WHY STUDY SEMIGROUPS?
John M. Howie
Lecture given to the New Zealand Mathematical Colloquium
(Received June 1986)

1. Introduction

Before tackling the question in my title I should perhaps begin by saying what a semigroup is. A non-empty set $S$ endowed with a single binary operation $\cdot$ is called a semigroup if, for all $x, y, z$ in $S$,

$$(xy)z = x(yz).$$

If in addition there exists $1$ in $S$ such that, for all $x$ in $S$,

$$1x = x1 = x$$

we say that $S$ is a semigroup with identity or (more usually) a monoid.

I shall be confining myself today to semigroups that have no additional structure. Thus, though semigroups feature quite prominently in parts of functional analysis, the algebraic structure of those semigroups is usually very straightforward and so they scarcely rate a mention in any algebraic theory. Equally, although they are often of greater algebraic interest, I shall say nothing about topological semigroups.

Let me begin by answering a slightly different question: Who studies semigroups? Section 20 in Mathematical Reviews is entitled 'Groups and generalizations' and has two 'leper colonies' at the end, called $20M$ Semigroups and $20N$ Other generalizations. In the Introduction to their book The algebraic theory of semigroups in 1961 Clifford and Preston [2] remarked that about thirty papers on semigroups per year were currently appearing. A brief look at recent annual index volumes of Mathematical Reviews shows that the current figures are

1982 321
1983 310
1984 311

Incidentally, comparable figures for 'other generalizations' are about one third of these. So it is clear that of all generalizations of the group concept the semigroup is the one that has attracted the most interest by far. I shall in due course hazard a guess as to why this is so.

Mathematicians are rightly a bit suspicious of theories whose only motive seems to be to generalize existing theories - and if the only motivation for semigroup theory were to examine group-theoretical results with a view to generalization, then I would have no very convincing answer to the question of my title. The test proposed by Michael Atiyah for a generalization - that it should have at least two distinct and interesting special cases - is a reasonable one provided it is not applied too dogmatically; and the semigroup concept passes the test, since from the outset semigroup theory drew its ideas partly from group theory and partly from ring theory.

For clearly every ring \((R,+,\cdot)\) is a semigroup if we simply neglect the operation \(+\). The converse is certainly not true: that is, there are semigroups \((S,\cdot)\) with zero on which it is not possible to define an operation \(+\) so as to create a ring \((S,+,\cdot)\). The easiest way to see this is to recall the known result that a ring \((R,+,\cdot)\) with the property that \(x^2 = x\) for all \(x\) in \(R\) (a Boolean ring) is necessarily commutative - i.e. satisfies \(xy = yx\) for all \(x\) in \(R\). Now let \(S = (A \times B) \cup \{0\}\), where \(A\), \(B\) are non-empty sets and \(0 \notin A \times B\), and make \(S\) into a semigroup with zero by defining

\[
(a,b)0 = 0(a,b) = 00 = 0,
\]

\[
(a_1,b_1)(a_2,b_2) = (a_1,b_2).
\]

Then \(S\) has the property that \(x^2 = x\) for all \(x\) in \(S\). On the other hand \(S\) is certainly not commutative and so cannot be made into a ring.

2. Pure mathematical reasons

Let me now turn to my question 'Why study semigroups?' I am a pure mathematician by instinct and so I begin by offering some pure mathematical reasons. I shall come later to what might be called 'applied mathematical' reasons.
My first point, not a fashionable one in these utilitarian times, is that they are fun, that they provide an elegant theory, with arguments that any mathematician can actively enjoy. Let me give an example. The concept of regularity, introduced for rings by no less a person than von Neumann, plays a much more central role in semigroup theory than it does in ring theory. We say that $(S,.)$ is regular if

$$ (\forall a \in S)(\exists x \in S) \quad axa = a. \quad (1) $$

As in ring theory, idempotent elements - elements $e$ such that $e^2 = e$ - are very important. From equation (1) we readily see that both $ax$ and $xa$ are idempotent.

Next, note that a semigroup $(S,.)$ is regular if and only if

$$ (\forall a \in S)(\exists a' \in S) \quad a'a = a, \quad a'aa' = a'. \quad (2) $$

It is clear that $(2) \implies (1)$. To show that $(1) \implies (2)$ simply take $a' = xax$; then from (1) we have

$$ a'a = a, \quad a'aa' = a'. $$

The element $a'$ is usually called an inverse of $a$, but it should be noted that this is a weaker concept of inverse than the one used in group theory: for example in the four element semigroup with Cayley table

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>a</td>
<td>b</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>b</td>
<td>a</td>
<td>b</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>c</td>
<td>c</td>
<td>d</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>d</td>
<td>c</td>
<td>d</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

it is easy to check that every element is an inverse of every other element.

Theorem 1. The following conditions on a regular semigroup $S$ are equivalent:

(1) Idempotents commute;
(2) Inverses are unique.

Proof. (1) ⇒ (2). Suppose that idempotents commute. Let \(a', a^*\) be inverses of \(a\). Then

\[
a' = a'aa' = a'a^*a^*a' = a'a^*a^*a' = a^*aa^*aa^* = a^*aa^* = a^*
\]

by (2).

(2) ⇒ (1). Let \(e, f\) be idempotents and let \(x\) be the unique inverse of \(ef\):

\[
ef xef = ef, \quad xefx = x.
\]

Then \(fxe\) is idempotent, since

\[
(fxe)^2 = f(xefx)e = fxe;
\]

and \(ef\) is an inverse of \(fxe\):

\[
(fxe)(ef)(fxe) = f(xefx)e = fxe,
\]

\[
(ef)(fxe)(ef) = efxef = ef.
\]

But an idempotent \(i\) is its own unique inverse \((iii = i, \ iii = i)\) and so \(ef = fxe\), an idempotent. Similarly \(fe\) is idempotent.

The unique inverse of \(ef\) is thus \(ef\) itself. On the other hand \(fe\) is an inverse of \(ef\), since

\[
(ef)(fe)(ef) = (ef)^2 = ef, \quad (fe)(ef)(fe) = (fe)^2 = fe.
\]

It follows that \(ef = fe\), as required.

That argument goes back to the early 1950s, to some fundamental work by Vagner [16, 17] and Preston [12, 13, 14]. A regular semigroup satisfying either one (and hence both) of the conditions in Theorem 1 is
called an *inverse semigroup,. A very impressive theory has been created for such semigroups, as is evidenced by the publication in 1984 of a 674-page monograph by Petrich [11] devoted entirely to semigroups of this kind.

But do semigroups of this kind occur 'in nature'? Let me give a not very well known example due to Schein [15] and McAlister [10]. Let $G$ be a group and let $K(G)$ be the set of all right cosets of $G$. This includes $G$ itself and also the cosets of the subgroup 1, which are effectively the elements of $G$. Define an operation $*$ on $K(G)$ by

$$Ha * Kb = (H \cap aK^{-1})ab.$$ 

This is a natural definition: it is not hard to check that $Ha * Kb$ is the smallest coset containing the product $HaKb$. [Certainly

$$HaKb = (HaK^{-1})ab \subseteq (H \cap aK^{-1})ab.$$ 

Conversely, suppose that $HaKb \subseteq P_a (\subseteq K(G))$. Then in particular $ab \in P_a$ and so $P_a = Pab$. Now

$$Hab \subseteq HaKb \subseteq Pab \text{ and so } H \subseteq P;$$

also

$$(aK^{-1})ab = aKb \subseteq HaKb \subseteq Pab$$

and so $aK^{-1} \subseteq P$. Thus $H \cap aK^{-1} \subseteq P$ and so

$$(H \cap aK^{-1})ab \subseteq Pab = P_a.$$ 

It is a routine matter to check that $*$ is an associative operation and that $(a^{-1}Ha)a^{-1}$ is an inverse of $Ha$ in the semigroup $(K(G),*)$. Now suppose that $Ha$ is idempotent:

$$Ha = Ha * Ha = (H \cap aH^{-1})a^2.$$ 

Then in particular $a^2 = 1a^2 \cdot Ha$; i.e. $a^2 = ha$ for some $h$ in $H$. 5
Hence $a = h \in H$ and so $Ha = H$. In fact the idempotents of $(K(G), *)$ are precisely the subgroups of $G$. For any two subgroups $H, K$ of $G$

$$H \ast K = H \vee K = K \ast H.$$  

Thus idempotents commute and so $(K(G), *)$ is an inverse semigroup.

The normal subgroups $N$ of $G$ are such that, for all $Ha$ in $K(G)$,

$$N \ast Ha = (N \vee H)a = (NH)a,$$

$$Ha \ast N = (H \vee aNa^{-1})a = (H \vee N)a = (NH)a,$$

and so are central idempotents in $(K(G), *)$. Conversely, if $N$ is a central idempotent then for all $a$ in $G$

$$Na = N \ast 1a = 1a \ast N = (1 \vee aNa^{-1})a = aN$$

and so $N$ is normal.

That is a slightly quaint example, but I mention it because I have the feeling that it has not yet been adequately exploited. The main reason that semigroups turn up in mathematics is that one is very often interested in self-maps of a set of one kind or another, and whenever $f, g, h$ are such maps it is automatically the case that

$$(f \circ g) \circ h = f \circ (g \circ h).$$

If the maps are bijections then the appropriate abstract idea is that of a group; if not then inevitably we must consider a semigroup. It is this connection with maps (arising from the associative axiom) that is the strongest reason why semigroups are more important both theoretically and in applications than the various non-associative generalizations of groups.

There is another pure mathematical reason for being interested in semigroups. It is possible to take a very general standpoint in algebra and to discuss a so-called $\Omega$-algebra $A$ having a family $\Omega = \{\omega_i : i \in I\}$ of operations, where $\omega_i : A^{n_i} \rightarrow A$ is an $n_i$-ary operation. [For example,
in a group one can take $I = \{1, 2\}$, $n_1 = 2$, $n_2 = 1$, with

$$[\omega_1(a_1, a_2) = a_1a_2, \quad \omega_2(a_1) = a_1^{-1}].$$

If $\phi : A \rightarrow B$ is a map between $\Omega$-algebras then we say that $\phi$ is a morphism if (for all $i$ in $I$ and for all $a_1, \ldots, a_{n_i}$ in $A$)

$$\phi(\omega_i(a_1, \ldots, a_{n_i})) = \omega_i(\phi(a_1), \ldots, \phi(a_{n_i})).$$

If we regard $\phi$ as applying to $A^{n_i}$ in an obvious way then we can express this property succinctly as a commuting condition

$$\phi \circ \omega_i = \omega_i \circ \phi. \quad (3)$$

A congruence on $A$ is an equivalence relation $\sim$ with the property that (for all $i$)

$$a_1 \sim a_1', \ldots, a_{n_i} \sim a_{n_i}' \Rightarrow \omega_i(a_1, \ldots, a_{n_i}) \sim \omega_i(a_1', \ldots, a_{n_i}'). \quad (4)$$

Consider the quotient set $A/\sim$, whose elements are equivalence classes

$$[a] = \{x \in A : x \sim a\}.$$

The congruence property means that $A/\sim$ inherits the $\Omega$-algebra structure from $A$ : we simply define

$$\omega_i([a_1], \ldots, [a_{n_i}]) = [\omega_i(a_1, \ldots, a_{n_i})] \quad (5)$$

and the compatibility condition (4) ensures that the definition makes sense. There is a natural map $\eta : A \rightarrow A/\sim$ defined by

$$\eta(a) = [a] \quad (a \in A)$$

and the definition (5) can be interpreted as saying that

$$\omega_i \circ \eta = \eta \circ \omega_i.$$
hence, comparing with (3), we see that \( \Phi \) is a morphism.

Now suppose that \( \phi \) is a morphism from \( A \) onto \( B \); we say that \( B \) is a morphic image of \( A \). Define \( \sim \) on \( A \) by the rule that

\[
a \sim a' \quad \text{if and only if} \quad \phi(a) = \phi(a').
\]

It is easy to verify that \( \sim \) is a congruence. The first isomorphism theorem for \( \Omega \)-algebras is then as follows:

**Theorem 2.** Let \( A, B \) be \( \Omega \)-algebras, let \( \phi : A \to B \) be a morphism with \( \text{im } \phi = B \), and let \( \sim \) be the congruence on \( A \) defined by (6).

Then there is an isomorphism \( \alpha : A/\sim \to B \) such that the diagram

\[
\begin{array}{ccc}
A & \overset{\phi}{\longrightarrow} & B \\
\downarrow{\sim} & & \downarrow{\alpha} \\
A/\sim & & \end{array}
\]

commutes.

This is, in one form or another, one the cornerstones of abstract algebra. It says in effect that an \( \Omega \)-algebra \( A \) carries its morphic images 'within itself' and that to reveal them we need only consider the quotients of \( A \) by its various congruences.

The result applies to groups, of course, but it is not usually stated in quite this way. This is because for a group \( A \) there is a one-one correspondence between congruences \( \sim \) and normal subgroups \( N \) given by

\[
N = \{ a \in A : a \sim 1 \}, \quad \text{or}
\]

\[
a \sim b \quad \text{if and only if} \quad ab^{-1} \in N.
\]

The quotient \( A/\sim \) is always denoted by \( A/N \). Similarly, for a ring \( A \) there is a one-one correspondence between congruences \( \sim \) and two-sided ideals \( I \) given by
\[ I = \{ a \in A : a \sim 0 \}, \text{ or} \]
\[ a \sim b \text{ if and only if } a - b \in I, \]

and the quotient \( A/\sim \) is always written \( A/I \).

In semigroups no such device is available and we must study congruences as such. So semigroups constitute the simplest, most manageable and most natural class of algebras to which the methods of universal algebra must be applied.

3. Applied Mathematical reasons

Let me now turn to less exalted reasons for studying semigroups. One of the striking aspects of semigroup conferences these days is that many of the participants, between a third and a half, at a guess, come from departments of computer science. The reason is that semigroups have found significant applications in the theory of automata, languages and codes.

If \( A \) is a non-empty set (an alphabet, as we often want to call it) then the set of all finite words

\[ w = a_1 a_2 \ldots a_n \]

in the alphabet \( A \) is a semigroup if we define multiplication by juxtaposition. Denote the length of \( w \) by \( |w| \). If we include the empty word \( 1 \) (with \( |1| = 0 \)) then we obtain a monoid, which we denote by \( A^* \). This is the free monoid generated by \( A \). The set of non-empty words in \( A^* \) is usually denoted by \( A^+ \). A subset of \( A^* \) is called a language.

Now let \( Q \) be a finite non-empty set and suppose that we have a map \( f : Q \times A \rightarrow Q \); we normally write \( f(q,a) \) simply as \( qa \) and think of \( A \) as 'acting' on \( Q \). The function \( f \) can be extended to \( Q \times A^* \) by defining (inductively)

\[ q1 = q(q \in Q) \]
\[ q(\omega a) = (q \omega)a \quad (q \in Q, \ \omega \in A^*, \ a \in A). \]

We say that \( A = (Q,f) \) is an \( A^*-automation \). (In the terminology of
Eilenberg [4] this is a complete deterministic automaton.) We may think of it as a very rudimentary machine whose states (the elements of \( Q \)) can be altered by various input (the elements of \( A \)).

Suppose now that among the elements of \( Q \) there is an element \( i \) which we call the initial state and that there is a subset \( T \) of \( Q \) called the set of terminal states. Let

\[
L = \{ w \in A^* ; \ iw \in T \} .
\]

Then we say that \( L \) is the language recognized by the automaton \( A \). A language \( L \) is called recognizable if there exists an automaton \( A \) recognizing \( L \).

Example. We can picture an automaton via its state graph. If \( A = \{ a, b \} \), \( Q = \{ 0,1,2,3 \} \), \( 0a = 0b = 0 \), \( 1a = 2a = 2 \), \( 3a = 0 \), \( 1b = 1 \), \( 2b = 3 \), \( 3b = 1 \), then we can draw the picture

Let \( i = 1 \), \( T = \{ 1,2,3 \} \). It is easy to see that 0 is a 'sink' state from which no escape is possible, and that

\[
1aba = 2aba = 0 .
\]

In fact \( L \), the language recognized by this automaton, consists of all words in \( A^* \) not containing \( aba \) as a segment.

If \( L_1 \), \( L_2 \subseteq A^* \) then \( L_1 . L_2 \) is defined as \( \{ w_1 w_2 : w_1 \in L_1 , w_2 \in L_2 \} \). If \( L \subseteq A^* \) then \( < L > \) denotes the submonoid of \( A^* \) generated by \( L \). Let \( F \) be the set of all finite subsets of \( A^* \). Then the set
Rat $A^*$ of rational subsets of $A^*$ is the set of subsets of $A^*$ obtained from $F$ by means of the operations of $U$ (finite union), $\cdot$ and $\prec$. This leads to an important characterization of recognizable subsets of $A^*$:

Theorem 3. (Kleene [7]). A language $L$ is recognizable if and only if it is rational.

Another characterization of recognizable languages is more algebraic in character. If $L$ is any subset of $A^*$ then the syntactic congruence $\sim$ of $L$ is the relation $\sim$ on $A^*$ defined by

$$w_1 \sim w_2 \iff [w_1 v \in L \iff w_2 v \in L]. \quad (7)$$

(I have not gone into the mathematical theory of grammar at all, but one can perhaps dimly see that this is saying that $w_1$ and $w_2$ are mutually interchangeable in the set of 'meaningful' sentences that constitutes $L$. Thus (very roughly speaking) if $u = \text{(the)}$, $v = \text{(sat on the mat)}$ then cat $\sim$ dog but cat $\not\sim$ black. So 'cat' and 'dog' are syntactically equivalent but 'cat' and 'black' are not.) It is easy to verify that $\sim$ is a congruence on $A^*$. Then $A^*/\sim$ is called the syntactic monoid of $L$ and is usually denoted by $M(L)$.

Theorem 4. $L$ is recognizable if and only if its syntactic monoid $M(L)$ is finite.

This is by no means as deep as the Kleene theorem. Indeed one way round it is virtually obvious. If $M(L)$ is finite we can define an action of $A$ on $M(L)$ by

$$f(\bar{w},a) = \bar{w}a = \bar{wa} \quad (\bar{w} \in M(L) = A^*/\sim)$$

making $A = (M(L),f)$ into an $A^*$-automation. If we then take $\bar{1}$ as initial state and $\bar{L} = \{\bar{w} : w \in L\}$ as the set of terminal states then $A$ recognizes $z \in A^*$ if and only if $\bar{1}z \in \bar{L}$, i.e. if and only if $\bar{z} \in \bar{L}$, i.e. if and only if there exists $w$ in $A^*$ such that $w \in L$ and $w \sim z$. But then $1w1 \in L$ and hence $z = 1z1 \in L$ by (7). We conclude that the automation $A = (M(L),f)$ recognizes $L$. 


I mention this proof (or rather half-proof) because it emphasised the very close links between automata and monoids. A further connection is provided by the theory of *codes*.

It is not the case that every submonoid of a free monoid is free. For example, in \( \{a,b\}\) consider the submonoid \( M = M^* \cup \{1\} \), where \( M^* = \{a^i : i \geq 2\} \). The base of \( M \) (i.e. the set \( M^* \setminus (M^*)^2 \) of indecomposable elements of \( M \)) is \( \{a^2,a^3\} \), but \( M \) is certainly not freely generated by \( \{a^2,a^3\} \). We say that a subset \( C \) of \( A^* \) is a *code* if it is the base of a free submonoid of \( M \). For example,

\[
C_1 = \{a^2,aba,ab^2,b\}, \quad C_2 = \{a^4,b,ba^2,ab,aba^2\}
\]

are codes. This means, for example, that any word in four letters - say \( x_3x_1x_4x_2x_3 \) - can be encoded unambiguously (in \( C_1 \), say) as

\[
ab^2a^2baba^2b^2.
\]

\( C_1 \) is in fact an example of a *prefix* code: any word in \( <C_1> = C_1^* \) can be decoded without hesitation by reading from the left. This is because \( C_1 \cap C_1A^+ = \emptyset \). By contrast if one tries to decode

\[
aba^5ba^2
\]

using \( C_2 \) the answer (unique since \( C_2 \) is a code) is \( x_5x_1x_3 \), but one might first have tried \( x_4x_1 \) or \( x_5x_1x_2 \) before reaching the correct solution.

That the theory of codes is intimately bound up with the theory of monoids is illustrated by.

**Theorem 5.** Let \( M \) be a submonoid of a free monoid \( A^* \) and let \( C \) be its base. Then \( C \) is a prefix code if and only if \( M \) satisfies

\[
u, \, u \in M \Rightarrow x \in M.
\]

(\( M \) is called left unitary.)
This has been a very sketchy introduction to a rich and rapidly growing area. Those whose appetites have been whetted should turn to Eilenberg [4,5], Lallement [8], Berstel and Perrin [1], and Lothaire [9].

REFERENCES


University of St. Andrews, SCOTLAND.