IX Verano de Investigación 1999

Harold V. McIntosh
Departamento de Aplicación de Microcomputadoras,
Instituto de Ciencias, Universidad Autónoma de Puebla,
Apartado postal 461, 72000 Puebla, Puebla, México.

June 5, 2000

Abstract

The Ninth Reasearch Summer was held in Mexico during July and August, 1999. The participants in my portion of this enterprise prepared individual reports of their summer's work, which can be consulted elsewhere. So here we survey some of the background ideas, topics discussed, and mention a few of the results obtained.

Contents

1 Introduction

Every year for the past several, the *Academy for Scientific Investigation* (now renamed the *Mexican Academy of Sciences*) and the *National Research Council* have sponsored some extracurricular activities destined to promote Science and the Image of Science in the republic. This year was the ninth such attempt, for which a group of students worked for two months on various projects, which can be consulted to learn about the individual projects. By contrast, some underlying themes of the activities carried out under our sponsorship are outlined below.

2 The Cross Ratio

Books on projective geometry depend heavily on the invariance of cross ratio with respect to linear fractional transformations, as do books on complex analysis. Verifying the invariance is customarily assigned as an exercise, but there seems to be almost no discussion of the sources of the cross ratio. As a result from the geometry of antiquity, its proof involves the comparison of similar triangles, but not much is said about the possible origins of those triangles in perspective drawings.

In our recent work the issue has come up in two places. First, the programs `PHOC` and `SHOC`, which are based on the books of E. A. Maxwell [4, 5], consist of a series of demonstrations based on exercises or theorems in the books. Since these rely heavily on the cross ratio, there is a natural element of curiosity regarding its origins. The most recent encounter with the elements of projective geometry has been in a course in mathematical methods. In the linear algrbra portion, there is a natural progression from linear spaces to affine spaces to projective spaces, especially during the development of group representation theory, where the cross ratio comes up in the study of projective invariants. In the complex analysis portion of the course, explaining the role of fractional linear transformations leads to the same interest in cross ratios.

Consider the projective line as an image of a two-dimensional linear algebra plane, and observe that determinants are multiplicative for linear transformations. Take a 2×2 matrix P containing the two columns of a basis and a transformation M. Then $|MP| = |M||P|$, so $|P|$ is not an invariant. But take another basis matrix Q for which $|MQ| = |M||Q|$. Then the quotient $|P|/|Q|$ *is* invariant, although its terms are not.

Carrying the result over to the projective line, write the matrices in projective form

$$P = \begin{bmatrix} xs & yt \\ s & t \end{bmatrix}, \quad Q = \begin{bmatrix} wu & zv \\ u & v \end{bmatrix},$$

and calculate the determinants. Note that w, x, y, z are the values of points on the projective line, whereas s, t, v, u are the multipliers lost by the projection. Then

$$\frac{|P|}{|Q|} = \frac{(x-y)st}{(w-z)uv}$$

Although this quotient is invariant enough, the multipliers would normally be unknown. Choosing more vectors and dividing once again would only create more multipliers, but the same multipliers can be kept by just rearranging the four vectors in the two bases. There are twenty four possibilities, the permutations of four objects, but only four of them produce the cancellation which would free the points from the multipliers.

The nice symmetry of the formula shows how to get the cancellation; note that each difference is multiplied by the product of its two multipliers. Making up a product of four multipliers in two

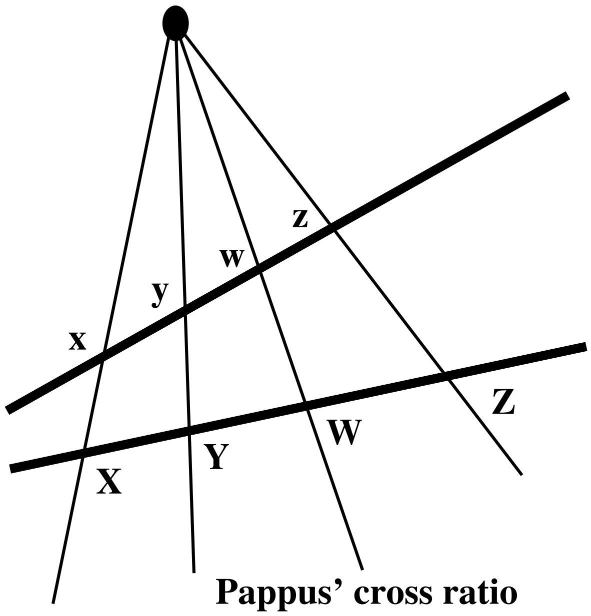


Figure 1: Pappus' construction for the invariance of the cross ratio. It would take a more cartesian form if the projection point were moved to the origin and one of the lines were $y = 1$. As a scheme for mapping lines, this is not a projection of a plane to a line; however it *is* a 1:1 mapping of one line to another. According to Hogben [6], the demonstration compares areas of triangles with common vertices at the point of projection, and bases on one or the other of the two lines being compared.

different ways would allow cancelling the unknown factors while retaining the differences in the points themselves.

So, make up two new matrices,

$$R = \begin{bmatrix} xs & zv \\ s & v \end{bmatrix}, \qquad S = \begin{bmatrix} wu & yt \\ u & t \end{bmatrix}.$$

The combination which we actually want is

$$\begin{aligned} \frac{|PQ|}{|RS|} &= \frac{(x-y)st\ (w-z)uv}{(x-z)sv\ (w-y)ut} \\ &= \frac{(x-y)(w-z)}{(x-z)(w-y)} \\ &= \left(\frac{x-y}{x-z}\right) / \left(\frac{w-y}{w-z}\right), \end{aligned}$$

which is the quotient of quotients of distances from the customary formula to be found in all textbooks.

2.1 Relative distance interpretation

The derivation of the cross ratio given in the last subsection was based on determinants in the plane, according to which it is really a result about areas, likewise in the plane. The drawing for Pappus's construction also uses a plane, although only as a device to illustrate a one-to-one reversible mapping between two lines gotten by drawing lines out from a focus, to see where they intersect a couple of lines.

Confined to the interior of a line, the invariance of the cross ratio tells something about trying to locate a point by specifying its relative distance from a pair of reference points. Under an affine transformation, which would be a combination of dilation and translation, that ought to suffice. By projective transformation, where the dilation is not uniform, it is the ratio of ratios which is invariant. In other words, if it is ρ times as far from w to x as it is from w to y, and it is σ times as far from z to x as it is from z to y (all of this being taken with due regard for sign), then the ratio of σ to ρ will always be the same, even when ρ and σ are not.

To give this a still more concrete interpretation, suppose that w is the midpoint between x and y. It probably isn't still the midpoint after projection, although if two points were halfway between, they would have to move together. Trisectors of an interval, w and z with ratios of $2:1$ and $1:2$ probably won't map into trisectors either, but the quotient 4 would have to be respected. And so on.

To summarize a long series of special cases, observe that although a cross ratio is unaffected by whatever projective transformation, its particular value still depends on the four points chosen. It is a reasonable question, given the cross ratio and three of the points, to ask for the fourth. Put

$$\sigma = \frac{w-z}{w-y}$$

to get, for cross ratio ϕ,

$$\begin{aligned}\phi &= \frac{(x-y)(w-z)}{(x-z)(w-y)} \\ x &= \frac{\phi z - y\sigma}{\phi - \sigma}\end{aligned}$$

The invariance of the cross ratio can be used much less explicitly. Suppose that it is desired to map points x_1, x_2, x_3 into points y_1, y_2, y_3, and to find the consequences for other points. Note that for whatever value of ϕ arising from x, x_1, x_2, x_3, there is always a y determined by the same ϕ and further values y_1, y_2, y_3. Using that common value of the cross ratio and unknown points x and y, we get

$$\begin{aligned}\phi &= \frac{(x-x_1)(x_3-x_2)}{(x-x_2)(x_3-x_1)} \\ \phi &= \frac{(y-y_1)(y_3-y_2)}{(y-y_2)(y_3-x_1)},\end{aligned}$$

so the equation for the mapping would be

$$\frac{(x-x_1)(x_3-x_2)}{(x-x_2)(x_3-x_1)} = \frac{(y-y_1)(y_3-y_2)}{(y-y_2)(y_3-y_1)}. \tag{1}$$

It can be checked by substitution and, if necessary, the use of l'Hospital's rule.

2.2 In a plane

Applying the foregoing ideas to a plane, suppose that three points are represented projectively. Then their determinant is

$$\begin{vmatrix} x_1 p & x_2 q & x_3 r \\ y_1 p & y_2 q & y_3 r \\ p & q & r \end{vmatrix} = pqr \begin{vmatrix} x_1 - x_3 & x_2 - x_3 \\ y_1 - y_3 & y_2 - y_3 \end{vmatrix},$$

which is the area of the triangle $\{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$ calculated vectorially with respect to the third vertex as an origin.

If there were a second triangle,

$$\begin{vmatrix} w_1 s & w_2 t & w_3 u \\ z_1 s & z_2 t & z_3 u \\ s & t & u \end{vmatrix} = stu \begin{vmatrix} w_1 - w_3 & w_2 - w_3 \\ z_1 - z_3 & z_2 - z_3 \end{vmatrix},$$

the quotient of these two determinants would be a three dimensional affine invariant, but not useful as a projective invariant because of the unknown multipliers.

Just as in the two-dimensional version, it should be possible to find a permutation of the six points for which the multipliers cancel, leaving a combination of determinants as an invariant.

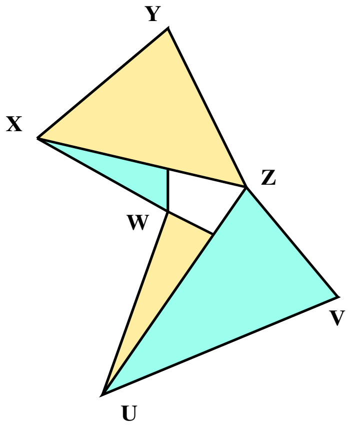


Figure 2: Area version of cross ratio invariance. The ratio of the areas of two triangles is the same before and after exchanging a pair of vertices, remaining so after all projective mappings.

One suggestion would be to exchange the points chosen as origins.

$$\frac{pqr \begin{vmatrix} x_1 - x_3 & x_2 - x_3 \\ y_1 - y_3 & y_2 - y_3 \end{vmatrix}}{str \begin{vmatrix} w_1 - x_3 & w_2 - x_3 \\ z_1 - y_3 & z_2 - y_3 \end{vmatrix}} \frac{stu \begin{vmatrix} w_1 - w_3 & w_2 - w_3 \\ z_1 - z_3 & z_2 - z_3 \end{vmatrix}}{pqu \begin{vmatrix} x_1 - w_3 & x_2 - w_3 \\ y_1 - z_3 & y_2 - z_3 \end{vmatrix}} \qquad (2)$$

A further suggestion would be to notice that of the six points, one stays fixed while the others get shuffled around. Therefore it could be made to coincide with one of the others, giving an invariant depending on five points rather than six, which could be of both symbolic and computational advantage.

2.3 The state of invariant theory

In fact, once the idea has been grasped that quotients of determinants are the fundamental invariants, and that the double quotient arises from removing unobservable quantities, the remainder of the search for invariants is pure routine. Even the symmetry groups of the cross ratios are seen to be a natural consequence of reducing the full permutation group relative to the ambiguity in arranging the cancellation.

The reason this topic got to be included in the category of "summer research" was dissatisfaction with the presentation of the cross ratio in textbooks. So, it was more historical research or a literature search than purely mathematical research, which is not to say that it is less interesting.

Years ago, when the <PLOT> programs were under development in the Institute for Nuclear Energy at Salazar, the convenience of including projective geometry for perspective views and hidden line drawings was apparent, and coincided with the recent publication of Maxwell's two books [4, 5]. Several other texts on projective geometry were available, purchased, and consulted. They all had one feature in common, that the invariance of the cross ratio was a given, which the readers were invited to verify, as an exercise. It can hardly be claimed that the cross ratio was an advanced topic which had been introduced in more elementary books, because many of these selfsame references were the basic texts.

Speculating on how Pappus encountered the invariance has to take into account that it was done long ago, and that whatever surviving details there may be are not in the literature which one has readily at hand, not even in the histories of mathematics. On the other hand, it is easy to surmise that geometers worked with intersections of families of lines, projections and what not, and wondered about such things as whether the midpoint of a line fell at the midpoint of its shadow. Of course, wondering and calculating are two different things.

It is also an interesting question, whether knowing about determinants and the fact that changes of basis multiply volumes by the determinant of the transformation, can be considered either as common or as elementary knowledge. Nevertheless, the idea should be familiar to anyone exposed to a course on linear algebra. So in that sense, it wasn't hard to devise an invariant, leaving open the question of whether that was how the result arose historically.

At this point, we had access to the Internet, so searching for "cross ratio" seemed to be a way to get more information, as did searching Barnes and Noble's online book catalog, for descriptions and conteent listings of books on projective geometry. The somewhat surprising result, for someone who has been out of touch with the field for years, was the degree to which Robotics and Computer Vision have adopted generalizations of the cross ratio as their stock in trade. Even so, without direct and convenient access to all the literature, it is hard to reconstruct even this recent history. That is, Who had the first understanding of the utility of the cross ratio? and From what sources did they obtain this insight?

Given that these fields are of active interest, there have been several conferences dedicated to the subject, and books written. Joseph L. Mundy has been one of the active researchers, and has edited two books [10, 11] which include articles on projective invariants among their contributions. Another good reference is *Pattern Classification and Scene Analysis* by Richard O. Duda and Peter E. Hart [9], dating from 1973.

One of the more interesting results of this literature search concerns the fact that determinantal ideas seem to have originated with A. F. Móbius as long ago as 1827. His book seems to have been reprinted, even to the extent of being listed by Barnes and Noble, but on trying to buy a copy, we found that even the reprint seems to be out of print. But to fix the chronology somewhat, his work antedates the use of matrices and the formalization of linear algebra, but falls in a time frame in which determinants and their uses were reasonably familiar.

Why this would not be mentioned in projective geometry books originating a century later may have had something to do with a change of emphasis, tending toward non-euclidean geometries and the axiomatic foundations of geometry.

3 Putzer's Method

While we were discussing the graphical representation of the 2×2 unimodular matrices as "quaternions" (that is, $\mathbf{i}^2 = -\mathbf{1}$, remaining squares $+\mathbf{1}$) and the variants of Euler's formula for $e^{i\phi}$ in quaternion form, an article appeared in *Journal of Mathematical Physics* [17] purporting to generalize Rodrigues' formula for some particular Lie groups using Putzer's method. This lead to looking up the method, wondering whether the scheme was actually correct, and finally to an understanding of Sylvester's formula for general (meaning non-normal as well as normal) matrices.

3.1 Background

Using the Dirac bra and ket notation for row and column vectors, the spectral theorem for matrices says that

$$M = \sum \lambda_i \frac{\mid i >< i \mid}{< i \mid\mid i >}, \tag{3}$$

for a matrix M with eigenvalues λ_i and with eigenvectors indexed accordingly. In truth this formula is not comprehensive, due to the possibility of vanishing denominators $< i \mid\mid i >$. That never happens for normal matrices, whose eigenvectors can even be normalized so that $< i \mid\mid i > = 1$, avoiding the necessity to include the denominator in the formula at all.

It is a short step from the spectral theorem to Sylvester's formula,

$$f(M) = \sum f(\lambda_i) \frac{\mid i >< i \mid}{< i \mid\mid i >}, \tag{4}$$

which facilitates the construction of functions of M under the most general conditions, such as the existence of a convergent powers series for f.

When a matrix is not normal, it can often be approximated by a parameterized normal matrix whose limit provides a formula involving derivatives of the function as well:

$$f(M) = \sum_{\lambda_i} \sum_{k=0}^{r_i} f^{(k)}(\lambda_i) N_i^{(k)}, \tag{5}$$

The formula can be established algebraically without taking limits and derivatives, although some preparation is required. Using limits illuminates the origin of the idempotent $N_i^{(0)}$ and nilpotent $N_i^{(k)}$, $k > 0$'s, but does not give the preferred derivation because of the wide variety of limits which can result. At the level of the spectral theorem the additional structure in the formula it is equivalent to recognizing the Jordan canonical form, and at the level of Sylvester's formula, treats matrix functions with full generality.

The full development of the Jordan form is considered to be an advanced topic which is avoided in contexts where it is not necessary (quantum mechanics, which thrives on Hermitean matrices, or introductory books on linear algebra), and often surrounded by much datail when it is actually needed. One of the best of these detailed treatments is Gantmacher's two volume presentation of matrix theory [31]. Although the Jordan form is developed very carefully, Sylvester's formula is left at the level of "Sylvester-Lagrange interpolation" and "elementary divisors;" but what is needed is a more explicit representation of the $N_i^{(k)}$'s and their multiplication table.

3.2 The Cayley-Hamilton theorem

The starting point for proving theorems about matrices is to establish the Cayley-Hamilton theorem, and to understand some of its corrolaries, such as theexistence and meaning of eigenvalues and eigenvectors. A good way to do this is to work with the resolvent,

$$\begin{aligned} R(\lambda) &= (\lambda \mathbf{1} - M)^{-1} \\ &= \sum_{i=0}^{n-1} A_i \lambda^i \end{aligned}$$

whose properties depend on the vanishing or nonvanishing of the determinant

$$\begin{aligned} \chi(\lambda) &= |\lambda \mathbf{1} - M| \\ &= \sum_{i=0}^{n-1} c_i \lambda^i \end{aligned}$$

which is well known to be a polynomial in λ with no more roots than the dimension of M, and exactly that number when multiplicity is taken into account. The reason that roots are involved is that eigenvectors, for which $MX = \lambda X$, have to avoid appearing to satisfy $X = (\lambda \mathbf{1} - M)^{-1} Y$ for $Y = 0$. They do so by being annihilated by the inverse of the resolvent, featuring the eigenvalues as locations where the resolvent itself cannot be formed.

The resolvent of any finite matrix has an explicit form which transforms an expression involving powers of λ into one involving powers of the matrix. Looking at the adjugate (which is almost the inverse) of $(\lambda I - M)$, observe that it is a polynomial of degree $n-1$ in λ because that is the maximum dimension of the cofactors and thus the maximum number of λs which could ever be multiplied together. Putting coefficients of λ^i together in a matrix called A_i, set

$$(\lambda I - M)^A = \sum_{i=0}^{n-1} \lambda^i A_i, \tag{6}$$

with a corresponding expansion of the characteristic polynomial

$$\chi(\lambda) = \sum_{i=0}^{n} c_i \lambda^i. \tag{7}$$

Then the equation

$$(\lambda I - M)(\lambda I - M)^A = \chi(\lambda) I$$

could be subjected to a series of transformations

$$\begin{aligned}
(\lambda I - M)\sum_{i=0}^{n-1}\lambda^i A_i &= \sum_{i=0}^{n} c_i\lambda^i I, \\
\sum_{i=0}^{n-1}\lambda^{i+1}A_i - \sum_{i=0}^{n-1}\lambda^i M A_i - \sum_{i=0}^{n} c_i\lambda^i I &= O, \\
\sum_{i=0}^{n}\{A_{i-1} - MA_i - c_i I\}\lambda^i &= O.
\end{aligned}$$

to get a result in which the matrix coefficient of each power of λ would have to vanish. That produces a chain of substitutions:

$$A_{i-1} = MA_i + c_i I. \tag{8}$$

The missing A_{-1}, as well as the nonexistent A_n would both have to be O. Written out in greater detail,

$$\begin{aligned}
A_{n-1} &= c_n I \\
A_{n-2} &= c_n M + c_{n-1} I \\
A_{n-3} &= c_n M^2 + c_{n-1} M + c_{n-2} I \\
&\cdots \\
A_0 &= c_n M^{n-1} + c_{n-1} M^{n-2} + \cdots + c_1 I \\
A_{-1} &= c_n M^n + c_{n-1} M^{n-1} + \cdots + c_1 M + c_0 I.
\end{aligned}$$

All these equations are readily summarized in one single matrix equation,

$$\begin{bmatrix} O \\ A_0 \\ \cdots \\ A_{n-3} \\ A_{n-2} \\ A_{n-1} \end{bmatrix} = \begin{bmatrix} c_0 & c_1 & c_2 & & c_{n-1} & c_n \\ c_1 & c_2 & c_3 & & c_n & . \\ & & & \cdots & & \\ c_{n-2} & c_{n-1} & c_n & & . & . \\ c_{n-1} & c_n & . & & . & . \\ c_n & . & . & & . & . \end{bmatrix} \begin{bmatrix} I \\ M \\ \cdots \\ M^{n-2} \\ M^{n-1} \\ M^n \end{bmatrix}. \tag{9}$$

In fact, $c_n = 1$, so that the last equation of the series asserts that $\chi(M) = O$, which is just the Cayley-Hamilton Theorem: a matrix satisfies its own characteristic equation.

3.3 Reduction of a power series modulo the characteristic polynomial

Of course, the matrix version of an analytic function could just be defined by its power series,

$$f(M) = \sum_{i=0}^{\infty} a_i M^i, \tag{10}$$

as long as the eigenvalues lay within its circle of convergence. The direct definition is unattractive unless the power series converges rapidly or perhaps can be summed symbolically. Usually the matrix would be diagonalized and functions of the eigenvalues calculated, a procedure presenting problems of its own.

Since a matrix satisfies its own characteristic equation,

$$M^n \quad = \quad \sum_{i=0}^{n-1} c_i M^i, \tag{11}$$

according to the Cayley-Hamilton theorem, it might seem that a good way to reduce any matrix function to a polynomial would be to reduce the power series modulo the characteristic polynomial. Although some effort might be required to get the new coefficients, the task of calculating high powers of a matrix could thereby be avoided.

The reduction can actually be performed by using the companion matrix of M,

$$\begin{bmatrix} \mathbf{0} & \mathbf{1} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ . & . & . & \dots & . & . \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{1} \\ c_0 & c_1 & c_2 & \dots & c_{n-2} & c_{n-1} \end{bmatrix} \begin{bmatrix} I \\ M \\ M^2 \\ \dots \\ M^{n-2} \\ M^{n-1} \end{bmatrix} = \begin{bmatrix} M \\ M^2 \\ M^3 \\ \dots \\ M^{n-1} \\ M^n \end{bmatrix} \tag{12}$$

Call C the matrix in this equation, which is a Kronecker product of the companion matrix with the $n \times n$ identity matrix, and V the vector formed by the first few powers of M, which is also a Kronecker product. Since M is an eigenvalue of C with eigenvector V, it follows

$$f(M)V \quad = \quad f(C)V. \tag{13}$$

Consequently, it is only necessary to read off the first row of $f(C)V$ to get $f(M)$ as a linear combination of the low powers of M. However, this result still requires summing a whole power series, so it may offer no advantages as a numerical procedure. The essence of Putzer's method, which is dedicated to calculating matrix exponentials, lies in obtaining e^{Ct} as the solution of a differential equation, then deducing e^{Mt}. While a meritorious route to solving to this particular application, the unwary might not notice that the Jordan decompositions for M and C need not coincide, and yet the method still yields a true result.

The companion matrix C is not a canonical form, because its Jordan blocks have dimension equal to the multiplicity of each eigenvalue, whereas those of M may decompose into smaller blocks in any way which is consistent with the overall multiplicity. They keep the same symmetric functions and hence the same companion matrix, even though equivalence through a change of basis would want to write the companion matrix as a direct sum of smaller companion matrices. On the other hand, C maximizes any other arrangement by containing all the other Jordan subspaces, which is why the numerical methods work; nilpotents in the full Sylvester's formula will all have expired by the time their presence could have caused any interference.

In subsequent articles some commentaries on the method were published, but the general context of canonical forms seems to have been overlooked in the pursuit of minor improvements and extensions. Since the coefficients of the companion matrix are symmetric functions of the eigenvalues of M, it might seem that the eigenvalues could be put to better use by just evaluating f rather than starting up a whole new procedure. However, all that can be avoided by observing that the symmetric functions can be obtained directly by using the traces of low powers.

Since those matrices are going to be used anyway, taking their traces represents little additional labor. There is a series of identities relating power sums to symmetric functions due to Newton,

$$\left.\begin{array}{rcl} c_{n-1} & = & -\text{Trace }(M), \\ 2c_{n-2} & = & \text{Trace }(M^2) - (\text{Trace }(M))^2 \\ 6c_{n-3} & = & (\text{Trace }(M))^3 + 2\text{ Trace }(M^3) - 3\text{ Trace }(M)\text{ Trace }(M^2) \\ \cdots & & , \end{array}\right\} \tag{14}$$

which could be used. The whole procedure has been consolidated into a single recursion by Faddeev (see [31] vol. 1, ch. IV, §5; [22]), by defining the recursive chain

$$\left.\begin{array}{rclrclrcl} N_1 & = & M & c_{n-1} & = & \text{Trace }(N_1) & A_{n-1} & = & N_1 - c_{n-1}I \\ N_2 & = & MA_{n-1} & c_{n-2} & = & \frac{1}{2}\text{ Trace }(N_2) & A_{n-2} & = & N_2 - c_{n-2}I \\ & \cdots & & & \cdots & & & \cdots & \\ N_{n-1} & = & MA_1 & c_1 & = & \frac{1}{n-1}\text{ Trace }(N_{n-1}) & A_1 & = & N_{n-1} - c_1 I \\ N_n & = & MA_1 & c_0 & = & \frac{1}{n}\text{ Trace }(N_n) & A_0 & = & N_n - c_0 I \end{array}\right\} \tag{15}$$

In other words, the symmetric functions of the roots, which are the coefficients in the characteristic polynomial, can be read off right away as traces of the auxiliary matrices N_i.

3.4 Sylvester's formula

It still remains to describe the transition from the reduction of the power series for f modulo the characteristic polynomial to Sylvester's formula. To do that it is convenient to factor the characteristic polynomial

$$\frac{\chi(\lambda)}{\chi(\lambda)} = 1 \tag{16}$$

in which the quotient $1/\chi(\lambda)$ is first resolved into partial fractions, and then multiplied by the numerator $\chi(\lambda)$ to get a sum of terms G_i which are characterized as arising from that part of the partial fraction depending upon the denominator $(\lambda - \lambda_i)$. If λ_i is simple,

$$G_i = \prod_{j\neq i}(\lambda - \lambda_j), \tag{17}$$

otherwise it might be multiplied by a polynomial in $(\lambda - \lambda_i)$ whose degree would be less than the multiplicity r_i. The point is that the $G_i(\lambda)$ resolve the identity, and so do their matrix counterparts:

$$\sum_i G_i(M) = \mathbf{1}. \tag{18}$$

Nevertheless each of them can be multiplied by a power of $(\lambda - \lambda_i)$ sufficient to create the characteristic polynomial, which vanishes for the matrix M, meaning that for some p,

$$(M - \lambda_i\mathbf{1})((M - \lambda_i\mathbf{1})^{p-1}G_i(M) = \mathbf{0} \tag{19}$$

$$M\left(((M - \lambda_i\mathbf{1})^{p-1}G_i(M)\right) = \lambda_i\left(((M - \lambda_i\mathbf{1})^{p-1}G_i(M)\right). \tag{20}$$

If this resolution of the identity is applied to the power series representation of a function, there results

$$f(M)\,\mathbf{1} \quad = \quad \sum_{n=0}^{\infty}\sum_{i}\frac{f^{(n)}(0)}{n!}M^n G_i(M). \tag{21}$$

After changing the order of summation, it would be preferable to have the expansion of f about each eigenvalue, to be consistent with the operation of each G_i on the terms of the series. The shift is accomplished by writing

$$f(\lambda_i + \lambda - \lambda_i) \quad = \quad \sum_{n=0}^{\infty}\frac{f^{(n)}(\lambda_i)}{n!}(\lambda - \lambda_i)^n; \tag{22}$$

whose immediate matricial counterpart is

$$f(M) \quad = \quad \sum_{n=0}^{\infty}\frac{f^{(n)}(\lambda_i)}{n!}(M - \lambda_i I)^n \tag{23}$$

$$\quad = \quad \sum_{n=0}^{m_i}\frac{f^{(n)}(\lambda_i)}{n!}(M - \lambda_i I)^n. \tag{24}$$

The series at each eigenvalue has been truncated, according to the power of $(M - \lambda_i I)$ needed as a multiplier of $G_i(M)$ to complete the characteristic polynomial. But this is nothing more than a derivation of equation 5, in which the derivatives of f replace the combinations of coefficients of f and elements of powers of the companion matrix.

The fact that the G_i omit terms in λ_i but have the full complement of remaining factors from χ ensures that $G_i G_j = 0$ for $i \neq j$; this together with the fact that the G_i resolve the identity ensures $G_i^2 = G_1$. We therefore have a set of commuting orthogonal idempotents, which are projectors onto the eigenvalue subspaces. We can even claim that

$$N_i^{(k)} \quad = \quad (M - \lambda_i I)^k G_i \tag{25}$$

are the nilpotents of equation 5, and that they satisfy

$$N_i^{(k)} N_i^{(\ell)} \quad = \quad N_i^{(k+\ell)} \tag{26}$$

The rows and columns of the matrices $N_i^{(k)}$ are eigenvectors and members of the Jordan chains by which the canonical form of M may be exhibited. They still have to be sorted out when there are several linearly independent eigenvectors belonging to the same eigenvalue, something which can be done with the help of the minimal polynomial. An unexpected aspect of Putzer's method lies in its avoidance of all those details on account of the fact that the general Sylvester's formula only requires the idempotents and nilpotents, without needing to identify the vectors in a Jordan basis explicitly.

3.5 A chain of references

Having outlined the theory behind the canonical form and Sylvester's formula, we review the series of articles which attracted out attention during the summer program (and beyond).

3.5.1 Leite and Crouch

The issue of the journal containing this article [18] arrived rather fortuitously at a time when a question about Rodriques' formula had been raised in an Internet discussion group, and the daily meetings of the summer program were discussing the representation of rotations, mainly because of the quaternion representation of matrices, its adaptation to unimodular matrices, and applications to vibrating chains, the solution of second order differential equations, and similar topics which have been a continuing source of interest.

The title of the article calls for the symbolic calculation of matrix exponentials, which was pertinent to our interests, using "Putzer's method," which was unfamiliar. Reading the article, it looked rather awkward for several reasons. It did not seem to fit smoothly into any physical publication sequence, cited a handful of references, and invoked *Heisenberg's* theory of the charge independence of nuclear forces. That may be correct, but one has the general feeling that the innovation was due to *Pauli*. We lack the immediate bibliographical resources to decide the issue, which in any event was only a minor allusion in the paper.

There was an evident proofreading error whereby a certain Lie group was assigned dimension 12 one place and 21 another. There was also a mention of "Rodrigues' formula" but it is not clear whether the reference was to Olinde Rodrigues (most likely) to whom the formula for the three dimensional rotation group is ascribed, or to a contemporary author named W. Rodrigues, Jr., coauthor of one of the articles cited in the bibliography. Generally, we do not have access to many of the articles cited by [17], to evaluate the doubt.

Finally, a referee is thanked for a reference to Putzer, which raises the question of whether the methods of the paper were somehow worked up, and then the title adjusted once the authors were apprised of the previous art. The reason for such a concern becomes evident when one goes to find out what Putzer's method is, and why it should be efficacious in evaluating Lie exponents. The article comments on the accessibility of the method. There is also some confusion, because Putzer's method is supposed to avoid eigenvalues and Jordan forms, yet attention is given to obtaining them as a preliminary to calculating the exponential, or maybe even to get the symmetric functions which are the coefficients of the charactristic polynomial. Some trace formulas are presented, but it is not mentioned whether those are the ones which go back to Newton.

Altogether, it is hard to render a judgement on reference [17] because of the inaccessibility to us of the majority of its references, upon which one usually depends to get the full benefit and understanding from reading a scientific article. Two important references which were available in our own library were to Putzer himself, and to the survey by Moler and van Loan.

3.5.2 Putzer

The central reference in this whole discussion was published by E. J. Putzer [18] in the *American Mathematical Monthly* in 1966 while he was working for an aircraft manufacturer. The title of the article, "Avoiding the Jordan canonical form in the discussion of linear systems with constant coefficients," states its intention quite concisely and in retrospect, the paper may have been composed more from an engineering background than one in physics and mathematics. The reason for this suspicion is that engineers are much more familiar with nonnormal matrices, which are never encountered when physicists use hermitean operators.

His paper is confined to the matrix exponential, making its point by showing that the solution of a differential equation of a certain order actually generates the required exponential. It is extremely easy to read too much into this procedure, which was the source of several months' confusion. Namely, the differential equation which is solved is the one worthy of the exponential of the companion matrix, not the matrix itself. The missed point: the first result is not in itself the desired exponential, but part of a two stage process which first generates the coefficients for the power series representation of the actual representation. The second step can include a matrix inversion and invites some subsidiary discussion; but it is important to realize that there are two steps, and that the Jordan normal form in the companion matrix majorizes the Jordan normal forms for all other matrices having the same symmetric functions.

3.5.3 Moler and van Loan

That the exponential is the function which Putzer's method computes is largely irrelevant. He probably had a need for that particular function, for which the differential equation allowed a self-consistent presentation of the method. However, the survey article written by Cleve Moler and Charles van Loan [19], "Nineteen dubious ways to compute the exponential of a matrix," concentrates exclusively on the exponential function while examining the strengths and weaknesses of a goodly number (19, to be exact) of schemes. Of course, one of them was the one which Putzer had published twelve years earlier. Given that differential equation integrators, using Runge-Kutta methods and others, show good performance, and that Putzer's process implies a sparse matrix, it counts as a viable way to get a matrix exponential.

The title of the article reflects the computers and program libraries available at the time it was written; the conclusion reached was that all of some seven categories of programs had strengths as well as weaknesses, leaving the choice of a particular method to depend on circumstances – none was clearly and unfailingly superior to all the others.

3.5.4 Kirchner

Shortly after Putzer's article appeared, and in direct reference to it, R. B. Kirchner's article [20] "An explicit formula for e^{At}," defined some interpolation polynomials and showed that they could be used to get Putzer's result, all the while satisfying the differential equation for an exponential. He defined interpolation polynomials

$$p_j(\lambda) \quad = \quad \prod_{i=0, i\neq j}^{n} (\lambda - \lambda_i)^{s_i} \tag{27}$$

together with their sum,

$$q(\lambda) \quad = \quad p_1(\lambda) + \cdots + P_k(\lambda) \tag{28}$$

and claimed that since $q(\lambda)$ has no roots in common with the characteristic polynomial, then by using the the Euclidean algorithm on the combination of q and characteristic polynomial, the matrix $q(A)$ is invertible. In other words, q is not necessarily a resolution of the identity, but it is the next best thing and can be used to derive an analogue of Sylvester's formula,

$$q(A)e^{AT} \quad = \quad \sum_{i=1}^{k} p_i(A) f_{s_i}((A - I\lambda_i)t)e^{\lambda_i t}. \tag{29}$$

The functions f_{s_i} are just the leading terms in the power series expansion for e^X and contribute the derivatives already seen in Sylvester's formula. In fact, the main difference between Kirchner's presentation and the one given in our own report lies in his use of powers to define the interpolation polynomials whilst our definition derived from a partial fraction decomposition and avoided the intervention of $q(A)$ in the formula.

This difference in approach is akin to the fololowing choice. Suppose that a confluent form of Lagrange interpolation is to be used at two points. The basis ploynomials at either point can consist of powere, or something else. Likewise the dual basis at the two points consist of derivatives or something else. Depending on the preference for derivatives or for powers at each point, four different interpolation formulas can result.

3.5.5 Leonard, Maki, and three others

In a "Classroom note" published in *Siam Review* in 1996, I. E. Leonard revived Putzer's method on the basis that students who knew some differential equation theory could obtain a matrix exponential without being encumbered with any excessive knowledge of linear algebra – just eigenvalues and maybe the Cayley-Hamilton throrem.

The article produced a quick reaction whose results finally appeared in print two years later. The editor [25] introduced three new notes by commenting that the Note of Eduardo Liz [24] showed that one of Leonard's matrices was a Wronskian, the Note of M. Kwapisz uses Putzer's method to calculate the power of a matrix by exhibiting the power as the solution of a finite difference equation rather than Putzer's differential equation for the exponential, and Shui-Hung Hou's Note [22] was a reminder that the relation between traces and symmetric polynomials is a byproduct of the formula for the resolvent.

3.5.6 and still more

If one's interest lies in solving differential equations or in calculating matrix exponentials, there appears to be a wealth of material.

Besides the references already described, we could mention a more complete follow-on to the articles of Putzer, Kirchner, and Leonard which was published by Saber N. Elaydi and William A. Harris, Jr. in 1998 [28] "On the computation of A^N." It explicitly mentions the resolution of the identity via partial fraction decomposition of equation W, and uses the Casorati matrix, which replaces Luz' Wronskian in the context of finite differences.

In 1975 Edward P. Fulmer [29] published "Computation of the matrix exponential," citing Kirchner as well as a still earlier reference, and showing how confluent Vandermonde matrices can be involved in the solution. In the article immediately following, "Explicit formulas for solutions of the second-order matrix differential equation $Y'' = AY$," Tom M. Apostol, [30] adapts the calculation to this frequently occurring equation.

Eugene Garfield's *Institute for Scientific Information* not only publishes the *Citation Index*, but offers to conduct searches and make reference trees for keywords selected by their clients. One could probably have an interesting time tracing down such keys as *Putzer's method* or *matrix exponential*, but the literature which we have encountered in our own library is probably a good sampling of the kinds of articles and their contents.

The subject is a popular one, and seems to be getting to be more so, judging by Java applets to be found on the Internet, and its inclusion in symbolic algebra systems such as Macsyma, Maple, or Mathematica. Looking in the opposite direction, one wonders what there is to be found in the literature of the first half of the century. The standard reference on matrix theory was Frazer, Duncan and Collar's *Elementary Matrices*, whose very title indicates an inclusion of general, and not just normal, matrices. However, it is only a fond memory, seeming to be both out of print and not to be found in the libraries to which we have access.

Those authors were aeronautical engineers and concerned with mechanical vibrations. Engineering problems are prone to encounter the Jordan form; one has only to recall that the case of critical damping is the important one for the vibration of mechanical systems. That means that one should probably look to the engineering, rather than quantum mechanical, literature for sympathetic discussions of matrices in their full generality (notwithstanding the fact that in parameter space, the "bad" matrices have measure zero). In that direction, a series of books by I. Gohberg (and in particular the latest [26], which builds on its predecessors), which concentrates on finite polynomials involving finite matrices, is worth consulting.

Of course, there are more traditional books on matrix theory, of which a classic is surely F. R. Gantmacher's *The Theory of Matrices*, [31]. Also worthy of note is A. I. Mal'cev's *Foundations of Linear Algebra*, [27]. They give very careful and exhaustive analyses of the normal forms. Unfortunatley it is just that attention to detail which has alienated the authors of all these other papers which wee have cited, which in turn, by concentrating exclusively on the matrix exponential, have obscured a much more general result.

4 Chaté-Manneville Automata

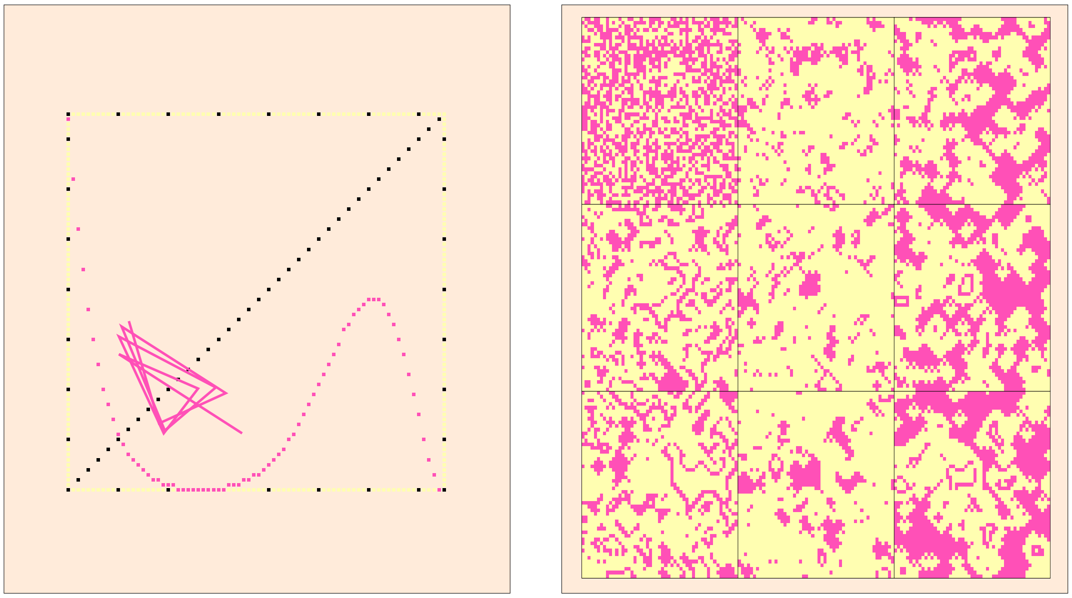


Figure 3: The density of states in Chaté-Manneville automata runs through cycles rather than approaching an equilibrium value. One of the simplest is the binary totalistic rule 33 in three dimensions with von Neumann neighborhoods. Left: a plane cross section of the evolution of a random initial configuration for nine generations. Right: the return map, which does not exactly follow the mean field density curve. In the background the diagonal can be seen as well as the mean field curve with an unstable equilibrium density of about 0.1.

When statistical methods began to be applied to the evolution of cellular automata, it was understood that there was a difference between periodic configurations for the automaton and periodicities in the statistical behavior of the automaton.

The long time behavior of a single configuration is one thing; it may repeat sooner or later exhibiting periodic behavior, although if the evolution consists of a neverending transient, it may still have convergent statistical properties. But it is another thing to examine statistical behavior averaged over all configurations, or all the members of a class of configurations. It has generally been expected that those averages would converge, and there was even an article published arguing to that effect. Shortly thereafter Hugues Chaté and Paul Manneville [71, 59] discovered some automata which, although they didn't exactly follow the return map of iteration theory, didn't tend toward a long term average either. The most prominent of them, discovered shortly after their original announcement by Jan Hemmingsson, was only three dimensional and followed a period of three in the density.

4.1 Mapping the mean field probabilities

4.2 Idempotency as a periodicity criterion

5 Reversible Automata

Reversible cellular automata are not only an interesting subject of study; they have practical applications in coding theory and cryptography. The classical study was published by Gustav A. Hedlund in 1969 [43], in the context of symbolic dynamics rather than automata theory. The paper contains several results, three of which are of central importance for cellular automata:

i) the transition rules of cellular automata are the continuous, shift-invariant mappings of the full shift,

ii) in order for every finite configuration to have at least one counterimage, all configurations must have the same number of counterimages, and

iii) there are indices multiplicative with respect to composition (the Welch indices) which characterize the reversibility of infinite configurations.

Recently Jarkko Kari [46] has published some articles characterizing reversible automata as those arising from block permutations, which is a significant advance in characterizing and classifying the reversible automata.

The first result practically defines celllular automata, although that was not the historical course of events. The others merit further discussion.

5.1 Uniform multilplicity

Strictly speaking, a *de Bruijn diagram* describes the sequences which are available in a shift register, one motivation for their study being the regognition of maximum length non-repeating sequences. The Hamiltonian paths which they contain serve that purpose quite nicely. However, it was realized that a variety of link labellings produced useful results in the theory of one dimensional cellular automara.

On the one hand, links can be taken to be neighborhoods. By using the states into which the neighborhoods map as alternative labels, paths through the diagram can be considered to generate the evolution of a configuration. Cellwise, evolution maps several states – the contents of the whole neighborhood – into a single state, but by always taking a neighbor from a fixed location (say the central element, if its length is odd) the mapping can be envisioned as taking place from cell to cell.

Shift-periodic configurations can be identified by Boolean labelling, *yes* or *no* (or more formally, `true` or `false`), according to whether the desired symmetry is realized. For example, left shift in a three neighbor automaton, given the evolution function φ, corresponds to the predicate:

$$p_{ls}(a,b,c) \quad \equiv \quad \varphi(a,b,c) = c. \tag{30}$$

Any path in the diagram following `true` labels will create a shifting configuration by simply recording the sequence of neighborhoods. Such labellings can be applied to the evolution generation after generation, while others will have a more limited duration; still, any predicate at all can be introduced to detect some property or other. Another in common use yields counterimages of constant configurations; given state s, take

$$p_s(a,b,c) \quad \equiv \quad \phi(a,b,c) = s. \tag{31}$$

Thinking that de Bruijn diagrams represent the evolution of automata works in reverse, because by following s sequence of evolved states, one knows the sequence of neighborhoods from which they arose. Finding a given path is equivalent to discovering the ancestor, but it is necessary to look around to find it. Sometimes there is none, when the configuration is part of the Garden of Eden.

The only way to find out whether there are paths of a given description is to make a systematic search. One way to do that is to use a *subset diagram*.

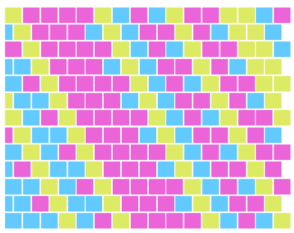
The idea of subset diagrams has a place in the theory of automata which ought to interest a historian, but suffice it to say that it is a device for conducting searches, creating a covering space for turning non-functions into functions, and much more. There are two subset diagrams, according their derivation from incoming links in the de Bruijn diagram, or outgoing links.

In the development of symbolic dynamics following Hedlund's ideas, Masakazu Nasu [49] showed that there was a homomorphic image of the de Bruijn diagram in the subset diagram, and that under a proper labelling of the states, consisted of an assemblage of trees. It is convenient to call this portion of the subset diagram a *Welch diagram*. On the other hand, these diagrams are based on the properties of maximal extensions, which were already studied by Hedlund which he attributed to L. R. Welch.

5.2 Diagrams

Having decided to represent cellular automata with the aid of paths in a diagram, not only is it found that there are several kinds of diagrams to work with, but that the properties of the diagrams are reflected in their topological connection matrices. In that respect, the study of cellular automata turns into an exercise in linear algebra, at least to the extent that such machinery as eigenvalues, eigenvectors, canonical forms and the representation of mappings by matrices is concerned.

In order to fix our attention on a concrete example, consider the reversible (3,1/2) cellular automaton numbered, in Wolfram's scheme, 14937.



	0	1	2
0	0	2	0
1	1	1	1
2	2	0	2

Figure 4: Left: Sample of the evolution of (3,1/2) automaton 14937, which generally consists of a right shift, with cyclic permutations occurring along the margins separating quiescent bands. Right: Matrix for the evolution rule: row label - left cell, column label - right cell, body - evolved cell.

Figure 4 shows a typical portion of an evolutionary sequence. The rule has been chosen to make every state quiescent, an option that is always available for reversible automata, making the remainder of the evolution stand out more clearly. Total quiescence means that the states run in order along the diagonal of the evolution matrix, which is shown over to the right in the same figure.

Uniform multiplicity is evident in the matrix when the number of times that a state is listed there is the same for all states; in this case three times each. That works out to once per row and once per column, but this example shows it is only by the average, and not always in detail.

Examination of the mean field probabilities is a standard procedure in the assay of any new automaton. For reversible automata, one suspects that concentrations of one state or another

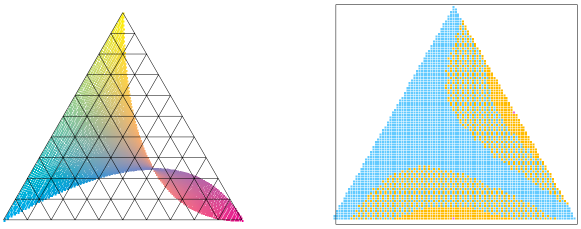


Figure 5: Mean field probabilities for one generation of evolution. Left: each set of probabilities $\pi = (p, q, r)$ is mapped into a new set $\pi' = (p', q', r')$ which is plotted at point π' with coloration π. Right: Contours of the vector norm $\| \pi - \pi' \|$ are shown as a function of π. Note that the probabilities don't change much.

will never be seen because that would go against the uniform multiplicity of ancestors. However, generation to generation exchanges of probabilities certainly occur because permutations, practically be definition, will not upset the balance of ancestors.

In the left portion of Figure 5 in which states 0 and 1 lie along the base in that order from left to right, state 2 sits at the apex. The rule tends to exchange 0 and 2, leaving 1 fixed. In the right half of the figure, a contour plot is shown in which fixed points of probability would manifest as zeroes. Without having some other rules available for comparison, it is only slightly evident that there is an extended saddle point encompassing the whole of the triangle of actual probabilities.

5.2.1 de Bruijn diagram

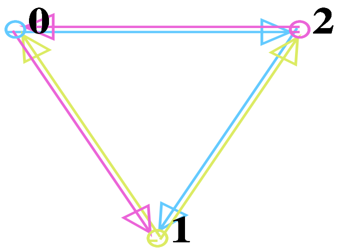


Figure 6: The de Bruijn diagram for (3,1/2) automata has three nodes and nine links. The self-links at quiescent nodes in this diagram are inconspicuous.

The basic objects are labelled de Bruijn diagrams, whose properties have been studied both numerically and symbolically. One of the first things to be done with the evolution-labelled diagram is to decompose its matrix into a sum of matrices according to the evolution. The uniform multiplicity principle will require each one of these fragments to be stochastic, as well as requiring minimum variance for the second moment matrix.

Consultation of the evolution table shown in Figure 4 reveals three de Bruijn fragments,

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \tag{32}$$

whose sum is the de Bruijn matrix:

$$D = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \tag{33}$$

which satisfies $D^2 = 3D$. This is a typical result for de Bruijn matrices because in a shift register of lemgth n, there is exactly one way to replace one sequence of n symbols by another; any sequence at all can be so transformed.

The table of powers for the individual fragments is

$$\begin{array}{c|ccc}
\text{power} & A & B & C \\ \hline
1 & \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} \\
2 & \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \\
3 & \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}
\end{array},$$

each of which satisfies $M^3 = M^2$. B even satisfies $B^2 = B$. Each matrix is idempotent, which is surprising at first sight, yet consistent with the rule of uniform multiplicity.

The fact that all these matrices are idempotent, a very suggestive relation which was noticed empirically while working out some particular examples. Inasmuch as it is not a common property of matrices, but central to understanding reversibility, it is worth exploring its origins in the algebraic properties of the connection matrices.

We can begin with one immediate consequence of the uniform multiplicity theorem by introducing the vector $<u|$ defined by

$$<u| = \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} \tag{34}$$

and recalling that the number of 1's in the de Bruijn fragments is the number of ancestors, which ought to be the same in each fragment. Products describe the ancestors of sequences of states corresponding to the choice of fragments that were multiplied together, within which the sum of the elements still represents the number of ancestors, always the same constant. The ancestor count for M has the simple matrix equivalent,

$$\nu = <u|M|u> \tag{35}$$

. Suppose now that we have the matrix form of the characteristic polyn0mial of one of these matrices,

$$\chi(M) \quad = \quad M^n + c_1 M^{n-1} + c_2 M^{n-2} + \ldots + c_n \mathbf{1} \tag{36}$$

. By forming the ancestor counting inner product, we conclude

$$\begin{aligned} <u|\chi(M)|u> &= <u|M^n|u> + c_1 <u|M^{n-1}|u> + c_2 <u|M^{n-2}|u> + \ldots + c_n <u|\mathbf{1}|u> && (37) \\ &= \nu\ (1 + c_1 + c_2 + \ldots + c_n) && (38) \\ &= \nu\ \chi(1). && (39) \end{aligned}$$

From this we conclude that any connection matrix obeying uniform multiplicity necessarily has 1 as an eigenvalue.

Common sense argues that the eigenvalues should not be greated than 1, just because small matrices cannot have large eigenvalues. If a connection matrix had an eigenvalue greater than 1 in absolute value, its powers would have ever greater eigenvalues, and eventually require at least one large Gerschgorin disk to hold them. That would imply some large matrix elements to provide an ample radius, which would go against the fact that the connection matrix only has positive (or zero) integer elements, and their sum is fixed, whatever the power, at the uniform value of the multiplicity.

There is then a question of possible eigenvalues in the range between zero and one, for which the answer seems to be that between these limits it would be hard to maintain the constancy of the ancestor count with all the different combinations of powers and eigenvalues that would be found. Beginning with Sylvester's formula in its full generality,

$$f(M) \quad = \quad \sum_{\text{eigenvalues}} \sum_{\text{multiplicity}} \frac{1}{j!} f^{(j)}(\lambda_i) N_i^j \tag{40}$$

the formula for the ancestor count of a repeating sequence,

$$<u|M^n|u> \quad = \quad \sum_{\text{eigenvalues}} \sum_{\text{multiplicity}} \frac{1}{j!} \lambda_i^{n-j} <u|N_i^j|u> \tag{41}$$

By varying the power n, we have a series of equations which relate the inner product of a constant vector whose components are the matrix elements of the idempotents and nilpotents, and a series of vectors whose components depend on varying powers of a fixed set of eigenvectors. Up to a certain point such a relationship may be feasible, but if it continues there must be restrictions which need to be examined. Of course, if different powers of the λ's were equal, only possible for $\lambda = 0$ or $\lambda = 1$, there would be no problem, and all such matrices would be idempotent or nilpotent.

5.2.2 the subset diagram

The vertices in the subset diagram are subsets of de Bruijn vertices. Each source subset is linked to the union of all the destinations of its elements. If there are no destinations for a given label, linkage goes to the empty set.

It is useful to introduce some symbolism to express these ideas, anticipating the eventual complexity of the constructions.

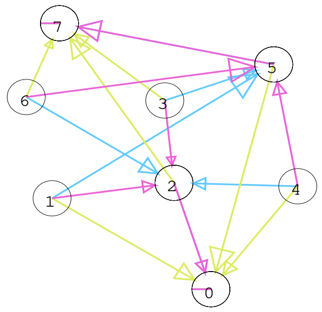


Figure 7: Since the de Bruijn diagram of a (3,1/2) automaton has three vertices, the subset diagram has eight vertices, ranked at four levels according to the size of the subset.

- link(i,j) A predicate, true when there is a link *from* i *to* j, a phraseology which inverts colloquial english usage somewhat.
- link(i) The set of all nodes to which the node i is linked:

$$\text{link}(i) \quad = \quad \{j|\text{link}(i,j)\} \tag{42}$$

- LINK(S) The set of all nodes in the de Bruijn diagram to which nodes in from the subset S are linked:

$$\text{LINK(S)} \quad = \quad \bigcup_{i \varepsilon S} \text{link}(i) \tag{43}$$

  Also note the interchangeability of points and their unit classes, lapsing from strict set theoretic syntax.

Having made these prearrangements, it is easier to connect the nodes in the subset diagram, using the word *LINK* instead of *link* because of the profusion of linkages that have to be dealt with. When suitable graphic facilities exist, such as when drawing the diagrams by hand, it is convenient to represent *link* linkages by thin arrows running between points, and *LINK* linkages by fat arrows running between subsets.

That is the reason for introducing a case-sensitive nomenclature which is more common in computer programming than in mathematical expositions. Following these precepts, it is time to define the linkage predicate for the subset diagram:

$$\text{LINK}(U,V) \quad \equiv \quad V = \bigcup_{u \varepsilon U} \text{NEXT}(u) \tag{44}$$

A detail which should be observed in this definition, which to a certain extent captures the essence of the subset diagram, is that there is one and only one set which can be called LINK(U).

There is *one* because in the worst case it is the empty set, and *only one* because of he insistence that it is the maximum set to which interelement connections via $NEXT$ can be formed. Consequently,

$$\mathrm{LINK}(U) \quad = \quad \bigcup_{u \varepsilon U} \mathrm{link}(u) \tag{45}$$

Leaves in the subset diagram have no incoming arrows, which could be a consequence of all the points in that node being leaves themselves; but it could also result from a lack of saturation. In other words, its points could have incoming arrows, emanating from sources which have additional destinations lying in some other subset.

Rootlets would have no outgoing links, which is impossible because such nodes would be connected to the empty set. But it is still possible to ask what kind has the empty set as its destination (which would have to be unique). Evidently unions of rootlets would be rootlets, just as unions of leaves would be leaves. Unlike the situation with leaves, however, there are no fat rootlets composed of anything other than thin rootlets.

5.2.3 the pair diagram

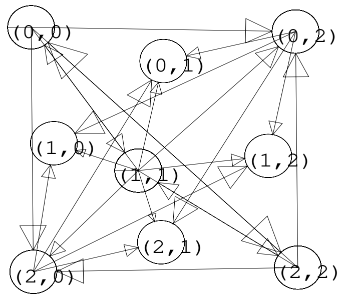


Figure 8: The de Bruijn diagram of a (3,1/2) automaton has three vertices, so the pair diagram for ordered pairs has 9 vertices. Sometimes it is useful to draw the pair diagram for unordered pairs.

Vertices in the pair diagram are pairs of vertices from the base diagram. Vertices are linked if all matching members of the pairs are linked, and with the same label, at that. Similar definitions apply to triple diagrams and so on. The connection matrix of the pair diagram is the second moment matrix of the de Bruijn fragments.

5.2.4 the Welch diagram

A Welch diagram is a subdiagram of the subset diagram, characterized for surjective automata as consisting of the largest subsets accessible from the unit classes. The common order of all its subsets is the *Welch index* for that automaton, left or right according to the handedness of the subset diagram. For any given label, the links bearing that label form trees when all states are quiescent.

The Welch diagram is important because it is an image of the de Bruijn diagram within the subset diagram. By image is meant that the same paths are encountered irrespective of the diagram, so let us suppose that M and N are the connection matrices of the two diagrams, and that L is a mapping from one diagram to the other. The requirement is then

$$ML = LN. \tag{46}$$

A good candidate for L is the membership matrix; the rows of L are indexed by nodes in the de Bruijn diagram, the columns in turn being indexed by subsets in the subset diagram. Consequently L is an $n \times 2^n$ matrix whose elements are the predicates

$$L_{i,S} \equiv i \varepsilon S \tag{47}$$

The equivalence of M and N then depends upon

$$\sum_S link(R,S)\ \&\ j\varepsilon S = \sum_k j\varepsilon S\ \&\ link(j,k) \tag{48}$$

5.3 The matrices

The symbolic connection matrix of the subset diagram is

	{012}	{01}	{02}	{12}	{0}	{1}	{0}	{}
{012}	01	.	.	.	.	.	.	.
{01}	.	.	0	.	.	.	.	.
{02}	0	.	.	.	.	.	.	.
{12}	.	.	.	.	.	0	.	.
{0}	.	.	0	.	.	.	.	.
{1}	.	.	.	.	.	.	.	0
{2}	.	.	.	.	.	.	0	.
{}	.	.	.	.	.	.	.	012

The connection matrix of the pair diagram is

	(00)	(01)	(02)	(10)	(11)	(12)	(20)	(21)	(22)
(00)	.	.	.	.	.	.	.		
(01)	.	.	.	.	.	.	.		
(02)	.	.	.	.	.	.	.		
(10)	.	.	.	.	.	.	.		
(11)	.	.	.	.	.	.	.		
(12)	.	.	.	.	.	.	.		
(20)	.	.	.	.	.	.	.		
(21)	.	.	.	.	.	.	.		
(22)	.	.	.	.	.	.	.		

5.4 Block permutations

6 Rule 110

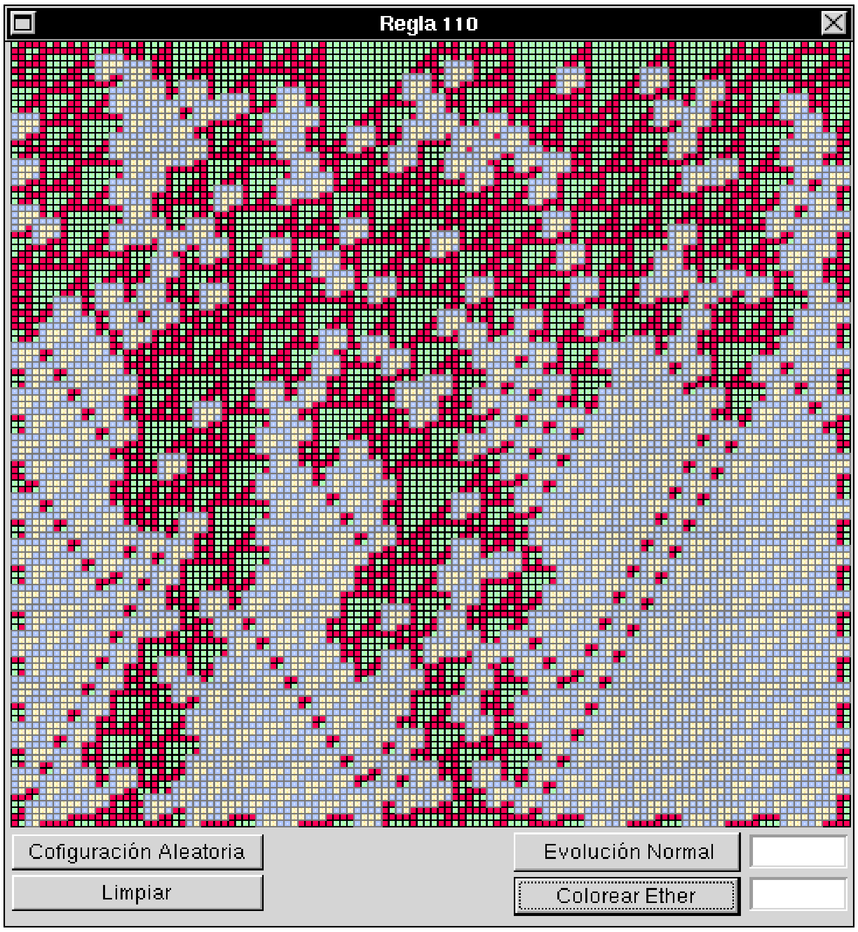


Figure 9: Rule 110 can be regarded as an exercise in tiling the plane with upside down right triangles, whose faults and dislocations are the gliders of cellular automaton theory. So far numerous gliders are known and many of their collisions have been studied.

When Stephen Wolfram surveyed the simplest one dimensional cellular automata and announced his four Classes, it seemed that there were none of the prized Class IV automata amongst the binary, first-neighbor automata. However, he thought that he saw possibilities in Rule 110, so much so that he dedicated Appendix 15 of his book [53] to some of its properties, including a collection of gliders. The rule attracted attention for other reasons, but little seems to have been done with it for a decade. Then, in the fall of 1998 at a meeting in Santa Fe, Matthew Cook is reported to have announced that the rule was computation universal.

There is little additional information concerning this announcement. Cook had maintained in Internet Web page dedicated to Rule 110, which was subsequently transferred to Hungary [88] and then retired. His presentation is scheduled for publication in the *Proceedings* of the CA98 meeting; further information may be available in Wolfram's forthcoming *A New Science* book.

There is some sporadic information of the Internet, with very little actually published in the literature. Li and Nordahl [91] were more interested interested in the long term state density in their 1992 paper, but Lindgren and Nordahl [92] were already concerned with the possibilities of universal computation in an article from 1990. They cautioned that one must be sure that it is the cellular automaton which is performing the computation, and that it is not the process by which its initial configuration is prepared which embodies the computation.

6.1 Subset diagram

Why should Rule 110 and its equivalents end up being so interesting? Of course Wolfram picked it out of his rogue's gallery of typical (3,1) evolution on the grounds of having evident structures, but they were messy enough, the more so against a textured background, that it would seem only the faithful took them seriously.

One evident feature which can be verified is the existence of membranes and macrocells. The membrane is the vertical spine where one or more triangles is stacked vertically, which is never violated until the quiescent space on its right disappears. Macrocells comprise the space trapped between the spines, but there are evanascent because their protection eventually collapses, sooner rather than later. In fact, not only does this transience tend to obscure their existence, but programming it is not such a simple task either.

Another characteristic of Rule 110 is that it presents a tiling problem for the plane, with integer right isosceles triangles of all sizes. However, it is a variant not found in the tiling literature, say in the comprehensive treatise of Grünbaum and Shepard [90].

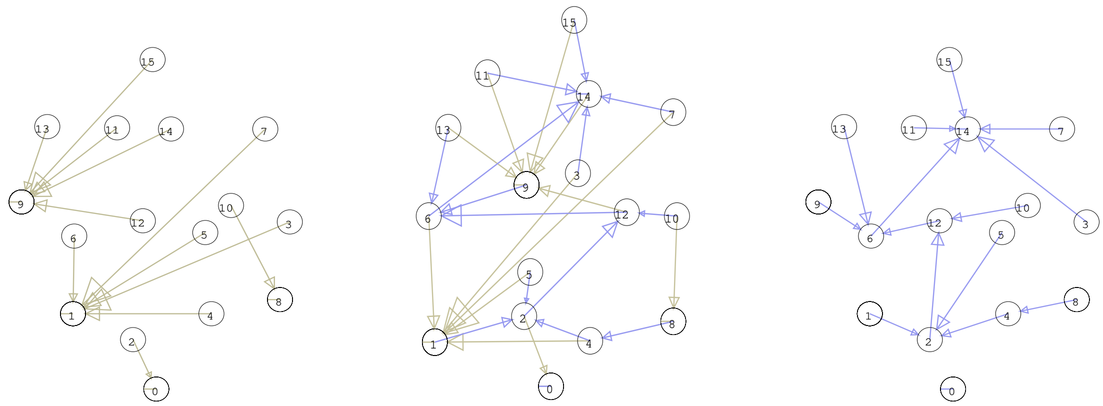


Figure 10: The subset diagram for Rule 110 shows that the shortest excluded word is 01010, and more typically the regular expression 00^*100^*10. Left: Only zero links are shown, from which four trees are apparent. Center: The full diagram, leaves and all. Note how the 1's free the paths from the 0^* traps. Right: Only one links are shown; aside from the empty set, there is one single tree rooted on the three-element node {01,10,11}.

Not only is the plane tiled by the right triangles; they are the only kind allowed and all cells in the zero state must be found inside one of them. One of the simplest prohibitions described by the subset diagram is that there cannot be two parallel columns of live cells surrounded by and separated by quiescent cells.

6.2 Mean field probabilities

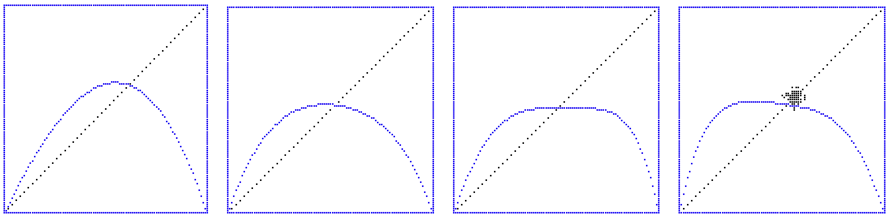


Figure 11: Mean field probabilities seem to converge to 60% which favors the T3 triangles.

A good place to begin the analysis of a cellular automaton is to consider its statistical behavior, insofar as it can be deduced from the rule of evolution.

For Rule 110, the table of evolution is

000	001	010	011	100	101	110	111
0	1	1	1	0	1	1	0

with the boolean formula

$$\varphi(a, b, c) = b \oplus c \oplus abc$$

given states a, b, c in three successive cells. According to the table. Langton's parameter for the automaton is $\lambda = 5/8$, or 0.625. The self-consistent mean field probability is given by

$$p \quad = \quad 2pq^2 + 3p^2q$$

whose positive root is $p = 0.612$ (the golden ratio) which, like Langton's parameter, requires slightly more than half of the cells to be alive.

Figure 11 shows the mean field curves for four successive generations of evolution. Superposed on the last curve are 200 Monte Carlo points which shows that actual experience, at least in fhe first generations, favors a value intermediate between Langton's value, the mean field value, and the *a priori* value of 50%, but in the vicintiy of all of them.

Constructing the evolution is a kind of tiling problem [90] for the plane, or at least a half-plane. But given the variety and specific shapes of the tiles it does not fit into one of the traditional categories. As a dissection problem it is akin to squaring the square or triangling the triangle [95].

Again, observation of evolution from random initial configurations shows a marked preference for T3 triangles (this nomenclature reflects the number of zero cells on the top line, or on the left edge, of the triangle). It is possible to tile the plane with triangles of any size, but for larger triangles some mixtures are needed.

There are two enantiomorphic tilings of the plane using T3 triangles exclusively, but one of them does not interact well with the rule forbidding prolongations of top edges (concretely, they are incompatible with $\varphi(1, 1, 1) = 0$). Cook called the other tiling the *ether*.

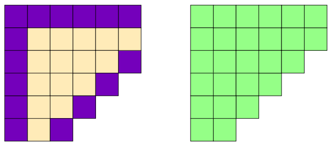


Figure 12: A T5 tile, illustrating triangles' canonical shape and from which the relative proportion of zeroes and ones may be deduced. Left: T5 decorated according to evolution in Rule 110. Right: An unadorned plane-tiling S5.

A reason for the predominance of the ether can probably be seen in the mean field probabilities, once the ratio of ones to zeroes in the individual tiles is examined.

Figure 12 shows a T5 tile, which is typical of all of them. To minimize overlapping and get a good tiling exercise, the right margin and the bottom margin are left off the tile, since they will always be provided by the abutting tile. Some additional overlap is inevitable since a neighboring triangle making contact along the hypotenuse always intersects the hypotenuse.

For plane tiling purposes, it is better to invent T0 triangles consisting of a single cell, and leave the diagonal off Tn triangles altogether. The new style can be called Sn, and by avoiding questions of overlap, give a more classical tiling exercise. In either event, we can elaborate the following table:

statistic	Tn	Sn
number of zero cells	$\frac{1}{2}n(n+1)$	$\frac{1}{2}n(n+1)$
number of one cells	$2n+1+(n-1)=3n$	$2n+1$
total area	$\frac{1}{2}n(n+7)$	$1+\frac{1}{2}n(n+5)$
proportion of ones	$6/(n+7)$	$(4n+2)(n^2+5n+2)$.

From this table it can already be seen that large triangles are at a disadvantage, although the disadvantage only obeys a 1/n law, and that there is an optimal approach to the mean field value at $n=3$. The following table describes this trend for Tn:

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$\frac{6}{n+7}$	0.75	0.67	0.60	0.55	0.50	0.46	0.43	0.40	0.375	0.35	0.33	0.26	0.30	0.28

If one confides in mean field theory at the 10% level, T1 is really out of range, T2 is not much better, nor are T5 nor anything further along the sequence. It is not excluded that some T2-T5 mixtures could be concocted, and it is seen that the hypotenusal overlaps won't affect the statistics much.

For reference, the absolute sizes of the first Tn's are shown in the following table.

	n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
T_n	$\frac{1}{2}n(n+7)$		4	9	15	22	30	39	49	60	72	85	99	114	130	147
S_n	$\frac{1}{2}n(n+5)+1$	1	4	8	13	19	26	34	43	53	64	76	89	103	118	134

The main purpose of listing the areas of these tiles lies in the packing problem which arises when the unit cell of a glider wants to be filled. Certain areas are available as cells, which are to be packed with tiles. The number of tiles is the total amount of diagonal available, and the S_0's are not freely available. Due to the prohibition against extending top edges, they can only occur singly, on diagonals. Furthermore, they occupy alternate diagonal sites, so that in the end roughly half the diagonal space availlable in a cell is dedicated to S_0's.

It is possible to speculate whether the properties of the mean field theory curves favor the T3 gliders, or whether the predominasnce of T3 gliders in later generations moved the mean field equilibrium. At least in the first few generations, where the influence of correlations is presumably smaller than in the later generations, both tiny and large triangles are at a disadvantage, it is still possible to feel that T3's are statistical favorites.

6.3 Gliders in the ether

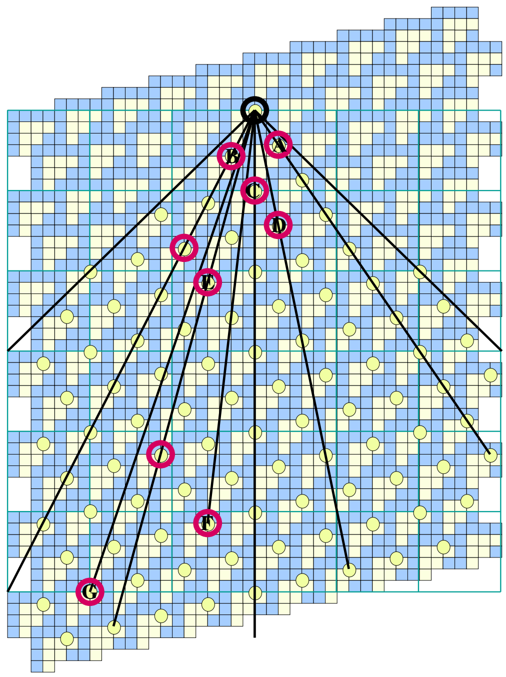


Figure 13: T3 tiles ought to be the most common, but one of the two enantiomers would violate the rule were it an extensive background. Gliders velocities correspond to fault lines in the T3 lattice. Possibilities are marked by small circles, Cook's by heavy circles.

Given the assumption that the ether is going to be the background against which glider theory is to be developed, their velocities can be read off as slopes of the fault lines in the ether lattice. That does not mean that there are not other lattices which admit gliders. We have searched for them and found several. One of them, with T5's, plays a role in Cook's extensible gliders, namely the E and the G series. We have tentatively called it an α tile. However, observations of evolution from random initial conditions show that the ether lattice is overwhelmingly, if not exclusively, favored. So that is the environment in which the potentialities for gliders should be examined.

It is worth emphasizing the subjectivity of the concept of a glider. Originally, in Conway's formulation of *Life*, they were recognizable discrete objects which moved against a quiescent background. In fact, the name "glider" reflects a peculiarity of that movement, that the object flipped about its line of motion as it progressed, reminiscent of glide planes in crystallography. Here there is no flipping, and the background is not quiescent, but those are quibbles.

Gliders find their place in de Bruijn diagrams given two or more cycles mutually interconnected. The pattern of one of them can be embedded in an environment created by the other; in *Life* one of those cycles is the self-loop of the quiescent node. If the connection is unidirectional, space is split in half rather than embedding; the structures are called *fuses*. If the two cycles are disconnected, there are two distinct patterns which cannot coexist shift-periodically.

Which is the glider and which is the background, and whether a thing is called a glider at all, dependends on appearance and aesthetics.

Figure 13 shows the fault lines in the ether lattice. On general principles, superluminal gliders are excluded, but not all the remaining possibilities turn out to be realized. Those whose de Bruijn diagrams are accessible to our computational facilities show the presence of only those which Cook announced. The absence of others depends on the lack of interconnection of configurations of the given shift periodicity to the subsets of the ether lattice having the same specifications.

There are selection rules; for example, nothing speedier than Cook's A and B gliders. Showing this has been set forth as a challange, but it is empirically true that the larger the largest triangle in a glider, the slower it moves. Slowness is related to the problem of filling the areas under the hypotenuses when large triangles are strung alongside each other. The triangle, the further down its next occurrence is likely to be found, promoting lethargy.

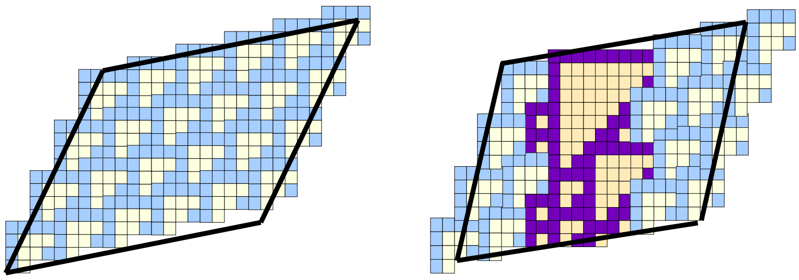


Figure 14: Gliders, having the same slope as a fault line in the ether lattice, can be inserted by dividing the lattice at the fault, translating the two halves to separate them, then inserting the glider strip. The insert has a width, which need not coincide with any strip taken out of the lattice.

Figure 14 shows a unit cell for one of Cook's E-gliders, which shifts left by 4 every 15 generations. The unit cell looks to be 19×15, the height being the period and the width being the distance necessary to resynchronize the ether lattice. However, the actual figure is 23×13 because the unit cell is a rhombus, so the dimensions cited are sections, not altitudes.

To reconcile the unit cell of 285 cells with tiles which it contains, consider

constituent tile	$T0$	$T1$	$T2$	$T3$	$T5$	$T8$
number per cell	20	7	2	12	1	1
area of each tile	1	4	9	15	30	60
raw total	20	28	18	180	30	60
diagonal per cell	0	0	1	2	4	7
diagonal total	0	0	2	24	4	7

Summing the areas of the tiles gives 336 cells. There are 23 tiles altogether with 37 diagonal cells, each of which has been covered twice in this accounting. The net total of 299 factors into 23×13, which is the area of the unit cell. The factors relate to the dimensions of the cell which, as a rhombus, has an effective height and an effective width.

6.4 Speed limits

Gliders in automata cannot have superluminal velocities because their configurations are determined by a function, and therefore deterministic once an initial neighborhood has been specified. This prohibition is inoperative at light velocity or less, but there may be other considerations limiting velocities. For example, in *Life* there are infinite patterns moving at light velocity, any of the ripples derived from the one-dimensional Rule 22. But restricting the extent of the live part of the configuration reduces the velocity, which was one of the first charasteristics mentioned when *Life* was first publicized.

Gliders relative to the ether seem to obey similar restrictions, evidences of which can be seen by examining Cook's list of known gliders. A's and B's are the fastest; A's and D's are the only ones with positive velocities. There are gliders which he called BBar, which have the same velocity as the B gliders, but which take a multiple of the B period to complete their own period. But there are no ABars, and no BDoubleBars.

His glider list, which others claim to have verified by their own search programs, is an empirical result. There are different kinds of search programs; one which we used to list cycles in Rule 22 up to length 34 should be able to detect Cook's gliders, but we have not tried to revive the older computers on which it was run.

Another approach is to filter the shift-periodic de Bruijn diagrams down to their ergodic nuclei, to see which combinations could be gliders. This approach is quite reliable, but the size of the diagrams to which it is to be applied are enormous; about ten generations is the limit for present computer speed and memory, although it could be stretched somewhat given sufficient incentive. But that region only contains A, B, C, and D gliders; BBar, E, F, and G lie on beyond, some of them quite a bit.

To describe a velocity as a class, including all te possible multiples of the basic period and the deviations which only return to the original form after many intervals, another approach is needed. One idea is shown in Figure 15, to take advantage of the little hook sitting on the margin of a strip containing a faster glider than the A. The consequences of filling the niche with tiles of different sizes can be followed down in one of those search trees which is so characteristic of applied Artificial Intelligence.

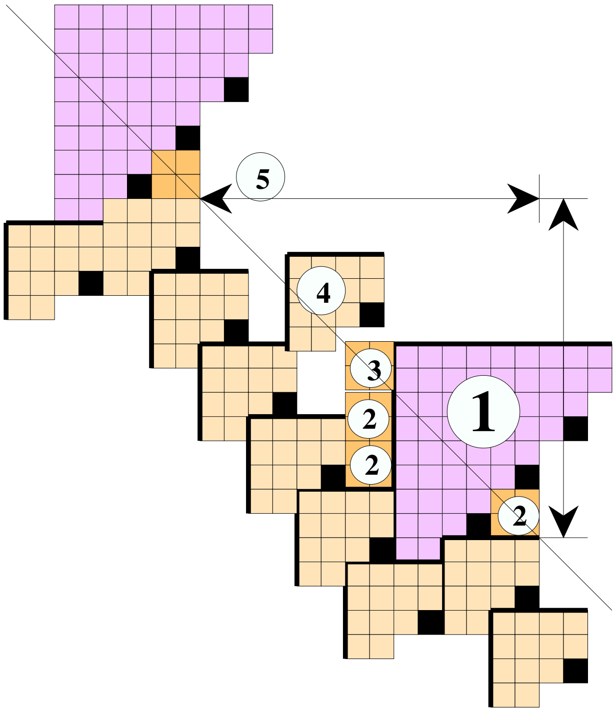


Figure 15: A possible speed limit. There is a transition region between a glider and the ambient lattice. If a glider is to move faster than an A glider, there will be an ether margin with a little hook in it, whose filling could have restrictive repercussions.

Two T0's would violate the prohibition against extending top margins, so the smallest viable candidate would be a T1 and if that ran back along the margin, the result would simply be an A glider. If something faster is required, that chain simply creates a new hook, so it is time to move on to the next possibility.

Figure 15 shows the initial consequences of starting with a T8 (1). That immediately forces three T1's (2), following which a fourth T1 can not be stacked alongside the T* (3). Backtracking, it is seen that a T3 could be inserted, based in the next column to the left (4). A new hook is created so it is time to begin again. Eventually the consequences of the original choice must be faced (5) in the form of a wall of some nature. If it is compatible with the procedure to date, the filling of the area above the lower margin of the glider strip is continued until the top margin is encountered.

Hopefully a decision will have been reached before the top margin is required, because there is really no way of knowing how wide a glider strip could be. For evidence, consider the infinitude of A polymers which is possible.

The search scheme depicted in Figure 15 may or may not succeed, but there is no denying that it is cumbersome. Figure 16 shows the ether margin from another perspective, which strongly invites

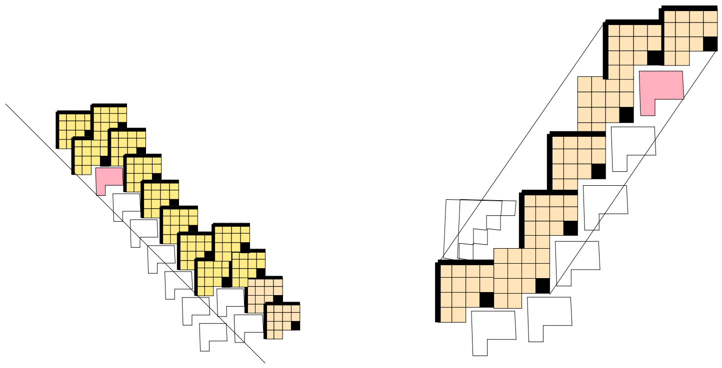


Figure 16: Taking advantage of the kink in the ether margin, and looking at it from the right angle, it becomes plausible that there are no right runners faster than A's nor leftrunners faster than B's.

the conclusion that nothing can be fitted into the nooks and crannies, rendering any search for faster gliders futile.

6.5 Glider collisions

Cook recognizes at least seven types of gliders, some of which have multiple variants, and all of which can be packed in various combinations. Glider technology consists in exploiting collisions. It is readily seen that there are fifty or more, actually in the hundreds or thousands, of binary collisions. From there on the numbers are much larger. Although that forebodes a mass of data, there is an element of hope.

Even though a dozen or two of gliders seems like a large number, that is far fewer than the number of possible collisions. Either collisions will produce multiple reaction products, or many collisions will end up producing identical results. Both alternatives coexist, but one of the fortunate combinations leaves the reactants unchanged, just that they chang places. That is a soliton reaction, useful for moving information from one place to another across intervening barriers.

Actually we didn't discover this for ourselves, but only as a result of a hint from Cook. That, and the general knowledge that it would eventually be necessary to examine the collisions in one way or another.

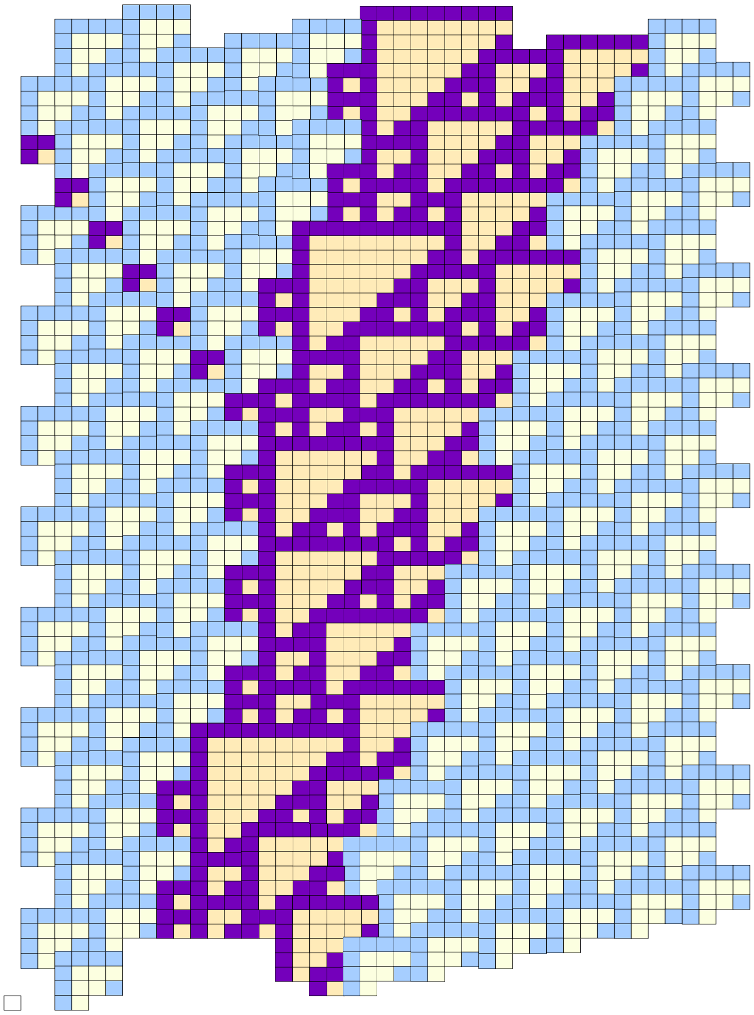


Figure 17: B collisions always promote E_n to E_{n+1}, but there are some A collisions which will restore the count, reducing E_n to E_{n-1} or at the bottom of the chain, E to C.

	E		EBar		F	
	Hi	Lo	Hi	Lo	Hi	Lo
C1	EBar F A	C2 D2 A	EBar C1	EBar C1	EBar C2	F C1
C2	2 B's 1 A	C1 EBar A	BBar 2 B's 1 A	3 B's	C1 F BBar	C1 F BBar
C3	3 B's 1 A	G A	F C2	4 B's	C1 C2	C2 F BBar

Figure 18: Some of the glider collisions involving the three different embedments of the C glider in the ether on the one hand, and members of the family E, EBar, and F in each of their two alignments relative to the C's. The labels are Cook's.

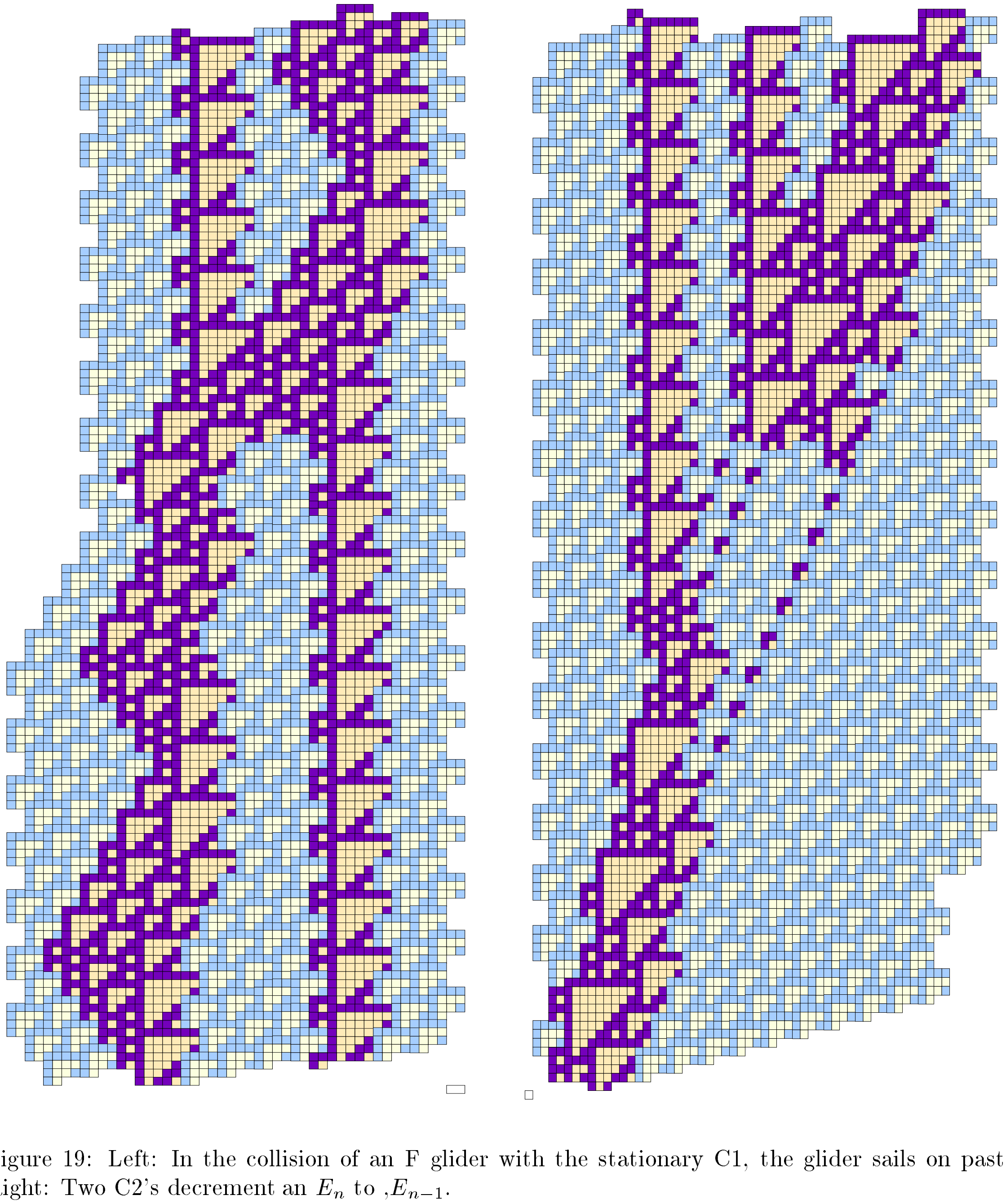


Figure 19: Left: In the collision of an F glider with the stationary C1, the glider sails on past. Right: Two C2's decrement an E_n to $,E_{n-1}$.

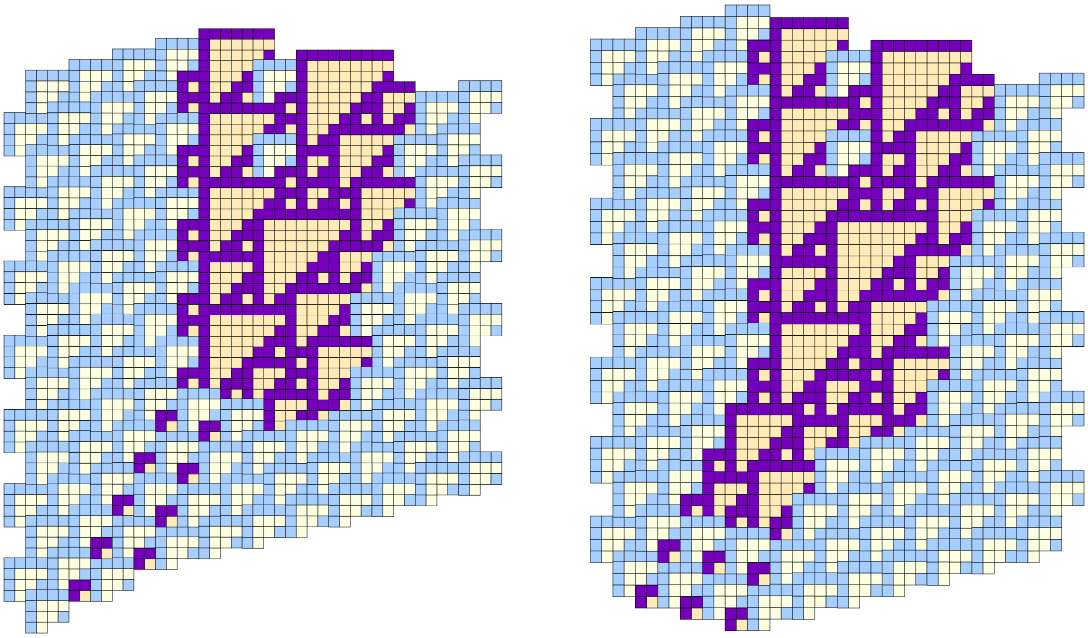


Figure 20: Left: $C2E_nHi$ collisions result in n B gliders which can interact with new C's to recreate E_{n-1} gliders. Right: $C3E_nHi$ collisions release $n+1$ gliders which can be used to restore the E_n.

C * En collisions

	C1	C2	C3
E1 hi	EBar, F, A	A, 2 B's	A, 5 B's
E1 lo	C2, D2, A	C1, EBar, A	G, A
E2 hi	EBar, F	2 B's	3 B's
E2 lo	C2, D2	C1, EBar	G
E3 hi	C2, EBar	3 B's	4 B's
E3 lo	A, 2 A dimers	A tetramer, 2 B's, BBar, F, C2	D1, EBar
E4 hi	D2, E	4 B's	5 B's
E4 lo	C1, E, 2B's	C3, EBar	
E5 hi		5 B's	6 B's
E5 lo			
E6 hi		6 B's	7 B's
E6 lo			

CEncollision June 24, 1999

Figure 21: A partial listing of the products of C and E collisions. All the $C2$ and $C3$ collisions in the high position are clean, releasing nothing more than a flotilla of B gliders.

Part of the understanding of E and G collisions is that the leading triangles are escorting α gliders in one constellation or another; when the escort is damaged by a collision, the result is a typical putrefaction of an α lattice, for which the rules of engagement can be worked out.

6.6 Nonemergence of large triangles

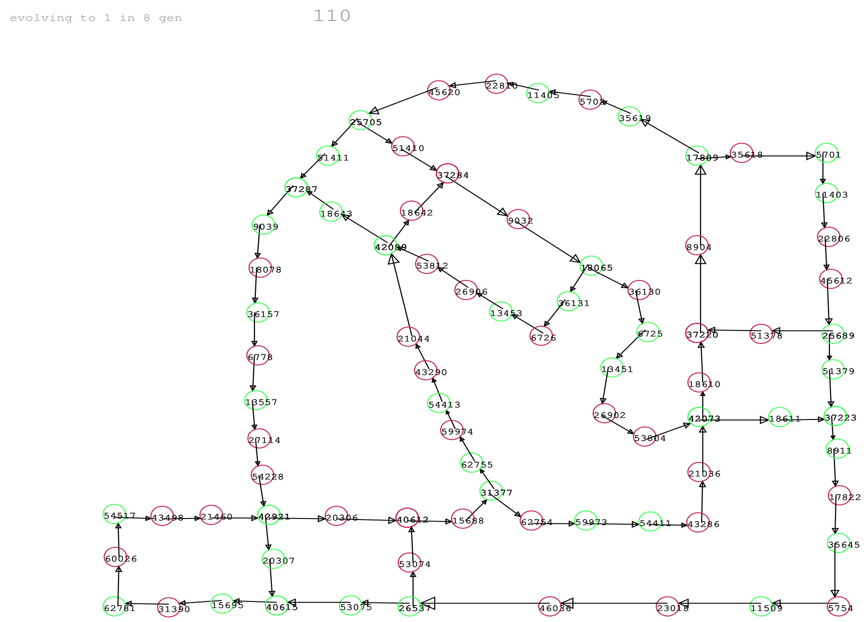


Figure 22: The de Bruijn diagram for evolution to ones in nine generations is null, but there are various cycles for evolution to ones in eight generations. Since ones are needed as borders of Tn's, this shows that there are no such triangles after eight generations. At the ninth generation, or any other, transients in the de Bruijn diagram are as much ancestors as any other part. Their length tells how big a triangle is still possile, but it is still necessary to consult the graph to know how different the transients are. Different components would yield separate and distinct ancestors, but within a component ancestors could coincide along part of their middles.

For small shift-period combinations it is feasible to scan the de Bruijn diagrams just to see the combinations, just as the small cycle diagrams yield useful information. Evolution to constancy is also readily available, as a consequence of which it was noticed that there was no kernel in the diagram for evolution to ones in nine generations, although the kernel of tghe diagram for evolution to ones in eight generations still has some structure, as seen in Figure 22.

That means that the nine generation diagram is acyclic. Since it is finite, there is an absolute limit to the length of configuration which can evolve to pure ones without repeating a portion. Since the next generation for any string of three or more ones would be pure zeroes, it follows that there is a limit on the size of triangles which can evolve from conditions other than already existing zeroes. Long sequences of zeroes always diminish, and new sequences only arise from sequences of ones.

De Bruijn diagrams are Hamiltonian, which means that paths can run through as many as all the vertices without repeating. For nine generations the number is huge: $2^{19} = 524,288$. Subdiagrams and their kernels need not remotely approach this limit, but as a bound it is still finite. In terms of specific cases, found by exhaustive search, the longest chains turn out to be 44 cells long. Since the kernel searching program can be made to report the number of links each time an outer layer of leaves (or rootlets) are removed, it is possible to get some idea of the form of the acyclic diagram. The table of computer listing is long and not necessarily interesting in its entirity; but here are some excerpts.

```
   ipass2i (nine generations - in links)
step  0, links =   281408, loss =        0, percentage =  0.00
step  1, links =   177048, loss =   104360, percentage = 37.08
step  2, links =   100354, loss =    76694, percentage = 43.32
...
step 41, links =      267, loss =       98, percentage = 26.85
step 42, links =      187, loss =       80, percentage = 29.96
step 43, links =       27, loss =      160, percentage = 85.56
step 44, links =        0, loss =       27, percentage = 100.00
maximum in length =       45
ipass2o (nine generations - out links)
maximum out length =       1
```

In this phase, the kernel program removes in-links, layer by layer, starting with a diagram containing 281,408 links altogether. Just over half the capacity of the de Bruijn diagram, it is just slightly more likely to evolve a one as a zero. For a while, about half the links are lost per cycle, reflecting the same tendency to slightly favor ones.

At the very end, there are still 27 links, which gives an idea of the number of chains, which is the number of rightmost neighborhoods. The program stops at step 25, when there are no more links. The second phase stops at once, there being no out links because there aren't any links at all.

The percentage lossage gives a rough idea of the branching of the acyclic binary graph. Half the content of a binary tree is in the leaves, but a minuscule part of an unbranched chain is in its leaf. Losing 85% of the leaves in step 43 implies that the termination of the ancestors didn't matter much.

For the sake of comparison, we present the same data, gotten by purging rootlets rather than leaves.

```
   ipass2i (nine generations - in links)
step  0, links =   281408, loss =        0, percentage =  0.00
step  1, links =   208120, loss =    73288, percentage = 26.04
step  2, links =   160912, loss =    47208, percentage = 22.68
step  3, links =   124096, loss =    36816, percentage = 22.88
...
step 41, links =       96, loss =       96, percentage = 50.00
step 42, links =       32, loss =       64, percentage = 66.67
step 43, links =       16, loss =       16, percentage = 50.00
step 44, links =        0, loss =       16, percentage = 100.00
maximum out length =       45
maximum in length =        1
```

The results are rather similar, the main difference being that there were only sixteen links at the end, so there is more of a tendency of distinct ancestors to begin similarly. In the initial somewhat more links survived. But in either case, there are about a score of ancestors of a string of 44 links. According to the system of naming triangles, it is T42 which we are discussing.

```
epass2i (five generations)
step  0, inlinks =     1058, loss =          0, or      0%
step  1, inlinks =      658, loss =        400, or     38%
step  2, inlinks =      357, loss =        301, or     46%
step  3, inlinks =      219, loss =        138, or     39%
step  4, inlinks =      184, loss =         35, or     16%
step  5, inlinks =      144, loss =         40, or     22%
step  6, inlinks =      140, loss =          4, or      3%
step  7, inlinks =      132, loss =          8, or      6%
step  8, inlinks =      121, loss =         11, or      8%
step  9, inlinks =      116, loss =          5, or      4%
maximum in length =       10
epass2i (five generations)
step  0, outlinks =      116, loss =          0, or      0%
step  1, outlinks =       85, loss =         31, or     27%
step  2, outlinks =       71, loss =         14, or     16%
step  3, outlinks =       60, loss =         11, or     15%
step  4, outlinks =       51, loss =          9, or     15%
step  5, outlinks =       44, loss =          7, or     14%
step  6, outlinks =       36, loss =          8, or     18%
step  7, outlinks =       28, loss =          8, or     22%
step  8, outlinks =       25, loss =          3, or     11%
step  9, outlinks =       22, loss =          3, or     12%
step 10, outlinks =       19, loss =          3, or     14%
step 11, outlinks =       16, loss =          3, or     16%
step 12, outlinks =       14, loss =          2, or     13%
step 13, outlinks =       12, loss =          2, or     14%
step 14, outlinks =       11, loss =          1, or      8%
step 15, outlinks =       10, loss =          1, or      9%
maximum out length =       16
```

```
epass2o (five generations)
step  0, outlinks =     1058, loss =          0, or      0%
step  1, outlinks =      844, loss =        214, or     20%
step  2, outlinks =      724, loss =        120, or     14%
step  3, outlinks =      600, loss =        124, or     17%
step  4, outlinks =      524, loss =         76, or     13%
step  5, outlinks =      464, loss =         60, or     11%
step  6, outlinks =      408, loss =         56, or     12%
step  7, outlinks =      360, loss =         48, or     12%
step  8, outlinks =      328, loss =         32, or      9%
step  9, outlinks =      296, loss =         32, or     10%
step 10, outlinks =      264, loss =         32, or     11%
step 11, outlinks =      220, loss =         44, or     17%
step 12, outlinks =      196, loss =         24, or     11%
step 13, outlinks =      168, loss =         28, or     14%
step 14, outlinks =      152, loss =         16, or     10%
step 15, outlinks =      140, loss =         12, or      8%
step 16, outlinks =      136, loss =          4, or      3%
step 17, outlinks =      132, loss =          4, or      3%
step 18, outlinks =      124, loss =          8, or      6%
maximum out length =       19
epass2i (five generations)
step  0, inlinks =      124, loss =          0, or      0%
step  1, inlinks =       62, loss =         62, or     50%
step  2, inlinks =       31, loss =         31, or     50%
step  3, inlinks =       18, loss =         13, or     42%
step  4, inlinks =       16, loss =          2, or     11%
step  5, inlinks =       14, loss =          2, or     13%
step  6, inlinks =       13, loss =          1, or      7%
step  7, inlinks =       12, loss =          1, or      8%
step  8, inlinks =       11, loss =          1, or      8%
step  9, inlinks =       10, loss =          1, or      9%
maximum in length =       10
```

Figure 23: The five generation de Bruijn diagram for evolution to the constant 1 is one single cycle of length 10, which is an interesting commentary for every evolution according to Rule 110. This cycle has transients, both incoming and outgoing, whose nature is indicated in the listings above. Transients can vary between 10 and 20 in length; the noncommutativity between the two tables is due to the existence of additional acyclic components. The information presented here is fairly indirect; were it worthwhile, the entire de Bruijn diagram could be displayed, although it is an artistic challenge with 1024 links.

Figure 24 shows the evolution of one T42, confirming the existence of such tiles. It also shows the failure of an attempt to construct a T43 in eight generations, although tiles of any length can be constructed for seven generations. The easiest way to get them is to read them off the seventh stage evolution to constancy, but of course there are many more ways to torm them, depending on the transients in the same diagram.

It is an interesting result that the ancestors of T41 cannot be pushed back beyond twelve generations, so that it is also a nonemergent tile. It remains unknown how large an emergent triangle can be, although in the course of experimentation we have compiled a concordance, showing gliders and glider collisions which we have actually observed. So far the largest tile resulting from a collision is T23. Such an origin means that it can be produced after an arbitrarily large number of generations, simply by adjusting the initial positions of the colliding gliders

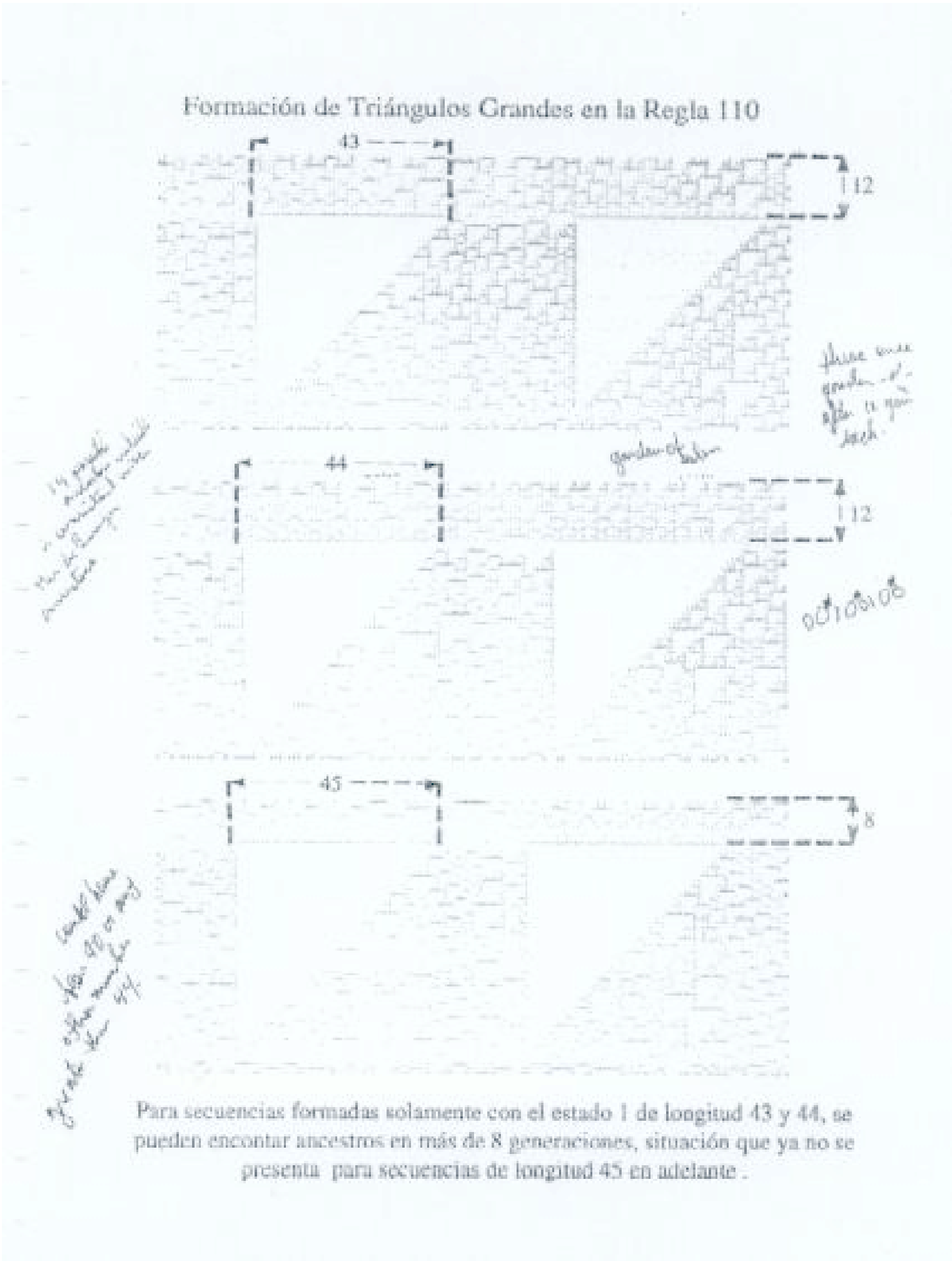


Figure 24: The largest triangle which can occur under natural evolution is T42, and that after no more than twelve generations. T41 can also arise after no more than twelve generations, but T43 and larger triangles can't ever make an appearance after the eighth generation unless they were there all along — Garden of Eden configurations. Up to and including eight generations, triangles of any size can be generated.

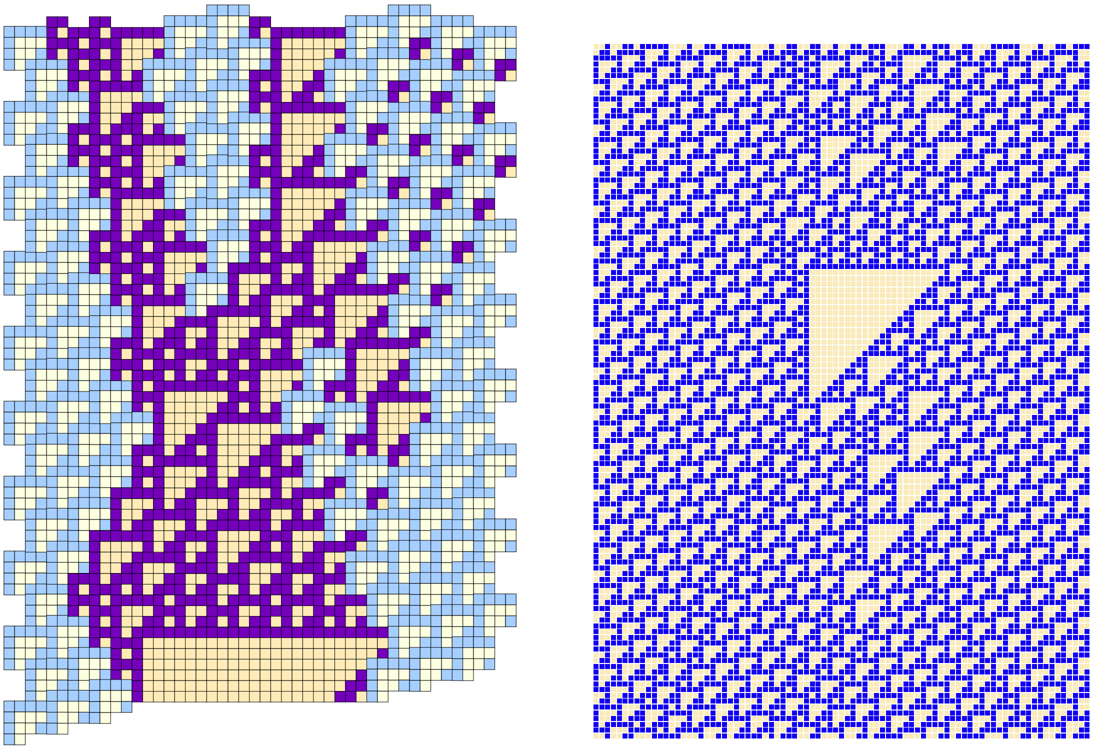


Figure 25: According to evolution governed by Rule 110, no triangle larger than T42 (which has 42×42 as its inner dimension) can evolve or persist for more than ten generations. The largest triangle which we have so far encountered in a glider collision is T23, as a consequence of which one can be formed after arbitrarily many generations. The range between T24 and T42 is still unexplored. Left: hand drawn evolution showing T23 as the result of a triple collision of a D2, C2, and B tetramer. Right: computer evolution confirming the correctness of the left hand panel.

7 Lorentz Contraction

The theory of the three dimensional rotation group has been pretty well studied and leads to some elegant interpretations of the parameters of the group. One of them uses arcs on the surface of a unit sphere, whereby the pole of the arc is the axis of rotation and the length of the arc is the angle of rotation. The arc can be positioned anywhere along its equator, but for the composition of rotations, the arcs should be placed tip to tail like vectors. The extreme points, together with the center of the sphere, define a new plane, and with it a new polar axis and arc of rotation. The curvature of the surface accounts for the noncommutativity of the composition.

7.1 Small matrices

There is a somewhat similar algebra applicable to 2×2 real matrices, which are of frequent occurence in the theory of second order linear differential equations and elsewhere. However, a quaternion basis is not strictly appropriate because the squares of two of the candidfates for basis vectors are $+\mathbf{1}$ rather than the $-\mathbf{1}$ characteristic of quaternions. The result is that all the theory resembles that of the rotation group except for the fact that it works with a Minkowski metric rather than the accustomed Euclidean metric, which in turn means that there is an effect similar to Lorentz contraction when working with geometric concepts.

Another representation of rotations considers them to be the product of two reflections. The reflecting planes should intersect along the axis of rotation, which is invariant under each of the two reflections and consequently their composite. The dihedral angle separating the planes should be half the angle of rotation, because under the reflection points end up just as far above the plane as they originated below it. Any pair of planes with the same axis and angular aperture will produce the same rotation, inviting the selection of a common intermediate plane, whose presence disappears by squaring when composing two individual rotations, always leaving two reflections to define a rotation.

Reflections aren't a normal part of quaternion lore, but both representations are a part of the common knowledge about rotations. Even so, it is not so easy to track down the historical origins of the representation, or even to be sure that it has not arisen spontaneously with different investigators in altogether different locations. For example, knowing that the exponential of an antisymmetric matrix is orthogonal usually sufficies to deduce Rodrigues' formula, the one which figured in our investigation of Putzer's method in Section 3. Going from that point to the vector representation is harder, making a historical analysis more significant if one were actually interested in origins.

Insofar as my own involvement in the representation is concerned, the question of multiplying 2×2 matrices to solve cascaded transmission line problems and later on to solve differential equations goes back to my graduate school days. Involvement with the rotation group goes almost as far back, both from an interest in group theory, and trying to express a certain practical problem in terms of spherical harmonics. Articles in the physical literature, particularly in *American Journal of Physics*, frequently spoke of Pauli matrices, quaternions, and other topics.

Before long, differences between the rotation group and the unimodular matrices emerge. The group $SL(2, R)$, as it is known, is locally compact but not compact, so it has infinite dimensional irreducible representations. These were studied by E. P. Wigner in the late 1930's [98], V. Bargmann a decade later [97], and are the principal focus of Serge Lang's book [96] of 1975. But is is harder to find treatments in the the group itself, rather than its representations. Over the years I have looked for a graphical representation equivalent to that of the rotation group. Although such a thing exists and seems to be relatively simple, it is just recently that a clear picture has emerged.

First we repeat some of the background details, taken from lecture notes and various places in the literature.

7.2 Quaternions versus elementary matrices

Let us start by reviewing unimodular algebra.

7.2.1 multiplication table

The best way to get at $2x2$ matrices, is to use quaternions. Starting from the natural basis for $2x2$ matrices,

$$\mathbf{e}_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{e}_{12} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{e}_{21} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{e}_{22} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

whose rule of multiplication is $\mathbf{e}_{ij}\mathbf{e}_{kl} = \delta_{jk}\mathbf{e}_{il}$, quaternion-like matrices can be defined by

$$\mathbf{1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{i} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \mathbf{j} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{k} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

In detail,

$$\mathbf{1} = \mathbf{e}_{11} + \mathbf{e}_{22}, \quad \mathbf{i} = \mathbf{e}_{12} - \mathbf{e}_{21}, \quad \mathbf{j} = \mathbf{e}_{12} + \mathbf{e}_{21}, \quad \mathbf{k} = \mathbf{e}_{11} - \mathbf{e}_{22},$$

all built from sums and differences, thereby retaining real matrices. Like quaternions, these matrices anticommute (except for the identity). The ostensible difference is that only one square is $-\mathbf{1}$, the others are $+\mathbf{1}$, changing Euler's formula from trigonometric to hyperbolic functions according to the sign.

The multiplication table is

	1	**i**	**j**	**k**
1	**1**	**i**	**j**	**k**
i	**i**	−**1**	**k**	−**j**
j	**j**	−**k**	**1**	−**i**
k	**k**	**j**	**i**	**1**

The usual way of performing algebraic operations on these matrices is to write a sum such as $a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ in the form $s + \mathbf{v}$, where $s = a\mathbf{1}$ and $\mathbf{v}$ is the rest of the sum. One can adopt the custom of not writing an explicit $\mathbf{1}$ in places where the meaning is clear.

7.2.2 splitting quaternions

Anyway, splitting the sum into scalar plus vector gives

$$(s + \mathbf{u})(t + \mathbf{v}) = st + s\mathbf{v} + t\mathbf{u} + (\mathbf{u}.\mathbf{v}) + (\mathbf{u} \times \mathbf{v}), \tag{49}$$

particular interest attaching to the case where s and t are zero, leaving the product of two vectors to take the form of a scalar plus a vector. This inner (or dot) product is not the usual one, but one using a Minkowski metric:

$$(\mathbf{u}\cdot\mathbf{v}) = -u_1v_1 + u_2v_2 + u_3v_3, \tag{50}$$

$$= \begin{bmatrix} u_1 & u_2 & u_3 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}. \tag{51}$$

Since the Minkowski inner product can be positive, negative, or zero, taking it for the square of a norm requires considering the sign, unless an imaginary norm is acceptable. So to define the norm of a vector, use the absolute value of the metric, by setting

$$|\mathbf{v}| = \sqrt{}\mathrm{abs}((\mathbf{v},\mathbf{v})), \tag{52}$$

which can vanish for a nonzero vector, and never forgetting the possible influence of the bypassed sign.

Since the special theory of relativity contains the best known uses of the Minkowslki metric, it is convenient to adopt the physical vocabulary according to which vectors whose squared norm is positive are spacelike, timelike for a negative square, and null when it is zero.

subsubsectionvector product

In turn the vector product differs slightly from its cartesian version, reading

$$\mathbf{u}\times\mathbf{v} = (u_3v_2 - u_2v_3)\,\mathbf{i} + (u_3v_1 - u_1v_3)\,\mathbf{j} + (u_1v_2 - u_2v_1)\,\mathbf{k} \tag{53}$$

$$= -\begin{vmatrix} u_2 & u_3 \\ v_2 & v_3 \end{vmatrix}\mathbf{i} + \begin{vmatrix} u_3 & u_1 \\ v_3 & v_1 \end{vmatrix}\mathbf{j} + \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix}\mathbf{k} \tag{54}$$

$$\tag{55}$$

$$= \begin{vmatrix} -\mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}. \tag{56}$$

This symbolic determinant works out well enough, differing from the classical formula only by the sign associated with $\mathbf{i}$. The notation makes it immediately apparent that $\mathbf{u}\times\mathbf{v} = -\mathbf{v}\times\mathbf{u}$ and that $\mathbf{u}\times\mathbf{u} = 0$.

Interestingly enough,

$$(\mathbf{u}\times\mathbf{v},\mathbf{w}) = \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}, \tag{57}$$

the minus sign in he inner product cancelling the minus in the vector product. So the formula for Euclidean volume can still be used to check for the linear dependence of vectors. The formula can be used to confirm that in the Minkowski metric, the vector product is Minkowski-orthogonal to its factors, in complete analogy to the Euclidean formula. However, if we are interested in the whole plane which is Minkowski-orthogonal to a given vector, it will be inclined relative to what would be expected from Euclidean orthogonality.

7.2.3 orthogonality

In a two-dimensional context, if the vector (x_1, x_2) is orthogonal to (x_1, x_2), we would have $x_1y_1 + x_2y_2 = 0$, or $x_2/x_1 = -y_1/y_2$ and the vectors would have negative reciprocal slopes. The minus sign in the Minkowski version gives them reciprocal slopes, making them mirror images in the diagonal. On the diagonal itself, the light cone, vectors are self-orthogonal, the same as being null. In three dimensions the diagonal is actually a cone, to which the orthogonal plane is tangent. Planes Minkowski-orthogonal to axes of other inclinations are filled with vectors of the opposite timeliness or spaceliness from the axis itself. That would seem to imply that the vector product of vectors of opposite timeliness would be null, but things are a little more complicated than that.

The triple vector product is not associative, but is multilinear in each of its factors. Again in analogy to the Euclidean formula,

$$(\mathbf{u} \times \mathbf{v}) \times \mathbf{w} \quad = \quad \mathbf{u}(\mathbf{v}, \mathbf{w}) - \mathbf{v}(\mathbf{u}, \mathbf{w}) \tag{58}$$

7.3 Quaternion inverse

The easiest way to get the inverse of a full quaternion $a + \mathbf{v}$ is to go back to its representation by elementary matrices, in the $2x2$ form

$$a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} \quad \to \quad \begin{bmatrix} a+d & b+c \\ -b+c & a-d \end{bmatrix}.$$

The determinant, $(a+d)(a-d) - (c+b)(c-b) = a^2 - d^2 + b^2 - c^2$, which is $a^2 - (\mathbf{v}, \mathbf{v})$, could be considered to be a candidate for the square of the norm of a full quaternion, in contrast to the norm of a mere vector. Under that assumption, the formula for the inverse of a $2x2$ matrix gives the inverse quaternion

$$\frac{1}{(a^2 - |\mathbf{u}|^2)} \begin{bmatrix} a-d & -b-c \\ b-c & a+d \end{bmatrix} \quad \to \quad \frac{1}{(a^2 - |\mathbf{u}|^2)}(a\mathbf{1} - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}).$$

which is almost the formula for inverting Hamilton's authentic quaternions, except for the way the norm is calculated.

Nevertheless, the difference is important, because null quaternions exist, just as well as null vectors, and they cannot be invertible. Hamiton's quaternions are invertible unless zero. Not only do they constitute a field, albeit noncommutative; they lie amongst the very few examples of algebraically and topologically complete infinite fields. The objects which we have just defined are not quaternions according to Hamilton's definition; neither are Hamiton's own quaternions taken with complex coefficients (look at $(\mathbf{i} - i\mathbf{j})^2$, which vanishes).

At least we have a quantity which decides the invertibility of a quaternion such as $a\mathbf{1} + \mathbf{v}$, multiplicative for being a determinant, and consistent with the definition of the vector norm relative to a sign choice.

$$\|a\mathbf{1} + \mathbf{v}\|^2 \quad = \quad a^2 - (\mathbf{v}, \mathbf{v})$$

There are null vectors in the Minkowski metric, which is to say, nonzero vectors with zero norm. What would a null vector look like? Consider $\sqrt{2}\mathbf{i} + \mathbf{j} + \mathbf{k}$.

$$\begin{bmatrix} 1 & 1+\sqrt{2} \\ 1-\sqrt{2} & -1, \end{bmatrix}$$

whose trace and determinant are both zero, yet it is not the zero matrix. It has to have the Jordan normal form with eigenvalue zero. Those are generally the null vectors; by satisfying $\mathbf{u}^2 = \mathbf{0}$, they are their own eigenvectors, and manifestly nilpotent.

A general form for a null vector could be to take

$$\rho(-\mathbf{i}\ +\mathbf{j}\ \cos(\phi)+\mathbf{k}\ \sin(\phi)) \tag{59}$$

7.4 Square roots

Since the square of a vector is not just a scalar, but actually the square of its Minkowski norm, it follows that it once divided by that norm, it is its own inverse, or that a unit vector is involutory. That is almost true, since the square of the vector can be negative; in that case the square root of the absolute value should be taken, and a minus sign should be appended to the inverse.

The unit quaternions are square roots of either unity or of minus one. Are there any others? According to Eq. 49, to get a root of unity needs

$$\begin{aligned}(s+\mathbf{u})^2 &= s^2+2s\mathbf{u}+(\mathbf{u}.\mathbf{u})\\ &= \mathbf{1},\end{aligned}$$

leading to two mutually exclusive alternatives

$$\begin{array}{rl} s^2=\mathbf{1} & \quad \mathbf{u}=0\\ (\mathbf{u}.\mathbf{u})=\mathbf{1} & \quad s=0.\end{array}$$

Besides the two expected scalar roots, any vector of unit norm fills the bill, infinitely many in all. A second glance at the derivation shows that any vector, the square of whose norm is -1 (such as $\mathbf{i}$) is a root of $-\mathbf{1}$, but that there are no (real) scalar roots.

As for square roots in general, the square root of any scalar follows the same line of reasoning with the exception that everything is scaled by the positive square root of the scalar when it has one. The general requirement for

$$s+\mathbf{u} = \surd(t\mathbf{1}+\mathbf{v})$$

would be

$$\begin{aligned}(s+\mathbf{u})^2 &= s^2+2s\mathbf{u}+(\mathbf{u}.\mathbf{u})\\ &= t\mathbf{1}+\mathbf{v},\end{aligned}$$

which would require in succession

$$\begin{aligned}\mathbf{u} &= \frac{\mathbf{v}}{2s}\\ s^2+\frac{(\mathbf{v}\cdot\mathbf{v})}{4s^2} &= t\\ s^2 &= \frac{1}{2}(t\pm\surd(t^2-(\mathbf{v}\cdot\mathbf{v}))\end{aligned}$$

If s were 0 then $\mathbf{v}$ would have to be zero, and only a scalar could have a vector square root. But non-vector quaternions can have quaternion roots, of which there would appear to be exactly four possible values for s, not all necessarily real. For example, $\surd\mathbf{i}=\pm(\mathbf{1}+\mathbf{i})/\surd 2$.

7.5 Vector exponential

An exponential is usually defined according to the traditional power series. Consider first the exponential of a vector:

$$\begin{aligned}
\exp(\mathbf{v}) &= \mathbf{1} + \mathbf{v} + \frac{1}{2!}\mathbf{v}^2 + \frac{1}{3!}\mathbf{v}^3 + \ldots && (60)\\
&= [\mathbf{1} + \frac{1}{2!}(\mathbf{v}\cdot\mathbf{v}) + \frac{1}{4!}(\mathbf{v}\cdot\mathbf{v})^2 + \ldots] + && (61)\\
&\quad \frac{\mathbf{v}}{\surd(\mathbf{v}\cdot\mathbf{v})}[\surd(\mathbf{v}\cdot\mathbf{v}) + \frac{1}{3!}\surd(\mathbf{v}\cdot\mathbf{v})^3 + \ldots] && (62)\\
&= \mathbf{1}\cosh(\surd(\mathbf{v}\cdot\mathbf{v})) + \frac{\mathbf{v}}{\surd(\mathbf{v}\cdot\mathbf{v})}\sinh(\surd(\mathbf{v}\cdot\mathbf{v})), && (63)\\
&= \mathbf{1}\cosh(|\mathbf{v}|) + \frac{\mathbf{v}}{|\mathbf{v}|}\sinh(|\mathbf{v}|), && (64)
\end{aligned}$$

a result analogous to Euler's formula. The exponential of a quaternion is not much more complicated, since any scalar which could be added would commute with the quaternion, so its exponential could just be set aside as a multiplying scalar factor. How much to set aside in the general case depends on satisfying the identity $\cosh^2(x) - \sinh^2(x) = 1$, but in general there is much to be said in favor of working with vectors of unit norm and treating norms separately.

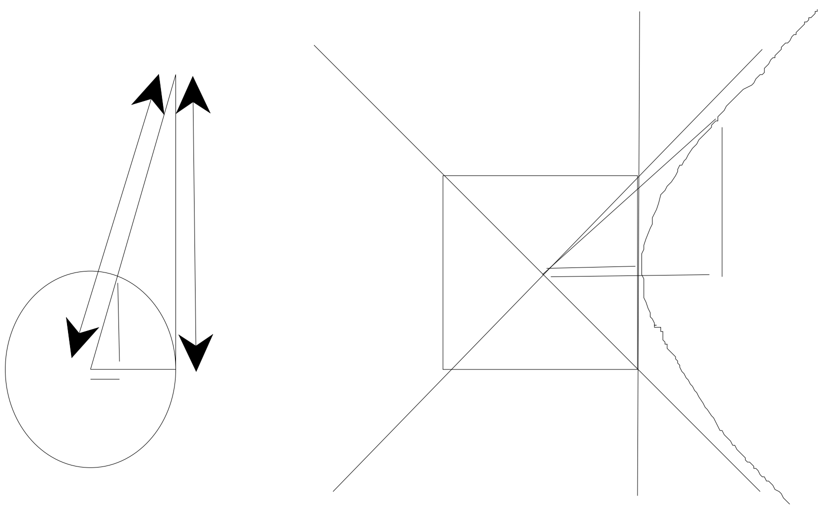


Figure 26: Arc length along the rectangular hyperbola defines hyperbolic trigonometry when the Lorentz metric is used.

The interesting properties of the exponential lie in its law of exponents. Notice that the angle, $\surd(\mathbf{v}\cdot\mathbf{v})$, is the norm of $\mathbf{v}$, and that imaginary quantities can be avoided by using trigonometric functions, such as should be done in association with the quaternion $\mathbf{i}$.

Consider, for unit vectors $\mathbf{u}$ and $\mathbf{v}$,

$$\exp(\alpha\mathbf{u})\exp(\beta\mathbf{v}) = (\mathbf{1}\cosh(\alpha) + \mathbf{u}\sinh(\alpha))(\mathbf{1}\cosh(\beta) + \mathbf{v}\sinh(\beta))$$

$$
\begin{aligned}
&= \mathbf{1}(\cosh(\alpha)\cosh(\beta) + \sinh(\alpha)\sinh(\beta)(\mathbf{u}\cdot\mathbf{v})) + \\
&\quad \mathbf{u}\sinh(\alpha)\cosh(\beta) + \mathbf{v}\cosh(\alpha)\sinh(\beta) + \\
&\quad (\mathbf{u}\times\mathbf{v})\sinh(\alpha)\sinh(\beta),
\end{aligned}
$$

and the prospects for seeing this as

$$
\exp(\gamma\mathbf{w}) = \mathbf{1}\cosh(\gamma) + \mathbf{w}\sinh(\gamma).
$$

Just define a new angle, $\cosh(\theta) = (\mathbf{u}\cdot\mathbf{v})$; then copy the two parts of the previous result:

$$
\begin{aligned}
\cosh(\gamma) &= \cosh(\alpha)\cosh(\beta) + \sinh(\alpha)\sinh(\beta)\cosh(\theta), \\
\mathbf{w}\sinh(\gamma) &= \mathbf{u}\sinh(\alpha)\cosh(\beta) + \mathbf{v}\cosh(\alpha)\sinh(\beta) + \\
&\quad (\mathbf{u}\times\mathbf{v})\sinh(\alpha)\sinh(\beta).
\end{aligned}
$$

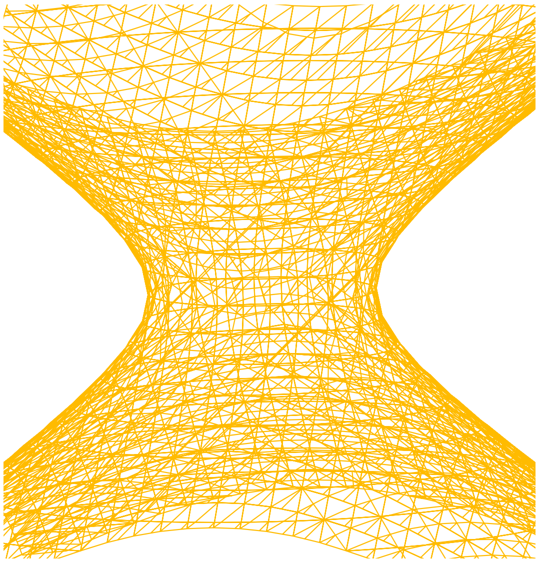


Figure 27: Quaternion exponentials and sequences of quaternion exponentials can be visualized in a nomogram based on the unit one-sheeted hyperboloid of radius squared -1.

7.6 Visualizing products

Addition of quaternions is not so much different from vector addition, but it would be nice to have a lore of logarithms by which products could be represented by sums. Due to the noncommutativity of the matrices involved, it could hardly be expected that simple addition would suffice. Nevertheless the noncommutativity is fairly mild with a trigonometric law of composition. For rotations, the result is a calculus of curved vectors on a spherical surface, whose analogue is what interests us here.

Scalar exponents give purely commutative results, so principal interest lies in the behavior of vector exponents, which results in a restriction to unimodular matrices. The product of two unimodular matrices is unimodular, so the system is closed; but a unimodular matrix is not a vector. Is there some way to have both? The answer lies in the observation that all vectors anticommute, for the matrices of $SL(2, R)$ as well as for Hamilton's quaternions. Therefore vor vectors $\mathbf{u}$ and $\mathbf{v}$,

$$e^{\phi \mathbf{u}} \mathbf{v} \quad = \quad \mathbf{v} e^{-\phi \mathbf{u}}, \tag{65}$$

and by a little cleverness,

$$e^{\phi \mathbf{u}} \mathbf{v} \quad = \quad e^{\frac{\phi}{2} \mathbf{u}} \mathbf{v} e^{-\frac{\phi}{2} \mathbf{u}}. \tag{66}$$

Since similarity transformation preserves traces, we conclude that not only is $|\mathbf{v}|$ unaltered by a unimodular multiplier, but also Trace($\mathbf{v}$). The difference between the trace and the determinant is the squared vector norm, leaving the useful result that vectors lying on a surface of given norm transform into vectors of the same norm as a consequence of $SL(2, R)$ multiplication. Therefore, there is a replacement for the unit sphere whose geometry was used in the study of rotations.

We need such a surface, that planes of all orientations will intersect it, letting exponential-representing arcs lie in actual intersecting planes, and so be joinable tip to tail. The hyperboloid of one sheet, the locus of vectors of squared norm -1, satisfies that requirement.

8 Episode I

It was hard not to notice this extraveganza of computer graphic techniques, so notice was taken.

9 Summary

- it is hard to read the de Bruijn diagrams. There is a graphing thesis. It needs to be interconnected to the NXLCAU programs.

- preparing the section of Rule 110 made it completely clear that Rule 110 is an exercise in tiling the plane with integer right triangles. They are something like polyominoes [89] except that the goal is not to cover the plane with certain objects of fixed area. Rather, the shape is fixed, the area variable, and a supplementary condition that top edges never continue one another. To do this, T0 triangles should be introduced, and diagonal cells should be withdrawn for all the others.

- the Faddeev computacion of the characteristic polynomial ought to go into the "ancestors" menu of the NXLCAU programs.

References

[1] Enrique Suárez Íñiguez, "Por qué la tesis de licenciatura debe desaparacer," *Ciencia* **50** 58-62 (Junio de 1999).

[2] Martín Aluja Schuneman, Rogelio Macías Ordíñez y Pedro Bosch Giral, "Por qué la tesis de licenciatura NO debe desaparacer," *Ciencia* **50** 45-49 (Septiembre de 1999).

[3] Shahen Hacyan, "Y usted, ¿cuántas citas tiene?" *Ciencia* **50** 54-57 (Junio de 1999).

Cross Ratio

[4] E. A. Maxwell, *The Methods of Plane Projective Geometry based on the use of General Homogeneous Coordinates*, Cambridge, at the University Press, 1963.

[5] E. A. Maxwell, *General Homogeneous Coordinates in Space of Three Dimensions*, Cambridge, at the University Press, 1961.

[6] Lancelot Hogben, *El Universo de los números*, Ediciones Destino, Barcelona, 1966, p. 145.

[7] Dirk J. Struik, *Lectures on Analytic and Projective Geometry*, Addison-Wesley Publishing Company, Inc., Cambridge, Massachusetts, 1953.

[8] Kenichi Kanatani, *Group-Theoretical Methods in Image Understanding*, Springer Verlag, Berlin, 1990 (ISBN 3-540-51253-5).

[9] Richard O. Duda and Peter E. Hart, *Pattern Classification and Scene Analysis,* John Wiley and Sons, New York, 1973 (ISBN 0-471-22361-1).

[10] Joseph L. Mundy and Andrew Zisserman (Eds.), *Geometric Invariance in Computer Vision,* The MIT Press, Cambridge, Massachussetts, 1992 (ISBN 0-262-13285-0).

[11] Joseph L. Mundy, Andrew Zisserman and David Forsyth (Eds.), "Application of Invariance in Computer Vision," *Lecture Notes in Computer Science 825,* Springer-Verlag, New York, 1994 (ISBN 0-387-58240-1).

[12] Thomas H. Reiss, "Recognizing Planar Objects Using Invariant Image Feartures," *Lecture Notes in Computer Science 676,* Springer-Verlag, New York, 1993 (ISBN 0-387-56713-5).

[13] T. Gevers and A. W. M. Smeulders, "Efficient recognition of 3-D rigid solid objects from 2-D projective images based on projective invariant descriptions," SPIE Volume 2357, pp. 296-302.

[14] Hazem F. Barakat and Edward M. Mikhail, "Object model construction by invariance and photogrammetry," SPIE Volume 3072, pp. 295-306.

[15] Eamon B. Barrett, Paul M. Payton, Peter J. Marra, and Michael H. Brill, "Geometric Interpretations of Algebraic Invariants in Images of 3D Scenes," SPIE Volume 3168 pp 286-299.

[16] Adolf Hurwitz and Richard Courant, *Vorlesungen über Allgemeine Funktionentheorie und Elliptische Funktionen*, Verlag von Julius Springer, Berlin, 1925.

Putzer's Method

[17] F. Silva Leite and P. Crouch, "Closed forms for the exponential mapping on matrix Lie groups based on Putzer's method," *Journal of Mathematical Physics* **40** 3561-3568 (1999).

[18] E. J. Putzer, "Avoiding the Jordan canonical form in the discussion of linear systems with constant coefficients," *American Mathematical Monthly* **73** 2-7 (1966).

[19] Cleve Moler and Charles van Loan, "Nineteen dubious ways to compute the exponential of a matrix," *SIAM Review* **20** 801-836 (1978).

[20] R. B. Kirchner, "An explicit formula for e^{At}," *American Mathematical Monthly* **74** 1200-1204 (1967).

[21] M. Kwapisz, "The power of a matrix," *SIAM Review* **40** 703-705 (1998).

[22] Shui-Hung Hou, "A simple proof of the Leverrier-Faddeev characteristic polynomial algorithm," *SIAM Review* **40** 706-709 (1998).

[23] I. E. Leonard, "The matrix exponential," *SIAM Review* **38** 507-512 (1996).

[24] Eduardo Liz, "A note on the matrix exponential," *SIAM Review* **40** 700-702 (1998).

[25] Jack Macki, "Three Notes on the Exponential of a Matrix and Applications," *SIAM Review* **40** 699-699 (1998).

[26] I. Gohberg, P. Lancaster, and L. Rodman, *Invariant Subspaces of Matrices with Applications*, John Wiley and Sons, Inc., New York, 1986 (ISBN 0-471-84260-5).

[27] A. I. Mal'cev, *Foundations of Linear Algebra*, W. H. Freeman and Company, San Francisco and London, 1963.

[28] Saber N. Elaydi and William A. Harris, Jr., "On the computation of A^N," *SIAM Review* **40** 965-971 (1998).

[29] Edward P. Fulmer "Computation of the matrix exponential," *American Mathematical Monthly* **82** 156-159 (1975).

[30] Tom M. Apostol, "Explicit formulas for solutions of the second-order matrix differential equation $Y'' = AY$," *American Mathematical Monthly* **82** 159-162 (1975).

[31] F. R. Gantmacher, *The Theory of Matrices*, volume 1, volume 2, Chelsea Publishing Company, New York, 1959.

Reversible Automata

[32] Jarkko Kari, "Representation of Reversible Cellular Automata with Block Permutations," *Mathematical Systems Theory* **29** 47-61 (1996).

[33] G. A. Hedlund, "Endomorphisms and automorphisms of the shift dynamical system," *Mathematical Systems Theory* **3** 320-375 (1969).

[34] Harold V. McIntosh, "Cellular Automata via de Bruijn diagrams," 1990.

[35] S. Amoroso and G. Cooper, "The Garden of Eden theorem for finite configurations," *Proceedings of the American Mathematical Society* **26** 158-164 (1970)

[36] S. Amoroso and Y. N. Patt, "Decision Procedures for Surjectivity and Injectivity of Parallel Maps for Tesselation Structures," *Journal of Computer and System Sciences* **6** 448-464 (1972).

[37] Mike Boyle, "Symbolic Dynamics and Matrices," in "Combinatorial and Graph-Theoretical Problems in Linear Algebra," IMA Volumes in Mathematics and its Applications, Vol. 50, Academic Press, New York, 1993, pp 1-38.

[38] Fan R. K. Chung, "Spectral Graph Theory," *CBMS Regional Conference Series in Mathematics Number 92*, American Mathematical Society, Providence, 1997.

[39] Dragos M. Cvetkovic, Michael Doob, and Horst Sachs, *Spectra of Graphs*, Academic Press, New York, 1979.

[40] P. Fatou, "Sur les équations fonctionelles (premiere mémoire)," *Bulletin de la Societe Mathematique de France* **47** 161-271 (1919).

[41] P. Fatou, "Sur les équations fonctionelles (deuxieme mémoire)," *Bulletin de la Societe Mathematique de France* **48** 33-94 (1920).

[42] P. Fatou, "Sur les équations fonctionelles (troisieme mémoire)," *Bulletin de la Societe Mathematique de France* **48** 208-314 (1920).

[43] G. A. Hedlund, "Endomorphisms and automorphisms of the shift dynamical system," *Mathematical Systems Theory* **3** 320-375 (1969).

[44] David Hillman "The structure of reversible one-dimensional automata," *Physica D* **52** 277-292 (1991).

[45] Jarkko Kari, "Reversibility and Surjectivity Problems of Cellullar Automata," *Journal of Computer and System Sciences* **48** 149-182 (1994).

[46] Jarkko Kari, "Representation of Reversible Cellular Automata with Block Permutations," *Mathematical Systems Theory* **29** 47-61 (1996).

[47] D. Koenig, *Theorie der endlichen und unendlichen Graphen*, Akademische Verlagsgesellschaft, Leipzig, 1936. Reprint: Chelsea Publishing Company, New York.

[48] Douglas Lind and Brian Marcus, *An Introduction to Symbolic Dynamics and Coding,* Cambridge University Press, Cambridge, 1995. (ISBN 0-521-55124-2 hardback).

[49] Masakazu Nasu, "Local Maps Inducing Surjective Global Maps of One Dimensional Tessellation Automata," *Mathematical Systems Theory* **11** 327-351 (1978).

[50] Stephen Wolfram, "Statistical mechanics of cellular automata," *Reviews of Modern Physics* **55** 601-644 (1983).

[51] Stephen Wolfram, "Universality and complexity in cellular automata," *Physica* **10D** 1-35 (1984).

[52] Stephen Wolfram, "Computation theory of cellular automata," *Communications in Mathematical Physics* **96** 15-57 (1984).

[53] Stephen Wolfram (Ed.), *Theory and Applications of Cellular Automata*, World Scientific Press, Singapore, 1986 (ISBN 9971-50-124-4 pbk).

Chaté-Manneville

[54] Bernard Barral, Hugues Chaté and Paul Manneville, "Collective behaviors in a family of high-dimensional automata," *Physics Letters A* **163** 279-285 (1992).

[55] R. Bidaux, N. Boccara, and H. Chaté, "Order of the transition versus space dimension in a family of cellular automata," *Physical Review A* **39** 3094-3105 (1989).

[56] Jan Hemmingsson, "A totalistic three-dimensional cellular automaton with quasiperiodic behaviour," *Physica A* **183** 255-261 (1992).

[57] J. Hemmingsson, A. Sørenson, H. Flyvbjerg and H. J. Herrmann, "What Synchronization?" *Europhysics Letters* **23** 629-634 (1993).

[58] H. Chaté and P. Manneville, "Evidence of Collective Behaviour in Cellular Automata," *Europhysics Letters* **14** 409-413 (1991).

[59] Hugues Chaté and Paul Manneville, "Collective Behaviors in Spatially Extended Systems with Local Interactions and Synchronous Updating," *Progress in Theoretical Physics* **87** 1-60 (1991).

[60] H. Chaté, A. Lemaître, Ph. Marcq, P. Manneville, "Non-trivial collective behavior in extensively-chaotic dynamical systems: an update," *Physica A* **224** 447-457 (1996).

[61] Hugues Chaté, Jérôme Losson, "Non-trivial collective behavior in coupled map lattices: A transfer operator perspective," *Physica D* **103** 51-72 (1997).

[62] Charles H. Bennett, G. Grinstein, Yu He, C. Jayaprakash and David Mukamel, "Stability of temporally periodic states of classical many body systems," *Physical Review A* **41** 1932-1935 (1990).

[63] Roger Bideaux, Nini Boccara and Hugues Chaté, "Order of transition versus space dimension in a family of cellular automata," *Physical Review A* **39** 3094-3105 (1989).

[64] P. -M. Binder and V. Privman, "Second-Order Dynamics in the Collective Temporal Evolution of Complex Systems," *Physical Review Letters* **68** 3830-3833 (1992).

[65] P. -M. Binder, "Domains and synchronization in high-dimensional cellular automata," *Physical Review E* **51** R839-R840 (1995).

[66] Tomas Bohr, G. Grinstein, Yu He, and C. Jayaprakash, "Coherence, Chaos, and Broken Symmetry in Classical, Many-Body Dynamical Systems," *Physical Review Letters* **58** 2155-2158 (1987).

[67] Hugues Chaté, "On the analysis of spatiotemporally chaotic data," *Physica D* **86** 238-247 (1995).

[68] Hugues Chaté, G. Grinstein and Lei-Han Tang, "Long-Range Correlations in Systems with Coherent (Quasi)periodic Oscillations," *Physical Review Letters* **74** 912-915 (1995).

[69] H. Chaté, A. Lemaître, Ph. Marcq, P. Manneville, "Non-trivial collective behavior in extensively-chaotic dynamical systems: an update," *Physica A* **224** 447-457 (1996).

[70] Hugues Chaté, Jérôme Losson, "Non-trivial collective behavior in coupled map lattices: A transfer operator perspective," *Physica D* **103** 51-72 (1997).

[71] H. Chaté and P. Manneville, "Evidence of Collective Behaviour in Cellular Automata," *Europhysics Letters* **14** 409-413 (1991).

[72] Hugues Chaté and Paul Manneville, "Collective Behaviors in Spatially Extended Systems with Local Interactions and Synchronous Updating," *Progress of Theoretical Physics* **87** 1-60 (1991).

[73] Hugues Chaté and Paul Manneville, "Emergence of Effective Low-Dimensional Dynamics in the Macroscopic Behaviour of Coupled Map Lattices," *Europhysics Letters* **17** 291-296 (1992).

[74] P. Fatou, "Sur les équations fonctionelles (premiere mémoire)," *Bulletin de la Societe Mathematique de France* **47** 161-271 (1919).

[75] P. Fatou, "Sur les équations fonctionelles (deuxieme mémoire)," *Bulletin de la Societe Mathematique de France* **48** 33-94 (1920).

[76] P. Fatou, "Sur les équations fonctionelles (troisieme mémoire)," *Bulletin de la Societe Mathematique de France* **48** 208-314 (1920).

[77] J. A. C. Gallas, P. Grassberger, H.J. Herrmann and P. Ueberholz, "Noisy collective behaviour in deterministic cellular automata," *Physica A* **180** 19-41 (1992).

[78] J. Gorodkin, A. Sørensen, and O. Winther, "Neural Networks and Cellular Automaton Complexity," *Complex Systems* **7** 1-24 (1994).

[79] G. Grinstein, "Stability of Nonstationary States of Many-Body Dynamical Systems," *Journal of Statistical Physics* **5** 803-815 (1988).

[80] Jens M. Houlrick, Itzhak Webman and Mogens H. Jensen, "Mean field theory and critical behavior of coupled map lattices," *Physical Review A* **41** 4210-4222 (1990).

[81] Jan Hemmingsson, "A totalistic three-dimensional cellular automaton with quasiperiodic behaviour," *Physica A* **183** 255-261 (1992).

[82] J. Hemmingsson and J. Herrmann, "On Oscillations in Cellular Automata," *Europhysics Letters* **23** 15-19 (1993).

[83] F. Jiménez-Morales and J. J. Luque, "Collective behavior of a probabilistic cellular automaton with two absorbing phases," *Physics Letters A* **181** 33-38 (1993).

[84] Anaël Lemaître, Hugues Chaté, and Paul Manneville, "Cluster Expansion for Collective Behavior in Discrete-Space Dynamical Systems," *Physical Review Letters* **77** 486-489 (1996).

[85] Jérôme Losson and Michael C. Mackey, "Coupled map lattices as models of deterministic and stochastic differential delay equations," *Physical Review E* **52** 115-128 (1995).

[86] Arkady S. Pikovsky and Jürgen Kurths, "Do Globally Coupled Maps Really Violate the Law of Large Numbers?" *Physical Review Letters* **72** 1644-1646 (1994).

[87] Y. Pomeau, "Periodic Behavior of Cellular Automata," *Journal of Statistical Physics* **70** 1379-1382 (1993).

Rule 110

[88] Matthew Cook, "Introduction to the activity of rule 110" (copyright 1994-1998 Matthew Cook) `http://w3.datanet.hu/ cook/Workshop/CellAut/Elementary/Rule110/110pics.html`

[89] Solomon W. Golomb, *Polyominoes* Charles Scribner's Sons, New York, 1965.

[90] Branko Grünbaum and G. C. Shepard, *Tilings and Patterns,* W. H. Freeman and Company, New York, 1987 (ISBN 0-7167-1193-1).

[91] Wentian Li and Mats G. Nordahl, "Transient behavior of cellular automaton rule 110," *Physics Letters A* **166** 335-339 (1992).

[92] Kristian. Lindgren and Mats. G. Nordahl, "Universal Computation in Simple One-Dimensional Cellular Automata," *Complex Systems* **4** 299-318 (1990).

[93] Harold V. McIntosh, "Wolfram's Class IV automata and a good Life," *Physica D* **45** 105-121 (1990).

[94] Grzegorz Rozenberg and Arto Salomaa, *Cornerstones of Undecidability,* Prentice Hall, New York, 1994 (ISBN 0-13-297425-8 pbk).

[95] W. T. Tutte, *Graph Theory As I Have Known It*, Clarendon Press, Oxford, 1998 (ISBN 0 19-850251-6).

Lorentz Contraction

[96] Serge Lang, SL_2(R), Addison Wesley Publishing Company, Reading, Massachussetts, 1975 (ISBN 0-201-04248-7)

[97] V. Bargmann, "Irreducible unitary representations of the Lorentz group," *Annals of Mathematics* **48** 568-640 (1947).

[98] E. Wigner, "On unitary representations of the inhomogeneous Lorentz group," *Annals of Mathematics* **40** 149-204 (1939).

Episode I

[99] Terry Brooks, *Star Wars: Episode I, the Phantom Menace*, Ballentine Publishing Company, 1999 (ISBN 0-345-42765-3).

[100] David West Reynolds, *Star Wars: Episode I, the Visual Dictionary,* D. K. Publishing Inc., New York, 1999 (ISBN 0-7894-4701-0).