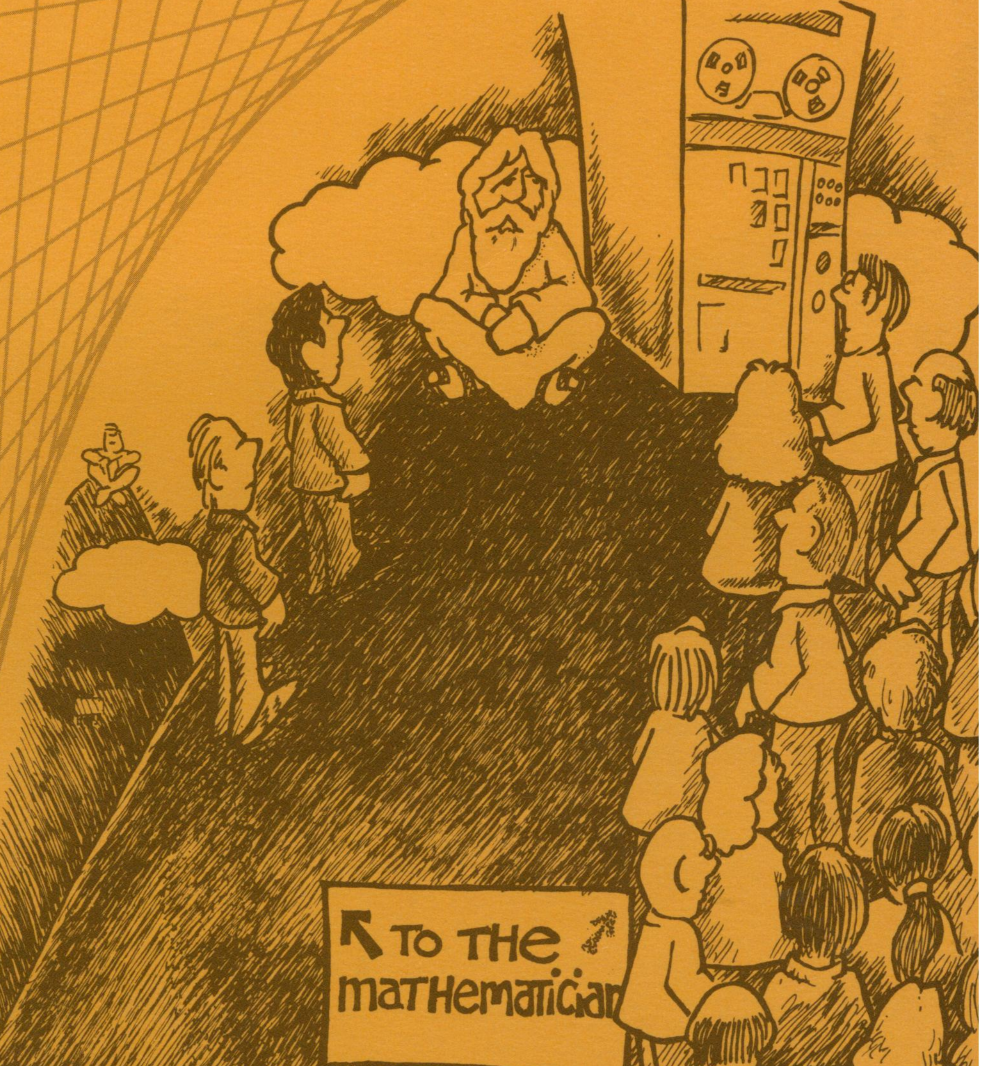
MATHEMATICS MAGAZINE



Vol. 53, No. 3
May 1980

GEOMETRY OF FABRICS • PROOFS BY COMPUTER
PROBABILITY EXAMPLES • MERCATOR'S SECANT

Mathematics Magazine

Volume 53, Number 3, May 1980
(ISSN—0025-570X)

COVER: Computers, a new source of proofs, undermine philosophical ideas of how mathematicians learn or know theorems. (See p. 131.)

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

EDITORIAL POLICY

Mathematics Magazine is a journal of collegiate mathematics designed to enrich undergraduate study of the mathematical sciences. The *Magazine* should be an inviting, informal journal emphasizing good mathematical exposition of interest to undergraduate students. Manuscripts accepted for publication in the *Magazine* should be written in a clear and lively expository style. The *Magazine* is not a research journal, so papers written in the terse "theorem-proof-corollary-remark" style will ordinarily be unsuitable for publication. Articles printed in the *Magazine* should be of a quality and level that makes it realistic for teachers to use them to supplement their regular courses. The editors especially invite manuscripts that provide insight into applications and history of mathematics. We welcome other informal contributions, for example, brief notes, mathematical games, graphics and humor.

New manuscripts should be sent to: Doris Schattschneider, Editor-Elect, Mathematics Magazine, Department of Mathematics, Moravian College, Bethlehem, Pennsylvania 18018. Manuscripts should be prepared in a style consistent with the format of *Mathematics Magazine*. They should be typewritten and double spaced on $8\frac{1}{2}$ by 11 paper. Authors should submit the original and one copy and keep one copy as protection against possible loss. Illustrations should be carefully prepared on separate sheets of paper in black ink, the original without lettering and two copies with lettering added; the printers will insert printed letters on the illustration in the appropriate locations.

ABOUT OUR AUTHORS

Thomas Tymoczko ("Computer, Proofs and Mathematicians: A Philosophical Investigation of the Four Color Proof") is Associate Professor of Philosophy at Smith College. His graduate work was in philosophy and logic (Harvard '72) following an undergraduate degree in mathematics (Harvard '65). Tymoczko maintains a philosophical interest in mathematics which he continues to hear about from his colleagues in the mathematics department at Smith. "Many interesting developments in mathematics seem to have some philosophical significance, but two things struck me about the Four Color Proof. One was the ease with which its philosophical significance could be developed, and the other was the depth of the challenge it posed to traditional philosophies of mathematics."

Branko Grünbaum and **Geoffrey C. Shephard** ("Satins and Twills: An Introduction to the Geometry of Fabrics") "discovered" fabrics somewhat accidentally while working on their forthcoming book *Tilings and Patterns*. During a lunch break, discussing the tiling aspect of a Scandinavian Christmas ornament, they observed that it is actually "woven" from thin strips of wood lying in three directions. This led to an inquiry into various mathematical aspects of weaves and fabrics. The topic turned out to be very interesting since non-trivial mathematical questions in geometry and combinatorics are here related to such activities as basket-making, weaving, and anthropology. The material on "2-way fabrics" presented in this issue is supplemented by soon to be published results on isonemal fabrics in which the strands are in three or more directions.

BUSINESS INFORMATION. Mathematics Magazine is published by the Mathematical Association of America at Washington, D.C., five times a year in January, March, May, September, and November. Ordinary subscriptions are $18 per year. Members of the Mathematical Association of America or of Mu Alpha Theta may subscribe at special reduced rates. Colleges and university mathematics departments may purchase bulk subscriptions (5 or more copies to a single address) for distribution to undergraduate students.

Subscription correspondence and notice of change of address should be sent to A. B. Willcox, Executive Director, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036. Back issues may be purchased, when in print, from P. and H. Bliss Co., Middletown, Connecticut 06457.

Advertising correspondence should be addressed to Raoul Hailpern, Mathematical Association of America, SUNY at Buffalo, Buffalo, New York 14214.

Second class postage paid at Washington, D.C., and additional mailing offices.

Computers, Proofs and Mathematicians: A Philosophical Investigation of the Four-Color Proof

Computer-assisted proofs illustrate the need for a more realistic philosophy of mathematics that allows for fallibility and empirical elements.

THOMAS TYMOCZKO
Smith College
Northampton, MA 01063

There are these hard facts. It is currently impossible to give a proof that four colors suffice to color any map without referring to computers. It is possible to describe a proof of the Four-Color Theorem that doesn't mention computers, but this description can apply equally well to valid proofs and to "invalid proofs." There is good evidence that this description applies to a valid proof, but this evidence refers to computers. Out of these hard facts we can attempt to weave something of philosophical significance. (For another attempt, see [11].)

Checking Proofs

Let's begin with an idea familiar to all mathematicians, the notion of checking a proof, or going over an argument and verifying that it establishes its conclusion. Checking proofs is a basic form of mathematical activity. It is something that any competent mathematician, even a good student, is supposed to be able to do. Checking is sometimes intermingled with other activities, like discovering proofs, but it often occurs in relative isolation. One gives a proof a final check before showing it to colleagues, one checks the proofs of colleagues, one reads articles and books and checks the proofs therein. Teachers check the proofs of students and vice-versa.

Mathematicians' ability to check proofs provides a simple idealized account of mathematical knowledge of theorems. How does a mathematician come to *know* a theorem? By checking a proof of it. Such knowledge is, of course, relative to certain background knowledge of axioms, rules of inference, and language, but such background knowledge won't matter much for our purposes here.

The simple ideal account must be complicated a bit to square with mathematical practice. Sometimes mathematicians claim to know a theorem because a teacher or colleague validated it for them. In such cases the first mathematician can be said to *borrow* knowledge of the theorem from another mathematician who has checked the proof—or who has borrowed it, in turn, from someone who has checked the proof, etc. Nevertheless, even in such cases mathematicians' knowledge of a theorem is ultimately based on the claim that some mathematician has checked a proof of it.

Thus the run-of-the-mill mathematical knowledge of theorems, lemmas, etc., rests on the run-of-the-mill activity of checking proofs. (Knowledge of axioms, on the other hand, is more controversial. Are axioms known intuitively on the basis of meaning, deduced from deeper principles, or arbitrarily chosen?)

Proofs and Purported Proofs

Let us look more closely at the idea of "checking a proof." If we're not careful, we'll find a too simple algorithm for checking proofs—when presented with a proof and asked to check it, answer, "Yes, it checks!" The trouble with this algorithm is that not only proofs get checked, but also misproofs, incomplete proofs, pseudoproofs. Call the whole lot—proofs and their assorted mimics—**purported proofs**. What are presented to mathematicians for checking are purported proofs. The process of checking distinguishes the proofs from their mimics: it verifies the proofs and reveals flaws in the mimics.

Thus "checking a proof" resolves into two components: the epistemological act of checking purported proofs and the logico-mathematical distinction among purported proofs between (real) proofs and counterfeits. (In **[11]**, I used the term "surveying" for checking a purported proof.)

Proofs themselves, more generally purported proofs, can be treated as objective facts or abstract patterns capable of existing independently of any checking subject (an image, the proof is there waiting to be discovered). The treatment of proofs as objective facts is carried out with great precision in the formal proofs of mathematical logic. However, it is not necessary to identify objective proofs with formal proofs in the logician's sense. One might be content to identify them with the non-formal but rigorous proofs published in mathematical journals and books.

If we view proofs as abstract patterns, it seems obvious that the *mere existence* of a proof of a theorem *guarantees* the truth of the theorem (modulo axioms, etc.). There aren't any "gaps" in the (real) proof; it is rigorous, indeed the standard of "rigor." In the case of formal proofs, the idea of rigor can be explained in terms of logical validity.

The degree of rigor for proofs is one of the main features that distinguishes mathematical proofs from arguments in the natural sciences. General scientific arguments do not purport to guarantee their conclusions. For one thing scientific arguments are usually supplemented by observations and/or experiments; and if the latter are flawed, the conclusion can be false no matter how good "the argument." Moreover, scientific arguments are usually probabilistic, investing their conclusions with a certain likelihood, often very high, but not with the apparent certainty we attribute to mathematical conclusions.

Philosophy of Mathematics

We have uncovered two plausible principles about proofs:

A. For someone to know a mathematical theorem, he or she must check a proof of it (modulo borrowing).

B. If a proof of a mathematical statement exists, then since a proof is rigorous, the statement is absolutely true (modulo axioms, etc.).

Put both principles together and we get a simple yet powerful picture of mathematical knowledge. If mathematicians know theorems via proofs and if proofs are logically rigorous, then mathematicians know the theorems of mathematics with absolute certainty. This picture is widely accepted among both philosophers and mathematicians. I suggest that the usual "isms" of philosophy of mathematics tend to assume this picture, e.g., Platonism, logicism, formalism, conventionalism, intuitionism. Its fundamental principles are conveniently summarized in Whitney and Tutte's advice to fellow mathematicians (see **[12]**) on the occasion of an early attempt at a computer-supplemented proof of the Four-Color Theorem that failed: "Anyone now having a proof that he wishes to be taken seriously would be well advised to write it out clearly and in full logical detail, so that any mathematician willing to spend enough time on it will be able to check it." In this paper my initial concern is not with the resultant picture of mathematical knowledge but with the two principles underlying it. Their conjunction, I will argue, is false.

The Four-Color Proof

I'll assume that readers know what the Four-Color Conjecture is, and that it has been recently elevated into a theorem (the Four-Color Theorem) on the basis of Appel, Haken and Koch's work (the Four-Color Proof). Since several good summaries of the proof are available (e.g., **[1]**, **[2]**, **[5]**, **[7]**), there is no need to provide one here. In any case, detailed knowledge of the Four-Color Proof is not necessary to follow the gist of my arguments. I will assume that some mathematicians *know* the Four-Color Theorem and that they know it on the basis of the Four-Color Proof. If principles A and B are jointly true, it must follow that the Four-Color Proof has been checked by some mathematician, and that this proof is rigorous. I would argue to the contrary that insofar as the Four-Color Proof has been checked, it is not rigorous and that insofar as it is rigorous, it has not been checked. Thus principles A and B are jointly false as long as mathematicians do know the Four-Color Theorem on the basis of the Four-Color Proof.

What has been Checked is not Rigorous

Take the Four-Color Proof to be that which has been checked by mathematicians, i.e., Appel, Haken, and Koch's paper, together with references and supplements. Does this provide a rigorous demonstration of the Four-Color Theorem? I think not, for an essential lemma in the Four-Color Proof is not justified by a presented proof. The lemma claims that each configuration in a certain set is D-reducible (or C-reducible). The grounds for this lemma are that suitably programmed computers delivered certain output when given certain input. However, such grounds are inherently fallible. The output could have been misread, the computers might have malfunctioned, the computers might have been misprogrammed, the programs might not have captured the mathematical intention. If any of these possibilities obtained, then, for all we know, the Four-Color Theorem might be false. Therefore, the Four-Color Proof does not eliminate all possibility that the Four-Color Theorem is false.

To be sure, the possibility for error is rather small. It does not preclude mathematicians from knowing the Four-Color Theorem any more than the possibility of error prevents scientists from knowing facts about the physical universe. But that small possibility of error does preclude mathematicians from knowing the Four-Color Theorem with absolute certainty. The proof is not rigorous.

This is an inevitable consequence of computer-supplemented proofs. To justify a lemma by computer verification is not the same as, nor is it much like, justifying a lemma by modus ponens. Mathematicians can "see" whether an application of modus ponens is valid or not, but they can't "see" whether an application of computer verification is. The computer application can fail for reasons not presented in the proof.

In point of fact, such a failure did occur when, in 1971, Shimamoto presented a purported proof of the Four-Color Theorem. This proof was logically rigorous and otherwise traditional except for one detail. A key lemma was justified not by proof, but on the grounds that a computer verified it. Furthermore, the key lemma of Shimamoto's proof is of the same form as that of Appel, Haken, and Koch's, that some configurations are D-reducible. This is not surprising because both purported proofs build on the work of H. Heesch who laid the foundation for computer assaults on the Four-Color Conjecture. In Shimamoto's case the computer supplement was wrong. Reprogramming the computer did not yield the same result; indeed, Whitney and Tutte proved without the aid of computers that the computer-based lemma is false. (For more detail on Shimamoto's purported proof, see Whitney and Tutte **[12]**.)

Many readers, I suspect, will in fact grant that essential use of computers in mathematical proof would violate some traditional standards of rigor. The fate of Shimamoto's purported proof is a logical possibility of any computer-supplemented proof, and it is a graphic demonstration of the perils of circumventing traditional rigor. Indeed, if we wished to defend the rigor of such proofs as the Four-Color Proof, the most plausible course is to deny that computers are essentially involved in them: it merely looks like they are. To this defense we now turn.

What is Rigorous hasn't been Checked

If we start with the idea that the Four-Color Proof is rigorous, we will want to identify this proof with a particular abstract pattern satisfying the usual requirements of rigorous proof. It should be a sequence of inevitable steps—no gaps—involving elementary logical operations (e.g., truth functions, quantification, sets) and the specific concepts, principles and methods of graph theory (e.g., graphs, configurations, D-reducibility, Kempe interchanges). If desired, we might equate the Four-Color Proof with some formal deduction in a sophisticated formalization of graph theory. In any case, it is obvious that no statement of a rigorous Four-Color Proof refers to computers and no step of inference is justified by computers. Computers disappear from the rigorous Four-Color Proof.

Let's concede that under this interpretation the Four-Color Proof is perfectly rigorous (but note that under this interpretation Shimamoto's proof could have been defended as perfectly rigorous). The proof pattern exists so the Four-Color Theorem is true. But for humans to know the Four-Color Theorem they must know the proof pattern. Someone must have checked it and recognized the proof. Even while granting the rigor of the Four-Color Proof, we can still ask who has checked it.

It has not been checked, not even by Appel, Haken, or Koch. Nor is it likely that any mathematician will check it. The problem is that the rigorous Four-Color Proof is extraordinarily long.

The rigorous Four-Color Proof starts with a set of configurations of order 10^3. Each configuration is listed separately; there is no general description defining the set. Although this is a rather large number of cases compared to the usual half-dozen cases in traditional proofs, it can be handled in the first part of the Four-Color Proof which shows the set to be unavoidable (every graph whose minimal vertex has five neighbors contains a configuration from the set). Appel and Haken give a general proof of this fact.

The real difficulty arises in showing that each configuration is D or C reducible. The rigorous Four-Color Proof must contain about 10^3 distinct subproofs, one for each case. A typical subproof of D-reducibility begins by examining each 4 coloring of the ring surrounding a given configuration. Since the rings contain, 11, 12, 13 or 14 vertices, the number of colorations to be examined in each case is of order 10^6 or 10^7. Each ring coloration must be matched with a 4 coloration of the configuration. Failing that, it must be modified by Kempe interchanges to another coloration which can be matched. Thus each subproof in the rigorous Four-Color Proof can be incredibly long and there are 10^3 such subproofs.

It is obvious that no mathematician has ever checked this proof and it is likely that none ever will.

Summary

Insofar as the Four-Color Proof has been checked, it is not rigorous and insofar as it is rigorous, it has not been checked. Thus, if mathematicians know the Four-Color Theorem on the basis of the Four-Color Proof—and I believe they do—then the two fundamental principles of proof which we began with are jointly false of the Four-Color Proof.

Before continuing with the main argument, let us pause to consider some objections that might be raised to the argument thus far.

Digression: Objections and Replies

A basic objection to the argument is the claim that there have always been recognized gaps between the checking and rigor of mathematical proofs. Thus the Four-Color Proof is not that unusual. The basic reply to this objection is that the recognized means for bridging these gaps do not apply to the Four-Color Proof so that thus it really is different in kind.

For example, a typical version of some traditional proof might fail to be rigorous because it involves borrowing the proof of some lemma from another mathematician. But there is no mathematician for Appel, Haken, and Koch to borrow from: there is only a computer.

Again, published proofs frequently contain omissions or sketches of arguments which, strictly speaking, violate rigor. On the other hand, the ordinary assumption is that the author and referee did verify the details, and that the reader could. The rigorous proof is checked by mathematicians, but not necessarily published. (Mathematicians are not supposed to leave proofs of lemmas as "exercises for the reader" when they themselves can't do the proofs!) This assumption, of course, fails for the Four-Color Proof.

Finally, one might try to explain the gap between rigor and checkability by distinguishing between informal proofs and formal proofs. Informal proofs get checked and published. Formal proofs are logically rigorous. An informal proof can be formalized by anyone who knows sufficient logic, although it frequently isn't. An informal proof of a theorem might be seen as a rigorous demonstration that a formal proof of the theorem exists.

While it is undeniably true that no formal Four-Color Proof has been checked, I have argued something much stronger, viz., that the informal Four-Color Proof has not been checked. It is not the details of reducing the proof to the primitive terms of mathematical logic that present the problem. All the details on configurations and colorations must be in the informal proof; they are what is to be formalized! Thus, whether the Four-Color Proof is a rigorous formal proof or a rigorous proof that can be formalized, it is not checkable.

So the traditional ways of explaining apparent gaps between checking and rigor do not apply to the Four-Color Proof. It is different in kind, and it drives a clear wedge between checking and rigor.

The Argument Resumed

If we are right so far, there are two plausible principles about proofs that don't come together on the Four-Color Proof, or computer-supplemented proofs in general. What can we make of this?

There's more than one way to approach mathematics from philosophy. The approach of this paper is to understand mathematics in its relation to the "doers" of mathematics. Such an understanding is called an epistemology of mathematics. Its basic question is "what is it to do (know) mathematics?" The "doers" of mathematics can be variously described—knowing subjects, rational minds, creative agents, human beings, trained mathematicians. However, the central model is composed of human beings, especially trained mathematicians. In this sense we can characterize the goal of the epistemology of mathematics as an attempt to understand mathematics as a human activity.

A more familiar approach to mathematics from philosophy is to concentrate on mathematical content, to view the central question as "What is mathematics?" This approach might be called ontological since it frequently focuses on the nature of mathematical entities. It considers questions like what are graphs? Are they geometrical configurations, abstract combinatorial patterns, complex sets, ideas, symbols? These various answers spin off into the traditional philosophies of mathematics.

These two approaches to mathematics, epistemological and ontological, are complementary not contradictory. Advances in one can shed light on the situation in the other. Mathematicians ordinarily approach mathematics from ontology, I suspect. They leave themselves out of their picture. But there are exceptions. A feature of intuitionistic mathematics is the requirement that the mathematician always remain conscious of his or her position in epistemological space.

Nevertheless, if we are to steer our way through the predicament posed by the Four-Color Proof, it is to epistemology of mathematics that we must turn.

Epistemology of Mathematics

If we focus directly on principles A and B, some simple revisions suggest themselves, revisions with significant consequences, however. We could preserve the ideal of rigorous proofs and concede that the Four-Color Proof has not been checked by mathematicians alone. But we could add a new way of checking rigorous proofs, via computers. Thus we get two different ways for mathematicians to check proofs, either directly as most proofs have been traditionally checked or indirectly using a computer. The indirect check brings a new kind of fallibility into mathematics, but that is the price of progress.

At this point our investigation could go either of two ways—inward into the details of computer supplemented proofs or outward in search of mathematical analogies to the computer. The first path would lead to questions about the nature and reliability of computing machines and programs, about the differences between Shimamoto's flawed effort, and the Four-Color Proof. What would count as a definitive check of the Four-Color Proof? Probably not organizing 10^3 mathematicians into a checking committee. K. Appel has suggested (in correspondence) that as home computers become more powerful, college freshmen will be able to check the computer-supplemented lemma. Such massive rechecking will increase our practical certainty, but will not eliminate computers from mathematics.

The second path leads to generalizations of computers, perhaps via the concept of "tool." Are there other tools in mathematics? (Some possibilities: pencil and paper, chalk and board, ruler and compass, language, formal system, logarithms, slide rules, etc. Both D. Isles and F. Pecchioni have suggested to me that *written* proofs, like computers, contribute empirical or a posteriori elements to mathematics.)

While it would be a worthwhile exercise to follow out the epistemological consequences of simply revising principles A and B, I feel it would be a mistake to restrict our attention to them. Floating behind these principles is a more basic framework for epistemology of mathematics, a model of the ideal case of mathematical knowledge. In the ideal, apart from borrowing, there is a single mathematician who is initially ignorant of a mathematical theory and finally knows it. The theory involves statements which this mathematician can understand and can come to know by privately checking proofs. It is with this framework that philosophers and mathematicians have tried to do epistemology. Rather than attempt piecemeal revisions of principles A and B, modern mathematicians and philosophers face a greater task in rethinking the basic framework for doing epistemology of mathematics. I think that the recognition that this task needs doing is perhaps the most important lesson of the Four-Color Proof, and I would like to conclude by repeating some suggestions that others have made to broaden the framework.

1. Drop the image of an isolated mathematician and admit that mathematical knowledge is public knowledge, the knowledge of a community of mathematicians. Recognizing this community will help us to recognize the communication among mathematicians that constitutes the doing of mathematics and grounds mathematical knowledge.

2. Drop the image of mathematicians as infallible. Recognizing the fallible nature of mathematical knowledge will help us to recognize the mechanisms by which mathematicians individually and in the community work to minimize error.
3. Once mathematics is recognized as fallible, there will be less difficulty in admitting tools as ingredients of mathematical knowledge supplementing proofs. Mathematical tools might provide an entry into the epistemology for the more general physical and cultural setting of the mathematical community.
4. On the other side of the epistemological coin, we can broaden our horizons beyond statements. In the first place a recognition of the richness of mathematical discourse—assertions, questions, conjectures, hypotheses, lemmas, criticisms, etc.,—is needed to describe mathematical communication. In the second place the foci of mathematical knowledge are often not individual statements but theories or even branches of mathematics. Reductionist epistemology holds that there must be one basic mathematical theory (be it logic, set theory, formal systems, etc.). But in fact since Godel's work we've known that no single reduction is satisfactory. Mathematics is not one, but many. There is always more than one mathematical theory in use at a given time. So we must consider more seriously the various relationships between mathematicians and the various overlapping theories being developed.
5. Finally, epistemology should abandon the idea of explaining how mathematicians go from knowing nothing to knowing modern mathematics. It is much more realistic to concentrate on mathematical development and to try to explain how mathematicians who already know x are able to learn y. We would thereby shift our sights from the one big unmanageable question to a lot of smaller, more tractable ones.

If these suggestions are vague and programmatic, they may nonetheless point the way to a more realistic framework for epistemology. Only in terms of such a realistic epistemology can we reach satisfying answers about checkability, rigor and computer supplements in proofs. More generally, only in a realistic epistemology will we achieve a philosophically satisfying understanding of mathematics that is faithful to actual mathematical practice. For some recent related attempts, see **[3]**, **[4]**, **[6]**, **[8]**, **[9]**, **[10]**, **[13]**.

Postscript

People sometimes object to the preceding argument on the grounds that at least in some cases they can be more certain of a computer based result than a result of a fellow mathematician—so what's the big deal about computer-supplemented proofs? If we examine this objection a little more closely, we can clarify the philosophical significance of the Four-Color Proof.

In brief, the objection is this: even though computers are fallible, on the mathematical matters at issue they are less fallible than mathematicians. So a mathematician might be more certain of a computer-supplemented result like the Four-Color Theorem than a result such as "all groups of odd order are solvable" which, while checked by mathematicians, has a very long and complex proof. In fact, most mathematicians borrow the complete proof of the latter theorem from Feit and Thompson. But they borrow only (the major) *part* of the Four-Color Proof from Appel, Haken, and Koch—part is borrowed from a computer and this can be viewed as an *increase* in certainty. Although this objection is not as obvious as it might first appear, rather than question it I would like to point out an implicit premise in it—the premise that mathematicians are fallible.

Obviously the objection assumes mathematicians are fallible. If they are not, if mathematicians are infallible, then it is simply wrong to claim that computers are less fallible than mathematicians.

On the other hand it is easy to establish that the great tradition in philosophy of mathematics ignores this assumption and proceeds as if mathematicians are infallible. (To assume infallibility one need not explicitly assume that mathematicians are infallible, one need only refrain from introducing the idea of mathematical mistake!) By "great tradition" I mean the rather coherent line of thought on mathematics developed by the Greeks (Euclid, Plato), advanced by modern philosophy beginning with Descartes and Leibniz through Kant, and dramatically advanced by the founders of modern mathematics such as Dedekind. Cantor, Frege, Russell, Hilbert, and Heyting. This tradition either ignores error or consigns it to nonmathematical causes (e.g., in Descartes' philosophy, mathematical error turns out to be a species of sin). In either case the effect is the same. Error is an accident, not part of the essence of mathematics: the mathematician qua mathematician is treated by philosophy as infallible. My claim is "checkable"—check the writings of the tradition for discussions of mathematical error!

The upshot of this is that the assumption that mathematicians are fallible, for all its common sense appeal, is philosophically very significant. It is not part of traditional theories of mathematical knowledge. It introduces a "new" topic to the philosophy of mathematics—error, uncertainty, fallibility. If philosophical reflection on the Four-Color Proof leads us to articulate and develop the assumption of mathematical fallibility, then I would say that the Four-Color Proof has proved its philosophical significance. But in fact it does more. The Four-Color Proof provides a positive argument for the inclusion of fallibility in the philosophy of mathematics.

Suppose we tried denying that mathematicians were fallible, would this keep error out of mathematics? No, not if mathematicians rely on computers to check proofs! Even if mathematicians qua mathematicians are infallible, computers are not. An actual mathematician might be regarded as a rational mind (the source of knowledge) caged in a physical body (the source of error). But present computers are just physical structures. Mathematicians' reliance on computers brings with it, in principle, the fallibility inherent in any physical science. In practice, it brings with it the special fallibility of computer science (computers are not just Turing machines and to the extent to which computers approximate Turing machines, computer programs do not approximate Turing algorithms). Thus, if for no other reason, fallibility is present in mathematics when mathematicians rely on computers.

Of course once we admit error into mathematics, it seems silly to recognize it only for computers. After all we began with the intuition that on occasion computers can be less fallible than mathematicians. We are returned to the conclusion of our original investigation. In order to understand the place of the Four-Color Proof in terms of mathematical knowledge, we have to reconstruct the theory of mathematical knowledge.

References

[1] K. Appel and W. Haken, The solution of the four-color map problem, Sci. Amer., 237 (October 1977) 108–121.

[2] F. Bernhart, A digest of the four-color problem, J. Graph Theory, 2 (1977) 207–225.

[3] P. J. Davis, Fidelity in mathematical discourse, Amer. Math. Monthly, 79 (1972) 252–262.

[4] R. DeMillo, R. Lipton, and A. Perlis, Social processes and proofs of theorems and programs, Comm. ACM, 22:5 (May, 1979) 271–280.

[5] W. Haken, An attempt to understand the four-color problem, J. Graph Theory, 1 (1977) 193–206.

[6] R. Hersh, Reviving the philosophy of mathematics, Advances in Math., 31 (1979) 31–50.

[7] P. Kainen and T. Saaty, The Four-Color Problem: Assaults and Conquest, McGraw-Hill, New York, 1977.

[8] I. Lakatos, Mathematics, Science and Epistemology, Cambridge Univ. Press, New York, 1978.

[9] I. Lakatos, Proofs and Refutations: The Logic of Mathematical Discovery, Cambridge Univ. Press, New York, 1976.

[10] H. Putnam, Mathematics, Matter and Methods, Philosophical Papers, vol. 1, Cambridge Univ. Press, New York, 1975.

[11] T. Tymoczko, The four-color problem and its philosophical significance, J. Philos., 76:2 (1979) 57–85.

[12] H. Whitney and W. T. Tutte, Kempe chains and the four-color problem, Utilitas Math., 2 (1972). Reprinted in Studies in Graph Theory, Part II, edited by D. R. Fulkerson, MAA Studies in Mathematics, vol. 12, pp. 378–413.

[13] R. L. Wilder, Evolution of Mathematical Concepts, Wiley, New York, 1968.

Satins and Twills: An Introduction to the Geometry of Fabrics

A mathematical investigation into patterns of weaving reveals subtle problems in combinatorics and geometry.

Branko Grünbaum
University of Washington
Seattle, WA 98195

Geoffrey C. Shephard
University of East Anglia
University Plain, Norwich NR4 7TJ

Weaving is one of the oldest activities of mankind and so it is hardly surprising that there exists a vast literature on the subject. But this literature is almost entirely concerned with the practical aspects of weaving; any treatment of the theoretical problem of designing fabrics with prescribed mathematical properties is conspicuously absent. And this is so in spite of the fact that many fabrics which are mathematically interesting were discovered empirically long ago by practitioners of the weaver's craft. One wonders how geometers can fail to be fascinated by the diagrams of fabrics that abound in the literature. Yet, so far as we are aware, the only papers that attempt to treat fabric design from a mathematical point of view are those of Lucas who published about a century ago, an isolated paper of Shorter which appeared in 1920, and a series of three papers by Woods published in 1935. All three authors were concerned principally with satins (or sateens), a type of fabric which we shall discuss in the third section of this paper.

The "geometry of fabrics", as we shall call it, involves ideas from elementary geometry, group theory, number theory and combinatorics. There is a large number of open problems, to some of which we shall draw attention in the following pages.

In order to make the subject manageable, it will be necessary to idealize the concept of a fabric. For example, we shall always assume that our fabrics are unbounded, that is, that they continue indefinitely in every direction. Thus edge-effects and selvedges (which are of great concern to the practical weaver) will be entirely ignored here. A fabric will consist of "strands" woven together and we shall only discuss those fabrics in which the strands are straight and lie in one of two directions, usually at right-angles to each other. Without these restrictions there are many other possibilities about which extremely little seems to be known.

As there is no accepted terminology, it will be necessary to begin by defining the words we shall use. A **strand** (see Figure 1(a)) is a doubly infinite open strip of constant width, that is, the set of points of the plane which lie strictly between two parallel straight lines. For purposes of visualization it is best to think of a strand as a strip of paper, or similar material of zero (or

(a) A strand (an open infinite strip) shaded to show its direction. (b) A layer of strands. Every point of the plane belongs to a single strand or to the boundaries of two strands.

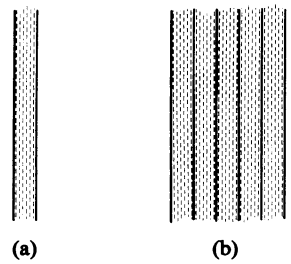
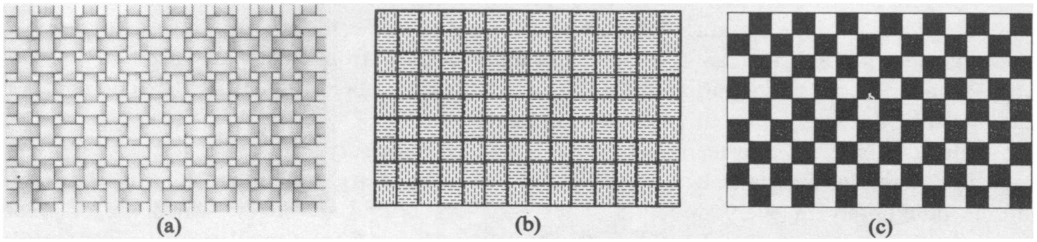


FIGURE 1.

negligible) thickness. In diagrams we shall sometimes use some method of shading, as in FIGURE 1(a), to indicate the direction of the strand. This will be essential for the correct interpretation of the diagram when only small portions of a strand are visible as in FIGURE 2(b). By a **layer** we mean a collection of congruent and disjoint parallel strands such that each point of the plane either belongs to (the interior of) one of the strands or is on the boundary of two adjacent strands (see FIGURE 1(b)).

The word **fabric** will be used in a mathematical sense to mean, roughly speaking, two layers of congruent strands in the same plane E such that the strands of different layers are nonparallel and they "weave" over and under each other in such a way that the fabric "hangs together." To be precise, "weaving" means that at any point P of E which does not lie on the boundary of a strand, the two strands containing P have a stated **ranking**, that is to say, one strand is taken to have precedence over the other, and this ranking is the same for each point P contained in both strands. This concept may be conveniently expressed by saying that one strand **passes over** the other, in accordance with the obvious practical interpretation. By saying that the fabric **hangs together** we mean that it is impossible to partition the set of all strands into two nonempty subsets so that each strand of the first subset passes over every strand of the second subset.

In FIGURE 2(b) we give a diagramatic representation of the commonest and most familiar of all fabrics, known variously as the **over-and-under**, **plain**, **calico** or **tabby weave**. Here the shading not only indicates the direction of each strand, but also shows which strand (a horizontal or a vertical one) passes over the other at each point of the plane. In order to avoid any possible misinterpretation we also give in FIGURE 2(a) a sketch of the same fabric. Here the strands have been "separated" for clarity—this diagram may be regarded as representing the "real" fabric corresponding to the "idealized" or "mathematical" fabric of FIGURE 2(b).

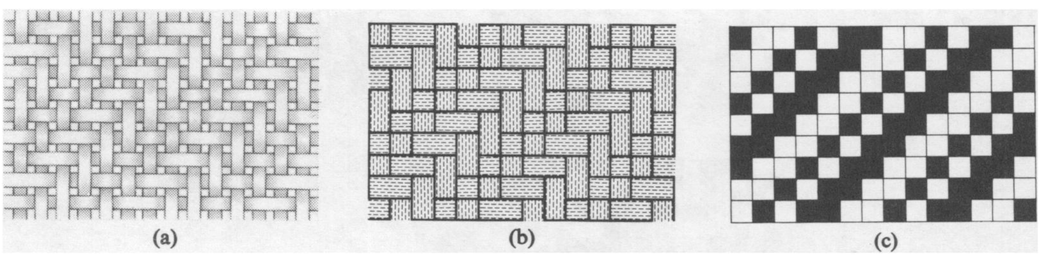


The most common and best-known type of fabric known as the over-and-under, plain, calico or tabby weave. (a) A sketch of the "real fabric." (b) The idealized fabric consisting of two superimposed layers of strands. (c) A design for the fabric. A white square means that a weft strand passes over a warp strand, and a black square that a warp strand passes over a weft strand.

FIGURE 2.

Traditionally the two layers in a fabric are called the **warp** and the **weft** (or **woof**). In a real fabric the warp runs lengthwise and the weft from side to side. In diagrams it is conventional to draw the warp vertically and the weft horizontally. Here it will be convenient to use the terms warp and weft in this sense.

A simple and convenient method of representing a fabric is by means of a **design** (also called a **diagram** or **draft** by some authors). This is constructed in the following way. We begin with the regular tiling of the plane by unit squares. Each square is the intersection of a row of squares (corresponding to a weft strand) and a column of squares (corresponding to a warp strand); according to the more usual convention we color the square white if the weft strand passes over the warp strand, and we color it black if the warp strand passes over the weft strand. Thus the design may be regarded as indicating the appearance that the fabric would have if the weft strands were colored white and the warp strands were colored black. For example, in FIGURE 2(c) we show a design for the plain weave; the design can be easily obtained from FIGURE 2(b) by replacing vertical shading by black and horizontal shading by white. Another example appears in FIGURE 3. The fabric shown in this figure is called a balanced twill of period six and is an example of a large class of fabrics of practical importance known as twills. A discussion of twills and their properties will be given in the next section. For an application of a computer to the drawing of fabric designs, see Huff [5].



A balanced twill of period six: (a) is a sketch of the "real fabric," (b) is the idealized fabric, and (c) is a design for this fabric.

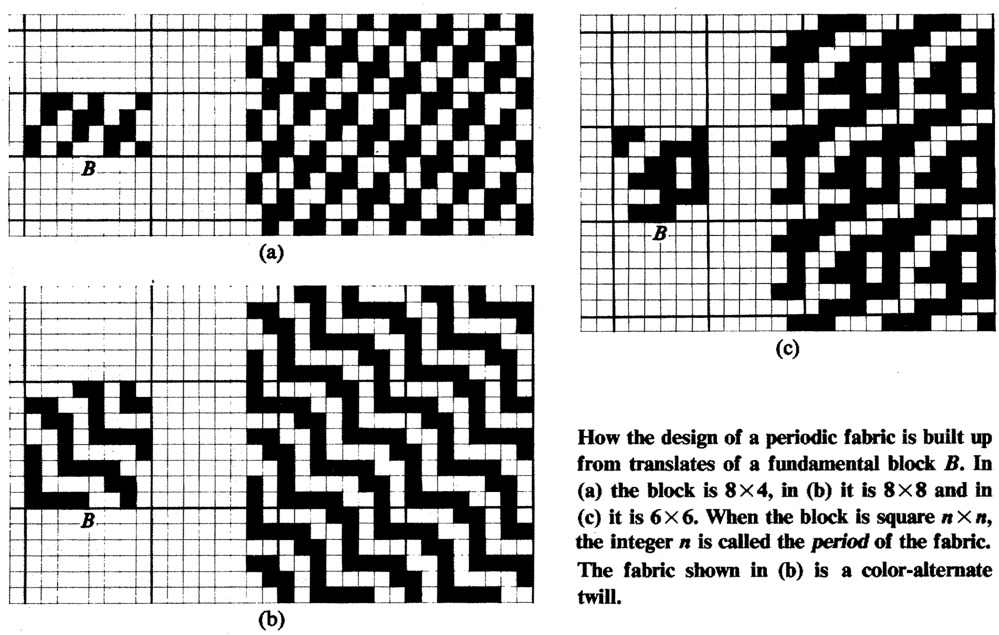
FIGURE 3.

Sometimes it is convenient to use coordinates for the squares in a design. We set up a coordinate system so that the centers of the squares lie on the standard integer lattice and then refer to the square with center (x, y) as the **(x, y)-square**.

By a **symmetry** of a fabric $\mathcal{F}$ in the plane E we mean any isometry that maps each strand of $\mathcal{F}$ into a strand of $\mathcal{F}$ and either preserves all the rankings or reverses them all. Thus it consists of an isometry σ in E (a translation, rotation, reflection, or glide-reflection) possibly followed by a reflection τ in the plane E. The operation τ reverses the ranking of the strands at each point, converting one strand which passes over another into one that passes under it. All the symmetries of a fabric $\mathcal{F}$ form a group under composition called the **symmetry group** of $\mathcal{F}$ and denoted by $S(\mathcal{F})$. The elements of $S(\mathcal{F})$ can be divided into two types: those which do not involve τ and so do not alter the rankings of the strands, and those which involve τ and reverse the rankings. The former type may be said to **preserve the sides** of the fabric and the set of all such forms a normal subgroup $S_0(\mathcal{F})$ of $S(\mathcal{F})$. The latter type may be said to **interchange the sides** of $\mathcal{F}$.

The design D of a fabric $\mathcal{F}$ also has a symmetry group $S(D)$ and each element of $S(D)$ corresponds to a symmetry of $\mathcal{F}$, though not necessarily of the same kind. For example, a translation in $S(D)$ corresponds to a translation in $S(\mathcal{F})$, but a rotation through 90° in $S(D)$ corresponds to a similar rotation of $\mathcal{F}$ combined with the reflection τ. Among others, the designs of satins in FIGURE 12(b) each possess 4-fold rotational symmetries. Each 90° rotation corresponds to a symmetry operation on the fabric which interchanges its sides. This relates to the fact that on one side of the fabric the warp strands are largely visible while on the other the weft strands predominate, and so accounts for the well-known property of satins that their two sides are often dissimilar in appearance. Other symmetries of a fabric $\mathcal{F}$ correspond to isometries which map D onto itself with the colors black and white interchanged. For example, for the design of FIGURE 3(c) there are rotations through 180° which reverse the direction of each row and map the design onto itself if the colors are interchanged. Such operations also correspond to symmetries of $\mathcal{F}$ which interchange its sides.

In almost all the fabrics $\mathcal{F}$ that we shall consider, $S_0(\mathcal{F})$ (and therefore $S(\mathcal{F})$) will contain translations in at least two nonparallel directions. The fabric will then be called **periodic**. The design of a periodic fabric can always be obtained from a **fundamental** $n \times m$ **block** of squares, suitably colored, by translations in horizontal and vertical directions through multiples of n and m units (see FIGURE 4). Although a fundamental block B determines the design of the fabric uniquely, in our diagrams it is usually convenient to show a larger part of the design and



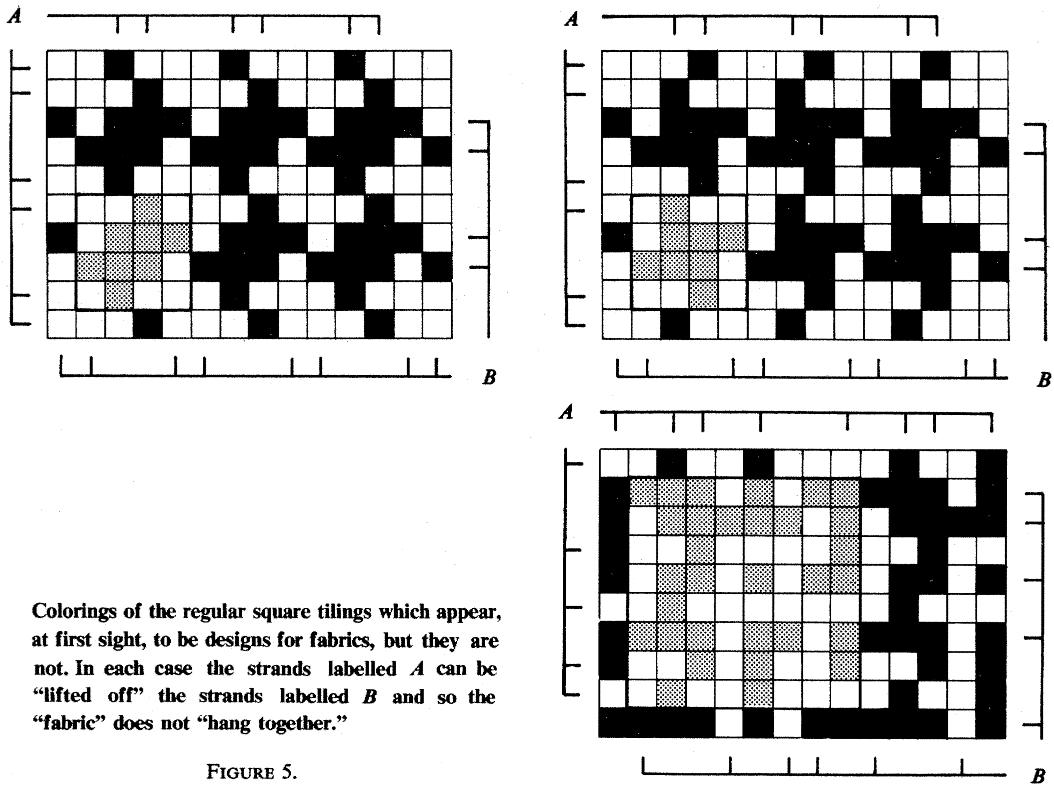
How the design of a periodic fabric is built up from translates of a fundamental block *B*. In (a) the block is 8×4, in (b) it is 8×8 and in (c) it is 6×6. When the block is square *n*×*n*, the integer *n* is called the *period* of the fabric. The fabric shown in (b) is a color-alternate twill.

FIGURE 4.

indicate a fundamental block B by using gray and white squares instead of black and white (see, for example, FIGURES 6(a), (d), (h), (i), (j), (k), (l), (m) and (n)). Occasionally, as in the case of certain sponge weaves (see FIGURES 6(b), (c), (e) and (f)), a fundamental block is too big to show on the diagram, and then it will be assumed that the reader will be able to "see" how the design can be continued from the part of it that is given.

The integers m and n (the sides of a fundamental block) are called the **periods** of the fabric. We shall mostly deal with the case where the fundamental block is square, so $m = n$ and the integer n is called the period. Notice that when we say that a fabric is of period n we do not preclude the possibility that it is also of period d where d is any divisor of n. Thus a plain weave is considered to be of period n where n is any even integer. Other terms in use for n are the **order** of the fabric and the **number of ends**. However, we shall not use these terms here.

Not every black and white coloring of a rectangular block of squares is a fundamental block in the design of a fabric, since the requirement that the fabric must "hang together" may be violated in ways that are not immediately apparent. For example, at first glance, the "designs" of FIGURE 5 seem to correspond to fabrics, but this is not so, for in each case the "fabric" will "fall apart." Intuitively the set of strands labelled A can be "lifted off" those labelled B since at each crossing the A strand passes over the B strand.



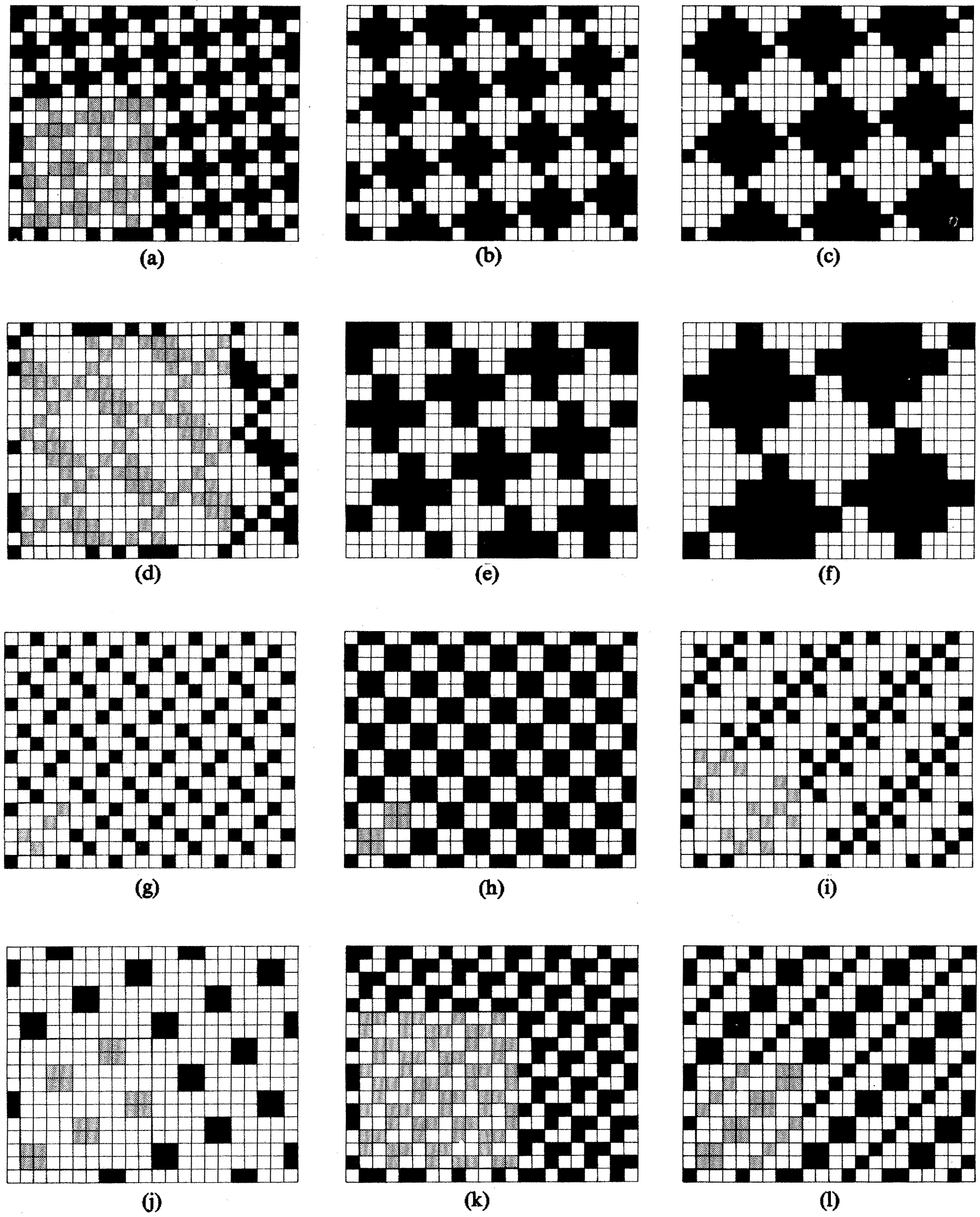
Colorings of the regular square tilings which appear, at first sight, to be designs for fabrics, but they are not. In each case the strands labelled A can be "lifted off" the strands labelled B and so the "fabric" does not "hang together."

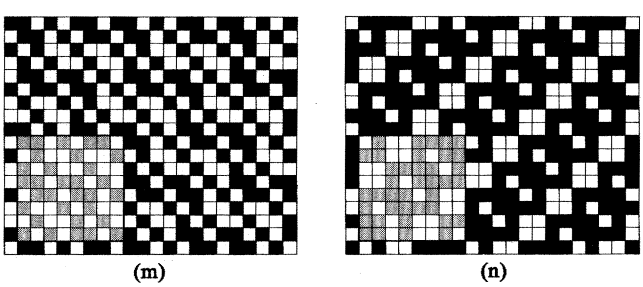
FIGURE 5.

An extremely important class of fabrics, both from a mathematical as well as a practical point of view, will be called "isonemal," a term derived from the Greek words ισος (the same) and νημα (a thread or yarn). A fabric $\mathcal{F}$ is **isonemal** if its symmetry group $S(\mathcal{F})$ is transitive on the strands of $\mathcal{F}$. In other words, for any two strands s_1 and s_2 there exists a symmetry of $\mathcal{F}$ that maps s_1 onto s_2. In terms of the design D of $\mathcal{F}$ this means that any row or column of squares in D can be mapped into any other row or column by either a symmetry of D, or by such a symmetry combined with interchange of the colors black and white. In FIGURE 6 we show examples of isonemal fabrics, and the profusion of possibilities will be apparent. Moreover, as will be seen from the caption, many of these are *actual fabrics* used by the textile industry. The reader to whom these ideas are unfamiliar should convince himself that the fabrics of FIGURE 6 are isonemal, while those of FIGURES 7 and 8 are not.

FIGURE 7 shows that even in fabrics which are not isonemal, it is possible for every strand to weave under and over the other strands in the same "pattern" or "sequence." A fabric with this property is called **mononemal**. The difference between the concepts of isonemality and mononemality can be explained by the observation that mononemality is local—it merely implies that every strand "looks alike"—whereas isonemality is global—the relationship of each strand to the totality of other strands must be the same. Again the reader is urged to verify for himself that the fabrics of FIGURE 7 are mononemal, while those of FIGURE 8 are not. Clearly every isonemal fabric is mononemal.

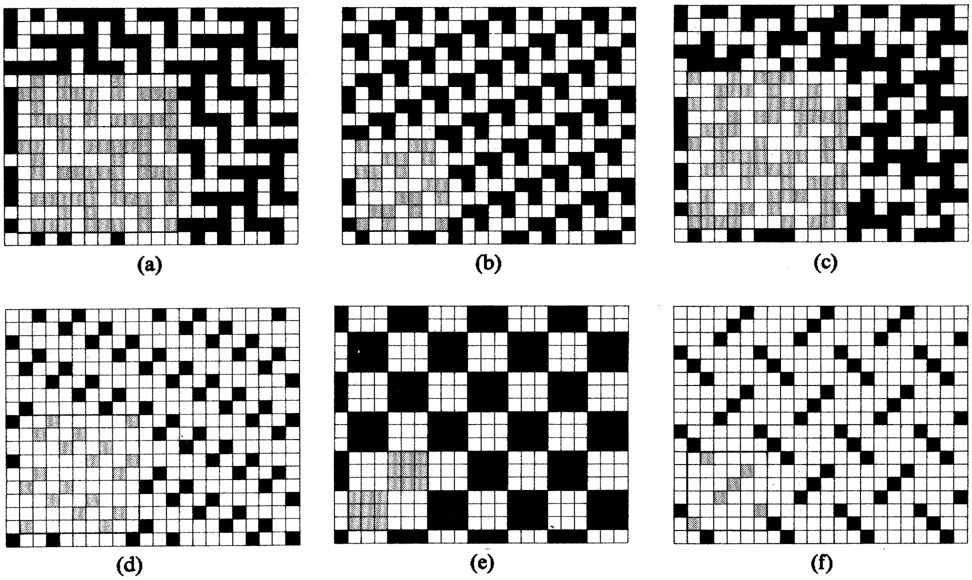
The above ideas lead to a classification of fabrics into three major types: isonemal (I), mononemal but not isonemal (M), and not mononemal (N). For some purposes a finer





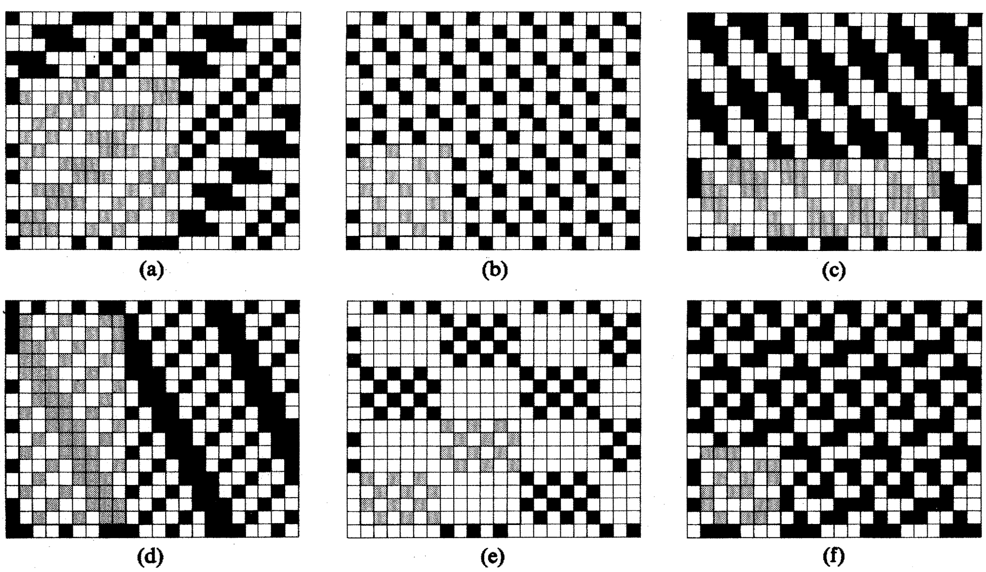
Examples of isonemal fabrics. (a), (b), (c) and (d) are sponge weaves, (e) and (f) are doubled sponge weaves, (g) is satinette, (h) is duck cloth or sailcloth, (j) is a doubled satin, and (l) is a figured twill. (Designs (a), (b), (d), (g), (h) and (l) are from Nisbet [c], FIGURES 232, 233, 240, 167, 21, and 177 respectively.)

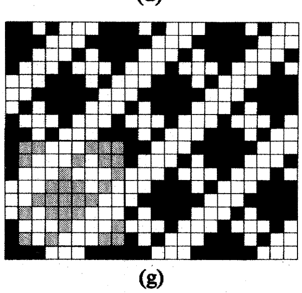
FIGURE 6.



Examples of fabrics which are mononemal but not isonemal. (e) is matt weave and (f) is rice weave. These latter two designs are from Nisbet [c] FIGURES 22 and 168. The first four designs are of fabrics which are both warp-isonemal and weft-isonemal and so are of type M1. The last two are of type M3.

FIGURE 7.





Examples of fabrics which are not mononemal. (a) and (g) are figured twills, (c) is a broken twill, (d) is a combined twill and (f) is a rearranged twill, being obtained from the twill shown in FIGURE 3 by rearranging the warp threads. (Designs (a), (c), (d), (f) and (g) are from Nisbet [c], FIGURES 186, 158, 133, 118 and 181 respectively.) Fabrics (a) and (b) are of type N1, (c) of type N2, (d) of type N3, (e) of type N4, (f) of type N5 and (g) of type N6.

FIGURE 8.

classification is useful and interesting. Referring to FIGURE 9, we see that it is possible for the symmetry group $S(\mathcal{F})$ of the fabric $\mathcal{F}$ to be transitive on the weft strands (each is mapped onto the next one above it by a "step" of 6 squares to the left or right), whereas $S(\mathcal{F})$ is *not* transitive on the warp strands (in fact these form three transitivity classes indicated by the letters X, Y and Z). We express this by saying that $\mathcal{F}$ is **weft-isonemal** but not **warp-isonemal**. On the other hand, the warp strands weave under and over the weft strands in the same pattern or sequence (one over, one under, one over, one under, and so on) and hence, by an obvious extension of the terminology, we may say that it is **warp-mononemal**. Just as an isonemal fabric is mononemal, so a weft-isonemal fabric such as that shown is also weft-mononemal. Hence we see that a fabric can be both warp-mononemal and weft-mononemal without being a mononemal fabric. On the other hand, every mononemal fabric must be both warp-mononemal and weft-mononemal. Similar remarks apply to isonemality; an isonemal fabric is necessarily both warp-isonemal and weft-isonemal, but the converse statement is not generally true.

This terminology enables us to classify fabrics into ten types (I, M1–M3 and N1–N6) as indicated in TABLE 1. The fabrics of FIGURE 6 are all of type I and in FIGURES 7 and 8 we show fabrics belonging to eight of the remaining nine classes. The type that is missing is M2 and we believe, but cannot prove, that no periodic fabrics of this kind exist. More precisely we conjecture that *every periodic mononemal fabric which is warp-isonemal is also weft-isonemal*. The reader may like to try to prove this conjecture; even if he does not succeed he will learn a great deal about the possible structures of different types of fabric.

The ten types of fabrics

Isonemal fabrics

Type I: necessarily warp I and weft I.

Mononemal, but not isonemal, fabrics

Type M1: warp I and weft I.
Type M2: warp I and weft M, *or* warp M and weft I.
Type M3: warp M and weft M.

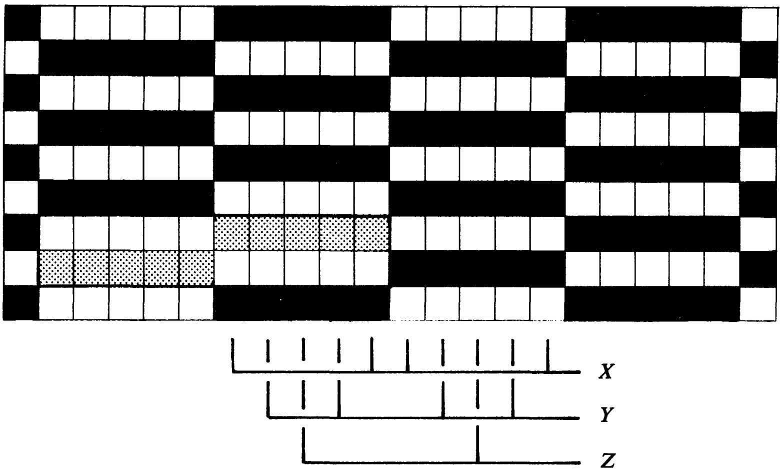
Fabrics which are not mononemal

Type N1: warp I and weft I.
Type N2: warp I and weft M, *or* warp M and weft I.
Type N3: warp I and weft N, *or* warp N and weft I.
Type N4: warp M and weft M.
Type N5: warp M and weft N, *or* weft N and warp M.
Type N6: warp N and weft N.

In the descriptions, I means isonemal, M means mononemal but not isonemal, and N means not mononemal. For example, the fabric whose design is shown in FIGURE 9 is not mononemal, but it is weft-isonemal and warp-mononemal. Hence it is of type N2.

TABLE 1.

Yet another kind of classification arises from considerations of balance; a periodic fabric is called **balanced** if a fundamental block contains equal numbers of black and white squares. Thus if a fundamental block is n by m, then at least one of n or m must be even, and then the block will contain $\frac{1}{2}nm$ white squares and $\frac{1}{2}nm$ black squares. Examples of balanced fabrics are given in FIGURES 2, 3, 5(a),(b), 6(a),(b),(c),(e),(f),(h),(k) and (m). In a balanced fabric equal "amounts" of warp and weft show on each side, a property which is desirable in certain practical applications. For a balanced isonemal fabric it is possible for $S_0(\mathcal{F})$ to be transitive on the strands of $\mathcal{F}$, though not all such fabrics have this property. For example $S_0(\mathcal{F})$ is transitive on the strands for the fabric of FIGURE 6(a) but not in the case of the fabrics of FIGURE 11.



A fabric which is weft-isonemal but not warp-isonemal. It is, however, warp-mononemal and so we see that it is of type N2. (Design from Nisbet [c], FIGURE 16.)

FIGURE 9.

We have already remarked on the great practical and theoretical importance of isonemal fabrics, and the rest of this paper will be devoted to these. The commonest kinds of isonemal fabrics are the twills and satins which will be discussed in the next two sections. After this we shall describe a general method for finding designs of certain kinds of isonemal fabrics of which the twills and satins are special cases. The method will yield many of the fabrics shown in FIGURE 6 (such as the sponge weaves and sailcloth) but not all. At present a completely general method of determining all possible isonemal fabrics of a given period is lacking, and the problem of enumerating such fabrics seems to be completely intractable.

In the above discussion, and also in the rest of this paper, we shall try to adopt terminology which conforms to that in use in the textile industry. Unfortunately this has not always been possible, for not only do authors disagree on the exact meanings of words, but in some cases they formulate their definitions so loosely that we were unable to understand, in a rigorous mathematical sense, exactly what is intended. At the end of the paper we give a list of references concerning the practical aspects of weaving. These are some that we have consulted, but apart from Nisbet [c], which gives a large number of very interesting examples and attempts to be comprehensive, they do not seem to us to be of particular interest or merit; the reader will easily find other works of equal usefulness in any large library or bookshop.

Twills

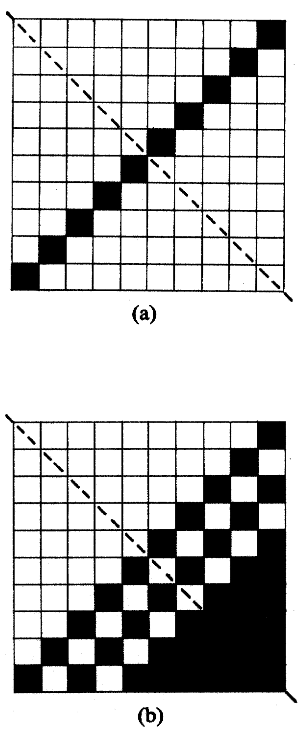
The plain weave of FIGURE 2 may be regarded as the simplest example of the class of fabrics known as twills. These can be easily described by the following scheme. Let $A=(a_i)_{-\infty}^{\infty}$ be any two-way infinite sequence of zeros and ones. A fabric $\mathcal{F}$ is an **A-twill** provided that in its design the (x,y)-square is colored black if $a_{y-x}=1$ and white if $a_{y-x}=0$, or, alternatively, that these relations hold after the design has been turned through 90°. Thus the plain weave is a twill with $A=(\ldots 0\ 1\ 0\ 1\ 0\ 1 \ldots)$ and in FIGURE 10 we show designs of A-twills with $A=(\ldots 0\ 0\ 0\ 1\ 0\ 0\ 0 \ldots)$, $(\ldots 0\ 0\ 0\ 1\ 0\ 1\ 0\ 1\ 1\ 1 \ldots)$, $(\ldots 1\ 0\ 0\ 1\ 0\ 0\ 1\ 0\ 0\ 1 \ldots)$ and $(\ldots 1\ 1\ 1\ 0\ 0\ 1\ 0\ 0\ 1\ 1\ 1\ 0\ 0\ 1\ 0\ 0\ 1\ 1\ 1 \ldots)$. Of these the last two are periodic and the last one is also balanced.

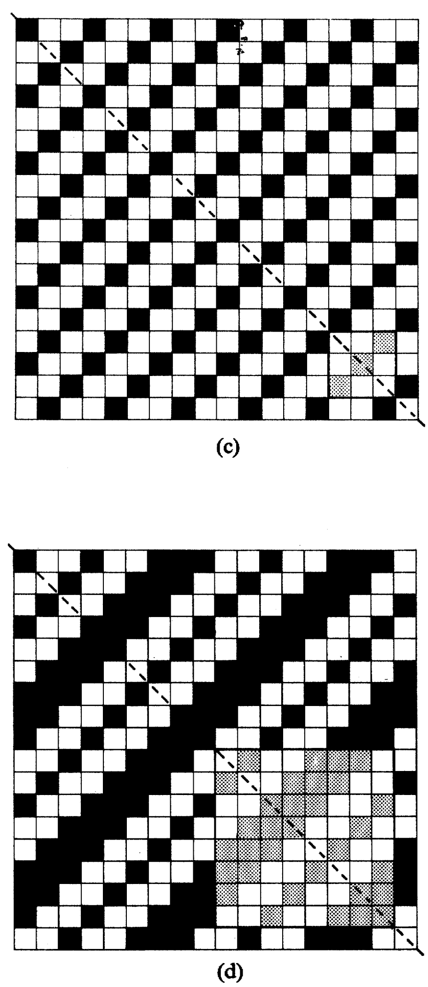
The colorings of any two rows in the design of a twill differ only in that one is shifted sideways relative to the other. If the rows are adjacent, then the upper row is obtained from the lower by a shift to the right through one unit or, as we shall say, a **1-step** to the right. A similar remark applies to the columns. It is this step-like structure that gives a twill its characteristic appearance—it is covered with diagonal stripes. In fact some authors extend the meaning of the word "twill" to include *any* fabric with a pronounced diagonal stripe, and Shorter **[11]** even suggests a numerical measure of "twilliness" which indicates the obviousness of such stripes.

Our first result is very simple.

THEOREM 1. *If A is a sequence of zeros and ones which contains at least two pairs of distinct neighbors, then the A-twill $\mathcal{F}$ is an isonemal fabric. Moreover $\mathcal{F}$ is periodic with period n if and only if A is periodic with period n, that is, if and only if $a_i = a_j$ whenever $i \equiv j$* (mod n).

The proof is rather trivial. The existence of two pairs of distinct neighbors is necessary for it is easily verified that $\mathcal{F}$ does not "hang together" if the sequence A is either constant or one of (...1 1 1 0 0 0...) or (...0 0 0 1 1 1...). These are the only sequences which have fewer than two pairs of distinct neighbors. Translations of the plane, corresponding to the shifts or steps mentioned above, show that $S(\mathcal{F})$ is transitive on the warp strands (warp-isonemal) and also on the weft strands (weft-isonemal). Clearly $S(\mathcal{F})$ also includes rotations by 180° in three dimensions about lines parallel to $x+y=0$ (one of which is shown dotted in FIGURE 10) and these interchange warp and weft strands, showing them to be equivalent. (Notice that this operation interchanges the sides of the fabric.) The proof of the second part of the theorem is apparent





Designs of some A-twills.
(a) A = (...0001000...)
(b) A = (...0001010111...)
(c) A = (...1001001001...)
(d) A = (...1110010011100100111...)
The twills shown in (c) and (d) are periodic with periods 3 and 8 respectively.

FIGURE 10.

from the diagrams, where the fact that $a_i = a_j$ when $i \equiv j \pmod{n}$ shows that both $x \to x+n$ and $y \to y+n$ are symmetries of the fabric which therefore has a fundamental block of size $n \times n$. Hence $\mathcal{F}$ is a periodic fabric with period n.

In practical applications the sequences A are invariably taken to be periodic, and there is a standard notation for these twills used by practical weavers. They are denoted by

$$\frac{c_1 \qquad c_2 \qquad \cdot \qquad \cdot \qquad c_p}{b_1 \qquad b_2 \qquad \cdot \qquad \cdot \qquad b_p}$$

when a period of A consists, in order, of b_1 zeros, c_1 ones, b_2 zeros, c_2 ones $,\ldots, b_p$ zeros, c_p ones (see [c, Chapter 3]). Thus the fabrics of FIGURES 10(c) and 10(d) are $\frac{1}{2}$ and $\frac{3\ 1}{2\ 2}$ twills, respectively. In particular we note that $\sum b_i + \sum c_i = n$, and that $\sum b_i = \sum c_i$ is a necessary and sufficient condition for the twill to be balanced.

It is an interesting combinatorial problem to determine the number $t(n)$ of distinct twills of any given period n. Here two twills are considered identical if one is the image of the other under an isometry. Equivalently, they are identical if their designs can be made to coincide by a rigid motion of the plane possibly followed by an interchange of the colors black and white. In order to give a formula for $t(n)$ we need to use Euler's phi-function $\phi(d)$ which is defined as the number of positive integers less than, and prime to, d (see [6, p. 120]). (Thus, for example, $\phi(12) = 4$ since $m = 1$, 5, 7 and 11 are the only integers satisfying $1 \leqslant m \leqslant 12$ and $\gcd(m, 12) = 1$.) It is well known that

$$\phi(d) = d\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right) \cdots \left(1 - \frac{1}{p_r}\right)$$

where $p_1, p_2, \ldots, p_r$ are the *distinct* prime factors of d. It is also convenient to introduce a function $\rho(n) = \frac{1}{2}(3 + (-1)^n)$ which takes the value 1 if n is odd and 2 if n is even. With this notation we can now state the result.

THEOREM 2. *The number $t(n)$ of distinct twills of period n is given by*

$$t(n) = 2^{\frac{1}{2}(n+\rho(n))-2} + \frac{1}{4n} \sum \phi(d)\rho(d) 2^{n/d} - 1,$$

where the summation in the second term is over all the positive integer divisors d of n.

n	symbol
2	$\frac{1}{1}$
3	$\frac{2}{1}$
4	$\frac{3}{1}$ $\frac{2}{2}$ $\frac{1\ 1}{1\ 1}$
5	$\frac{4}{1}$ $\frac{3}{2}$ $\frac{2\ 1}{1\ 1}$
6	$\frac{5}{1}$ $\frac{4}{2}$ $\frac{3\ 1}{1\ 1}$ $\frac{2\ 2}{1\ 1}$ $\frac{3}{3}$ $\frac{2\ 1}{2\ 1}$ $\frac{1\ 1\ 1}{1\ 1\ 1}$
7	$\frac{6}{1}$ $\frac{5}{2}$ $\frac{4\ 1}{1\ 1}$ $\frac{3\ 2}{1\ 1}$ $\frac{4}{3}$ $\frac{3\ 1}{2\ 1}$ $\frac{2\ 2}{2\ 1}$ $\frac{2\ 1\ 1}{1\ 1\ 1}$
8	$\frac{7}{1}$ $\frac{6}{2}$ $\frac{5\ 1}{1\ 1}$ $\frac{4\ 2}{1\ 1}$ $\frac{3\ 3}{1\ 1}$ $\frac{5}{3}$ $\frac{4\ 1}{2\ 1}$ $\frac{3\ 2}{2\ 1}$ $\frac{3\ 1\ 1}{1\ 1\ 1}$ $\frac{2\ 2\ 1}{1\ 1\ 1}$ $\frac{4}{4}$ $\frac{3\ 1}{3\ 1}$ $\frac{2\ 2}{3\ 1}$ $\frac{2\ 2}{2\ 2}$ $\frac{2\ 1\ 1}{2\ 1\ 1}$ $\frac{1\ 2\ 1}{2\ 1\ 1}$ $\frac{1\ 1\ 1\ 1}{1\ 1\ 1\ 1}$

Symbols for all distinct twills of period n up to $n = 8$.

TABLE 2.

The proof of this theorem involves standard methods, and we only give a brief outline here. The reader familiar with Pólya's Theorem and its applications, see [6, Chapter 5], will appreciate that $t(n)$ can be interpreted as the number of bracelets of n beads, each black or white, subject to the additional proviso that two bracelets are not to be counted as distinct if one can be obtained from the other by interchange of the two colors. The cycle index of the corresponding group of the bracelet is

$$\frac{1}{2n}\left(ns_1 s_2^{(n-1)/2} + \sum \phi(d) s_d^{n/d}\right)$$

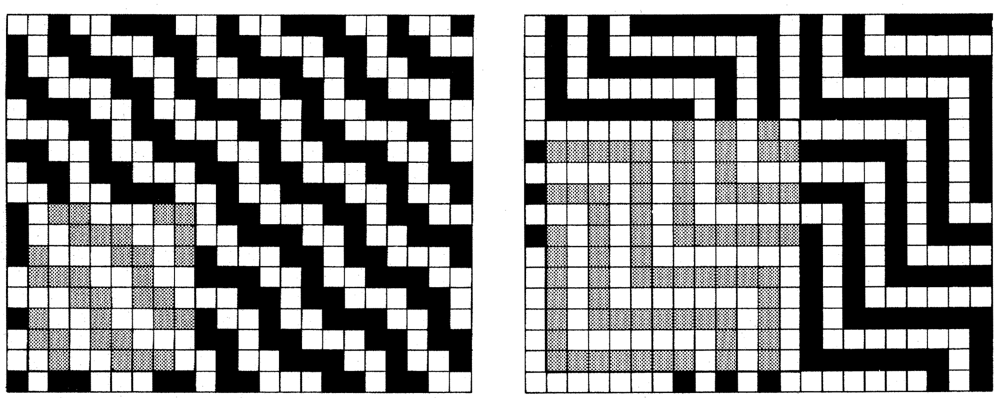
or

$$\frac{1}{2n}\left(\tfrac{1}{2}ns_1^2 s_2^{(n-2)/2} + \tfrac{1}{2}ns_2^{n/2} + \sum \phi(d) s_d^{n/d}\right)$$

according as n is odd or even. From this the result follows by application of a generalization of Pólya's Theorem (see [6, p. 157]). The function $\rho(n)$ is introduced to enable the values for both even and odd n to be written in one compact formula.

Theorem 2 yields, for example, $t(1)=0$, $t(2)=1$, $t(3)=1$, $t(4)=3$, $t(5)=3$, $t(6)=7$, $t(7)=8$ and $t(8)=17$. These figures can be easily verified by actual construction (see TABLE 2).

In FIGURE 4(b) and FIGURE 11 we show three isonemal fabrics which are not twills but are what we shall call **color-alternate twills**. In these, each row is obtained from the one below it by a 1-step to the right and an interchange of colors black and white. In other words, starting from a



(a) (b)

Two color-alternate twills. Each row is obtained from the one below it by a 1-step to the right and an interchange of the colors black and white.

FIGURE 11.

two-way infinite sequence A, the *odd* rows of the fabric are colored in the same way as for an A-twill (as described at the beginning of this section) while the *even* rows are obtained by reversing the colors in the corresponding rows of the A-twill. However, unlike the "ordinary" twills, in a color-alternate twill the sequence A has to be chosen very carefully if the resulting fabric is to be isonemal. We shall explain how this can be done in the fourth section of this paper.

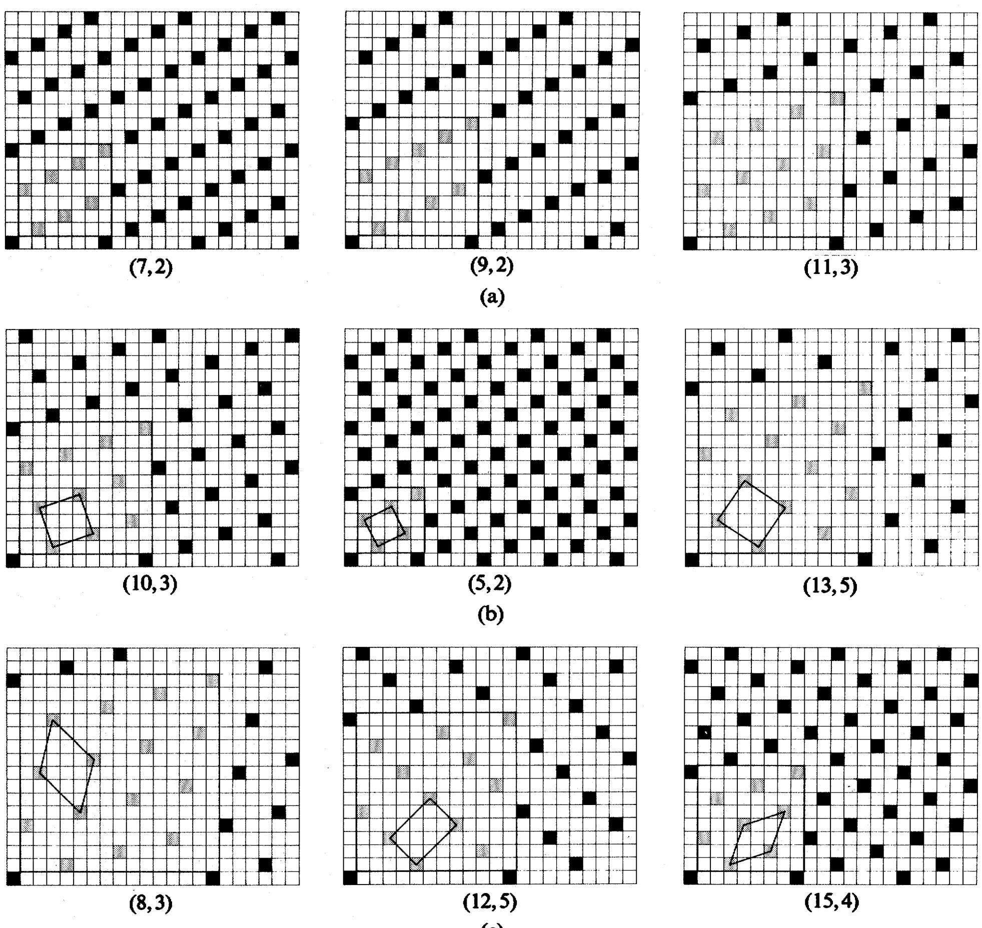
It is strange that color-alternate twills seem to have been rarely, if ever, used in practice, and we can find no record of them in the literature. They have a characteristic and attractive appearance which may be described as a modified herring-bone effect.

Satins

The next type of fabric to be discussed is known as a **satin** or **sateen**. (Authorities differ on the distinction between the meanings of these two words.) An (n,s)-**satin** is a periodic fabric of period n, in which the fundamental $n \times n$ block contains just one black square in each row, and the position of that square is displaced from one row to the next above it by a step of s units to the right (an s-step) (see FIGURE 12). Alternatively, an (n,s)-satin can be defined as one for which the (x,y)-square in the design is colored black if and only if $sy \equiv x \pmod{n}$. Of course, exactly similar considerations will apply if the roles of the colors black and white are interchanged.

We observe that unless s is prime to n the resultant satin does not "hang together," and that there is no loss of generality in assuming that $1 < s < \frac{1}{2}n$. The left inequality arises from the fact that $s = 1$ corresponds to the twills $\frac{1}{n-1}$ discussed in the previous section, and the right inequality because any (n,s)-satin is a mirror-image of an $(n,n-s)$-satin.

An easy counting argument (see Shorter [11]) shows that if an (n,s)-satin is rotated counterclockwise through 90° (interchanging warp and weft) then we obtain either an (n,t)-satin or an $(n,n-t)$-satin, where t is the (unique) solution of the congruence, $st \equiv \pm 1 \pmod{n}$ that satisfies $1 < t < \frac{1}{2}n$. By putting $s = t$ we obtain the following fundamental result.



Examples of (n,s)-satins. The pair (n,s) is indicated near each diagram: (a) mononemal (not isonemal) satins, (b) square isonemal satins, (c) symmetric isonemal satins. Of the three latter that are shown, two are rhombic and one is rectangular.

FIGURE 12.

THEOREM 3. *An* (n,s)*-satin is isonemal if and only if* $s^2 \equiv \pm 1 \pmod{n}$.

The two cases lead to isonemal satins with essentially different properties. If $s^2 \equiv -1 \pmod{n}$, then the satin is said to be **square**, and the symmetry group of the design contains 4-fold rotations but no reflections or glide-reflections (see FIGURE 12(b)). The name "square" comes from the fact that the centers of the black squares form a lattice of which one of the fundamental parallelograms is a square. (In FIGURE 12(b) we have indicated one such square for each of the three fabrics.) These square isonemal satins are characterised by the fact that -1 is a quadratic residue modulo n.

If, on the other hand, $s^2 \equiv +1 \pmod{n}$ then the satin is called **symmetric**. Its symmetry group contains reflections and 2-fold rotations but no 4-fold rotations (see FIGURE 12(c)). The name arises since each design is symmetric in a line parallel to $x=y$. The symmetric satins are characterised by the fact that $+1$ is a quadratic residue modulo n. Symmetric satins can be divided into two classes, **rectangular** and **rhombic** (or **diamond**) satins, according to the possible shapes of the fundamental parallelograms of the lattices of centers of the black squares. (In FIGURE 12(c) it will be seen that the first and third fabrics are rhombic, while the second is rectangular.) Woods [**12**, p. T307] briefly discusses rectangular and rhombic satins which are not necessarily isonemal. It is easy to distinguish between these two classes:

THEOREM 4. *A symmetric isonemal satin is rectangular if n is even and* $s^2 \equiv +1 \pmod{2n}$. *Otherwise it is rhombic.*

We give an outline of the proof of this theorem, leaving the details to be filled in by the reader. As before, we refer to each square in the satin by its coordinates, so the (x,y)-square is colored black if and only if $sy \equiv x \pmod{n}$. Let us consider the family of parallel lines $x+y=$constant that contain black squares. One of these is $x+y=0$, and another is $x+y=s+1$ (because the $(s,1)$-square is black). The line immediately to the right of $x+y=0$ is $x+y=d$, where $d=\gcd(s+1,n)$. The satin will be rectangular if and only if the $(\frac{1}{2}d,\frac{1}{2}d)$-square is black (see FIGURE 12(c)) which implies that d must be even, so n is even and s is odd. But the $(\frac{1}{2}d,\frac{1}{2}d)$-square is black if and only if $\frac{1}{2}ds \equiv \frac{1}{2}d \pmod{n}$, that is, $\frac{1}{2}d(s-1) \equiv 0 \pmod{n}$. Substituting for d we get the equivalent condition $n|\gcd(\frac{1}{2}(s^2-1), \frac{1}{2}n(s-1))$, or $n|\frac{1}{2}(s^2-1)$ (because s is odd and therefore $\frac{1}{2}n(s-1)$ is necessarily a multiple of n). This can be rewritten as the stated condition $s^2 \equiv 1 \pmod{2n}$, so completing the proof of the theorem.

In order to enumerate the isonemal satins for a given n we have to determine the number of solutions of the congruence and inequality $s^2 \equiv \pm 1 \pmod{n}$, $1 < s < \frac{1}{2}n$, and this is easily achieved using known results on quadratic residues (see, for example, Bachmann [**2**, pp. 172, 187, 198]).

THEOREM 5. *For a given n the number of distinct isonemal (n,s)-satins is $u(n)+v(n)$, where $u(n)$ is the number of square satins and $v(n)$ is the number of symmetric satins. If* $n = 2^{\alpha} p_1^{\beta_1} p_2^{\beta_2} \cdots p_j^{\beta_j}$ *is the factorization of n into distinct primes* $2, p_1, p_2, \ldots, p_j$, *then*

$$u(n) = \begin{cases} 0 & \text{if } \alpha \geqslant 2 \text{ or if } p_i \equiv 3 \pmod 4 \\ & \text{for some } i \text{ with } 1 \leqslant i \leqslant j, \\ 2^{j-1} & \text{if } \alpha \leqslant 1 \text{ and if } p_i \equiv 1 \pmod 4 \\ & \text{for } i = 1, 2, \ldots, j, \end{cases}$$

and

$$v(n) = \begin{cases} 2^{j-1}-1 & \text{if } \alpha = 0 \text{ or } \alpha = 1, \\ 2^{j}-1 & \text{if } \alpha = 2, \\ 2^{j+1}-1 & \text{if } \alpha \geqslant 3. \end{cases}$$

Thus, for example, the smallest n for which there exist two distinct isonemal square satins is 65, the satins being (65, 8) and (65, 18). There is also a (rhombic) symmetric (65, 14)-satin, and 65 is the smallest n for which both square and symmetric satins exist. The smallest value of n for which there exists more than one symmetric satin is 24, and the satins are (24, 5), (24, 7) and (24, 11). Two of these are rhombic and one is rectangular, see TABLE 3.

The number of distinct mononemal (but not isonemal) satins of period n can also be found (see Lucas [8]).

n	s	n	s	n	s
5	2a	42	5,11,13c	73	2,3,4,5,6,7,8,10,11,13,14,
7	2	43	2,3,4,5,6,8,9,10,12,15		15,16,17,19,25,27a,31
8	3b	44	3,5,7,13,21c	74	3,5,7,9,11,13,19,23,31a
9	2	45	2,4,7,8,14,19b	75	2,4,7,8,11,13,14,17,26b,29
10	3a	46	3,5,7,11,17	76	3,5,7,9,13,21,23,27,37c
11	2,3	47	2,3,4,5,6,7,9,10,11,13,15	77	2,3,4,5,6,8,9,10,12,15,16,18,20,25,34b
12	5c	48	5,7b,11,17c,23b	78	5,7,17,19,25c,29
13	2,3,5a	49	2,3,4,5,6,9,13,17,18,20	79	2,3,4,5,6,7,8,9,11,12,14,15,18,19,23,
14	3	50	3,7a,9,13,19		27,28,29,32
15	2,4b	51	2,4,5,7,8,11,16b,20	80	3,7,9b,11,13,17,19,31c,39b
16	3,7b	52	3,5,7,9,11,25c	81	2,4,5,7,8,11,13,14,17,26,31,32,35
17	2,3,4a,5	53	2,3,4,5,6,7,8,10,11,12,14,	82	3,5,7,9a,11,13,17,21,23,31
18	5		17,23a	83	2,3,4,5,6,7,8,9,10,11,13,16,17,18,19,
19	2,3,4,7	54	5,7,13,17		20,22,24,27,30
20	3,9c	55	2,3,4,6,7,12,13,16,19,21b	84	5,11,13c,19,25,29c,41c
21	2,4,8b	56	3,5,9,13b,15c,17,27b	85	2,3,4,6,7,8,9,11,13a,16b,18,22,23,24,
22	3,5	57	2,4,5,7,10,11,13,16,20b		26,29,38a
23	2,3,4,5,7	58	3,5,7,9,11,15,17a	86	3,5,7,9,11,13,15,21,25,27
24	5b,7c,11b	59	2,3,4,5,6,7,8,9,11,14,18	87	2,4,5,7,8,10,13,14,16,17,19,23,28b,37
25	2,3,4,7a,9		19,24,25	88	3,5,7,9,13,15,17,19,21b,23c,43b
26	3,5a,7	60	7,11c,13,19c,29c	89	2,3,4,5,6,7,8,9,12,13,14,16,17,20,23,
27	2,4,5,8	61	2,3,4,5,6,7,8,9,11a,13,16,		24,25,27,28,29,34a,36
28	3,5,13c		17,21,22,24	90	7,11,17,19c,23,29
29	2,3,4,5,8,9,12a	62	3,5,7,11,13,15,23	91	2,3,4,5,6,8,9,11,12,16,19,20,22,25,
30	7,11c	63	2,4,5,8b,10,11,13,17,20		27b,31,32,36
31	2,3,4,5,7,11,12	64	3,5,7,11,15,19,23,31b	92	3,5,7,9,11,15,17,19,21,33,45c
32	3,5,7,15b	65	2,3,4,6,7,8a,9,12,14b,17,	93	2,4,5,7,8,10,11,13,14,16,19,22,25,32b,
33	2,4,5,7,10b		18a,19,21		34
34	3,5,9,13a	66	5,7,17,23c,25	94	3,5,7,9,11,13,15,23,33,35,39
35	2,3,4,6b,8,11	67	2,3,4,5,6,7,8,9,10,12,13,	95	2,3,4,6,7,8,9,11,13,14,17,18,23,29,
36	5,11,17c		14,16,18,23,29		31,39b,41,42
37	2,3,4,5,6a,7,8,	68	3,5,7,9,11,13,19,33c	96	5,7,11,13,17b,23,31c,47b
	10,13	69	2,4,5,7,8,11,13,19,20,22b,	97	2,3,4,5,6,7,8,9,10,11,13,17,18,19,20,
38	3,5,7,9		28		21,22a,23,25,26,28,30,33,35
39	2,4,5,7,14b,16	70	3,9,11,13,17,29c	98	3,5,9,13,17,19,25,27,37,41
40	3,7,9c,11b,19b	71	2,3,4,5,6,7,8,11,15,16,17,	99	2,4,5,7,8,10b,13,16,17,19,23,28,29,32,
41	2,3,4,5,6,9a,11,12,		20,21,22,23,26,28		40
	13,16	72	5,7,11,17c,19b,23,35b	100	3,7,9,13,17,19,27,29,39,49c

A list of all the (n,s)-satins with $n < 100$. If the value of s is followed by a, b or c, then the satin is isonemal. The letter a means that the satin is square, b that it is rhombic and c that it is rectangular.

TABLE 3.

THEOREM 6. *The number $w(n)$ of mononemal (but not isonemal) satins of period n is given by* $w(n)=\frac{1}{4}[\phi(n)-2u(n)-2v(n)-2]$, *where ϕ is Euler's phi-function.*

The proof of Theorem 6 depends upon the observation that to each mononemal (but not isonemal) satin of period n correspond exactly four distinct integers less than, and prime to, n, namely s, $n-s$, t and $n-t$ in the notation used above. On the other hand, isonemal satins of either kind correspond to two such integers, and the correction -2 arises from the exclusion of the plain weave.

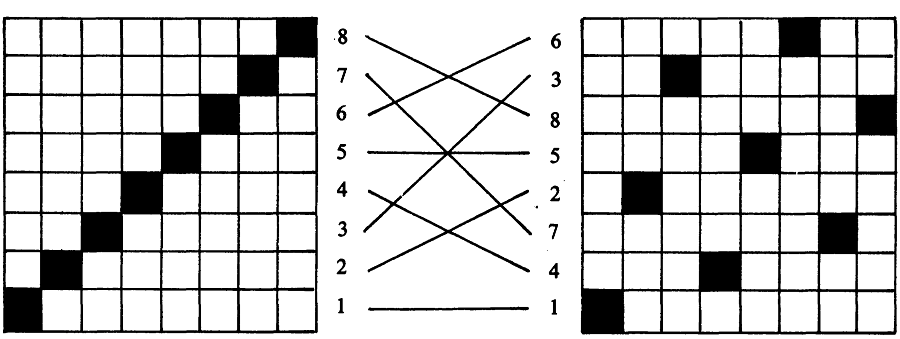
In TABLE 3 we list all possible satins, both mononemal and isonemal, for values of $n \leqslant 100$. This extends the table in Lucas [9], besides giving additional information. (Note that the corresponding table in Lucas [8] contains many errors.) We are indebted to M. G. Shephard for help with the computations needed in the preparation of TABLE 3. Examination of this table reveals an arithmetical curiosity concerning satins. Let $n'>1$ be any divisor of n and s' be the remainder on dividing s by n'. Then if the (n,s)-satin is square, or symmetric, then the (n',s')-satin is also square, or symmetric, respectively (compare Woods [12, p. T306]). Thus, for example, as the $(48,17)$-satin is symmetric, so is the $(12,5)$-satin, and as the $(85,38)$-satin is square, so is the $(17,4)$-satin. The proof of this result is an elementary exercise.

We remark, in conclusion, that a related family of isonemal fabrics can be constructed from the isonemal square satins by the process of "doubling." For this we replace every square in the design by a 2×2 block of squares all colored in the same way. Thus the fabric shown in FIGURE 6(j) is obtained by doubling the $(5,2)$-satin of FIGURE 12(b).

Twillins

Nisbet [c] remarks on the fact that a fundamental block of an (n,s)-satin can be obtained from that of an $\frac{1}{n-1}$ twill by suitably rearranging the rows (weft strands) or columns (warp strands) (see FIGURE 13). He then generalizes this procedure by starting from any twill, and so obtains a class of fabrics called **rearranged twills**. For example, FIGURE 8(f) shows a fabric obtained by rearranging the warp strands of the twill $\frac{1\,2}{2\,1}$ of FIGURE 3. Most rearranged twills are, as in this example, not isonemal (or even mononemal) and this suggests the following interesting combinatorial problem: *How can one determine the twills $\mathcal{T}$ and the rearrangements of the weft (or warp) strands of $\mathcal{T}$, which lead to isonemal fabrics?* Any isonemal fabric which can be constructed in this way will be called a **twillin** being a generalization of both a twill and a satin. The purpose of this section is to explain how *all* twillins of a given period n can be found.

Let us begin by considering which permutations of the weft strands of a twill are admissible. Let B_t be a fundamental $n\times n$ block for the twill $\mathcal{T}$, and let B_f be a fundamental block obtained from B_t by permuting its rows. Without loss of generality we may suppose that the first row of B_f is the first row of B_t, and the second row of B_f is the $(s+1)$st row of B_t $(1<s<n-1)$. Then this



Rearranging the rows (weft strands) of a twill so as to form a satin.

FIGURE 13.

second row is obtained from the first by an s-step to the right (see FIGURE 13 for the case $s=3$, $n=8$). If the resulting fabric is to be weft-isonemal, then it is clearly necessary that every other row of B_f is obtained from the previous row by an s-step to the right with the same value of s. Thus the permutation of the rows may be written

$$\begin{pmatrix} 1 & 1+s & 1+2s & \cdots & 1+(n-1)s \\ 1 & 2 & 3 & \cdots & n \end{pmatrix}$$

where the integers in the top row are reduced modulo n. Note that s must be prime to n since otherwise the numbers in the top row of the above array would not be distinct and so we would not have a proper permutation. We shall use the term (n,s)-twillin for an isonemal fabric of period n constructed by applying s-steps to the rows of B_t in this way.

Now let us consider how the twill $\mathfrak{F}$ can be chosen so that the resulting fabric is isonemal. In order to illustrate the method we shall consider in detail the special case $n=8$, $s=3$. We begin with an 8×8 block of squares and number the squares in the first (lowest) row with the integers $1,2,\ldots,8$. These numbers are repeated in the other rows using 3-steps to the right between adjacent rows. The resulting 8×8 array will be called an $(8,3)$-**number square** (see FIGURE 14). We observe that every row and every column of this square contains all the integers $1,2,\ldots,8$ just once—this is a consequence of the fact that s was chosen prime to n. Our objective is to convert this number square into a fundamental block for a twillin by replacing each integer by a color (black or white).

For a design produced in this way to be mononemal it is necessary and sufficient that the sequence of colors in each column (warp strand) should either be the same as the sequence of colors in each row (weft strand) or should be so after the colors black and white have been interchanged. This can be achieved in several ways. Let us begin, for example, by seeing if it is possible for the first column of the number square (read upwards) to correspond to the first row. We write these thus,

first row:	1	2	3	4	5	6	7	8
first column:	1	6	3	8	5	2	7	4

and note that the coloring will be the same if 2 and 6, and also 4 and 8, represent the same color. To put it another way, if the above scheme is thought of as representing a permutation and this permutation is written in cycle notation (1)(2 6)(3)(4 8)(5)(7), then all integers in the same cycle must represent the same color. Let us label the six cycles, $A,B,\ldots,F$ and make the corresponding substitutions in the number square of FIGURE 14. We obtain the block labelled I in FIGURE 15. This will be called an $(8,3)$-**letter square**. By construction it has a property corresponding to mononemality, namely that if the plane is covered by translates of this block so as to form a tiling $\mathfrak{T}$ of square tiles labelled with the letters $A,B,\ldots,F$, then each row and column of tiles in $\mathfrak{T}$ will contain the same sequence of letters in the same order. In fact, in this case, a much stronger property corresponding to isonemality also holds; the symmetry group of $\mathfrak{T}$ (that is, the group of isometries which map each tile of $\mathfrak{T}$ onto one bearing the same letter) is transitive on the rows and columns of $\mathfrak{T}$. This transitivity property is a consequence of the fact that $3^2\equiv 1 \pmod 8$: It will always occur whenever we are constructing (n,s)-fabrics with $s^2\equiv \pm 1 \pmod n$. The proof of this assertion follows in exactly the same way as Theorem 3.

In the letter square we now substitute the colors black or white for each of the letters $A,B,\ldots,G$. However we do this, subject only to the overriding condition that the resultant fabric must "hang together," we will obtain a fundamental $n\times n$ block for an $(8,3)$-twillin. Hence if we can determine all possible letter squares, it will be possible to obtain the designs of all $(8,3)$-twillins by systematic substitution of colors for letters.

At first sight it appears that there are a great number of possibilities, but many of these can be eliminated immediately. For one thing, we can ignore all twills and satins since these have already been described and enumerated in the previous two sections, and for another, very many designs are repeated. We do not, of course, consider as distinct fundamental blocks that can be

4	5	6	7	8	1	2	3
7	8	1	2	3	4	5	6
2	3	4	5	6	7	8	1
5	6	7	8	1	2	3	4
8	1	2	3	4	5	6	7
3	4	5	6	7	8	1	2
6	7	8	1	2	3	4	5
1	2	3	4	5	6	7	8

The (8,3)-number square.

FIGURE 14.

D	*E*	*B*	*F*	*D*	*A*	*B*	*C*
F	*D*	*A*	*B*	*C*	*D*	*E*	*B*
B	*C*	*D*	*E*	*B*	*F*	*D*	*A*
E	*B*	*F*	*D*	*A*	*B*	*C*	*D*
D	*A*	*B*	*C*	*D*	*E*	*B*	*F*
C	*D*	*E*	*B*	*F*	*D*	*A*	*B*
B	*F*	*D*	*A*	*B*	*C*	*D*	*E*
A	*B*	*C*	*D*	*E*	*B*	*F*	*D*

I

D	*A*	*D*	*E*	*B*	*A*	*B*	*C*
E	*B*	*A*	*B*	*C*	*D*	*A*	*D*
B	*C*	*D*	*A*	*D*	*E*	*B*	*A*
A	*D*	*E*	*B*	*A*	*B*	*C*	*D*
B	*A*	*B*	*C*	*D*	*A*	*D*	*E*
C	*D*	*A*	*D*	*E*	*B*	*A*	*B*
D	*E*	*B*	*A*	*B*	*C*	*D*	*A*
A	*B*	*C*	*D*	*A*	*D*	*E*	*B*

II

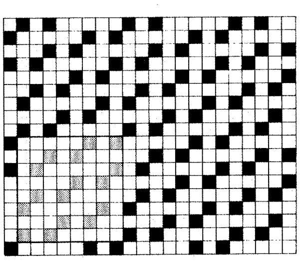
The two (8,3)-letter squares that can be derived from the number square shown in FIGURE 14.

FIGURE 15.

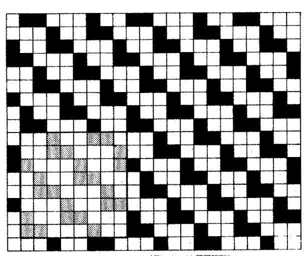
made to coincide by cyclic rearrangement or reversal of the rows or columns, or by interchanging the colors black and white.

From the letter square I of FIGURE 15 substituting colors for letters yields only four twillins, namely those shown in FIGURES 16(a), (b), (c) and (d). Below each diagram we have indicated an allocation of colors by the simple device of separating the letters which represent each of the two colors by a hyphen. By way of example, we note that for the letter square I, *A-BCDEF* is a satin, *B-ACDEF* is a twill, and *AB-CDEF* is identical with *AD-BCEF* and also with *ADEF-BC* since each of these is an isometric image of the others.

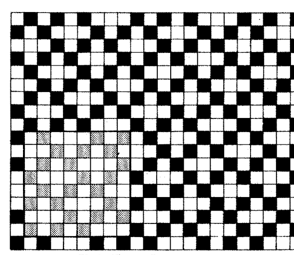
To construct other letter squares, instead of identifying the colors in the first row with those in the first column, we do so after applying a cyclic permutation to the latter. Equivalently we may use any of the other seven columns in the number square. Further possibilities arise if we read the columns *downwards* instead of *upwards*, so eight more cases need to be considered. Some of these lead to letter squares that can only represent twills (for example if we identify 1 2 3 4 5 6 7 8 with 4 7 2 5 8 3 6 1) and we can reject these. Systematic investigation of the possibilities leads to just two letter squares, namely those shown in FIGURE 15. (The second of



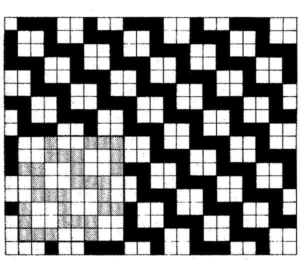
(a) I *AC-BDEF*



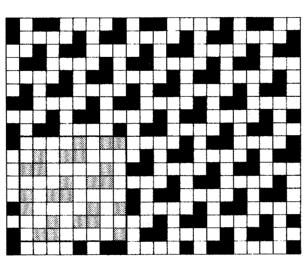
(b) I *AB-CDEF*



(c) I *ACE-BDF*



(d) I *ABC-DEF*



(e) II *BC-ADE*

Five designs for (8,3)-twillins derived from the letter squares of FIGURE 15. In each case we have indicated one possible method of substituting colors for letters.

FIGURE 16.

these comes from identifying the first row 1 2 3 4 5 6 7 8 with the second column 5 8 3 6 1 4 7 2 of the number square read downwards.) These lead in turn to the five fundamental blocks of FIGURE 16; for each we have indicated one possible allocation of colors.

There is still another possibility which we have not yet considered. The fabric can be isonemal if there is an isometry which maps the rows of the design onto the columns with colors reversed. This cannot happen in the case of an (8,3)-twillin, but does occur for (10,3)-twillins (see FIGURE 17). Identifying the first row of the number square with the second column (read upwards), we obtain

$$\begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 2 & 9 & 6 & 3 & 10 & 7 & 4 & 1 & 8 & 5, \end{array}$$

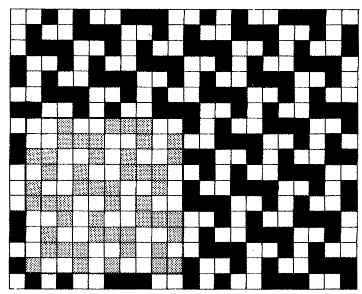
leading to the permutation that can be written as three cycles (1 2 9 8)(3 6 7 4)(5 10). Instead of allocating the same color to all the numbers in a cycle, we do so alternately; this is possible since all the cycles are of even length. Thus we write A for 1 and 9, and A' (the opposite color) for 2 and 8, and so on. This leads to the letter square of FIGURE 17(b), and in FIGURE 17(c) we show an example of the fundamental block of a design obtained from this.

4	5	6	7	8	9	10	1	2	3
7	8	9	10	1	2	3	4	5	6
10	1	2	3	4	5	6	7	8	9
3	4	5	6	7	8	9	10	1	2
6	7	8	9	10	1	2	3	4	5
9	10	1	2	3	4	5	6	7	8
2	3	4	5	6	7	8	9	10	1
5	6	7	8	9	10	1	2	3	4
8	9	10	1	2	3	4	5	6	7
1	2	3	4	5	6	7	8	9	10

(a)

B'	C	B'	B	A'	A	C'	A	A'	B
B	A'	A	C'	A	A'	B	B'	C	B'
C'	A	A'	B	B'	C	B'	B	A'	A
B	B'	C	B'	B	A'	A	C'	A	A'
B'	B	A'	A	C'	A	A'	B	B'	C
A	C'	A	A'	B	B'	C	B'	B	A'
A'	B	B'	C	B'	B	A'	A	C'	A
C	B'	B	A'	A	C'	A	A'	B	B'
A'	A	C'	A	A'	B	B'	C	B'	B
A	A'	B	B'	C	B'	B	A'	A	C'

(b)



(c)

The construction of a (10,3)-twillin for which the symmetry operations which map rows into columns interchange the colors black and white. The fabric in (c) is given by the coloring A-BC. It is also given by the coloring B-AC, while the sponge weave of FIGURE 6(c) is given by the coloring C-AB.

FIGURE 17.

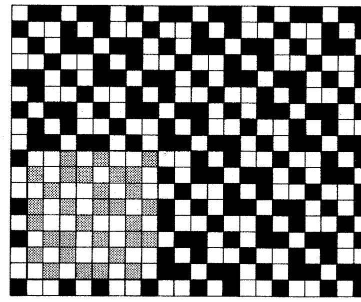
An analogous method to that described above can be used to construct **color-alternate twillins.** (The color-alternate twills mentioned at the end of Section 2 are examples of these with the value $s=1$.) A color-alternate (n,s)-twillin is defined as an isonemal fabric of period n in which each row is obtained from the one below it by an s-step to the right and an interchange of the colors black and white. For example, let us consider the case $n=8$ and $s=3$. Starting from the number square of FIGURE 14 we add primes to the integers in alternate rows (to signify that the colors are interchanged) (see FIGURE 18(a)) and then construct a permutation as before by comparing one of the columns with a row. In the example shown in FIGURE 18 we have identified the first row with the first column read upwards to obtain the permutation (1)(2 6′)(3)(4 8′)(5)(7). Again allocate letters $A, B, \ldots, G$ to these six cycles, and so obtain the letter-square of FIGURE 18(b). As usual, a prime indicates that the colors must be reversed; thus B is substituted for 2 and 6′, which means that B' is substituted for 2′ and 6. From the letter square any allocation of colors to the various letters (subject only to trivial restrictions) will yield an isonemal fabric, that is, a color-alternate twillin. An example of such a fabric is given in FIGURE 18(c) along with the corresponding allocation of colors. As before, there are many possibilities to be explored, though the number of distinct fabrics obtained in this way is not large.

4′	5′	6′	7′	8′	1′	2′	3′
7	8	1	2	3	4	5	6
2′	3′	4′	5′	6′	7′	8′	1′
5	6	7	8	1	2	3	4
8′	1′	2′	3′	4′	5′	6′	7′
3	4	5	6	7	8	1	2
6′	7′	8′	1′	2′	3′	4′	5′
1	2	3	4	5	6	7	8

(a)

D′	*E′*	*B*	*F′*	*D*	*A′*	*B′*	*C′*
F	*D′*	*A*	*B*	*C*	*D*	*E*	*B′*
B′	*C′*	*D′*	*E′*	*B*	*F′*	*D*	*A′*
E	*B′*	*F*	*D′*	*A*	*B*	*C*	*D*
D	*A′*	*B′*	*C′*	*D′*	*E′*	*B*	*F′*
C	*D*	*E*	*B′*	*F*	*D′*	*A*	*B*
B	*F′*	*D*	*A′*	*B′*	*C′*	*D′*	*E′*
A	*B*	*C*	*D*	*E*	*B′*	*F*	*D′*

(b)



(c)

The construction of a color-alternate (8,3)-twillin. The design of (c) is given by the coloring *ABDEF-C*.

FIGURE 18.

The constructions described above for twillins and color-alternate twillins are very simple and lead to many attractive designs for fabrics. For example, the designs of FIGURES 16(c), (d), 17 and 18 seem to be especially pleasing and we find it hard to believe that they have not been used by some practical weaver—yet we can find no mention of them in the literature.

Although the above method enables us to construct all twillins and color-alternate twillins, it does not lead to a solution of the problem of enumerating these fabrics. In fact, since the same twillin can arise in many different ways and there seems to be no way of deciding just how many such ways, the enumeration problem seems completely intractable. And, of course, one must remember that, as we remarked earlier, the twillins and color-alternate twillins compose only a small part of the large class of isonemal fabrics.

New Viewpoints and Questions

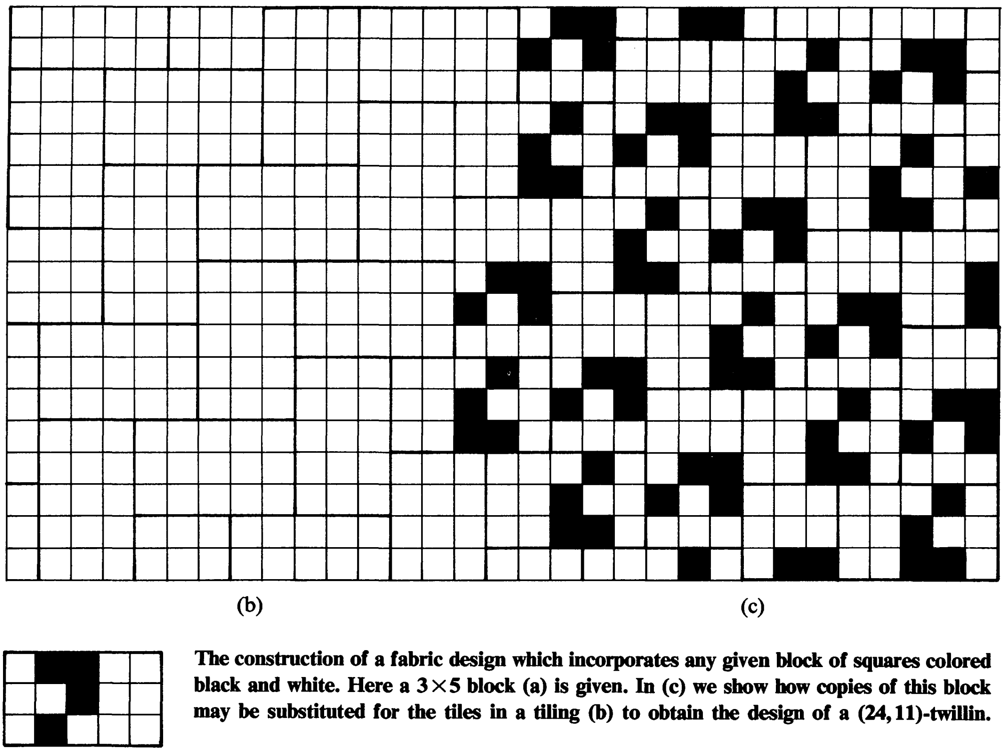
In this section our outlook changes. Instead of constructing isonemal fabrics with a given period, we suppose that we are given *any* block Q of squares colored black and white, and we ask whether this block can form part of the design of an isonemal fabric. The following result, which implies that the answer is in the affirmative, hints at the enormous number of isonemal fabrics of a given period that exist.

THEOREM 7. *Let p and q be relatively prime integers. Then any $p \times q$ rectangular block B of black and white squares is part of the design of a twillin of period $2pq$.*

If a block Q of squares is not of the required shape, we can apply the theorem by determining the smallest $p \times q$ block B (with p and q relatively prime) that contains it. In particular, if Q is square of side k, then we can take $p = k$, $q = k+1$, and the period of the twillin is then $2k(k+1)$.

The constructive proof of Theorem 7 is very simple. It is based on the existence of isohedral tilings $\mathfrak{T}$ by rectangular tiles (see FIGURE 19(b) for tiles of size 3×5). Given any 3×5 block B such as that in FIGURE 19(a), we replace each tile of $\mathfrak{T}$ by a copy of B as shown. It is easily verified that the resulting design is that of a fabric (in that it "hangs together") unless B consists entirely of black or white squares, and that the fabric is a twillin of period $2pq$. The exceptional (monochromatic) blocks are easily dealt with by considering suitable satins.

It is probable that, in general, the period $2pq$ cannot be greatly reduced, since each strand must contain copies of the p rows of B (each of length q) and of the q columns of B (each of length p). However, for small values of p and q, better results may be obtained by *ad hoc* methods. For example, every 2×2 block Q is contained in the design of either a twill $\frac{1}{2}$ or the duck weave (FIGURE 6(h)). These are of periods 3 and 4 respectively, which improves the estimate $2(2^2+2) = 12$ given by the theorem. We do not know the minimum period for 3×3



(b) (c)

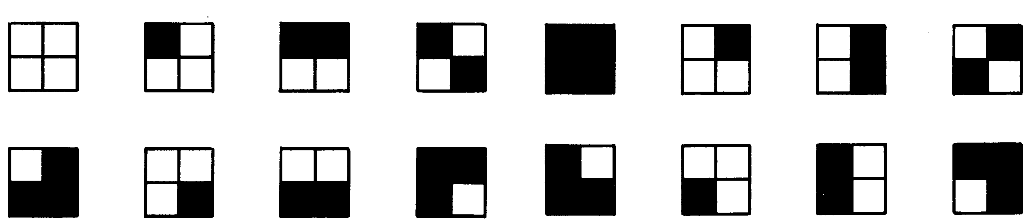
(a)

The construction of a fabric design which incorporates any given block of squares colored black and white. Here a 3×5 block (a) is given. In (c) we show how copies of this block may be substituted for the tiles in a tiling (b) to obtain the design of a (24,11)-twillin.

FIGURE 19.

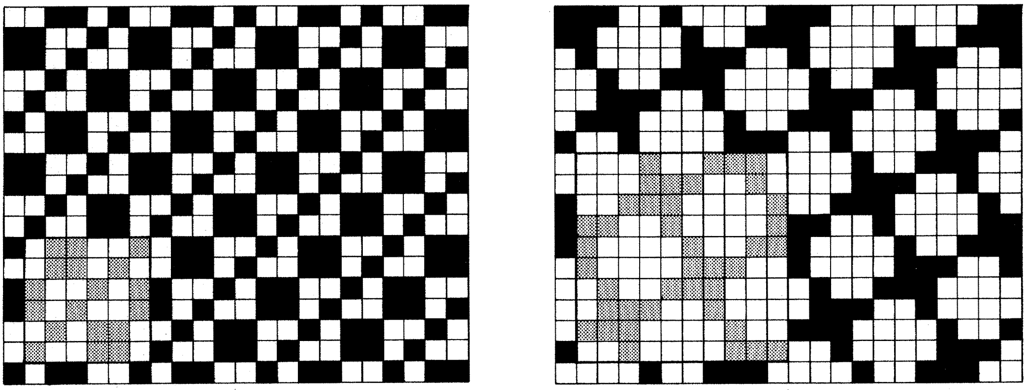
blocks, though we should not be surprised if it can be reduced to a value much less than that $(2(3^2+3)=24)$ given by the theorem.

A related but apparently very difficult problem is to determine for each k the minimal period of an isonemal fabric $\mathcal{F}$ which is k-**universal** in that it contains every $k\times k$ block of squares colored black and white in all possible ways. That universal fabrics exist is an easy consequence of THEOREM 7, for we need only "stack" all possible $k\times k$ blocks together to form a large block Q, and then proceed as before. But the period of the fabric obtained in this way is clearly wildly larger than necessary. There are two possible interpretations of this problem. Of the 16 possible colorings of a 2×2 block (see FIGURE 20) only four are essentially distinct in the sense that all the others can be obtained from these four by a rigid motion or an interchange of colors. Four such essentially different blocks are shown in the top row of the diagram. We can ask either for *all* sixteen blocks to occur in the design of a fabric $\mathcal{F}$ (in which case we shall call $\mathcal{F}$ **strongly**



The sixteen different 2×2 blocks of black and white squares. Only four blocks are essentially different (for example, the first four in the top row); all the other blocks can be obtained from these by a rigid motion or, possibly, by interchange of colors.

FIGURE 20.



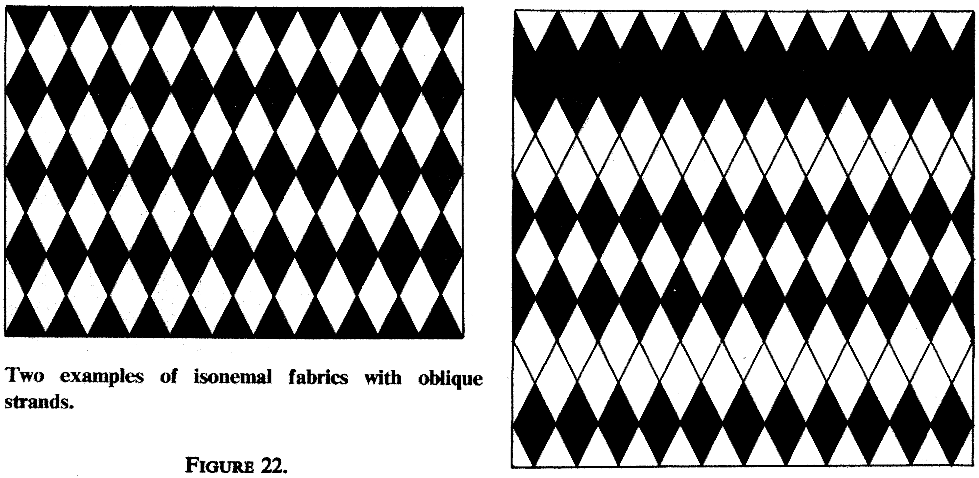
(a) (b)

Designs of two 2-universal fabrics. That shown in (b) is a strongly universal twillin. It is believed that these universal fabrics have the smallest possible periods (6 for 2-universal and 10 for strongly 2-universal).

FIGURE 21.

2-universal) or for only four essentially distinct ones. (An example of a fabric which is 2-universal in this latter sense is shown in FIGURE 6(n).) The theory of pantactic squares (see Astle [1] and Bouwkamp et al. [3]) is clearly relevant to the discovery of strongly universal fabrics, and we remark on the curious fact that the minimum period of a strongly 2-universal fabric *only just* fails to be four! The "design" of FIGURE 5(b) is strongly 2-universal, but unfortunately does not represent a fabric that "hangs together." The strongly 2-universal fabric of least period that we have been able to find has period 10 (see FIGURE 21(b)), while for a 2-universal fabric (not strongly 2-universal) the corresponding period is 6 (see FIGURE 21(a)). For 3-universal fabrics we have no results, or conjectures of any kind, and this remains a completely open field for investigation.

Throughout the whole of this paper we have restricted attention to fabrics in which the warp and weft strands are perpendicular to each other. This is not necessary, and there exist isonemal fabrics with oblique strands (see FIGURE 22). In fact every fabric whose design admits, as a symmetry, reflection in a line parallel to $x+y=0$ or $x-y=0$ remains isonemal if its strands are made oblique. Thus all the twills, the symmetric isonemal satins, and many of the twillins, can be made into "oblique fabrics."



Two examples of isonemal fabrics with oblique strands.

FIGURE 22.

We conclude with some general remarks. It is clear that the material in this paper is only the beginning of a large subject; generalizations in many directions are possible and most of these are completely unexplored. Why is class M2 of mononemal fabrics empty? How many distinct (n,s)-twillins exist for small values of n and s? What are the possible symmetry groups of each of the ten types of fabric? Are there any interesting 2-isonemal fabrics (those in which the strands form two transitivity classes under the operations of the symmetry group) apart from the mononemal satins and those that can be obtained by "doubling" any isonemal fabric?

There is no need to restrict attention to the plane. For example a fabric in the shape of a torus can be constructed from two sets of "annular" strands, or even from just two strands if these are allowed to "spiral" round the torus. Recently Jean J. Pedersen has constructed isonemal fabrics on polyhedral surfaces [**10**], but there still remain many open problems concerning fabrics on manifolds and other surfaces in three dimensional space.

Yet another possibility is to investigate fabrics in which the strands lie in more than two directions. (Practical examples of these occur in basketry.) Some results on such fabrics are already known and will be described in a forthcoming paper by the authors. We can already say that in this case a large number of new isonemal fabrics exists.

We are grateful to the referees for suggesting several improvements to this paper, and to Paul J. Campbell for drawing our attention to the work of Lucas [**7**, **8**, **9**] and Woods [**12**] concerning fabrics. Both these authors mention some earlier literature, mainly concerned with satins, but this seems to be almost inaccessible.

The material in this paper is based on work supported in part by the National Science Foundation Grant No. MCS77-01629 AOI.

References

Literature on Fabrics and Weaving

[a] A. Albers, On Weaving, Wesleyan Univ. Press, Middletown, Conn., 1965.

[b] T. W. Fox and W. A. Hanton, Article on "Weaving" in Encyclopaedia Brittanica, 14th ed., London and New York, 1929.

[c] H. Nisbet, Grammar of Textile Design, 3rd ed., Benn, London, 1927.

[d] G. H. Oelsner, A Handbook of Weaves, Dover, New York.

[e] J. J. Pizzuto and P. L. d'Alessandro, 101 Fabrics, Analyses and Textile Dictionary, Textile Press, New York, 1952.

[f] J. H. Strong, Foundations of Fabric Structure, National Trade Press, London, 1953.

[g] W. Watson, Textile Design and Colour, 6th ed., Longmans, London, 1954.

Mathematical References

[1] B. Astle, Pantactic squares, Math. Gaz., 49 (1965) 144–152.

[2] P. Bachmann, Niedere Zahlentheorie, Part 1, Teubner, Leipzig, 1921.

[3] C. J. Bouwkamp, P. Jansen and A. Koene, Note on pantactic squares, Math. Gaz., 54 (1970) 348–351.

[4] L. Fejes Tóth, Regular Figures, Pergamon, New York, 1964.

[5] Karen E. Huff, Computer weaving: Modern technology confronts an ancient craft, Artist and Computer, edited by Ruth Leavitt, Harmony Books, New York, 1976.

[6] C. L. Liu, Introduction to Combinatorial Mathematics, McGraw-Hill, New York-Toronto-London-Sydney, 1968.

[7] E. Lucas, Application de l'Arithmétique à la Construction de l'Armure des Satins Réguliers, Paris, 1867.

[8] E. Lucas, Principii fondamentali della geometria dei tessuti, L'Ingegneria Civile e le Arti Industriali, 6 (1880) 104–111, 113–115.

[9] E. Lucas, Les principes fondamentaux de la géometrie des tissus, Compte Rendu de l'Association Française pour l'Avancement des Sciences, 40 (1911) 72–88.

[10] Jean J. Pedersen, Regular isonemal fabrics on polyhedral surfaces (to appear).

[11] S. A. Shorter, The mathematical theory of sateen arrangement, Math. Gaz., 10 (1920) 92–97.

[12] H. J. Woods, The geometrical basis of pattern design, Textile Institute of Manchester Journal, 26 (1935) (Sect. 2, Transactions), Part I—Points and line symmetry in simple figures and borders, T197–T210; Part II —Nets and sateens, T293–T308; Part III—Geometrical symmetry in plane patterns, T341–T357.

NOTES

An Application of Geography to Mathematics: History of the Integral of the Secant

V. Frederick Rickey
Bowling Green State University
Bowling Green, OH 43403

Philip M. Tuchinsky
Ford Motor Company
Engineering Computing Center
Dearborn, MI 48121

Every student of the integral calculus has done battle with the formula

$$\int \sec\theta\, d\theta = \ln|\sec\theta + \tan\theta| + c. \tag{1}$$

This formula can be checked by differentiation or "derived" by using the substitution $u = \sec\theta + \tan\theta$, but these ad hoc methods do not make the formula any more understandable. Experience has taught us that this troublesome integral can be motivated by presenting its history. Perhaps our title seems twisted, but the tale to follow will show that this integral should be presented not as an application of mathematics to geography, but rather as an application of geography to mathematics.

The secant integral arose from cartography and navigation, and its evaluation was a central question of mid-seventeenth century mathematics. The first formula, discovered in 1645 before the work of Newton and Leibniz, was

$$\int \sec\theta\, d\theta = \ln\left|\tan\left(\frac{\theta}{2} + \frac{\pi}{4}\right)\right| + c, \tag{2}$$

which is a trigonometric variant of (1). This was discovered, not through any mathematician's cleverness, but by a serendipitous historical accident when mathematicians and cartographers sought to understand the Mercator map projection. To see how this happened, we must first discuss sailing and early maps so that we can explain why Mercator invented his famous map projection.

From the time of Ptolemy (c. 150 A.D.) maps were drawn on rectangular grids with one degree of latitude equal in length to one degree of longitude. When restricted to a small area, like the Mediterranean, they were accurate enough for sailors. But in the age of exploration, the Atlantic presented vast distances and higher latitudes, and so the navigational errors due to using the "plain charts" became apparent.

The magnetic compass was in widespread use after the thirteenth century, so directions were conveniently given by distance and compass bearing. Lines of fixed compass direction were called **rhumb** lines by sailors, and in 1624 Willebrord Snell dubbed them **loxodromes**. To plan a journey one laid a straightedge on a map between origin and destination, then read off the compass bearing to follow. But rhumb lines are spirals on the globe and curves on a plain chart —facts sailors had difficulty understanding. They needed a chart where the loxodromes were represented as straight lines.

It was Gerardus Mercator (1512–1594) who solved this problem by designing a map where the lines of latitude were more widely spaced when located further from the equator. On his famous world map of 1569 ([**1**], p. 46), Mercator wrote:

> In making this representation of the world we had...to spread on a plane the surface of the sphere in such a way that the positions of places shall correspond on all sides with each other both in so far as true direction and distance are concerned and as concerns correct longitudes and latitudes... . With this intention we have had to employ a new proportion and a new arrangement of the meridians with reference to the parallels. ... It is for these reasons that we have progressively increased the degrees of latitude towards each pole in proportion to the lengthening of the parallels with reference to the equator.

Mercator wished to map the sphere onto the plane so that both angles and distances are preserved, but he realized this was impossible. He opted for a conformal map (one which preserves angles) because, as we shall see, it guaranteed that loxodromes would appear on the map as straight lines.

Unfortunately, Mercator did not explain how he "progressively increased" the distances between parallels of latitude. Thomas Harriot (c. 1560–1621) gave a mathematical explanation in the late 1580's, but neither published his results nor influenced later work (see [**6**], [**11**]-[**15**]). In his *Certaine Errors in Navigation*... [**22**] of 1599, Edward Wright (1561–1615) finally gave a mathematical method for constructing an accurate Mercator map. The Mercator map has its meridians of longitude placed vertically and spaced equally. The parallels of latitude are horizontal and unequally spaced. Wright's great achievement was to show that the parallel at latitude θ should be stretched by a factor of $\sec\theta$ when drawn on the map. Let us see why.

FIGURE 1 represents a wedge of the earth, where AB is on the equator, C is the center of the earth, and T is the north pole. The parallel at latitude θ is a circle, with center P, that includes arc MN between the meridians AT and BT. Thus BC and NP are parallel and so angle $PNC=\theta$. The "triangles" ABC and MNP are similar figures, so

$$\frac{AB}{MN}=\frac{BC}{NP}=\frac{NC}{NP}=\sec\theta,$$

or $AB=MN\sec\theta$. Thus when MN is placed on the map it must be stretched horizontally by a factor $\sec\theta$. (This argument is not the one used by Wright [**22**]. His argument is two dimensional and shows that $BC=NP\sec\theta$.)

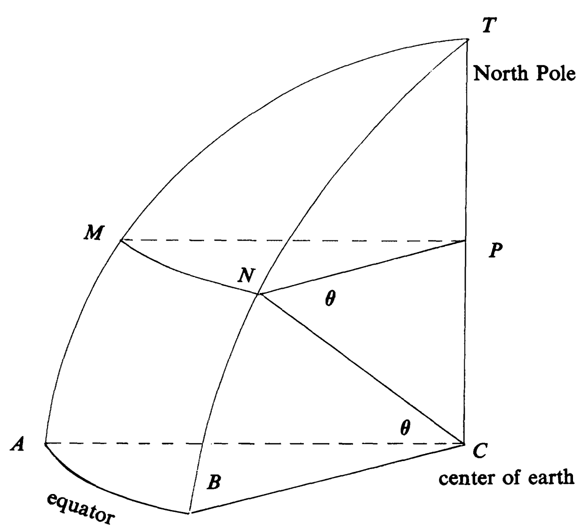


FIGURE 1.

Suppose we can construct a map where angles are preserved, i.e., where the globe-to-map function is conformal. Then a loxodrome, which makes the same angle with each meridian, will appear on this map as a curve which cuts all the map's meridians (a family of parallel straight lines) at the same angle. Since a curve that cuts a family of parallel straight lines at a fixed angle is a straight line, loxodromes on the globe will appear straight on the map. Conversely, if loxodromes are mapped to straight lines, the globe-to-map function must be conformal.

In order for angles to be preserved, the map must be stretched not only horizontally, but also vertically, by $\sec\theta$; this, however, requires an argument by infinitesimals. Let $D(\theta)$ be the distance *on the map* from the equator to the parallel of latitude θ, and let dD be the infinitesimal change in D resulting from an infinitesimal change $d\theta$ in θ. If we stretch vertically by $\sec\theta$, i.e., if

$$dD = \sec\theta\, d\theta$$

then an infinitesimal region on the globe becomes a similar region on the map, and so angles are preserved. Conversely, if the map is to be conformal the vertical multiplier must be $\sec\theta$.

Finally, "by perpetuall addition of the Secantes," to quote Wright, we see that the distance on the map from the equator to the parallel at latitude θ is

$$D(\theta) = \int_0^\theta \sec\theta\, d\theta.$$

Of course Wright did not express himself as we have here. He said ([2], pp. 312–313):

> the parts of the meridian at euery poynt of latitude must needs increase with the same proportion wherewith the Secantes or hypotenusae of the arke, intercepted betweene those pointes of latitude and the aequinoctiall [equator] do increase. ... For...by perpetuall addition of the Secantes answerable to the latitudes of each point or parallel vnto the summe compounded of all former secantes,...we may make a table which shall shew the sections and points of latitude in the meridians of the nautical planisphaere: by which sections, the parallels are to be drawne.

Wright published a table of "meridional parts" which was obtained by taking $d\theta = 1'$ and then computing the Riemann sums for latitudes below 75°. Thus the methods of constructing Mercator's "true chart" became available to cartographers.

Wright also offered an interesting physical model. Consider a cylinder tangent to the earth's equator and imagine the earth to "swal [swell] like a bladder." Then identify points on the earth with the points on the cylinder that they come into contact with. Finally unroll the cylinder; it will be a Mercator map. This model has often been misinterpreted as the cylindrical projection (where a light source at the earth's center projects the unswollen sphere onto its tangent cylinder), but this projection is not conformal.

We have established half of our result, namely that the distance on the map from the equator to the parallel at latitude θ is given by the integral of the secant. It remains to show that it is also given by $\ln|\tan(\frac{\theta}{2} + \frac{\pi}{4})|$.

In 1614 John Napier (1550–1617) published his work on logarithms. Wright's authorized English translation, *A Description of the Admirable Table of Logarithms*, was published in 1616. This contained a table of logarithms of sines, something much needed by astronomers. In 1620 Edmund Gunter (1581–1626) published a table of common logarithms of tangents in his *Canon triangulorum*. In the next twenty years numerous tables of logarithmic tangents were published and so were widely available. (Not even a table of secants was available in Mercator's day.)

In the 1640's Henry Bond (c. 1600–1678), who advertised himself as a "teacher of navigation, survey and other parts of the mathematics," compared Wright's table of meridional parts with a log-tan table and discovered a close agreement. This serendipitous accident led him to conjecture that $D(\theta) = \ln|\tan(\frac{\theta}{2} + \frac{\pi}{4})|$. He published this conjecture in 1645 in Norwood's *Epitome of Navigation*. Mainly through the correspondence of John Collins this conjecture became widely known. In fact, it became one of the outstanding open problems of the mid-seventeenth century, and was attempted by such eminent mathematicians as Collins, N. Mercator (no relation), Wilson, Oughtred and John Wallis. It is interesting to note that young Newton was aware of it in 1665 **[18]**, **[21]**.

The "Learned and Industrious *Nicolaus Mercator*" in the very first volume of the *Philosophical Transactions* of the Royal Society of London was "willing to lay a *Wager* against any one or more persons that have a mind to engage... *Whether the Artificial* [logarithmic] *Tangent-line be the true Meridian-line*, yea or no?" (**[9]**, pp. 217–218). Nicolaus Mercator is not, as the story is often told, wagering that he knows more about logarithms than his contemporaries; rather, he is offering a prize for the solution of an open problem.

The first to prove the conjecture was, to quote Edmund Halley, "the excellent Mr. *James Gregory* in his *Exercitationes Geometricae*, published *Anno* 1668, which he did, not without a long train of Consequences and Complication of Proportions, whereby the evidence of the Demonstration is in a great measure lost, and the Reader wearied before he attain it" (**[7]**, p. 203). Judging by Turnbull's modern elucidation **[19]** of Gregory's proof, one would have to agree with Halley. At any rate, Gregory's proof could not be presented to today's calculus students, and so we omit it here.

Isaac Barrow (1630–1677) in his *Geometrical Lectures* (Lect. XII, App. I) gave the first "intelligible" proof of the result, but it was couched in the geometric idiom of the day. It is especially noteworthy in that it is the earliest use of partial fractions in integration. Thus we reproduce it here in modern garb:

$$\begin{aligned}
\int \sec\theta\, d\theta &= \int \frac{1}{\cos\theta}\, d\theta \\
&= \int \frac{\cos\theta}{\cos^2\theta}\, d\theta \\
&= \int \frac{\cos\theta}{1-\sin^2\theta}\, d\theta \\
&= \int \frac{\cos\theta}{(1-\sin\theta)(1+\sin\theta)}\, d\theta \\
&= \frac{1}{2}\int \frac{\cos\theta}{1-\sin\theta} + \frac{\cos\theta}{1+\sin\theta}\, d\theta \\
&= \frac{1}{2}[-\ln|1-\sin\theta| + \ln|1+\sin\theta|] + c \\
&= \frac{1}{2}\ln\left|\frac{1+\sin\theta}{1-\sin\theta}\right| + c \\
&= \frac{1}{2}\ln\left|\frac{1+\sin\theta}{1-\sin\theta}\cdot\frac{1+\sin\theta}{1+\sin\theta}\right| + c \\
&= \frac{1}{2}\ln\left|\frac{(1+\sin\theta)^2}{1-\sin^2\theta}\right| + c \\
&= \frac{1}{2}\ln\left|\frac{(1+\sin\theta)^2}{(\cos\theta)^2}\right| + c \\
&= \ln\left|\frac{1+\sin\theta}{\cos\theta}\right| + c \\
&= \ln|\sec\theta + \tan\theta| + c.
\end{aligned}$$

We became interested in this topic after noting one line of historical comment in Spivak's excellent *Calculus* (p. 326). As we ferreted out the details and shared them with our students, we found an ideal soapbox for discussing the nature of mathematics, the process of mathematical discovery, and the role that mathematics plays in the world. We found this so useful in the classroom that we have prepared a more detailed version for our students [17].

References

The following works contain interesting information pertaining to this paper. The best concise source of information about the individuals mentioned in this paper is the excellent *Dictionary of Scientific Biography*, edited by C. C. Gillespie.

[1] Anonymous, Gerard Mercator's Map of the World (1569), Supplement no. 2 to Imago Mundi, 1961.

[2] Florian Cajori, On an integration ante-dating the integral calculus, Bibliotheca Mathematica, 3rd series, 14(1915) 312–319.

[3] H. S. Carslaw, The story of Mercator's map. A chapter in the history of mathematics, Math. Gaz., 12(1924) 1–7.

[4] Georgina Dawson, Edward Wright, mathematician and hydrographer, Amer. Neptune, 37(1977) 174–178.

[5] Jacques Delevsky, L'invention de la projection de Mercator et les enseignements de son histoire, Isis, 34(1942) 110–117.

[6] Frank George, Hariot's meridional parts, J. Inst. Navigation, London, 21(1968) 82–83.

[7] E. Halley, An easie demonstration of the analogy of the logarithmick tangents to the meridian line or sum of secants: with various methods for computing the same to the utmost exactness, Philos. Trans., Roy. Soc. London, 19(1695–97) 202–214.

[8] Johannes Keuning, The history of geographical map projections until 1600, Imago Mundi, 12(1955) 1–24.

[9] Nicolaus Mercator, Certain problems touching some points of navigation, Philos. Trans., Roy. Soc. London, 1(1666) 215–218.

[10] E. J. S. Parsons and W. F. Morris, Edward Wright and his work, Imago Mundi, 3(1939) 61–71.

[11] Jon V. Pepper, Hariot's calculation of the meridional parts as logarithmic tangents, Archive for History of Exact Science, 4(1967) 359–413.

[12] ______, A note on Hariot's method of obtaining meridional parts, J. Inst. Navigation, London, 20(1967) 347–349.

[13] ______, The study of Thomas Hariot's manuscripts, II: Hariot's unpublished papers, History of Science, 6 (1967) 17–40.

[14] ______, Hariot's earlier work on mathematical navigation: theory and practice. With an appendix, 'The early development of the Mercator chart,' in Thomas Hariot: Renaissance Scientist, Clarendon Press, Oxford, 1974, John W. Shirley, editor, pp. 54–90.

[15] D. H. Sadler and Eva G. R. Taylor, The doctrine of nauticall triangles compendious. Part I–Thomas Hariot's manuscript (by Taylor). Part II–Calculating the meridional parts (by Sadler), J. Inst. Navigation, London, 6(1953) 131–147.

[16] Eva G. R. Taylor, The Haven-Finding Art, Hollis and Carter, London, 1971.

[17] P. M. Tuchinsky, Mercator's World Map and the Calculus, Modules and Monographs in Undergraduate Mathematics and its Applications (UMAP) Project, Education Development Center, Newton, Mass., 1978.

[18] H. W. Turnbull, editor, The Correspondence of Isaac Newton, Cambridge Univ. Press, 1959–1960, vol. 1, pp. 13–16, and vol. 2, pp. 99–100.

[19] H. W. Turnbull, James Gregory Tercentenary Memorial Volume, G. Bell & Sons, London, 1939, pp. 463–464.

[20] D. W. Waters, The Art of Navigation in England in Elizabethan and Early Stuart Times, Yale Univ. Press, New Haven, 1958.

[21] D. T. Whiteside, editor, The Mathematical Papers of Isaac Newton, vol. 1, Cambridge Univ. Press, 1967, pp. 466–467, 473–475.

[22] Edward Wright, Certainė Errors in Navigation, Arising either of the ordinaire erroneous making or vsing of the sea Chart, Compasse, Crosse staffe, and Tables of declination of the Sunne, and fixed Starres detected and corrected, Valentine Sims, London, 1599. Available on microfilm as part of Early English Books 1475–1640, reels 539 and 1018 (these two copies from 1599 have slightly different title pages). The preface and table of meridional parts have been reproduced as "Origin of meridional parts," International Hydrographic Review, 8(1931) 84–97.

Equally Likely, Exhaustive and Independent Events

MILTON W. LOYER
Montana State University
Bozeman, MT 59717

When introducing the terms (1) disjoint $(A_i \cap A_j = \varnothing)$, (2) equally likely $(P(A_i) = P(A_j))$, (3) exhaustive $P(\cup A_i) = 1)$, and (4) independent $(P(A_i \cap A_j) = P(A_i A_j) = P(A_i)P(A_j))$, most elementary statistics texts (e.g., Huntsberger and Billingsley **[2]**) note that no set of nontrivial events can simultaneously possess properties (1) and (4). Although one can quickly construct a set of n events satisfying properties (1)–(3) (for example, the sample space for tossing a fair n-sided die), the existence of sets of n events possessing properties (2)–(4), depends upon exactly what is meant by the independence of n events.

To see that no nontrivial set of 2 events can simultaneously satisfy properties (2)–(4), consider equally likely events A_1 and A_2 each with probability $p < 1$. (We dismiss the trivial case of $A_1 = A_2 =$ the sample space.) If the events are independent, so $P(A_1 A_2) = p^2$, they cannot be exhaustive, since $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 A_2) = 2p - p^2 < 1$ for $p < 1$. Conversely, if the events are exhaustive, they cannot be independent, since $P(A_1 A_2) = P(A_1) + P(A_2) - P(A_1 \cup A_2) = 2p - 1 \neq p^2$ for $p < 1$.

If $n > 2$, however, it is possible to specify the concept of independence in a way that allows events other than those with probability 1 to possess properties (2)–(4). Feller **[1]**, for example, calls the events $A_1, A_2, \ldots, A_n$ **mutually independent** if for all combinations $1 \leqslant i < j < k < \cdots \leqslant n$ the multiplication rules

$$
\begin{aligned}
&P(A_i A_j) = P(A_i)P(A_j) \\
&P(A_i A_j A_k) = P(A_i)P(A_j)P(A_k) \\
&\cdots \\
&P(A_1 A_2 \cdots A_n) = P(A_1)P(A_2)\cdots P(A_n)
\end{aligned}
$$

apply. If all but the last of these rules apply, the events are called $n-1$ mutually independent.

For n equally likely events, we let $P(A_i) = p$ for $i = 1, 2, \cdots, n$. If property (4) is restricted to mutual independence, no set of nontrivial events can possess properties (2)–(4) since the probability that none of the events occur must by (3) be zero, yet by (4) and (2) it is $(1-p)^n : 0 = P(\varnothing) = \Pi[1 - P(A_i)] = (1-p)^n$. Hence $p = 1$, the trivial case. When n is two, mutual independence is the only independence possible, so no set of two nontrivial events can possess properties (2)–(4).

However, if property (4) is taken to mean pairwise (rather than mutual) independence, sets of $n = 3$ nontrivial events satisfying properties (2) and (3) are indeed possible. Suppose A_1, A_2 and A_3 are equally likely events, each with probability p. The Venn diagram giving all eight combinations of these three events is given in FIGURE 1.

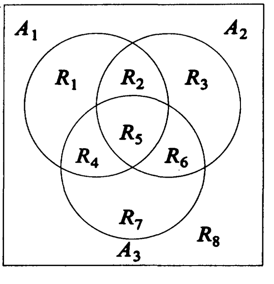


FIGURE 1.

Let t_s denote the probability that exactly s of the events occur in one specific way:

$$t_0 = P(R_8) \qquad\qquad = P(\text{no event occurs})$$
$$t_1 = P(R_1) = P(R_3) = P(R_7) = P(\text{exactly one specific event occurs})$$
$$t_2 = P(R_2) = P(R_4) = P(R_6) = P(\text{exactly two specific events occur})$$
$$t_3 = P(R_5) \qquad\qquad = P(\text{all three events occur}).$$

The pairwise independent and exhaustive properties then require

$$t_2 = P(R_2) = P(R_4) = P(R_6) = p^2 - t_3$$
$$t_1 = P(R_1) = P(R_3) = P(R_7) = p - 2(p^2 - t_3) - t_3$$
$$3P(R_1) + 3P(R_2) + P(R_5) = 3[p - 2(p^2 - t_3) - t_3] + 3[p^2 - t_3] + t_3 = 1,$$

from which it follows that

$$t_0 = 0$$
$$t_1 = p^2 - 2p + 1$$
$$t_2 = -2p^2 + 3p - 1$$
$$t_3 = 3p^2 - 3p + 1.$$

For confirmation, we check properties (2)–(4):

$$P(\cup A_i) = 3t_1 + 3t_2 + t_3 = 1 \quad \text{(exhaustive)}$$
$$P(A_i) = t_1 + 2t_2 + t_3 = p \quad \text{(equally likely)}$$
$$P(A_i A_j) = t_2 + t_3 = p^2 \quad \text{(pairwise independent).}$$

Since $p^2 - t_3 = t_2 \geqslant 0$, it follows that $t_3 \leqslant p^2$. Hence $3p^2 - 3p + 1 \leqslant p^2$ which implies that $1/2 \leqslant p \leqslant 1$. The first and second derivative tests indicate that an absolute minimum of $t_3 = 3p^2 - 3p + 1$ on $[1/2, 1)$ occurs at $p = 1/2$, where $t_3 = 1/4$. So $p \geqslant 1/2$ (or equivalently, $t_3 \geqslant 1/4$) is a necessary condition for the existence of $n = 3$ sets that are equally likely, exhaustive and pairwise (but not mutually) independent. FIGURE 2 shows such sets constructed on [0, 1] for $p = 1/2$ and $p = 2/3$.

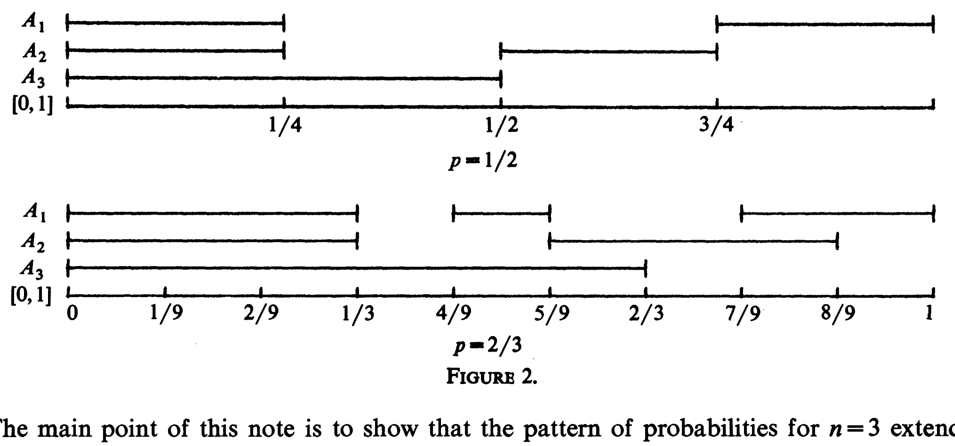


FIGURE 2.

The main point of this note is to show that the pattern of probabilities for $n = 3$ extends to any number of sets.

THEOREM 1. *Equally likely, exhaustive and* $n-1$ *mutually independent events occur if and only if* $t_0 = 0$ *and*

$$t_{n-k} = p^{n-k}(1-p)^k + (-1)^{k+1}(p-1)^n \tag{1}$$

for $k = 0, 1, 2, \ldots, n-1$.

Proof. We imitate the proof for the case $n=3$. The $n-1$ mutual independence requires that

$$t_{n-1}=p^{n-1}-t_n$$
$$t_{n-2}=p^{n-2}-2t_{n-1}-t_n$$
$$\cdots$$

and, in general,

$$t_{n-k}=p^{n-k}-\sum_{j=1}^{k-1}\binom{k}{j}t_{n-j}-t_n \quad \text{for} \quad k=1,2,\ldots,n-1. \tag{2}$$

The exhaustive property requires that

$$P(\cup A_i)=\Sigma P(A_i)-\Sigma P(A_iA_j)+\cdots+(-1)^{n+1}P(A_1A_2\cdots A_n)=1.$$

When coupled with the $n-1$ mutual independence property, this becomes

$$1=\binom{n}{1}p-\binom{n}{2}p^2+\cdots+(-1)^{n+1}t_n. \tag{3}$$

But (3) implies

$$(-1)^n t_n=\sum_{k=0}^{n-1}\binom{n}{k}p^k(-1)^{k-1},$$

which gives $t_n=p^n-(p-1)^n$. This formula, together with (2), yields our desired result (1), since (1), that is, $t_{n-k}=p^{n-k}(1-p)^k+(-1)^{k+1}(p-1)^n$, is the solution of the recurrence relationship (2):

$$p^{n-k}-\sum_{j=1}^{k-1}\binom{k}{j}t_{n-j}-t_n=p^{n-k}-\sum_{j=1}^{k-1}\binom{k}{j}\left[p^{n-j}(1-p)^j+(-1)^{j+1}(p-1)^n\right]-p^n+(p-1)^n$$

$$=p^{n-k}-\sum_{j=0}^{k-1}\binom{k}{j}\left[p^{n-j}(1-p)^j+(-1)^{j+1}(p-1)^n\right]$$

$$=p^{n-k}-p^n\sum_{j=0}^{k}\binom{k}{j}[(1-p)/p]^j+p^{n-k}(1-p)^k+(p-1)^n\sum_{j=0}^{k}\binom{k}{j}(1)^{k-j}(-1)^j+(-1)^{k+1}(p-1)^n$$

$$=p^{n-k}-p^n[(1-p)/p+1]^k+p^{n-k}(1-p)^k+(p-1)^n(1-1)^k+(-1)^{k+1}(p-1)^n$$

$$=p^{n-k}-p^{n-k}+p^{n-k}(1-p)^k+0+(-1)^{k-1}(p-1)^n=p^{n-k}(1-p)^k+(-1)^{k+1}(p-1)^n=t_{n-k}.$$

To verify sufficiency, we calculate each requirement (2)–(4). Exhaustion:

$$P(\cup A_i)=\sum_{j=0}^{n-1}\binom{n}{j}\left[p^{n-j}(1-p)^j+(-1)^{j+1}(p-1)^n\right]$$

$$=\sum_{j=0}^{n-1}\binom{n}{j}p^{n-j}(1-p)^j-(p-1)^n\sum_{j=0}^{n-1}\binom{n}{j}(1)^{n-j}(-1)^j$$

$$=(p+1-p)^n-(1-p)^n-(p-1)^n\left[(1-1)^n-(-1)^n\right]$$

$$=1-(1-p)^n+(1-p)^n=1.$$

Equally likely:

$$P(A_i)=\sum_{j=0}^{n-1}\binom{n-1}{j}\left[p^{n-j}(1-p)^j+(-1)^{j+1}(p-1)^n\right]$$

$$=p\sum_{j=0}^{n-1}\binom{n-1}{j}p^{n-1-j}(1-p)^j+(p-1)^n\sum_{j=0}^{n-1}\binom{n-1}{j}(1)^{n-2-j}(-1)^{j+1}$$

$$=p(p+1-p)^{n-1}+(p-1)^n(1-1)^{n-1}=p.$$

$N-1$ mutual independence:

$$P(A_i \cdots A_j) = \sum_{j=0}^{k} \binom{k}{j} t_{n-j} \qquad \text{for } n-k \text{ events, where } k=1,2,\ldots,n-2$$

$$= \sum_{j=0}^{k} \binom{k}{j} \left[p^{n-j}(1-p)^j + (-1)^{j+1}(p-1)^n \right]$$

$$= p^{n-k} \sum_{j=0}^{k} \binom{k}{j} p^{k-j}(1-p)^j + (p-1)^n \sum_{j=0}^{k} \binom{k}{j} (1)^{k-j-1}(-1)^{j+1}$$

$$= p^{n-k}(p+1-p)^k + (p-1)^n(1-1)^k = p^{n-k}.$$

COROLLARY. *For $n \geqslant 3$ events to be equally likely, exhaustive and $n-1$ mutually independent, it is necessary that*

(i) $p \geqslant 1/2$ for all values of n, and

(ii) $t_n = p^n - (p-1)^n \geqslant (1/2)^{n-1}$ for n odd.

Proof. For n even $t_n \geqslant 0$ gives $p^n-(p-1)^n \geqslant 0$ which implies $p \geqslant 1/2$; for n odd, $t_n \leqslant t_{n-1}$ gives $p^n-(p-1)^n \leqslant p^{n-1}$ which also implies that $p \geqslant 1/2$. This proves (i). The second condition just describes the minimum value of $t_n = p^n-(p-1)^n$ on $[1/2, 1)$: the first and second derivative tests indicate that for odd n an absolute minimum occurs at $p=1/2$, where $t_n=(1/2)^{n-1}$.

FIGURE 3 shows $n=4$ sets constructed on $[0,1]$ for $p=1/2$ and $p=2/3$.

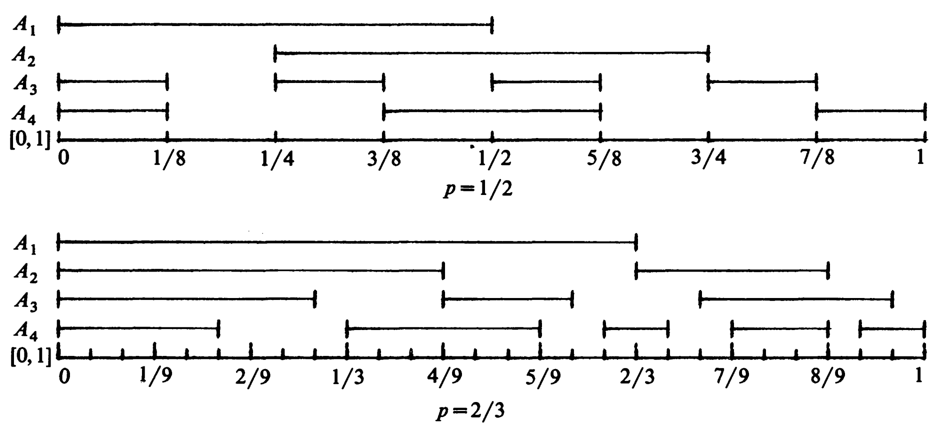


FIGURE 3.

The discussion so far has tacitly assumed continuous spaces. Discrete spaces deserve special consideration. Clearly, one can construct n equally likely, exhaustive and $n-1$ mutually independent sets using $\sum_{k=0}^{n-1}\binom{n}{k} = 2^n-1$, points, $\binom{n}{k}$ points for each $t_{n-k}, k=0,1,\ldots,n-1$, although not all points in such an example will necessarily have the same probability. The following theorem shows that only when $p=1/2$ are there smaller examples.

THEOREM 2. *For $n \geqslant 3$ sets that are equally likely, exhaustive and $n-1$ mutually independent, $t_{n-k}=0$ if and only if $p=1/2$ and $k+n$ is even. Moreover, the minimum number of discrete points that is necessary and sufficient to construct such sets is 2^{n-1}.*

Proof. $0=t_{n-k}=p^{n-k}(1-p)^k+(-1)^{k+1}(p-1)^n=(1-p)[p^{n-k}+(-1)^{n+k+1}(1-p)^{n-k}]$ is true if and only if (i) $p=1$, the trivial case, or (ii) $p^{n-k}=(-1)^{n+k}(1-p)^{n-k}$, which has no solution when $k+n$ is odd and which is true if and only if $p=1/2$ when $k+n$ is even. The minimum number of discrete points, which can be achieved only at $p=1/2$, is given by $\Sigma\binom{n}{k}$, summed only over the terms where $n+k$ is odd. This is

$$\sum_{k=0}^{n-1}\binom{n}{k}(1/2)[1-(-1)^{n+k}]=(1/2)\sum_{k=0}^{n-1}\binom{n}{k}-(1/2)(-1)^n\sum_{k=0}^{n-1}\binom{n}{k}(-1)^k$$
$$=(1/2)\sum_{k=0}^{n}\binom{n}{k}-1/2-(1/2)(-1)^n\sum_{k=0}^{n}\binom{n}{k}(-1)^k+1/2=(1/2)2^n-(1/2)(-1)^n(1-1)^n$$
$$=2^{n-1}.$$

Wong [3] gives a procedure for using 2^{n-1} points to construct n equally likely sets that are $n-1$ mutually independent and for which $p=1/2$ and $t_n=(1/2)^{n-1}$. For n odd his sets are also exhaustive, and illustrate the achieving of the above lower bounds for p, t_n and the number of points in the construction.

References

[1] W. Feller, An Introduction to Probability Theory and Its Applications, vol. 1, 2nd ed., Wiley, New York, 1957.

[2] D. V. Huntsberger and P. Billingsley, Elements of Statistical Inference, 4th ed., Allyn and Bacon, Boston, Mass., 1977.

[3] C. K. Wong, A note on mutually independent events, Amer. Statistician, 26:2 (1972) 27–28.

Cauchy and Economic Cobwebs

Elwyn H. Davis

Pittsburg State University

Pittsburg, KS 66762

A well-known theorem of mathematical economics known as the Cobweb Theorem relates price and quantity of production when there is a time lag between obtaining price information to determine production level and actual marketing of a product. This type of lag occurs in agricultural markets where, for example, one year's wheat price may influence farmers in deciding how much wheat to plant for next year. It turns out that the proof of this theorem is an interesting and easily understood application of Cauchy's Mean Value Theorem.

We can model the Cobweb situation with the aid of two functions, s and d. The supply function s gives the quantity q producers are willing to produce for price p. The demand function d relates the price consumers are willing to pay to the quantity on the market. Figure 1 gives typical supply and demand curves. Their point of intersection E is called the equilibrium point; its coordinates will be denoted by (p_E, q_E).

The Cobweb situation, pictured in Figure 1, begins with some initial price p_0 for a commodity. The supply function then gives the amount q_1 of that commodity that producers will be willing to produce for the next period. Since d typically has an inverse (as it does in Figure 1), $p_1=d^{-1}(q_1)$ is the price investors will be willing to pay for the amount q_1. This price then determines a new quantity q_2 from $q_2=s(p_1)$. Then $p_2=d^{-1}(q_2)$, etc. This iteration yields two sequences $\{p_n\}$ and $\{q_n\}$ defined by the equations $q_n=s(p_{n-1})$ and $p_n=d^{-1}(q_n)$. The picture (see Figure 1) formed by the lines involved in tracing the numbers $p_0, q_1, p_1, q_2, \ldots$ explains the name "cobweb."

number of discrete points, which can be achieved only at $p=1/2$, is given by $\Sigma\binom{n}{k}$, summed only over the terms where $n+k$ is odd. This is

$$\sum_{k=0}^{n-1}\binom{n}{k}(1/2)[1-(-1)^{n+k}]=(1/2)\sum_{k=0}^{n-1}\binom{n}{k}-(1/2)(-1)^n\sum_{k=0}^{n-1}\binom{n}{k}(-1)^k$$
$$=(1/2)\sum_{k=0}^{n}\binom{n}{k}-1/2-(1/2)(-1)^n\sum_{k=0}^{n}\binom{n}{k}(-1)^k+1/2=(1/2)2^n-(1/2)(-1)^n(1-1)^n$$
$$=2^{n-1}.$$

Wong [3] gives a procedure for using 2^{n-1} points to construct n equally likely sets that are $n-1$ mutually independent and for which $p=1/2$ and $t_n=(1/2)^{n-1}$. For n odd his sets are also exhaustive, and illustrate the achieving of the above lower bounds for p, t_n and the number of points in the construction.

References

[1] W. Feller, An Introduction to Probability Theory and Its Applications, vol. 1, 2nd ed., Wiley, New York, 1957.

[2] D. V. Huntsberger and P. Billingsley, Elements of Statistical Inference, 4th ed., Allyn and Bacon, Boston, Mass., 1977.

[3] C. K. Wong, A note on mutually independent events, Amer. Statistician, 26:2 (1972) 27–28.

Cauchy and Economic Cobwebs

Elwyn H. Davis

Pittsburg State University

Pittsburg, KS 66762

A well-known theorem of mathematical economics known as the Cobweb Theorem relates price and quantity of production when there is a time lag between obtaining price information to determine production level and actual marketing of a product. This type of lag occurs in agricultural markets where, for example, one year's wheat price may influence farmers in deciding how much wheat to plant for next year. It turns out that the proof of this theorem is an interesting and easily understood application of Cauchy's Mean Value Theorem.

We can model the Cobweb situation with the aid of two functions, s and d. The supply function s gives the quantity q producers are willing to produce for price p. The demand function d relates the price consumers are willing to pay to the quantity on the market. Figure 1 gives typical supply and demand curves. Their point of intersection E is called the equilibrium point; its coordinates will be denoted by (p_E, q_E).

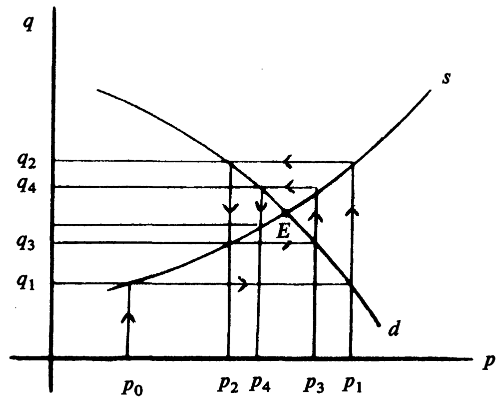
The Cobweb situation, pictured in Figure 1, begins with some initial price p_0 for a commodity. The supply function then gives the amount q_1 of that commodity that producers will be willing to produce for the next period. Since d typically has an inverse (as it does in Figure 1), $p_1=d^{-1}(q_1)$ is the price investors will be willing to pay for the amount q_1. This price then determines a new quantity q_2 from $q_2=s(p_1)$. Then $p_2=d^{-1}(q_2)$, etc. This iteration yields two sequences $\{p_n\}$ and $\{q_n\}$ defined by the equations $q_n=s(p_{n-1})$ and $p_n=d^{-1}(q_n)$. The picture (see Figure 1) formed by the lines involved in tracing the numbers $p_0, q_1, p_1, q_2, \ldots$ explains the name "cobweb."

Typically, supply curves have positive slopes, as higher prices induce higher production. Demand curves usually have negative slopes as higher prices usually result in less demand for a product. The purpose of the Cobweb Theorem is to give conditions which are sufficient to guarantee that the p_n converge to p_E and that the q_n converge to q_E. It turns out that the essential property of the supply and demand curves is that demand be more sensitive to price than supply. In terms of the functions $s(p)$ and $d(p)$, this can be stated as $|s'(p)|<|d'(p)|$. Note that it does not matter if the slopes are positive or negative; demand might even increase with price, and equilibrium will still occur if demand increases fast enough.

The key to the proof is the Cauchy Mean Value Theorem, which states that if f and g are differentiable on the closed interval $[a,b]$, and $g'(x)\neq 0$ for all x in $[x,b]$ then

$$\frac{f(b)-f(a)}{g(b)-g(a)}=\frac{f'(c)}{g'(c)}$$

for some c between a and b. This will enable us to prove the entirely reasonable observation that since $|s'(p)|<|d'(p)|$, p_n is closer to p_E than is p_{n-1}: since d is steeper than s, d rises or falls a given distance more rapidly than does s (see FIGURE 1).



Economic Cobweb

FIGURE 1.

COBWEB THEOREM. *Let s and d be real-valued functions of the real variable* $p>0$, *and suppose that the graphs of s and d intersect at the point* (p_E, q_E) *where* $q_E>0$. *Let I be a closed interval centered at* p_E *on which s and d have nonvanishing continuous derivatives. Define sequences* $\{p_n\}$ *and* $\{q_n\}$ *by letting* p_0 *be any element of I,* $q_n=s(p_{n-1})$ *and* $p_n=d^{-1}(q_n)$ *for* $n\geq 1$. *If* $|s'(p)|<|d'(p)|$ *for all p in I, then* $\lim_{n\to\infty}p_n=p_E$ *and* $\lim_{n\to\infty}q_n=q_E$.

Proof. Since $s'(p)$ and $d'(p)$ are nonzero on I, then each function s and d has an inverse and we have $(s^{-1})'(q)=1/s'(p)$ and $(d^{-1})'(q)=1/d'(p)$. By our hypothesis that $|s'(p)|<|d'(p)|$, we have $|(d^{-1})'(q)|<|(s^{-1})'(q)|$. So by Cauchy's Mean Value Theorem, we have

$$\frac{|p_E-p_1|}{|p_E-p_0|}=\frac{|d^{-1}(q_E)-d^{-1}(q_1)|}{|s^{-1}(q_E)-s^{-1}(q_1)|}=\frac{|(d^{-1})'(z)|}{|(s^{-1})'(z)|}$$

for some z between q_E and q_1. Thus, $|p_E-p_1|<|p_E-p_0|$. This implies that p_1 is in I. A similar argument shows that $|p_E-p_2|<|p_E-p_1|$, so p_2 is in I. Hence, by induction, $|p_E-p_n|<|p_E-p_{n-1}|$, so each p_n is in I. This means that all of the p_n lie in the same closed interval I.

Since $d'(p)\neq 0$ on I and s' and d' are continuous on I, the function f defined by $f(p)=|s'(p)|/|d'(p)|$ is continuous on I. Thus at some point u in I, f has a maximum value M which by hypothesis is less than 1. In other words, $f(u)=|s'(u)|/|d'(u)|=M<1$. Again applying Cauchy's Mean Value Theorem, we have

$$\frac{|q_E - q_n|}{|q_E - q_{n-1}|} = \frac{|s(p_E) - s(p_{n-1})|}{|d(p_E) - d(p_{n-1})|} = \frac{|s'(c)|}{|d'(c)|}$$

for some c between p_E and p_{n-1}, and so

$$|q_E - q_n| = \frac{|s'(c)|}{|d'(c)|} \cdot |q_E - q_{n-1}| \leq M|q_E - q_{n-1}|.$$

Thus, for each $n \geq 2$, $|q_E - q_n| \leq M^{n-1}|q_E - q_1|$. But since $M < 1$, $\lim_{n\to\infty}|q_E - q_n| = \lim_{n\to\infty} M^{n-1}|q_E - q_1| = 0$. Hence, $\lim_{n\to\infty} q_n = q_E$. And since d^{-1} is continuous on the image of I, $\lim_{n\to\infty} p_n = p_E$.

The situation portrayed by the Cobweb Theorem, which has price converging to an equilibrium price, is described by economists by the phrase, "stable equilibrium." They mean by this that if a small extraneous disturbance occurs in the market, eventually price will again converge to p_E. So not only is $\lim_{n\to\infty} p_n = p_E$, but a large disturbance must occur to keep price from approaching p_E. The disturbance would have to be large enough to remove price from the interval I on which the hypotheses of the theorem hold. Such a disturbance might be a depression, drought, or large recession.

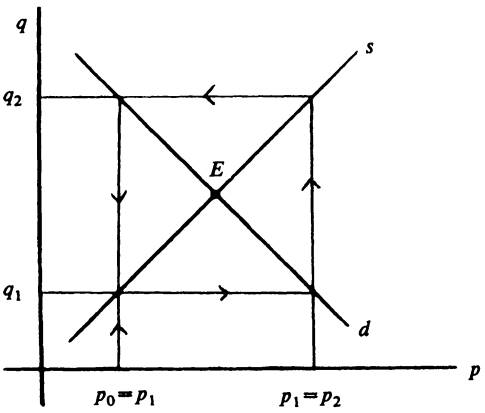


FIGURE 2.

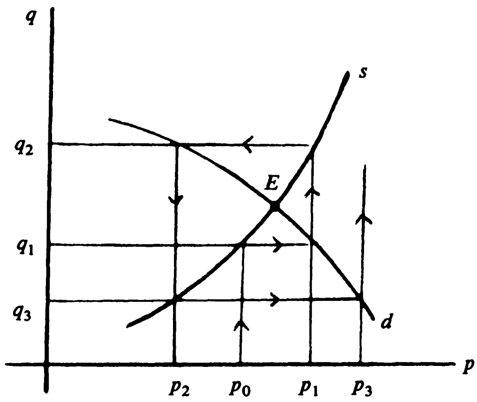


FIGURE 3.

There are two cases of unstable equilibrium which are easy to describe. If $|s'(p)| = |d'(p)|$ for all p, the sequence $\{p_n\}$ alternates between the values p_1 and p_2, and $\{q_n\}$ alternates between q_1 and q_2 (see FIGURE 2); whereas if $|s'(p)| > |d'(p)|$, the sequences $\{p_n\}$ and $\{q_n\}$ diverge (see FIGURE 3). These are examples of boom and bust cycles, the first somewhat controlled and the last disastrously out of control. If price is moved away from equilibrium in either of these two cases, it will never return.

References

[1] William J. Baumol, Economic Dynamics, 3rd ed., Macmillan, 1970, pp. 111–115.

[2] E. Bender, An Introduction to Mathematical Modeling, Wiley, 1978, pp. 57–59.

[3] N. S. Buchanan, A reconsideration of the Cobweb Theorem, J. Polit. Econom., 47 pp. 67–81. Reprinted in R. V. Clemence, ed., Readings in Economic Analysis, vol. 1, Addison-Wesley, 1950, pp. 46–60.

[4] Alpha Chiang, Fundamental Methods of Mathematical Economics, 2nd ed., McGraw-Hill, 1974, pp. 562–566.

[5] A. A. Harlow, The Hog Cycle and the Cobweb Theorem, J. Farm Econom., 42 (Nov. 1960) 842–853.

[6] James M. Henderson and Richard E. Quandt, Microeconomic Theory: A Mathematical Approach, 2nd ed., McGraw-Hill, 1971, pp. 142–149.

[7] H. H. Liebhafsky, The Nature of Price Theory, The Dorsey Press, Homewood, Ill., 1968, pp. 327–329.

[8] Walter Nicholson, Microeconomic Theory, The Dryden Press, Hinsdale, Ill., 1978, pp. 317–321.

The Rotating Table

TED LEWIS
STEPHEN WILLARD
University of Alberta
Edmonton, Alberta, Canada T6G 2G1

Imagine a square table at each corner of which is a deep well. Hidden from view at the bottom of each well is a drinking glass which may be either upright or inverted. You ("the player") may place one or both hands into any wells desired and may adjust the glasses found there in any way you wish. After you have done this, if all glasses around the table are in the same state, a bell rings and you win. Otherwise, the table is rotated (with you blindfolded) and you are allowed again to select wells and adjust glasses. The question: can you force the bell to ring in a finite number of moves?

This problem can be generalized in an obvious way: Assume a polygonal table with n wells and a player (the "bell ringing octupus") with $n_0<n$ hands. We will show here that, in this general setting, a player can force the objective of ringing the bell in a finite number of moves if and only if he has been provided with at least $[(p-1)/p]n$ hands, where p is the largest prime divisor of n.

The problem for the table with 4 wells has been discussed recently by Gardner (**[2]**, **[3]**). In **[3]** he reports that R. L. Graham and P. Diaconis have shown that the n-well game can be won by a player with $n-2$ hands if and only if n is nonprime. It is also suggested that, for composite n, fewer than $n-2$ hands may be required. That this is indeed the case follows from the result we have stated above.

In what follows, we will establish the necessity of $[(p-1)/p]n$ hands for the n-well table, and then the sufficiency of $[(p-1)/p]n$ hands. We will also remark on the total number of moves required to win the game (expanded to encompass a discussion of the "total energy" required to win), and on a further generalization of the game to one in which the glasses are replaced by objects, each of which can assume any of h different states.

It will be convenient to think of the n-well game as being played on a regular n-gon ("vertex" = "well") between two contestants: the **player**, who has been provided with $n_0<n$ hands, and **fate**, who initially sets the glasses in the wells and who at all times knows the state of every glass. A move of the game proceeds as follows: The player places his hands over the wells

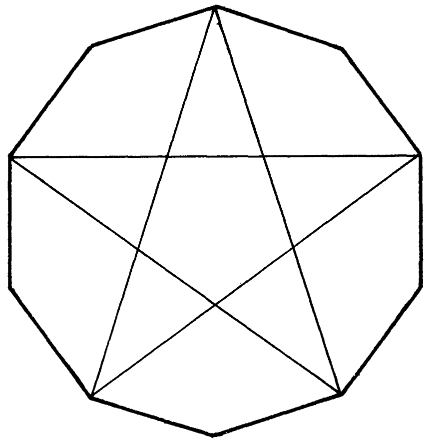


FIGURE 1(a)

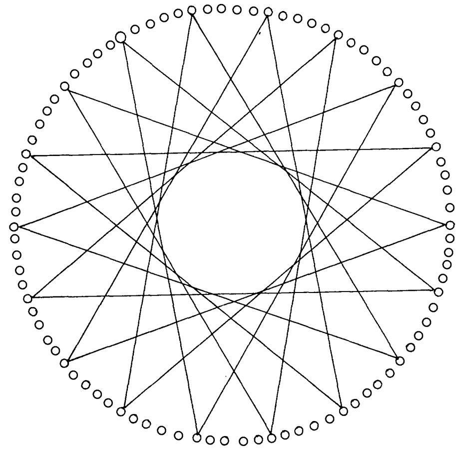


FIGURE 1(b)

in the pattern he intends to use; fate then rotates the table to any position fate wishes, and finally the player drops his hands into the wells provided by fate and completes the move. In this context, we will succeed in showing that fate has a strategy which will force a loss on a player with too few hands, and that a player with enough hands can force a win against any strategy employed by fate.

To show that $[(p-1)/p]n$ hands are necessary, we will need some terminology. For any integer k with $1 \leqslant k \leqslant n$, an $[{}^{n}_{k}]$–**gon** is obtained by choosing every kth vertex (clockwise) of the regular n-gon, beginning and ending with the same vertex (which may require several trips around the n-gon). Two vertices of an $[{}^{n}_{k}]$-gon are said to be **successive** if one is followed immediately by the other in the order in which the $[{}^{n}_{k}]$-gon is traversed. (Notation similar to this has been used by Coxeter ([**1**] and O'Daffer and Clemens [**4**] in connection with star polygons.)

Evidently, an $[{}^{n}_{k}]$-gon has $n/(k,n)$ vertices, where (k,n) denotes the greatest common divisor of k and n, and these vertices are the vertices of a regular $n/(k,n)$-gon, although successive vertices of the $[{}^{n}_{k}]$-gon are not necessarily adjacent vertices of the regular $n/(k,n)$-gon. FIGURE 1(a) shows the $[{}^{10}_{4}]$-gon in the 10-gon, while FIGURE 1(b) depicts the $[{}^{90}_{35}]$-gon in the 90-gon.

THEOREM 1. *Let n be any natural number, let p be the largest prime divisor of n, and let n_0 be any natural number less than $[(p-1)/p]n$. Then the n-well game cannot be won by a player with n_0 hands.*

Proof. Suppose first that n is prime. Fate will begin by setting the glasses so that not all are in the same state. Since the player has $n-2$ (or fewer) hands, his intended hand pattern for any move will include two unchecked wells. Fate notes these for the current move, and that one is located, say, k wells clockwise from the other (where $1 \leqslant k < n$). Since n and k are relatively prime, the $[{}^{n}_{k}]$-gon includes all n vertices of the table. Since not all glasses are in the same state, two glasses at successive vertices of the $[{}^{n}_{k}]$-gon will differ. Fate rotates the table to place these glasses in the unchecked positions in the player's hand pattern, thus preventing a win on this, or indeed any, move.

If now $n = p \cdot l$, where p is a prime, and if the player has fewer than $(p-1)\cdot l$ hands, fate initially sets the glasses so that those on the vertices of some regular p-gon P are not all in the same state. Since $n = p \cdot l$, there are l distinct $[{}^{n}_{l}]$-gons each with p vertices; thus any hand pattern must leave two unchecked vertices in some $[{}^{n}_{l}]$-gon Q. But the $[{}^{n}_{l}]$-gons are all regular p-gons, so fate can rotate the table to make P and Q coincide in such a manner that, as before, the two unchecked glasses are in different states.

Having established that no winning strategy is available to a player of the n-well game who has fewer than $[(p-1)/p]n$ hands, we will show that the same game can, in fact, be won by a player with $[(p-1)/p]n$ hands.

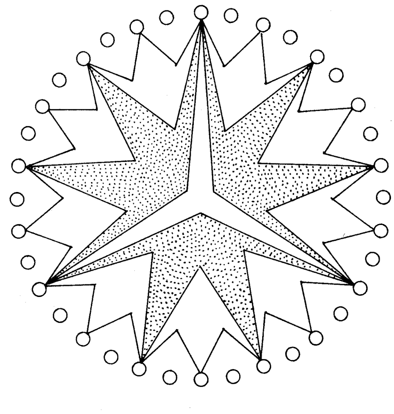


FIGURE 2

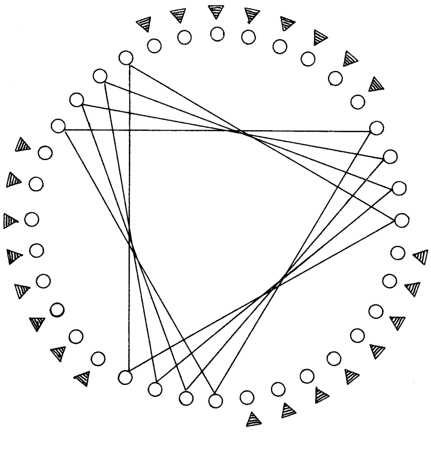


FIGURE 3

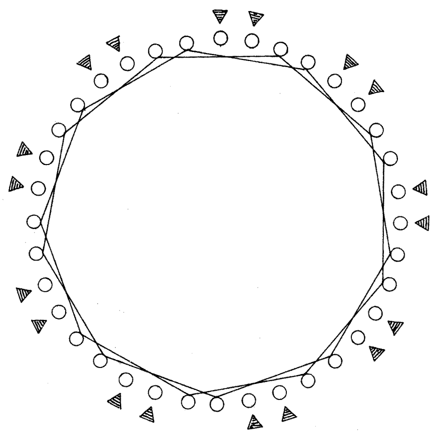


FIGURE 4

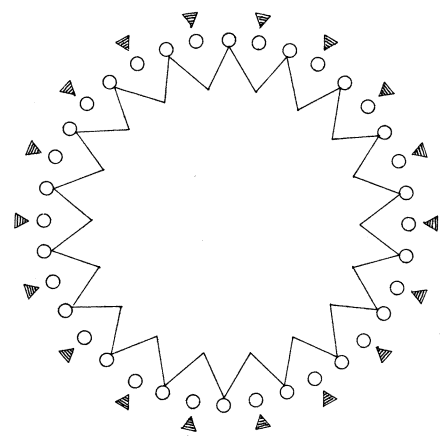


FIGURE 5

To provide an algorithm, we will factor n in such a way that no factor is larger than p. Any such factorization will do, but it proves efficient (for reasons we will discuss later) to write: $n = p_0 p_1 \cdots p_m$, where $p = p_0 \geq p_1 \geq \cdots \geq p_m$, and where each of $p_1, p_2, \ldots, p_m$ is as large as possible.

In our algorithm, certain $[{n \atop l}]$-gons play a primary role, and we will find it convenient to reserve for them special names. For $0 \leq k \leq m$, then by a ***k*-star** we mean any $[{n \atop l_k}]$-gon, where $l_k = p_k \cdots p_m$. FIGURE 2 shows the 36-well table with a 3-star (18 vertices), an embedded 2-star (9 vertices), and an embedded 1-star (3 vertices).

The moves (or hand patterns) used by the player will be specified by describing the positions of the wells (vertices) being left unchecked. For $0 \leq k \leq m$, the ***k*-move** is the pattern which leaves unchecked the vertices of c_k consecutive k-stars, where

$$c_k = \begin{cases} 1, & \text{if } k = m, \\ p_{k+1} p_{k+2} \cdots p_m, & \text{if } 0 \leq k < m, \end{cases}$$

(and where k-stars are **consecutive** if one can be obtained from the other by shifting each vertex one space clockwise). For example, in the 36-well game, a 1-move contacts all wells except those at the vertices of 4 consecutive 1-stars (FIGURE 3), a 2-move contacts all wells except those at the vertices of 2 consecutive 2-stars (FIGURE 4), and a 3-move contacts the vertices of a 3 star, while missing the vertices of a 3-star (FIGURE 5).

Here are some useful facts about the k moves. The first two are easy to check, while the third requires some explanation.

1. Proceeding in a clockwise manner around the table, the k-move alternately contacts $l_k - c_k$ consecutive vertices and misses c_k consecutive vertices.
2. The k-move requires $[(p_k - 1)/p_k]n$ hands (so the player always has sufficiently many hands).
3. If the s-move is applied for $s = m, m-1, \ldots, r$ in order, the only vertices left unchecked are the vertices of some r-star.

To verify the third fact, it will suffice to show that the vertices from a $(k+1)$-star which are unchecked by a k-move form a k-star. But $c_k = l_{k+1}$, so a blank space in the k-move has one vertex in common with a given $(k+1)$-star, and the vertices of the $(k+1)$-star thus falling in blank spaces fall in the same position in each blank space (since they are separated by multiples of l_{k+1} vertices). Hence the blanked vertices of the $(k+1)$-star form a k-star, as claimed.

It will be convenient, hereafter, to refer to an upright glass as being in **state** 0, and an inverted glass as being in **state** X. We will say the table is in an ***r*-phase**, $0 \leq r \leq m$, if glasses belonging to the same r-stars are uniformly in the same state. Thus the 36-well table in FIGURE 6(a) is in a 1-phase, since every 1-star (one is pictured) finds glasses in the same state. If in addition, there is precisely one r-star whose glasses are in one state, while all other glasses are in the opposite state, the table is said to be in a **principal r-phase**, and the vertices of the maverick r-star are referred to as **principal vertices**. In this situation, a k-star containing all of the principal vertices is called a **principal k-star** (where, of course, $r \leqslant k \leqslant m$). In FIGURE 6(b), the 36-well table is in a principal 2-phase, and the principal 2-star and principal 3-star are identified.

THEOREM 2. *If the table is in an r-phase, the player can produce either a win or a principal r-phase.*

Proof. Apply in order moves $m, m-1, \ldots, r$, each time turning all contacted glasses upright. Since the table is in an r-phase, if move r does not produce a win, the table must be in a principal r-phase, by Fact 3.

COROLLARY 3. *Whatever beginning position fate selects, the player can produce either a win or a principal 0-phase.*

Proof. Any position is a 0-phase.

Now the player's objective is to effect the passage from a principal r-phase to a principal s-phase for some $s > r$. (The principal m-phase can be won with move m, so this procedure will produce a win.) Supposing that the principal vertices have state X, the **basic strategy** he employs may be described as follows: *Perform in sequence moves $m, m-1, \ldots, r+1$, each time converting all contacted glasses to state X unless on some move the principal vertices are contacted, in which case change the state of these principal glasses to* 0.

The following two theorems will show that this basic strategy produces the desired conversion (from a principal r-phase) when $p_r \geqslant 3$. When $p_r = 2$ $(= p_{r+1} = \cdots = p_n)$, the strategy required is essentially the same but involves a coding procedure in move $r+2$, so is dealt with separately.

It is to be noted that the strategy is workable only if the player can recognize the principal vertices if and when fate presents them to him. This is obviously true on move m, and the following theorem asserts that it will remain the case for the other moves:

THEOREM 4. *Suppose that the basic strategy is employed through move $k+1$, where $m \geqslant k+1 \geqslant r+2$, and suppose that no move has encountered the principal vertices. Then the only k-star which fails to have all vertices in the same state is the principal k-star.*

An inductive proof of Theorem 4 follows neatly from Fact 3 together with the observations that move j looks at consecutive j-stars and each j-star is a subset of a $j+1$-star.

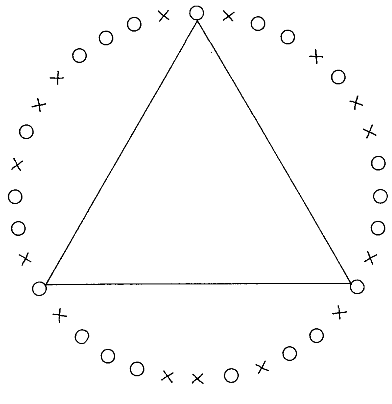


FIGURE 6(a)

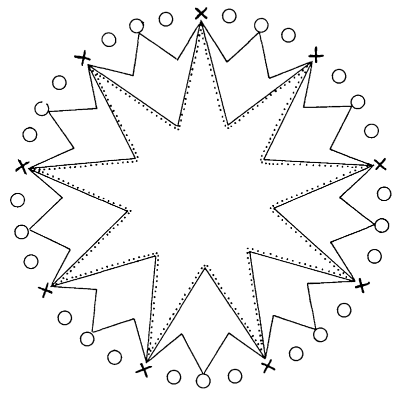


FIGURE 6(b)

THEOREM 5. *If $p_r \geqslant 3$, the basic strategy will enable the player to produce a principal s-phase for some $s > r$.*

Proof. If fate were to show the principal vertices on any of the moves, that move would produce the principal s-phase for an $s > r$. If fate were to prevent the player from contacting the principal vertices, we claim that the player can win the game outright, or else produce the principal $r+1$-phase by performing move r and adopting the following strategy:

(i) If fate does not show the principal vertices, win by changing all glasses in state 0 to X;
(ii) if fate does show the principal vertices, obtain the principal $r+1$-phase by changing the principal vertices to state 0.

To check that this works, note that after move $r+1$, the only $r+1$-star that contains glasses of state 0 is the principal $r+1$-star (Fact 3). Further, move r will see all but some r-star of the principal $r+1$-star. This means that each group of consecutive hands sees either $p_r - 1$ glasses of state 0 (i.e., misses the principal r-star) or $p_r - 2$ glasses of state 0. Since $p_r > 2$, it is now obvious that the player can employ the strategy we describe.

Our proof shows clearly that the basic strategy requires that $p_r \geqslant 3$ and furthermore, that if fate refuses to show the principal vertices prior to move r, then $(m+1) - r$ moves are needed to obtain a principal $r+1$-phase.

For the case where $p_r = 2$, all factors following p_r are also 2's. The strategy that should be employed when $m = r+1$ is given, essentially, in Gardner's article [3]. In our terminology (assuming the principal r-phase with principal vertices of state X): perform move m and if principal vertices are missed, change every other contacted glass to state X. Now perform move $m-1$ (= move r), and change all contacted glasses to the opposite state. If this does not win, perform move m, change all contacted glasses to opposite state and win.

An alternate strategy is used when $m \geqslant r+2$: *Perform, in sequence, move $k, k = m, m-1, \ldots, r+2$, each time switching all glasses encountered to state X unless on some move fate shows the principal vertices, in which case change the principal vertices to state* 0.

Assuming that the principal vertices are not contacted, the strategy on move $(r+1)$ is to code the table as follows:

(i) Switch the state of each glass located immediately clockwise to the contacted state 0 glasses, or if this is not possible,
(ii) switch the state of each glass located immediately counterclockwise to the contacted state 0 glasses.

Of course, if on move $(r+1)$ the player contacts the principal glasses he changes them to state 0. If he does not contact the principal glasses, then one of (i) or (ii) above must be possible: All glasses of state 0 are contained in the principal $(r+2)$ star, and move $(r+1)$ sees all of this

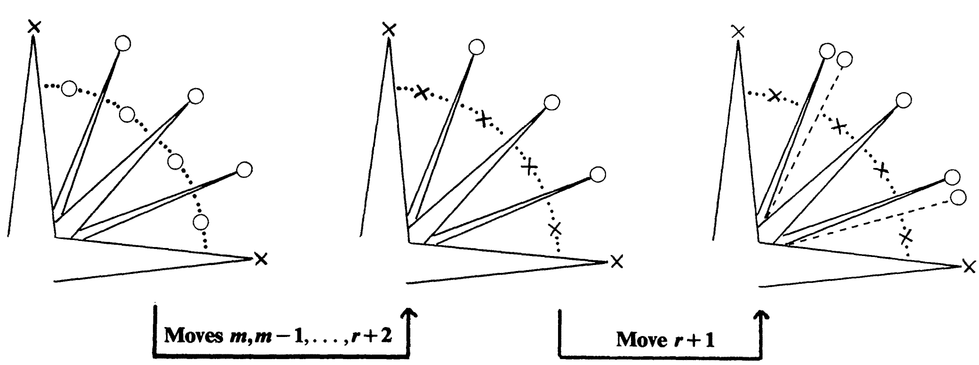


FIGURE 7

$(r+2)$-star except some $(r+1)$ star. Moreover, since $m \geqslant r+2$, the hands on move $(r+1)$ form groups of length at least 2.

FIGURE 7 depicts the alternate strategy through move $(r+1)$, assuming principal vertices are never seen. To obtain the reduction to an $(r+1)$-phase, the player would now perform move r. Here, he must contact either the isolated glasses of state 0 or the principal vertices of state X, and the coding (together with the memory of which of (i) or (ii) applies) will identify these glasses. He obtains the $(r+1)$-phase by, in the former case, switching the isolated 0 to an X, and in the latter, by switching the principal vertices to state 0.

We remark that move m would now enable us to obtain the principal $(r+1)$-phase, and this can be incorporated with the first move for reducing the $(r+1)$-phase to the $(r+2)$-phase. Thus, $(m+1)-r$ moves are required to change from a principal r-phase to a principal $r+1$ phase.

This completes the demonstration that the n-well game can be won with $[(p-1)/p]n$ hands. It is interesting to note that the number of moves required to win the n-well game depends only on the size of the selected factorization $n=p \cdot p_1 \cdots p_m$ of n, and is given (after a simple counting procedure) by

$$T(n)=\frac{m^2+5m+4}{2}.$$

Note that, although the presence of one or more factors of 2 creates difficulties with our strategy which require special treatment (the coding procedure) the moves actually used do not change, hence the number of moves is unaffected.

It might be of more interest to calculate, for a given number n, the "total energy" required to win. This can be measured as the total number of vertices contacted during the course of the game. The counting now, though still routine, is more elaborate since the various moves require different numbers of hands. The final figure (with $p=p_0$ for convenience) is

$$E(n,m,p_1,\ldots,p_m)=n\left[T(m)-\sum_{k=0}^{m}\frac{k+2}{p_k}\right].$$

The critical variable in this energy computation is m. The fewer factors used, the smaller E will be. Thus the general procedure for selecting a factorization would entail determining those factorizations with a minimum number of factors, and then selecting from among these the one which minimizes E. The procedure we suggested earlier, which requires that $p_1,p_2,\ldots,p_m$ be chosen recursively by choosing p_k to be the largest product of prime factors of $n/p_0p_1\cdots p_{k-1}$ which is less than p_0, will usually, but not always, produce the minimum energy solution. For example, it selects $37\cdot 32\cdot 11\cdot 11\cdot 11\cdot 5$ when $37\cdot 22\cdot 22\cdot 22\cdot 20$ is more efficient. We have so far failed in a search for an algorithm which will recursively produce a factorization of n (by factors less than p_0) with fewest factors. This seems to be an interesting problem in its own right.

Finally, we can now easily analyze the more general version of the n-well game in which each well contains an object which can be in any one of h states (say, $a_1,a_2,\ldots,a_h$). Try first, via Theorem 2 and Corollary 3, to change each object to state a_1. This either wins or produces a two-state game with state 0 being a_1 and state X being one of a_2 through a_h. If we continue play on the assumption that $X=a_2$, either we will win or else (we do not see X and) end with a two-state game in which C is known to be one of a_3 through a_h. It is now clear how to continue.

References

[1] H. S. M. Coxeter, Introduction to Geometry, Wiley, New York, 1961.

[2] Martin Gardner, About rectangling rectangles, parodying Poe and many another pleasing problem, Mathematical Games, Sci. Amer., vol. 240, no. 2 (1979) 16–24.

[3] Martin Gardner, On altering the past, delaying the future and other ways of tampering with time, Sci. Amer. vol. 240, no. 3 (1979) 21–30.

[4] Phares G. O'Daffer and Stanley B. Clemens, Geometry: An investigative approach, Addison-Wesley, Menlo Park (1976).

PROBLEMS

DAN EUSTICE, Editor
LEROY F. MEYERS, Associate Editor
The Ohio State University

Proposals

To be considered for publication, solutions should be mailed before December 1, 1980.

1096. An electric timer runs continuously when the current is on, but the outlet from the timer transfers that current (the timer is "on") only during a preset time interval in a 24-hour day.

Suppose n electric timers are connected in series (the kth timer is plugged into the outlet of the $(k-1)$st timer) and the kth timer is set to be on for a random time interval of length L_k. If an electric light is plugged into the last timer in the series, what is the expected value for the length of time the light will be on during a 24-hour day? [*Peter Ørno, The Ohio State University.*]

1097. Show that real orthogonal $n\times n$ matrices can be represented as points in a Euclidean space of $n(n-1)/2$ dimensions and, as such, form the boundary of a convex region. [*I. J. Good & D. R. Jensen, Virginia Polytechnic Institute and State University.*]

1098. Player A flips $n+1$ coins and keeps n of the coins to maximize the number of heads. Player B flips n coins. The maximum number of heads wins, with ties awarded to Player B. Which player should win and what is the probability of winning? (See Problem 1071, this MAGAZINE, March 1979.) [*Peter Ørno, The Ohio State University.*]

1099. Let A be an $n\times n$ complex matrix and $f(t)$ a polynomial with complex coefficients such that $f(AA^*)=0$. Give an elementary proof that $f(A^*A)=0$. [*Anon, Erewhon-upon-Spanish River.*]

1100. Suppose that $F(x)$ is a power series (finite or infinite) with rational coefficients and $A_k=\int_0^1 x^kF(x)\,dx$ for integers $k\geqslant 0$.

(i) If all the A_k's are rational, must $F(x)$ be a polynomial?

(ii) Does there exist an $F(x)$ such that all the A_k's except one are rational?

(iii) Does there exist an $F(x)$ such that all the A_k's except $A_{P(k)}$, $k=0,1,2,\ldots,$ are rational, where $P(k)$ is an integer valued polynomial, e.g., $P(k)=2k$?

(iv)* Given a finite indexed set $A_{m(k)}$, $k=0,1,2,\ldots,n$, does there exist an $F(x)$ such that all the A_k's except the $A_{m(k)}$'s are rational? [*M. S. Klamkin & M. V. Subbarao, University of Alberta.*]

ASSISTANT EDITORS: DON BONAR, *Denison University;* WILLIAM A. MCWORTER, JR., *The Ohio State University. We invite readers to submit problems believed to be new. Proposals should be accompanied by solutions, when available, and by any information that will assist the editors. Solutions to published problems should be submitted on separate, signed sheets. An asterisk (*) will be placed by a problem to indicate that the proposer did not supply a solution. A problem submitted as a Quickie should be one that has an unexpected succinct solution. Readers desiring acknowledgment of their communications should include a self-addressed stamped card. Send all communications to this department to Dan Eustice, The Ohio State University, 231 W. 18th Ave., Columbus, Ohio 43210.*

Solutions

Matrix of Integers **January 1979**

1063. Let M be an $n \times n$ matrix of integers whose inverse is also a matrix of integers. Prove that the number of odd entries in M is at least n and at most $n^2 - n + 1$, and that these are the best possible bounds. [*D. A. Moran, Michigan State University.*]

Solution: If M has fewer than n odd entries there must be a row with all even elements. Suppose it is the ith row. Then the ith row in the product MM^{-1} will have all even elements since the product and sum of even numbers is again even. However, the ith element in the ith row must be 1 (since $MM^{-1} = I_n$) which is not even. Contradiction.

If M has more than $n^2 - n + 1$ odd entries, it must have two distinct rows with all odd elements. Say the ith and jth rows are the all odd rows. Then the ith and jth rows of MM^{-1} must be congruent componentwise (mod 2). However the ith element of the ith row must be 1 while the ith element in the jth row must be 0 (since $MM^{-1} = I_n$). Clearly $1 \not\equiv 0 \pmod 2$. Contradiction.

These bounds are actually reached by $M = I_n$

and by
$$M = \begin{bmatrix} 1 & 1 & 1 & 1 & \ldots 1 \\ 1 & 0 & 1 & 1 & \ldots 1 \\ 1 & 1 & 0 & 1 & \ldots 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 & \ldots 0 \end{bmatrix}.$$

The first case is obvious since $I_n^{-1} = I_n$. In the second case

$$M^{-1} = \begin{bmatrix} 2-n & 1 & 1 & \ldots & 1 & 1 \\ 1 & -1 & 0 & \ldots & 0 & 0 \\ 1 & 0 & -1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & \ldots & -1 & 0 \\ 1 & 0 & 0 & \ldots & 0 & -1 \end{bmatrix}$$

which is easy to show.

MICHAEL FINN, student
Lake Braddock Secondary School
Burke, VA

Also solved by Duane M. Broline, F. J. Flanigan, Richard A. Gibbs, Eli L. Isaacson, Jinku Lee, Gadi Moran (Israel), William Myers, Michael Raship, J. M. Stark, Edward T. H. Wang (Canada), Joseph Weening, and the proposer.

Famous Formula **January 1979**

1064. For each positive integer n, define

$$L(n) = \int_0^\infty \left(\frac{\sin x}{x}\right)^n dx.$$

It is well known that $L(1) = L(2) = \pi/2$.

(a) Find $L(3)$, $L(4)$, and $L(5)$.

(b*) Is there a formula for $L(n)$ for general n? [*Edward T. H. Wang, Wilfrid Laurier University.*]

Solution I: $L(n)$ can be evaluated by contour integration. By this method we find that

$$L(3)=\frac{3}{8}\pi,\quad L(4)=\frac{1}{3}\pi\quad\text{and}\quad L(5)=\frac{115}{384}\pi.$$

The evaluation of $L(5)$ is shown below and is followed by an outline of the derivation of formulas for $L(n)$.

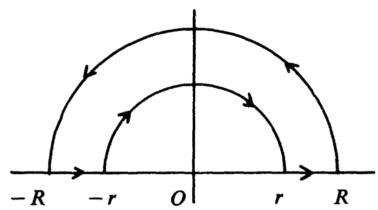
We can derive the identity

$$16\sin^5\theta=\sin 5\theta-5\sin 3\theta+10\sin\theta$$

from De Moivre's Theorem. Let $F_5(z)=e^{5iz}-5e^{3iz}+10e^{iz}$ and $f_5(z)=F_5(z)+P_2(z)$ where $P_2(z)=-5z^2-6$, z representing a complex variable. Then the function $f_5(z)/z^5$ is analytic in the whole plane except for a simple pole at $z=0$ where the residue is $115/12$. Let C be the contour shown below. Then we have

$$\int_C \frac{f_5(z)}{z^5}dz=\int_{-R}^{-r}\frac{f_5(x)}{x^5}dx+\int_r^R\frac{f_5(x)}{x^5}dx+I(r)+I(R)=0$$

in which $I(r)$ and $I(R)$ represent contour integrals along the semicircles of radii r and R respectively. After replacing x by $-x$ in $\int_{-R}^{-r}(f_5(x)/x^5)dx$ we can combine the first two integrals on the right hand side to obtain



$$2i\int_r^R\frac{16\sin^5 x}{x^5}dx=-I(r)-I(R).$$

Taking limits as $r\to 0$ and $R\to\infty$ it is easy but tedious to show that $I(r)\to -\frac{115}{12}\pi i$ and $I(R)\to 0$. (See, for example, R. V. Churchill, *Complex Variables and Applications,* 2nd ed., page 172, problem #11 and page 173, problem #14). Thus we have

$$\int_0^\infty\frac{\sin^5 x}{x^5}dx=\frac{115}{384}\pi.$$

The derivation of formulas for $L(n)$ follows the same ideas as outlined above. The identities

$$(-1)^{n/2}2^{n-1}\sin^n\theta=\sum_{k=0}^{n/2-1}(-1)^k\binom{n}{k}\cos(n-2k)\theta+(-1)^{n/2}\cdot\frac{1}{2}\binom{n}{n/2}\quad\text{if } n \text{ is even,}$$

$$(-1)^{(n-1)/2}2^{n-1}\sin^n\theta=\sum_{k=0}^{(n-1)/2}(-1)^k\binom{n}{k}\sin(n-2k)\theta\quad\text{if } n \text{ is odd,}$$

appearing in Hobson, E. W., *A Treatise on Plane and Advanced Trigonometry,* 7th ed., suggest that we define

$$f_n(z)=P_{n-3}(z)+\begin{cases}\displaystyle\sum_{k=0}^{n/2-1}(-1)^k\binom{n}{k}e^{i(n-2k)z}; & n \text{ even,}\\ \displaystyle\sum_{k=0}^{(n-1)/2}(-1)^k\binom{n}{k}e^{i(n-2k)z}; & n \text{ odd,}\end{cases}$$

in which $P_{n-3}(z)$, a polynomial of degree $n-3$, is determined so that $f_n(z)/z^n$ is analytic in the plane except for a simple pole at $z=0$, where the residue is $f_n^{(n-1)}(0)/(n-1)!$. Integrating along the contour shown above and taking limits as $r\to 0$ and $R\to\infty$, we will find that $I(R)\to 0$ and

$I(r) \to -\pi i$ Residue $(f_n(z)/z^n)$. This gives

$$L(n)=\begin{cases} \dfrac{(-1)^{n/2}\pi i^n}{2^n(n-1)!}\displaystyle\sum_{k=0}^{n/2-1}(-1)^k\binom{n}{k}(n-2k)^{n-1} & \text{if } n \text{ is even,}\\ \dfrac{(-1)^{(n-1)/2}\pi i^{n-1}}{2^n(n-1)!}\displaystyle\sum_{k=0}^{(n-1)/2}(-1)^k\binom{n}{k}(n-2k)^{n-1} & \text{if } n \text{ is odd.}\end{cases}$$

ROBERT L. YOUNG
Cape Cod Community College

Solution II: Let $f(x)=\sin^m x$ where $m>1$. Integration by parts gives the formula

$$\int_0^\infty \frac{f(x)}{x^m}\,dx=\frac{1}{m-1}\int_0^\infty \frac{f'(x)}{x^{m-1}}\,dx.$$

Repeating the process $m-2$ times we find

$$\int_0^\infty \frac{f(x)}{x^m}\,dx=\frac{1}{(m-1)!}\int_0^\infty \frac{f^{(m-1)}(x)}{x}\,dx. \tag{1}$$

Now we have the identity

$$\sin^{2n}x=\frac{1}{2^{2n-1}}\sum_{k=0}^{n-1}(-1)^{n-k}\binom{2n}{k}\cos(2n-2k)x+\frac{1}{2^{2n}}\binom{2n}{n}.$$

Differentiate this $2n-1$ times and use (1) with $m=2n$ to obtain

$$L(2n)=\frac{\pi}{2^{2n}(2n-1)!}\sum_{k=0}^{n-1}(-1)^k\binom{2n}{k}(2n-2k)^{2n-1} \tag{2}$$

since

$$\int_0^\infty \frac{\sin(2n-2k)x}{x}\,dx=\int_0^\infty \frac{\sin t}{t}\,dt=\frac{\pi}{2}.$$

Similarly, by differentiating the identity

$$\sin^{2n-1}x=\frac{1}{2^{2n-2}}\sum_{k=0}^{n-1}(-1)^{n-k-1}\binom{2n-1}{k}\sin(2n-2k-1)x,$$

$2n-2$ times and using (1), we find

$$L(2n-1)=\frac{\pi}{2^{2n-1}(2n-2)!}\sum_{k=0}^{n-1}(-1)^k\binom{2n-1}{k}(2n-1-2k)^{2n-2}. \tag{3}$$

Both formulas (2) and (3) can be written as follows:

$$L(m)=\frac{\pi}{2^m(m-1)!}\sum_{0<k<m/2}(-1)^k\binom{m}{k}(m-2k)^{m-1}.$$

Taking $m=3, 4, 5, 6$ we find

$$L(3)=\frac{3\pi}{8}, L(4)=\frac{\pi}{3}, L(5)=\frac{115\pi}{384}, L(6)=\frac{11\pi}{40}.$$

TOM M. APOSTOL
California Institute of Technology

Also solved by Ed Adams, E. M. Beesley, Gerald Bergum, Bruce C. Berndt, R. P. Boas, Paul Bracken (Canada), Paul F. Byrd, L. Carlitz, Michael J. Dixon, Henry E. Fettis, M. R. Gopal, Andrew P. Guinand (Canada), John P. Hoyt, Hans Kappus (Switzerland), B. Margolis (France), Nadim Nasir (Saudi Arabia), Reinhard Razen (Austria), *Otto G. Ruehr, Lajos Takács, Michael Vowe and the proposer. Several solvers of part (b) supplied references. The list of references is as follows:*

D. Bierens de Haan, Nouvelles Tables d'Intégrales Définies, Hafner, New York, 1957, pp. 217 and 222.

Joseph Edwards, A treatise on the Integral Calculus, vol, 2, Chelsea, New York, p. 209.

Erdelyi, Magnus, Oberhettinger, and Tricomi, Tables of Integral Transforms, vol. 1, McGraw-Hill, 1954, p. 20.

G. M. Fichtenholtz, Differential- und Integralrechnung, VEB Deutscher Verlag der Wissenschaften, Berlin, 1966, p. 60.

I. S. Gradšhteyn and I. M. Ryžik, Tables of Integrals, Series and Products, Academic Press, 1965, section 3.82.

N. I. Lobachevski, Collected Works, vol. 5, Moscow, 1951, pp. 303–312.

Oberhettinger, Tabellen zur Fourier Transformation, Springer, 1957, p. 19.

O. Schlömilch, Über das bestimme Integral $\int_0^\infty \frac{\sin^p x}{x^a}\,dx$, Zeitschrift für Mathematik und Physik, vol. 5, 1860, p. 286.

Zero and Ones — January 1979

1065. A is an $n+1$ by $n+1$ matrix; its $(1,1)$th element is 0 and all others are 1. Find a formula for the elements of A^k when $k \geqslant 2$. [*H. Kestelman, University College, London.*]

Solution: We see by induction that

$$A^k = \begin{bmatrix} a_k & b_k & \cdots & b_k \\ b_k & c_k & \cdots & c_k \\ \vdots & \vdots & & \vdots \\ b_k & c_k & \cdots & c_k \end{bmatrix}$$

where $a_1 = 0$, $b_1 = 1$, and $c_1 = 1$, and $a_{k+1} = n \cdot b_k, b_{k+1} = n \cdot c_k, c_{k+1} = b_k + n \cdot c_k$. Thus the sequence $\{b_k\}$ satisfies the recursion $b_{k+2} - n \cdot b_{k+1} - n \cdot b_k = 0$. The characteristic polynomial $x^2 - nx - n$ has the distinct roots

$$r_1 = \frac{1}{2}(n + \sqrt{n^2+4n}\,), r_2 = \frac{1}{2}(n - \sqrt{n^2+4n}\,)$$

so we can write $b_k = d_1 \cdot r_1^k + d_2 \cdot r_2^k$. Since $b_1 = 1$ and $b_2 = n$, we can solve two simultaneous linear equations to get $d_1 = (\sqrt{n^2+4n}\,)^{-1}, d_2 = -(\sqrt{n^2+4n}\,)^{-1}$. Therefore

$$b_k = \frac{1}{\sqrt{n^2+4n}}(r_1^k - r_2^k),$$

$$a_k = \frac{n}{\sqrt{n^2+4n}}(r_1^{k-1} - r_2^{k-1}),$$

$$\text{and} \quad c_k = \frac{b_{k+1}}{n} = \frac{r_1^{k+1} - r_2^{k+1}}{n\sqrt{n^2+4n}}.$$

JOSEPH S. WEENING, student
Princeton University

Also solved by Duane M. Broline, Michael Finn, Clifford H. Gordon, I. P. Goulden (Canada), Kenneth Michael Gustin, G. A. Heuer, Eli L. Isaacson, Ralph Jones, Lew Kowarski, Roger B. Nelsen, Joseph O'Rourke, Lal Sabharwa, Michael Raship, David Singmaster (England), J. M. Stark, Gerald Thompson, Michael Vowe, Ken Yocum, and the proposer. A solver from Munich had an unreadable signature.

The Last 1 — March 1979

1066. Consider the following children's game ("clock"): k copies of well-shuffled cards numbered $1,2,3,\ldots,L$ are distributed in boxes labeled $1,2,3,\ldots,L$, with exactly k cards per box. (For the game as normally played, a standard deck of playing cards is used, with $L = 13$ and $k = 4$.) At the start of the game the top card in box 1 is drawn. If the value of this card is j

$(j=1,2,3,\ldots,L)$ we proceed to box j, draw the top card and go to the box so numbered, draw the top card, and so on. The objective of the game (a 'win') is to draw all cards from every box before being directed to an empty box. Characterize all winning distributions of cards, and find the probability of a win. [*Eric Mendelsohn & Stephen Tanny, University of Toronto.*]

Solution: Observe that there is a correspondence between the hands as they are dealt in the L piles and the order of the cards as they turn up during the play of the game. The key idea is this: the winning sequences are those (of length kL) which end in a 1 (presuming that the first card drawn is from the first pile). To see this, observe that after the first card is turned over, there are only $k-1$ cards left in pile 1. Therefore, the kth 1 drawn will lead to an empty pile; so to win, this 1 has to appear last in the sequence. As for the other piles, the game can never end in them, because only k cards direct the play to any one pile, and all piles except the first have k cards in them after the first card is turned over. Thus, the probability of winning the game is the same as the proportion of sequences (of length kL) which end in 1, and this is obviously $1/L$.

DAVID KLEINER, student
St. Olaf College

Also solved by Michael W. Ecker, B. R. Johnson (Canada), R. S. Stacy (Germany), Roberta S. Wenocur, and the proposers.

Shortest Chord sans Calculus **March 1979**

1067. Problem P. M. 11 on the first William Lowell Putnam Competition, April 16, 1938, was to find the length of the shortest chord that is normal to the parabola $y^2=2ax, a>0$, at one end. A calculus solution is quite straightforward. Give a completely "non-calculus" solution. [*M. S. Klamkin, University of Alberta.*]

Solution: Let t be a real parameter, $t>0$. Then $P[t^2/(2a),t]$ is a point on that part of the parabola $y^2=2ax$ which lies above the x-axis.

A line with slope m passing through P has equation

$$y-t=m\left[x-(t^2/2a)\right]. \tag{1}$$

Without use of the calculus, the tangent line to the parabola at point P is found by requiring that line (1) have a double point of intersection with the parabola at P. Solving (1) for y and substituting into $y^2=2ax$ gives a quadratic equation in x whose discriminant can be written in the form $(2a-2mt)^2$. Requiring that this discriminant be zero gives $m=(a/t)$ as slope of the tangent line to the parabola $y^2=2ax$ at the point $P[t^2/(2a),t], t>0$. By trigonometry the slope of the normal line to this parabola at P is $-t/a$, and so the normal line at P has equation

$$y-t=(-t/a)\left[x-(t^2/2a)\right]. \tag{2}$$

Solving simultaneously (2) and $y^2=2ax$ gives that the normal line to the parabola $y^2=2ax$ at the point $P[t^2/(2a),t]$, $t>0$, also intersects the parabola at the point $Q[(t^2+2a^2)^2/(2at^2), -t-(2a^2/t)]$.

Denote by L the length of the chord $\overline{PQ}$, normal to $y^2=2ax$ at P. Using the formula for the distance between two points and factoring, it is easily obtained that

$$L^2=(a^6/t^4)\left[4(t/a)^2+1\right]\left[(t/a)^2-2\right]^2+27a^2. \tag{3}$$

Now (3) shows that $L^2\geqslant 27a^2$ and that for $t>0$, the shortest chord normal to the parabola at one end is obtained when $t=\sqrt{2}\,a$, and the length of this shortest chord is $3\sqrt{3}\,a$.

J. M. STARK
Lamar University

Also solved by Mangho Ahuja, J. Dou (Spain), Frank Eccles, Michael Goldberg, Hans Kappus (Switzerland), Russell Lyons, V. N. Murty, R. S. Stacy (Germany), and the proposer.

1068. Given a simple closed curve S, let the "navel" of S denote the envelope of the family of lines that bisect the area within S.

(a) If S is a triangle, find sharp upper and lower bounds for the ratio of the area within the navel of S to the area within S.

(b)* If S bounds a convex set, find a sharp upper bound for this ratio.

(c)* If S is arbitrary, find a sharp upper bound for this ratio. [*James Propp, Great Neck, New York.*]

Solution: (a) Let A_S, A_N denote the area of triangle $S=ABC$ and its navel, respectively. Let f be an affine transformation mapping S into $S^*=f(S)$. If N^* is the navel of triangle S^*, then clearly $N^*=f(N)$ and $A_N/A_S=A_{N^*}/A_{S^*}$. Since any two triangles are affine images of each other, the above ratio must be the same for all triangles. Hence we may consider, without loss of generality, the special triangle with vertices $A=(0,0), B=(4,0), C=(0,4)$, say.

To construct a line B_1C_1 cutting off a triangle AB_1C_1 with area $A_S/2$, choose B_1 on AB with $AB_1 \geqslant B_1B$, fix D on AC such that BD is parallel to B_1C and take C_1 to be the midpoint of AD.

Putting $B_1=(t,0)$, $2 \leqslant t \leqslant 4$, the family of segments B_1C_1 is given by

$$y=-(8/t^2)(x-t), \quad 2 \leqslant t \leqslant 4. \tag{1}$$

Eliminating t from the pair of equations (1) and

$$\frac{\partial y}{\partial t}=+\frac{16x}{t^3}-\frac{8}{t^2}=0, \tag{2}$$

we see that the envelope of the family (1) is the hyperbolic arc

$$y=2/x, \quad 1 \leqslant x \leqslant 2. \tag{3}$$

In the same fashion the envelopes of the families of lines C_2A_2, A_3B_3, cutting off triangles A_2BC_2, A_3B_3C respectively, having area $A_S/2$, turn out to be hyperbolic arcs, symmetric to one another with respect to the line $y=x$, namely

$$x=4-y-2/y, \quad 1 \leqslant y \leqslant 2 \tag{4}$$

and

$$y=4-x-2/x, \quad 1 \leqslant x \leqslant 2. \tag{5}$$

Thus from (3)–(5) we get by a straightforward integration

$$A_N=\frac{1}{2}-3\int_1^2\left(3-x-\frac{2}{x}\right)dx=6\log 2-4. \tag{6}$$

Finally,

$$A_N/A_S=(3\log 2-2)/4 \approx .01986. \tag{7}$$

HANS KAPPUS
Switzerland

Comment: Essentially the same problem as part (a) had been proposed by me some years ago. It is E 1228 (Amer. Math. Monthly, 63 (1956) 491). A similar problem was again proposed as E 1343 (Amer. Math. Monthly, 65 (1958) 774).

My conjecture is that the solution for part (b) is the same as for part (a), namely $(3\log 2-2)/4$. That is, the triangle is the "most asymmetric" convex set. I do not have a proof—only a sketchy outline. As for part (c), I am inclined to conjecture again that the same upper bound will hold but I have even less solid grounds for this statement.

VIKTORS LINIS
University of Ottawa

Part (a) also solved by Jordi Dou (Spain) and the proposer.

REVIEWS

Paul J. Campbell, Editor
Beloit College

Pierre Malraison, Editor
Control Data Corp.

Assistant Editor: Eric S. Rosenthal, Princeton University. Articles and books are selected for this section to call attention to interesting mathematical exposition that occurs outside the mainstream of the mathematics literature. Some reviews of books are adapted from the Telegraphic Reviews in the American Mathematical Monthly.

Wolfe, Philip, *The "Ellipsoid algorithm,"* Science 208 (18 April 1980) 240, 242.

An informative, concise letter reporting on progress and prospects for the much heralded new algorithm for linear programming introduced last year by the Russian mathematician L.G. Khachian. Wolfe proposes to term this type of algorithm EA (for ellipsoid algorithm), one version of which is the Khachian variant--which itself will not be of practical value. "One can, however, say this for EA: it solved a significant theoretical problem, and can be used to solve others; it may still be practically useful for difficult nonlinear problems."

Friedman, Judith, *Group theory,* Scientific American 242 (May 1980) 82-84.

A brief report in the Science and the Citizens column of the recent announcements by Robert Griess and others that two of the largest proposed sporadic groups, the "monster" F_1 and the Janko group J_4, have been proven to really exist. This work goes a long way towards completing the classification of the finite simple groups. "The end is clearly in sight."

An Agenda for Action: Recommendations for School Mathematics of the 1980's, NCTM, 1980.

A forceful set of recommendations in response to what NCTM terms a crisis in mathematics education, caused by a public preoccupation with tests of basic skills that has blinded people to other far more serious problems. The report recommends in part that the notion of basic skill be broadened to include more than computational skill, and that problem solving replace computation as the focus of all school mathematics. The report is available for free from NCTM.

Ralston, Anthony and Shaw, Mary, *Curriculum '78--Is computer science really that unmathematical?* Communications Assoc. Comp. Machinery 23 (February 1980) 67-70.

"Inevitably,...the fundamental principles and theories can only be understood through the medium of mathematics...we focus on the place of mathematics in the computer science curriculum and try to show how badly Curriculum '78 fails in this respect." The authors, computer scientists, instead urge two years of discrete mathematics, followed by an introduction to calculus in the junior year.

Thomsen, Dietrick E., *Making music--fractally,* Science News 117 (23 March 1980) 187, 190.

If loudness and melodic fluctuations are analyzed together, a wide variety of music types exhibit fractal structure. More structured than white noise, but less so than Brownian (or $1/f^2$) noise, artificially-generated fractal (or 1/f) noise can be surprisingly similar to real music.

Pycior, Helena M., *Benjamin Peirce's linear associative algebra,* Isis 70 (December 1979) 537-551.

"Mathematics is the science which draws necessary conclusions." So begins Peirce's epoch-making 1870 work. In it he also marks the introduction of zero divisors, idempotents, and nilpotents; contains 150 examples of associative algebras; and offers a theological argument to legitimate his new algebras. The article sets his achievement in perspective.

Jones, W.F.B., *A computer program to play Othello,* Mathematical Spectrum 12 (1979/80) 17-24.

A dated catalogue of game-playing programs introduces this article. The main work is an empirical investigation of parameter weights in an evaluation function for the game best known as Reversi.

Meyer, Paul R. (Ed.), Papers in Mathematics, Annals of the New York Academy of Sciences, V. 321, 1979; 101 pp, $20.

Nine fine short papers, at the level of *Mathematics Magazine,* treating nonassociative multiplication, undecidable topological statements, efficient algorithms, physics of wind instruments, and other topics.

Moser, Jürgen, *Hidden symmetries in dynamical systems,* American Scientist 67 (November-December 1979) 689-695.

Survey of recent progress.

Malina, Frank J. (Ed.), Visual Art, Mathematics, and Computers: Selections from the Journal *Leonardo,* Pergamon, 1979; xiv + 325 pp, $55.

More than 50 essays from the last ten years, by authors from all over the world, on: mathematics and art, computer art, mathematical themes in art, and accounts of human-computer interaction in fashioning art. There are hundreds of figures, including 15 color plates. The authors are physicists, engineers, computer scientists, and artists--except for two contributions by Frank Harary, mathematicians are (surprisingly?) under-represented.

Bers, Lipman, *et al., An unpublished reply,* Science 208 (4 April 1980) 6.

Quotes paragraph from autobiographical article in *Uspekhi Mat. Nauk* by L.S. Pontryagin, in which the latter expressed unmistakably anti-Semitic remarks and characterized Nathan Jacobson as "a mediocre scientist but an aggressive Zionist." P.S. Aleksandrov, editor of the publication, has refused to publish a reply by Jacobson. The 10 very distinguished authors of this letter to *Science* deplore this unprecedented attack and lack of opportunity for rebuttal.

Bentley, Jon Louis, *An introduction to algorithm design,* Dr. Dobb's Journal of Computer Calisthenics & Orthodontia 5:44 (April 1980) 5-15.

Examples of slow and fast algorithms for subset testing, substring searching, matrix multiplication, fast Fourier transform, and public-key cryptosystems.

Saaty, Thomas L., *The U.S.-OPEC energy conflict: The payoff matrix by the analytic hierarchy process*, International Journal of Game Theory 8 (1979) 225-234.

Saaty's analytic hierarchy process employs eigenvalues of matrices of ranks and their reciprocals. As in his articles on conflict in Northern Ireland, terrorism, corporate planning, and the future of higher education, the method proves a useful tool in analyzing conflict situations.

Anderson, Norman H., *Algebraic rules in psychological measurement*, American Scientist 67 (September-October 1979) 555-563.

Treats the problem of finding true linear scales of subjective sensation.

Unguru, Sabetai, *History of ancient mathematics: Some reflections on the state of the art*, Isis 70 (December 1979) 555-565.

Greek mathematics: geometry, or algebra disguised as geometry? Geometry, argues Unguru, standing against Freudenthal, van der Waerden, and Weil.

Westfall, Richard S., *Newton's marvelous years of discovery and their aftermath: Myth versus manuscript*, Isis 71 (March 1980) 109-121.

"For whatever it is worth, Newton's manuscripts do not indicate that anything special happened in Woolsthorpe." So concludes the author about Newton's 1664-1666 absence from Cambridge, asserting a continuity of effort and achievement before, during, and after: no "miraculous burst of creativity."

Moser, W., Problems in Discrete Geometry, 4th ed. (from the author), 1979; 23 pp.

Collection of 28 research problems, all understandable. Sample: (Grünbaum) What is the smallest number of vectors determined by n points in the plane? in E_3? References for the problems include 5 citations of *Mathematics Magazine.*

Kapur, J.N., *Some problems in biomathematics*, International Journal of Mathematical Education in Science and Technology 9 (1978) 287-306.

Grab-bag of mathematical models in biology and medicine.

Pielou, E.C., Biogeography, Wiley, 1979; ix + 351 pp.

Exciting interdisciplinary survey of enormous breadth, aimed at the level of senior undergraduates, and developed with the help of a cornucopia of mathematical concepts and techniques. Topics include biological implications of plate tectonics, evolutionary consequences of biogeographical change, the results of the last ice age, marine and island biogeography, geoecology, dispersion of species, and variations among species and within a species' range. A vast number of figures illustrate the text. "...a model that does *not* fit is often more apparent than real, whereas a bad fit usually permits the rejection of false hypothesis and to that extent constitutes a gain in knowledge.

Guest, Julius, New Shapes: A Collection of Computer-Generated Designs, R.A. Vowels (93 Park Drive, Parkville 3052, Victoria, Australia), 1979; v + 174 pp, $12 (+ $1.20 postage).

Attractively printed designs produced by Algol-60 programs on a digital plotter; the programs are printed in the appendix. Almost all of the designs are centrally symmetric, so as to remind an American of the popular linkage drawing toy "Spirograph."

NEWS & LETTERS

MATHEMATICS TEACHERS URGE REFORM OF SCHOOL MATHEMATICS PROGRAMS

Responding to what it terms a "crisis" in school mathematics, the National Council of Teachers of Mathematics has sent forth at its annual meeting in Seattle, Washington, a series of policy recommendations for mathematics education in the 1980's. Shirley Hill, Council President, challenged parents, policy makers, and the general public to address three major problems that contribute to the crisis:

- School mathematics is not keeping pace with the changing needs dictated by developing technologies;
- Most students are not taking as much mathematics in high school as they will need for their future careers;
- The present shortage of qualified mathematics teachers is increasing dramatically, largely due to greater professional and financial rewards in other technological careers.

"Policy makers are not confronting the deepest problems," says Hill, "because the public and its representatives have been diverted by a fixation on test scores."

The Council recommends that mathematics programs at all levels concentrate on problem solving, not just on acquiring techniques, and that the scope of basic knowledge in mathematics be expanded to include skills essential for the future, not merely those required for present needs. "Skills are tools, and their importance rests in the needs of the times."

In such a future-directed curriculum many skills formerly considered basic become obsolete. The Council cites as one example the continuing stress in elementary classrooms on multiplication and division of large numbers, even though all current work is done on calculators and computers. Indeed, the Council urges that mathematics programs take full advantage of calculators and computers at all grade levels, and that computer literacy become part of the education of every student.

The Council also urges that three years of mathematics in grades 9-12 should be required of all high school graduates, and cites evidence from a survey of parents supporting this recommendation. (In many states now only one year is required.) To help bring about this change, they urge colleges to stop awarding college credit for courses covering mathematics ordinarily taught in high school. This practice, the teachers believe, encourages students to take only the minimum requirement in high school.

GÖDEL, ESCHER, BACH WINS PULITZER PRIZE

Douglas Hofstadter, Professor of Computer Science at Indiana University, received a 1980 Pulitzer prize for his innovative book *Gödel, Escher, Bach: An Eternal Golden Braid* (Basic Books, 1979) which the Pulitzer committee called a work of "mathematical philosophy."

FAMOUS MATHEMATICIANS FEATURED AT AUGUST ANN ARBOR MEETING

The annual summer meeting of the Mathematical Association of America will feature two internationally famous mathematicians: John Conway of Trinity College, Cambridge University, and L.K. Hua of the Institute of Mathematics, Academia Sinica. The titles of their talks will be "Pensively penetrating Penrose's pentapieces" and "Some personal experiences for popularization of mathematics in China," respectively.

The meeting will be held from Monday, August 18 until Wednesday, August 20 on the campus of the University of Michigan in Ann Arbor, Michigan. Professor George Andrews of Pennsylvania State University will deliver the Earle Raymond Hedrick Lectures on Monday morning and afternoon, and on Tuesday morning.

1980 U.S.A. MATHEMATICAL OLYMPIAD

The ninth annual U.S.A. Mathematical Olympiad, which took place on May 6, 1980, consisted of the following five problems.

1. A two pan balance is inaccurate since its balance arms are of different lengths and its pans are of different weights. Three objects of different weights A, B and C are each weighed separately. When placed on the left hand pan, they are balanced by weights A_1, B_1, C_1, respectively. When A and B are placed on the right hand pan, they are balanced by A_2 and B_2, respectively. Determine the true weight of C in terms of A_1, B_1, C_1, A_2 and B_2.

2. Determine the maximum number of different three term arithmetic progressions which can be chosen from a sequence of n real numbers $a_1 < a_2 < \dots < a_n$.

3. Let

$$F_r = x^r \sin(rA) + y^r \sin(rB) + z^r \sin(rC) ,$$

where x, y, z, A, B, C are real and $A+B+C$ is an integral multiple of π. Prove that if $F_1 = F_2 = 0$, then $F_r = 0$ for all positive integral r.

4. The inscribed sphere of a given tetrahedron touches each of the four faces of the tetrahedron at their respective centroids. Prove that the tetrahedron is regular.

5. If $1 \geq a,b,c \geq 0$, prove that

$$\frac{a}{b+c+1} + \frac{b}{c+a+1} + \frac{c}{a+b+1} + (1-a)(1-b)(1-c) \leq 1 .$$

NATIONAL EDUCATIONAL COMPUTING CONFERENCE

The second annual National Educational Computing Conference will be held June 23-25 in Norfolk, Virginia. The Conference is intended to provide a forum for discussion among individuals at all levels and from all institutions with interest in educational computing.

Of special interest to mathematicians and mathematics educators are tutorial sessions on PASCAL, presentations on symbolic computer mathematics, and descriptions of computer-assisted courses in calculus, finite mathematics and biomathematics.

Further information on the conference and on possible travel support can be obtained from Professor Gerald Engel, NECC Conference Chairman, Computer Science Department, Christopher Newport College, 50 Shoe Lane, Newport News, Virginia 23606.

PI MU EPSILON STUDENT CONFERENCE

The seventh annual undergraduate conference of Pi Mu Epsilon will be held September 26-27, 1980 at Miami University, Oxford, Ohio, in conjunction with the Statistics Conference being held at the same time (see p. 60 of the January issue of this *Magazine*).

Conference planners seek undergraduate papers on any topic in mathematics, statistics or computing; presentations should be from fifteen to thirty minutes in length.

Undergraduates wishing to present a paper should submit title, preferred date, time length, and abstract by September 26 to Professor Milton D. Cox, Department of Mathematics and Statistics, Miami University, Oxford, Ohio 45056.

A SURVEY OF COMPUTER-BASED INSTRUCTION

A Consortium Interest Survey of those interested in or using computers or researching their use in mathematics instruction in high school and college courses is being distributed by Dr.

Ronald H. Wenger, Associate Professor of Mathematics and Associate Dean of the College of Arts and Science at the University of Delaware. Please write to him for a copy of this brief survey or to tell him of your own activities. His address is: 123 Memorial Hall, Univeristy of Delaware, Newark, Delaware 19711 (302-738-2351).

PROBLEM SOLVING COURSES

A subcommittee of the MAA Committee on the Teaching of Undergraduate Mathematics plans a survey of problem solving courses in mathematics at the secondary and undergraduate levels. Its chair, Alan Schoenfeld, says that the job of the subcommittee is to prepare a report which describes the "state of the art" in problem solving courses, lists available resources for teaching problem solving (and possibly creates some such resources), and makes recommendations regarding the place of problem solving in the curriculum and ways to teach it. The subcommittee plans to distribute a questionnaire to persons teaching such a course. Anyone with ideas as to what should be on the questionnaire, about useful resources, or about possible contributions the subcommittee might make, should get in touch with Alan H. Schoenfeld, Mathematics Department, Hamilton College, Clinton, NY 13323.

1980 NSF-CBMS REGIONAL RESEARCH CONFERENCES

The National Science Foundation has granted through the Conference Board of the Mathematical Sciences nine Regional Research Conferences for college and university mathematics teachers for the summer and fall of 1980. Each Regional Conference features ten lectures by a distinguished guest expert; subsequently a monograph by the principal lecturer based on his Regional Conference lectures normally appears in the Regional Conference Series in Mathematics published by the American Mathematical Society or in the Regional Conferences Series in Applied Mathematics published by the Society for Industrial and Applied Mathematics.

Date	Host Institution	Subject	Lecturer
June 16-20	Bowling Green State University	Jackknife and Bootstrap Methods in Statistics	E. Efron
June 16-20	University of Texas, Arlington	Nonlinear Functional Analysis and Applications to Differential Equations	J. Hale
June 23-27	SUNY, Albany	Brown-Peterson Homology	S. Wilson
July 7-11	Oklahoma State University	Asymptotic Sequential Estimation and Testing	M. Woodroofe
July 7-11	Washington University	Affine and Projective Structures on Complex Manifolds	S. Kobayashi
July 21-25	Rensselaer Polytechnic Institute	Mathematical Modeling of the Hearing Process	C. Steele
Jan. 12-16 1981	Pomona College	Global Topological Methods in Applied Mathematics	J. Yorke
Aug. 18-22	Emory University	Homology and Dynamical Systems	J. Franks
Dec. 15-19	Tulane University	Harmonic Maps	J. Eells

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, N.W.
Washington, DC 20036

MATHEMATICS MAGAZINE VOL. 53, NO. 3, MAY 1980