CHAPTER 2

How the Internet Developed

The development of the Internet has been a contingent matter, under constraints of technology, human resources, budgets, politics, the Cold War, business needs, and philosophical fashion. Many histories of the Internet focus on the individuals that made a difference, and, for ease of making the narrative, this book will often succumb to that temptation. However, it is important to understand that most innovations require groups of people working together, and don't appear from nowhere; the role of women in particular tends to get airbrushed from the histories.[1]

THE INTERNET'S PREHISTORY

The Internet's development began as an attempt to improve on batch processing for early stand-alone computers, whose operating systems (OSs) ran a 'batch' of programs in order, so that at any time a single program had control of the machine. This seemed wasteful for many purposes (not least because as computers sped up, the time taken to make the transition from one program to the next in the batch became a higher percentage of computer time), prompting research into parallel running of programs, and ultimately remote access via individual workstations, so that each 'computer' became in effect a very small network centred around a processing hub that did the computational work on demand. From here, the idea of linking computers themselves, using the pre-existing telephone network, was a fairly natural suggestion. J. C. R. Licklider, in a position of some power as director of the Information Processing Techniques Office at the Advanced Research Projects Agency (ARPA) in the US Department of Defense in the early 1960s, was able to fund research into the idea that ultimately one could connect all computers

Four Internets. Kieron O'Hara and Wendy Hall, Oxford University Press. © Oxford University Press 2021.
DOI: 10.1093/oso/9780197523681.003.0002

in such a way, to create the ARPANET network, which eventually went live in 1969, with an initial link between terminals at Doug Engelbart's laboratory at the Stanford Research Institute (now SRI International) and Len Kleinrock's at the University of California, Los Angeles.[2]

Its key technology was *packet switching*, an efficient and robust means for transferring data across a network. The sender initiates it by chopping the data into packets of uniform size, and attaching metadata to the packets about how to reassemble them, and giving their destination in a computer network. The sender pushes the packets into a network of connected computers, which, upon receiving a packet, check its metadata and forward it to another computer nearer its destination. Packets would not necessarily take the same route across the network, and as the network became more complex they would all take different routes. At the destination node, when all the packets were received, the receiving computer would use the metadata to reassemble the original data. The ARPANET connected geographically clustered networks via their host computers (e.g. at a particular university), using the telephone network, and so its ambitions at this stage were as a purely national proof of concept. The Network Information Center (NIC), run by Elizabeth Feinler, was required to keep track of addresses and coordinate ARPANET's operation as it grew.[3]

Packet switching produced a communication network that was robust to attack—remove one or more nodes from the network, and the network would still operate, as long as it remained connected (note that this implies redundancy in the network—there should be several routes from A to B), ensuring that communication could continue even if military strikes took out some communications hubs. Packet switching was not proprietary to the ARPANET. It was also promoted by scientists in the United Kingdom, who were impressed by its efficiency; if particular connections become congested, packets can easily be routed around them, thereby allowing speedy and resilient commercial use of data. Sadly, reliant on the government-run monopoly communications infrastructure, the British suggestion foundered. One interesting aspect of the development of the Internet is the reluctance of telecommunications monopolies such as the UK General Post Office and AT&T to engage with it in its formative years. Britain's telecoms network was split off in 1980, and became a private company in 1984; it is intriguing to imagine how the Internet might have developed differently had a commercial motive been introduced in the United Kingdom at this stage. Europe was finally connected to the ARPANET in 1973.[4]

Using packet switching across a network established three important ideas of how information could best be transferred. First, the nodes would take as little interest as possible in the packets, ignoring their contents entirely and examining only their metadata. Second, when a network joined others, it would commit to forwarding packets appropriately on behalf of other networks, and thereby doing some *pro bono* work. And third, perhaps most important, an *end-to-end* principle was established, whereby the bulk of the data processing happened at the sending and receiving nodes, minimizing the processing required en route. This really was innovative; other commonly used networks, such as the telephone network, do relatively little work at the ends (the telephones, in this case), so that the networks themselves perform at least some vital functions. The end-to-end principle improves efficiency by reducing congestion at bottlenecks.

It is important to remember that other networks flourished alongside the ARPANET, many but not all in the United States. Licklider's global vision was hamstrung by the incompatible protocols of these networks, which therefore led separate and parallel existences. Although the contents of the packets are irrelevant to a packet-switching network, the metadata needs to be understood across the network, interpreted and acted upon consistently. Hence each networked computer needs compatible protocols to express the metadata, and the appropriate actions that follow its interpretation.

THE INTERNET ARRIVES

The work of Vinton Cerf and Robert Kahn for the now-renamed DARPA got around the lack of standards by defining a protocol that would work on heterogeneous networks, thus being compatible with as many other protocols as possible, exploiting the end-to-end property of the networks by ignoring as far as possible their idiosyncrasies. The result of their work was TCP/IP, finalized in 1978, which became the approved ARPANET protocol in 1983, and subsequently was adopted by most of the leading networks. All ARPANET hosts were running TCP/IP by the end of 1983, and that year is often taken as the beginning of the Internet, when TCP/IP became the means to connect ARPANET with other packet-switching networks such as the Packet Radio Network and the Packet Satellite Network (although DARPA continued to fund the creation of TCP/IP implementations of important OSs such as UNIX, to the tune of tens of millions of dollars, so we shouldn't think of the Internet even at that stage as a done deal).[5]

The beauty of TCP/IP is thus that it connects networks without constraining them, so it acts as a *lingua franca*, which we take as the defining aspect of what we call 'the Internet'. The network is not controlled by anyone, so any network or computer able to follow TCP/IP could join; in other words, it is a *permissionless* system. By the end of the 1980s, Europe was firmly integrated, with the help of research groups such as that of Cerf's collaborator Peter Kirstein of University College, London, which connected to the ARPANET in 1973 and adopted TCP/IP early, in 1982.[6] The ARPANET was decommissioned in 1990, leaving the Internet as a worldwide, self-standing network of networks. Cerf and Kahn anticipated the importance of their innovation with a newly minted term: 'Internet', signifying the *Internetwork* as a network of networks, and *internetworking* as the practice of joining networks together.[7]

The US government still played an important part, as the National Science Foundation's NSFNET acted as the skeleton of the newly minted Internet (other important backbones sponsored by the government included the Department of Energy's ESNET and NASA's NSINET).[8] The voluntary open standards defining the Internet, particularly TCP/IP, were developed by the Internet Engineering Task Force (IETF), also originally set up by the US government in 1986. Government involvement made commercial innovation difficult, for both legal and political reasons, and so the NSF pushed the Internet into the private sector (the IETF was handed over to the non-profit Internet Society [ISOC] in 1993). Privatization was a key decision, and the future of the Internet was influenced by the way it happened. It might simply have been transferred to a major company like AT&T or IBM, but the decision was taken to open it up to what became known as *Internet Service Providers* (ISPs).[9]

The role of ISPs in addressing illustrates the delicate balancing acts in Internet governance, and the way that technical issues require institutional implementation as well as technological solutions. For example, let's consider how data gets to its final destination. The main identifier introduced by IP is the numerical IP address, which specifies a single—but not necessarily permanent—destination for data; most addresses are assigned dynamically to computers, so they are more akin to a piece of rental property than a permanent address. A single IP address is typically occupied by different computers over time, but by only a single one at a time, and hence is always unique. The assignment is done by *registries* that allocate them in batches to ISPs, and by the ISPs that go on to assign smaller groups to networks of computers which have paid for the service. Dynamic addressing leads to greater efficiency, preventing a scarcity of available addresses while many stand vacant. You only use as many addresses as needed at any particular time.

It also sets a few puzzles. Firstly, the management of IP addresses by senders and receivers of data in the network is only straightforward as long as they stay constant in between a request for information and its receipt. When combined with the mechanism of packet switching, a coordination problem will arise, as sending a packet through the network depends on knowing where all the addresses are, and crucially that they haven't changed, during its journey. Imagine posting a letter without knowing for sure that your correspondent will remain at the address long enough to receive it. The Internet gets round this by assigning part of the address to a sub-network, and having devices called routers which use these for deciding where to send packets. Routers and address registries must constantly inform each other about changes to assignments, so that means a lot of extra information having to flow around the network to maintain packet-switching schedules, so that the 'dumb' network with its end-to-end principle has to do a bit of work after all, keeping track of which computer is referred to by which IP address.

Secondly, the ability of the network to do this constrains its ability to scale. One aim of the decentralized Internet architecture is to allow new people to join the network easily, but there is an obvious problem if there aren't enough IP addresses to go round. This was an issue with earlier versions, and a difficult (and not yet complete) upgrade was required from Internet Protocol version 4 (IPv4), whose addresses are 32 bits in length (i.e. there are 2^{32} of them, about four and a quarter billion), to IPv6, 128 bits long, which became a draft IETF standard as long ago as 1998. We will return to this issue later.

A third issue is that if the amount of information that routers have to process for packet switching increases too quickly, they will not be able to plan routes efficiently enough. Technologically, this problem is under control, but the cost of the solution is that organizations can't take their blocks of IP addresses with them if they change their ISPs, and unused IP address blocks can't be traded. These mean that addressing can't be as efficient as it could be, as the cost of efficient addressing would be inefficiency of packet switching.[10]

ISPs jointly cooperate on a gateway to exchange data between them, to give users a seamless experience of the Internet—a system known as *peering*, located at physical structures known as *Internet Exchange Points*, or IXPs. The exchange of data at IXPs is largely free, their upkeep paid for by charging the ISPs for access, which pay from the revenue they receive from their customers. This turns out to be essential for the free movement of data round the network; had the IXPs demanded payment per bit of data exchanged, that would have had the effect of putting a meter on the dataflow, and ISPs would have tried to save money by inhibiting the flow of data through IXPs. Furthermore, routing decisions for data would have had to take the number of exchanges into account, making them more complex, and requiring even more intelligence in the network. As peering is effectively charged at a flat rate, there are no economic disincentives for rich dataflow.[11]

The story is not exclusively technical. The early steps of the Internet were taken in response to uses and demands of relatively small but influential groups of people. Most obviously, engineers in universities, companies, and government agencies shaped the network, but not only through the designs they created, protocols they wrote, and funding decisions they made. The Internet was woven into this research culture at an early stage, and the researchers used it to send messages, converse, and pass on insights, papers, and code, and this usage also informed the design; researchers designed the network to do what they wanted it to do, which is what they were already using it to do. Bulletin Boards allowed like-minded groups, interested in topics such as *Star Trek*, Buddhism, or feminism, as well as the mechanics of programming themselves, to get together, chat, and play. Early businesses also chipped in.[12]

Thus the medium became a space where a reasonably *simpático* set of well-behaved and culturally aligned people congregated, even if separated by geography, and the Internet gradually came to signify both a technically linked net of machines and a socially linked network of people. Such people collaborated in designing and developing the new space, and a principle of consensus emerged, exemplified by the RFC (Request For Comments), which would be placed in the documentation for a new program or specification, inviting positive and negative feedback. An RFC is not a formal constraint, but invites discussion, amendment, and improvement, and is a means of demonstrating consensus among a wide group. And because they were like-minded, culturally similar, often government-funded, and generally acting in good faith, security and censorship hardly seemed pressing (John Gilmore, one of the founders of the Electronic Frontier Foundation, pointed out in 1993 that 'the Net interprets censorship as damage and routes around it'.).[13] Bad behaviour took relatively asinine and apparently harmless forms, such as flaming, where a discussion would degenerate into argument and insult, often written in capital letters with lots of exclamation marks. But as the Internet grew, new populations supplanted the small group of engineers, behaviour deteriorated, and bad faith (or simply different interests) needed to be factored in. With commerce would come cybercrime; with e-government would come spying and cyberwarfare.[14]

The Internet is often conceived using immaterial metaphors—'cloud', 'cyberspace', 'Web', 'virtual'—or liminal ones—'surfing', 'superhighway', 'portal'. But it is very much a real-world phenomenon with important political, social, and economic effects. We can sketch its basic principles, and see how they led to the extraordinary technology in place today. In this section we have looked at some technical matters, some institutional matters, and the social factors involved in its creation. The Internet Architecture Board Network Working Group asserted in 1996 that 'the community believes that the goal is connectivity, the tool is the Internet Protocol, and the intelligence is end to end rather than hidden in the network'.[15] We might add the importance of decentralization, the ease of joining the network, and the technology of packet switching.

APPLICATIONS ON THE INTERNET

Because no permission is needed, it is relatively straightforward to create an application to which access is gained via the Internet; the application *sits on* the Internet, as it were, using it to transfer data. Some applications have grown to such size and pervasiveness that they are often mistaken for the Internet, and for many might be the only access to it. Examples include the World Wide Web and Facebook; the former was the main vector for the massive growth of the Internet in the early 1990s, while the latter certainly would like to be the gateway to the Internet in parts of the developing world (see Chapter 9). Yet we mustn't confuse the application with the delivery method, like confusing the train with the track. The Internet delivers the application, and its lightweight design has meant that it has been able to scale up, and deliver ever more data to ever more users, continuing to support these applications, even while it grows. The current scale of the Internet must be beyond the wildest of dreams of all but the most visionary of its pioneers.

The Web was the brainchild of physicist Tim Berners-Lee, who wanted an application to enable scientists to see each other's papers and data and collaborate, using the Internet protocols. The basic technology, *hypertext*, had already been widely theorized by thinkers such as Ted Nelson; a non-linear conception of text whereby links in the text could take the reader to an arbitrary text or graphic elsewhere in the document, in a different version of the document, or even in a different document altogether.[16] Berners-Lee's idea was not only to put the hypertext system on the Internet to facilitate access and use, but also to make it an open, permissionless, system describable with open and transparent standards, and requiring neither licence nor fee to exploit the intellectual property of the protocols. The key protocols were the Hypertext Transfer Protocol (HTTP) that governs how hypertext documents get transferred across the Web, the Hypertext Markup Language (HTML), which describes the structure and layout of those documents, and the notion of a Uniform Resource Identifier (URI) or Locator (URL), that identifies a Web resource unambiguously, and also gives an indication of where the resource is held (commonly called a Web address).[17] A link created in one document to a second document utilizes the Web address of the second document, and clicking on the blue hyperlink in the first causes the Web browser to resolve (seek out) the destination address, and to present the resource it found there.

Ultimately, the Web evolved into a giant platform on which many further applications have been created—its openness supporting the spectacular growth of networks that would be inhibited in a closed or proprietary system. Its open standards are coordinated by a non-profit organization called the World Wide Web Consortium (W3C).

The growth of the Web was facilitated not only by W3C standards-setters, but also by people innovating on it, using W3C standards, to increase its utility. Mosaic was the first browser to display the hypertext documents on the Web as single entities, including the graphics, and so became the first commonly used access point. A browser turns the Web into a seamless experience, as if all the documents it displays were stored on one's own hard drive. Mosaic facilitated the growth of the Web, and inspired other browsers, including Netscape Navigator, Internet Explorer, Google Chrome, Firefox, and Safari.

The growth of the Web resulted in a wider user base, outgrowing the culture of academics and scientists sharing static documents, exposing its limitations.

Users wanted to personalize Web resources, conduct conversations, interact commercially (e-commerce), remember where they had been online, collaborate with others. So-called Web 2.0 arrived, with the development of *HTTPS* (a securely encrypted version of HTTP), *cookies* (which allowed websites to store current and previous states of their interactions with users, *JavaScript* (a programming language for webpages), browsers which were adaptable by *plugins* (for example, introducing audio and video), *RSS* (which kept readers updated about changes to websites), *APIs* (application programming interfaces, which allowed other websites and organizations access to a site's data), *wikis* (sites whose content could be modified collaboratively by users), and so on. None of these, and certainly not their combination, was planned from the top down; they may have been developed by specialized standards groups, but they were all responses to demand. Nevertheless, these technologies opened up the Web to innovations such as e-commerce, blogs, and complex games, and none of them required permission to develop; Web technology patent rates soared.[18] Web 2.0 is a heterogeneous set of tools and ideas, but between them they created a read-write Web sitting on top of a read-only Web 1.0 of static documents, which itself sits on top of the Internet transport network.[19]

Openness meant that collective intelligence could be realized by the artful aggregation of data, like a market arriving at a price through the balance of supply and demand. Peer review systems in e-commerce, such as eBay's seller ratings and Amazon's product reviews, allowed the bootstrapping of reputational trust for would-be purchasers, while other sites enabled the aggregation of lots of small chunks of information into, say, an encyclopaedia (such as Wikipedia) or a database of movie credits and reviews (such as IMDb, the Internet Movie Database).[20] Most spectacularly, Google's PageRank algorithm aggregated the information about which documents are linked on the Web. No single link creator need consider anything more than where it would be helpful to send a reader, but aggregation of all these decisions by PageRank resulted in the Google search engine, which became the market leader (partly through its brilliant algorithm, and partly because of the large quantity of data it generates about search, thanks to its popularity).[21]

One more development of the Web has been a move from linking documents (rendered as webpages), to linking data, via languages such as the Resource Description Framework (RDF), which uses URIs to describe and link objects and the relationships between them. In its more ambitious form, this ideal was called the *Semantic Web*, while the less ambitious desire merely to create links between pieces of data gives us the *Linked Data Web*.[22] The original Web linked entire documents, but the original vision was always meant to be extended to connect multimedia data, with 'the evolution of objects from being principally human-readable documents to contain more machine-oriented semantic information, allowing more sophisticated processing'.[23] The Linked Data Web now allows individual pieces of data to be linked, so, for example, it is possible to automate the extraction of all the data about a particular object (via its URI) expressed in RDF from a large series of documents (or, put another way, RDF enables someone to turn data distributed over the Web into a structure analogous to a single spreadsheet). This concept enables a number of important applications, such as Google's *knowledge graph*, a giant database of entities, concepts, and links that populates the 'infoboxes' positioned to the right of Google search pages.

Another important application is the cloud. Cloud computing is the pooling of data storage and processing capabilities, maintained by cloud providers such as Apple, Microsoft, and Amazon, accessible to users from any Internet-connected device. Data centres make data available to users as they need it; the data might be private data (such as photographs or audio files), software, or computing infrastructure. The arrangement means that programs and data storage no longer represent up-front costs for a user, but are delivered as services, allowing economies of scale while users pay only for the computing power they use. The cloud is made possible by high-speed, high-reliability networks, and can be public, private, commercial, or in-house (so one organization can run a cloud for all its computing needs). It raises interesting issues of jurisdiction (the data centres which make up the cloud hold data in a particular nation—does its government get access to the data?), security (how should the data be protected?), privacy (what data is legal to keep?), and liability (who is responsible if it goes missing?).

The cloud was in turn important for the revolution of mobile computing via smartphones. Mobile devices, including larger tablet computers, as well as the sensors that are crucial for the so-called Internet of Things, have the potential to attract billions more to the Internet, as they do not require the hefty investment in hardware or reliable electricity supplies needed for desktop computing. Mobile devices can access programs and data (input and output) in a cloud when needed, giving them impressive functionality, despite being small with limited power supplies.

All of these developments have cumulatively made the way for what for many is the ultimate application of the Internet. *Social media*, linked in particular to social networking sites (SNSs) such as Facebook[24] and Instagram,[25] as well as former SNSs such as Friendster and MySpace, and special-purpose ones like LinkedIn, have turned out to be extremely important applications. They allow users to construct profiles of themselves which act as points of contact, avatars of the real-world individual, to connect their profiles with those of others, and to communicate, interact, and share (often user-generated) content.[26] 'Classic' SNSs share features with other types of sites, such as gaming, microblogging, and media-sharing sites, where networks of contacts and friends also help determine the user experience.[27] The graphs of these social networks are valuable to users, to SNSs, and to third parties (such as advertisers), as are the data trails that interactions leave behind.[28]