Loop diassociativity has no finite basis

Warren D. Smith
WDSmith@fastmail.fm

January 30, 2005

Abstract — **We prove that no finite set of equations characterizes diassociativity in either finite or infinite loops, settling a 52-year-open problem.**

Contents

1 Introduction

1.1 History

Power-associativity has no finite equational basis in finite or infinite *loops*. Trevor Evans and B.H. Neumann [15] in 1953 gave a beautifully simple argument for that on their page 349, but it actually merely shows a lesser claim. §2 will provide the additional argumentation required to fill their gap.

On the other hand, in (not necessarily associative) *rings* with characteristic prime to 30, there *is* a finite equational basis for power-associativity [1], namely $x \cdot xx = xx \cdot x$ and $xx \cdot xx = x(x \cdot xx)$.

Evans and Neumann then asked whether *diassociativity* has a finite equational basis in loops. (Again, it does in rings: alternative, Moufang, and diassociative rings all are the same thing, according to pages 35-37 and 342-345 of [49].) The same problem was re-posed by M.Kinyon, K.Kunen, and J.D.Phillips [22] in 2001, and has been billed as among the top 5 unsolved problems in loop theory.[1]

We shall solve it here. There is no finite equational basis for loop-diassociativity, whether the loops are finite or infinite, or commutative or not.

1.2 Background on Loops

A *loop* is a set L equipped with a binary operation ab such that

1. There exists an identity element e so $ex = xe = x$ for all $x \in L$ and
2. There exists a unique solution x to $yx = z$ (usually denoted $x = y\backslash z$) and to $xy = z$ (usually denoted $x = z/y$).

(Colloquially: "a loop is a non-associative group.") Sometimes the loop operation is regarded as multiplication (in which case we usually call the identity 1), other times it is regarded as addition (in which case we usually call the identity 0). We shall use both notations in this paper.

We shall call a loop L *power-associative* if x^n is unambiguous for all integer n and[2] all $x \in L$. It is *diassociative* if any two elements of L generate a subgroup.

An *equational basis* for a property P is a set of equations which are both implied by, and imply, property P. Some other important properties P that a loop could obey include:

Moufang The Moufang property[3] $(x \cdot yz)x = xy \cdot zx$. Equivalent to obeying both the left-Bol $x(y \cdot xz) = (x \cdot yx)z$ and right-Bol $x(yz \cdot y) = (xy \cdot z)y$ properties.
Lalt the left-alternative law $x \cdot xy = xx \cdot y$;
Ralt the right-alternative law $yx \cdot x = y \cdot xx$;
Flex the flexible law $xy \cdot x = x \cdot yx$;
LIP the left-inverse-property $(1/x) \cdot xy = y$;
RIP the right-inverse-property $yx \cdot (x\backslash 1) = y$;
Antiaut the law $1/(xy) = (1/y)(1/x)$ of antiautomorphic inverses.

[1] Charles University in Prague, Czechoslovakia maintains an online list of the big ones.

[2] Warning: This power-associativity definition differs slightly from the one in the companion paper [44].

[3] There are 4 Moufang identities, all equivalent by lemma 3.1 p.115 of [6]. The other three are $x(yz \cdot x)x = xy \cdot zx$, $(xy \cdot z)y = x(y \cdot zy)$, and $y(z \cdot yx) = (yz \cdot y)x$.

We shall say that a loop obeying the Lalt, Ralt, and flexible laws is *alternative*[4]. A loop obeying the LIP, RIP, and Anti-aut laws (any two imply the third, in loops) has the *inverse property.* Finally, a loop which is both alternative and inverse property is *IP-alternative.* Note that each of these properties has a finite equational basis.

Much of the research in loop theory has taken the view that since loops are very much like groups, except for not being associative, the important question is just what weakened-associativity properties suffice to force loops to obey your favorite group-theorem (or perhaps some weakened version of it). Hence the presumed importance of investigating loop diassociativity (this paper) as well as other associative-resembling loop laws such as the Moufang, Bol, and alternative laws (famous older papers).

"Moufang's theorem" of 1933 states that Moufang loops are diassociative. (Indeed section VII.4 page 117 of [6] proves the stronger statement that in a Moufang loop, if $ab \cdot c = a \cdot bc$ then a, b, c generate a subgroup.)

Robinson's theorem of 1966 [38] states that left-Bol loops (and right-Bol loops also) are power-associative.

However, the reverse implications are false because of the 10-element counterexample in figure 1.1.

*	0	1	2	3	4	5	6	7	8	9
0	0	1	2	3	4	5	6	7	8	9
1	1	0	3	2	7	9	8	4	6	5
2	2	3	0	1	9	8	7	6	5	4
3	3	2	1	0	8	7	9	5	4	6
4	4	7	9	8	0	6	5	1	3	2
5	5	9	8	7	6	0	4	3	2	1
6	6	8	7	9	5	4	0	2	1	3
7	7	4	6	5	1	3	2	0	9	8
8	8	6	5	4	3	2	1	9	0	7
9	9	5	4	6	2	1	3	8	7	0

Figure 1.1. Unique 10-element Steiner loop. Diassociative and commutative and $xx = e$, but neither left- nor right-Bol. The 12 Steiner triples are the rows, columns, and generalized diagonals of $\left(\begin{smallmatrix}123\\456\\789\end{smallmatrix}\right)$. ▲

*	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G	H
0	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G	H
1	1	2	0	8	5	7	3	4	6	F	G	H	A	B	9	E	C	D
2	2	0	1	6	7	4	8	5	3	E	C	D	G	H	F	9	A	B
3	3	8	6	4	0	1	7	2	5	C	D	E	H	F	G	A	B	9
4	4	5	7	0	3	8	2	6	1	H	F	G	9	A	B	D	E	C
5	5	7	4	1	8	6	0	3	2	D	E	C	F	G	H	B	9	A
6	6	3	8	7	2	0	5	1	4	G	H	F	B	9	A	C	D	E
7	7	4	5	2	6	3	1	8	0	A	B	9	E	C	D	H	F	G
8	8	6	3	5	1	2	4	0	7	B	9	A	D	E	C	G	H	F
9	9	F	E	D	G	C	H	B	A	0	8	7	5	3	2	1	4	6
A	A	G	C	E	H	D	F	9	B	7	0	8	2	5	3	6	1	4
B	B	H	D	C	F	E	G	A	9	8	7	0	3	2	5	4	6	1
C	C	A	G	F	B	H	9	D	E	3	2	5	1	4	6	8	0	7
D	D	B	H	G	9	F	A	E	C	5	3	2	6	1	4	7	8	0
E	E	9	F	H	A	G	B	C	D	2	5	3	4	6	1	0	7	8
F	F	E	9	B	C	A	D	G	H	1	4	6	7	8	0	2	5	3
G	G	C	A	9	D	B	E	H	F	6	1	4	0	7	8	3	2	5
H	H	D	B	A	E	9	C	F	G	4	6	1	8	0	7	5	3	2

Figure 1.2. 18-element loop. IP-alternative and power-associative, but neither left-Bol, right-Bol, nor diassociative. Indeed, it disobeys $x(yy \cdot y) = (xy \cdot y)y$. (A counterexample loop related to, and perhaps even isomorphic to, this one was first found by J.D.Phillips.)

Raoul E. Cawagas helped me "beautify" this loop. All subloops are subgroups. The unique subgroup with 9 elements, namely $\{012345678\}$, happens to be both abelian (isomorphic to $C_3 \times C_3$) and a normal subloop. All subloops arise from C_6's, C_3's, C_2's, and their direct products, and therefore all are abelian groups of orders 2,3,6,9, with the exception of one *non*abelian subgroup $\{0789AB\}$, which is a dihedral group D_6, and also is a normal subloop. ▲

Plainly, diassociative loops are both power-associative and IP-alternative. But again the reverse implication is not valid, this time due to the 18-element counterexample in figure 1.2.

1.3 Background on the prime number theorem and two of its generalizations

The *prime number theorem* (PNT) states that the number of primes less than x is asymptotic to $x/\ln x$. This was originally conjectured by C.F.Gauss in about 1792. The PNT holds in certain nonrigorous probabilistic models of the prime numbers. This was perhaps what Gauss had in mind. Gauss's student G.F.B.Riemann showed how to prove the PNT rigorously under the assumption of the "Riemann hypothesis" (RH) that all the nonreal zeroes of the Riemann zeta function have real part 1/2. Although the RH remains unproved, Hadamard and de la Vallée Poussin in 1896 independently showed a large enough zero-free region to prove the PNT without depending on any unproven assumptions.

The *prime number theorem in arithmetic progressions* states that the number of primes below x in an arithmetic progression of the form $an + b$ with a, b fixed and $\gcd(a, b) = 1$,

[4] Some other authors have defined this word differently.

[5] The reader is warned that N.G. {Tsch, Tch, Ch, Č}ebotar{ev, eff, öw, ěv}'s name has been transliterated into English in many different ways. He proved his theorem in about 1926. A modern proof is in chapter 7.3 of [35]; its history is discussed in note 19 p.415-416. It is also discussed in [23] and [24]. Results considerably stronger even than the full Chebotarev theorem are available, e.g. under generalized Riemann hypotheses [4] [18] [26] [30] [36] [37] [41] [46] [47] [48].

is asymptotic to $x/(a \ln x)$. The proof [5] involves Dirichlet L-functions instead of the Riemann zeta function.

Let $P(y)$ be a fixed monic polynomial, with integer coefficients, of y. *Chebotarev's density theorem*[5] states that the number of primes $p < x$ such that the complete factorization of $P(y)$ mod p assumes any particular fixed form (for example: the form $9 = 1+1+1+2+4$ meaning that the degree-9 polynomial P is a product of 3 linears, 1 quadratic, and 1 quartic) compatible with (i.e. a refinement of) the form of its full factorization over the integers (e.g. in the same example: $9 = 3+6$ corresponding to a product of a cubic and sextic) also is asymptotic to $Cx \ln x$, where C is a positive constant depending on the polynomial. The proof now involves Artin L-functions instead of Dirichlet's.

The *generalized Riemann hypothesis* states that all of the nonreal zeros of these L-functions lie on the line with real part 1/2. Again this remains unproven, but in all cases large enough zero-free regions have been established to prove the appropriate generalized PNT.

More recently, "elementary" proofs of the PNT have been found [11][25][29] i.e. which employ neither complex analysis nor zeta functions. This was first done by A.Selberg and P.Erdös in about 1949. However, as far as I know, nobody has yet produced an elementary proof of the Chebotarev density theorem. Furthermore, problems such as "are there an infinite number of primes of the form n^2+1?" and "are there an infinite number of twin primes (pairs p, $p+2$, both prime)?" remain unresolved.

1.4 Background on the Amitsur-Levitski theorem and the Schwartz-Zippel lemma

The *Amitsur-Levitski theorem* [2][40] says that there are no degree-k identities obeyed by $N \times N$ matrices, if $k < 2N$, except for the ones implied by associativity and distributivity. (Here the inequality $k < 2N$ is tight since there *is* an extra identity of degree $2N$ found by Amitsur and Levitski.) It is totally understood what those identities are since as everyone knows there is a simplification-to-fully-expanded canonical-form procedure to verify or disprove any such identity.

Thus, by making N large enough, any such putative "extra" identity will be violated in the multiplicative group generated by two generic real $N \times N$ matrices.

The *Schwartz-Zippel lemma* [42][50] says that for any set I of alleged polynomial (or rational) identities that hold in the real numbers,

1. They also hold in $\mathbb{Z}_p$, the integers modulo a prime p, if p is sufficiently large
2. If I is not a valid identity, then random variable values will cause it to be violated in $\mathbb{Z}_p$ with a probability that goes to 1 as p is made large.

Thus, the previous paragraph about the "group generated by two generic real $N \times N$ matrices," is also valid for the "group generated by two random $N \times N$ matrices over $\mathbb{Z}_p$, with probability→ 1 when p is chosen sufficiently large and prime."

2 Evans, Neumann, and power-associativity in loops

Theorem 1 (Evans & Neumann [15] p.349). *No finite set of power-associativity statements (that $x^n = x^n$ for some finite set of positive integers n and parenthesizations of the two sides) suffice to imply power-associativity in loops, not even in finite and/or commutative loops.*

Proof: Let the maximum degree n of the power-associativity statements be D. Construct the $(2N) \times (2N)$ Cayley +table of the additive group of integers mod $2N$. Circle all the entries in that table which arise from powers$\le D$ of the group elements. There will then be $\le 2(D+1)N$ circled entries among the $4N^2$ total entries. Now find 4 uncircled entries in the table forming the vertices of an axis-oriented square of sidelength N. This square is specified by choosing its upper left corner, and if this corner is chosen at random, then the expected number of circled entries among the 4 entries that are vertices of the square, is $\le 2(D+1)/N$. This goes to 0 as $N \to \infty$. Since there must exist a square at or below the expectation value, there must exist one with no vertices circled if N is sufficiently large, namely if $N > 2(D+1)$. Find it. Now take its 4 vertex-entries and permute them as follows:

$$\begin{pmatrix} a & b \\ b & a \end{pmatrix} \to \begin{pmatrix} b & a \\ a & b \end{pmatrix}. \qquad (1)$$

The result is the +table of a non-power-associative, but still commutative, loop obeying all the power-associativity statements in the putative "basis." □

Remark. It also is possible to modify the Evans-Neumann proof to make their loops incorporate violations of commutativity: alter a 4×4 submatrix of the abelian cyclic group +table, rather than a 2×2 one. Or just take the direct product of their commutative loop with some non-commutative group. The latter idea will also be applicable to the diassociativity result of the present paper.

The Evans and Neumann argument, although valid for the purpose of proving theorem 1, is insufficient for the purpose of proving power-associativity has no finite equational basis. That is because they forgot to consider the possibility of basal equations involving *more* than 1 variable.

Furthermore, since loops involve three operators \, /, and $\cdot$, really the basal equations should be permitted to include them also. It is equivalent, but more convenient, to restrict ourselves to pure-polynomial basal equations but allowing them to be linked via the word "where." For example, the "left inverse property" that $x \cdot yz = z$ *where* $xy = 1$ is expressed by two logically linked polynomial equations; the latter equation says x is the left-inverse of y. We shall allow using "where" in this way in all bases considered in the remainder of this paper; so doing allows eliminating all occurences of / and \ in the basis. It is easy to see that allowing this does not affect the Evans and Neumann proof above (where now expressions for x^n are allowed to be of the form, e.g., x^{n+3}/x^3) except that the meaning of "sufficiently large N" may now be larger.

Theorem 2 (Repair to Evans-Neumann). *Any equation not of the form $x^n = x^n$ (with different parenthesizations of its two sides) cannot be a member of any basis for loop power-associativity.*

Part I of two-part proof: No equation of the form $x^a = x^b$ for nonnegative unequal integers a, b can be a member of a basis for power-associativity. That is because it is falsified in the additive group of reals, or of integers, and hence (by Schwartz-Zippel) also in the integers mod P for any sufficiently large prime P. □

Before we can prove part II of the proof of theorem 2, we need to discuss latin squares. A *latin square* is an $N \times N$ array of numbers from the set $\{1, 2, 3, \dots, N\}$ with no repeated number in any row or column. An *incomplete* $r \times r$ latin square based on N is an $r \times r$ array of either of numbers from $\{1, 2, 3, \dots, N\}$ or blank entries, with no repeated number in any row or column.

The simplest embedding theorem for latin squares[6] [14] states that any incomplete $r \times r$ latin square based on N may be completed (by filling in the blanks and adding $N - r$ extra rows and columns) to yield an $N \times N$ latin square, *provided* $N \geq 2r$ (and this inequality is best possible if $r \geq 4$). The same reference also proves this theorem if the latin squares are required to be the multiplication tables defining a *loop*.

Part II of proof of theorem 2: If the equation I involves at least *two* variables x, y (and cannot be re-expressed with fewer[7], e.g. is not just a statement of form $(xy)^a = (xy)^b$) then we shall explain how to construct a power-associative loop in which I is false.

The $3P - 2$ elements of the loop are 0 and x_c where x is a nonzero integer mod P and c is a color from the set {red,blue,green}. The loop operation $*$ is defined as follows:

1. $0 * 0 = 0$.
2. $x_c * 0 = 0 * x_c = x_c$.
3. $x_c * y_c = (x + y)_c$.
4. if $a \neq b$ then $x_a * y_b = (L_{xy})_c$ where c is the third color and L is a $(P-1) \times (P-1)$ latin square on the nonzero integers modulo P.

The fact that this *is* a loop may be confirmed by showing that all right-division problems have a unique solution (the proof for left-division is the same only mirrored). We have $0/x_c = -x_c$, $x_c/0 = x_c$, $x_c/y_c = (x - y)_c$, and if $a \neq c$ then $z_c/y_b = x_a$ where a is the third color and x is chosen so that $L_{xy} = z$.

The latin square L is irrelevant to power-associativity since all powers of anything always have the same color.

We now may use the previously mentioned latin square embedding theorem to construct a suitable latin square L in which I is false: if I wants some particular $pq = s$ to be true we simply make the pqth entry of the incomplete latin square be something other than s (which plainly is always possible if the square is large enough compared to the size of the identity I) and then complete the square. □

Side Remark. Embedding theorems for latin squares can also be applied to loop theory in other ways. For example, J.D.Phillips asked what the possible sizes of commutative exponent-2 loops were. His computer search showed that an n-element loop of this type exists for all *even* $n \leq 10$, but no odd ones. That suggested

Theorem 3 (Commutative exponent-2 loops have an even number of elements). *A loop obeying $xy = yx$, $1x = x1 = 1$, and $xx = 1$ must have an even number of elements.*

Proof: This is an easy consequence of a much more general known theorem: Hoffman [19] gave necessary and sufficient conditions for completing incomplete commutative $n \times n$ latin squares with prescribed diagonals. We may use his necessary conditions to rule out the case of odd n: following the notation in the mathematical review of [19] and always letting y denote an element of L with $y \neq 1$, put $r = n$, $c(1) = d(1) = n$, $t(1) = 0$ and $t(y) = d(y) = 0$ and $c(y) \geq 2$. Then one of Hoffman's necessary conditions gives $d(y) + t(y) = 0 \equiv n (\text{mod } 2)$, which rules out odd n. □

One might then ask whether theorem 3 has any generalization to loops with $xxx = 1$ (exponent 3), or non-commutative loops, but apparently the answers are *no*, because there is a unique 5-element loop obeying $xx = 1$ (given in the book [20] or readily constructed by hand), and because figure 2.1 gives a 7-element commutative and power-associative loop with exponent 3.

*	0	1	2	3	4	5	6
0	0	1	2	3	4	5	6
1	1	2	0	4	3	6	5
2	2	0	1	5	6	3	4
3	3	4	5	6	1	2	0
4	4	3	6	1	5	0	2
5	5	6	3	2	0	4	1
6	6	5	4	0	2	1	3

Figure 2.1. 7-element loop found by J.D.Phillips using `mace4`: $xx \cdot x = x \cdot xx = 1$, power-associative, commutative (and hence flexible), but not Lalt, Ralt, LIP, RIP, nor antiaut. ▲

Clearly, any diassociative loop obeys the alternative laws, the inverse property, and power-associativity, but as we have seen, the reverse need not be the case.

The question now arises whether any IP-alternative loop is automatically power-associative.

The answer is no, although the smallest counterexample is remarkably large. I found it by a combination of human reasoning and exhaustive computer searching with W.McCune's powerful backtrack search program `mace4`.

Theorem 4 (IP-alternativity does not imply power-associativity). *There is a unique loop with ≤ 35 elements which obeys the alternative laws and the inverse property but is not power-associative. It is commutative, has 27 elements, and is given in figure 2.3.*

Proof: Let PA(n) denote the statement that in a (multiplicative) loop, x^n is unambiguous. We first show that in an Alt loop which is PA(d) for all d with $0 \leq d < a + b$,
Tool#1. If $a, b > 0$ then $x^a x^b = x^b x^a$.
Tool#2. If $0 \leq a < b$ then $x^a x^b = x^{2a} x^{b-a}$.
(Proofs: flexibility, combined with the inductive PA assumption that x^k is unambiguous for all $k < a + b$, shows

[6] Many extensions of this theorem are available, e.g. to latin cubes, e.g. see [19][39][43][9].

[7] We shall discuss this notion more detail in the proof of theorem 7.

$x^c \cdot x^d x^c = x^c x^d \cdot x^c$ and now let $a = c$ and $b = c + d$ to get tool#1. Left-alternativity and the inductive PA assumption give $x^a \cdot x^a x^b = x^a x^a \cdot x^b$, proving tool#2.) Now PA(1) and PA(2) always hold. In any flexible loop $x^2 x^1 = x^1 x^2$ so we have PA(3). In any Alt loop $x^3 x^1 = x^1 x^3 = x^2 x^2$ so we have PA(4). In any Alt loop $x^1 x^4 = x^2 x^3 = x^3 x^2 = x^4 x^1$ so we have PA(5). However, we hit a snag at PA(6): all our tools give us is that $x^1 x^5 = x^2 x^4 = x^4 x^2 = x^5 x^1$, while $x^3 x^3$ might conceivably differ; there might be two different equivalence classes of 6th powers. If we temporarily ignore that snag and continue on, we find that the possible equivalence classes for PA(n) for $n = 2, 3, \ldots, 36$ (at each n under the assumption of PA(d) for all d with $0 < d < n$) are given in table 2.2.

We now asked `mace4`[8] to search for n-element Alt loops for $2 \leq n \leq 150$ disobeying $x^3 x^3 = x^1 x^5$. It reported that it had completed an exhaustive search, finding that no such loops exist.

In retrospect, that was not surprising, because in fact, in Lalt and Ralt loops, $x^3 x^3 = x^1 x^5$. To prove that, first realize that[9]

$$x \cdot x(x \cdot xy) \underset{L}{=} xx(xx \cdot y) \underset{L}{=} (xx \cdot xx)y \underset{L}{=} (x \cdot x[xx])y. \qquad (2)$$

Now use this leftmost and rightmost sides of this identity with $y = x \cdot xx$ to get

$$x(x \cdot x[x \cdot xx]) = (x[x \cdot xx])(x \cdot xx) \underset{R}{=} x \cdot (x \cdot xx)(x \cdot xx). \qquad (3)$$

Finally left-cancel the x's from the leftmost and rightmost sides of that identity to get $x^1 x^5 = x^3 x^3$.

n	equiv classes for $x^a x^b$ as set of allowable a
2	{1}
3	{1}
4	{1}
5	{1}
6	{1, 3}
7	{1}
8	{1}
9	{1, 3}
10	{1, 5}
11	{1}
12	{1, 3}
13	{1}
14	{1, 7}
15	{1, 3, 5}
16	{1}
17	{1, 3}
18	{1, 3, 9}
19	{1}
20	{1, 5}
21	{1, 3, 7}
22	{1, 11}
23	{1}
24	{1, 3}
25	{1, 5}
26	{1, 13}
27	{1, 3, 9}
28	{1, 7}
29	{1}
30	{1, 3, 5, 15}
31	{1, 3, 5}
32	{1}
33	{1, 3, 5, 11}
34	{1, 3, 17}
35	{1, 5, 7}
36	{1, 3, 9}

Figure 2.2. (Computer-generated) table of equivalence classes resulting from tools #1 and #2. ($n = a + b$.) ▲

So we next asked `mace4` to search for n-element IP-Alt loops for $2 \leq n \leq 38$ disobeying $x^1 x^8 = x^3 x^6$. It reported that it had completed an exhaustive search, finding that the only such loops have either 27 or 36 elements, and the 27-element loop is *unique*, namely the one in figure 2.3.

(If we instead do not require IP and just ask `mace4` to seek n-element Alt loops for $2 \leq n \leq 23$ disobeying $x^1 x^8 = x^3 x^6$, then it reports that there are exactly two such loops[10] both with 21 elements, and both commutative.)

If we now continue on through the PA-equivalence-class table, asking `mace4` next for the n-element Alt loops for $2 \leq n \leq 100$ disobeying $x^1 x^9 = x^5 x^5$, $x^1 x^{13} = x^7 x^7$, $x^1 x^{14} = x^3 x^{12} = x^5 x^{10}$, $x^1 x^{16} = x^3 x^{14}$, $x^1 x^{17} = x^3 x^{15} = x^9 x^9$, $x^1 x^{17} = x^5 x^{15}$, $x^1 x^{20} = x^3 x^{18} = x^7 x^{14}$, $x^1 x^{21} = x^{11} x^{11}$, $x^1 x^{23} = x^3 x^{21}$, $x^1 x^{24} = x^5 x^{20}$, or $x^1 x^{25} = x^{13} x^{13}$, it always reports that no such loop exists.

[8] As supplied, `Mace4` had a built-in loop-size limit of 100, which we changed to 200 by recompiling the code. W.McCune, the author of `mace4`, had not imagined that it could exhibit such tremendous power that it could perform exhaustive searches over all 100-element loops. However, the constraints in our present problem are so severe and enable so much pruning of the backtracking that `mace4` is able to handle all loop sizes ≤ 150 in under 10 minutes.

[9] "$\underset{L}{=}$" means "= as may be realized from the Lalt law.

[10] One of them is given in the companion paper [44]. They may be isomorphic.

Finally, $x^1x^{11} = x^3x^9$ is violated by our unique 27-element IP-Alt loop, but by no other IP-alt loops with ≤ 35 elements. (It also is violated by both the 21-element alt loops, but no others with ≤ 23 elements.) □

*	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	R	S
0	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	R	S
1	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	R	S	0
2	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	R	S	0	1
3	3	4	5	6	7	8	J̈	A	B	M̈	D	E	F	G	H	9̈	K	L	C̈	N	P	Q	R	S	0	1	2
4	4	5	6	7	8	9	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	R	S	0	1	2	3
5	5	6	7	8	9	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	R	S	0	1	2	3	4
6	6	7	8	J̈	A	B	C	D	E	Q̈	G	H	9̈	K	L	M	N	P	F̈	R	S	0	1	2	3	4	5
7	7	8	9	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	R	S	0	1	2	3	4	5	6
8	8	9	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	R	S	0	1	2	3	4	5	6	7
9	9	A	B	M̈	D	E	Q̈	G	H	J	K	L	3̈	N	P	6̈	R	S	0	1	2	C̈	4	5	F̈	7	8
A	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	R	S	0	1	2	3	4	5	6	7	8	9
B	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	R	S	0	1	2	3	4	5	6	7	8	9	A
C	C	D	E	F	G	H	9̈	K	L	3̈	N	P	Q	R	S	0	1	2	M̈	4	5	6	7	8	J̈	A	B
D	D	E	F	G	H	J	K	L	M	N	P	Q	R	S	0	1	2	3	4	5	6	7	8	9	A	B	C
E	E	F	G	H	J	K	L	M	N	P	Q	R	S	0	1	2	3	4	5	6	7	8	9	A	B	C	D
F	F	G	H	9̈	K	L	M	N	P	6̈	R	S	0	1	2	3	4	5	Q̈	7	8	J̈	A	B	C	D	E
G	G	H	J	K	L	M	N	P	Q	R	S	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
H	H	J	K	L	M	N	P	Q	R	S	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G
J	J	K	L	C̈	N	P	F̈	R	S	0	1	2	M̈	4	5	Q̈	7	8	9	A	B	3̈	D	E	6̈	G	H
K	K	L	M	N	P	Q	R	S	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G	H	J
L	L	M	N	P	Q	R	S	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G	H	J	K
M	M	N	P	Q	R	S	0	1	2	C̈	4	5	6	7	8	J̈	A	B	3̈	D	E	F	G	H	9̈	K	L
N	N	P	Q	R	S	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G	H	J	K	L	M
P	P	Q	R	S	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G	H	J	K	L	M	N
Q	Q	R	S	0	1	2	3	4	5	F̈	7	8	J̈	A	B	C	D	E	6̈	G	H	9̈	K	L	M	N	P
R	R	S	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q
S	S	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	R

Figure 2.3. The unique (< 36)-element IP-Alt loop which is not power-associative (since 1+8=9≠J=3+6). Entries $a+b$ not agreeing with integer addition mod 27 have been decorated with umlauts (M̈ versus M). Note that these exceptions occur only on the index-3 subgrid and that the diagonal entries $a+a$, the first row and column $0+a = a+0$, and the antidiagonal $(-a)+a = 0$ never are umlauted. This loop has 27 elements and is commutative, Lalt, Ralt, Flexible, LIP, RIP, antiaut, but not power-associative, L-Bol, nor R-Bol. ▲

The 27-element loop in the preceding proof plainly exhibits a very interesting structure – and both of the 21-element loops mentioned (but not exhibited) there also have the same kind of structure. Understanding this structure is the key to all the results that will follow that will ultimately settle the loop-diassociativity-basis problem. Indeed, all of the constructions that will follow are successively more and more refined variants of the same construction idea. That idea, in an abstract form from which all inessential details have been removed, is this:

Lemma 5 (Construction of IP-alt but not power-associative loops). *Let $(\mathbb{Z}_N, +)$ be the additive (cyclic) group of integers modulo some composite number N. Let $(H, +)$ be a nontrivial subgroup (necessarily cyclic with $1 < |H| < N$). Let π be a permutation of the elements of H. Define $(H, \circ)$ to be the new group isomorphic to $(H, +)$ via π, that is $x \circ y = \pi\{\pi^{-1}(x) + \pi^{-1}(y)\}$. Choose π so that $x + x = x \circ x$ for all $x \in H$, so that the two-sided inverse x^{-1} is the same in both groups [i.e. so $\pi(-x) + \pi(x) = 0$], and so that $(H, +) \neq (H, \circ)$, i.e. so that there are $x, y \in H$ with $x + y \neq x \circ y$. Further assume $x \notin H$ implies $x + x \notin H$, or equivalently that the index of H in $\mathbb{Z}_N$ is odd. Define a new operation $*$ on $\mathbb{Z}_N$ as follows:*

$$x * y \overset{\text{def}}{=} \begin{cases} x \circ y & \textit{if } x, y \in H \\ x + y & \textit{otherwise.} \end{cases} \tag{4}$$

Then $(\mathbb{Z}_N, *)$ is an IP-alternative but not power-associative loop.

Proof: To show flexibility $x*(y*x) = (x*y)*x$: if $x, y \in H$ this is obvious since H is a group. If $x \notin H$ then $x*(y*x) = x*(y+x) = x+(y+x) = (x+y)+x = (x+y)*x = (x*y)*x$.

To show Lalt $(x*x)*y = x*(x*y)$: If $x, y \in H$ this is again obvious since H is a group. if $x \notin H$ then $x*x = x+x \notin H$ and so $(x*x)*y = (x+x)*y = (x+x)+y = x+(x+y) = x*(x*y)$. Finally if $x \in H$ and $y \notin H$ then $(x*x)*y = (x \circ x)*y = (x+x)+y = x+(x+y) = x*(x*y)$ since $x \circ x = x+x$.

Ralt is shown in the same (but mirrored) way.

To show LIP: x^{-1} agrees in $(\mathbb{Z}_N, +)$ and in our loop. Now to show $x^{-1}*(x*y) = y$: If $x, y \in H$ the property is obvious since H is a group. If $x \notin H$ we have $x^{-1}*(x*y) = x^{-1}+(x+y) = y$. If $x \in H$, $y \notin H$ we have $x^{-1}*(x*y) = y$ since $x*y = x+y \notin H$.

RIP is shown in the same (but mirrored) way.

Finally, to show that power-associativity does not hold: Let g be a generator of G and h of H. Since $(H, \circ) \neq (H, +)$ there are exponents a, b with $h^a * h^b = h^a \circ h^b \neq h^a + h^b = h^{a+b}$. Writing $h = g^c$ we have that powers of g are ambiguous. □

Example. One may readily check that π defined by[11]

x	0	1	2	3	4	5	6	7	8
$\pi(x)$	0	2	4	3	8	1	6	5	7

obeys $\pi(x+x) = \pi(x) = \pi(x)$ and $\pi(-x) = -\pi(x)$ both mod 9; and $x \circ y = \pi\{\pi^{-1}(x) + \pi^{-1}(y)\}$ does not coincide with $(\mathbb{Z}_9, +)$ since $1 \circ 2 = \pi(5+1) = 6 \neq 3$. Therefore this choice of π using as H the 9-element subgroup of $\mathbb{Z}_{27}$ consisting of the multiples of 3 yields an IP-alternative 27-element loop that is not power-associative; in fact it is the loop in figure 2.3.

Theorem 6 (An infinite set of different IP-alternative but not power-associative loops). *Let P be a prime such that the set $\{+2, -2, -1\}$ does not suffice to generate the multiplicative group of integers modulo P (for example if ± 2 both are squares modulo P). There are an infinite number of such primes. Then there is a loop with $N = 3P$ elements which is both IP-alternative and commutative but not power-associative.* [12]

Proof: The fact that an infinite (indeed, constant density) suitable subset of primes exist is a consequence of the Chebotarev density theorem using $P(y) = (y^2 - 2)(y^2 + 2)$. The ones less than 100 are $P = \{17, 31, 41, 43, 73, 89, 97\}$ and 423 among the first 1000 primes work.

We employ the construction of the Lemma using as H the P-element cyclic subgroup of $\mathbb{Z}_N$ consisting of the multiples of 3.

[11] Equivalently, $\pi(x) = 2x$ mod 9, except that $\pi(x) = x$ if x is a multiple of 3.

[12] It is also possible to construct IP-alternative but not power-associative loops which feature *non*commutativity, by, e.g. taking the direct product of one of the present theorem's loop constructions with a nonAbelian group.

Why does a suitable permutation π exist? Because the *multiplicative* group of integers mod P is not generated by $\{+2,-2,-1\}$, they therefore generate some subgroup. Then the permutation can interchange some of the cosets of this subgroup without destroying the agreement of doubling and of negation in our subgroup, with those in the integers mod $3P$. (Think of the nonzero residues mod P as being the vertices of a graph whose edges link x to $2x$ and $-x$ modulo P. The different cosets correspond to *disconnected* components of this graph, and thus interchanging them cannot destroy any relationship corresponding to a graph edge.) Conversely, if $\{+2,-2,-1\}$ *do* generate the full multiplicative group mod P, then there is no freedom to permute its elements in any manner and no suitable π exists. □

Remarks. As a check, I have explicitly constructed loops of orders $51 = 3 \cdot 17$ and $93 = 3 \cdot 31$ that arise as the first two cases of the construction in this proof. We have already mentioned our exhaustive computer searches verifying that all IP-alt loops with N elements are power associative if $N \le 35$ and $N \neq 27$. These forbidden N include $9 = 3\times3$, $15 = 3\times5$, $21 = 3\times7$, and $33 = 3\times11$, all of which arise from primes P *not* of the form specified in theorem 6, c.f. the last sentence of the proof.

Of course, the same sort of construction as in our proof can work even for *non*prime[13] P. Thus our 27-element loop in figure 2.3 arose from $P = 9$, and I have also constructed loops of this sort with $36 = 3 \cdot 12$ and $45 = 3 \cdot 15$ elements. Also, both the 21-element alternative (but *not* IP) loops mentioned in the proofs of theorem 4 are of the same sort, except that there the prime $P = 7$ is required *only* to have multiplicative group not generated by 2, since the need for -1 and therefore -2 arise from the inverse property, which we here aren't imposing. (Verification: $\{2, 4 = 2^2, 1 = 2^3\}$ indeed is a subgroup of $\{1,2,3,4,5,6\}$ mod 7.) We do not claim to understand the full set of (possibly nonprime) integers P that are permitted and the full set of permitted permutations π. Therefore the present theorem has restricted itself to the (easier to understand) cases where P is prime.

3 Negative solution of loop diassociativity basis problem

This section will prove that there is no finite basis[14]for diassociativity in loops.

3.1 There is no finite basis for diassociativity in (possibly infinite) loops, even commutative ones

Our proof proceeds in two steps:

Theorem 7 (The only things the basis could contain). *Any equation which is not equivalent to an obvious diassociativity statement (e.g. necessarily involving ≥ 3 variables, or if it involves ≤ 2 variables with different orderings of or numbers of those two variables in the products on the left and right side of the identity) is falsified by some diassociative loop.*

Corollary 8 (What the basis must be). *Hence if there is a finite equational basis for loop-diassociativity, it must consist solely of (≤ 2)-variable statements, and hence must simply be all the diassociativity statements of degree$\le D$, for some finite D.*

Theorem 9 (Insufficiency of putative basis). *Any set of (≤ 2)-variable equations is insufficient even to imply power-associativity in (possibly infinite) loops, even in commutative ones.*

Theorem 10 (Main result$_\infty$). *There is no finite equational basis for loop-diassociativity, even in commutative loops.*

Proof: Corollary of the last two theorems. □

Note. We call theorem 10 "main result$_\infty$" since it allows the loops to be infinite. For the extension of the main result to cover *finite* loops, see §3.2. Indeed §3.2 largely obsoletes the present section; nevertheless the present section has not been discarded because it illustrates an interesting technique and because I doubt I would have thought of §3.2 if I had not first discovered theorem 10. For discussion of possible further extension to allow power-associativity, see §3.3.

Proof of theorem 7. First we need to make the theorem statement more precise so we know exactly what we are proving. The reason this is tricky may be comprehended by considering a few examples. The identity

$$(ab)c \cdot (c \cdot ab) = ab \cdot c(c \cdot ab)$$

is not really a 3-variable identity because a and b only occur in the combination ab, which we may rename d, getting

$$dc \cdot cd = d(c \cdot cd),$$

which is an obvious disassociativity statement. Similar, but slightly more difficult, is

$$[(a \cdot bb)c \cdot c(a \cdot bb)]f = [(ab \cdot b)c \cdot c(ab \cdot b)]f$$

which by right-cancelling the f's, renaming $d = a \cdot bb$, and then realizing that $d = ab \cdot b$ is an obvious diassociativity consequence, also is seen to be an obvious diassociativity consequence.

So: by an "obvious diassociativity statement" we shall mean an identity of the form $abaabbb = abaabbb$, i.e. equating two products of two variables only, both products having the same length and with the variables in them occurring in the same order, *but* the two sides possibly could be parenthesized differently.

We call two identities "diassociativity equivalent" if one may be converted into the other either by

1. renaming some common subexpression, or
2. using an obvious diassociativity statement to convert some subexpression of one, which involves only two variables, into another expression that is a product of the same two variables in the same order, but perhaps parenthesized differently, or

[13] We have concentrated on prime P merely because primes are easier to understand – the multiplicative structure mod P is a cyclic group for prime P – and because of the availability of the Chebotarev density theorem (which is about primes) for seeing an infinite set of examples exist.

[14] I.e., there is no finite set of polynomial equations implied by and implying diassociativity. As usual we shall allow the polynomial equations to be logically linked by means of the word "where," which shall not affect our reasoning.

3. Cancelling some common multiple,
4. Multiplying both sides by some common multiple, or
5. some finite chain of steps of the preceding four forms.

Similarly two identities are "diassociativity-commutativity equivalent" if one may be converted into the other by the above plus perhaps using commutativity.

An identity which has the fewest variables among all its diassociativity-equivalent forms, will be said to be "minimal."

Then our more precise statement of theorem 7 is that
(i) any 1-variable or 2-variable identity which is not an obvious diassociativity statement is violated in some diassociative loop.
(ii) any minimal identity with ≥ 3 variables is violated in some diassociative loop.

Proof of i: There are two subcases:

1. The two words of x's and y's on each side of the identity have different numbers of x's and/or y's – for example
$$xxyy = xxxy.$$
Any such identity is violated in the additive subgroup of $\mathbb{R}$ generated by 1 and $\sqrt{2}$. Therefore no such identity is allowed to be in a diassociativity basis.
2. The two words have the same numbers of x's and y's on both sides, but they occur in a different order, for example
$$xyxxxy = xxyxyx.$$
By Amitsur-Levitski, if $N > 2k$ then any such k-term identity is violated in the multiplicative group generated by two generic real $N \times N$ matrices. Therefore no such identity is allowed to be in a diassociativity basis.

Proof of ii: For a long time it puzzled me how to show the seemingly "clear" fact that there cannot be any (≥ 3)-variable minimal identity satisfied in all finite diassociative loops. The solution strategy is, for each such putative identity, to construct a diassociative loop violating that identity[15]. The beautiful construction that solves this puzzle was found buried deep in a rather obscure paper by Hart and Kunen (it is theorem 5.2 in [17]) where it was used for a different purpose.

The Hart-Kunen construction. Let p be a prime. Consider the additive group $(\mathbb{Z}_p^n, +)$ of integer n-vectors modulo p. We shall define a loop-operation $*$ on the p^n elements of this set.

For each 2D subspace T of $\mathbb{Z}_p^n$ (defined by 2 nonzero points, not multiples of each other mod p) pick a "radial" map Ψ_T. More precisely, this map is defined by, for each 1D subspace S of T, (defined by 1 nonzero point) choosing a nonzero "stretching constant" $C_{ST} \in \mathbb{Z}_p$; then $\vec{x} \in T$ is mapped to $C_{ST}\vec{x}$ where S is the unique 1D subspace containing $\vec{x}$ if $\vec{x} \neq \vec{0}$. Meanwhile $\vec{0}$ is of course mapped to itself. Then
$$\vec{x} * \vec{y} \overset{\text{def}}{=} \Psi_T^{-1}[\Psi_T(\vec{x}) + \Psi_T(\vec{y})] \qquad (5)$$
where T is the 2D subspace containing both $\vec{x}$ and $\vec{y}$. Note this uniquely defines T unless $\vec{x}$ and $\vec{y}$ are linearly dependent in which case we get $\vec{x} * \vec{y} = \vec{x} + \vec{y}$ regardless of which T's, S's, and C_{ST}'s are chosen. Note: in the above, all arithmetic is mod p.

Since EQ 5 is a group isomorphism on T, it preserves diassociativity, commutativity, and indeed the validity of every 2-variable identity valid in $\mathbb{Z}_p$. Nevertheless, there is tremendous freedom here since for each ordered pair of a 2D subspace T and a 1D subspace S within it, we may choose the corresponding stretching constant C_{ST} completely arbitrarily. That freedom is enough to destroy every (≥ 3)-variable identity except for those that are really just disguised (≤ 2)-variable obvious diassociativity-commutativity statements, i.e. except for those in which we can never leave a 2D subspace. (And by considering the direct product of our loop with the group generated by two generic $N \times N$ real matrices, N sufficiently large, we may remove commutativity while leaving associativity unaffected.) Indeed, if the prime p and the dimension n are chosen large enough and the C_{ST} are independently chosen at random from $\{1, 2, 3, \ldots, p-1\}$, we claim that it is obvious that the probability any particular such identity is violated, tends to 1. Essentially, once the two sides of the "identity" get out of the same 2D subspace, they in general never can return to a common 2D subspace again. (For an explicit verification of that in the case of the associativity identity $xy \cdot z = z \cdot yz$, see the end of the proof of theorem 5.2 in [17].) □

Remarks. In the proof of theorem 7ii it also is possible to work, not over $\mathbb{Z}_p$, but instead over the real numbers $\mathbb{R}$, in which case the cardinality of our loop, and the number of freely alterable C_{ST}'s, are both continuum-infinite instead of finite, and the probability of identity-violation is 1. This whole construction may be viewed as a way of converting putative (≥ 3)-variable identities to "identities" over $\mathbb{R}$ involving a much larger number of variables once all the C_{ST}'s are included. The point is that every departure of the input identity from being a pure consequence of diassociativity is associated with its own private C_{ST}, and so by freely varying these C_{ST}'s slightly we can cause the output identity obviously to be false.

It would also be possible (and some may prefer this) to *define* the set of identities that hold in the $\mathbb{R}$-based loop construction regardless of the C_{ST} choices to *be* the "diassociativity-commutativity-equivalent identities" – and regard them alone as permissible members of a diassociativity basis.

Finally, those readers (primarily logicians?) who feel there is still something lacking in the proof of theorem 7ii are urged to consult the alternate proof presented at the end of this section.

Proof of theorem 9. Consider some 2-variate polynomial equation arising from diassociativity, for example the statement (in an additive loop) that
$$A + B + A + A + A$$
is unambiguous. What does this imply about power-associativity? In a multiplicative loop in which we (maximally optimistically) already have power-commutativity

[15] A large class of such identities are ruled out by appropriate Steiner loops, which may be constructed by using either "embedding theorems," or theorems about "independent sets," in Steiner triple systems [3] [8] [12] [27] [28]. But unfortunately Steiner loops obey $xx = e$ where e is the identity, so that such identities as $xx \cdot (yy \cdot zz) = (xx \cdot yy) \cdot zz$ are never violated by them.

$x^c x^d = x^d x^c$ and in which we already know there is power-associativity for all powers less than $4a+b$, this would imply that

$$x^{4a}x^b = x^{3a}x^{a+b} = x^{2a}x^{2a+b} = x^a x^{3a+b}$$

The point is that this is a set of equivalences among different ways of factoring x^p into two smaller powers of x, where in the present case $p = 4a+b$. Any such 2-term factoring $x^p = x^c x^d$ may be represented for short by its lefthand power alone, here c. The first 4 in our set of equivalences shows that

$$a \sim 2a \sim 3a \sim 4a$$

and considering also power-commutativity (if we assume it), we have also

$$a \ \sim \ p - a.$$

The point is that this set of equivalences is always among certain *nonzero integer linear multiples* of a modulo p. In the present case the multiples are $\{1, 2, 3, 4, -1\}$. (If rational multiples arise, we multiply them by their least common denominator to convert them to all-integer form. For example $2a/3 \sim a/2$ would become $4a \sim 3a$.)

No matter what finite set of 2-variable loop-equations we started with, we always get in this way a finite set of equivalences among integer linear multiples of a single variable, modulo p.

Now, consider p to be some suitable large prime and consider the *graph* G_p whose $p-1$ vertices are the nonzero residues mod p. Two vertices x and y of this graph are joined by an edge if there is some equivalence relation $x \sim y$, i.e. if $k_1 x = k_2 y \pmod{p}$ for some 2-tuple of fixed nonzero integers (k_1, k_2) arising from our identity-set as described above. Because we only have a finite set of loop-equations, our k_1's and k_2's all are *bounded*. Denote the least upper bound on all these k_j by K.

The statement that our loop is power-associative, would be compatible with the statement that each of these graphs is *connected*, i.e. that every factorization $x^p = x^c x^d$ is equivalent (via some path of equivalences) to every other, for each $p \geq 1$. However, we shall show that there are an infinite set of primes p such that G_p is a disconnected graph, which will ultimately allow violating power-associativity.

If p is prime, then G_p is the Cayley graph of the abelian multiplicative *group* of integers modulo p using as generators multiplications by the k_1/k_2 and k_2/k_1.

As is well known (it was first proven by Galois), the full multiplicative group of integers (and it *is* a group) mod p is cyclic with $p-1$ elements. Our Cayley graph may have more than one connected component. For example if p is odd, but all our k_j happened to correspond to an *even* number of "steps along the $(p-1)$-cycle," i.e. happened all to be *squares* modulo p, then we would have two connected components, one of the "even numbered" and the other of the "odd numbered" vertices along the $(p-1)$-cycle (for a numbering scheme where vertex $x = g^\ell$ is numbered by its discrete logarithm ℓ rather than by x; here g is a generator, or "primitive root" modulo p). That is, the squares and non-squares mod p would be in different connected components of G_p.

We *claim* that, for any fixed finite set of rational numbers mod p, there always exists an infinite set of primes p such that all of these rational numbers taken together only suffice to generate a strict subgroup of the full multiplicative group mod p; in fact such that all the numerators and denominators of all the rational numbers all are *squares* mod p. The claim seems "intuitively obvious" because the "probability" a number is a square mod p is $1/2$ (unless it is an obvious square such as 9, in which case the probability is 1 and our argument only becomes more true) so that "therefore," the probability that all of our numbers are squares exceeds 2^{-K}, which is a fixed constant. So if primes and quadratic residues behave "randomly enough" we would expect a constant fraction $\geq 2^{-K}$ of all primes will do, proving the theorem.

This intuition may instantly be made rigorous (although perhaps not with the numbers $1/2$ and 2^{-K}, but instead with certain other positive constants playing their roles) by invoking the Chebotarev density theorem. In fact, all we need is a statement considerably weaker than, and easier to prove than, the full Chebotarev theorem. That is because Chebotarev was concerned with factorization of a general but fixed univariate polynomial with integer coefficients mod p. He proved that that every possible kind of factorization would occur (e.g., for a degree-4 monic polynomial irreducible over $\mathbb{Z}$, factorizations of degrees $1+1+1+1$, $1+1+2$, $1+3$, $2+2$, and 4 all would occur), each on sets of primes p of nonzero-constant density. But we here are instead considering a very *special* polynomial of form $P(x) = \prod_i (x^2 - r_i)$. For us the situation is actually far simpler since the quadratic residue behaviors of primes p with respect to any fixed set S of integer putative residues are periodic, that is, depend only on p mod N for some large but fixed modulus N determined by S. (This may be seen using Gauss' quadratic reciprocity laws.) One may now merely employ the theorem (often ascribed to Dirichlet) that there are an infinite number of primes in every arithmetic progression $P = ak + b$ with the constants a, b obeying $\gcd(a, b) = 1$. In other words, the full power of Chebotarev was not needed.

Now our theorem will follow from our claim, by considering the infinite commutative *free loop with one generator* obeying the identities in our basis and the ones they imply, but no others. Let us be specific about how we construct this loop. Denote the elements of the loop by the integers (of both sign). Let the (two-sided) inverse of any element whose integer label is x be the one whose integer label is $-x$. Call the identity element 0. Let the loop operation be called $+$. Define $2 = 1+1$, $3 = 1+2$, $4 = 1+3$,... Now we *demand* that $a + b \neq p$ (where the integer labels a and b *do* sum to p) for some suitable a such that a and 1 are not in the same connected component of G_p, and where p is some suitable sufficiently large prime. The purpose of this demand is to prevent power-associativity. Now, to construct the +table of the loop, we simply proceed by means of an "infinite backtrack search." That is: we order the entries of the +table in some manner (such as spiralling outward from the center at $0+0$) and we fill them in with numbers in that order. Each time we fill in a +table entry $x+y$ (assuming we have already filled in some finite number of entries) that, via our identity-set, yields a finite number of consequences about other entries in the +table, thus either enabling them to be instantly filled in also, or yielding a contradiction. In the latter case we "backtrack" – go back and erase our choice of $x+y$, replacing it with the next available choice ("next" according to this ordering of the integers:

$0, 1, -1, 2, -2, 3, -3, \ldots$) and try again, up until the choice CNM where C is some sufficiently large constant, N is the number of entries filled in so far, and M is the maximum absolute value among them[16]. If all choices fail, that means we have to backtrack one further back into the past.

The backtracking choices conceptually form a tree of possibilities for the loop's +table (each tree node is a table entry to be filled in, and its child subtrees arise from the different choices at that node) and the backtrack procedure is systematically examining root-down paths through that tree. Our goal is to prove the backtracking cannot be stopped.

For each $N > 0$, we know that there some root-down path in this tree N choices long. That is because the only way no such path could exist, i.e. the only way the backtracking could be forced to halt, would be if there were some contradiction – which would have to arise from a finite chain of linear equivalences connecting $a + b = p$ to $1 + (p - 1) = p$, i.e. showing $a \sim \cdots \sim 1$. Why is this the only possible way a contradiction could arise? Since we know that diassociativity in loops is logically consistent, the only possible way a contradiction could arise would have to be from the only other axiom we have supplied the backtracker, namely, that $x^a x^b \neq x^1 x^{p-1}$ for some x for some fixed pre-chosen a, b with $a + b = p$. But: since we know that a and 1 are in *different* connected components of the equivalence graph G_p, no such chain, and hence no such contradiction, is possible. The backtracking cannot be halted[17].

Now by Denes König's "infinity lemma" (discussed in section 2.3.4.3 of [21] we know that any tree with a root-down path of length N for every positive integer N, must be an infinite tree, containing an infinite root-down path. Therefore, there is a way to fill in the entire infinite +table, and furthermore for any given entry of that table, this backtracking procedure will ultimately spit out the right value for that entry and never backtrack to revise it again, forever after. Perhaps the procedure will do this extremely slowly, e.g. superexponential time using the CNM bound, but if that bound is not valid then still in finite time although conceivably the amount of time cannot be bounded by any computable function – the slowness does not matter for the purpose of proving existence. □

Remark. Our proof technique involving an "infinite backtrack tree" turns out to have a distinguished provenance. It was first used by Kurt Gödel in his 1929 PhD thesis [16] to prove the "Gödel completeness theorem" in logic. (Of course, Gödel used considerably different language, since the concept of a "backtracking algorithm" had not yet been invented, but the idea was the same. See, e.g., pages 124, 128, and 141 in [13]. Also proven in [34][10].) This states, essentially, that any sound set of axioms in first order logic (i.e. which do not provably lead to a contradiction) is instantiated by some model, possibly of countably-infinite-size.

Alternate proof of theorem 7ii: It is important to note [10] that Gödel's completeness theorem for first order logic remains valid whether the sets of axioms, variable names, constant names, predicates, and functions used to define the first order language are finite or *countably infinite.*

Regard diassociativity as being defined by a countably infinite set of axioms, namely, all the obvious diassociativity statements. These axioms plus the loop axioms lead, via finite chains of logical deductions, to certain consequences C. Plainly no such consequence can be a minimal identity with (≥ 3) variables. Now it is an immediate consequence of the Gödel completeness theorem, that *there is no other statement (i.e not in C) valid in every (at most countably infinite) diassociative loop.* Any attempt to produce such a statement would be confronted with a countermodel produced by the usual proofs of Gödel's theorem. □

3.2 Extension to finite loops

The previous proofs, especially of theorem 9, seemed to depend on the infinite size of the constructed loops, and left the hope that diassociativity in *finite* loops might have a finite equational basis. That hope is destroyed by:

Theorem 11 (Main result$_\mathrm{F}$). *There is no finite equational basis for diassociativity in finite loops, even in commutative ones.*

Proof: The proof will actually be surprisingly easy, at least for the reader familiar with the ideas of the previous proofs.

We employ theorems 7 and 9 as before, but these both have to be modified to make all the loops become finite. In the case of theorem 7 this modification is easy: it suffices to replace the two generic real matrices with random integer matrices over GF_p for sufficiently large prime p, and the additive group generated by 1 and $\sqrt{2}$ by the additive group $\mathbb{Z}_p$ of integers modulo some sufficently large prime p. (To see why this must work with overwhelming probability as $p \to \infty$, employ the Schwartz-Zippel lemma.)

Theorem 9 is the difficult case. For any given finite set S of obvious diassociativity statements, we are going to construct an infinite family of commutative but *not* power-associative finite loops which satisfy all the identities in S. Let $D \geq 2$ be the maximum degree of (i.e. number of terms on each side of) any diassociativity statement in S and let I be the number of statements in S.

Let P and Q be[18] sufficiently large primes. Choose P such that $\{\pm 1, \pm 2, \pm 3, \ldots, \pm D^{I+D}\}$ do not suffice to generate the multiplicative group of integers mod P, e.g. because all these numbers are squares mod P. As before, the Chebotarev density theorem assures us that an infinite (indeed, constant density) set of such primes P exists, for any fixed I and D. Our loop will have $N = PQ$ elements.

[16] Actually, conceivably this upper bound CNM is not large enough, but our backtracking procedure below can be altered to make it double C each time a backtrack is needed, in which case that objection, even if valid, is made moot.

[17] Incidentally, it is is important to note that because our loop obeys the inverse property, left-division and right-division are simply multiplying on the left or right by x^{-1}, which by construction is the loop element with integer label $-x$. Thus we automatically know the x/y and $x\backslash y$ tables. It is considerably trickier to construct infinite loops not obeying the inverse property. There is one such construction in the companion paper [44].

[18] Making P and Q be primes may not be necessary, but it definitely makes the proof easier. Choosing P prime causes the multiplicative structure mod P to be easy to understand. (Multiplicative structure modulo primes is a cyclic group, i.e. is maximally simple.) We also want to use the Chebotarev density theorem – which is about primes – to show an infinite set of possible P exist. Choosing Q to be a large prime prevents any small-length sum of a's and b's (not both multiples of Q) from being a multiple of Q and thus getting into the subgroup, except for possibly in a *unique* way $k_1 a + k_2 b$, where k_1 and k_2 are small integers dependent on a and b and unique up to taking common small-integer multiples.

Let the loop operation be $*$ and let 0 be the identity. Let the $*$table of the loop (telling us what $a*b$ is) agree with $+$ in the additive group of integers mod N *except* that if both

1. Both a and b are nonzero multiples of Q.
2. The sum is *not* of the form $a+b$ with the integers a, b related by $k_1 a = k_2 b$ mod N where k_1 and k_2 are integers, not both 0, and both with absolute values $\leq D^{I+D}$.

then disagreement is permitted. Our loop will have a P-element subgroup consisting of the multiples of Q. This all is achieveable by making $*$ inside that subgroup agree with the additive group of integers mod P (ignoring the overall multiplication by Q) but with the nonzero elements *permuted* away from sorted order, i.e. $a*b = \pi^{-1}(\pi(a)+\pi(b))$ if a, b are both in the subgroup. We now demand that the permutation π must respect condition 2, i.e. $a*b = a+b$ if loop-elements a and b happen to be related by $k_1 a = k_2 b$ mod N where k_1 and k_2 are integers, not both 0, and both with absolute values $\leq D^{I+D}$.

Such a permutation π exists, as in the proof of theorem 6 due to the way P was chosen. (The point is that the P-vertex graph of integers mod P with x joined to y by an edge if $k_1 x = k_2 y$ for k_1 and k_2 integers, not both 0, and both with absolute values $\leq D^{I+D}$, is *not connected.*)

Why will this loop not be power-associative? Because the permutation was nontrivial so that *some* $a+b$ with a,b both multiples of Q will not agree with integer addition mod N.

Let our two variables be called a and b. We shall from here on assume $Q > D^{I+D}$.

It is then not possible to sum b with itself more than D^{I+D} times by using our identities (even logically linked with the use of the word *where* as in footnote 14; just one identity acting alone cannot even sum b with itself more than D times) hence no achievable sum of b's can be in the subgroup if b is outside it. If a is in the subgroup but b is not then similarly no achievable sum of a's and b's can enter the subgroup. However, if both a and b lie outside the subgroup, then some combined sum such as $c = a+a+b+b+b$ might be in the subgroup. Due to Q being a large prime, this c will necessarily be unique (up to small-rational multiples) among achievable sums.

Why will this loop obey all the diassociativity statements in S? If both variables are in the subgroup, this is trivial since groups are power associative (even after isomorphisms π have been applied to them). If a is in the subgroup but b is not then all $+$'s will operate on at least one summand not in the subgroup, or will be a sum of small-rational numbers of pure a's, hence either way will agree with integer addition mod N, and hence will satisfy S.

If both of the variables avoid the subgroup (but possibly some sum of them is in it) then all the $*$'s on both sides of the identity will agree with addition in the additive group of integers mod N (and hence will trivially satisfy S) if either

1. One or more summands is not in the subgroup;
2. Both summands are in the subgroup – because then the two summands necessarily will satisfy some linear relation with small integer coefficients k_1, k_2.

□

Remarks. If linking the statements in S via the use of the word "where" is disallowed, then it is impossible to use a long string of 'where's to make an exponentially long sum (e.g. $d+d$ *where* $d = c+c$ *where* $c = b+b$ *where* $b = a+a$). In that case the uses of the expression D^{I+D} in the proof could be replaced simply by D.

This sort of construction sometimes also will work for non-prime P, Q, although we do not claim any full understanding of which. Again the essential core of the argument may be abstracted in the form of a lemma:

Lemma 12 (Abstractified essence). *Let $(\mathbb{Z}_N, +)$ be the additive (cyclic) group of integers modulo some composite number N. Let $(H, +)$ be a nontrivial subgroup (necessarily cyclic with $1 < |H| < N$). Let π be a non-identity permutation of the elements of H. Define $(H, \circ)$ to be the new group isomorphic to $(H, +)$ via π, that is $x \circ y = \pi\{\pi^{-1}(x) + \pi^{-1}(y)\}$. Choose π so that $x+y = x \circ y$ for all $x, y \in H$ such that $k_1 x = k_2 y$ mod N for any nonzero k_1, k_2 with $|k_1|, |k_2| < K$. Further assume $x, y \notin H$ implies that at most one number of the form $k_1 x + k_2 y$ mod N (up to integer multiples, where again k_1,k_2 are integers with $|k_1|, |k_2| < K$) is in H or equivalently that the index of H in $\mathbb{Z}_N$ is relatively prime to all numbers below K. Define a new operation $*$ on $\mathbb{Z}_N$ as follows:*

$$x * y \overset{\text{def}}{=} \begin{cases} x \circ y & \text{if } x, y \in H \\ x + y & \text{otherwise.} \end{cases} \tag{6}$$

Then $(\mathbb{Z}_N, *)$ is a commutative loop which is not power-associative but obeys every obvious diassociativity statement with $< K$ occurences of each variable on each side of the equation, and thus every diassociativity statement of degree$\leq K$.

Example. A 75-element loop arises by letting G be $\mathbb{Z}_{75}$, letting H consist of the 15 elements within G which happen to be multiples of 5, and letting $\pi(x)$ map every $x \in H$ to itself except for interchanging 25 and 50. It therefore is commutative and obeys all degree$\leq$ 5 diassociativity statements such as $a(bb \cdot b) = ab \cdot bb$ and $(aa \cdot ba)b = a(ab \cdot ab)$.

The computer tells us that this loop is the unique IP-alternative 75-element loop whose loop operation $a*b$ agrees with integer addition mod 75 for all a, b which are not both nonzero multiples of 5, but which disobeys $x^5 x^{25} = x^1 x^{29}$ when $x = 1$.

3.3 Extension to finite power-associative loops?

The previous results might suggest that somehow the essence of why diassociativity has no finite basis, is merely Evans and Neumann's fact that power-associativity has no finite basis. So one might conjecture that diassociativity *does* have a finite basis in *power-associative* loops. But the 18-element loop in figure 1.2 makes me conjecture the opposite!

4 Open questions

1. Is there a finite equational basis for diassociativity in power-associative loops?

2. Can all identities that hold in all Moufang loops be understood, and can an efficient simplification-to-canonical-form procedure be constructed which will reduce any polynomial, via Moufang-loop identities, thus enabling quick verification or refutation of any identity in generic Moufang loops?

5 Acknowledgements

W.McCune (Argonne Nat'l Lab.) wrote the programs `mace4` [32] and `otter` [33] which, along with my own program `loopbeaut 1.0`, played an important role in this research. `Mace4` is often thousands of times faster than McCune's previous, and more widely circulated, model-finding program `mace2`.

J.D.Phillips (Wabash College, Indiana) first found, by using `mace4`, an 18-element IP-alternative loop and a 21-element alternative and commutative loop, neither of which are diassociative. He showed me these loops and how to use `mace4`, thus stimulating this investigation.

Petr Vojtechovsky (Denver University, Colorado) spotted a serious error in an earlier draft, and made various critiques and rewriting suggestions.

For those reasons, and since a good fraction of this paper was developed while carrying on email conversations with Phillips and/or Vojtechovsky, I proposed that their names be added as secondary authors, but with me being the primary author. They felt, however, that they should either be full authors or not authors at all, with the result that I am now listed as the sole author. (Indeed, one of them did not even want me to acknowledge him, but I have disregarded that request.)

After this paper was completed, M.K.Kinyon pointed out Clark's 2-page long 1970 paper *Diassociative groupoids are not finitely based* [7]. Clark's proof contains the same kind of hole that we pointed out in Evans & Neumann's power-associativity proof: Clark just takes it for granted that only (what we have called) "obvious diassociativity statements" can form a diassociativity-basis. Assuming that, he then argues that, if n is made large enough, then no finite collection of diassociative identities can transform an expression of form $(ab^n a)b$ (where we have intentionally left the inner parenthesization unspecified) to any other form. "Q.E.D."

There is also a second gap in Clark's argument: he does not explicitly construct any counterexample groupoid – he merely assumes that, given his preceding argument about $(ab^n a)b$, one "obviously" must exist. (The present paper's result, of course, supercedes Clark's.)

References

[1] A.A.Albert: Power-associative rings, Trans. Amer. Math. Soc. 64 (1948) 552-593.

[2] A.S.Amitsur & J.Levitski: Minimal identities for algebras, Trans. Amer. Math. Soc. 105 (1962) 202-221.

[3] L.D.Andersen, A.J.W.Hilton, E.Mendelsohn: Embedding partial Steiner triple systems, Proc. London Math'l. Soc. 41 (1980) 557-576.

[4] N.C.Ankeny: The least quadratic non-residue, Annals Math. 55 (1952) 65-72.

[5] Tom M. Apostol: Introduction to analytic number theory, Springer-Verlag (5th ed.) 1995.

[6] Richard Hubert Bruck: A survey of binary systems, Springer-Verlag 1958, third corrected printing 1971 (Ergebnisse der Math. #20).

[7] David M. Clark: Diassociative groupoids are not finitely based, J.Australian Math'l. Soc. 11 (1970) 113-114.

[8] C.J. Colbourn & Alexander Rosa: Triple systems, Clarendon Press, Oxford 1999.

[9] J.Denes & A.D.Keedwell (eds.); Latin squares, North Holland 1991 (Annals Discr. Math.#46).

[10] Vilnis Detlovs & Karlis Podnieks: Introduction to Mathematical Logic, electronically available textbook at University of Latvia, `http://www.ltn.lv/∼podnieks/mlog/ml.htm`. Chapter 4 covers completeness theorems.

[11] H.G. Diamond: On elementary methods in the study of the distribution of prime numbers, Bull. Amer. Math. Soc. 7 (1982) 553-589.

[12] Jean Doyen & R.M.Wilson: Embedding of Steiner triple systems, Discr. Math. 5 (1973) 227-239.

[13] Herbert B. Enderton: A mathematical introduction to logic, Academic Press 1972.

[14] Trevor Evans: Embedding incomplete Latin squares, Amer. Math'l. Monthly 67 (1990) 958-961.

[15] Trevor Evans & B.H.Neumann On varieties of groupoids and loops, J.London Math. Soc. 28 (1953) 342-350.

[16] Kurt Gödel: Die Vollständigkeit der Axiome des logischen Funktionen-kalküls, Monatsh. für Mathematik und Physik 37 (1930) 349-360.

[17] Joan Hart & Kenneth Kunen: Single axioms for odd exponent groups, J.Automated Reasoning 14 (1995) 383-412.

[18] D.R.Heath-Brown: Artin's conjecture for primitive roots, Quart. J. Math. (Oxford, ser. 2) 37,145 (1986) 27-38.

[19] D.G. Hoffman: Completing incomplete commutative latin squares with prescribed diagonals, Europ. J. Combinatorics 4,1 (1983) 33-35. [Same theorem also proved by L.D.Anderson: Annals Discrete Math. 15 (1982) 33-35, according to the review of Hoffman's paper in mathematical reviews #84h:05024.]

[20] D.R. Hughes & F.C. Piper: Projective Planes, Springer-Verlag, New York, 1973.

[21] D.E.Knuth: Fundamental algorithms, vol. 1 of Art of Computer Programming, Addison-Wesley Reading Mass 1972.

[22] M.K.Kinyon, K.Kunen, J.D.Phillips: A Generalization of Moufang and Steiner Loops, Algebra Universalis 48,1 (2002) 81-101.

[23] J.C.Lagarias & A.M.Odlyzko: Effective versions of the Chebotarev density theorem, pp. 409-464 in Algebraic Number Fields (ed. Albrecht Frohlich), Academic Press, 1977.

[24] "The Chebotarev density theorem," e-available notes by Hendrik W. Lenstra from an Oberwohlfach lecture. `http://math.berkeley.edu/∼jvoight/notes/oberwolfach/Lenstra-Chebotarev.pdf`

[25] N. Levinson: A motivated account of an elementary proof of the prime number theorem, Amer. Math. Monthly 76 (1969) 225-245.

[26] S.Li & C.Pomerance: On generalizing Artin's conjecture to composite moduli, J.Reine Angew. Math. 556 (2003) 205-224.

[27] Charles C. Lindner: A partial Steiner triple system of order n can be embedded in a Steiner triple system of order $6n + 3$, J.Combin.Theory A 18 (1975) 349-351.

[28] C.C. Lindner & C.A. Rodger: General embedding therorems for partial latin squares, Bull.Inst.Combinat.Applic. 5 (1992) 81-99

[29] Wen Chao Lu: On the elementary proof of the prime number thorem with a remainder term, Rocky Mtn. Math. J. 29,3 (1999) 979-1053.

[30] K.R.Matthews: A generalization of Artin's conjecture for primitive roots, Acta Arith. 29,2 (1976) 113-146.

[31] W.McCune: Son of Bird Brain (automated deduction system demonstration web page), `http://www-unix.mcs.anl.gov/AR/sobb/`.

[32] William W. McCune: Mace 2.0 reference manual and guide, `cs.SC/0106042`

[33] William W. McCune: Otter 3.3 reference manual, `cs.SC/0310056`

[34] Elliot Mendelson: Introduction to mathematical logic, Lewis Publishers Inc. 4th ed. 1997.

[35] W. Narkiewicz: Elementary and analytic theory of algebraic numbers, second ed. Springer 1990.

[36] R.Odoni: Some global norm density results obtained from an extended Cebotarev density theorem, pp. 485-495 in Algebraic Number Fields, A.Frohlich (ed.), Academic Press, 1977.

[37] A.Paszkiewicz & A.Schinzel: On the least primitive root modulo a prime, Math. of Comput. 71, 239 (2002) 1307-1321.

[38] D.A.Robinson: Bol loops, Trans.Amer.Math.Society 123 (1966) 341-354.

[39] C.A.Rodger: Embedding an incomplete latin square in a latin square with a prescribed diagonal, Discrete Math. 51,1 (1984) 73-89.

[40] S.Rosset: A new proof of the Amitsur-Levitski identity, Israel J. Math. 23,2 (1976) 187-188.

[41] M.Rubinstein & P.Sarnak: Chebyshev's bias, J.Experimental Math. 3,3 (1994) 173-197.

[42] J.T.Schwartz: Fast probabilistic algorithms for verification of polynomial identities, J. ACM 27,4 (1980) 701-717.

[43] B.Smetaniuk: A new construction of latin squares I. A proof of the Evans conjecture, Ars Combinatorica 11 (1981) 155-172.

[44] Warren D. Smith: Inclusions among diassociativity-related loop properties, available at `http://math.temple.edu/~wds/homepage/works.html`.

[45] Warren D. Smith: Quaternions, octonions, and now, 16-ons and 2^n-ons; New kinds of numbers. Book, under review for publication.

[46] P.J.Stephens: An average result for Artin's conjecture, Mathematika 16 (1969) 178-188.

[47] S.S.Wagstaff Jr.: Pseudoprimes and a generalization of Artin's conjecture, Acta Arith. 41,2 (1982) 141-150.

[48] Clifton T. Whyburn: The density of power residues and nonresidues in subintervals of $[1, \sqrt{p}]$, Acta. Arith. 14 (1967/8) 113-116.

[49] K.A. Zhevlakov, A.M. Slin'ko, I.P. Shestakov, A.I. Shirshov: Rings that are nearly associative, Academic Press 1982.

[50] R.Zippel: Probabilistic algorithms for sparse polynomials, EUROSAM '79, Springer (Lecture Notes in CS#72) 216-226.