs.

The complication of the magnetically confined fusion program seems to be that all the hydrodynamic dimensionless constants (about eight) as well as several electromagnetic and nuclear dimensionless constants are intimately connected in the process of transforming a low density, low temperature plasma to a higher density, very high temperature plasma. Since, as we shall

now indicate, the cost involved in, or the time required for the understanding of the nature of a process characterized by N interacting dimensionless constants can be expected to grow exponentially with N, we should not be surprised with the slow progress in the field of magnetically confined fusion.

Let N be the number of dimensionless constants required to characterize a process. Then an experimental program must sample $n_1 \times n_2 \times \cdots \times n_N$ points in the N-dimensional space of characterization. The cost of the program P should be proportional to the number of sampling tests; i.e.,

$$P = kn_1 \times n_2 \times \cdots \times n_N = k \exp N\left\{\frac{1}{N} \sum_{j=1}^{N} \log n_j\right\}.$$

Hence, if we define λ to be the average value of the logarithm of the number of observations for each dimensionless constant,

$$P = k \exp N\lambda$$

as was suggested to be the case.

The genius of individual inventors sometimes allows them to cut costs and time by going directly to the correct regime of the dimensionless constant of interest without conducting model tests over a broad range. A probabilistic argument similar to that given above indicates[49] that the probability of an individual's being identified as a genius by going "directly to the point" in the development of a technology that involves N connected dimensionless constants decreases exponentially with N.

Certain social situations and environmental processes might also depend on a large number of dimensionless constants. The understanding of these processes is not exempt from the tyranny of many dimensionless constants; nor is an attempt to make policies exempt, without a considerable insight into the manner in which a change in a single dimensionless constant influences others. Just as the enthusiast for magnetically confined nuclear fusion knows how he *would like* to solve the energy problem, so the enthusiast for social and environmental reform knows how he *would like* to make our lives full of harmony and beauty. Unfortunately, both of these classes of enthusiasts remain dreamers until the tyranny of many dimensionless constants is overcome.[50]

5. On some problems arising in the use of computers for the investigation of complex systems. Most modern students of subjects cultivated by mathematical models become adept in the employment of computing machines for the investigation of their models. A commonly used modeling style exploits sets of coupled rate equations (often nonlinear) for variables of interest. It is not unusual to attempt to describe complex systems by a dozen or more coupled differential equations. For computer analysis the rate equations are approximated by difference equations. Experienced investigators are generally aware of problems that arise from the sensitivity of the nature of the solution of the rate equations to the numerical choice of rate constants (the coefficients of the dependent variables) in the differential equations. In some cases a small change in the rate constants may lead to qualitatively different solutions of the

equations. This may result in several possible consequences—one may have discovered an important "instability" or "phase transition" in the phenomenon or one may have the wrong interpretation of a calculation because the rate constants were not properly chosen. Even in a system of five linear equations the required twenty-five rate constants may be difficult to find.

Since the determination of rate constants by experiment or observation of external events is generally costly, a discipline of *sensitivity analysis*[51] has evolved to develop strategies for deducing the most critical (or sensitive) constants in the process. Knowing these, more attention (or money) might be devoted to their precise estimation at the expense of paying less attention to those constants that are less relevant for the characterization of a model.

The purpose of this brief section is to direct attention to another problem concerning the analysis of nonlinear models that is not sufficiently recognized by some investigators. When a differential equation is converted to a difference equation, for certain initial conditions the difference equation may have qualitatively different solutions. We shall exhibit this for even the most simple case, that of the logistic equation

$$dx/dt = x(1 - x) \quad \text{if } x > 0 \text{ and } x(0) > 0. \tag{39}$$

It will be shown that the natural difference equation representation

$$(x_{n+1} - x_n)/h = x_n(1 - x_n), \qquad 0 < h < 1 \tag{40}$$

with $x_n \equiv x(hn)$ and $hn = t$ has solutions that are qualitatively different from those of (39) when $x(0) > 1/h$. Then a difference equation will be constructed whose solutions at the points $x(hn)$ lie precisely on the continuous solution of (39).

The usual difference equation does not always apply.

We now show through an examination of the logistic equation (39) that in the process of converting a nonlinear differential equation to a difference equation, solutions of the difference equation exist which differ qualitatively from those of the differential equation. The solution of (39) follows from separation of variables or by linearization through the transformation $y = x^{-1}$. It is found that

$$x(t) = x_0\{x_0 + (1 - x_0)e^{-t}\}^{-1}. \tag{41}$$

When $x_0 = x(0) > 0$ and $t > 0$ there are only two forms for this solution of (39):

(a) $x(t) \to 1$ in a monotone increasing manner from $x(0)$ when $0 < x(0) < 1$; and

(b) $x(t) \to 1$ in a monotone decreasing manner from $x(0)$ when $x(0) > 1$. There are no periodic solutions of (39) and $x(t)$ never becomes negative when $x(0) > 0$ and $t > 0$.

The difference equation (40), which may also be written as a recurrence formula

$$x_{n+1} = x_n + hx_n(1 - x_n), \tag{42}$$

would commonly be used to construct a numerical solution of (39). A detailed analysis of (42) for the regime $x_0 > 0$ and $0 < h < 1$ is provided in the Appendix (see also reference 50), where it is shown that four distinct ranges of initial conditions yield four distinct classes of solutions of (40). They are:[50]

(a) $0 < x_0 < 1$. The set of values $\{x_n\}$ derived from (42) form a monotone increasing set of numbers that approach 1 from below as $n \to \infty$.

(b) $1 < x_0 < h^{-1}$. The set of values $\{x_n\}$ form a monotone decreasing set of numbers that approach 1 from above as $n \to \infty$.

(c) $h^{-1} < x_0 < 1 + h^{-1}$. Here the set of values $\{x_n\}$ is not monotone; x_1 drops to between 0 and 1 and succeeding x_n's form a monotone increasing sequence approaching 1 from below.

(d) $x_0 > 1 + h^{-1}$. Here x_1 becomes negative and the succeeding x_n's remain negative, forming a monotone decreasing unbounded sequence.

Clearly only in cases (a) and (b) do the solutions of (40) mimic the solutions of (39). Hence in the range of initial conditions $x_0 > h^{-1}$ the solution of the difference equation bears no resemblance to that of the differential equation.*

An alert calculator understanding the structure of (39) would soon observe his error when he chose $x_0 > h^{-1}$ and correct for it by perhaps replacing h by $h/2$. However, suppose he was investigating a pair of coupled equations, say

$$\dot{x}_1 = a_{11}x_1 + a_{12}x_2 + c_{11}x_1^2 + c_{12}x_1x_2 + c_{13}x_2^2, \tag{45a}$$

$$\dot{x}_2 = a_{21}x_1 + a_{22}x_2 + c_{21}x_1^2 + c_{22}x_1x_2 + c_{23}x_2^2. \tag{45b}$$

Then he might not realize that the recurrence formulae derived by setting

$$\dot{x}_j = \left[x_j^{(n+1)} - x_j^{(n)}\right]/h \tag{46}$$

might not yield solutions that mimic those of (45), a set with twenty-seven classes of solutions (45) depending upon the rate constants and initial conditions.

Now suppose an investigator not to be alert to the fact that in the act of approximating (39) by (40) one might obtain a form of solution inappropriate to (39). Furthermore, suppose that of a total range that might be reasonable for a choice of x_0, the interval $(0, h^{-1})$ represented a fraction f. Then the probability that our uninformed investigator choosing an initial x_0 at random would produce a numerical solution of (40) mimicking the solution of the differential equation (39) would be a fraction f.

Even a more sophisticated analyst might be ignorant of an expected form for the solution of (45). Suppose that the user of a difference equation constructed from (45) by employing (46) were concerned with initial conditions in the

---

* In the case $h > 1$ many remarkable sequences have been generated from (42). If one sets

$$x_n = (1 + h^{-1})y_n \quad \text{and} \quad c = h + 1 \tag{43}$$

(42) becomes

$$y_{n+1} = cy_n(1 - y_n). \tag{44}$$

By proper choice of c and x_0 one may find periodic sequences, random sequences, bifurcation solutions, etc. Indeed a whole modern branch of mapping theory has centered around the recurrence (44) [cf. (42), (43), and (44)].[52,53]

range (A_1, B_1) for x_1 and (A_2, B_2) for x_2. Furthermore, suppose that in a fraction f_1 of the x_1 range, f_2 of the x_2 range, the solution of the difference equations representing (45) would mimic those of (45). Then the probability that a pair of initial values $[x_1(0), x_2(0)]$ would yield solutions of the difference equations properly mimicking the solution of (45) would be $f_1 f_2$ with each $0 < f < 1$. The argument immediately generalizes to the case of n coupled nonlinear equations.

It is doubtful that even an investigator clever enough to understand the expected behavior of solutions of (45) could comprehend the full variety of solutions available to the generalization of (45) to a set of nonlinear equations for a half-dozen variables. He would then be at the mercy of the computing machine. It is quite likely that for n nonlinear equations, the probability that the solution of the difference equations constructed from the rule (46) would mimic the solution of the generalization of (45) would be of the order of $f_1 f_2 \cdots f_n$ (with each $0 < f_j < 1$), so that in the regime of large n the probability of the machine solution of the difference equation resembling the solution of the differential equation could become vanishingly small.

One might argue that by decreasing the time interval h between recurrence steps each f_j could be increased. However this would be accomplished at the expense of increasing the number of iterations (probably exponentially in h^{-1}) required for the calculation of trajectories associated with a preassigned total process duration. Let the number of coupled variables become large, say of the order of five or more. Then the number of iterations required to follow a trajectory sufficiently long for the full effect of nonlinearities to properly exhibit themselves could become so large that round-off errors in the calculation might continue the apparent nature of the trajectories. The random statistical character of the round-off errors could effect computed trajectories in a manner similar to that of a random noise source upon natural trajectories in real life situations. Situations can be imagined with round-off errors acting as a "heat bath" with the driver variables achieving an equilibrium with the computer "heat bath."

Appendix[50]

We seek the qualitative characteristics of the solution to the recurrence formula

$$x_{n+1} = h\left[(1 + h^{-1}) - x_n\right] x_n \tag{A1}$$

when

$$x_0 > 0, \qquad 0 < h < 1 \tag{A2}$$

in four regimes of initial values

$$\begin{aligned} &\text{(a) } 0 < x_0 < 1, && \text{(c) } h^{-1} < x_0 < 1 + h^{-1}, \\ &\text{(b) } 1 < x_0 < h^{-1}, && \text{(d) } 1 + h^{-1} < x_0. \end{aligned} \tag{A3}$$

The character of the solution to (A1) in these regimes can be developed from the following properties of the mapping

$$\text{(A4)} \qquad \begin{aligned} y &= x[1 + h(1 - x)] \\ &= hx[(1 + h^{-1}) - x], \qquad 0 < h < 1. \end{aligned}$$

(i) If $0 < x < 1$, then $y > x$, since $1 + h(1 - x) > 1$.

(ii) If $x < 0$, then $y < x$, since $1 + h(1 - x) > 1$ and both x and y are negative; then $|y| = |x|[1 + h(1 - x)] > |x|$ as required.

(iii) $dy/dx > 0$ for all $0 < x < 1$, so that y is an increasing function of x in that range. To prove this statement, we first note that y is a continuous function of x and that

$$dy/dx = (1 - hx) + h(1 - x).$$

Then $dy/dx > 0$ since $(1 - x) > 0$ and $hx < 1$. Since $y(1) = 1$, this also implies

(iv) If $0 < x < 1$, then $y < 1$;

(v) If $x > 1$, then $y < x$ (this follows from (A4) and $h - hx < 0$ or $1 + h(1 - x) < 1$);

(vi) If $1 < x < h^{-1}$, then $y > 1$.

For proof, we note first that $y(1) = 1$ and $y(h^{-1}) = 1$. At the midpoint of the x interval $x = (1 + h^{-1})/2 = (1 + h)/2h$ so that

$$y([1 + h]/2h) = (1 + h^{-1})^2/4 > 1.$$

Inasmuch as $y(x)$ is a continuous quadratic function of x, it cannot achieve the value unity at any other choice of x than $x = 1$ and h^{-1}. Therefore, $y > 1$ if $1 < x < h^{-1}$, as required.

We are now in a position to find x_n for various initial regimes of x_0.

(a) If $0 < x_0 < 1$, then from (i) and (ii), by letting $x \equiv x_n$ and $y \equiv x_{n+1}$, $y(x_0 < x_1 < x_2 < \cdots < 1)$ is a monotone increasing function of n which never exceeds 1.

(b) If $1 < x_0 < h^{-1}$, then from (v) and (vi) $y(x_0 > x_1 > x_2 > \cdots > 1)$ is a monotone decreasing function of n, never becoming < 1.

In the above two regimes, the solutions of the difference equation (A1) have the same character as those of the differential equation (39).

Now consider the remaining cases:

(c) Let $h^{-1} < x_0 \leqslant 1 + h^{-1}$. We may then write $x_0 = h^{-1} + \varepsilon$, with $0 < \varepsilon \leqslant 1$. Hence $x_1 = (1 - \varepsilon)(h\varepsilon + 1) = 1 - (1 - h)\varepsilon - h\varepsilon^2 < 1$. The smallest value achievable for x_0 in the present regime corresponds to $\varepsilon = 1$, at which point $x_1 = 0$. Hence, if $h^{-1} < x_0 < 1 + h^{-1}$, then $0 < x_1 < 1$. Combining the above inequalities with (i), we find $0 < x_1 < 1 < x_0$ and $x_1 < x_2 < x_3 < \cdots < 1$. Therefore, in the present regime, x_n is not a monotone function of n since it first decreases with increasing n and later increases with increasing n, unlike the solutions of the differential equations.

(d) Let $x_0 > 1 + h^{-1}$. Then $h(x_0 - 1) > 1$ or $h(1 - x_0) < -1$, so that $x_1 = x_0[1 + h(1 - x_0)] < x_0(1 - 1) = 0$. Then from (ii), $\cdots < x_3 < x_2 < x_1 < 0$, so that all x_n for $n \geqslant 1$ are negative. Hence, when $x_0 > 1 + h^{-1}$, x_n

eventually becomes and remains negative for succeeding n, yielding a form of solution of the difference equation (A1) which is not obtained from the logistic differential equation.

The conclusion to be drawn from the above analysis is that there are regimes of initial conditions which yield solutions of difference equations which are qualitatively different from the classes of solutions of the differential equations that they were intended to mimic. Incidentally, several authors[54] have studied (A1) for $h \geqslant 1$ and have found it to have periodic solutions for some special initial conditions, another class not extant in the solutions of (39). There is no reason why the difference equations derived from coupled sets of nonlinear differential equations should not have classes of solutions that are not obtained from the differential equations.

There is another nonlinear difference equation which claims to be the more appropriate representative of (39) than (A1) for the regime $x > 0$ and $t \geqslant 0$ in that its solutions completely mimic the solutions (41). The more appropriate alternative equation is, for $0 < h < 1$,

$$x_{n+1} = \frac{x_n}{1 + h(1 - x_n)}. \tag{A5}$$

It might be argued that, if the right-hand side of (A5) is expanded in powers of h and only the term of $O(h)$ is retained, then (A5) would become equivalent to (A1). This is true only for certain regimes of initial conditions and corresponding values of h. Equation (A5) has the charm that it is easily solved in closed form by letting $y_n = 1/x_n$. Then (A5) becomes linear

$$y_{n+1} = y_n(1 - h) + h$$

or

$$(y_{n+1} - 1) = (y_n - 1)(1 - h)$$

so that

$$y_n = 1 + (y_0 - 1)(1 - h)^n$$

and

$$x_n = x_0\left[x_0 + (1 - x_0)(1 - h)^n\right]^{-1}, \tag{A6}$$

which has the same qualitative monotonicity properties as defined by (41), if $0 < h < 1$. Since $t = nh$, n does not have to be very large to make the approximation

$$(1 - h)^n = \left(1 - \frac{t}{n}\right)^n \sim e^{-t},$$

as required to pass from (A6) to (41) as $h \to 0$. When $h = 2$, (A5) has solutions of period 2 as is evident from (A6).

References

1. I. Fisher, *The application of mathematics to the social sciences*, 7th Gibbs Lecture (December 1929); reprinted in Bull. Amer. Math. Soc. **36**, 225 (1930).

2. J. K. Galbraith, *The Great Crash*, Houghton Mifflin, 1954.

3. P. Handler, President, National Academy of Sciences, Minutes of the Annual NAS Meeting (April 1981). [Not available to the general public.]

4. L. Boltzmann, Wein. Bericht. II, 373 (1877).

5. G. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); ibid, 623 (1948).

6. E.W. Montroll, *The Study of Information: Interdisciplinary Messages*, (F. Machlup and U. Mansfield, eds.), John Wiley, New York, 1983.

7. R. E. Chandler, R. Herman, and E. W. Montroll, Oper. Res. **6**, 165 (1958).

8. R. Herman, E. W. Montroll, R. B. Potts, and R. W. Rothery, Oper. Res. **7**, 86 (1959).

9. D. C Gazis, R. Herman, and R. B. Potts, Oper. Res. **7**, 499 (1959).

10. E. W. Montroll, *Recent Advances in Engineering Science*, Proc. 1st Symp. on Engr. Appl. of Random Function Theory and Problems, (Bagdonoff and Kozin, eds.), John Wiley & Sons, New York, 1963, p. 231.

11. E. W. Montroll, Proc. Nat. Acad. Sci. U.S.A. **78**, 7839 (1981).

12. H. Greenberg, Oper. Res. **7**, 79 (1959).

13. Sears and Roebuck Corp. Catalogues, 1900–1984.

14. D. L. Cohn, *The Good Old Days*, Simon & Schuster, New York, 1940.

15. B. Emmet and T. E. Jenck, *Catalogues and Counters*, Univ. of Chicago Press, Chicago, Ill. 1950.

16. R. Herman and E. W. Montroll, *Statistical Mechanics and Statistical Methods in Theory and Applications*, (U. Landman, ed.), Plenum, New York, 1977, p. 785.

17. M. F. M. Osborne, Oper. Res. **7**, 145 (1959).

18. E. W. Montroll, *Nonlinear Equations in Abstract Spaces*, (V. Lakshmikantham, ed.), Academic Press, New York, 1978, p. 161.

19. Amer. Automobile Assoc., *Automobile Facts and Figures*, Ann. Vols.

20. Scientific American (Nov. 1978), p. 94A.

21. W. Badger, *Mathematical Models as a Tool for the Social Sciences* (B. J. West, ed.), Gordon and Breach, New York, 1980, p. 87.

22. R. Gibret, *Les inégalités economiques*, Paris, 1931.

23. L. R. Klein, *An Introduction to Econometrics*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

24. I. B. Kravis, *The Structure of Income, Some Quantitative Essays*, Univ. of Pennsylvania Press, Philadelphia, 1962.

25. U. S. National Resources Comm., *Consumer Incomes in the U. S.* 1935–1936, U. S. Gov't. Print. Ofc., Washington, D. C., 1939.

26. H. T. Davis, *Theory of Econometrics*, Principia Press, Bloomington, IN, 1941.

27. E. W. Montroll and W. Badger, *Introduction to Quantitative Aspects of Social Phenomena*, Gordon and Breach, New York, NY, 1974.

28. E. W. Montroll and M. F. Shlesinger, Proc. Nat. Acad. Sci. U.S.A. **79**, 3380 (1982).

29. U. S. Dept of Commerce, Bureau of the Census, *Historical Statistics of the U. S., 1790–1957.*

30. U. S. Dept. of Commerce, Bureau of the Census, *Historical Statistics of the U. S., Colonial Times to 1970.*

31. E. T. Newell, *The Numismatist*, May issue (1916). Reprinted in *Selections from the Numismatist, Ancient and Medieval Coins*, Whitman, Racine, WI, 1960.

32. S. Smiles, *The Lives of the Engineers*, reprinted by MIT Press, Cambridge, MA, 1853. (See especially chapters on James Brindley.)

33. P. F. Verhulst, Mem. Acad. Royale de Belgique **28**, 1 (1844).

34. T. C. Fisher and R. H. Pry, Tech. Forecasting Soc. Changes **3**, 75 (1971).

35. C. Marchetti and N. Nakicenovic, *The Dynamics of Energy Systems and the Logistic Substitution Model*, Int. Inst. for Appl. Sys. Anal., Laxenburg, Austria, 1980.

36. R. Herman and E. W. Montroll, Proc. Nat. Acad. Sci. U.S.A. **69**, 3019 (1972).

37. E. W. Montroll, Proc. Nat. Acad. Sci. U.S.A. **75**, 4633 (1978).

38. T. G. Gillmore, *Modern Ship Design*, Naval Inst. Press, Annapolis, MD, 1975.

39. F. C. Nix and W. Shockley, Rev. Mod. Phys. **10**, 1 (1938).

40. T. H. Von Karmen, *Aerodynamics*, Cornell Univ. Press, Ithaca, N. Y., 1954.

41. C. V. L. Smith, *Advances in Electronics and Electron Physics*, vol. 4 (L. Marton, ed.), 1952, p. 157.

42. M. V. Wilkes, *Proc. Conf. on Digital Computers*, Nat. Physical Lab., England, 1953.

43. A. D. Booth, *Proc. Conf. on Digital Computers*, Nat. Physical Lab., England, 1953.

44. R. Keys, *Science* **196**, 945 (1977).

45. B. Mitchell, *Depression Decade*, Vol. IX of *The Economic History of the United States*, Rinehart & Co., New York, 1947, p. 128.

46. L. Prandtl and O. G. Tietjens, *Applied Hydro- and Aero-Mechanics*, McGraw-Hill, Englewood Cliffs, N. J. (Dover, reprint, 1957).

47. J. Dugan, *The Great Iron Ship*, Harper, New York, 1953.

48. J. W. R. Taylor (ed.), *The Lore of Flight*, Crescent Books, New York, 1974.

49. E. W. Montroll, *Proc. of EPA Meeting on Mathematical Modeling* (J. Fisher, ed.), in press (1985).

50. E. W. Montroll, *Mathematical Biology: A Conference on Molecular Science*, Pergamon Press, New York, 1981, p. 179.

51. R. I. Cukier, H. B. Levine, and K. E. Shuler, J. Comp. Phys. **26**, 1 (1978).

52. P. Collet and J. P. Eckmann, *Iterated maps on the interval as dynamical systems*, Birkhäuser, Basel, 1980.

53. E. Ott, Rev. Mod. Phys. **53**, 643 (1981).

54. S. Smale and R. F. Williams, J. Math. Biol. **3**, 1 (1977).

INSTITUTE FOR PHYSICAL SCIENCE AND TECHNOLOGY, UNIVERSITY OF MARYLAND, COLLEGE PARK, MARYLAND 20742

DIVISION OF APPLIED NONLINEAR PROBLEMS, LA JOLLA INSTITUTE, 3252 HOLIDAY COURT, SUITE 208, LA JOLLA, CALIFORNIA 92037 (Address of Bruce J. West)