WHEN THE ELECTRON FALLS APART

In condensed matter physics, some particles behave like fragments of an electron.

Philip W. Anderson

It is ironic that, in the year when we celebrate the centenary of the discovery of the electron, the most exciting developments in the theory of electrons in solids have to do with the "fractionalization" of the electron—the discovery of particles that behave as though the electron had broken apart into three or five or more pieces each containing one-third or one-fifth of its charge, or into separate particles, one containing its charge and one its spin. (See figure 1.) No longer is the quantum theory of solids confined to the boring old electron; we now have a remarkable variety of fractional parts of electrons: composite fermions, composite bosons, spinons and holons, in addition to the heavy electrons, quasiparticles and small polarons of older stages of condensed matter theory.

These developments of course have nothing to do with the internal structure of the electron as an elementary particle; we are told that the electron is one of the few particles that is so far seen as truly elementary in that sense. But they do mean that in describing the low-energy excitations of solids, condensed matter theorists have had the opportunity to play with many of the interesting new concepts that modern field theory has developed to deal with strings, non-Abelian gauge fields and the like, as well as with some interesting new ideas of our own. As has often been the case in the past, the mathematical structure of modern condensed matter theory is exhibiting parallel development with elementary particle theory.

In fact, right from the start of the quantum theory of solids in the 1920s, it was found useful to invent new particles to describe the collective behavior of the large number of electrons that occupy the energy bands of solids. The most useful and ubiquitous of these is the "hole," invented by Heisenberg already in 1926 to describe nearly full atomic shells, which exhibit a reversed sign of spin-orbit coupling. Rudolph Peierls's arguably greatest contribution to physics was to extend this concept to the energy bands in solids, and to realize that the missing particles in a nearly full band would behave dynamically like positively charged holes. His application of this was to the "anomalous" Hall effect in metals. The Hall effect is the transverse voltage that develops from sideways acceleration of a moving electron by a magnetic field. The force on the moving particle is

$$\mathbf{F} = e \mathbf{v} \times \mathbf{B},$$

and clearly which way the voltage deviates from the current is a measure of the sign of e. It had long been puzzling that in metals on the right side of the periodic table, e often came out positive. This idea of the hole was soon taken up by semiconductor physicists, with enormous consequences for technology, as we all know. (And, of course, it was taken up by Dirac, in his subsequent relativistic theory of the electron, in which he originally postulated the positron as a hole in a filled band of negative-energy solutions of his equation.)

The hole is not fractionalized, unless one views -1 as a fraction, but it has in common with the soliton-like excitations discussed below that collective behavior of the whole band is responsible. It is the Pauli exclusion principle (together with basic principles of symmetry) that forces a filled shell or band to be electronically inert and to have zero net velocity, so that the shell or band with one missing particle responds in many ways like the negative of a real particle.

BCS quasiparticles

The concept of the hole plays an important role in the next example of a bizarre particle, the quasiparticle of the Bardeen–Cooper–Schrieffer (BCS) theory of superconductivity. One may think of the ground state of the BCS theory as a coherent linear combination of states with different numbers of electron pairs, in which the members of the pairs have equal but opposite spin and momentum, and hence zero total spin and momentum, but charge 2e. Thus, when one adds an electron of momentum k and spin \(\frac{1}{2}\), one arrives at a state equivalent to one that could be reached by subtracting an electron of momentum \(-k\) and spin \(\frac{1}{2}\) from such a pair, leaving behind its partner. Thus, in this case, particle states—the exact eigenstates of the system—are linear combinations of electrons and holes, with opposite charge but the same momentum and spin.

Again, it is the background of the entire Fermi sea of electrons that absorbs the missing quantum number, so there is nothing really there but physical electrons with
charge \(-e\) and spin \(\frac{1}{2}\). Nonetheless, quasiparticles can do odd things: One may demonstrate, as Walter J. Tomaseh did, quantum interference fringes between electrons and holes from the same region of the Fermi surface (William MacMillan and I supplied the theory). At an interface between a normal metal and a superconductor, electrons in the normal metal can be reflected by the interface as holes, leaving behind the net charge \(2e\), which is carried off by the superconducting condensate. This Andreev scattering is the main current-carrying process at such an interface, and hence is responsible for our ability to insert current into a superconductor.\(^4\)

**One-dimensional Hubbard model**

The superconducting case, bizarre as it is, still gives us one new quasiparticle state corresponding to one single old free electron state, so while the electron’s properties are profoundly modified, we have not yet managed to take it apart into separate pieces. As far as I can tell, this was first envisaged by Elliott Lieb and Fred Wu\(^5\) in a spectacular theoretical paper published in 1968 under the misleading title “Absence of Mott Transition in [the One-Dimensional Hubbard Model].” The Mott transition is a theory of the metal–insulator transition for certain types of materials. John Hubbard had proposed his model of a local repulsive interaction in a tight-binding band as an appropriate system for modeling magnetism and metal–insulator transitions (it was very similar to a model I had put forward in 1959 to describe the Mott insulating antiferromagnet), and the subject of Mott metal–insulator transitions had (and has) continued to be rife with confusion and controversy. Lieb and Wu’s *tour de force* was an exact formal solution of the one-dimensional version of this rather realistic model of interacting, spin-degenerate electrons, for all parameter sets and sizes (including the \(N \to \infty\) thermodynamic limit, which is relevant for describing real solids).

Whether they made the point given in their title is moot—what they showed was that the case with exactly one electron per site is always an insulator, and all other occupancies are metals, and whether that is a “Mott transition” or not is a question that is semantic, not substantive—but the real kicker came when they described the excitation spectrum. They wrote that “we find three types of excitations,” which can be interpreted as two types of spinless charge excitations (holes and particles) and one type of neutral spin excitations. Please note that none of these are at all like electrons! Lieb and Wu referred back to similar phenomena in a 1963 paper by Lieb and Werner Liniger, and in a 1967 paper by C. N. Yang, but none of those are as conclusive or deal with physical models, though both papers indeed pioneered the mathematical techniques used.

It is an understatement to suggest that the significance of this discovery was not appreciated for many (perhaps twenty) years. For the first time, it was shown that in a particular interacting many-electron system, the single added electron or quasiparticle could be simply not a stable component of the system: that it could and would decay into two or a number of constituent parts, and that the system’s true, exact excitation spectrum is not electron-like at all.

There was a second feature, again foreshadowed in many papers going back even to early work by Sin-itiro Tomonaga and Joaquin Luttinger in the early 1950s: In spite of breaking apart, the electron still has a Fermi surface (in this one-dimensional case, two “Fermi points”), which are the only points in electron momentum space where excitation energies vanish—and these Fermi points remain at the *same* locations in momentum space. This last property is known as Luttinger’s theorem and it is surprising that it occurs for this bizarre system: The proof of Luttinger’s theorem uses perturbation theory methods that should not be valid in this case.

Let us not suppose the Hubbard model is an isolated instance of charge–spin separation. In 1974, Alan Luther and his collaborators\(^6\) (following ideas of Dieter and Ursula Schotte) developed a much simpler and more generally applicable, but less rigorous, technique called bosonization, which allows many variants of the original model to be treated, and the above phenomenon of “charge–spin separation” seems to be almost ubiquitous for physical one-dimensional systems. Bosonization is essentially the description of a fermionic entity as a function of bosonic
operators—the requisite anticommutators typically arise from an interplay between normal ordering and the bosonic commutators. Groups in the USSR also developed techniques relying on resummation of perturbation theory, by use of which Igor Dzialoshinsky and Anatoly Larkin in 1974 even gave a formal description of exactly how the electron breaks up—that is, they gave a formal Green's function, or propagator, for the electron, valid in the asymptotic long-time limit.

The remarkable thing to me has not been the burgeoning of beautiful theoretical work on these one-dimensional systems, which continues to this day, but the failure of most of the theoretical work to connect with experiments on real one-dimensional or quasi one-dimensional systems. As far as I know, only in the past year has a single instance of experimental results from an interplay between normal ordering and the bosonic commutators resulted from just such a system with attractive interactions between electrons mediated by phonons: the case of polyacetylene, culminating in the classic paper of Wu-Pei Su, Robert Schrieffer and Alan Heeger in 1979.

The polymer chain based on acetylene takes two forms. The interesting one is called trans-polyacetylene and, in the ground state, is an insulator in spite of having just one odd \( \pi \) electron in its valence band per molecule, because the electrons couple strongly to the molecular displacements. They form stronger bonds—that is, they displace the atoms toward each other—between successive pairs of atoms along the chain, giving us what is called a charge-density wave (CDW), which doubles the periodicity over which the chain repeats. This may be drawn schematically as shown in figure 2. The pair bonds in the figure each imply two electrons but in fact the electrons are by no means totally localized in the regions between the closer pairs of carbons. (See the article on CDWs by Robert E. Thorne in PHYSICS TODAY, May 1996, page 42.) By doubling the periodicity, the tight-binding energy band opens up an energy gap at the momentum vector \( \pi/a \) corresponding to the double periodicity. (See figure 3.)

Now, we ask, what happens if, starting from one end of the chain, we pair atoms 0 and 1, 2 and 3 and so on, and from the other end we pair atoms \( 2n + 1 \) and \( 2n \) and so on, so that they don’t match in the middle? The odd atom—say number 4—may keep its electron and be neutral, but have a spin; if an electron is added or subtracted, however, the atom is spinless but charged! (See figure 4.)

Schrieffer and his coworkers showed that these two types of solitons existed and were locally stable, and that each was quite extended and mobile, so that doping polyacetylene could lead to a sample with charged mobile carriers but no mobile spins, for instance. These carriers

**Trans-polyacetylene**

Lieb and Wu and their successors dealt primarily with the case of repulsive, Coulomb interactions among electrons, because of their interest in the Mott transition. In the Hubbard model the repulsive interactions are represented by a repulsive potential \( U \). Attractive interactions between electrons are invariably the consequence of coupling to lattice vibrations, and the frequencies of these vibrations, or phonons, introduce a second, lower-frequency dynamical scale. The “attractive \( U \)” Hubbard model is an oxymoron, and one must always include the dynamics of the phonons or possibly other attractive mechanisms to arrive at realistic results. The second—and first

**FIGURE 2.** **TRANS-POLYACETYLENE** has alternating double and single bonds (green) between its carbon atoms (black). Although the electrons are not as localized as this diagram implies, the C=C bonds are shorter than the C-C bonds and the electron densities form a Peierls charge-density wave along the chain.

**FIGURE 3.** **ENERGY BANDS IN POLYACETYLENE** develop gaps because of the charge-density wave. **a.** In the absence of a CDW, there is a single tight-binding energy band, corresponding to the periodicity \( a \) (the distance between successive carbon atoms). **b.** The CDW doubles the period of the chain to \( 2a \) and an energy gap opens up at the momentum \( \pi/a \) that corresponds to this period.
are quantum particles, in a real sense. Although this proposal initially met with considerable resistance, the experimental existence of free solitons with quantum numbers not corresponding to those of an electron—that is, fractionalized electrons—has been checked out experimentally for this system. As Schrieffer also pointed out, if the CDW does not have double periodicity but triple or more, the charged version of the soliton—a so-called discommensuration—could carry fractional charge, but this has not been confirmed in any experimental one-dimensional case that I know of.

Fractional quantum Hall effect

In 1982 the most exciting and spectacular example of electron fractionalization appeared in physics through an entirely experimental breakthrough: the discovery of the fractional quantum Hall effect by Daniel C. Tsui, Horst Stormer and Arthur C. Gossard at Bell Laboratories.10 A few years earlier, Klaus von Klitzing, Gerhard Dorda and Michael Pepper had discovered the (integer) quantum Hall effect through measuring the Hall effect in high magnetic fields on the two-dimensional electron gas in a high-mobility surface inversion layer on silicon. Sharp plateaus of the Hall voltage as a function of field occur at quantized values of the Hall constant,

$$\phi_0 = \frac{hc}{e}.$$  

That is, for every unit of flux there are $v$ electrons; in the case of $v = \frac{1}{2}$, for example, there are three flux quanta per electron.

Laughlin is again responsible for the correct explanation of these fractional values.12 He pointed out that his explanation for the integer quantum Hall effect could be recast by describing a filled quantized Landau level as an incompressible electron fluid because once a Landau level is filled with exactly one electron per flux quantum, the next electron must go into the next higher energy level, which is $\hbar \omega_c = \hbar B/mc$ higher in energy. He then ingeniously showed that he could also make a plausible many-body wavefunction that is an incompressible electron quantum liquid at $\frac{1}{3}, \frac{1}{5}, \ldots, \frac{1}{2n+1}$ of a filled Landau level.

To get an idea how he could do this, notice that the two-dimensional coordinates, $x_j$ and $y_j$, of an electron can be combined into the single complex coordinate

$$Z_j = x_j + iy_j.$$  

A wavefunction in the lowest Landau level can be described (aside from a fixed single-particle factor) as any polynomial function of $Z$, that is not a function containing any factors of $Z^n$, the complex conjugate of $Z$. In any many-body wavefunction that satisfies the exclusion principle, no two particles can be at the same point, so the function must necessarily have a factor $(Z_j - Z_l)$ for any pair of particles $j$ and $l$, and a filled Landau level actually has one factor $(Z_j - Z_l)$ for every pair.

Now, one could satisfy the exclusion principle also with factors $(Z_j - Z_l)^n$ (the exponent must be odd for fermions). If one does just that for every pair, it turns out that the magnetic field must be three times as big for the same number of particles. But such a wavefunction will be very much favored if the particles repel each other at short range, as of course they do. With ingenious and extensive numerical calculations, Laughlin showed the stability of the fractional states—that is, the states with $(Z_j - Z_l)^n$ factors for odd $n$ and the usual quantized values of the flux.

These states behave as though they are made up of fractional electrons. But can we construct a fractional electron state? Laughlin also solved this problem in his initial paper by using the following ingenious construction.

He began with a wavefunction where all the electron pairs have $(Z_j - Z_l)^3$ factors. If an infinitesimal solenoid
with one quantum of magnetic flux were to be added at some point $x_j + 1/3 = Z_0$, every electron would have to precess around it so the wavefunction has to have a new factor $(Z_j - Z_0)$ for each $j$. The electrons are all pushed out a little bit by the new flux, and it turns out that the hole created by their outward motion contains just one-third of an electron, so one has a soliton with charge $+1/3$. By removing a flux quantum (or introducing a negative one) one may also add one-third of an electron. For a long time, it was not clear whether one had established a physical meaning for these fractional charges, but extensive and patient effort by experimenters such as Tsui and his coworkers has left us quite sure they exist.

In addition to these fractional states, a zoo of more or less complicated nonclassical excitations of these extraordinarily perfect two-dimensional electron gases has been postulated and in most cases demonstrated in the course of the intervening years. An ingenious scheme pioneered by Steven Girvin and Allan MacDonald is to associate a fictitious flux of the sort described above with a particle. For each quantum of fictitious flux, one changes the sign of the statistics, so that one or three quanta plus an electron act like a boson, and the plateaus (at which $\rho_{xx} = 0$) act like Bose condensates of the fictitious particle, when one compensates the external field with an odd number of fictitious quanta. But at $v = 1/3$, the compensation occurs for two quanta, and the particle is a fermion again. (See PHYSICS TODAY, June 1994, page 21.) Present theoretical ideas (due to Nicholas Read, Bertrand Halperin, Patrick Lee and others) propose that these fermions have an actual Fermi surface, which may have been observed.

Another type of fractionalized particle occurs in the so-called Hall edge states, states that travel around the boundary of a quantum Hall fluid. They behave like a special kind of one-dimensional system that, in the fractional $v$ case, has fractional solitons as its basic excitations. (See PHYSICS TODAY, June 1994, page 21, and September 1996, page 19.)

**High-$T_c$ cuprates**

The glorious successes of the theory for the fractional quantum Hall effect contrast with the bitter controversy and lack of consensus associated with the high-$T_c$ cuprates. Nonetheless, a substantial number of theorists and experimenters in this field agree that it is likely that in the cuprate normal metal (above $T_c$) the transport properties cannot be understood with a conventional quasiparticle theory. The most striking of many unusual observations is the observation of two transport relaxation times $\tau$ in yttrium barium copper oxide (as well as other systems in which good single crystal measurements are available). The evidence is shown in figure 6.

It is a well-known feature of optimally doped cuprates that the resistivity $\rho$ in the plane of the cuprate (which in simple kinetic theory is $\rho = m/n e^2 \tau$, where $n$ is the total number of electrons of effective mass $m$, and $\tau$ is the mean free time) is roughly proportional to the absolute temperature, with mysteriously little or no intercept or "residual resistance." The coefficient $m$ can be determined from infrared measurements, so we actually find that

$$\frac{h}{\tau} = kT$$

(to within 50% or so, where $\tau$ is the relaxation time as determined by resistivity. This rate, $1/\tau$, is a very large relaxation rate. (From infrared measurements, one finds that, at high frequencies, $h/\tau = \omega_c$, also.)

There is a second simple way of determining $\tau$, from the Hall effect. The angle through which an electron's momentum precesses in a relaxation time $\tau_H$ is the Hall angle between current and voltage in a magnetic field, and is

$$\theta_H = \omega_c \tau_H = \frac{eB}{mc} \tau_H$$

where $\omega_c$ is the cyclotron frequency, the frequency of the precession. But when we measure $\theta_H$ in a wide variety of samples, we find, quite reliably,

$$\theta_H^{-1} \propto \tau^{-1} \propto T^2.$$  

So there are two separate relaxation times in the cuprate planes, one of about $h/kT$ for accelerations parallel to the electronic momentum (and hence perpendicular to the Fermi surface), and one, somewhat longer and propor-

**FIGURE 5. QUANTUM HALL EFFECT** (upper curve) and magnetoresistance (lower curve) in a GaAs-Ga(Al)As heterojunction. Plateaus occur in the Hall resistivity $\rho_{xx}$ at values $h/ve^2$, where $v$ is an integer or an odd-denominator fraction. Zeros or smaller dips in the longitudinal resistivity $\rho_{xx}$ coincide with these plateaus. The scale along the top of the figure indicates the filling factor (v) at the corresponding magnetoresistance (lower curve) in a GaAs-Ga(Al)As heterojunction. Plateaus occur in the Hall resistivity $\rho_{xx}$ at values $h/ve^2$, where $v$ is an integer or an odd-denominator fraction. Zeros or smaller dips in the longitudinal resistivity $\rho_{xx}$ coincide with these plateaus. The scale along the top of the figure indicates the filling factor (v) at the corresponding magnetoresistance (lower curve) in a GaAs-Ga(Al)As heterojunction. Plateaus occur in the Hall resistivity $\rho_{xx}$ at values $h/ve^2$, where $v$ is an integer or an odd-denominator fraction. Zeros or smaller dips in the longitudinal resistivity $\rho_{xx}$ coincide with these plateaus. The scale along the top of the figure indicates the filling factor (v) at the corresponding magnetoresistance (lower curve) in a GaAs-Ga(Al)As heterojunction.
As we remarked earlier, the transport theory of composite (fluctuation-dissipation theory) do not apply. These stand-electrons is unfamiliar, and standard methods (such as lard methods have served us faithfully over many years, ard methods have served us faithfully over many years, and it is very hard, particularly for young physicists who have never known anything else, to abandon the appar-ently trustworthy certainty of the textbook formulas. Misa-