§ 1. Introduction and early happenings

In this essay we wish to embark on the telling of a story which, almost certainly, stands only at its beginning. We shall discuss the links between one very old subject, *physics*, and a much newer one, *topology*. Physics, being so much older, has a considerably longer history than does topology. After all the bulk of topology didn't even exist before the beginning of the twentieth century. However, despite this disparity of antiquity between the two subjects, we shall still find it worthwhile to examine the situation before the twentieth century. We should also remind the reader that the term topology is not usually found in the literature before 1920 or so (cf. footnote 5 of this article), one finds instead the Latin terms *geometria situs* and *analysis situs*.

From graph theory to network theory

Perhaps the earliest occurrence of some topological noteworthiness was Euler's solution [**1**] of the *bridges of Königsberg problem* in 1736. This was nothing if not a physical problem though one with a rather amusing aspect. The problem was this: in the city of Königsberg[1] were seven bridges and the local populace were reputed to ask if one could start walking at any one of the seven so as to cross all of them precisely once and end at one's starting point. Euler had the idea of associating a graph with the problem (cf. fig. 1). He then observed that a positive answer to this question requires the vertices in this graph to have an even number of edges and so the answer is *no*[2].

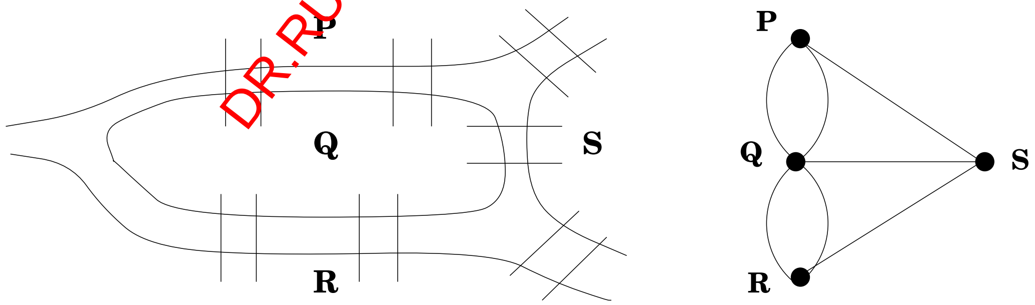

Fig. 1. **The seven bridges of Königsberg and its associated graph with seven edges**

In devising his answer Euler gave birth to what we now call graph theory and, in so doing, was considering one of the first problems in combinatorial topology; and though graph theory could once be viewed as a combinatorial study of the topology of one dimensional complexes it is now an independent subject in its own right.

Among the parts of physics with close connections to graph theory is network theory. The earliest connection occurs in the work of Kirchhoff [**1**] in 1847 who, as well as formulating his two famous laws for electric circuits, made use of a graph theoretic argument to solve the resulting equations for a general electrical network.

From these beginnings the links between graph theory and physics have strengthened over the centuries.

Electromagnetic theory and knots

In the nineteenth century we encounter a more substantial example of a physical phenomenon with a topological content. The physics concerned electromagnetic theory while the topology concerned the linking number of two curves.

In electromagnetic theory the magnetic field **B** produced by a current I passing through a wire obeys the pair of equations

$$\nabla \times \mathbf{B} = \mu_0 \, \mathbf{J}, \qquad \nabla . \mathbf{B} = 0 \tag{1.1}$$

where **J** is the current density. Now let us briefly summarise what is entailed in deriving the famous Ampère law. We integrate **B** round a closed curve C which bounds a surface S, then we have

$$\begin{aligned} \int_C \mathbf{B} \cdot \mathbf{dl} &= \int_S \nabla \times \mathbf{B} \cdot \mathbf{ds} \\ &= \mu_0 \int_S \mathbf{J} \cdot \mathbf{ds} \\ &= \mu_0 I \end{aligned} \tag{1.2}$$

where in the last step we use the fact that the integral $\int_S \mathbf{J} \cdot \mathbf{ds}$ gives the total current passing through S and this will be precisely I if (as we assume) the wire cuts the surface S just once. However the wire could cut the surface more than once—say n times. When this is the case at each cut the the integral receives a contribution of $\mp \mu_0 I$ depending on whether **J** is parallel or anti-parallel to **ds** at the cut. Thus, if $m\mu_0 I$ denotes the algebraic sum of all these n contributions then the most general statement for Ampère's law in this case is

$$\int_C \mathbf{B} \cdot \mathbf{dl} = \mu_0 m I, \quad m \in \mathbf{Z} \tag{1.3}$$

This integer m is a familiar feature of textbook calculations of the magnetic field due to a solenoid.

Now let us take an alternative route to calculating $\int_C \mathbf{B} \cdot \mathbf{dl}$. If we introduce the vector potential **A** where

$$\mathbf{B} = \nabla \times \mathbf{A} \tag{1.4}$$

and impose the Coulomb gauge condition

$$\nabla . \mathbf{A} = \mathbf{0} \tag{1.5}$$

then, to find $\mathbf{B}$, we only have to solve, the equation

$$\nabla^{\mathbf{2}} \mathbf{A} = -\mu_0 \mathbf{J} \tag{1.6}$$

and this has the solution

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int_{\mathbf{R}^3} \mathbf{d^3 r'} \frac{\mathbf{J}(\mathbf{r'})}{|\mathbf{r} - \mathbf{r'}|} \tag{1.7}$$

But if the wire forms a closed curve C', say, with infinitesimal element of length $\mathbf{dl'}$ then $\mathbf{J}$ has support only on the wire and so is related to the current I by

$$\mathbf{J}(\mathbf{r'}) \mathbf{d^3 r'} = I \, \mathbf{dl'} \tag{1.8}$$

This allows us to express $\mathbf{B}$ as a line integral around C' giving us

$$\begin{aligned} \mathbf{B}(\mathbf{r}) &= \frac{I\mu_0}{4\pi} \nabla \times \int_{\mathbf{C'}} \frac{\mathbf{dl'}}{|\mathbf{r} - \mathbf{r'}|} \\ &= -\frac{I\mu_0}{4\pi} \int_{C'} \frac{(\mathbf{r} - \mathbf{r'}) \times \mathbf{dl'}}{|\mathbf{r} - \mathbf{r'}|^3} \end{aligned} \tag{1.9}$$

Now we introduce a second curve C and integrate $\mathbf{B}$ round C thereby obtaining

$$\int_C \mathbf{B} \cdot \mathbf{dl} = -\frac{I\mu_0}{4\pi} \int_C \int_{C'} \frac{(\mathbf{r} - \mathbf{r'}) \times \mathbf{dl'} \cdot \mathbf{dl}}{|\mathbf{r} - \mathbf{r'}|^3} \tag{1.10}$$

Perusal of 1.3 and 1.10 together shows us immediately that

$$-\frac{1}{4\pi} \int_C \int_{C'} \frac{(\mathbf{r} - \mathbf{r'}) \times \mathbf{dl'} \cdot \mathbf{dl}}{|\mathbf{r} - \mathbf{r'}|^3} = m \tag{1.11}$$

For comparison with the work of Gauss below we wish to have this formula in a completely explicit form with all its coordinate dependence manifest and so we write

$$\mathbf{r} = (x, y, z), \quad \text{and} \quad \mathbf{r'} = (x', y', z') \tag{1.12}$$

giving

$$-\frac{1}{4\pi} \int \int \frac{(x'-x)(dydz' - dzdy') + (y'-y)(dzdx' - dxdz') + (z-z')(dxdy' - dydx')}{[(x'-x)^2 + (y'-y)^2 + (z'-z)^2]^{\frac{3}{2}}} = m \tag{1.13}$$

That this is a topological statement is clear and the integer m is actually the *linking number* of the two curves C and C' and so electromagnetic theory has provided with an explicit formula for the linking number and so an early result in knot theory.

This result 1.13 was known to Gauss in *exactly* the form that we have presented it above. Gauss's work on this matter also came from work (in 1833) on electromagnetic theory and is to be found in his *Nachlass* (Estate) cf. Gauss [**1**] where on p. 605 one finds the formula (which I quote entirely unchanged from the printed version available in Gauss [**1**] although one should remember that the original is handwritten rather than printed)

ZUR ELECTRODYNAMIK. 605

[4.]

Von der *Geometria Situs*, die LEIBNITZ ahnte und in die nur einem Paar Geometern (EULER und VANDERMONDE) einen schwachen Blick zu thun vergönnt war, wissen und haben wir nach anderthalbhundert Jahren noch nicht viel mehr wie nichts.

Eine Hauptaufgabe aus dem *Grenzgebeit* der *Geometria Situs* und der *Geometria Magnitudinis* wird die sein, die Umschlingungen zweier geschlossener oder unendlicher Linien zu zählen.

Es seien die Coordinaten eines unbestimmten Punkts der ersten Linie x, y, z; der zweiten x', y', z' und

$$\int\int \frac{(x'-x)(dydz'-dzdy')+(y'-y)(dzdx'-dxdz'+(z-z')(dxdy'-dydx')}{[(x'-x)^2+(y'-y)^2+(z'-z)^2]^{\frac{3}{2}}} = V$$

dann ist dies Integral durch beide Linien ausgedehnt

$$= 4m\pi$$

und m die Anzahl der Umschlingungen.

Der Werth ist gegenseitig, d. i. er bleibt derselbe, wenn beide Linien gegen einander umgetauscht werden. 1833. Jan. 22.

This translates[3] to

ON ELECTRODYNAMICS. 605

[4.]

Concerning the *geometria situs*, foreseen by LEIBNITZ, and of which only a couple of geometers (EULER and VANDERMONDE) were allowed to catch a glimpse, we know and have obtained after a hundred and fifty years little more than nothing.

A main task (that lies) on the border between *geometria situs* and *geometria magnitudinis* is to count the windings of two closed or infinite lines.

The coordinates of an arbitrary point on the first line shall be x, y, z; (and) on the second x', y', z', and (let)

$$\int\int \frac{(x'-x)(dydz'-dzdy')+(y'-y)(dzdx'-dxdz'+(z-z')(dxdy'-dydx')}{[(x'-x)^2+(y'-y)^2+(z'-z)^2]^{\frac{3}{2}}} = V$$

then this integral carried out over both lines equals

$$= 4m\pi$$

[3] I am greatly indebted to Martin Mathieu for providing me with this translation.

and m is the number of windings.

This value is shared, i.e., it remains the same if the lines are interchanged.

1833. Jan. 22.

We see on comparing Gauss's formula with 1.13 that they are exactly the same modulo the fact that his integer m is minus our m. Gauss's remarks[4] above also show that he understood the topological nature of his result which he quotes without reference to any electromagnetic quantities; in addition he bemoans the paucity of progress in topology in a manner which makes clear that he thinks that there is much to be discovered eventually.

Maxwell was also aware of Gauss's result and mentions it in Maxwell [**1**] when he discusses the conditions for the single valuedness of a function defined by a line integral. It is clear that he realises the need for a topological restriction on the domain of definition of the function. On p. 17 of Maxwell [**1**] he says

> *There are cases, however in which the conditions*
>
> $$\frac{dZ}{dy} - \frac{dY}{dz} = 0, \quad \frac{dX}{dz} - \frac{dZ}{dx} = 0, \quad \frac{dY}{dx} - \frac{dX}{dy} = 0,$$
>
> *which are those of* $Xdz + Ydy + Zdz$ *being a complete differential, are satisfied throughout a certain region of space, and yet the line-integral from* A *to* P *may be different for two lines, each of which lies wholly within that region. This may be the case if the region is in the form of the ring, and if the two lines from* A *to* P *pass through opposite segments of the ring .... We are here led to considerations belonging to the Geometry of Position, a subject which, though its importance was pointed out by Leibnitz and illustrated by Gauss, has been little studied. The most complete treatment of this subject has been given by J. B. Listing.*[5]

Then on p. 43 of Maxwell [**2**] Maxwell refers to Gauss's linking number formula and says

> *It was the discovery by Gauss of this very integral, expressing the work done on a magnetic pole while describing a closed curve in presence of a closed electric current, and indicating the geometrical connexion between the two closed curves, that led him to lament the small progress made in the Geometry of Position since the time of Leibnitz, Euler and Vandermonde. We have now some progress to report, chiefly due to Riemann, Helmholtz, and Listing.*

Maxwell also includes a figure for which the linking number of two oppositely oriented curves is zero.

Knots, vortices and atomic theory

The nineteenth century was to see another discussion of knots and physics before it came to an end. This was in the theory of vortex atoms proposed by Lord Kelvin (alias W. H.

[4] Additional references on the history of knot theory are Epple [**1–2**] of which Epple [**2**] also discusses the work above of Gauss on knots and, in this connection, prints a fragment from one of Gauss's notebooks which show that Gauss spent some time thinking about what the current literature now calls braids. I am indebted to Ioan James for sending me a preprint of Epple [**2**].

[5] The work of Listing referred to by Maxwell is Listing [**1**]. We note that Listing was the first to use the term *Topology* (actually "Topologie", since he wrote in German) in a letter to a friend in 1836, cf. the detailed account of this on pp. 41–42 of Pont [**1**], cf. also Listing [**2**].

Thomson) in 1867; cf. Thomson [**1**] for his paper on atoms as vortices and Thomson [**2–4**] for his work on vortices themselves including references to knotted and linked vortices.

Kelvin was influenced by an earlier fundamental paper by Helmholtz (1858) [**1**] on vortices, and a long and seminal paper of Riemann (1857) [**1**] on Abelian functions[6].

His idea really was that an atom was a kind of vortex. He was sceptical about the chemists espousal of the Lucretius atom, (in Thomson [**1**]) he says:

> *Lucretius's atom does not explain any of the properties of matter without attributing them to the atom itself .... The possibility of founding a theory of elastic solids and liquids on the dynamics of closely-packed vortex atoms may be reasonably anticipated.*

and later in the same article

> *A full mathematical investigation of the mutual action between two vortex rings of any given magnitudes and velocities passing one another in any two lines, so directed that they never come nearer one another than a large multiple of the diameter of either, is a perfect mathematical problem; and the novelty of the circumstances contemplated presents difficulties of an exciting character. Its solution will become the foundation of the proposed new kinetic theory of gases.*

A significant part of the *"difficulties of an exciting character"* referred to by Kelvin above concerned the topological nature of vortices[7] i.e. that they can be knotted and that several may be linked together (cf. fig. 2).

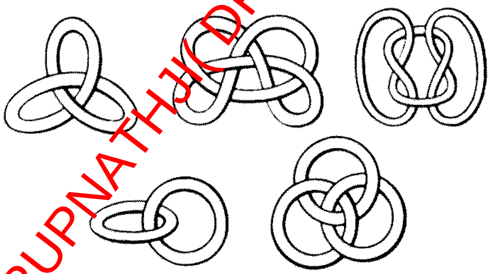

Fig. 2. **Some of the vortex tubes considered by Kelvin in Thomson [2].**

[6] Actually, in Thomson [**2**], Kelvin specifically refers to section 2 of this paper which is topological: it discusses multiple connectedness for what we nowadays refer to as Riemann surfaces. This section bears the title: *Lehrsätze aus der analysis situs für die Theorie der Integrale von zweigliedrigen vollständigen Differentialien* or *Theorems from analysis situs for the theory of integrals over total differentials of functions of two variables.*

[7] It seems clear that Kelvin thought of these vortex tubes as knotted tubes of the ether for the opening sentences of Thomson of [**2**], though they do not mention the word ether, describe an ideal fluid of that kind. One reads: *The mathematical work of the present paper has been performed to illustrate the hypothesis, that space is continuously occupied by an incompressible frictionless liquid acted on by no force, and that material phenomena of every kind depend solely on motions created in this liquid.* His belief in the ether lasted much further into the future cf. Thomson [**5**] which is dated 1900; this of course is to be expected as special relativity still lay five years ahead.

Atiyah [**1**] has summarised very well the main points that Kelvin considered to be in favour of his theory. Hence we provide a brief paraphrase of that summary here. The stability of atoms would be be accounted by the stability under deformation of the topological type of knots. The large variety of different knot types can accommodate all the different elements. Vibrational oscillations of knots could be the mechanism for atomic spectral lines.

With regard to this latter point about spectral lines Kelvin even gives (Thomson [**1**]), a rough upper bound for the rotation period of a sodium vortex based on the frequency of its celebrated yellow emission line.

Kelvin's contemporary Tait was thereby stimulated to do extensive work on knot theory and to begin work on classifying knots cf. Tait [**1**] where numerous complex knots are copiously illustrated as well as discussed. However, despite a lot of work, many of his results remained unproven and were christened the *Tait conjectures*. It is clear now that these conjectures were out of range of the mathematical techniques of his day; but many of them were finally disposed of in the 1980's by the work of Jones [**1**] which we shall come to in section 6. This work caused a resurgence in knot theory in mathematics and also coincided with a renewal of the physicists interest in knots. This is a rather special and interesting story cf. section 6.

The advent of Poincaré

As the nineteenth century drew to a close topological matters were considerably enlivened by the work of Poincaré. This work had some strong connections with physics as we shall now explain.

Poincaré's interest in physics endured throughout his all too short life but began with Newtonian dynamics and, in particular, with the three body problem.

The Swedish mathematician Mittag-Leffler, founder and chief editor of the journal *Acta Mathematica*, was the prime mover in the organisation of an international mathematical competition held to celebrate the 60th birthday of King Oscar II of Sweden and Norway. The King was well disposed towards science and allowed a competition to be announced in 1885. Four problems were proposed but entrants (who were, in theory, anonymous) were also at liberty to choose their own topic[8]. The judges were Hermite, Mittag-Leffler and Weierstrass. Poincaré was declared the winner in 1889—the King having officially approved the result on January 20th the day before his 60th birthday—there were eleven other entrants.

Poincaré had chosen to work on the first problem and in particular the three body problem[9]. His prize winning memoir[10] bore the title *Sur le problème de trois corps et les*

[8] In brief the four problems (which were contributed by Hermite and Weierstrass) were the n-body problem, an analysis of Fuchs theory of differential equations, a problem in non-linear differential equations and a final (algebraic) problem on Fuchsian functions.

[9] Poincaré actually studied the restricted three body problem. This meant that he took the first mass to be large, the second small and positive and the third negligible. He also took the first two masses to have a circular orbit about their common centre of mass while the third moved in the plane of the circles.

[10] The prize winning entry was to be published in *Acta Mathematica* and this was eventually done in Poincaré [**1**]. However the published version of Poincaré's work differs in some important respects from the entry that he submitted. This is because of an error discovered before publication by Phragmén. The mathematical and historical details of this particular story are available in the excellent book Barrow-Green [**1**].

équations de la dynamique.

Before working on this problem Poincaré had worked on the theory of differential equations and had concentrated on obtaining *qualitative* results. A key viewpoint he adopted and exploited was *geometric*—he thought of the solutions as defining geometric objects: e.g. *curves.* He then quickly obtained results of a topological nature.

For example he studied the singular points of these equations on surfaces of genus p and introduced the notions of saddle points, nodes and foci to classify these singularities—in French he used the terms *cols*, *noeuds* and *foyers* respectively—then using C, F or N to denote the type of the singularity he proved (cf. Poincaré [**5**] and [**2-6**] (1880–1886) that

$$N + F - C = 2 - 2p \tag{1.14}$$

which one recognises immediately as the index of a vector field being equated to the Euler-Poincaré characteristic of the Riemann surface.

As evidence of the stimulus that physics gave to investigations we quote from an analysis Poincaré prepared in 1901 of his own scientific work. This was published after his death (Poincaré [**7**]).

> *Pour étendre les résultats précedénts aux équations d'ordre supérieur au second, il faut renoncer à la représentation géometrique qui nous a été si commode, à moins d'employer le langage de l'hypergéometrie à n dimensions .... Ce qu'il y a de remarquable, c'est que le troisième et le quatrième cas, c'est à dire ceux qui corresponde à la stabilité, se rencontrent précisément dans les équations générales de la Dynamique .... Pour aller plus loin, il me fallait créer un instrument destiné à remplacer l'instrument géometrique qui me faisait défaut quand je voulais pénétrer dans l'espace à plus de trois dimensions. C'est la principale raison qui m'a engagé à aborder l'étude de l'Analysis situs; mes travaux à ce sujet seront exposés plus loin dans une paragraphe spécial.*

that is

> *To extend the preceding results to equations of higher than second order, it is necessary to give up the geometric representation which has been so useful to us unless one employs the language of hypergeometry of n dimensions .... What is remarkable is that in the third and fourth case, that is to say those that correspond to stability, are found precisely in the general equations of dynamics .... To go further it was necessary to create a tool designed to replace the geometric tool which let me down when I wanted to penetrate spaces of more than three dimensions. This is the principal reason which led me to take up the study of Analysis situs; my work on this subject will be expounded further down in a special paragraph.*

As yet another insight into the way Poincaré was thinking about celestial mechanics we quote the following which is taken from the introduction to his paper *Analysis situs* of 1895 (Poincaré [**8**]); this is the first of his epoch making papers on topology.

> *D'autre part, dans une série de Mémoires insérés dans le* Journal de Liouville, *et intitulés:* Sur les courbes définies par les équations différentielles, *j'ai*

> *employé l'Analysis situs ordinaire à trois dimensions à l'étude des équations différentielles. Les mêmes recherches ont été poursuivies par M. Walther Dyck. On voit aisément que l'Analysis situs géneralisée permettrait de traiter de même les équations d'ordre supérieur et, en particulier, celles de la Mécanique céleste.*

that is

> *On the other hand, in a series of memoirs in* Liouville's Journal[11] *under the title* On curves defined by differential equations, *I have used ordinary analysis situs in three dimensions in the study of differential equations. The same research has been followed by Mr. Walther Dyck. One easily sees that generalised analysis situs would permit the treatment of equations of higher order and, in particular, those of celestial mechanics.*

Poincaré's work opened a new chapter in celestial mechanics; the strong topological content of his papers on differential equations lead directly to his papers Poincaré [**8**–**18**] on *analysis situs* which gave birth to the subjects of algebraic and differential topology.

Poincaré's geometric theorem

Poincaré left unproved at the time of his death a (now) famous result usually referred to as *Poincaré's geometric theorem.* This theorem has both physical and topological content. Shortly before he died, Poincaré wrote [**18**] in order to describe the theorem and the reasons for his believing it to be true.

In the first paragraph (of Poincaré [**18**]) he says

> *Je n'ai jamais présenté au public un travail aussi inachevé; je crois donc nécessaire d'expliquer en quelques mots les raisons qui m'ont déterminé à le publier, et d'abord celles qui m'avaient engagé à l'entreprendre. J'ai démontré il y a longtemps déjà, l'existence des solutions périodiques du problème des trois corps; le résultat laissait cependant encore à désirer; car si l'existence de chaque sorte de solution était établie pour les petites valeurs des masses, on ne voyait pas ce qui devait arriver pour des valuers plus grandes, quelles étaient celles de ces solutions qui subsistaient et dans quel ordre elles disparaissaient. En réfléchissant à cette question, je me suis assuré que la réponse devait dépendre de l'exactitude ou la fausseté d'un certain théorème de géometrie dont l'énoncé est très simple, du moins dans le cas du problème restreint et des problèmes de Dynamique où il n'y a que deux degrés de liberté.*

that is

> *I have never presented to the public such an incomplete work; I believe it necessary therefore to explain in a few words the reasons which have decided me to publish it, and first of all those (reasons) which had led me to undertake it. I proved, a long time ago now, the existence of periodic solutions to the three body problem; however the result left something to be desired; for if the existence of each sort of solution were established for small values of the masses, one couldn't*

[11] Poincaré uses the term Liouville's Journal to refer to the *Journal de Mathématiques* as found, for example, in Poincaré [**3**].

> *see what might happen for larger values, (and) what were those (values) of those solutions which persisted and in what order they disappeared. On reflecting on this question, I have ascertained that the answer ought to depend on the truth or falsity of a certain geometrical theorem which is very simple to state, at least in the case of the restricted (three body) problem and in dynamical problems with only two degrees of freedom.*

The geometric theorem is a *fixed point theorem*: it states that a continuous, one to one, area preserving map f from an annulus $0 < a \leq r \leq b$ to itself has a pair of fixed points (f is also required to have the property that it maps the two boundary circles in opposite senses). When applied to the restricted three body problem it proves the existence of infinitely many periodic solutions.

Topology immediately enters because Poincaré's index theorem can easily be used to show—as Poincaré himself pointed out (Poincaré [**18**])—that there must be an even number of fixed points; hence it is sufficient to prove that f has at least one fixed point.

In 1913, Birkhoff, who was to be a prime mover and founder of the new subject of *dynamical systems*, proved the theorem cf. Birkhoff [**1**].

After Poincaré the pursuit of problems with a joint dynamical and topological content was taken up by Birkhoff, Morse, Kolmogorov, Arnold and Moser and many, many others. We shall have something to say about these matters later on in section 8.

§ 2. A quiescent period

Dirac and Schwarzschild

In the early twentieth century physics was undergoing the twin revolutions of quantum theory and special and general relativity. These revolutions imported much new mathematics into physics but topology, though itself growing at an explosive rate, did not figure prominently in the physics of this story.

Nevertheless two papers on physics of this period, do have a topological content, and are worth noting. In the first case this content is implicit and in the other it is explicit. These papers are, respectively, that of Schwarzschild in 1916 (Schwarzschild [**1**]) on solutions to Einstein's equations and that of Dirac in 1931 (Dirac [**1**]) on magnetic monopoles.

We shall discuss Schwarzschild's paper in section 3 but for the moment we want to deal with Dirac's paper because its topological content is explicit from the outset and because it has proved to be so influential.

Dirac's magnetic monopoles

Dirac begins his paper with some comments on the rôle of mathematics in physics which are both philosophical and somewhat prophetic. He says

> *The steady progress of physics requires for its theoretical formulation a mathematics that gets continually more advanced. This is only natural and to be expected. What however was not expected by the scientific workers of the last century was the particular form that the line of advancement of the mathematics would take, namely, it was expected that the mathematics would get more and*

> *more complicated, but would rest on a permanent basis of axioms and definitions, while actually the modern physical developments have required a mathematics that continually shifts its foundations and gets more abstract. Non-euclidean geometry and non-commutative algebra, which were at one time considered to be purely fictions of the mind and pastimes of logical thinkers, have now been found to be very necessary for the description of general facts of the physical world. It seems likely that this process of increasing abstraction will continue in the future and that advance in physics is to be associated with a continual modification and generalisation of the axioms at the base of the mathematics rather than with a logical development of any one mathematical scheme on a fixed foundation.*

The subject matter of Dirac's paper was both topological and electromagnetic: he found that there were magnetic monopole solutions to Maxwell's equations but that the magnetic charge μ_0 of the monopole had to quantised; in addition there were two striking facts about this quantisation one mathematical and one physical. The mathematical novelty was that the quantisation was not due to the discreteness of the spectrum of an operator in Hilbert space but rather to topological considerations. The physical novelty was that the existence of even one of these monopoles would imply the quantisation of electric charge, something not hitherto achieved[12].

Dirac considered carefully the phase of a wave function $\psi(x, y, z, t)$ of a particle in quantum mechanics. If A and γ are the amplitude and phase respectively then we have

$$\psi = Ae^{i\gamma} \tag{2.1}$$

Once ψ is normalised to unity, in the usual way, there remains a freedom to add a constant to the phase γ without altering the physics of the particle. Dirac wanted to exploit this fact and argue that the absolute value of γ has no physical meaning and only phase *differences* matter physically. In Dirac [**1**] he wrote

> *Thus the value of γ at a particular point has no physical meaning and only the difference between the values of γ at two different points is of any importance.*

He immediately introduces a generalisation, saying

> *This immediately suggests a generalisation of the formalism. We may assume that γ has no definite value at a particular point, but only a definite difference in value for any two points. We may go further and assume that this difference is not definite unless the two points are neighbouring. For two distant points there will then be a definite phase difference only relative to some curve joining them and different curves will in general give different phase differences. The total change in phase when one goes round a closed curve need not vanish.*

Dirac now does two more things: he finds that this change in phase round a closed curve will give rise to an ambiguity unless it takes the same value for all wave functions, and he goes on to equate this phase change to the flux of an electromagnetic field. For a wave

[12] Electric charge, not being the eigenvalue of any basic operator, is not quantised by the mechanism that quantises energy and angular momentum.

function in three *spatial* dimensions, to which Dirac specialises, this flux is just that of a *magnetic field*. Yet closer scrutiny of the situation forces the consideration of the possibility of ψ vanishing (which it will do generically along a line in three dimensions) about which Dirac says

> *There is an exceptional case, however, occurring when the wave function vanishes, since then its phase does not have a meaning. As the wave function is complex, its vanishing will require two conditions, so that in general the points at which it vanishes will lie along a line. We call such a line a* nodal line.[13]

Dirac now finds that to get something new he must relax the requirement that phase change round a closed curve be the same for all wave functions; he realises that it is possible to have it differ by integral multiples of 2π for different wave functions.

The final stage of the argument is to compute the flux of the magnetic field $\mathbf{B}$ through a closed surface S allowing also that nodal lines may lie totally within S or may intersect with S.

Each nodal line is labelled by an integer n which one detects by integrating round a small curve enclosing the line. If n_i are the integers for the nodal lines inside, or intersecting with, S the n_i are then related to the magnetic flux $\int_S \mathbf{B} \cdot \mathbf{ds}$ by

$$\sum_i 2\pi n_i + \frac{2\pi e}{hc} \int_S \mathbf{B} \cdot \mathbf{ds} = 0 \tag{2.2}$$

where h , c and e are Planck's constant, the velocity of light and the charge on the electron respectively.[14] Now if the nodal lines are closed they always cross S an even number of times and hence contribute zero to $\sum_i 2\pi n_i$ when one takes account of the sign of contributions associated with incoming and outgoing lines. Hence $\sum_i 2\pi n_i$ is only non-zero for those lines having end points within S. Finally Dirac then surrounds just *one* of these end points with a small surface S so that 2.2 immediately implies that magnetic flux emanates from this end point so that it is the location of a *magnetic monopole*. Such an end point will also be a singularity of the electromagnetic field.

Now if a single electric charge produces a field $\mathbf{E}$ then the size of its charge q is given by

$$q = \frac{1}{4\pi} \int_S \mathbf{E} \cdot \mathbf{ds} \tag{2.3}$$

where S is a small surface enclosing the charge. So, by analogy, the magnetic charge μ_0 of a magnetic monopole is defined by writing

$$\mu_0 = \frac{1}{4\pi} \int_S \mathbf{B} \cdot \mathbf{ds} \tag{2.4}$$

[13] In the current literature nodal lines are called Dirac strings

[14] Informally the reason that the RHS of 2.2 is zero is because it is the limit of the change in phase round a closed curve C as C shrinks to zero. C bounds a surface S so that when C has finally shrunk to zero S becomes closed.

Applying this to our small surface S and using 2.2 this gives at once the result that [15]

$$\begin{aligned} 2\pi n + \frac{8\pi^2 e}{hc}\mu_0 = 0 \\ \Rightarrow e\mu_0 = -\frac{nhc}{4\pi}, \quad n \in \mathbf{Z} \end{aligned} \tag{2.5}$$

We see that the magnetic charge μ_0 has a quantised strength with the fundamental quantum being

$$\frac{hc}{4\pi e} \tag{2.6}$$

which we note is inversely proportional to the electric charge e. Dirac noted the key fact that the mere existence of such a monopole means that electric charge is quantised. In Dirac **[1]** we find the passage

> *The theory also requires a quantisation of electric charge, since any charged particle moving in the field of a pole of strength μ_0 must have for its charge some integral multiple (positive or negative) of e, in order that the wave functions describing the motion may exist.*

Dirac's remarkable paper encouraged physicists to consider particles which are simultaneously magnetically and electrically charged—such particles are called *dyons*. The set of possible electric and magnetic charges form a skew lattice covering an entire $\mathbf{R}^2$ cf. fig. 3. For convenience from now on we denote magnetic charge by g rather than Dirac's μ_0 and in doing so we change to more conventional units of magnetic and electric charge: in these units $\hbar = c = 1$ and the charge q on the electron is given by $q^2/(4\pi) = 1/137$. Dirac's condition 2.5 now reads

$$\frac{eg}{2\pi} = n \in \mathbf{Z} \tag{2.7}$$

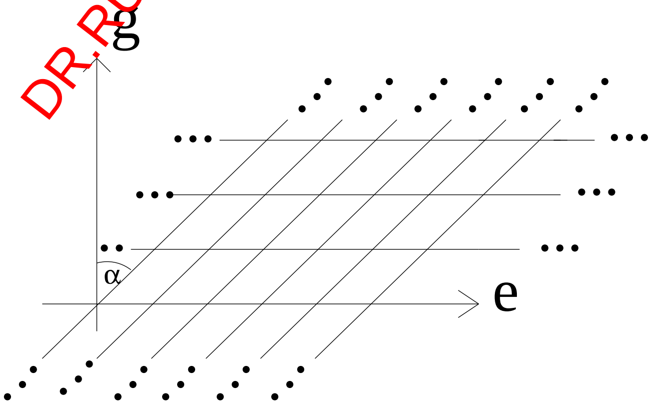

Fig. 3. **The lattice of possible magnetic and electric charges**

[15] When reading Dirac **[1]** one should be aware that the symbol h denotes Planck's constant divided by 2π; nowadays this quantity is usually denoted by $\hbar$.

The quantisation condition for dyons is a little more complicated than that for particles with only one type of charge. It has been studied independently by Schwinger [**1**] and Zwanziger [**1–2**]. They found that if (e_1, g_1) and (e_2, g_2) represent the electric and magnetic charges of a pair of dyons then

$$\frac{(e_1 g_2 - e_2 g_1)}{2\pi} = n \in \mathbf{Z}$$

and these values we represent on the lattice of fig. 3. If the angle α of fig. 3 is precisely zero then the lattice becomes rectangular rather than skew: this is practically forced if the theory is CP invariant since CP acts on (e, g) to give $(-e, g)$ (actually $\alpha = \pi/4$ is also allowed). Hence the size of α is a measure of CP breaking which is experimentally found to be small.

For monopoles coming from *non-Abelian* gauge theories serious consideration must be given to a specific topological mechanism for this CP breaking. We shall expand on this remark in section 4.

The topology of Dirac's monopoles

It is very instructive to examine Dirac's work from the point of view of topology. The mathematical setting is that of connections and curvatures on fibre bundles together with the vital calculational tool of characteristic classes; this latter was only at the very beginning of its development when Dirac [**1**] was published.

The standard physical setting is as follows. In electromagnetic theory the electric and magnetic fields form the components of the Maxwell field tensor $F_{\mu\nu}$ according to[16]

$$\begin{aligned} &E_i = F_{i0}, \quad B_i = \frac{1}{2}\epsilon_{ijk}F_{jk} \\ &F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \end{aligned} \tag{2.8}$$

However geometrically speaking the $F_{\mu\nu}$ are also the components of a curvature tensor for the gauge potential A_μ. This suggests immediately that one uses the curvature 2-form F and the connection 1-form A about which we know that

$$\begin{aligned} &F = \frac{1}{2}F_{\mu\nu}dx^\mu \wedge dx^\nu, \quad A = A_\mu dx^\mu \\ &F = dA \end{aligned} \tag{2.9}$$

Now because the electric field $\mathbf{E}$ is expressed as a gradient, while the magnetic field $\mathbf{B}$ is expressed as a curl, then it is natural to associate $\mathbf{E}$ to a 1-form and $\mathbf{B}$ to a 2-form. Hence we define the forms E and B by writing

$$E = E_i dx^i, \quad B = \frac{1}{2}F_{ij}dx^i \wedge dx^j \tag{2.10}$$

[16] In our notation x^0 denotes the time coordinate, summation is implied for repeated indices and we use the convention that Greek and Latin indices run from 0 to 3 and 1 to 3 respectively.

This means that the curvature 2-form F is expressible as

$$F = dx^0 \wedge E + B \tag{2.11}$$

and Maxwell's equations simply assert the closure of $*F$ i.e.

$$d * F = 0 \tag{2.12}$$

where the $*$ denotes the usual Hodge dual with respect to the standard flat Minkowski metric on $\mathbf{R}^4$.

But recall that Dirac specialises to a three dimensional situation by taking a time independent electromagnetic field. We see at once from 2.11 that, in the absence of time (or at a fixed time), F pulls back to a 2-form on $\mathbf{R}^3$ which is just the magnetic field 2-form B. Thus, denoting for convenience this 2-form on $\mathbf{R}^3$ by $F_{|_{\mathbf{R}^3}}$, we have

$$F_{|_{\mathbf{R}^3}} = B \tag{2.13}$$

It is important to note that the magnetic field in the guise of $F_{|_{\mathbf{R}^3}}$ is a curvature on $\mathbf{R}^3$ but the same is not true of the electric field. So particularly for a *static* electromagnetic field the magnetic and electric fields are geometrically *very different*.

Dirac's monopole has to be singular on $\mathbf{R}^3$ so its curvature $F_{|_{\mathbf{R}^3}}$ is a curvature defined on $\mathbf{R}^3 - \{0\}$ with associated connection form $A_i dx^i$. Now the gauge invariance of the monopole is the standard possibility of replacing A by $A + df$ where f is a function on $\mathbf{R}^3 - \{0\}$, i.e. the gauge group is the Abelian group $U(1)$.

So, in bundle theoretic terms, we have a connection on a $U(1)$ bundle P, say, over $\mathbf{R}^3 - \{0\}$; but $\mathbf{R}^3 - \{0\}$ is $S^2 \times \mathbf{R}^+$ ($\mathbf{R}^+$ is the positive real axis) and so homotopy invariance means we might as well consider P to be a $U(1)$ bundle over S^2. Such a P has an integral first Chern class $c_1(P)$ given in terms of its curvature form F by the standard formula

$$c_1(P) = \int_{S^2} \frac{F}{2\pi}, \quad c_1(P) \in \mathbf{Z} \tag{2.14}$$

We can now illustrate everything by doing a concrete calculation: If we use spherical polar coordinates[17] (r, θ, ϕ) and take

$$A = \frac{C}{2}(1 - \cos(\theta))d\phi, \quad C \text{ a constant} \tag{2.15}$$

[17] Were we to use Cartesian coordinates (x, y, z) we would obtain

$$A = \frac{C}{2r}\frac{(xdy - ydx)}{(z + r)}$$

Note that this expression has a *genuine singularity* at $r = 0$ and a *coordinate singularity* at $z + r = 0$—i.e. the negative z-axis—this latter singularity corresponds to Dirac's nodal line or the Dirac string. This coordinate singularity can be shifted to the positive z-axis by the gauge transformation $f = \tan^{-1}(y/x)$ yielding the connection

$$A + df = \frac{C}{2r}\frac{(xdy - ydx)}{(z - r)}$$

then the connection A has curvature

$$F = \frac{C}{2}\sin(\theta)\,d\theta \wedge d\phi \equiv \frac{C}{r^3}\epsilon_{ijk}x^i dx^j \wedge dx^k \tag{2.16}$$

Now F of course is the same as the magnetic field B or $F_{|_{\mathbf{R}^3}}$ introduced above but the integrality of the Chern class of P gives

$$\begin{aligned} \int_{S^2} \frac{C}{4\pi}\sin(\theta)\;d\theta \wedge d\phi = n \in \mathbf{Z} \\ \Rightarrow C = n \end{aligned} \tag{2.17}$$

so that the constant C is quantised, and this ensures that P is well defined. This then is Dirac's quantisation condition 2.5 in units where magnetic charges take integral values (i.e. units of size $hc/4\pi e$).

Aharonov and Bohm

After 1931, and the appearance of Dirac's paper, neither physics nor topology stood remotely still but they largely went along separate ways. To pinpoint a significant instance of an interaction between topology and physics we pass forward nearly three decades to 1959 and the paper of Aharonov and Bohm [**1**].

The Aharonov–Bohm effect [**1**] is a phenomenon in which the non-triviality of a gauge field A is measurable physically even though its curvature F is zero. Moreover this non-triviality is *topological* and can be expressed as a number n, say, which is a global topological invariant.

To demonstrate this effect physically one arranges that a *non simply connected* region Ω of space has zero electromagnetic field F. We are using the same notation as in the discussion above of Dirac's magnetic monopole, i.e we have

$$F = dA \tag{2.18}$$

and

$$F = \frac{1}{2}F_{\mu\nu}dx^\mu \wedge dx^\nu, \;\; \text{and} \;\; A = A_\mu dx^\mu \tag{2.19}$$

where the x_μ are local, coordinates on Ω.

Given this F and Ω one can devise an experiment in which one measures a diffraction pattern associated with the parallel transport of the gauge field A round a non-contractible loop C in Ω.

A successful experiment of precisely this kind was done by Brill and Werner [**1**] in 1960. The experimental setup—cf. fig. 4—was of the Young's slits type with electrons replacing photons and with the addition of a very thin solenoid. The electrons passed through the slits and on either side of the solenoid and an interference pattern was then detected. The interference pattern is first measured with the solenoid off. This pattern is then found to *change* when the solenoid is switched on even though the electrons always pass through a region where the field F is zero.

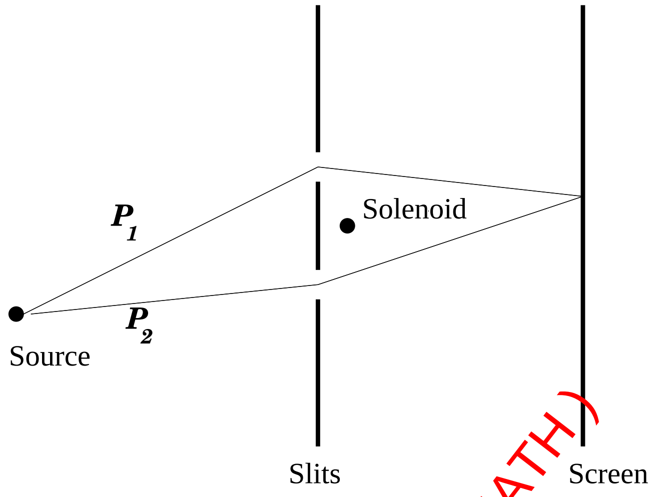

Fig. 4. **A schematic Aharonov–Bohm experiment**

This result is quite a dramatic demonstration of the fact that the connection, or gauge field, A is a more fundamental object than the electromagnetic field, or curvature F; all the more so since the key point is a topological one[18].

A topological explanation is easy to provide and begins with the notion of parallel transport round *closed* curves. The action of parallel transport on multilinear objects, viewed either as vectors, spinors, tensors etc. or sections of the appropriate bundles, is via the operator $PT(C)$ where

$$PT(C) = \exp\left[\int_C A\right]$$

Differential topology provides us immediately with the means to see that $PT(C)$ is non-trivial. The argument goes as follows: $F = dA$ so that the vanishing of F gives us

$$\begin{aligned} dA &= 0 \\ \Rightarrow A &= df \quad \text{locally} \end{aligned}$$

Hence A determines a de Rham cohomology class [A] and we have

$$[A] \in H^1(\Omega; \mathbf{R})$$

It is clear from Stokes' theorem that integral $\int_C A$ only depends on the homotopy class of the loop C; in addition the loop C determines a homology class $[C]$ where

$$[C] \in H_1(\Omega; \mathbf{R})$$

[18] It is possible that this result would not have been a complete surprise to Maxwell because, as is shown in our first quotation from him above, he was well aware of the necessity for topological considerations, and even knew something of their nature, when a function w as defined by a line integral.

This means that the integral $\int_C A$ (which we can take to be the number n) is just the dual pairing $(\bullet, \bullet)$ between cohomology and homology i.e.

$$([A], [C]) = \int_C A$$

Mathematically speaking we realise the solenoid by a cylinder L so that

$$\begin{aligned} \Omega &= \mathbf{R^3} - L \\ \Rightarrow H^1(\Omega; \mathbf{R}) &= H^1(\mathbf{R^3} - L; \mathbf{R}) \\ &= \mathbf{Z} \end{aligned}$$

In the experiment referred to above the loop C is the union of the electron paths P_1 and P_2, cf. fig. 4.

The holonomy group element $PT(C)$ is also of central interest elsewhere. It occurs in the study of the adiabatic periodic change of parameters of a quantum system described by Berry [**1**] (1984) cf. also Simon [**1**]. This phenomenon is known as Berry's phase. A relevant topological application here is an explanation of the quantum Hall effect, cf. Morandi [**1**].

The importance of non-trivial flat connections for physics extends to the non-Abelian or Yang–Mills case as we shall see in due course below.

It is now time to discuss topological matters concerning general relativity.

§ 3. Topology, general relativity and singularities—black holes and the big bang

Chandrasekhar and gravitational collapse

In 1965 Penrose [**1**] made the biggest breakthrough since Chandrasekhar [**1**] (1931) in understanding the nature of gravitational collapse: He proved the first theorem which showed that singularities of the gravitational field are a generic feature of gravitational collapse. Moreover Penrose's methods were topological.

Before saying anything more about Penrose's paper we need a brief sketch of some of the salient features of gravitational collapse.

Gravitational collapse is something that is worth investigating for very massive objects such as stars. This simple sounding idea is that, for a sufficiently massive body, the attractive force of gravity may be strong enough to cause it to start to *implode.*

To find something massive enough we have to choose a stellar object such as a star. Now, for a young active star, the burning of the nuclear fuel causes enough outward pressure to counteract all its gravitational inward pressure. However, since the nuclear fuel will eventually be used up this line of thought suggests that one calculate what gravity can do once it is not opposed by the nuclear reactions. In 1931 Chandrasekhar [**1**] published his celebrated paper on this matter.

He took a star of mass M to be a relativistic gas at temperature T obeying a relativistic equation of state. With the star's nuclear fuel all spent its cooling and contraction under gravity was opposed by the degeneracy pressure of the electrons produced by Fermi-Dirac statistics. However Chandrasekhar found that this pressure could not resist gravity if M was greater than about 1.4 solar masses (in standard notation one writes this as $1.4M_{\odot}$ where $M_{\odot}$ denotes the mass of the Sun). On the other hand for M less than $1.4M_{\odot}$ the star should cool and contract to what is called a white dwarf. Hence, for stars heavier than $1.4M_{\odot}$, unless something special intervened—for example a mechanism causing matter to be ejected during cooling until the Chandrasekhar limit is eluded—gravitational collapse is predicted.

No one (and this included Einstein) was very comfortable with this result but it resisted all the attempts made to get round it or even to disprove it. White dwarfs as ultimate fates of cooling stars were then supplemented by neutron stars.

Neutrons stars are so dense that their protons and electrons have combined to form neutrons; these neutrons then have a degeneracy pressure which resists the gravitation of the cooling star just as the electrons do in a white dwarf. The same ideas about the maximum mass of white dwarfs apply to neutron stars which then also have a maximum mass, this varies from about $2M_{\odot}$ to $3M_{\odot}$, the precise value depending on one's knowledge of the nuclear force, or strong interactions, at high densities.

Stars which are heavy enough[19] are thought not to end up in the graveyard of white dwarfs or neutrons stars but instead continue their collapse and form black holes[20].

The attitude to results about gravitational collapse as such as Chandrasekhar's **[1,2]** was that they were properties of the unrealistically high degree of symmetry of the solutions: collapse was not expected in the real Universe where such symmetry would not be found. This was also largely the attitude taken to the much later result of Oppenheimer and Snyder **[1]** (1939). This was a a paper (in which spherical symmetry was assumed) that produced the new result that a star undergoing gravitational collapse cut itself off from external observation as it contracted through a certain critical radius—the Schwarzschild radius. It contained, too, the facts about time asymptotically slowing to zero for an external observer but not for an observer moving with the star:

> *The total time of collapse for an observer co-moving with the star is finite .... an external observer sees the star asymptotically shrinking to its gravitational radius.*

The Schwarzschild metric plays a central part in understanding gravitational collapse and we shall now sketch some of its main properties.

[19] Apparently it may be possible for a large amount of matter to be shed by stars as they collapse: a star may even need a mass M greater than $20M_{\odot}$ in order to be forced to gravitationally collapse. However there are stars known with mass M ranging up to $100M_{\odot}$ so we do expect gravitationally collapsed stars to exist. For more details cf. chapter 9 of Hawking and Ellis **[1]**.

[20] Black holes, in the sense of *dark stars* from which light cannot escape, were discussed in Newtonian physics by Michell **[1]** (1784) and Laplace **[1]** (1799) (cf. appendix A of Hawking and Ellis **[1]** for a translation of Laplace **[1]**). They took light to obey a corpuscular theory and computed the size of a star whose escape velocity was greater than that of light. For other details of interest cf. the article by Israel in Hawking and Israel **[1]** and the excellent book for the layman by Thorne **[1]**.

The Schwarzschild metric

Schwarzschild [**1**] derived the form of a (spatially) spherically symmetric metric. With spherical polar coordinates (r, θ, ϕ), and time t, it is determined by the line element

$$ds^2 = -\left(1 - \frac{2m}{r}\right) dt^2 + \left(1 - \frac{2m}{r}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \tag{3.1}$$

This metric is meant to represent the empty space-time outside a spherically symmetric body of mass m[21]. We see at once a singularity at $r = 0$ and one at $r = 2m$. However the singularity at $r = 0$ is *genuine*—for example the Riemann curvature tensor diverges there—but the singularity at $r = 2m$ is only a *coordinate singularity* and disappears in an appropriately chosen coordinate system. Incidentally this is precisely analogous to the two singularities we encountered for the Dirac monopole: one at $r = 0$ and one at $z + r = 0$; we found that $r = 0$ was a real singularity but that $z + r = 0$ was only a coordinate singularity.

However the benign nature of the hypersurface $r = 2m$ was not realised for many years and it was usually misleadingly referred to as the "Schwarzschild singularity". This special value of r is called the *Schwarzschild radius* of the mass m. Schwarzschild [**2**] himself[22] went to the trouble of quoting the value for the Sun: it is very small, namely 3km.

This smallness led to the conviction that $r = 2m$ was irrelevant in practice because such a value of r lay deep down in the interior of any realistic body. Hence comfort was derived from the fact that the Schwarzschild metric was always used to describe the gravitational field in the empty space *outside* the star where r was always much bigger than the Schwarzschild radius. This was fine if a star never started to collapse but not otherwise.

The cosmologist Lemaître did notice in 1933 that $r = 2m$ was not a real singularity but this seems to have gone unnoticed or not been appreciated for a long time. In Lemaître [**1**] we find the words

> *La singularité du champ de Schwarzschild est donc une singularité fictive*

i.e.

> *The singularity of the Schwarzschild field is therefore a fictitious singularity*

It is clear, though, that an $(r = const, t = const)$ surface does change its character precisely when r passes through the value $2m$: For $r > 2m$ such a surface is timelike

[21] One might think of the Schwarzschild metric as a solution to the one body problem in general relativity; unfortunately things are worse than in Newtonian gravity where the three body problem was so hard to solve analytically, in general relativity the two body problem has not (yet!) been solved analytically; this means that accurate approximation methods must be used to treat important problems such as binary stars. As regards the Newtonian three body problem we should add that it was finally solved analytically in 1909 by Sundman [**1**], cf. also Sundman [**2,3**] and Barrow-Green [**1**], p. 187. Unfortunately, though the Sundman solution is a great triumph, his solution is a convergent series whose convergence rate is incredibly slow: apparently some 10^p terms, with p measuring in the millions, would be needed for practical work.

[22] Since this is a historical article I just add that this presumably was Schwarzschild's last article as he died of an illness while on the Russian front in 1916; he was born in 1876.

while for $r < 2m$ it is spacelike. This does mean that there is *something* special about the Schwarzschild radius, the question is just what is this something? Penrose was able to provide the answer and use it to make a breakthrough in understanding gravitational collapse. The point is that, for $r < 2m$, one can have what Penrose called a *trapped surface* and these we now consider.

Penrose and trapped surfaces

Gravitational collapse still refused to go away and in the early 1960's with the discovery of gigantic energy sources dubbed quasars the subject again became topical. It was suggested (other more conventional explanations didn't seem to fit) that the energy source of a quasar came from the gravitational collapse of an immensely massive object of mass $10^6 M_\odot$–$10^9 M_\odot$. Presumably such a collapse would not be spherically symmetric, for example one would expect there to be non-zero angular momentum. All this increased the need to study the possibility of gravitational collapse in general, i.e. without without any assumption of a special symmetry, spherical or otherwise.

Fortunately this is precisely where Penrose's topologically obtained result comes to the rescue. Penrose deduced that gravitational collapse to a space-time singularity was inevitable given certain reasonable conditions, and these conditions did *not* require any assumption about symmetry.

In Penrose [**1**] (1965) we find the statements

> *It will be shown that, after a certain critical condition has been fulfilled, deviations from spherical symmetry cannot prevent space-time singularities from arising .... The argument will be to show that the existence of a trapped surface implies—irrespective of symmetry—that singularities develop.*

Special attention has to be given to providing a definition of a singularity which is both mathematically and physically reasonable. In brief geodesic completeness is used as the basis for the definition of a singularity[23] of the space-time manifold $\mathcal{M}$; to see the significance of such completeness just consider that if a particle travelled along an *incomplete* timelike geodesic then it could disappear suddenly from $\mathcal{M}$ in a finite time.

It is impossible to give a detailed account here of the arguments so we shall only outline them; for a proper account cf. Hawking and Ellis [**1**].

As the matter constituting the star contracts it passes through its Schwarzschild radius $r = 2m$ and after this has happened the matter lies totally within a spacelike sphere S^2 (cf. fig. 5). This S^2 is what is called a trapped surface; technically it is closed, compact, spacelike,

[23] In relativity, since the metric is Lorentzian, geodesic completeness exists in three varieties: null, timelike and spacelike. To be singularity free both null and timelike geodesic completeness are demanded of $\mathcal{M}$; spacelike completeness is not required because physical motion does not take place along spacelike curves. In addition to geodesic completeness one also requires a causality condition and a non-negative energy condition: The causality condition is usually stated as the absence in $\mathcal{M}$ of any closed timelike curves—causes always precede effects. The non-negative energy condition is that, if $T_{\mu\nu}$ is the energy momentum tensor, then $T_{\mu\nu}t^\mu t^\nu \geq 0$ everywhere in $\mathcal{M}$ for all timelike vectors t^μ—in the rest frame of such a t^μ this becomes the statement that the energy $T_{00} \geq 0$ which is just another way of saying that gravity is always attractive.

two dimensional and has the property that null geodesics which intersect it orthogonally converge in the future.

The space-time manifold $\mathcal{M}$ is the future time development of an initial *non-compact* Cauchy hypersurface.

Figure 5 shows a space-time diagram of the collapse. In perusing the figure the reader should bear in mind that *one spatial dimension is suppressed* and that the circular symmetry of the diagram is there only for aesthetic reasons; the whole point being that no symmetry is assumed. The initial Cauchy hypersurface is represented by the plane at the bottom of the diagram.

The argument then computes the degree of an appropriate map which shows that null geodesic completeness implies that the future of the trapped surface is *compact.* However this is incompatible with the fact the initial Cauchy hypersurface is *non-compact*; this contradiction forces $\mathcal{M}$ to have a singularity.

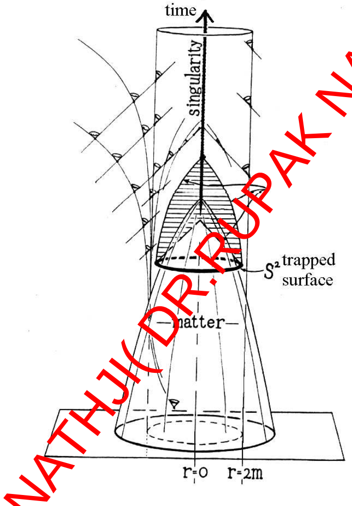

Fig. 5. **Gravitational collapse and a trapped surface** S^2

Some insight into the importance of a trapped surface can be obtained from a physical discussion: Normally if light is emitted radially outwards from all points on the surface S, say, of a sphere then it creates an outward moving spherical wave with surface S'; furthermore S' has a *bigger area* than S. However if S is a trapped surface then one finds that S' has a *smaller area* than S; this corresponds to the fact that gravity bends the light back and stops it escaping from the region inside $r = 2m$. Hence, as time progresses, S evolves to a smaller and smaller surface which eventually becomes a singularity, cf. again, fig. 5. The three surface $r = 2m$ is called the (absolute) *event horizon* of the collapse.

The use of the term black hole to describe such singularities is due to Wheeler who coined it in 1968, cf. Thorne [**1**]. The entire present day Universe may have originated in a past singularity known as the *big bang* a possibility for which there is considerable experimental evidence nowadays. This has resulted in the big bang being taken very seriously. However without the use of topological methods to convince one that singularities are generic under certain reasonable conditions the big bang would have been much more difficult to take seriously.

A further important paper on singularities was Hawking and Penrose [1] (1970); the general situation is discussed at great length and in full detail in Hawking and Ellis [1].

To round things off we point out that an important consequence of this work on singularities is that the set of solutions to the general relativistic hyperbolic Cauchy problem, which are destined to evolve into singularities, form a set of positive measure in an appropriate topology.

Just as topology was becoming a permanent bed fellow of relativity it also began to play a rôle in the Yang–Mills or non-Abelian gauge theories, these theories having moved to centre stage in elementary particle theory. This was to be an even more important event for topology as it has led to a genuinely two sided interaction between theoretical physics and mathematics [24] We begin this story in the next section.

§ 4. Topology and Yang–Mills theory—the latter day explosion

The rise of gauge invariance in the 1970's

There is no doubt that a principal factor in the rise of topology in physics is due to the rise to supremacy [25] of gauge theories in physics. Topology entered, in the main, via gauge theories: physicists learned that gauge theories had a formulation in terms of fibre bundles; they learned too that much useful cohomological data was possessed by these bundles.

Nielsen, Olesen, Polyakov and 't Hooft

An early important result of this period—which we may take to be post the papers of 't Hooft [1] and 't Hooft and Veltman [1] (1971–72)—is on magnetic monopoles in *non-Abelian* gauge theories. Two independent papers 't Hooft [2] and Polyakov [1] produced the first *non-Abelian monopole* (now referred to as the 't Hooft–Polyakov monopole). An earlier paper by Nielsen and Olesen [1] on magnetic vortices in superconductors was an important influence: In 't Hooft [2] the author opens with

> *The present investigation is inspired by the work of Nielsen et al.* [1]*, who found that quantized magnetic flux lines, in a superconductor, behave very much like the Nambu string* [2].

[24] There was also some work in the 1960's using algebraic topology to study singularities of Feynman integrals, cf. Hwa and Teplitz [1] (1966), Froissart [1] (1966) and Pham [1] (1967) but this has not continued to any great extent.

[25] This, of course, is another story and it is not our task to tell it here. However it is important for the reader to be aware that the interest in non-Abelian gauge theories was rekindled almost overnight with the vital proof of the renormalisability of non-Abelian gauge theories by 't Hooft [1] and 't Hooft and Veltman [1] (1971–72). This led to the resurrection of earlier papers on the subject and to the so called *standard model* with gauge group $SU(2) \times U(1)$ of weak and electromagnetic interactions and to QCD or *quantum chromodynamics*, with gauge group $SU(3)$, the favoured model for the confined quarks believed to be responsible for the strong interactions.

The 't Hooft–Polyakov monopole, like the Dirac monopole, is a static object and lives in $\mathbf{R}^3$; however, unlike the Dirac monopole, it has *no singularity* at the origin and is regular *everywhere* in $\mathbf{R}^3$. indeed as 't Hooft [**1**] says

> *Our way for formulating the theory of magnetic monopoles avoids the introduction of Dirac's string* [**3**].

The magnetic charge g of the monopole is topologically quantised and is inversely proportional to q where q is the electric charge of a heavy gauge boson in the theory. The topology enters through the boundary condition at infinity in $\mathbf{R}^3$. We shall now attempt to elucidate this by supplying some of the details of the mathematical setting.

The topology of monopole boundary conditions

Monopoles are static, finite energy, objects which give the critical points of the energy of an appropriate system of fields defined on a three dimensional Riemannian manifold M. In fact the usual choice for M is the non-compact space $\mathbf{R}^3$. Analysis on a non-compact M introduces some technical difficulties but these have not proved insurmountable.

The physical system studied consists of a Yang–Mills G-connection A, with curvature F, and a Higgs scalar field ϕ transforming according to the adjoint representation of G. If the Hodge dual with respect to the metric on M is denoted by $*$, then the energy E of the system is given by

$$E = \frac{1}{2}\int_M \left\{-tr(F \wedge *F) - tr(d_A\phi \wedge *d_A\phi) + \lambda * (|\phi|^2 - C^2)^2\right\} \tag{4.1}$$

where $d_A\phi$ is the covariant exterior derivative of the Higgs field, tr denotes the trace in the Lie algebra $\mathfrak{g}$ of G and $|\phi|^2 = -2tr(\phi^2)$. The field equations for the critical points of this system are difficult to solve explicitly (indeed the 't Hooft–Polyakov monopole is constructed numerically) but many solutions are available in what is called the Prasad–Sommerfield limit (cf. Prasad and Sommerfield [**1**]) where the scalar potential term vanishes. The energy is then

$$\begin{aligned} E \equiv E(A,\phi) &= -\frac{1}{2}\int_M \{tr(F \wedge *F) + tr(d_A\phi \wedge *d_A\phi)\} \\ &= \frac{1}{2}\{\|F\|^2 + \|d_A\phi\|^2\} \\ &= \frac{1}{2}\{\|F \mp *d_A\phi\|^2 \pm 2 < F, *d_A\phi >\} \end{aligned} \tag{4.2}$$

This shows that the absolute minima of E are attained when the pair (A,ϕ) satisfy

$$F = \mp * d_A\phi \tag{4.3}$$

which is the celebrated Bogomolny equation (Bogomolny [1]). The expression[26]

$$< F, *d_A\phi >= -\int_M tr(F \wedge d_A\phi) \tag{4.4}$$

is the absolute minimum and looks like a topological charge.

Now suppose that $M = \mathbf{R}^3$ furnished with the Euclidean metric and also set $G = SU(2)$. For the energy E to converge and to make the field equation problem well posed we must specify boundary conditions at infinity. A standard boundary condition for ϕ is

$$\lim_{r \to \infty} |\phi| \longrightarrow (C + O(r^{-2})) \tag{4.5}$$

where r is the distance from the origin in $\mathbf{R}^3$. The integral 4.4 can now be non-zero and it does have a topological interpretation which forces it to take discrete values. More precisely, if k is an integer, then

$$\frac{1}{4\pi C} < F, *d_A\phi >= k \tag{4.6}$$

This integer is the magnetic charge and can be thought of as the Chern class of a $U(1)$ bundle over a two sphere which is S^2_∞, the two sphere at infinity; setting $C = 1$ and using Stokes' theorem, we have

$$k = \frac{1}{4\pi}\int_{\mathbf{R}^3} tr(F \wedge d_A\phi) = \frac{1}{4\pi}\int_{S^2_\infty} tr(F\phi) \tag{4.7}$$

The condition $|\phi| = 1$ on the boundary defines an S^2 inside the Lie algebra $su(2)$ and $(F\phi)$, when evaluated at infinity, becomes the $U(1)$-curvature of a bundle over S^2_∞ and k is its Chern class. Alternatively, one can write k as the winding number, or degree, of a map $\hat{\phi} : S^2_\infty \to S^2_{su(2)}$, giving

$$\begin{aligned} \hat{\phi} : S^2_\infty &\longrightarrow S^2_{su(2)} \\ x &\longmapsto \hat{\phi}(x) = \frac{\phi}{|\phi|} \\ \text{and} \qquad k &= -\frac{1}{2\pi}\int_{S^2_\infty} tr(\hat{\phi}\, d\hat{\phi} \wedge d\hat{\phi}) \end{aligned} \tag{4.8}$$

This boundary integer k is the only topological invariant associated with the monopole system; the $SU(2)$ bundle over $\mathbf{R}^3$ is topologically trivial since $\mathbf{R}^3$ is contractible.

Dyons and CP breaking

In section 2 we promised to return to the subject of dyons and a *topological* mechanism for CP breaking. This CP breaking requires a non-Abelian monopole since it comes from the

[26] Throughout this article the inner product $< \omega, \eta >$ between Lie algebra valued p-forms ω and η on a Riemannian manifold M has the standard definition: i.e.

$$< \omega, \eta >= -\int_M tr(\omega \wedge *\eta >$$

presence in the action of a multiple of the second Chern class $c_2(P)$ given by (e and θ are real constants and for convenience we assume that the gauge group is $SU(N)$)

$$\frac{\theta e^2}{16\pi^2} tr(F \wedge F) = \frac{\theta e^2}{2} c_2(P) \tag{4.9}$$

This has been discussed by Witten [**1**] and it is immediate that such a term is CP non-invariant. The electric charge of a dyon now involves the θ parameter: one finds that for dyons with magnetic charge $g = 2\pi n_0/e$, $n_0 \in \mathbf{Z}$ their electric charge q obeys the formula

$$q = \left(n - \frac{\theta}{2\pi}\right) e,\, n \in \mathbf{Z} \tag{4.10}$$

This has the interesting feature that the *electric* charge of a dyon is not a *rational* multiple of a fundamental electric charge unless the CP violating Chern class coefficient θ is zero ($\theta = \pi$ is also allowed but may be too large experimentally). In terms of fig. 3 above it means that the angle α vanishes when θ is zero.

Still further insight into the rôle played by monopoles in quantum field theory has been obtained by combining the electric and monopole charges into a single *complex* parameter $e + ig$; we shall discuss this in section 8.

Gauge theories in four dimensions: Instantons

Topology came even more to the fore in Yang–Mills theories with the publication in 1975 by Belavin et al. [**1**] of topologically non-trivial solutions to the *Euclidean* Yang–Mills equations in four dimensions. Such solutions have come to be called *instantons*[27].

Of fundamental importance for these solutions to the Euclidean Yang–Mills equations is that instantons are at the same time *non-perturbative* and *topological*.

The Euclidean version of a quantum field theory is obtained from the Minkowskian version by replacing the Minkowski time t by it. The relation between the two theories is supposed to be one of analytic continuation in the Lorentz invariant inner products $x_\mu y^\mu$; allowing these inner products to be complex is the simple way to pass from one theory to the other. The existence of such a continuation makes tacit certain assumptions which require proof; significant progress in this technical matter was made in 1973–75, cf. Osterwalder and Schrader [**1–3**], and Streater [**1**].

The term instanton, though not quite precise, is often generalised to refer to a critical point of finite Euclidean action for any quantum field theory. Such solutions to the equations of motion—for this is what these critical points are—are closely related to quantum mechanical tunnelling phenomena. This property of instantons quickly attracted great interest because tunnelling amplitudes are not calculable perturbatively but require a knowledge of the theory for large coupling as well as small.

This opening of the door into the room of non-perturbative techniques was a noteworthy event and we shall see below that topology was a key ingredient to picking the lock on this door.

[27] Quite a few early papers on the subject used the less attractive term pseudoparticle instead of instanton but luckily this usage was short-lived.

A good account of this is to be found in Coleman **[1]** (1977) who showed his pleasure at the progress made in his opening sentences:

> *In the last two years there have been astonishing developments in quantum field theory. We have obtained control over problems previously believed to be of insuperable difficulty and we have obtained deep (at least to me) insights into the structure of the leading candidate for the theory of strong interactions, quantum chromodynamics.*

In Coleman **[1]** there is also an account of an important paper ('t Hooft **[3]**) which used instantons to solve an outstanding problem known as "the $U(1)$ problem", thereby imbuing the fledgling instantons with considerable status. We shall now give a brief summary of some of the more salient features of an instanton in the Yang-Mills case.

Profile of an instanton

Our life can be made a little easier by choosing a very specific setting: we have a non-Abelian gauge theory with G a compact simple Lie group and action

$$S \equiv S(A) = \|F\|^2 = -\int_M \operatorname{tr}(F \wedge *F) \tag{4.11}$$

with M a closed four dimensional orientable Riemannian manifold and $*$ the Hodge dual with respect to the Riemannian metric on M. Instantons are those A which correspond to critical points of S; however we shall specialise the term here to mean only *minima* of S.

First we should obtain the Euler–Lagrange equations of motion, i.e. the equation for the critical points. Let A be an arbitrary connection through which passes the family of connections

$$A_t = A + ta \tag{4.12}$$

Expanding in the vicinity of $t = 0$ gives

$$\begin{aligned} S(A_t) &= < F(A), F(A) > + t\frac{d}{dt} < F(A_t), F(A_t) >|_{t=0} + \cdots \\ \text{and} \qquad F(A_t) &= F(A) + t(da + A \wedge a + a \wedge A) + t^2 a \wedge a \\ &= F(A) + t d_A a + t^2 a \wedge a \\ \Rightarrow S(A_t) &= \|F(A)\|^2 + t\{< d_A a, F(A) > + < F(A), d_A a >\} + \cdots \\ &= S(A) + 2t < F(A), d_A a > + \cdots \end{aligned} \tag{4.13}$$

A is a critical point if

$$\left.\frac{dS(A_t)}{dt}\right|_{t=0} = 0 \tag{4.14}$$

That is, if

$$\begin{aligned} < F(A), d_A a > &= 0 \\ \Rightarrow < d_A^* F(A), a > &= 0 \\ \Rightarrow d_A^* F(A) &= 0, \quad \text{since } a \text{ is arbitrary} \end{aligned} \tag{4.15}$$

However, $F(A) = dA + A \wedge A$ also satisfies the Bianchi identity $d_A F(A) = 0$ and so we have the pair of equations

$$d_A F(A) = 0, \qquad d_A^* F(A) = 0 \tag{4.16}$$

This is similar to the condition for a form ω to be harmonic, which is

$$d\omega = 0, \qquad d^*\omega = 0 \tag{4.17}$$

It should be emphasised, though, that the Yang–Mills equations are not linear; thus they really express a kind of non-linear harmonic condition.

The most distinguished class of solutions to the Yang–Mills equations $d_A^* F(A) = 0$ is that consisting of those connections whose curvature is self-dual or anti-self-dual.

To see how such solutions originate we point out that with respect to our inner product on 2-forms d_A^* has the property that

$$d_A^* = - * d_A * \tag{4.18}$$

so that the Yang–Mills equations become

$$d_A * F(A) = 0 \tag{4.19}$$

Thus if $F = \mp * F$ the Bianchi identities immediately imply that we have a solution to the Yang–Mills equations—we have managed to solve a non-linear second order equation by solving a non-linear first order equation.

It is also easy to see that these critical points are all *minima* of the action S; here are the details. First we (orthogonally) decompose F into its self-dual and anti-self-dual parts F^+ and F^-, giving

$$\begin{aligned} F &= \frac{1}{2}(F + *F) + \frac{1}{2}(F - *F) \\ &= F^+ + F^- \\ \Rightarrow S &= \|(F^+ + F^-)\|^2 \\ &= \|F^+\|^2 + \|F^-\|^2 \end{aligned} \tag{4.20}$$

where the crossed terms in the norm contribute zero.

The *topological type* of the instanton is classified by the second Chern class $c_2(F) \in H^2(M; \mathbf{Z})$ of the bundle on which the connection A is defined: Taking G to be the group $SU(N)$ and evaluating $c_2(F)$ on M we obtain the integer

$$c_2(F)[M] = \frac{1}{8\pi^2} \int_M tr(F \wedge F) \in \mathbf{Z} \tag{4.21}$$

The *instanton number*, k, is defined to be minus this number so we find that

$$\begin{aligned} k &= -\frac{1}{8\pi^2} \int_M tr(F \wedge F) \\ &= -\frac{1}{8\pi^2} \int_M tr\{(F^+ + F^-) \wedge (F^+ + F^-)\} \\ &= \frac{\|F^+\|^2 - \|F^-\|^2}{8\pi^2} \end{aligned} \tag{4.22}$$

The inequality $(a^2+b^2) \geq |a^2-b^2|$ shows that, for each k, the absolute minima of S are attained when

$$S = 8\pi^2|k| \tag{4.23}$$

and this corresponds to $F^{\mp}=0$ or equivalently

$$F = \mp * F \tag{4.24}$$

and we have the celebrated self-dual and anti-self-dual conditions. Changing the orientation of M has the effect of changing the sign of the $*$ operation and so interchanges F^+ with F^-.

Up to now, although we have not mentioned it, for algebraic convenience we have set the coupling constant of the theory equal to unity. But to understand anything non-perturbative the coupling must be present so we now temporarily cease this practice. Denoting the coupling constant by g (the context should prevent any confusion with magnetic charge) the action S is given by

$$S \equiv S(A) = \frac{1}{g^2}\|F\|^2 = -\frac{1}{g^2}\int_M tr(F \wedge *F) \tag{4.25}$$

Hence if A is an instanton then we immediately have

$$S(A) = \frac{8\pi^2|k|}{g^2},\ k \in \mathbf{Z} \tag{4.26}$$

Finally the corresponding quantum mechanical amplitude is $\exp[-S]$ so that we have

$$\exp[-S(A)] = \exp\left[-\frac{8\pi^2|k|}{g^2}\right] \tag{4.27}$$

which we see at once is an *inverse* power series in g^2; moreover topology is uppermost for we note that for this inverse power series to exist the instanton number k must be non-zero.

The mathematicians take a strong interest

The pace of instanton research increased towards the end of the 1970's due in part to a keen interest being taken in the problems by some highly able and gifted mathematicians. As we shall see below this attack on the problems by two distinct groups was to prove highly beneficial to both physics *and* mathematics. In fact some particularly choice fruits of these labours fell into the garden of the mathematicians.

Thus far we have stressed the topological nature of the connections of Yang–Mills theory: the relevant mathematical structure is a fibre bundle and together with this comes cohomological characteristic class data giving discrete numerical invariants such as the instanton number k. However, for instantons, there remains a more prosaic object to study namely the non-linear partial differential equation for the instanton A itself, i.e. the self-duality equation

$$F = *F \tag{4.28}$$

or, more explicitly,

$$\partial_\mu A^a_\nu - \partial_\nu A^a_\mu + igf^{abc}A^b_\mu A^c_\nu = \frac{1}{2}\epsilon_{\mu\nu\alpha\beta}(\partial^\alpha A^{a\beta} - \partial^\beta A^{a\alpha} + igf^{abc}A^{b\alpha}A^{c\beta}) \tag{4.29}$$

A key change of viewpoint on the self-duality equation changed the focus away from differential equations; this was the breakthrough made by Ward [**1**] (1977).

Ward showed that the solution of the self-duality partial differential equation was equivalent to the construction of an appropriate vector bundle. His paper (Ward [**1**]) gives a brief summary at the beginning

> *In this note we describe briefly how the information of self-dual gauge fields may be "coded" into the structure of certain complex vector bundles, and how the information may be extracted, yielding a procedure by which (at least in principle) all self-dual solutions of the Yang–Mills equations may be generated. The construction arose as part of the programme of twistor theory* [**3**]*; it is the Yang–Mills analogue of Penrose's "non-linear graviton" construction* [**4**]*, which relates to self-dual solutions of Einstein's vacuum equations.*

This discarding of the differential equation and its encoding into the transition functions of certain vector bundles immediately made the problem of more interest and accessibility to mathematicians. Atiyah and Ward [**1**] showed how the problem was equivalent to one in algebraic geometry; there then followed a complete solution to the problem for $M = S^4$ by Atiyah, Drinfeld, Hitchin and Manin [**1**], the situation for other four manifolds is treated in Atiyah, Hitchin and Singer [**1**].

Atiyah, who was to become a key figure in many subsequent developments of joint interest to mathematicians and physicists describes his introduction to Yang–Mills theories as follows (taken from the preface to Atiyah [**2**])

> *My acquaintance with the geometry of Yang–Mills equations arose from lectures given in Oxford in Autumn 1976 by I. M. Singer, and I am very grateful to him for arousing my interest in this aspect of theoretical physics.*

It was not long before a large body of both mathematicians and physicists were working on a large selection of problems related in some way to Yang–Mills theories. The next breakthrough was in mathematics rather than in physics and we turn to this in the section that follows.

§ 5. The Yang–Mills equations and four manifold theory

Donaldson's work

In the 1980's interest in instantons continued strongly but there was a most striking result proved by Donaldson [**1**] (1983) which used the Yang-Mills instantons to make a fundamental advance in the topology of four manifolds.

Donaldson's result concerned simply connected compact closed four manifolds M. We shall now give a short account of some of the result's main features so that the reader may be better able to appreciate its significance.

In topology one distinguishes three types of manifold M: topological, piecewise-linear and differentiable (or smooth) which we can denote when necessary by M_{TOP}, M_{PL} and

M_{DIFF} respectively. There are topological obstacles to the existence of PL and DIFF structures on a given topological manifold M. The nature of these obstacles is quite well understood in dimension 5 and higher but, in dimension 4, the situation is quite different and much more difficult to comprehend. It is for this dimension that Yang–Mills theories and Donaldson's work have made such an important contribution.

On the subject of the importance of Yang–Mills theories for obtaining these results Donaldson and Kronheimer **[1]** (p. 27) have said the following in favour of Yang–Mills theory.

> *These geometrical techniques will then be applied to obtain the differential–topological results mentioned above. It is precisely this departure from standard techniques which has led to the new results, and at present there is no way known to produce results such as these which does not rely on Yang–Mills theory.*

Donaldson and simply connected four manifolds

We consider here compact closed four manifolds M. For a simply connected four manifold M, $H_1(M;\mathbf{Z})$ and $H_3(M;\mathbf{Z})$ vanish and the non-trivial homological information is concentrated in the middle dimension in $H_2(M;\mathbf{Z})$. A central object then is the *intersection form* defined by

$$q(\alpha,\beta) = (\alpha \cup \beta)[M], \qquad \alpha,\beta \in H_2(M;\mathbf{Z}) \tag{5.1}$$

with $\cup$ denoting cup product so that $(\alpha \cup \beta)[M]$ denotes the integer obtained by evaluating $\alpha \cup \beta$ on the generating cycle $[M]$ of $H_4(M;\mathbf{Z})$ on M. Poincaré duality implies that the intersection form is always *non-degenerate* over $\mathbf{Z}$ and so has $\mathit{det}\, q = \mp 1$—q is then called unimodular. Also we refer to q, as *even* if all its diagonal entries are even, and as *odd* otherwise. A very powerful result of Freedman **[1]** can now be called on—the intersection form q very nearly determines the homeomorphism class of a simply connected M, and actually only fails to do so in the odd case where there are still just two possibilities. Further *every* unimodular quadratic form occurs as the intersection form of some manifold.

The relevant theorem is

Theorem (Freedman **[1]** (1982)) *A simply connected 4-manifold M with even intersection form q belongs to a unique homeomorphism class, while if q is odd there are precisely two non-homeomorphic M with q as their intersection form.*

An illustration of the impressive nature of Freedman's work is readily available. Recollect that the Poincaré conjecture in four dimensions is the statement that any homotopy 4-sphere, S_h^4 say, is actually *homeomorphic* to the standard sphere S^4. Now S^4 has trivial cohomology in two dimensions so its intersection form q is the zero quadratic form which we write as $q = \emptyset$. But S_h, having the same homotopy type as S^4, has the same cohomology as S^4. So any S_h^4 also has intersection form $q = \emptyset$. But Freedman's result says that for a simply connected M with even q there is only *one homeomorphism class* for M, therefore S_h^4 *homeomorphic* to S^4 and we have established the conjecture. Incidentally this means that the Poincaré conjecture has now been proved for all n except $n = 3$—the case originally proposed by Poincaré.

Now we come to Donaldson's work which concerns smoothability of four manifolds; one should also note that, when q is a definite quadratic form, a choice of orientation can always render q *positive* definite. Then we have the following theorem

Theorem (Donaldson [**1**] (1983)) *A simply connected, smooth 4-manifold, with positive definite intersection form q has the property that q is always diagonalisable over the integers to $q = diag(1, \ldots, 1)$*

Immediately one can go on to deduce that no simply connected, 4-manifold for which q is even and positive definite can be smoothed! For example the Cartan matrix for the exceptional Lie algebra e_8 is given by

$$E_8 = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 2 \end{pmatrix} \tag{5.2}$$

Freedman's result guarantees that there is a manifold M with intersection form $q = E_8 \oplus E_8$. However Donaldson's theorem forbids such a manifold[28] from existing smoothly. Before Donaldson's work surgery techniques had been extensively used to try to construct smoothly the manifold with intersection form $E_8 \oplus E_8$. We can now see that these techniques were destined to fail.

In fact, in contrast to Freedman's theorem, which allows *all* unimodular quadratic forms to occur as the intersection form of some topological manifold, Donaldson's theorem says that in the positive definite, smooth, case only *one* quadratic form is allowed, namely the identity I.

One of the most striking aspects of Donaldson's work is that his proof uses the Yang–Mills equations. We can only outline what is involved here, for more details cf. Donaldson and Kronheimer [**1**], Freed and Uhlenbeck [**1**] and Nash [**1**].

In brief then let A be a connection on a principal $SU(2)$-bundle over a simply connected 4-manifold M with positive definite intersection form. If S is the usual Euclidean Yang–Mills action S of 4.11 one has

$$S = \|F\|^2 = -\int_M tr(F \wedge *F) \tag{5.3}$$

Now given one instanton A which minimises S one can perturb about A in an attempt to find more instantons. When this is done the space of all instantons can be fitted together to form a global *moduli space* of finite dimension. For the instanton with $k = 1$ which provides

[28] The reader may wonder why we did not discuss the four manifold with the simpler intersection form $q = E_8$. This manifold of course exists. It is not smoothable but this fact is due to a much older result of Rohlin [**1**] (1952) concerning smoothability and the signature of q. Rohlin's theorem only provides a necessary condition for smoothability, this is that the signature of an even q must be divisible by 16. The lack of sufficiency of this condition is shown by the example of $q = E_8 \oplus E_8$ since one can verify that the signature of q in this case is divisible by 16.

the absolute minimum of S, this moduli space $\mathcal{M}_1$, say, is a non-compact space of dimension 5, with singularities.

We can now summarise the logic that is used to prove Donaldson's theorem: there are very strong relationships between M and the moduli space $\mathcal{M}_1$; for example, let q be regarded as an $n \times n$ matrix with precisely p unit eigenvalues (clearly $p \leq n$ and Donaldson's theorem is just the statement that $p = n$), then $\mathcal{M}_1$ has precisely p singularities which look like cones on the space $\mathbf{C}P^2$. These combine to produce the result that the 4-manifold M has the same topological signature $Sign\,(M)$ as p copies of $\mathbf{C}P^2$; now p copies of $\mathbf{C}P^2$ have signature $a - b$ where a of the $\mathbf{C}P^2$'s are oriented in the usual fashion and b are given the opposite orientation. Thus we have

$$Sign\,(M) = a - b \tag{5.4}$$

Now the definition of $Sign\,(M)$ is that it is the signature $\sigma(q)$ of the intersection form q of M. But since, by assumption, q is positive definite $n \times n$ then $\sigma(q) = n = Sign\,(M)$. So we can write

$$n = a - b \tag{5.5}$$

However, $a + b = p$ and $p \leq n$ so we can assemble this information in the form

$$n = a - b, \qquad p = a + b \leq n \tag{5.6}$$

but one always has $a + b \geq a - b$ so now we have

$$n \leq p \leq n \Rightarrow p = n \tag{5.7}$$

and we have obtained Donaldson's theorem.

§ 6. Physics and knots revisited—the Jones polynomial

Three manifolds and Floer, Jones and Witten

In section one we discussed knots in our material on the nineteenth century. It is now time to return to this subject.

In 1985 Jones [**1**] made a great step forward in knot theory by introducing a new polynomial invariant of knots (and links), now known as the *Jones polynomial* and denoted by $V_L(t)$, where L denotes the knot or link and t is a real variable. Knot invariants of this kind had proved hard to find: the original one was that of Alexander [**1**] (1928), denoted by $\Delta_L(t)$.

The Jones polynomial originates in certain finite dimensional von Neumann algebras which Jones denotes by A_n. A point of physical interest here is that, as Jones observed in his paper, D. Evans pointed out that some representations of these A_n's had already been constructed in the physics literature in statistical mechanics, the relevant reference (which Jones gives) being Lieb and Temperley [**1**]. The statistical mechanics concerns the Potts and ice-type models, cf. Baxter [**1**]. This leads one to speculate that the combinatorial structure of some models in statistical mechanics has a *topological origin*; this does seem to be borne out by subsequent work.

The Jones polynomial proved powerful enough to decide many of the longstanding Tait conjectures on knots which we referred to in section one. The next event of joint topological and physical interest was a result by Witten [**2–3**] (1988) which gave a completely new formulation (and generalisation) of the Jones polynomial in terms of a certain kind of quantum field theory, nowadays known as a *topological quantum field theory*.

A vital ingredient in this whole story is the work of Floer [**1–2**] (1987–88) on a new homology invariant of three manifolds constructed from considerations of gauge theory and instantons. We shall meet this work again in section 8. We mention it now because of its influence on subsequent work. For the moment we just need to inform the reader that Floer considers the critical point behaviour of the function f where f depends on an $SU(2)$ connection A: f is simply the Chern–Simons function obtained by integrating the Chern–Simons secondary characteristic class over a closed three manifold M. If $\mathcal{A}$ denotes the space of all connections A, we have

$$\begin{aligned} f : \mathcal{A} &\longrightarrow \mathbf{R} \\ A &\longmapsto f(A) \\ \text{with} \qquad f(A) &= -\frac{1}{8\pi^2}\int_M tr(A\wedge dA + \frac{2}{3}A\wedge A\wedge A) \end{aligned} \tag{6.1}$$

Then from a very clever study of the Morse theory of this function f, whose domain is the *infinite dimensional* space $\mathcal{A}$, Floer obtains new homology groups $HF(M)$ known as *Floer* homology groups associated to the three manifold M.

In 1987 Atiyah [**3**] speculated that there was a relation between Floer's work and the Jones polynomial: Towards the end of this paper we find the following

> *Finally let me list a few of the major problems that are still outstanding in the area. .... More speculatively, I would like to end with*
> *4) Find a connection with the link invariants of Vaughan Jones [11].*
>
> *As circumstantial evidence that this is reasonable I will list some properties shared by Floer homology and the Jones polynomial.*
>
> *(i) both are subtle 3-dimensional invariants,*
> *(ii) they are sensitive to orientation of 3-space (unlike the Alexander polynomial),*
> *(iii) they depend on Lie groups: $SU(2)$ in the first instance but capable of generalisation,*
> *(iv) there are 2-dimensional schemes for computing these 3-dimensional invariants,*
> *(v) whereas the variable in the Alexander polynomial corresponds to $\pi_1(S^1)$, the variable in the Jones polynomial appears to be related to $\pi_3(S^3)$, the origin of "instanton numbers",*
> *(vi) both have deep connections with physics, specifically quantum field theory (and statistical mechanics).*

In 1988 Witten [**2**] rose to this challenge and found the relation that Atiyah had suspected existed. The content of Witten [**3**] is described by its author with a certain amount of understatement. He says

> *In a lecture at the Hermann Weyl Symposium last year* [**1**]*, Michael Atiyah proposed two problems for quantum field theorists. The first problem was to give a physical interpretation for Donaldson theory. The second problem was to find an intrinsically three dimensional definition of the Jones polynomial of knot theory. I would like to give a flavour of these two problems.*

Our next task is to have a look at the methods that Witten used.

Topological quantum field theories

A topological quantum field theory (also called simply a *topological field theory*) is one which, at first sight, may seem trivial physically: it has an action with no metric dependence. The absence of a metric means that there are no distance measurements or forces and so no conventional dynamics. The Hamiltonian $\mathcal{H}$ of the theory has only zero eigenstates and the Hilbert space of the theory is usually finite dimensional. The theory can, however, be non-trivial: its non-triviality is reflected in the existence of *tunnelling* between vacua.

The particular action chosen by Witten for obtaining the Jones polynomial was the well known Chern–Simons action S given by

$$S = \frac{ik}{4\pi g^2}\int_M tr(A \wedge dA + \frac{2}{3}A \wedge A \wedge A),\ k \in \mathbf{Z} \tag{6.2}$$

where A is an $SU(2)$ connection or gauge field and M is a closed, compact three dimensional manifold. The partition function for this quantum field theory is $Z(M)$ where

$$Z(M) = \int \mathcal{DA}\, \exp\left[-\frac{ik}{4\pi g^2}\int_M tr(A \wedge dA + \frac{2}{3}A \wedge A \wedge A)\right] \tag{6.3}$$

This partition function itself is an invariant—the *Witten invariant*—of the three manifold M; however, at present, we want to study knots: knots enter in the following way. Consider a closed curve C embedded in M so as to form a knot, K, say. Now ones takes the connection A, parallel transports it around C and constructs the holonomy operator $PT(C)$ as described above in the discussion of the Aharonov–Bohm effect in section 2. This time we have a non-Abelian connection and to obtain a gauge invariant operator we must take the trace of $PT(C)$ giving what is called a *Wilson line*; we denote this by $W(R, C)$ where

$$W(R, C) = tr\, P \exp\left[\int_C A\right] \tag{6.4}$$

and R denotes the particular representation carried by A. There is a natural correlation function associated with this knot namely

$$< W(R, C) > = \frac{1}{Z(M)}\int \mathcal{DA}\, W(R, C) \exp\left[-\frac{ik}{4\pi g^2}\int_M tr(A \wedge dA + \frac{2}{3}A \wedge A \wedge A)\right] \tag{6.5}$$

and Witten [**2–3**] shows that this determines the Jones polynomial $V_K(t)$ of the knot.

Further if one has not *one* curve C but a number of them, say $C_1, \ldots C_p$ then one has a p component link L, say, whose Jones polynomial $V_L(t)$ is determined by a multiple correlation function of p Wilson lines given by

$$\begin{gathered}< W(R_1,C_1)\cdots W(R_p,C_p) > = \frac{1}{Z(M)}\int \mathcal{D}\mathcal{A} \\ W(R_1,C_1)\cdots W(R_p,C_p)\exp\left[-\frac{ik}{4\pi g^2}\int_M tr(A\wedge dA+\frac{2}{3}A\wedge A\wedge A)\right]\end{gathered} \tag{6.6}$$

A nice thing that happens if we step backwards slightly to the *Abelian* case is that one can recover the Gauss linking number: In the Abelian case A is just a $U(1)$ connection and $S(A)$ becomes only quadratic in A giving (we have set $g=1$)

$$S(A)=\frac{ik}{4\pi}\int_M tr(A\wedge dA) \tag{6.7}$$

so that

$$\begin{gathered}< W(R_1,C_1)\cdots W(R_p,C_p) > = \\ \frac{1}{Z(M)}\int \mathcal{D}\mathcal{A}\, W(R_1,C_1)\cdots W(R_p,C_p)\exp\left[-\frac{ik}{4\pi}\int_M tr(A\wedge dA)\right]\end{gathered} \tag{6.8}$$

The quadratic action, together with the exponential dependence on A of a Wilson line, allows the entire integrand to be written as a Gaussian after completing the square. The calculation of the functional integral rests just on the calculation of a Green's function which, for $M=S^3$ (to which we now specialise), is an elementary computation. The result is that

$$\begin{gathered}< W(n_1,C_1)\cdots W(n_p,C_p) > = \\ \exp\left[\frac{i}{4k}\epsilon_{ijk}\sum_{l,m=1}^{p} n_l n_m \int_{C_l} dx^i \int_{C_m} dy^j \frac{(x-y)^k}{|x-y|^3}\right]\end{gathered} \tag{6.9}$$

with x^i and y^j local coordinates on the knots C_l and C_m.

We easily recognise the basic integral

$$\frac{\epsilon_{ijk}}{4\pi}\int_{C_l} dx^i \int_{C_m} dy^j \frac{(x-y)^k}{|x-y|^3} \tag{6.10}$$

in 6.9 as the linking number of Gauss that we met in section 1; and we note that we have met it again in a physical context. Incidentally for another physical context in which the linking number appears cf. Wilczek and Zee [**1**] (1983); in this paper a connection is made between the spin-statistics properties of particles and topology [29]

In the non-Abelian case we can also obtain a quadratic functional integral by studying the limit of small coupling g, or, completely equivalently, the limit of large k. For any

[29] On this latter topic there is more work. For some examples cf. Balachandran et al. [**1**], Berry and Robbins [**1**], Finkelstein and Rubinstein [**1**], Mickelsson [**1**], Tscheuschner [**1**], and references therein.

topological field theory such a limit has considerable significance. When this limit is evaluated for this theory one obtains another differential topological invariant: the Ray–Singer torsion of the connection A on M, cf. Witten [**2–3**].

Numerous topological field theories are now studied in the current literature, we shall meet another one in section 7; indeed the whole notion of a topological quantum field theory has been axiomatised in Atiyah [**4**].

Finally we point out some of the new and more general features of the Witten approach to the Jones polynomial: Witten's definition is intrinsically three dimensional and not dependent on any two dimensional arguments for its validation. The group $SU(2)$ of the connection A is not obligatory—it can be replaced by another Lie group G, say $G = SU(N)$; for the appropriate representation of $SU(N)$ this gives rise to a two variable generalisation of the Jones polynomial cf. Freyd, et al. [**1**]. There is an immediate generalisation to knots in *any* three manifold M rather than the classical case of knots in S^3 (i.e. compactified $\mathbf{R}^3$). Invariants for three manifolds themselves immediately arise and so the theory is not really just one of knots (i.e. embeddings).

§ 7. Yang–Mills and four manifolds once more

Donaldson again: polynomial invariants for four manifolds

In the 1990's more progress was made in four dimensions with another result of Donaldson [**3**] (1990); actually some of these results were announced considerably earlier in 1986 by Donaldson in his Field's medal address (cf. Donaldson [**2**]).

In section 5 we described Donaldson's use of the moduli space $\mathcal{M}_1$ to derive smoothability results about 4-manifolds. The space $\mathcal{M}_1$ only contains instantons with instanton number k equal to one. In addition to this, by using all values of k there exist moduli spaces $\mathcal{M}_k$, $k = 1, 2, \ldots$ for instantons of any instanton number k. Donaldson's new invariants use all of the $\mathcal{M}_k$ and in the process one obtains powerful differential topological invariants of simply connected 4-manifolds. Donaldson [**3**] begins with

> *The traditional methods of geometric topology have not produced a clear picture of the classification of smooth 4-manifolds. This gap has been partially bridged by methods using Yang–Mills theory or gauge theory. Riemannian manifolds carry with them an array of moduli spaces—finite dimensional spaces of connections cut out by the first order Yang–Mills equations. These equations depend on the Riemannian geometry of the 4-manifold, but at the level of homology we find properties of the moduli spaces which do not change when the metric is changed continuously. Any two metrics can be joined by a path, so by default, these properties depend only on the differential topology of the 4-manifold, and furnish a mine of potential new differential topological invariants.*

The "mine of potential new differential topological invariants", as Donaldson modestly puts it, is a reference to his new polynomial invariants. He goes on, in the same paper, to say

> *Here we use infinite families of moduli spaces to define infinite numbers of invariants for simply connected manifolds with b_2^+ odd*[30] *and greater than 1.*

[30] The number b_2^+ is defined to be the rank of the positive part of the intersection form.

These invariants are distinguished elements in the ring:

$$S^*(H^2(X))$$

of polynomials on the cohomology of the underlying 4-manifold X.

Equivalently, they can be viewed as symmetric multi-linear functions:

$$q : H_2(X; \mathbf{Z}) \times \cdots \times H_2(X : \mathbf{Z}) \longrightarrow \mathbf{Z}$$

.... Certainly one of the most striking facts is that we get infinitely many invariants for a single manifold. Discovering to what extent these are independent (i.e. whether there are strong universal relations between them) appears to be an interesting target for future research.

We just want to mention some results that have been obtained with the Donaldson invariants which serve to show that they are non-trivial and important. To be able to do this we must introduce some notation.

Let M be a smooth, simply connected, orientable Riemannian four manifold without boundary and A be an $SU(2)$ connection which is anti-self-dual[31] so that

$$F = - * F \tag{7.1}$$

Then the dimension of the moduli space $\mathcal{M}_k$ is the integer

$$\dim \mathcal{M}_k = 8k - 3(1 + b_2^+) \tag{7.2}$$

A Donaldson invariant $q_d(M)$ is a symmetric integer polynomial of degree d in the 2-homology $H_2(M; \mathbf{Z})$ of M

$$q_d(M) : \underbrace{H_2(M) \times \cdots \times H_2(M)}_{d \text{ factors}} \longrightarrow \mathbf{Z} \tag{7.3}$$

Given a certain map m (cf. Donaldson [**3–4**] or Nash [**1**])

$$m : H_2(M) \to H^2(\mathcal{M}_k) \tag{7.4}$$

we use m to define by $q_d(M)$ by using de Rham cohomology and differential forms. Setting $d = \dim \mathcal{M}_k/2$ we define $q_d(M)$ by

$$\begin{aligned} q_d(M) : H_2(M) \times \cdots \times H_2(M) &\longrightarrow \mathbf{Z} \\ a_1 \times \cdots \times a_d &\longmapsto \int_{\overline{\mathcal{M}}_k} m(a_1) \wedge \cdots \wedge m(a_d) \end{aligned} \tag{7.5}$$

where $\overline{\mathcal{M}}_k$ denotes a compactification of the moduli space. We see that the $q_d(M)$ are symmetric integer valued polynomials of degree d in $H^2(M)$, i.e. $q_d(M) \in Sym^d\,(H_2(M)) \subset$

[31] We take anti-self-dual connections rather than self-dual connections so as to follow Donaldson's sign conventions.

$S^*(H(M))$; also, since $d = \dim \mathcal{M}_k/2 = (8k - 3(1 + b_2^+))/2$, we now understand why Donaldson required b_2^+ to be odd.

Now the Donaldson invariants are, *a priori*, not very easy to calculate since they require detailed knowledge of the instanton moduli space. However, if M is a complex algebraic surface, a positivity argument shows that

$$q_d(M) \neq 0, \qquad \text{for } d \geq d_0 \tag{7.6}$$

with d_0 some integer—in other words the $q_d(M)$ are all non-zero when d is large enough. Conversely, if M can be written as the connected sum

$$M = M_1 \# M_2 \tag{7.7}$$

where M_1 and M_2 both have $b_2^+ > 0$ then

$$q_d(M) = 0, \qquad \text{for all } d \tag{7.8}$$

The $q_d(M)$ are *differential* topological invariants rather than topological invariants; this means that they have the potential to distinguish homeomorphic manifolds which have distinct diffeomorphic structures. An example where the $q_d(M)$ are used to show that two homeomorphic manifolds are not diffeomorphic can be found in Ebeling **[1]**. A possible a physical context for this result can be found in Nash **[2]**, cf. too, Libgober and Wood **[1]** for some earlier work related to Ebeling **[1]** which was done before the $q_d(M)$ were defined.

The $q_d(M)$ can also be obtained from a topological quantum field theory as we shall now see below.

Another topological field theory

In 1988 Witten showed how to obtain the $q_d(M)$ as correlation functions in a BRST supersymmetric topological field theory. We shall only give a brief statement of facts to give the reader some idea of what sort of action and physical fields are involved; for a full account, cf. Witten **[4]**.

The action S for the theory is given by

$$\begin{aligned} S = \int_M d^4x\sqrt{g}\, tr\, \{\, &\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{4}F^*_{\mu\nu}F^{\mu\nu} + \frac{1}{2}\phi D_\mu D^\mu \lambda + iD_\mu \psi_\nu \chi^{\mu\nu} \\ &- i\eta D_\mu \psi^\mu - \frac{i}{8}\phi[\chi_{\mu\nu}, \chi^{\mu\nu}] - \frac{i}{2}\lambda[\psi_\mu, \psi^\mu] - \frac{i}{2}\phi[\eta, \eta] - \frac{1}{8}[\phi, \lambda]^2 \,\} \end{aligned} \tag{7.9}$$

where $F_{\mu\nu}$ is the curvature of a connection A_μ and $(\phi, \lambda, \eta, \psi_\mu, \chi_{\mu\nu})$ are a collection of fields introduced in order to construct the right supersymmetric theory; ϕ and λ are both spinless while the multiplet $(\psi_\mu, \chi_{\mu\nu})$ contains the components of a 0-form, a 1-form and a self-dual 2-form respectively. The significance of this choice of multiplet is that the anti-instanton version of the instanton deformation complex used to calculate $\dim \mathcal{M}_k$ contains precisely these fields. Even though S contains a metric its correlation functions are independent of the metric g so that S can still be regarded as a topological field theory. This can be shown to follow from the fact that both S and its associated energy momentum tensor $T \equiv (\delta S/\delta g)$ can be written as BRST commutators $S = \{Q, V\}$, $T = \{Q, V'\}$ for suitable V and V'—cf. Witten [**4**].

With this theory it is possible to show that the correlation functions are independent of the gauge coupling and hence we can evaluate them in a small coupling limit. In this limit the functional integrals are dominated by the classical minima of S, which for A_μ are just the instantons

$$F_{\mu\nu} = -F^*_{\mu\nu} \tag{7.10}$$

We also need ϕ and λ to vanish for irreducible connections. If we expand all the fields around the minima up to quadratic terms and do the resulting Gaussian integrals, the correlation functions may be formally evaluated. Let us consider a correlation function

$$< P >= \int \mathcal{DF} \, \exp[-S] \, P(\mathcal{F}) \tag{7.11}$$

where $\mathcal{F}$ denotes the collection of fields present in S and $P(\mathcal{F})$ is a polynomial in the fields. Now S has been constructed so that the zero modes in the expansion about the minima are the tangents to the moduli space $\mathcal{M}_k$; thus, if the $\mathcal{DF}$ integration is expressed as an integral over modes, all the non-zero modes may be integrated out first leaving a *finite dimensional* integration over $\dim \mathcal{M}_k$. The Gaussian integration over the non-zero modes is a Boson–Fermion ratio of determinants, a ratio which supersymmetry constrains to be ∓ 1 since Bosonic and Fermionic eigenvalues are equal in pairs. This amounts to expressing $< P >$ as

$$< P >= \int_{\mathcal{M}_k} P_n \tag{7.12}$$

where P_n is an n-form over $\mathcal{M}_k$ and $n = \dim \mathcal{M}_k$. If the original polynomial $P(\mathcal{F})$ is chosen in the correct way then calculation of $< P >$ reproduces evaluation of the Donaldson polynomials.

The next breakthrough in the topology of four manifolds came from physics and was due to Seiberg and Witten (1994) (cf. Seiberg and Witten [**1–2**] and Witten [**5**]) and it is the next topic to which we turn.

Physics again: Seiberg–Witten theory and four manifolds

In Donaldson [**4**] we find the most upbeat introduction to a review article on the Seiberg–Witten equations; it gives some idea of the excitement and the power of the methods associated with this latest breakthrough.

> *Since 1982 the use of gauge theory, in the shape of the Yang–Mills instanton equations, has permeated research in 4-manifold topology. At first this use of differential geometry and differential equations had an unexpected and unorthodox flavour, but over the years the ideas have become more familiar; a body of techniques has built up through the efforts of many mathematicians, producing results which have uncovered some of the mysteries of 4-manifold theory, and leading to substantial internal conundrums within the field itself. In the last three months of 1994 a remarkable thing happened: this research was turned on*

> *its head by the introduction of a new kind of differential–geometric equation by Seiberg and Witten: in the space of a few weeks long–standing problems were solved, new and unexpected results were found, along with simpler new proofs of existing ones, and new vistas for research opened up. This article is a report on some of these developments, which are due to various mathematicians, notably Kronheimer, Mrowka, Morgan, Stern and Taubes, building on the seminal work of Seiberg [S] and Seiberg and Witten [SW].*

We shall say a little about both the physics and the mathematics relating to the Seiberg–Witten equations; however we shall make the remarks about the mathematics here but leave the remarks about the physics until section 9 where they fit in more naturally.

Seiberg and Witten's work allows one to produce another physical theory, in addition to Witten [**4**], with which to compute Donaldson invariants. In Witten [**4**], as just described above, Donaldson theory is obtained from a twisted $N = 2$ supersymmetric Yang–Mills theory. Seiberg and Witten produce a duality which amounts to an equivalence between the *strong coupling* limit of this $N = 2$ theory and the *weak coupling* limit of a theory of *Abelian* monopoles. This latter theory is much easier to compute with leading (on the mathematical side) to the advances described in Donaldson [**4**] and Witten [**5**].

If we choose an oriented, compact, closed, Riemannian manifold M then the data we need for the Seiberg–Witten equations are a connection A on a line bundle L over M and a "local spinor" field ψ. The Seiberg–Witten equations are then

$$\partial\!\!\!/_A \psi = 0, \qquad F_A^+ = -\frac{1}{2}\,\overline{\psi}\,\Gamma\psi \tag{7.13}$$

where $\partial\!\!\!/_A$ is the Dirac operator and Γ is made from the gamma matrices Γ_i according to $\Gamma = \frac{1}{2}[\Gamma_i, \Gamma_j]dx^i \wedge dx^j$. We call ψ a local spinor because global spinors may not exist on M; however orientability guarantees that a *spin*$_{\mathbf{c}}$ structure does exist and ψ is the appropriate section for this *spin*$_{\mathbf{c}}$ structure. We note that A is just a $U(1)$ Abelian connection and so $F = dA$, with F^+ just being the self-dual part of F.

We shall now have a brief look at one example of a new result using the Seiberg-Witten equations. The equations clearly provide the absolute minima for the action

$$S = \int_M \left\{ |\partial\!\!\!/_A \psi|^2 + \frac{1}{2}|F^+ + \frac{1}{2}\,\overline{\psi}\,\Gamma\psi|^2 \right\} \tag{7.14}$$

If we use a Weitzenböck formula to relate the Laplacian $\nabla_A^* \nabla_A$ to $\partial\!\!\!/_A^* \partial\!\!\!/_A$ plus curvature terms we find that S satisfies

$$\begin{aligned}\int_M \left\{ |\partial\!\!\!/_A \psi|^2 + \frac{1}{2}|F^+ + \frac{1}{2}\,\overline{\psi}\,\Gamma\psi|^2 \right\} &= \int_M \left\{ |\nabla_A \psi|^2 + \frac{1}{2}|F^+|^2 + \frac{1}{8}|\psi|^4 + \frac{1}{4}R|\psi|^2 \right\} \\ &= \int_M \left\{ |\nabla_A \psi|^2 + \frac{1}{4}|F|^2 + \frac{1}{8}|\psi|^4 + \frac{1}{4}R|\psi|^2 \right\} + \pi^2 c_1^2(L)\end{aligned} \tag{7.15}$$

where R is the scalar curvature of M. The action now looks like one for monopoles—indeed in Witten [**5**], Witten refers to 7.13 as "the monopole equations". But now suppose that

R is *positive* and that the pair (A, ψ) is a solution to the Seiberg–Witten equations: then the LHS is zero and all the integrands on the RHS are positive so the solution must obey $\psi = 0$ and $F^+ = 0$. It turns out that if M has $b_2^+ > 1$ then a perturbation of the metric can preserve the positivity of R but change $F^+ = 0$ to be plain $F = 0$ rendering the connection A flat. Hence, in these circumstances, the solution (A, ψ) is the trivial one. This means that we have a new kind of vanishing theorem in four dimensions.

Theorem (Witten [**3**] (1994)) *No four manifold with $b_2^+ > 1$ and non-trivial Seiberg–Witten invariants admits a metric of positive scalar curvature.*

We referred just now to the Seiberg–Witten invariants and unfortunately we cannot define them here. However we do want to say that they are rational numbers a_i and there are formulae relating the Donaldson polynomial invariants q_d to the a_i.

Many more new results have been found involving, for example, symplectic and Kähler manifolds, cf. Donaldson [**4**]; the story, however, is clearly not at all finished.

§ 8. Dynamics and topology since Poincaré

Dynamical systems and Morse theory

In this section we want to return to Poincaré and consider that part of his topological legacy which sprang from his work on dynamics. We shall only be able to look at two areas and these are the theory of *dynamical systems* and *Morse theory*. This is, of necessity, somewhat selective, nevertheless these two subjects do represent mainstream developments which descend directly from Poincaré's work on dynamics and topology[32]. It should also be borne in mind that there is a large overlap between the two subjects.

It is still true that the n-body problem attracts much attention from mathematicians, including those using topological techniques. A few references of interest here are Smale [**2–3**] (1970) and Saari and Xia [**1**] (1996).

Dynamical systems

Poincaré's pioneering work on celestial mechanics prepared the way for the present day subject of dynamical systems with Birkhoff as the actual founder. In this subject one studies an immense diversity of sophisticated mathematical problems usually no longer connected with celestial or Newtonian mechanics.

A very rough idea of what is involved goes as follows: Recall first that the celestial mechanics of n bodies has a motion that is described by a set of differential equations together with their initial data. One then varies the initial data and asks how the motion changes.

Now the modern mathematical setting is to view the orbits of the n bodies as integral curves for their associated differential equations. Then one regards the *qualitative study* of the orbits as being a study of the *global geometry* of the space of integral curves as their

[32] Some more detailed historical material, of relevance here, is that of Dahan-Dalmédico [**1–2**].

initial conditions vary smoothly. Integral curves $\gamma(t)$ are associated with vector fields $V(t)$ via the differential equation

$$\frac{d\gamma(t)}{dt} = V(\gamma(t)) \tag{8.1}$$

Hence one is now studying the vastly more general subject of the global geometry of the space of flows of a vector field V on a manifold M.

Two notions play a distinguished part in the theory of dynamical systems: closed integral curves and singular points. It is natural to regard two flows on M as *equivalent* if there is a homeomorphism of M which takes one flow into the other; one can also insist that this homeomorphism is smooth, i.e. a diffeomorphism. Finally an equivalence class of flows in the homeomorphic sense is a *topological dynamical system*, and one in the diffeomorphic sense is *a smooth, or differentiable, dynamical system.*

As we explained in section 1 Birkhoff proved Poincaré's geometric theorem in 1913; a subsequent piece of work of great importance and influence was Birkhoff's proof of what is called his *ergodic theorem* in 1931, cf. Birkhoff [**2**].

The subsequent blossoming of ergodic theory can be dated from this time. Ergodic theory originates largely in nineteenth century studies in the kinetic theory of gases. However it has now been axiomatised, expanded, refined and reformulated so that it has links with many parts of mathematics as well as retaining some with physics.

Some dynamical systems exhibit ergodic behaviour, a notable class of examples being provided by *geodesic flow* on surfaces of constant negative curvature. This involves too the study of the flows by a discrete encoding known as symbolic dynamics, use of one dimensional interval maps cf. Bedford et al. [**1**]. Classical and quantum chaos, and the distinction between the two, are also studied in this context.

A vast body of the theory of dynamical systems concerns *Hamiltonian systems.* These of course have their origin in ordinary dynamics but exist now in a much wider context. To have a Hamiltonian system M must be even dimensional, possess a Hamiltonian function

$$H : M \longrightarrow \mathbf{R} \tag{8.2}$$

and have a closed non-degenerate symplectic form 2-form ω appropriately related to H. The perturbation theory of Hamiltonian systems underwent an enormous development in the 1950's and 1960's with the work particularly of Kolmogorov, Arnold and Moser and the creation of what is known as KAM theory (cf. Broer et al. [**1**]).

Gradient dynamical systems were used by Thom [**1–3**] (1969–71) in his work on what is now called *Catastrophe theory.* Thom took the system

$$\frac{d\gamma(t)}{dt} = grad\, V(\gamma(t)) \tag{8.3}$$

where V is a potential function. Thom classified the possible critical points of V into seven types known as the seven elementary catastrophes; he then proposed to use these dynamical systems as models for the behaviour of a large class of physical, chemical and biological systems. In many cases the models are not at all adequate, nevertheless, there are some successes. However the seminal nature of Thom's work is clear though as it is the beginning of the classification theory for singularities. In this connection there are the two results of Arnold [**1–2**] (1973, 1978) which closely relate the classification of singularities to the Weyl groups of the various compact simple Lie groups.

Morse theory: the topology of critical points

The aim in Morse theory is to study the relation between critical points and topology. More specifically one extracts topological information from a study of the critical points of a smooth real valued function

$$f : M \longrightarrow \mathbf{R} \tag{8.4}$$

where M is a compact manifold usually without boundary. For a suitably behaved class of functions f there exists quite a tight relationship between the number and type of critical points of f and topological invariants of M such as the Euler–Poincaré characteristic, the Betti numbers and other cohomological data. This relationship can then be used in two ways: one can take certain special functions whose critical points are easy to find and use this information to derive results about the topology of M; on the other hand, if the topology of M is well understood, one can use this topology to infer the existence of critical points of f in cases where f is too complex, or too abstractly defined, to allow a direct calculation.

Taking a function f the equation for its critical points is

$$df = 0 \tag{8.5}$$

We *assume* that all the critical points p of f are non-degenerate; this means that the Hessian matrix Hf of second derivatives is invertible at p, or

$$det\, Hf(p) \neq 0 \quad \text{where} \quad Hf(p) = \left[\partial^2 f/\partial x^i \partial x^j \big|_p \right]_{n\times n} \tag{8.6}$$

Each critical point p has an index λ_p which is defined to be the number of *negative* eigenvalues of $Hf(p)$. We can then associate to the function f and its critical points p the Morse series $M_t(f)$ defined by

$$M_t(f) = \sum_{\text{all } p} t^{\lambda_p} = \sum_i m_i t^i \tag{8.7}$$

The topology of M now enters via $P_t(M)$: the Poincaré series of M. We have

$$P_t(M) = \sum_{i=0}^{n} \dim H^i(M;\mathbf{R}) t^i = \sum_{i=0}^{n} b_i t^i \tag{8.8}$$

The fundamental result of Morse theory is the statement that

$$M_t(f) \geq P_t(M) \tag{8.9}$$

from which so many things follow, to mention just one simple example

$$m_i \geq b_i \tag{8.10}$$

showing that the number of critical points of index i is bounded below by the Betti number b_i.

Successful applications of Morse theory in mathematics are impressive and widespread; a few notable examples are the proof by Morse [**1**] (1934) that there exist infinitely many geodesics joining a pair of points on a sphere S^n endowed with any Riemannian metric, Bott's [**1–2**] (1956, 1958) proof of his celebrated periodicity theorems on the homotopy of Lie groups, Milnor's construction [**1**] (1956) of the first exotic spheres, and the proof by Smale [**1**] (1961) of the Poincaré conjecture for $\dim M \geq 5$.

Supersymmetric quantum mechanics and Morse theory

In 1982 Witten [**6**] constructed a quantum mechanical point of view on Morse theory which has proved very influential. It also provides a point of departure for the Floer theory discussed below. In summary Witten gives a quantum mechanical proof of the Morse inequalities on M; however an important extra feature is that the cohomology of M is also explicitly constructed.

Witten takes as Hamiltonian $\mathcal{H}$: the Hodge Laplacian on forms, i.e. one has

$$\mathcal{H} = \bigoplus_{p\geq 0} \Delta_p \equiv \bigoplus_{p\geq 0} (dd^* + d^*d)_p \tag{8.11}$$

The Bosons and Fermions of the supersymmetry are the spaces H^B and H^F formed by even and odd forms while the supersymmetry algebra is generated by two operators Q_1 and Q_2 which are constructed from d and d^*. The appropriate definitions are

$$\begin{aligned} Q_1 &= (d + d^*) \qquad & Q_2 &= i(d - d^*) \\ H^B &= \bigoplus_{p\geq 0} \Omega^{2p}(M) \qquad & H^F &= \bigoplus_{p\geq 0} \Omega^{2p+1}(M) \end{aligned} \tag{8.12}$$

A Morse function f is now incorporated into the model without changing the supersymmetry algebra by replacing d by d_t where

$$d_t = e^{-ft} d e^{ft} \tag{8.13}$$

It is a routine matter to verify that this conjugation of d by e^{ft} leaves the algebra unchanged. The proof of the Morse inequalities rests on an analysis of the spectrum of the associated Hamiltonian, which is now

$$\mathcal{H}_t = d_t d_t^* + d_t^* d_t = \bigoplus_{p\geq 0} \Delta_p(t) \tag{8.14}$$

One needs additional physics to carry out the rest of the work. It turns out that this all comes from the consideration of quantum mechanical tunnelling between critical points of f. One considers *grad* f as a vector field on M and then studies the integral curves of this vector field, that is the solutions $\gamma(t)$ of the differential equation

$$\frac{d\gamma(s)}{ds} = -grad\, f(\gamma(s)) \tag{8.15}$$

We can give no more details here but, as has been emphasised by Witten [**6**], these ideas are applicable in quantum field theory as well as in quantum mechanics. In that case one has to deal with functions in infinite dimensions and it was not long before a significant result along these lines emerged; this was the work of Floer **[1–2]** (1988) which we now examine.

Floer homology and Morse theory

In section 6, cf. 6.1, we referred to Floer's work and his Morse theoretic study of the function

$$\begin{aligned} f : \mathcal{A} &\longrightarrow \mathbf{R} \\ A &\longmapsto f(A) \\ \text{with} \qquad f(A) &= -\frac{1}{8\pi^2}\int_M tr(A \wedge dA + \frac{2}{3}A \wedge A \wedge A) \end{aligned} \tag{8.16}$$

We review now some of the details.

The critical points of f are given by

$$df(A) = 0 \tag{8.17}$$

where the exterior derivative is now taken to be acting in the space $\mathcal{A}$ of $SU(2)$ connections on M. If A is such a critical point then we can write $A_t = A + ta$ and obtain

$$f(A_t) = f(A) - \frac{t}{4\pi^2}\int_M tr(F(A) \wedge a) + \cdots \tag{8.18}$$

Hence we can conclude that

$$df(A) = -\frac{F(A)}{4\pi^2} \tag{8.19}$$

and so the critical points of the Chern–Simons function are the *flat connections* on M.

As long as $\pi_1(M) \neq 0$ then flat connections on M are not trivial, since they can have non-zero holonomy round a non-trivial loop on M. The holonomy of each flat connection is an $SU(2)$ element parametrised by a loop on M; in this way it defines a representation of $\pi_1(M)$ in $SU(2)$ and the space of *inequivalent* such representations is the quotient

$$Hom\,(\pi_1(M), SU(2))/Ad\,SU(2) \tag{8.20}$$

Having found a critical point Morse theory requires us to calculate its index and so we must also calculate the Hessian of f: the snag is that this gives an operator which is unbounded from below rendering the index formally infinite. This is not entirely unexpected since we are working in infinite dimensions.

Floer gets round this very cleverly by realising that he only needs a *relative index* which he can compute via spectral flow and the Atiyah–Singer index theorem. He takes two critical points A_P and A_Q in $\mathcal{A}$ and joins then with a steepest descent path $A(t)$; i.e. a path which obeys the equation

$$\frac{dA(t)}{dt} = -grad\, f(A(t)) \tag{8.21}$$

with *grad* denoting the gradient operator on the space $\mathcal{A}$. The consequence of all this for the Morse theory construction is that he is able to construct a homology complex and associated homology groups $HF_p(M)$. However the topology of the situation dictates that the relative Morse index of f is only well defined $mod\,8$. This means that $HF_p(M)$ are graded $mod\,8$ and one only obtains eight homology groups: $HF_p(M)$, $p = 0, \ldots 7$; for more details cf. Nash [**1**].

Morse theory has also been successfully applied to other problems in Yang–Mills theory. Some important papers are Atiyah and Bott [**1**] (1982) on Yang–Mills theories on Riemann surfaces where an equivariant Morse theory was required, and Taubes [**1–2**] (1988, 1985) on pure Yang–Mills theory and Yang–Mills theory for monopoles. In all these examples one has to grapple with the infinite dimensionality of the quantum field theory.

Knots again

Vassil'ev [**1–2**] (1990) has developed an approach to knot theory using singularity theory. Vasil'ev constructs a huge new class of knot invariants and we shall now give a sketch of what is involved.

A knot is a smooth embedding of a circle into $\mathbf{R}^3$. So a knot gives a map

$$f : S^1 \longrightarrow \mathbf{R}^3 \tag{8.22}$$

so that f belongs to the space $\mathcal{F}$ where $\mathcal{F} = Map\,(S^1, \mathbf{R}^3)$. Not all elements of $\mathcal{F}$ give knots since a knot map f is not allowed to self-intersect or be singular. Let Σ be the subspace of $\mathcal{F}$ which contains either self-intersecting or singular maps, then the subspace of knots is the *complement*

$$\mathcal{F} - \Sigma \tag{8.23}$$

Now any element of Σ can be made smooth by a simple one parameter deformation, hence Σ is a *hypersurface* in $\mathcal{F}$ and is known as the *discriminant.* As the discriminant Σ wanders through $\mathcal{F}$ it skirts along the edge of the complement $\mathcal{F} - \Sigma$ and divides it into many different connected components. Clearly knots in the same connected component can be deformed into each other and so are equivalent (or isotopic).

Now any knot *invariant* is, by the previous sentence, a function which is *constant* on each connected component of $\mathcal{F} - \Sigma$. Hence the task of constructing all (numerical) knot invariants is the same as finding all functions on $\mathcal{F} - \Sigma$ which are constant on each connected component. But topology tells us at once that this is just the 0-cohomology of $\mathcal{F} - \Sigma$. In other words

$$H^0(\mathcal{F} - \Sigma) = \text{The space of knot invariants} \tag{8.24}$$

Vasil'ev provides a method for computing most, and possibly all, of $H^0(\mathcal{F} - \Sigma)$. There are connections, too, to physics cf. Bar-Natan, [**1**].

§ 9. Strings, mirrors and duals

String theory, supersymmetry and unification of interactions

String theory has, by now, a fairly long history very little of which we can mention here. We shall, in the main, limit ourselves to remarks which relate in some way to topology.

Topology enters string theory at the outset because a moving string sweeps out a two dimensional surface and, in quantum theory, all such surfaces must be summed over whatever their topology. This leads to the Polyakov expression for the partition function Z of the Bosonic string which contains a sum over the genera p of Riemann surfaces Σ. One has

$$Z = \sum_{genera} \int \mathcal{D}g\mathcal{D}\phi \, \exp\left[-\frac{1}{2}\int_\Sigma < \partial\phi, \partial\phi >_g\right] = \sum_{p=0}^{\infty} Z_p \tag{9.1}$$

where the functional integral is over all metrics g on Σ and the string's position ϕ. We have $\phi \equiv \phi^\mu(x^1, x|^2)$, $\mu = 1, \ldots d$ so the ϕ^μ can be thought of as specifying an embedding of Σ in a d-dimensional space-time M. It is well known that the theory is only conformally invariant when $d = 26$: the critical dimension; however if one includes Fermions then this critical dimension changes to $d = 10$.

A string theory is also a two dimensional conformal field theory and this latter subject is very important for mathematics as well as physics. It has been axiomatised in a very fruitful and influential way by Segal [**1**] (1988). It involves important representation theory of infinite dimensional groups such as $LG = Map(S^1, G)$ and $Diff(S^1)$. We regret that we have been unable to trace its history in this article because of lack of space; its importance is immediately apparent when one reads the literature on string theory as well as that of many statistical mechanical models, it is also a key notion used in calculation and conceptual work in string theory. Finally, closely connected to conformal field theories, are the subjects of Kac–Moody algebras, vertex operator algebras and quantum groups; these all have close connections with physics but, although they have topological aspects, their algebraic properties are more prominent and this, as well as considerations of space, is another reason why we have had to omit them from this essay. Conformal field theory is also a vital ingredient in the surgery argument used in Witten [**2–3**] to compute the Jones polynomial and, viewed from this standpoint, conformal field theories can be seen to provide a link between the infinite dimensional representational theory just mentioned and the topology of two and three manifolds.

String theory came into its own with the incorporation of supersymmetry in the early 1980's cf. Green, Schwarz and Witten [**1–2**].

Quantum field theories with chiral Fermions sometimes exhibit a pathological behaviour when coupled to gauge fields; gauge invariance may break down, this is referred to as an anomaly. Such anomalies, too, have played a major part in shaping present state of string theory; a key paper here is that of Green and Schwarz [**1**] (1984) who discovered a remarkable anomaly cancellation mechanism which thereby singles out five distinguished supersymmetric $d = 10$ string theories. Anomalies also have an important topological aspect involving the Atiyah–Singer index theorem for families of Dirac operators. We have not had space to discuss this here, cf. Nash [**1**] for more details.

The low energy limit of a string theory is meant to be a conventional quantum field theory which should be a theory that describes all known interactions including gravity. Supersymmetric string theories do succeed in including gravity and so offer the best chance so far for a quantum theory of gravity as well as, perhaps, for an eventual unified theory of all interactions.

The five string theories singled out by the work of Green and Schwarz [**1**] are all supersymmetric and contain Yang–Mills fields. These five 10 dimensional theories are denoted by type I, type IIA, type IIB, $E_8 \times E_8$ heterotic and $SO(32)$ heterotic.

Only four of the ten dimensions of space-time M are directly observable; so the remaining six are meant to form a small (i.e. small compared with the string scale) compact six dimensional space M_6, say. The favoured physical choice for M_6 is that it be a three complex dimensional *Calabi–Yau* manifold—this means that it is complex manifold of a special kind: it is Kähler with holonomy group contained in $SU(3)$. The favoured status of Calabi–Yau manifolds has to do with what is called *mirror symmetry* which we now briefly review as it is of considerable interest to both mathematicians (for example algebraic geometers) and physicists.

Mirror symmetry

Mirror symmetry refers to the property that Calabi–Yau manifolds come in dual pairs which were conjectured to give equivalent string theories. These dual pairs are then called *mirror manifolds*. Mirror symmetry can also be profitably thought of as providing a *transform*. In other words a difficult problem on one Calabi–Yau manifold may be much easier, but equivalent to, one on its dual. We give an example of a result obtainable this way below; it concerns the number of curves of prescribed degree and genus on a Calabi–Yau manifold.

The term *mirror manifolds* was coined in Greene and Plesser [**1**] (1990). This paper also contains the first, and as yet only, known construction of such dual pairs.

If M and N are Calabi–Yau manifolds of complex dimension n ($n = 3$ in the string theory cited above), and if $h^{(p,q)}(M)$ denote the Hodge numbers[33] of a complex manifold M, then mirror symmetric pairs satisfy

$$h^{(p,q)}(M) = h^{(n-p,q)}(N) \tag{9.2}$$

The term mirror symmetry originates in the fact that this represents a reflection symmetry about the diagonal in the Hodge diamond formed by the $h^{(p,q)}$'s. This reflection property of the Hodge numbers is not sufficient to ensure that M and N are mirror manifolds, one must prove that the associated conformal field theories are also identical, cf. Greene and Plesser [**1-2**] for more information.

The introduction (by Greene, Vafa and Warner [**1**]) of manifolds which are complete intersections in *weighted* projective spaces is an important part of this story, cf. too, Candelas, Lynker and Schimmrigk [**1**]; previously complete intersections had been studied in ordinary projective space, cf., for example, Green, Hübsch and Lütken [**1**] and references therein.

A bold prediction of mirror symmetry (following from the equality of the three point functions on a Calabi–Yau M and its mirror) concerned the numbers n of curves of genus g and degree d on a Calabi—Yau manifold; these numbers could be read off from a rather sophisticated instanton calculation, cf. Candelas, Ossa et al. [**1**] (1991) and the articles in Yau [**1**] for more details. The startling nature of these predictions can be partly appreciated by browsing the following table (this is half of a table appearing in Bershadsky et al. [**1**]). The particular Calabi–Yau manifold is three dimensional and is a certain quotient by $\mathbf{Z}_5 \times \mathbf{Z}_5 \times \mathbf{Z}_5$ of a quintic hypersurface in $\mathbf{CP}^4$ whose equation in homogeneous coordinates is

$$z_1^5 + z_2^5 + \cdots + z_5^5 = 0 \tag{9.3}$$

[33] Hodge numbers are the dimensions of the various Dolbeault cohomology groups of M, i.e. $h^{(p,q)}(M) = \dim H_{\overline{\partial}}^{(p,q)}(M) = \dim H^q(M; \Omega^p)$ where Ω^p is the sheaf of holomorphic p-forms on M.

Degree	$g = 0$	$g = 1$
n=0	*5*	*50/12*
n=1	*2875*	*0*
n=2	*609250*	*0*
n=3	*317206375*	*609250*
n=4	*242467530000*	*3721431625*
n=5	*229305888887625*	*12129909700200*
n=6	*248249742118022000*	*31147299732677250*
n=7	*295091050570845659250*	*71578406022880761750*
n=8	*375632160937476603550000*	*154990541752957846986500*
n=9	*503840510416985243645106250*	*324064464310279585656399500*
$\vdots$	$\vdots$	$\vdots$
large n	$a_0 n^{-3}(\log n)^{-2} e^{2\pi n\alpha}$	$a_1 n^{-1} e^{2\pi n\alpha}$

Table showing numbers of curves of genus g on a quintic hypersurface as predicted by mirror symmetry

Mathematical verifications of these spectacular, physically obtained, numbers were at first only available for small values of the degree. However there has now been a beautiful confirmation of all of them by Givental [**1**] (1996); this work also extends the ideas to non-Calabi–Yau manifolds; a key paper, whose results are used in Givental [**1**], is that of Kontsevich [**1**] (1995).

The solutions methods employed for these problems involving curve counting, or *enumerative geometry*, are closely connected with another development of joint physical and mathematical interest: this is the subject of *quantum cohomology*. Quantum cohomology originates in quantum field theory. One considers a quantum cohomology ring $H^*_q(M)$ which is a natural deformation of the standard cohomology ring $H^*(M)$ of a manifold M. The q in $H^*_q(M)$ is a real parameter which can be taken to zero and, when this is done, one recovers the standard cohomology ring $H^*(M)$; thus $q \to 0$ represents the *classical limit*, one also recognises similarities with the deformations of Lie algebras known as quantum groups. For some physical and mathematical background material on this highly interesting new area cf. Morrison and Plesser [**1**] and Kontsevich and Manin [**1**] and references therein.

Dyons again and the tyranny of dualities

Several new kinds of duality emerged from about 1994 onwards. Their origin can be traced back to the subject of dyons and in particular to a paper of Montonen and Olive [**1**] (1977); on the other hand the present work in the subject is due in great part to the papers of Seiberg and Witten [**1–2**] and the insights they offer into strong coupling problems such as the celebrated conundrum of the mechanism for quark confinement.

Montonen and Olive proposed a duality between a theory of magnetic monopoles and one containing gauge fields. They were motivated in part by a semiclassical analysis of dyons and gauge fields. Let g denote the coupling of the gauge field[34] then a dyon with (magnetic,

[34] Previously we used g to denote magnetic charge, since this conflicts with our present choice we shall go back to Dirac's notation and use μ for magnetic charge.

electric) quantum numbers (n, m) has mass M given by

$$M^2 = V^2 \left(n^2 + \frac{16\pi^2}{g^4} m^2 \right) \tag{9.4}$$

where V is the Higgs vacuum expectation value. If we interchange g with $1/g$, n with m, and V and $4\pi V/g^2$ then M is invariant; there is also an exchange of the gauge group G with $\widehat{G}$: a group with weight lattice dual to that of G.

Montonen and Olive astutely observed this invariance and boldly conjectured, with accompanying reasons, that it led to a duality possessed by a full quantum theory of dyons. They proposed that the two quantum field theories passed between by the interchanges above were really dual: i.e. it was only necessary to calculate one of them to obtain full knowledge of the other. This is a very dramatic and attractive conjecture because the interchange of g with $1/g$ is an exchange of strong and weak coupling: i.e. the (intractable) strong coupling limit of one theory ought to be calculable as the (tractable) weak coupling limit of a theory with appropriately altered spectrum and quantum numbers. Osborn [**1**] (1979) found that an $N = 4$ supersymmetric $SU(2)$ gauge theory was the best candidate for which the conjecture might hold; nevertheless Seiberg and Witten's results are for an $N = 2$ theory.

It transpires that the electric and magnetic charges e, μ and the CP breaking angle θ live more naturally together as the single complex variable

$$e + i\mu = e_0(n + \tau), \qquad \text{where } \tau = \frac{\theta}{2\pi} + \frac{4\pi i}{g^2} \tag{9.5}$$

A general point on the dyon lattice of Fig. 3. is now given by

$$e_0(m\tau + n), \qquad \text{where } m, n \in \mathbf{Z} \tag{9.6}$$

This addition of the angle θ allows the interchange (or $\mathbf{Z}_2$) symmetry of Montonen and Olive to be promoted to a full $SL(2, \mathbf{Z})$ symmetry under which

$$\tau \mapsto \frac{a\tau + b}{c\tau + d}, \qquad \text{with } \begin{cases} a, b, c, d \in \mathbf{Z} \\ ad - bc = 1 \end{cases} \tag{9.7}$$

Now the conventional string viewpoint of the physics is that this $SL(2, \mathbf{Z})$ symmetry, and the mirror symmetry, are only a low energy manifestation of a richer symmetry of the full string theory. There is now considerable successful work in this direction which involves string dualities known by the symbols S, T and U. Two theories which are S dual have the property that their weak and strong coupling limits are equivalent, two theories which are T dual have the property that one compactified on a large volume is equivalent to the other compactified on a small volume, finally U duality corresponds to a theory compactified on a large, or small, volume being equivalent to another at strong, or weak, coupling respectively.

Certain pairs of the five basic superstring theories are thought to be dual to one another in this framework. It is conjectured that these five theories are just different manifestations of a single eleven dimensional theory known as M theory.

Black hole postscript

It has also recently become of interest to study what happens when strings have ends that move on p-dimensional membranes (called Dirichlet branes or D-branes because of the boundary condition imposed at their ends). This has made possible a study of black holes in this string setting. Included in this study is the ability to calculate some of their quantum properties and their entropy (Strominger and Vafa [**1**] (1996)). The topological nature of space-time here makes contact with the non-commutative geometry of Connes [**1**].

§ 10. Concluding remarks

In the twentieth century, some time after the early successes of the new quantum mechanics and relativity, quantum field theory encountered difficult mathematical problems. Efforts to solve these problems led to the birth of the subject of axiomatic quantum field theory. In this approach the main idea was to tackle the formidable problems of quantum field theory head on using the most powerful mathematical tools available; the bulk of these tools being drawn from analysis.

It is now evident that the way forward in these problems is considerably illuminated if, in addition to analysis, one uses differential topology. We have also seen that this inclusion of topology has produced profound results in mathematics as well as physics.

Not since Poincaré, Hilbert and Weyl took an interest in physics has such lavish attention been visited on the physicists by the mathematicians. The story this time is rather different: In the early part of the twentieth century the physicists imported Riemannian geometry for relativity, thereby of course accelerating its rise to be an essential pillar of the body of mathematics; Hilbert spaces were duly digested for quantum mechanics as was the notion of symmetry in its widest possible sense leading to a systematic use of the theory of group representations.

The Riemannian geometry used by physicists in relativity was first used implicitly in a local manner; but it was inevitable that global issues would arise eventually. This, of course, entails topology and so more mathematics has to be learned by the physicist but the singularity theorems of general relativity more than justify the intellectual investment required.

Thus far, then, the gifts, were mainly from the mathematicians to the physicists. For the last quarter of the twentieth century things are rather different, the physicists have been able to give as well as to receive. The Yang–Mills equations have been the source of many new results in three and four dimensional differential geometry and topology.

In sum, the growing interaction between topology and physics has been a very healthy thing for both subjects. Their joint futures look very bright. It seems fitting that we should leave the last word to Dirac [**1**]

> *It seems likely that this process of increasing abstraction will continue in the future and that advance in physics is to be associated with a continual modification and generalisation of the axioms at the base of the mathematics rather than with a logical development of any one mathematical scheme on a fixed foundation.*

References

Aharonov Y. and Bohm D.

1. Significance of electromagnetic potentials in quantum theory, *Phys. Rev.*, **115**, 485–491, (1959).

Alexander J. W.

1. Topological invariants of knots and links, *Trans. Amer. Math. Soc.*, **30**, 275–306, (1928).

Arnold V. I.

1. Normal forms of function of functions close to degenerate critical points, The Weyl groups A_k, D_k, E_k and Lagrangian singularities, *Funct. Anal. Appl.*, **6**, 254–272, (1973).
2. Critical points of functions on manifolds with boundary, the simple Lie groups B_k, C_k and F_4., *Russ. Math. Surv.*, **33**, 99–116, (1978).

Atiyah M. F.

1. *The geometry and physics of knots*, Cambridge University Press, (1990).
2. *Geometry of Yang-Mills Fields*, Accademia Nazionale dei Lincei, (1979).
3. *New invariants of 3 and 4 dimensional manifolds*, Symposium on the mathematical heritage of Hermann Weyl, May 1987, edited by: Wells R. O. , Amer. Math. Soc., (1988).
4. Topological quantum field theories, *Inst. Hautes Études Sci. Publ. Math.* , **68**, 175–186, (1989).

Atiyah M. F. and Bott R.

1. The Yang-Mills equations over Riemann surfaces, *Phil. Trans. Roy. Soc. Lond. A*, **308**, 523–615, (1982).

Atiyah M. F., Hitchin N. J., Drinfeld V. G. and Manin Y. I.

1. Construction of Instantons, *Phys. Lett.*, **65A**, 185–187, (1978).

Atiyah M. F., Hitchin N. J. and Singer I. M.

1. Self-duality in four dimensional Riemannian geometry, *Proc. Roy. Soc. Lond. A.*, **362**, 425–461, (1978).

Atiyah M. F. and Ward R. S.

1. Instantons and Algebraic Geometry, *Commun. Math. Phys.*, **55**, 117–124, (1977).

Balachandran A. P., Daughton A., Gu Z.-C., Sorkin R. D., Marmo G., and Srivastava A. M.

1. Spin-statistics theorems without relativity or field theory, *Int. J. Mod. Phys.*, **A8**, 2993–3044, (1993).

Bar-Natan D.

1. Vassiliev and quantum invariants of braids. The interface of knots and physics, *Proc. Sympos. Appl. Math.*, **51**, 129–144, (1996).

Barrow-Green J.

1. *Poincaré and the three body problem*, American Mathematical Society, (1997).

Baxter R. J.

1. *Exactly solved models in statistical mechanics*, Academic Press, (1982).

Bedford T., Keane M. and Series C. (eds.)

1. *Ergodic theory, symbolic dynamics and hyperbolic spaces*, Oxford University Press, (1991).

Belavin A. A., Polyakov A. M., Schwarz A. S. and Tyupkin Y. S.

1. Pseudoparticle solutions of the Yang-Mills equations, *Phys. Lett.*, **59B**, 85–87, (1975).

Berry M. V.

1. Quantal phase factors accompanying adiabatic changes, *Proc. Roy. Soc. London A*, **392**, 45–57, (1984).

Berry M. V. and Robbins J. M.

1. Indistinguishability for quantum particles: spin, statistics and the geometric phase, *Proc. Roy. Soc. Lond. A*, **453**, 1771–1790, (1997).

Bershadsky M., Cecotti S., Ooguri H. and Vafa C.

1. Kodaira-Spencer theory of gravity and exact results for quantum string amplitudes, *Commun. Math. Phys.*, **165**, 311–428, (1994).

Birkhoff G. D.

1. Proof of Poincaré's geometric theorem, *Trans. Amer. Math. Soc.*, **14**, 14–22, (1913).
2. Proof of the ergodic theorem, *Proc. Nat. Acad. Sci. U. S. A.*, **17**, 656–660, (1931).

Bogomolny E. B.

1. Stability of Classical solutions, *Sov. J. Nucl. Phys.*, **24**, 861–870, (1976).

Bott R.

1. An application of Morse theory to the topology of Lie groups, *Bull. Soc. Math. France*, **84**, 251–281, (1956).
2. The stable homotopy of the classical groups, *Ann. Math.*, **70**, 313–337, (1959).

Brill D. R. and Werner F. G.

1. Significance of electromagnetic potentials in the quantum theory in the interpretation of electron fringe interferometer observations, *Phys. Rev. Lett.*, **4**, 344–347, (1960).

Broer H. W., Hoveijn, Takens F. and van Gils S. A.

1. *Nonlinear dynamical systems and chaos*, Birkhauser, (1995).

Candelas P., Lynker M. and Schimmrigk R.

1. Calabi–Yau manifolds in weighted P_4, *Nucl. Phys.*, **B341**, 383–402, (1990).

Candelas P., de la Ossa X. C., Green P. S. and Parkes L.

1. A pair of Calabi–Yau manifolds as an exactly soluble superconformal theory, *Nucl. Phys.*, **B359**, 21–74, (1991).

Chandrasekhar S.

1. The maximum mass of ideal white dwarfs, *Astrophys. J.*, **74**, 81–82, (1931).
2. The highly collapsed configurations of a stellar mass, *Mon. Not. R. Astron. Soc.*, **95**, 207–225, (1935).

Coleman S.

1. *The uses of instantons*, The whys of subnuclear physics: Erice summer school 1977, edited by: Zichichi A., Plenum Press, (1979).

Connes A.

1. *Noncommutative geometry*, Academic Press, (1994).

Dahan-Dalmédico A.

1. La renaissance des systèmes dynamiques aux Etats-Unis après la deuxième guerre mondiale: l'action de Solomon Lefschetz, *Rend. Circ. Mat. Palermo (2) Suppl. No 34*, , 133–166, (1994).
2. Le difficile héritage de Henri Poincaré en systèmes dynamiques; Henri Poincaré: science et philosophie (Nancy, 1994), *Publ. Henri-Poincaré-Arch.*, , 13–33, (Akademie Verlag, Berlin, 1996).

Dirac P. A. M.

1. Quantised singularities in the electromagnetic field, *Proc. Roy. Soc. Lond.*, **A133**, 60–72, (1931).

Donaldson S. K.

1. An Application of Gauge Theory to Four Dimensional Topology, *J. Diff. Geom.*, **18**, 279–315, (1983).

2. *The Geometry of 4-Manifolds*, Proc. of the International Congress of Mathematicians, Berkeley 1986, edited by: Gleason A. M., Amer. Math. Soc., (1987).
3. Polynomial invariants for smooth four manifolds, *Topology*, **29**, 257–315, (1990).
4. The Seiberg–Witten equations and 4-manifold topology, *Bull. Amer. Math. Soc.*, **33**, 45–70, (1996).

Donaldson S. K. and Kronheimer P. B.

1. *The geometry of four manifolds*, Oxford University Press, (1990).

Ebeling W.

1. An example of two homeomorphic, nondiffeomorphic complete intersection surfaces, *Inventiones Math.*, **99**, 651–654, (1990).

Epple M.

1. Branch points of algebraic functions and the beginnings of modern knot theory , *Historia Math.*, **22**, 371–401, (1995).
2. Orbits of asteroids, a braid and the first link invariant, *Math. Intell.*, , , (to appear, 1998).

Euler L.

1. Solutio problematis ad geometriam situs pertinentis, *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, **8**, 128–140, (1736).

Finklestein D. and Rubinstein J.

1. Connection between spin, statistics and kinks, *J. Math. Physics*, **9**, 1762–1779, (1968).

Floer A.

1. A Relative Morse Index for the Symplectic Action, *Comm. Pure Appl. Math.*, **41**, 393–407, (1988).
2. An Instanton Invariant for 3-Manifolds, *Commun. Math. Phys.*, **118**, 215–240, (1988).

Freedman M. H.

1. The topology of 4-dimensional manifolds, *Jour. Diff. Geom.*, **17**, 357–453, (1982).

Freed D. S. and Uhlenbeck K. K.

1. *Instantons and Four-Manifolds*, Springer-Verlag, (1984).

Freyd P., Yetter D.; Hoste J.; Lickorish W. B. R., Millet K.; and Ocneanu A.

1. A new polynomial invariant of knots and links, *Bull. Amer. Math. Soc.*, **12**, 239–246, (1985).

Froissart M.

1. *Applications of algebraic topology to physics*, Mathematical theory of elementary particles, edited by: Goodman R. and Segal I., M. I. T. Press, (1966).

Gauss C. F.

1. Zur mathematischen theorie der electrodynamischen Wirkungen (1833), *Werke. Königlichen Gesellschaft der Wissenschaften zu Göttingen*, **5**, 605, (1877).

Givental A. B.

1. Equivariant Gromov–Witten invariants, *Internat. Math. Res. Notices*, **13**, 613–663, (1996).

Green M. B. and Schwarz J. H.

1. Anomaly cancellations in supersymmetric $D = 10$ gauge theory and superstring theory, *Phys. Lett.*, **149B**, 117–122, (1984).

Green P., Hübsch T. and Lütken C. A.

1. All the Hodge numbers for all Calabi–Yau complete intersections, *Class. Quant. Grav.*, **6**, 105–124, (1989).

Green M. B., Schwarz J. H. and Witten E.

1. *Superstring Theory vol. 1*, Cambridge University Press, (1987).
2. *Superstring Theory vol. 2*, Cambridge University Press, (1987).

Greene B. R. and Plesser M. R.

1. Duality in Calabi-Yau moduli space, *Nucl. Phys.*, **B338**, 15–37, (1990).
2. *An introduction to mirror manifolds*, Essays on Mirror Manifolds, edited by: Yau S.-T., International Press, Hong Kong, (1992).

Greene B. R., Vafa C. and Warner N. P.

1. Calabi–Yau manifolds and renormalization group flows, *Nucl. Phys.*, **B324**, 371–390, (1989).

Hawking S. W. and Ellis G. F. R.

1. *The large scale structure of space-time*, Cambridge University Press, (1973).

Hawking S. W. and Israel W. (eds.)

1. *Three hundred years of gravitation*, Cambridge University Press, (1987).

Hawking S. W. and Penrose R.

1. The singularities of gravitational collapse and cosmology, *Proc. Roy. Soc. (London)*, **A314**, 529–548, (1970).

Helmholtz H. L. F.

1. Ueber Integrale der hydrodynamischen gleichungen welche den wirbelbewegungen entsprechen, *Jour. für die reine und ang. Math.*, **55**, 25–55, (1858).

Hwa R. C. and Teplitz V. L.

1. *Homology and Feynman integrals*, W. A. Benjamin, (1966).

Jones V. F. R.

1. A Polynomial invariant for knots via Von Neumann algebras, *Bull. Amer. Math. Soc.*, **12**, 103–111, (1985).

Kirchhoff G.

1. Ueber die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanisher Ströme gefürt wird, *Ann. der Physik und Chemie*, **72**, 497–508, (1847).

Kontsevich M.

1. Enumeration of rational curves via torus actions, *Progr. Math.*, **129**, 335–368, (in *The moduli space of curves (Texel Island 1994), Birkhäuser, Boston, MA* 1995).

Kontsevich M. and Manin Y.

1. Gromov-Witten classes, quantum cohomology and enumerative geometry, *Commun. Math. Phys.*, **164**, 525–562, (1994).

Laplace P. S.

1. Proof of the theorem that the attractive force of a heavenly body could be so large that light could not flow out of it, *Geographische Ephemeriden, verfasset von Einer Gesellschaft Gelehrten*, **I**, , (1799).

Lemaître G.

1. L'univers en expansion, *Ann. Soc. Sci. (Bruxelles)*, **A53**, 51–85, (1933).

Libgober A. S. and Wood J. W.

1. Differentiable structures on complete intersections—I, *Topology*, **21**, 469–482, (1982).

Lieb E. H. and Temperley H. N. V.

1. Relations between the percolation and colouring problem and other graph-theoretical problems associated with regular planar lattices: Some exact results for the percolation problem, *Proc. Roy. Soc. Lond.*, **A322**, 251, (1971).

Listing J. B.

1. Der Census räumlicher Complexe oder Veallgemeinerung des Euler'schen Satzes von den Polyedern, *Abhandlungen der königlichen Gesellschaften zu Göttingen*, **10**, 97–180, (1861).
2. Vorstudien zur Topologie, *Göttinger Studien*, , 811–875, (1847).

Maxwell J. C.

1. *A treatise on electricity and magnetism vol. I, (1873) ed. 3 (1904)*, Oxford University Press, (1904).
2. *A treatise on electricity and magnetism vol II, (1873) ed. 3 (1904)*, Oxford University Press, (1904).

Michell J. (Rev.)

1. On the means of discovering the distance, magnitude, etc., of the fixed stars, in consequence of the diminution of their light, in case such a diminution should be found to take place in any of them, and such other data should be procured from observations, as would be further necessary for that purpose, *Phil. Trans. Roy. Soc. (London)*, **74**, 35–57, (1784).

Mickelsson J.

1. Geometry of spin and statistics in classical and quantum mechanics, *Phys. Rev.*, **D3**, 1375–1378, (1984).

Milnor J.

1. On manifolds homeomorphic to the 7-sphere, *Ann. Math.*, **64**, 399–405, (1956).

Montonen C. and Olive D.

1. Magnetic monopoles as gauge particles, *Phys. Lett.*, **B72**, 117–120, (1977).

Morandi G.

1. *Quantum Hall effect. Topological problems in condensed matter physics*, Bibliopolis, (1988).

Morrison D. R. and Plesser R. M.

1. Summing the instantons: quantum cohomology and mirror symmetry in toric varieties, *Nucl. Phys.*, **B440**, 279–354, (1995).

Morse M.

1. *Calculus of variations in the large*, Amer. Math. Soc. Colloq. Publ., (1934).

Nash C.

1. *Differential Topology and Quantum Field Theory*, Academic Press, (1991).
2. A comment on Witten's topological Lagrangian, *Mod. Phy. Lett. A*, **7**, 1953–1958, (1992).

Nielsen H. and Olesen P.

1. Vortex line models for dual strings, *Nucl. Phys.*, **B61**, 45–61, (1973).

Oppenheimer J. R. and Snyder H.

1. On continued gravitational contraction, *Phys. Rev*, **56**, 455–459, (1939).

Osborn H.

1. Topological charges for $N = 4$ supersymmetric gauge theories and monopoles of spin 1, *Phys. Lett.*, **83B**, 321–326, (1979).

Osterwalder K. and Schrader R.

1. Axioms for Euclidean Green's functions, *Commun. Math. Phys.*, **31**, 83–112, (1973).

2. Axioms for Euclidean Green's functions 2, *Commun. Math. Phys.*, **42**, 281–305, (1975).
3. *Springer Lecture notes in physics* **25***: Constructive quantum field theory*, Springer, (1973).

Penrose R.

1. Gravitational collapse and space-time singularities, *Phys. Rev. Lett.*, **14**, 57–59, (1965).

Pham F.

1. *Introduction à l'étude topologique des singularitiés de Landau. Mémorial des Sciences Mathématiques, Fasc. 164*, Gauthier-Villars, (1967).

Poincaré H.

1. Sur le problème de trois corps et les équations de la dynamique, *Acta Mathematica*, **13**, 1–270, (1890).
2. Sur les courbes définies par une équation différentielle, *C. Rend. Acad. Sc.*, **90**, 673–675, (1880).
3. Mémoire sur les courbes définies par une équation différentielle, *J. de Math.*, **7**, 375–422, (1881).
4. Mémoire sur les courbes définies par une équation différentielle, *J. de Math.*, **8**, 251–296, (1882).
5. Sur les courbes définies par les équations différentielle, *J. de Math.*, **1**, 167–244, (1885).
6. Sur les courbes définies par les équations différentielle, *J. de Math.*, **2**, 151–217, (1886).
7. Analyse des travaux scientifiques de Henri Poincaré, faite par lui même, *Acta Math.*, **38**, 1–135, (1921).
8. Analysis situs, *J. Éc. Polyt.*, **1**, 1–121, (1895).
9. Sur l'Analysis situs, *C. R. Acad. Sc.*, **115**, 633–636, (1892).
10. Sur les nombres de Betti, *C. R. Acad. Sc.*, **128**, 629–630, (1899).
11. Complément à l'Analysis situs, *Rend. Circ. Matem. Palermo*, **13**, 285–343, (1899).
12. Second complément à l'Analysis situs, *Proc. Lond. math. Soc.*, **32**, 277–308, (1900).
13. Sur l'Analysis situs, *C. R. Acad. Sc.*, **133**, 707–709, (1901).
14. Sur certaines surfaces algébriques; troisième complément à l'Analysis situs, *Bull. Soc. Math. Fr.*, **30**, 49–70, (1902).
15. Sur la connexion des surfaces algébriques, *C. R. Acad. Sc.*, **133**, 969–973, (1901).
16. Sur les cycles des surfaces algébriques; quatrième complément à l'Analysis situs, *J. Math. pures et appl.*, **8**, 169–214, (1902).
17. Cinquième complément à l'Analysis situs, *Rend. Circ. Matem. Palermo*, **18**, 45–110, (1904).
18. Sur un théorème de géométrie, *Rend. Circ. Matem. Palermo*, **33**, 375–407, (1912).

Polyakov A. M.

1. Spectrum of particles in the quantum field theory, *JETP Lett.*, **20**, 194–195, (1974).

Pont J.-C.

1. *La topologie algébrique des origines à Poincaré*, Presses Universitaires de France, (1974).

Prasad M. K. and Sommerfield C. M.

1. Exact classical solution for the 't Hooft monopole and the Julia-Zee dyon, *Phys. Rev. Lett.*, **35**, 760–762, (1975).

Riemann B.

1. Theorie der Abel'schen Functionen, *Jour. für die reine und ang. Math.*, **54**, , (1857).

Rohlin V. A.

1. New results in the theory of 4 dimensional manifolds, *Dok. Akad. Nauk. U. S. S. R.*, **84**, 221–224, (1952).

Saari D. G. and Xia Z. (eds)

1. *Hamiltonian mechanics and celestial mechanics*, Amer. Math. Soc., (1996).

Schwarzschild K.

1. Über das Gravitationsfeld eines Masses nach der Einsteinschen Theorie, *Sitzungberichte Königlich Preuss. Akad. Wiss., Physik-Math. Kl.*, , 189–196, (1916).
2. Über das Gravitationsfeld einer Kugal aus inkompressibler Flüssigkeit nach der Einsteinschen Theorie, *Sitzungberichte Königlich Preuss. Akad. Wiss., Physik-Math. Kl.*, , 424–434, (1916).

Schwinger J.

1. Sources and magnetic charges, *Phys. Rev.*, **173**, 1536–1544, (1968).

Segal G.

1. *Two dimensional conformal field theories and modular functors*, I. A. M. P. Congress, Swansea, 1988, edited by: Davies I., Simon B. and Truman A., Institute of Physics, (1989).

Seiberg N. and Witten E.

1. Electric-magnetic duality, monopole condensation, and confinement in $N = 2$ supersymmetric Yang–Mills theory, *Nucl. Phys.*, **B426**, 19–52, (1994; Erratum-ibid., **B430**, 485–486, 1994).
2. Monopoles, duality and chiral symmetry breaking in $N = 2$ supersymmetric QCD, *Nucl. Phys.*, **B431**, 484–550, (1994).

Simon B.

1. Holonomy, the quantum adiabatic theorem and Berry's phase, *Phys. Rev. Lett.*, **51**, 2167–2170, (1983).

Smale S.

1. Generalised Poincaré's conjecture in dimensions greater than four, *Ann. Math.*, **74**, 391–406, (1961).
2. Topology and mechanics I, *Inventiones Math.*, **10**, 305–331, (1970).
3. Topology and mechanics II, *Inventiones Math.*, **11**, 45-64, (1970).

Streater R. F.

1. Outline of axiomatic relativistic quantum field theory, *Rep. Prog. Phys.*, **38**, 771–846, (1975).

Strominger A. and Vafa C.

1. Microscopic origin of the Bekenstein–Hawking entropy, *Phys. Lett.*, **B379**, 99–104, (1996).

Sundman K. F.

1. Nouvelles recherche sur le problème des trois corps, *Acta Soc. Sci. Fenn.*, **35**, 1–27, (1909).
2. Recherches sur le problème des trois corps, *Acta Soc. Sci. Fenn.*, **34**, 1–43, (1907).
3. Memoire sur le problème des trois corps, *Acta Mathematica*, **36**, 105–179, (1912).

Tait P. G.

1. *Scientific papers (On knots I, II, III) 273–437*, Cambridge University Press, (1898).

Taubes C. H.

1. A framework for Morse theory for the Yang-Mills functional, *Inventiones Math.*, **94**, 327–402, (1988).
2. Min-Max theory for the Yang-Mills-Higgs Equations, *Commun. Math. Phys.*, **97**, 473–540, (1985).

Thom R.

1. Topological models in biology, *Topology*, **8**, 313–335, (1969).
2. *Stabilité structurelle et morphogenèse*, Benjamin, (1972).

3. *Modèles mathématiques de la morphogenèse*, Acad. Naz. Lincei, Pisa, (1971).

Thomson W. H. (Lord Kelvin)

1. On vortex atoms, *Proc. Roy. Soc. Edin.*, **34**, 15–24, (1867).
2. On vortex motion, *Trans. Roy. Soc. Edin.*, **25**, 217–260, (1869).
3. Vortex statics, *Proc. Roy. Soc. Edin.*, **1875–76 session**, , (1875).
4. *Mathematical and physical papers volume IV*, Cambridge University Press, (1910).
5. On the duties of ether for electricity and magnetism, *Phil. Mag.*, **50**, 305–307, (1900).

’t Hooft G.

1. Renormalization of massless Yang–Mills fields, *Nucl. Phys.*, **B33**, 173–199, (1971).
2. Magnetic monopoles in unified gauge theories, *Nucl. Phys.*, **B79**, 276–284, (1974).
3. Computation of the quantum effects due to a four dimensional pseudoparticle, *Phys. Rev.*, **D14**, 3432–3450, (1976).

’t Hooft G. and Veltman M.

1. Regularization and renormalization of gauge fields, *Nucl. Phys.*, **B44**, 189–213, (1972).

Thorne K. S.

1. *Black holes and time warps: Einstein’s outrageous legacy*, W. W. Norton, (1994).

Tscheuschner R. D.

1. Topological spin-statistics relation in quantum field theory, *Int. Jour. Theor. Phys.*, **28**, 1269–1310, (1989).

Ward R. S.

1. On self-dual gauge fields, *Phys. Lett.*, **61A**, 81–82, (1977).

Wilczek F. and Zee A.

1. Linking numbers, spin and statistics of solitons, *Phys. Rev. Lett.*, **51**, 2250–2252, (1983).

Witten E.

1. Dyons of charge $e\theta/2\pi$, *Phys. Lett.*, **86B**, 283–287, (1979).
2. *Some geometrical applications of quantum field theory*, I. A. M. P. Congress, Swansea, 1988, edited by: Davies I., Simon B. and Truman A., Institute of Physics, (1989).
3. Quantum field theory and the Jones polynomial, *Commun. Math. Phys.*, **121**, 351–400, (1989).
4. Topological quantum field theory, *Commun. Math. Phys.*, **117**, 353–386, (1988).
5. Monopoles and four-manifolds, *Math. Res. Lett.*, **1**, 769–796, (1994).
6. Supersymmetry and Morse theory, *J. Diff. Geom.*, **17**, 661–692, (1982).

Yau S.-T. (ed.)

1. *Essays on Mirror Manifolds*, International Press, Hong Kong, (1992).

Zwanziger D.

1. Exactly soluble nonrelativistic model of particles with both electric and magnetic charge, *Phys. Rev.*, **176**, 1480–1488, (1968).
2. Exactly soluble nonrelativistic model of particles with both electric and magnetic charge, *Phys. Rev.*, **176**, 1480–1488, (1968).