Discrete Applied Mathematics 19 (1988) 113-128
North-Holland

THRESHOLD PHENOMENA IN RANDOM STRUCTURES

Joel E. COHEN

The Rockefeller University, 1230 York Avenue, New York, NY 10021-6399, USA

Received 18 May 1985
Revised 28 January 1986

The physical theory of phase transition explains sudden changes of phase in materials that undergo gradual changes of some parameter like temperature. There are analogs of phase transition in the theory of random graphs, initiated by Erdös and Rényi. This paper gives a nontechnical but precise account, without proofs, of some of the beautiful discoveries of Erdös and Rényi about threshold phenomena in graphs, describes an application of their methods to interval graphs, and gives some examples of threshold phenomena under other definitions of randomness and in combinatorial structures other than graphs. The paper offers some speculations on possible applications of random combinatorial structures to telecommunications, neurobiology, and the origin of life.

> A rich man commissioned three experts, a veterinarian, an engineer, and a theoretical physicist, to find out what made the best race horses. After a few years they reported their results. The vet concluded from genetic studies that brown horses were the fastest. The engineer found that thin legs were optimal for racing. The theoretical physicist asked for more time to study the question because the case of the spherical horse was proving extremely interesting.
>
> Aharon Katchalsky

> No one is exempt from talking nonsense;
> the only misfortune is to do it solemnly.
>
> Montaigne

1. Introduction

How does it happen that ordinary water, superficially well behaved as its temperature is raised from 1° to 99° C, abruptly changes to steam and remains steam as its temperature rises above 100° C? Sudden changes of phase in response to gradual changes of some parameter such as temperature or pressure are widespread among materials. The physical theory of phase transitions is devoted to explaining such changes.

In the mathematical models of this theory, a phase transition appears only in the limit as the number of particles or interacting units in the system becomes very large. The large number of elements is crucial to the possibility of an aburpt change in overall quality as a function of smooth changes of a parameter.

There are analogs to phase transition in the theory of random graphs of Erdös and Rényi [11]. This theory appears not to be widely known except among specialists in probabilistic combinatorics. Kennedy [17] proposes using a modification of the Erdös-Rényi theory of random graphs to explain properties of water. My purpose in Section 2 is to give a nontechnical but precise account of selected results of the beautiful theory of Erdös and Rényi. I will eschew proofs altogether. I will then (in Section 3) describe some recent results of applying the method of Erdös and Rényi to the problem of finding the probability that a large random graph is an interval graph. This problem arises in diverse life sciences. In Section 4, I will give examples to show that threshold phenomena like those discovered by Erdös and Rényi arise under other definitions of randomness and in combinatorial structures other than graphs. Finally, in Section 5, I will offer some pure speculation on possible applications of random combinatorial structures to telecommunications, neurobiology, and the origin of life. I emphasize that Section 5 is speculative to avoid discrediting the empirically detailed applications of the theory of Erdös and Rényi in Section 3. Section 5 is to biology as the rich man's theoretical physicist is to horse racing.

2. Erdös and Rényi

Erdös and Rényi [11] need no interpreter: their exposition is as beautiful as their results. In this account, I read between the lines of their proofs in order to highlight some of their conclusions.

A graph is a set of some positive number n of labelled points or vertices $P_1, \ldots, P_n$ and a set of some positive number N of edges, which are distinct unordered pairs $\{P_i, P_j\}$ with $i \neq j$. Parallel edges and edges from a point to itself are excluded.

There are $\binom{n}{2} = n(n-1)/2$ possible edges in a graph on n points. The number of graphs with N edges on n points is the number $C_{n,N}$ of ways of choosing N edges from the $\binom{n}{2}$ possible edges. A random graph $G_{n,N}$ with n points and N edges is defined as one chosen by regarding each of the $C_{n,N}$ graphs as equiprobable.

One graph is a subgraph of a second if the set of points of the first is a subset of the set of ponts of the second and if the set of edges of the first is a subset of the set of edges of the second.

Now suppose that the number n of points of a random graph $G_{n,N}$ gets very large, i.e., increases beyond any finite bound. Consider the subgraphs of $G_{n,N}$ under various assumptions about the number N of edges.

To take a trivial case first, suppose that as n increases, the number N of edges is always bounded above by some fixed finite constant greater than 2. The proportion of all possible graphs in which any two edges have a common point will grow smaller and smaller. In the limit, the probability that a random graph with a bounded number of edges contains two edges with a common point is 0.

Could the number N of edges increase without bound as the number n of points gets arbitrarily large so that, in a random graph $G_{n,N}$, every edge would still be an isolated edge with probability approaching 1? Yes: all that is required is that $N = o(n^{1/2})$, meaning that, in the limit as n gets large, the ratio of N to the square root of n approaches 0.

Let $\lim_n$ mean the limit as n approaches infinity. If $\lim_n N/n^{(1/2)-\varepsilon} = c$, where c and ε are any positive constants independent of n, then, with a probability that approaches 1, no two edges of $G_{n,N}$ will have a point in common. However, if $\lim_n N/n^{1/2} = c$, where again c is a positive constant that does not depend on n, $G_{n,N}$ will contain a tree of order 3 (three points linked by two edges) with a probability that approaches a positive limit depending on c. (More generally, a tree of order k is a connected graph with k points and $k-1$ edges such that none of its subgraphs is a cycle. A cycle of order k is a cyclic sequence of k edges of a graph such that every two consecutive edges and only these have a common vertex. A graph is connected if every pair of its points belongs to some sequence of edges, called a path, such that every two consecutive edges and only these have a point in common.) If $\lim_n N/n^{1/2} = \infty$ (for example, if $\lim_n N/n^{(1/2)+\varepsilon} = c$, where c and ε are positive constants), then a random graph $G_{n,N}$ will contain a tree of order 3 with a probability that approaches 1 in the limit as n increases. Provided $(1/2)+\varepsilon < 2/3$, almost no random graph $G_{n,N}$ will contain a tree of order 4 or larger or, for that matter, any connected subgraph with 4 or more points.

It is natural to call $n^{1/2}$ a threshold function for trees of order 3. With increasing n, if N (the number of edges) increases more slowly than $n^{1/2}$, trees of order 3 occur asymptotically with probability 0: for practical purposes, not at all. If N increases faster than $n^{1/2}$, trees of order 3 occur asymptotically with probability 1: for practical purposes, with certainty.

In my view, the most surprising finding of Erdös and Rényi is just that threshold functions exist and can be explicitly calculated for many fundamental properties of graphs. Having illustrated the meaning of a threshold function for trees of order 3, I now give the definition of a threshold function $A(n)$ corresponding to any property A of a graph. A function $A(n)$ that tends monotonically to $+\infty$ as n increases without bound is a threshold function for property A if the probability $P_{n,N}(A)$ that a random graph $G_{n,N}$ has the property A satisfies:

$$\begin{aligned} \lim_n P_{n,N}(A) &= 0 \quad \text{if } \lim_n N/A(n) = 0, \\ &= 1 \quad \text{if } \lim_n N/A(n) = +\infty. \end{aligned}$$

The following facts about threshold functions are consequences of a more general theorem of Erdös and Rényi:

The threshold function for the property that a random graph contains a tree of order k is $n^{(k-2)/(k-1)}$, for $k = 3, 4, \ldots$.

The threshold function for the property that a graph contains a cycle of order k is n, for $k=3,4,\ldots$.

The threshold function for the property that a graph contains a complete subgraph of order $k\geq 3$ is $n^{2(k-2)/(k-1)}$. (A complete graph of order k is a set of k points together with all $\binom{k}{2}$ possible edges on those points.)

The threshold function for the property that a graph contains a subgraph consisting of $a+b$ points $P_1,\ldots,P_a,Q_1,\ldots,Q_b$, and of all ab edges $\{P_i,Q_j\}$ is $n^{2-[(a+b)/(ab)]}$. (Such a graph is called a saturated even subgraph of type (a,b).) When $a=1$ and the total number of points in the saturated even subgraph of type $(1,b)$ is $k=1+b$, $2-[(a+b)/(ab)]=(k-2)/(k-1)$ and the threshold function reduces to that for a tree of order k, as desired.

To see these same facts from another point of view, consider the subgraphs of a very large random graph $G_{n,N}$ with N on the order of n^z. Suppose z increases gradually from 0 to 2. For z up to but not including $z=1/2$, almost all graphs contain only isolated edges or edgeless subgraphs. When z passes through 1/2, large random graphs suddenly contain trees of order 3 with probability 1. Such trees may also be viewed as saturated even subgraph of type (1,2). When z reaches 2/3, trees of order 4 suddenly appear, and these include saturated even subgraphs of type (1,3). As z gets closer and closer to 1, trees of larger and larger order appear, including saturated even subgraphs of type $(1,b)$ for larger and larger values of b. As long as $N=o(n)$, $G_{n,N}$ is the union of disjoint trees with asymptotic probability equal to 1. *Exactly when z passes through the value* 1, *even though z is changing smoothly, the asymptotic probability of cycles of all orders changes from* 0 *to* 1. Cycles of order 3 can also be viewed as complete graphs of order 3, and cycles of order 4 can also be viewed as saturated even subgraphs of type (2,2). When z passes 7/6, saturated even subgraphs of type (2,3) pass from probability 0 to probability 1, followed at $z=5/4$ by saturated even subgraphs of type (2,4). At $z=4/3$ complete graphs on 4 points appear simultaneously with saturated even subgraphs of type (3,3). As z continues to increase, saturated even subgraphs of larger and larger type and complete graphs of larger and larger order continue to appear. For even k, saturated even subgraphs of type $(k/2,k/2)$ appear at a value of $z=2(k-2)/k$ smaller than the value of $z=2(k-2)/(k-1)$ at which complete graphs with the same number of points appear. As z approaches 2, almost every random graph approaches the complete graph on n points.

Erdös and Rényi derive much more detailed information about the asymptotic probability distributions of the numbers of trees and cycles when the number of edges in a large random graph is close to the number of edges specified by the threshold function. I will give one example of an asymptotic probability distribution in the next section.

In addition to finding the threshold functions and the asymptotic probability distribution functions for important classes of subgraphs of random graphs, Erdös and Rényi investigate global properties of the large random graph $G_{n,N}$ in the sensitive region where $z=1$; that is, they consider the behavior of random graphs where

$\lim_n N/n = c$ for various values of the positive constant c. Each time I reread these theorems, I have the feeling that a miracle has just passed before my eyes.

An isolated subgraph G' of a graph G is defined as a subgraph such that if any edge of G has one or both endpoints belonging to G', then the edge also belongs to G'. Let $V_{n,N}$ be the number of points of a random graph $G_{n,N}$ that belong to an isolated tree contained in $G_{n,N}$, and let $E(\cdot)$ be the expected value of the random variable $(\cdot)$. Then when $\lim_n N/n = c$,

$$\begin{aligned} \lim_n E(V_{n,N})/n &= 1, && \text{for } c \leq 1/2, \\ &= x(c)/2c, && \text{for } c > 1/2, \end{aligned}$$

where $x(c)$ is the only root in the open interval $(0, 1)$ of the equation $xe^{-x} = 2ce^{-2c}$. ($x(c)$ can be computed using an infinite series.) For $c > 1/2$, the graph of $x(c)/2c$ roughly resembles an exponentially decaying function that drops from 1 asymptotically toward 0. (Erdös and Rényi give a picture.) Thus, in the limit, $E(V_{n,N})/n$ changes suddenly from a constant 1 to a sharply falling fraction as c passes beyond 1/2.

Recall that the threshold function for the appearance of cycles of all orders is n. Let $H_{n,N}$ denote the number of all cycles contained in the random graph $G_{n,N}$. Then, when $\lim_n N/n = c$,

$$\begin{aligned} \lim_n E(H_{n,N}) &= -(1/2)\log(1-2c) - c - c^2, && \text{for } c < 1/2, \\ E(H_{n,N}) &\sim (1/4)\log n, && \text{for } c = 1/2. \end{aligned}$$

Here $\sim$ means that the ratio of the quantities on the right and left approaches 1 as n increases. Thus for $c < 1/2$, the average number of all cycles remains bounded as n gets arbitrarily large, but increases without bound when $c = 1/2$. For $0 < c < 1/2$, with asymptotic probability 1, all components of $G_{n,N}$ are either trees or components containing exactly one cycle. (A component of a graph is a connected, isolated subgraph of the graph. The number of points belonging to the component is called the size of the component.)

If $S_{n,N}$ denotes the number of components of $G_{n,N}$ and $\lim_n N/n = c$, then

$$\begin{aligned} E(S_{n,N}) &= n - N + \mathrm{O}(1), && 0 < c < 1/2, \\ &= n - N + \mathrm{O}(\log n), && c = 1/2, \\ \lim E(S_{n,N})/n &= (1/(2c))(x(c) - (x(c))^2/2) && c > 1/2, \end{aligned}$$

where $x(c)$ is the same as before and $a(n) \doteq \mathrm{O}(b(n))$ means that $|a(n)|/b(n)$ is bounded as n increases. Here the bound on the O(1) term depends only on c. Equivalently, $\lim_n E(S_{n,N})/n = 1 - c$ for $c \leq 1/2$ but $\lim_n E(S_{n,N})/n$ decreases slower than linearly for $c > 1/2$.

I conclude this feast of phenomena with a double jump that even Erdös and Rényi, who must have been at home among such wonders, considered "one of the most striking facts concerning random graphs''. Let $R_{n,N}$ be the size of the largest component of $G_{n,N}$. When $\lim_n N/n = c$, $R_{n,N}$ is of order

$$\begin{aligned} &\log n, && \text{for } 0<c<1/2,\\ &n^{2/3}, && \text{for } c=1/2,\\ &n, && \text{for } c>1/2. \end{aligned}$$

More precisely, for $c>1/2$, and for any positive constant ε,

$$\lim_n P(|R_{n,N}/n - G(c)| < \varepsilon) = 1.$$

$G(c)$, the asymptotic fraction of all points belonging to the 'giant component', is given by $G(c) = 1 - x(c)/(2c)$ and $x(c)$ is as before. In this case ($c>1/2$), neglecting $o(n)$ points, $G_{n,N}$ consists, with asymptotic probability 1, only of isolated trees and of a single giant component whose size is asymptotically $G(c)n$. The number of the trees of order k is approximately $(n/(2c))k^{k-2}(2ce^{-2c})^k/(k!)$. As c increases, the giant component absorbs one isolated tree after another. The larger the tree, the larger the risk of absorption.

3. Interval graphs

In several areas of the life sciences, it is desirable to know the probability that a random graph, in the sense of Erdös and Rényi, is an interval graph. A graph G with a finite number n of points $P_1, \ldots, P_n$ and distinct undirected edges $\{P_i, P_j\}$, $i \neq j$, is an interval graph if, for each point P_i, there is a non-empty interval S_i of the real line such that $\{P_i, P_j\}$ is an edge of G if and only if S_i and S_j overlap, or have non-empty intersection. Komlós [7] has shown how the methods of Erdös and Rényi can be extended to calculate the probability that a random graph is an interval graph in the limit as n gets arbitrarily large. The asymptotic results are useful for finite numbers of points. Here I sketch the results and give two examples of how the question arises [7].

The possibility of applying the methods of Erdös and Rényi to find the asymptotic probability that a random graph is an interval graph depends on a characterization of interval graphs in terms of forbidden induced subgraphs. A subgraph G' of a graph G is an induced subgraph of G if there is an edge between two points of G' whenever there is an edge between two points in G. A graph G is an interval graph if and only if G contains no induced subgraph belonging to any of five specified classes of graphs. Four of these five classes of forbidden induced subgraphs contain cycles, and therefore have threshold functions that are not of smaller order than n. One of the five classes of forbidden induced subgraphs is a tree on seven points. From the results of Erdös and Rényi, the threshold function for the appearance of this tree as an isolated subgraph of a large random graph $G_{n,N}$ is $n^{5/6}$. Recall that a subgraph G' was defined as isolated if all edges of G, one or both endpoints of which belong to G', belong to G'. Thus every isolated subgraph G' is an induced subgraph G', but not conversely. So a random graph is an interval graph with probability 1 in the limit of large n if the number N of edges is of smaller order of magnitude than $n^{5/6}$ and is an interval graph with asymptotic probability 0 if N is of magnitude larger than $n^{5/6}$.

Now suppose $\lim_n N/n^{5/6} = c$. The probability that a random graph $G_{n,N}$ is an interval graph is asymptotically $\exp(-32c^6/3)$. This function $\exp(-32c^6/3)$ illustrates a class of functions called threshold distribution functions $F(c)$ by Erdös and Rényi. They discovered that among the structural properties A of graphs for which threshold functions $A(n)$ exist, there are some for which there also exists a probability distribution function $F(c)$ that is the limit of the probability that a random graph possesses property A as $\lim_n N/A(n) = c$. Erdös and Rényi computed threshold distribution functions for a variety of properties.

For large n and N, as long as N^6/n^5 is not orders of magnitude greater than 1, a more refined estimate of the asymptotic probability that a random graph $G_{n,N}$ is an interval graph is

$$\exp\left(-\binom{n}{7}(7!/6)p^6(1-p)^{15}\right), \quad \text{where } p = N\bigg/\binom{n}{2}.$$

Similarly precise formulas can be derived by the same methods for a variety of graphs related to interval graphs [8].

To determine how large n must be for these asymptotic formulas to be close to the truth, we generated 100 random graphs on a computer for each of several values of n, found the proportion of these graphs that were interval graphs, and compared the proportions with the probabilities given by the asymptotic theory [7]. For $n = 200$, the deviations between the Monte Carlo proportions and the asymptotic probabilities could be attributed to sampling fluctuations. For $n = 100$, the asymptotic theory was not too close to the Monte Carlo proportions.

The probability that a random graph is an interval graph is needed for statistical inference in biology. When graphs are observed to be interval graphs, it is desired to know how likely it is that these graphs would be interval graphs by chance alone. I give two examples.

Benzer, a biologist at the California Institute of Technology, is one of two independent inventors of interval graphs [2]. He wanted to know whether the genetic fine structure of a virus called T4 could be linear. Using $n = 19$ different clones of viruses with mutations in the rII region of their genetic material, he performed all possible $\binom{19}{2}$ recombination experiments and found $N = 61$ pairwise overlaps of the mutant regions. The graph with one point for each mutant clone and an edge corresponding to each overlap of two mutant regions was an interval graph.

Substituting $n = 19$ and $N = 61$ into $N = n^z$ gives approximately $z = 1.40$. Benzer's graph falls in the region where interval graphs would occur with probability 0 among random graphs if the asymptotic theory were relevant. The asymptotic threshold distribution function gives the probability that a random graph $G_{19,61}$ is an interval graph as exp(−221942) (and not 10^{-51}, as is mistakenly asserted in [7, p. 113]; I thank M. Golumbic for catching this error). The Monte Carlo studies for $n=10$ and $n=40$ confirm that the probability is very small that $G_{n,N}$ is an interval graph. All the genetic and physical evidence collected since 1959 has not altered the conclusion that genetic fine structure is linear in the rII region of bacteriophage T4.

In ecology, a 'trophic niche overlap graph' (which I originally called a 'competition graph', a name that has stuck among graph theorists) has a point for each kind of organism in some set and an edge between two points if there is some item of diet that both of the corresponding kinds of organisms eat. For example, a fish community on the rocky shore of Lake Nyasa has $n=28$ consumers and $N=256$ dietary overlaps. Here $z=(\log 256)/(\log 28)=1.66$ is even further into the region where the asymptotic theory says that interval graphs occur with probability 0. This and other natural communities have overlap graphs that are interval graphs. There is likely to be some special structure in the organization of diets among consumers that live together [6]. At least with the ways of assigning probabilities that have been used so far, this corner of nature appears to live in a set of measure 0. Various explanations of this observation have been proposed [10].

4. Other definitions, other structures

The threshold phenomena discovered by Erdös and Rényi also arise under other definitions of a random graph and in combinatorial structures other than graphs.

Another definition of a random graph, for example, requires a fixed number p, $0<p<1$. Define a random graph G_n on n points as one in which each edge $\{P_i, P_j\}$, $i\neq j$, occurs with probability p independently of all other edges. Erdös and Rényi mention that many of their threshold results hold true under this second definition as well as under the first definition of a random graph used in Sections 2 and 3.

Now define a clique to be a maximal complete subgraph, that is, a complete subgraph that is not contained in a bigger complete subgraph. Then, given $\varepsilon>0$, when n is large enough, almost every random graph G_n contains a clique with k points, where

$$(1+\varepsilon)(\log n)/\log(1/p)<k<(2-\varepsilon)(\log n)/\log(1/p),$$

but does not contain a clique with fewer than $(1-\varepsilon)(\log n)/\log(1/p)$ or more than $(2+\varepsilon)(\log n)/\log(1/p)$ points [5,23,24]. Thus, according to this theorem, in a perfectly random high school with $n=1000$ students, where any two given students have a one in ten ($p=0.1$) chance of knowing each other, cliques of 4 and 5 students are almost certain to exist, but not cliques of fewer than 3 or more than 6 students. (Though Bollobás and Erdös [5] give no quantitative information about how large n must be for their results to apply, I assume in this example that $n=1000$ is large enough.)

In view of the importance of combinatorial structures other than graphs in science and mathematics, it is reassuring that threshold phenomena, in the limit of large size, and explicitly calculable threshold functions are not restricted to graphs. Two results of Bollobás and Erdös [5] extend immediately to hypergraphs, which are structures with many applications [9].

A threshold also arises in a problem of Ulam. Let Q_n be any permutation of the first n positive integers. The integer k, $1 \le k \le n$, is permuted to $Q_n(k)$, $1 \le Q_n(k) \le n$. $L(Q_n)$ is the length of the longest increasing sequence in a random permutation Q_n, where a random permutation is one chosen with equal probability from the $n!$ possible permutations. How does $L(Q_n)$ behave as n gets large [18,21,30]?

For any $\varepsilon > 0$,

$$\lim_n P[2(1-\varepsilon) < L(Q_n)/n^{1/2} < 2(1+\varepsilon)] = 1.$$

Thus, in the limit of large n, almost every random permutation has an increasing sequence of length r if $r < 2n^{1/2}$ and almost no random permutation has an increasing sequence of length r if $r > 2n^{1/2}$.

These examples show that thresholds are not a peculiarity of a special definition of randomness nor a peculiarity of graphs. Threshold phenomena occur in a variety of random combinatorial structures in the limit of large size.

5. Some speculations

Erdös and Rényi [11] observed that "the evolution of graphs may be considered as a rather simplified model of the evolution of certain communication nets (railway, road or electric network systems, etc.) of a country or some other unit. (Of course, if one aims at describing such a real situation, one should replace the hypothesis of equiprobability of all connections by some more realistic hypothesis.)" They suggested that graphs with different types of points and different types of edges might yield "fairly reasonable models of more complex real growth processes (e.g. the growth of a complex communication net consisting of different types of connections, and even of organic structures of living matter, etc.)".

In each of the following speculations, the graph theory or the equiprobability assumed by Erdös and Rényi require elaboration. The threshold theorems for these models remain to be discovered. That such theorems may exist is strongly suggested by the existence of limit theorems for random graphs and random directed graphs having unequal edge probabilities [19,20].

A natural way to view the telephone network of the United States is to treat each subscriber as a point of a large graph and each interconnection as one edge. Initially, there were many small independent telephone companies. Gradually more and more of these companies become connected to the Bell System. Now interconnection with the Bell System is almost universal, both in the United States and worldwide [1,13].

One might consider, as a rough model, a Poisson distribution of central telephone exchanges. The intensity of the Poisson process might vary in space and time with population density and economic indicators. All subscribers connected to a given exchange would correspond to points in a complete subgraph. Complete subgraphs could grow by accretion of individual subscribers and by connections between central exchanges. Both accretion and interconnection could be modeled by random processes. One might calibrate such models against the quantitative details of the early history of American telephone companies, if it were possible to obtain credible data in the form required to estimate model parameters. It would then be interesting to see whether such models, like the graph-theoretic models of Erdös and Rényi, predict the discontinuous emergence of a 'giant component' that corresponds to the Bell System.

In referring to "organic structures of living matter", Erdös and Rényi may well have had the brain in mind. If so, their hint is being taken to heart, so to speak, by neurobiologists only very slowly. The first reference to Erdös and Rényi in the neurobiological compendia and papers available to me appears in a manuscript of Bienenstock [3]. Bienenstock [4] proposes and investigates numerically a dynamic brain model that generalizes the Ising model of statistical mechanics to allow for randomly changing edges (or interactions) between sites (or neurons). He uses the random graph model of Erdös and Rényi as a null model against which to measure the emergence of structure.

To make one possible interpretation of the graph theory slightly more explicit, consider a large number of neurons. (Here 'large' means only large enough to make the asymptotic theory relevant, which may be far fewer than the estimated 10^{10} neurons of the human brain.) Suppose that the fraction of all pairs of neurons that were functionally connected gradually increased during phylogenetic or ontogenetic development. If (contrary to all the evidence on the specificity of neuronal connections) these connections were made at random, as defined by Erdös and Rényi, then when the number of connections exceeded the number of neurons, cycles of all orders (less than the number of neurons) would pass from asymptotic probability 0 to asymptotic probability 1. The existence of cycles might be associated with significant changes in the functioning of the nervous system. For example, cycles of neurons have been proposed as the physiological basis of short term memory. Similarly, perhaps the increasing extent to which brains dominate nervous systems in phylogeny could be modeled formally by the growth of the giant component in a random graph. Again, the interesting question is how many of the known quantitative details are consistent with the theory of Erdös and Rényi or with some other quantitative theory of random structure.

This simple interpretation of the graph-theoretic model may require at least five improvements. First, it may be more useful to identify the points of a randomly connected graph not with neurons but with synapses between neurons and to view an edge of such a graph as an interaction between synapses. Second, since neuronal connections are typically oriented or directed, an extension of the method of Erdös and Rényi to directed graphs might prove necessary. Third, since the large-scale architecture of a vertebrate brain is clearly not random, it will be necessary either to replace equiprobable interconnections of elements by probability assignments that reflect known anatomy or to narrow the application of the model to regions where equiprobable connections are plausible. Fourth, it may be less useful to model interconnections between neural units (neurons or synapses) as all-or-none than to model them as graded. The effect of changing discrete to graded connections on the properties of a random graph in the limit of large size is not clear. Fifth, for an understanding of neural functioning, it may be necessary to replace the simple points of graph theory by some kind of computing element. Neurobiologists should be able to suggest other improvements, each of which will challenge mathematicians.

Randomly constructed nets of elements that compute randomly chosen logical functions have been simulated [14] for another purpose, to which I turn next, but could be interpreted as neural models. Other than the results described in Section 3 on random permutations, which arise as an unrealistic special case of these models, I know of no threshold theorems for random computing nets. Kauffman [16] provides a useful and tantalizing recent review of models and numerical phenomena. MacDonald [22] reviews the work of Kauffman and the use of random directed graphs, mentioned above in Section 3, in ecology.

My final fantasy here concerns the origin of life. If ecology has fundamental problems, the origin of life must be one of them. "The sequence of events between the time when only the mixture of organic precursors existed in the early oceans and the time when the first living cell appeared, 3.1 billion [10^9] or more years ago, is still unclear. It is the only portion of the entire chain of events constituting biological evolution that is not yet understood. It is a crucial step, for it marks the transition from the nonliving to the living system. Somehow the organic molecules of the primitive ocean were assembled into that complex unit of life, the cell." (Oliver [25, p. 19]).

Oliver assesses too kindly the present understanding of biological evolution since the appearance of the first living cells, but focuses attention usefully on an even greater gap in understanding. According to his view (not accepted e.g. by those who believe life originated on a clay matrix), one may take as explained or explicable a primordial soup of organic precursors. "We visualize the primitive ocean containing in dilute solution a wide variety of organic compounds suitable as precursors for living systems. The environment is fairly stable over millions of years. As the compounds degrade they are replaced by more of their kind falling into the sea from the atmosphere, and they are modified at the surface by ultraviolet radiation." [25, p. 19]. How does this soup become transformed into an ensemble of self-reproducing systems?

According to Kauffman [14, p. 465], "One can little doubt that the earliest proto-organisms aggregated their [chemical] reaction nets at random in the primeval seas... Evolution, therefore, probably had as its initial substrate the behavior of randomly aggregated [chemical] reaction nets."

Kauffman studied the deterministic trajectories of randomly constructed automata. Each automaton contained a fixed number of elements, each of which had a fixed number K of 0 or 1 inputs that were outputs of other elements in the automaton. Each element computed its state, 0 or 1, at time t, according to a Boolean function, chosen initially at random and then fixed, of the K inputs at time t. The K elements of the automaton that were inputs to each given element were also chosen initially at random and then fixed. The state of the automaton at time t was the vector of states of its elements at t. At $t+1$, the outputs, if any, of an element were its state, either 0 or 1, at t.

Kauffman [14] identified each element as a gene and each automaton as a cell. He found parallels between the simulated behavior of his random automata and observations of cellular metabolism. He subsequently developed these parallels in much greater detail [15].

The trajectories of Kauffman's autonomous, deterministic automata must inevitably enter cycles. In computer simulations, Kauffman [14] studied the typical lengths of such cycles as a function of the number K of inputs per element and the number of elements per automaton. For example, when $K=2$, he found that the typical cycle length increased approximately as $n^{0.3}$, for numbers n of elements in the range from 10^3 to 10^4. Here may be raw material for a threshold theorem. An exact theory of the asymptotic behavior of such automata remains to be developed.

To avoid the mathematical uncertainty, one might turn to the graph theory of Erdös and Rényi. Rössler [26, pp. 407–408] pointed out that aspects of prebiological evolution might be explained by the theorems of Erdös and Rényi. He did not identify in detail the points, edges, and probability assignments of Erdös and Rényi with the observable features of biochemical systems. One biochemical interpretation of the theory would be to pretend that each point of a graph stands for a chemical species, and that each edge stands for a reversible chemical reaction between two chemical species. Because a graph represents a binary relation, it is a fine model for a soup of isomers undergoing isomerization reactions. But isomerizations are of much less biochemical interest than chemical cycles, such as

$$A+B \leftrightharpoons AB,$$

$$AB+C \leftrightharpoons ABC,$$

$$ABC \leftrightharpoons A+BC,$$

$$BC \leftrightharpoons B+C,$$

in which each step is associated with a collision or dissociation. For such cycles of reactions, graph theory seems an inadequate language.

An alternative approach is to provide a combinatorial structure appropriate to chemistry, tentatively to assign probability distributions to this structure, and then to explore asymptotic threshold phenomena. With this approach one might hope at least to interest biochemists in the assignment of the probabilities and in interpreting any resulting theorems, since the fundamental units of the theory will be the nuts and bolts of biochemists' daily work.

The first step in this approach has been taken by Sellers [27]; see also [28, 29, 12]. I will suggest the flavor of his approach by describing in his terms the illustrative chemical cycle given above. First, some formalism.

In the example, there are three ultimate components, A, B and C. In probabilistic developments, the ultimate components could be fixed or could be sampled from a larger set of possible ultimate components. The ultimate components are free generators of a composition space C_0, which contains sums of ultimate components.

In the example, there are 6 chemical species, A, B, C, AB, BC, and ABC. A function δ_1 maps each chemical species into its expression in composition space C_0. E.g., $\delta_1(ABC)=A+B+C$, $\delta_1(BC)=B+C$, $\delta_1(A)=A$. In probabilistic developments, the chemical species could be fixed or could be sampled from all possible chemical species with the given set of ultimate components. Other possible chemical species include A_2, AC, B_2C_3 and so on.

The chemical species are the generators of a reaction space C_1. The points of C_1 are what appear on the two sides of a chemical equation, with the convention that what goes in on the left of a chemical equation takes a minus sign and what comes out on the right takes a plus sign.

An elementary mechanism, denoted in general by $j\times k$, for any two chemical species j and k, is assigned to a point in the reaction space C_1 by a function δ_2 according to $\delta_2(j\times k)=-j-k+jk$. Thus $\delta_2(j\times k)=0$ means $j+k=jk$. In the example, the four chemical equations can be rewritten in terms of four elementary mechanisms as

$$\begin{aligned}\delta_2(A\times B)&=0,\\ \delta_2(AB\times C)&=0,\\ \delta_2(-A\times BC)&=0,\\ \delta_2(-B\times C)&=0.\end{aligned}$$

In probabilistic developments, the elementary mechanisms could be fixed or could be sampled from the set of possible elementary mechanisms, given the chemical species. Other possible elementary mechanisms include $A\times C$, $B\times B$, $ABC\times C$, and so on.

The elementary mechanisms are the free generators of a mechanism space C_2, whose points are sums or differences of elementary mechanisms. The conversion of one mechanism $j\times k$ to another mechanism $h\times j+hj\times k-h\times jk$ is called a catalyzation and is denoted $h\times j\times k$, where h, j, k, hj, jk, and hjk are chemical species, and h is the catalyst for the reaction $j\times k$. A function δ_3 maps each catalyzation into the difference between any two points in C_2 that are related to each other by the catalyzation. Thus $\delta_3(h\times j\times k)=-j\times k+h\times j+hj\times k-h\times jk$. Any mechanism z is a cycle if $\delta_3(z)=0$, and Sellers proves that every cycle is a linear combination of cycles of the form $\delta_3(h\times j\times k)=0$. In our example, the entire cycle of chemical equations may be expressed concisely as $\delta_3(A \times B \times C) = 0$. In probabilistic developments, the catalyzations could be fixed or could be sampled from the set of possible catalyzations, given the elementary mechanisms. Other possible catalyzations using the same ultimate components are $A \times C \times B$, $A \times B \times B$, $A \times B_2 \times C$, and so on.

What is the profit of this (and more!) formality? Sellers sought to enumerate all possible combinations, subject to some constraints, of elementary mechanisms that would 'explain' a given mechanism. In so doing, he produced mathematically intelligible language for discussing chemical reaction systems.

Now it becomes possible to ask meaningfully: What is the distribution of the lengths of the cycles? How do the answers to these questions vary as one increases the fraction of all possible chemical species that are actual chemical species, given a set of ultimate components, or as one increases the fracton of all possible elementary mechanisms that are actual elementary mechanisms, given a set of chemical species? The answers to these questions depend on the probability distributions chosen. In this choice a knowledge of thermodynamics must play a role, if the answers to the questions just asked are to relate to reality. Mathematicians and scientists will need to collaborate in the analysis of these complicated structures.

The pot of gold that waits at the end of this rainbow is threshold laws like those found by Erdös and Rényi for random graphs. In particular, suppose that the probability that any given potential elementary mechanism actually occurs were to increase with time as a result, for example, of an increasing number of chemical species capable of acting as catalytic agents or enzymes in the primordial soup. Suppose also that there were a threshold function for the simultaneous appearance of cycles of all orders. Some of these cycles might be negative feedback cycles. Others might be positive feedback cycles. When the ratio of actual to potential elementary mechanisms passed smoothly through this threshold, one might suddenly observe an enormous increase in the number of positive feedback cycles. No special law would have to be invoked to explain why all the cycles necessary to the sustained growth of a self-replicating system would appear simultaneously. Natural selection acting among these competing chemical systems could then, in principle, lead to the organization of cells.

Is this program for studying the transition to life pie in the sea? Ultimately, only colleagues more expert than I am in the physical and chemical details can say. My hope is that this account will embolden these colleagues by making them aware of some surprising phenomena that mathematics can explain without magic.

Acknowledgements

I am particularly grateful for the encouragement and tolerance of Mark Kac, who died before this paper saw print. M. Marois provided an opportunity to present an initial version of this paper at a conference of the Institut de la Vie held in Vienna, July 1979. He then kindly held the manuscript without publishing it for six years.

E. Bienenstock, Lothar Collatz, Jack D. Cowan, Ronald F. Fox, Mark Kac, John W. Kennedy, Otto E. Rössler, Peter H. Sellers, Sol Spiegelman, Eric Wolman and a referee provided numerous suggestions, references and papers. This work was supported in part by the U.S. National Science Foundation grants DEB74-13276, DEB80-11026, and BSR84-07461 and by the hospitality of Mr. and Mrs. William T. Golden.

References

[1] A.T. and T. Long Lines, The World's Telephones; a Statistical Compilation as of January, 1978 (A.T. and T. Long Lines Overseas Administration, Bedminster, NJ, 1978).

[2] S. Benzer, On the topology of the genetic fine structure, Proc. Nat. Acad. Sci. USA 45 (1959) 1607-1620.

[3] E. Bienenstock, A statistical mechanics approach to the correlation theory of brain function and a numerical study of a related matrix differential equation, Manuscript, Université de Paris-Sud, February 1984.

[4] E. Bienenstock, Dynamics of central nervous system, in: J.P. Aubin and K. Sigmund, eds., Dynamics of Macrosystems (Laxenburg, Austria, September 1984) (Springer, New York, in press).

[5] B. Bollobás and P. Erdös, Cliques in random graphs, Math. Proc. Camb. Phil. Soc. 80 (1976) 419-427.

[6] J.E. Cohen, Food Webs and Niche Space (Princeton University, Press, Princeton, NJ, 1978).

[7] J.E. Cohen, János Komlós, and Thomas Mueller, The probability of an interval graph, and why it matters, Proc. Symp. Pure Math. 34 (Amer. Math. Soc., Providence, RI, 1979) 97-115.

[8] J.E. Cohen, The asymptotic probability that a random graph is a unit interval graph, indifference graph, or proper interval graph, Discrete Math. 40 (1982) 21-24.

[9] L. Collatz, Typen von Hypergraphen innerhalb und ausserhalb der Mathematik, in: L. Collatz, G. Meinardus, W. Wetterling, eds., Numerische Methoden bei graphentheoretischen und kombinatorischen Problemen, Band 2 (Birkhäuser, Basel, 1979) 37-65.

[10] D.L. DeAngelis, W.M. Post, and G. Sugihara, eds., Current Trends in Food Web Theory: Report on a Food Web Workshop (North Carolina, 1982), ORNL-5983 (Oak Ridge National Laboratory, Oak Ridge, TN, October 1983).

[11] P. Erdös and A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci. 5 (1960) 17-61.

[12] J. Happel and P.H. Sellers, Analysis of the possible mechanisms for a catalytic reaction system, Advances in Catalysis 32 (1983) 273-323.

[13] B.A. Hart, Geographical areas serviced by Bell and independent telephone companies in the United States, U.S. Dept. of Commerce Office of Telecommunications Rept. 73-1, Feb. 1973 (Washington, DC, U.S. Government Printing Office, 1973).

[14] S.A. Kauffman, Metabolic stability and epigenesis in randomly constructed genetic nets, J. Theoret. Biol. 22 (1969) 437-467.

[15] S.A. Kauffman, The large scale structure and dynamics of gene control circuits: an ensemble approach, J. Theoret. Biol. 44 (1974) 167-190.

[16] S.A. Kauffman, Emergent properties in random complex automata, Physica 10D (1984) 145-156.

[17] J.W. Kennedy, Icycles - I. Random graphs, physical transitions, polymer gels and the liquid state, in: The Theory and Applications of Graphs (Kalamazoo, MI, 1980) (Wiley, New York, 1981) 409-429.

[18] J.F.C. Kingman, Subadditive processes, in: P.-L. Hennequin, ed., Ecole d'Eté de Probabilités de Saint-Flour V-1976, Lecture Notes in Math. 539 (Springer, New York, 1976) 168-223.

[19] I.N. Kovalenko, On the theory of random graphs, Kibernetika (Kiev) 4 (1971) 1-4. [In Russian.]

[20] I.N. Kovalenko, The structure of a random directed graph [Russian], Teor. Verojatnost. i Mat. Statist. 6 (1972). [English transl.: The structure of a random directed graph, Theory of Probability and Math. Statistics 6 (1975) 83-92.]

[21] B.F. Logan and L.A. Shepp, A variational problem for random Young tableaux, Adv. Math. 26 (1977) 206-222.

[22] N. MacDonald, Trees and Networks in Biological Models (Wiley, New York, 1983).

[23] D. Matula, The employee party problem, Notices A.M.S. 19 (1972) A-382.

[24] D. Matula, The largest clique size in a random graph. Tech. Rep. Dept. of Computer Sci. (Southern Methodist Univ., Dallas, 1976).

[25] B.M. Oliver, ed., Project Cyclops: A Design Study of a System for Detecting Extraterrestrial Intelligent Life, Rev. ed. CR 114445 (National Aeronautics and Space Administration/Ames Research Center, Code LT, Moffett Field, CA 94035, 1973).

[26] O.E. Rössler, Chemical automata in homogeneous and reaction-diffusion kinetics, in: M. Conrad, W. Guttinger, M. Dal Cin, eds., Physics and Mathematics of the Nervous System, Lecture Notes in Biomath. 4 (Springer, New York, 1974) 399-418.

[27] P.H. Sellers, Combinatorial analysis of a chemical network, J. Franklin Inst. 290 (1970) 113-130.

[28] P.H. Sellers, An introduction to a mathematical theory of chemical reaction networks, Arch. Rat. Mech. Anal. 44 (1971) 23-40; 44 (1972) 376-386.

[29] P.H. Sellers, Combinatorial classification of chemical mechanisms, SIAM J. Appl. Math. 44 (1984) 784-792.

[30] A.M. Versik and S.V. Kerov, Asymptotic behavior of the Plancherel measure of the symmetric group and the limit form of Young tableaux [Russian], Dokl. Akad. Nauk SSSR 233(6) (1977) 1024-1027. [English transl.: Soviet Math. Dokl. 233 (1977) 527-531.]