From the book *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*.
By David Easley and Jon Kleinberg. Cambridge University Press, 2010.
Complete preprint on-line at http://www.cs.cornell.edu/home/kleinber/networks-book/

Chapter 13

The Structure of the Web

Up to this point in the book, we've considered networks in which the basic units being connected were people or other social entities, like firms or organizations. The links connecting them have generally corresponded to opportunities for some kind of social or economic interaction.

In the next several chapters, we consider a different type of network, in which the basic units being connected are pieces of information, and links join pieces of information that are related to each other in some fashion. We will call such a network an *information network*. As we will see, the World Wide Web is arguably the most prominent current example of such a network, and while the use of information networks has a long history, it was really the growth of the Web that brought such networks to wide public awareness.

While there are basic differences between information networks and the kinds of social and economic networks that we've discussed earlier, many of the central ideas developed earlier in the book will turn out to be fundamental here as well: we'll be using the same basic ideas from graph theory, including short paths and giant components; formulating notions of power in terms of the underlying graph structure; and even drawing connections to matching markets when we consider some of the ways in which search companies on the Web have designed their businesses.

Because the Web plays such a central role in the modern version of this topic, we begin with some context about the Web, and then look further back into the history of information networks that led up to the Web.

Draft version: June 10, 2010

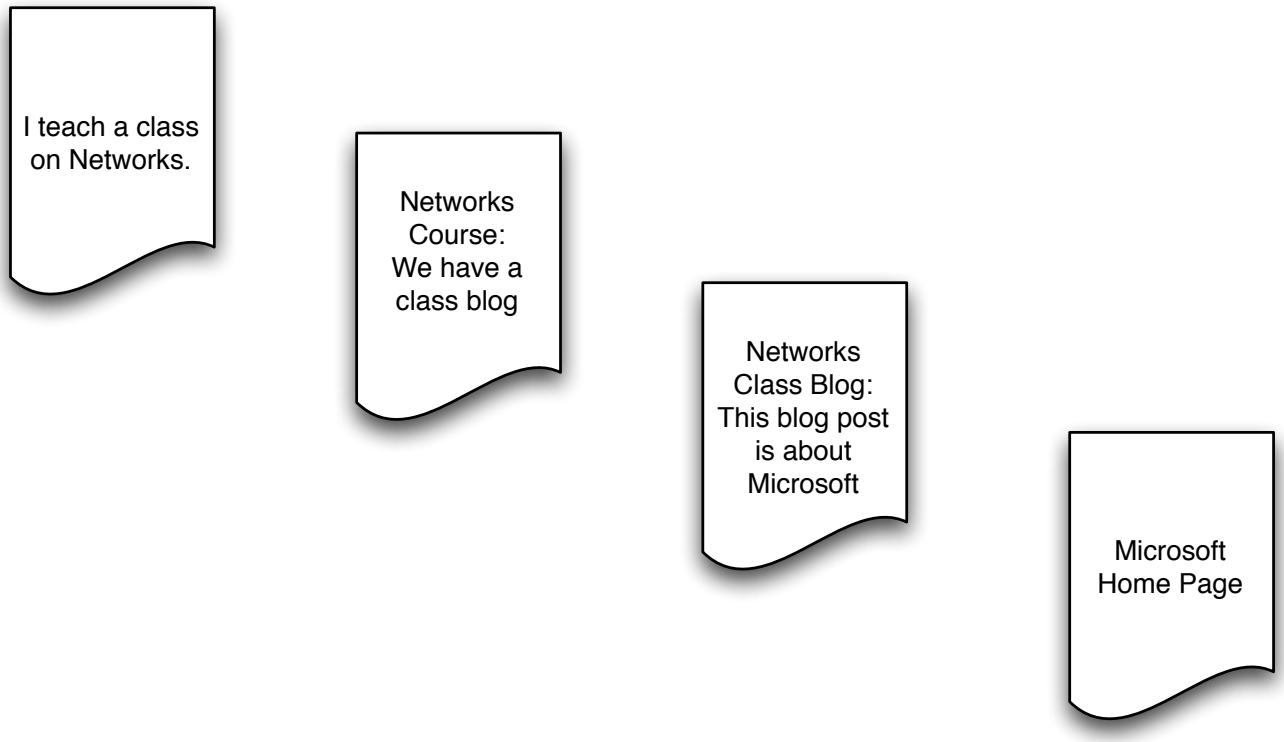

Figure 13.1: A set of four Web pages.

13.1 The World Wide Web

If you're reading this book, it's likely that you use the Web on a daily basis. But since the Web is so enmeshed in the broader information infrastructure of the world (including the Internet, wireless communication systems, and the global media industry), it's actually useful to think a bit about what the Web is and how it came about, starting from first principles.

At a basic level, the Web is an application developed to let people share information over the Internet; it was created by Tim Berners-Lee during the period 1989-1991 [54, 55]. Although it is a simplification, we can view the original conception and design of the Web as involving two central features. First, it provided a way for you to make documents easily available to anyone on the Internet, in the form of *Web pages* that you could create and store on a publically accessible part of your computer. Second, it provided a way for others to easily access such Web pages, using a *browser* that could connect to the public spaces on computers across the Internet and retrieve the Web pages stored there.

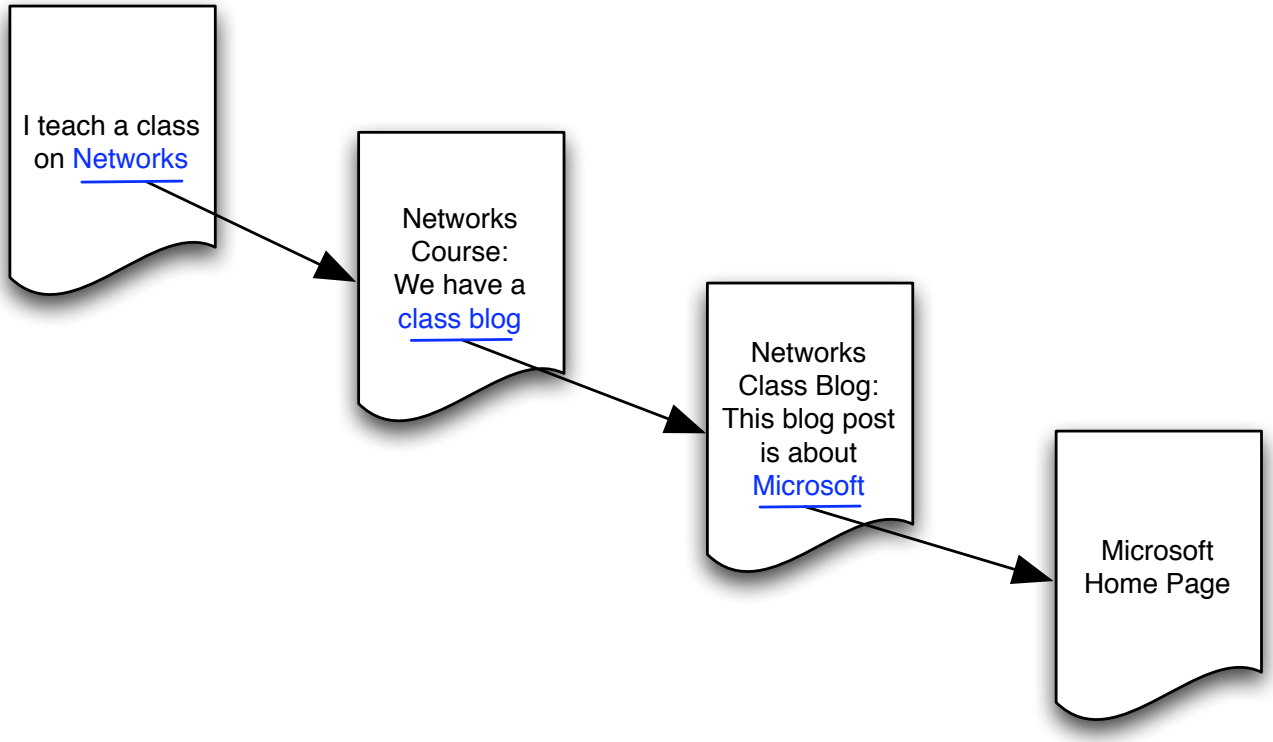

Figure 13.2: Information on the Web is organized using a network metaphor: The links among Web pages turn the Web into a directed graph.

To a first approximation, this is still how we experience the Web today: as a sequence of Web pages rendered inside a browser. For example, Figure 13.1 shows a set of four separate Web pages: the home page of a college instructor who teaches a class on networks; the home page of the networks class he teaches; the blog for the class, with a post about Microsoft listed at the top; and the corporate home page for Microsoft. Because of the underlying design, we can think of these pages both as part of a single coherent system (the Web), but also as files that likely reside on four separate computers, controlled by several different and completely independent organizations, and made publically accessible by a now-universal consensus to participate in the protocols of the Web.

Hypertext. Beyond these basic features, there is a crucial design principle embedded in the Web — the decision to organize the information using a network metaphor. This is what turns the set of Web pages from Figure 13.1 into the "web" of Web pages in Figure 13.2: in writing a Web page, you can annotate any portion of the document with a virtual link to

another Web page, allowing a reader to move directly from your page to this other one. The set of pages on the Web thereby becomes a graph, and in fact a directed graph: the nodes are the pages themselves, and the directed edges are the links that lead from one page to another.

Much as we're familiar with the idea of links among Web pages, we should appreciate that the idea to organize Web pages as a network was both inspired and non-obvious. There are many ways to arrange information — according to a classification system, like books in a library; as a series of folders, like the files on your computer; even purely alphabetically, like the terms in an index or the names in a phone directory. Each of these organizational systems can make sense in different contexts, and any of them could in principle have been used for the Web. But the use of a network structure truly brings forth the globalizing power of the Web by allowing anyone authoring a Web page to highlight a relationship with any other existing page, anywhere in the world.

The decision to use this network metaphor also didn't arise out of thin air; it's an application of a computer-assisted style of authoring known as *hypertext* that had been explored and refined since the middle of the twentieth century [316, 324]. The motivating idea behind hypertext is to replace the traditional linear structure of text with a network structure, in which any portion of the text can link directly to any other part — in this way, logical relationships within the text that are traditionally implicit become first-class objects, foregrounded by the use of explicit links. In its early years, hypertext was a cause passionately advocated by a relatively small group of technologists; the Web subsequently brought hypertext to a global audience, at a scale that no one could have anticipated.

13.2 Information Networks, Hypertext, and Associative Memory

The hypertextual structure of the Web provides us with a familiar and important example of an information network — nodes (Web pages in this case) containing information, with explicit links encoding relationships between the nodes. But the notion of an information network significantly predates the development of computer technology, and the creators of hypertext were in their own right motivated by earlier networks that wove together large amounts of information.

Intellectual Precursors of Hypertext. A first important intellectual precursor of hypertext is the concept of *citation* among scholarly books and articles. When the author or authors of a scholarly work wish to credit the source of an idea they are invoking, they include a citation to the earlier paper that provides the source of this idea. For example, Figure 13.3 shows the citations among a set of sociology papers that provided some of the

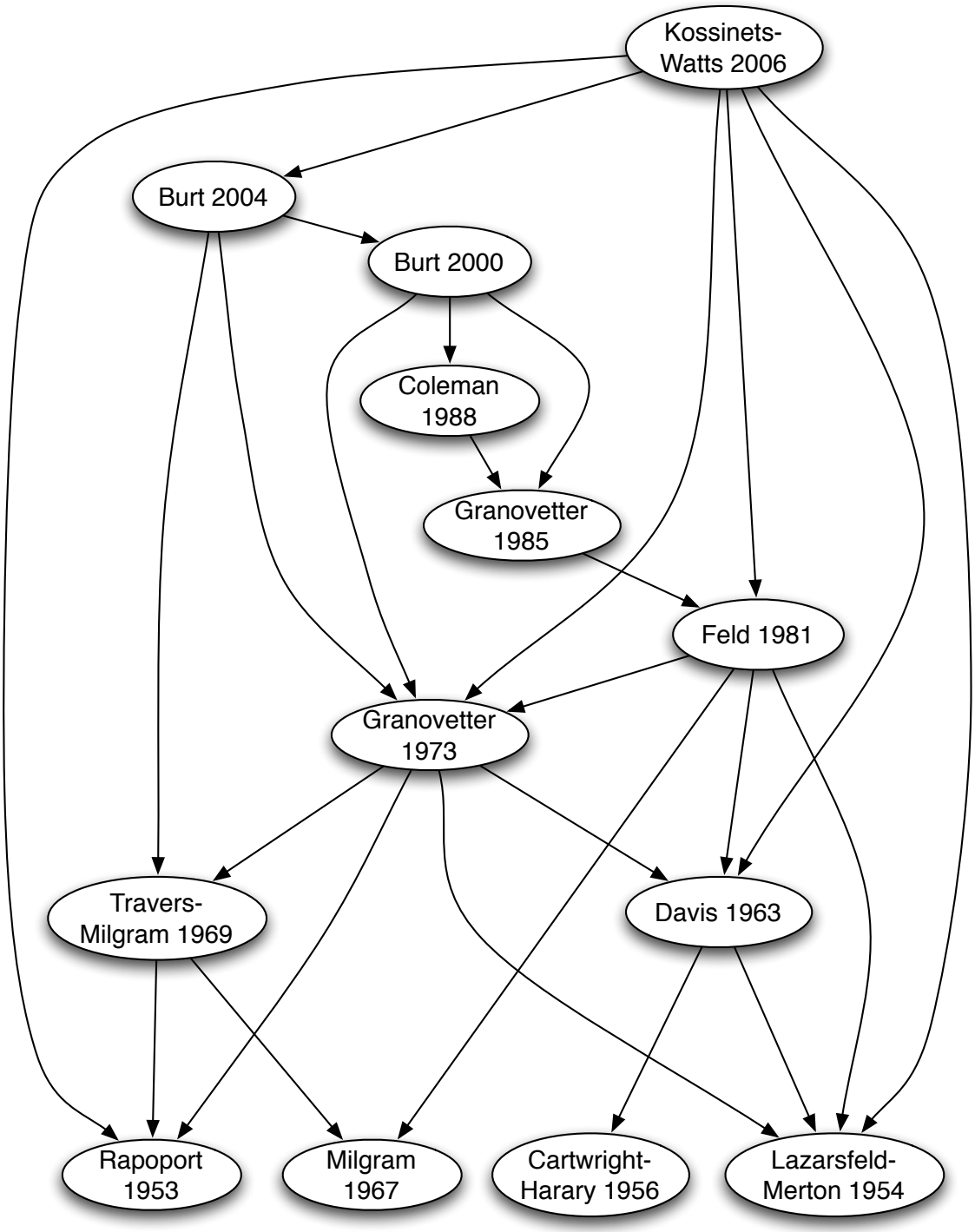

Figure 13.3: The network of citations among a set of research papers forms a directed graph that, like the Web, is a kind of information network. In contrast to the Web, however, the passage of time is much more evident in citation networks, since their links tend to point strictly backward in time.

key ideas in the first part of this book. (At the bottom of this figure are seminal papers on — from left to right — triadic closure, the small-world phenomenon, structural balance, and homophily.) We can see how work in this field — as in any academic discipline — builds on earlier work, with the dependence represented by a citation structure. We can also see how this citation structure naturally forms a directed graph, with nodes representing books and articles, and directed edges representing citations from one work to another. The same structure arises among patents, which provide citations to prior work and earlier inventions; and among legal decisions, which provide citations to earlier decisions that are being used as precedents, or are being distinguished from the present case. Of course, the example in Figure 13.3 is a tiny piece of a much larger directed graph; for instance, Mark Granovetter's 1973 paper on the strength of weak ties has been cited several thousand times in the academic literature, so in the full citation structure we should imagine thousands of arrows all pointing to this single node.

One distinction between citation networks and the Web is that citations are governed much more strongly by an underlying "arrow of time." A book, article, patent, or legal decision is written at a specific point in time, and the citations it contains — the edges pointing outward to other nodes — are effectively "frozen" at the point when it is written. In other words, citations lead back into the past: if paper X cites paper Y, then we generally won't find a citation from Y back to X for the simple reason that Y was written at a time before X existed. Of course, there are exceptions to this principle — two papers that were written concurrently, with each citing the other; or a work that is revised to include more recent citations — but this flow backward in time is a dominant pattern in citation networks. On the Web, in contrast, while some pages are written once and then frozen forever, a significant portion of them are evolving works in progress where the links are updated over long periods of time. This means that while links are directed, there is no strong sense of "flow" from the present into the past.

Citation networks are not the only earlier form of information network. The cross-references within a printed encyclopedia or similar reference work form another important example; one article will often include pointers to other related articles. An on-line reference work like Wikipedia (even when viewed simply as a collection of linked articles, independent of the fact that it exists on the Web) is structured in the same way. This organizing principle is a clear precursor of hypertext, in that the cross-referencing links make relationships among the articles explicit. It is possible to browse a printed or on-line encyclopedia through its cross-references, pursuing serendipitious leads from one topic to another.

For example, Figure 13.4 shows the cross-references among Wikipedia articles on certain topics in game theory, together with connections to related topics.[1] We can see, for example,

[1]Since Wikipedia changes constantly, Figure 13.4 necessarily represents the state of the links among these articles only at the time of this writing. The need to stress this point reinforces the contrast with the "frozen" nature of the citations in a collection of papers such as those in Figure 13.3.

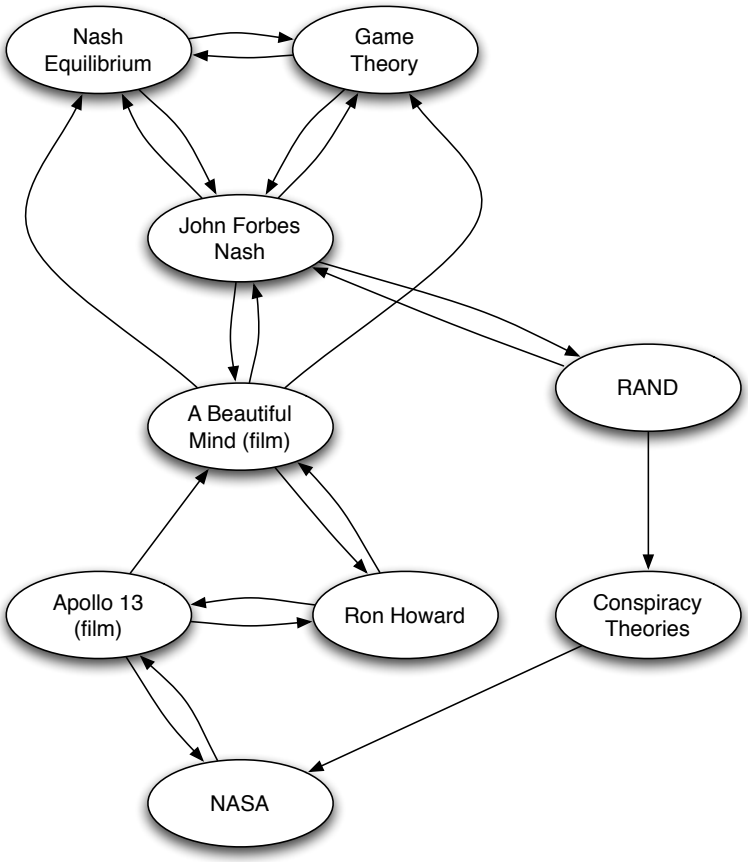

Figure 13.4: The cross-references among a set of articles in an encyclopedia form another kind of information network that can be represented as a directed graph. The figure shows the cross-references among a set of Wikipedia articles on topics in game theory, and their connections to related topics including popular culture and government agencies.

how it's possible to get from the article on Nash Equilibrium to the article on NASA (the U.S. National Aeronautics and Space Administration) by passing through articles on John Nash (the creator of Nash equilibrium), *A Beautiful Mind* (a film about John Nash's life), Ron Howard (the director of *A Beautiful Mind*), *Apollo 13* (another film directed by Ron Howard), and finally on to the article about NASA (the U.S. government agency that managed the real Apollo 13 space mission). In short: Nash equilibrium was created by someone whose life was the subject of a movie by a director who also made a movie about NASA. Nor is this the only short chain of articles from Nash equilibrium to NASA. Figure 13.4 also contains a sequence of cross-references based on the fact that John Nash worked for a period of time at RAND, and RAND is the subject of several conspiracy theories, as is NASA. These short paths between seemingly distant concepts reflect an analogue, for information

networks, of the "six degrees of separation" phenomenon in social networks from Chapter 2, where similarly short paths link apparently distant pairs of people.

Indeed, browsing through chains of cross-references is closely related to the stream-of-consciousness way in which one mentally free-associates between different ideas. For example, suppose you've just been reading about Nash equilibrium in a book, and while thinking about it during a walk home your mind wanders, and you suddenly notice that you've shifted to thinking about NASA. It may take a bit of reflection to figure out how this happened, and to reconstruct a chain of free-association like the one pictured in Figure 13.4, carried out entirely among the existing associations in your mind. This idea has been formalized in another kind of information network: a *semantic network*, in which nodes literally represent concepts, and edges represent some kind of logical or perceived relationship between the concepts. Researchers have used techniques like *word association studies* (e.g. "Tell me what you think of when I say the word 'cold' ") as a way to probe the otherwise implicit structure of semantic networks as they exist in people's minds [381].

Vannevar Bush and the Memex. Thus, information networks date back into much earlier periods in our history; for centuries, they were associated with libraries and scholarly literature, rather than with computer technology and the Internet. The idea that they could assume a strongly technological incarnation, in the form of something like the Web, is generally credited to Vannevar Bush and his seminal 1945 article in the *Atlantic Monthly*, entitled "As We May Think" [89]. Written at the end of World War II, it imagined with eerie prescience the ways in which nascent computing and communication technology might revolutionize the ways we store, exchange, and access information.

In particular, Bush observed that traditional methods for storing information in a book, a library, or a computer memory are highly *linear* — they consist of a collection of items sorted in some sequential order. Our conscious experience of thinking, on the other hand, exhibits what might be called an *associative memory*, the kind that a semantic network represents — you think of one thing; it reminds you of another; you see a novel connection; some new insight is formed. Bush therefore called for the creation of information systems that mimicked this style of memory; he imagined a hypothetical prototype called the *Memex* that functioned very much like the Web, consisting of digitized versions of all human knowledge connected by associative links, and he imagined a range of commercial applications and knowledge-sharing activities that could take place around such a device. In this way, Bush's article foreshadowed not only the Web itself, but also many of the dominant metaphors that are now used to think about the Web: the Web as universal encyclopedia; the Web as giant socio-economic system; the Web as global brain.

The fact that Vannever Bush's vision was so accurate is not in any sense coincidental; Bush occupied a prominent position in the U.S. government's scientific funding establish-

ment, and his ideas about future directions had considerable reach. Indeed, the creators of early hypertext systems explicitly invoked Bush's ideas, as did Tim Berners-Lee when he set out to develop the Web.

The Web and its Evolution. This brings us back to the 1990s, the first decade of the Web, in which it grew rapidly from a modest research project to a vast new medium with global reach. In the early phase of this period, the simple picture in Figure 13.2 captured the Web's essential nature: most pages were relatively static documents, and most links served primarily *navigational* functions — to transport you from one page to another, according to the relational premise of hypertext.

This is still a reasonable working approximation for large portions of the Web, but the Web has also increasingly outgrown the simple model of documents connected by navigational links, and it is important to understand how this has happened in order to be able to interpret any analysis of the Web's structure. In the earliest days of the Web, the computers hosting the content played a relatively passive role: they mainly just served up pages in response to requests for them. Now, on the other hand, the powerful computation available at the other end of a link is often brought more directly into play: links now often trigger complex programs on the computer hosting the page. Links with labels like "Add to Shopping Cart," "Submit my Query," "Update my Calendar," or "Upload my Image," are not intended by their authors primarily to transport you to a new page (though they may do that incidentally as part of their function) — such links exist to activate computational transactions on the machine that runs the site. Here's an example to make this concrete. If we continued following links from the Microsoft Home Page in the example from Figure 13.2, we could imagine taking a next step to the on-line shopping site that Microsoft hosts for its products. From this page, clicking on a link labeled "Buy Now" next to one of the featured products would result in a charge to your credit card and the delivery of the product to your home in the physical, off-line world. There would also be a new page providing a receipt, but the purpose of this last "Buy Now" link was not primarily to transport you, hypertextually, to a "receipt page"; rather, it was to perform the indicated transaction.

In view of these considerations, it is useful to think of a coarse division of links on the Web into *navigational* and *transactional*, with the former serving the traditional hypertextual functions of the Web and the latter primarily existing to perform transactions on the computers hosting the content. This is not a perfect or clear-cut distinction, since many links on the Web have both navigational and transactional functions, but it is a useful dichotomy to keep in mind when evaluating the function of the Web's pages and links.

While a lot of content on the Web now has a primarily transactional nature, this content still remains largely linked together by a navigational "backbone" — it is reachable via relatively stable Web pages connected to each other by more traditional navigational links.

This is the portion of the Web we will focus on in our analysis of its global structure. Sorting out what should belong to this navigational backbone and what shouldn't is ultimately a type of judgment call, but fortunately there is a lot of experience in making and even codifying such judgments. This is because distinguishing between navigational and transactional links has long been essential to Web search engines, when they build their indexes of the available content on the Web. It's clearly not in a search engine's interest to index, for the general public, every receipt from an on-line purchase that every user of the Web has ever made, or every query result for available airline flight times or product specifications that every Web user ever has made. As a result, search engines have developed and refined automated rules that try to assess whether the content they are collecting is relatively stable and intended for public consumption, and they tend to collect content that is reachable via navigational links. We will implicitly be following such working definitions when we talk about the structure of the Web; and when we discuss empirical data on large sets of Web pages in Section 13.4, it will be based on collections assembled by search engines according to such rules.

13.3 The Web as a Directed Graph

Viewing social and economic networks in terms of their graph structures provides significant insights, and the same is true for information networks such as the Web. When we view the Web as a graph, it allows us to better understand the logical relationships expressed by its links; to break its structure into smaller, cohesive units; and — as we will see in Chapter 14 — to identify important pages as a step in organizing the results of Web searches.

To begin with, it is important to note two things. First, in discussing the graph structure of the Web, we will be following the plan outlined at the end of Section 13.2 and focusing on its navigational links. As we observed in that discussion, the navigational links still form the bulk of the Web's structural backbone, despite the increasing richness of Web content as a whole.

Second, we need to appreciate that the fundamentally *directed* nature of the Web makes it different from many of the networks we've considered thus far. Recall that in a directed graph, the edges don't simply connect pairs of nodes in a symmetric way — they point *from* one node *to* another. This is clearly true on the Web: just because you write a blog post and include a link to the Web page of a company or organization, there is no reason to believe that they will necessarily reciprocate and include a link back to the blog post.

This distinction between directedness and undirectedness is an important aspect of the difference between social and information networks; an analogy here is to the difference between the global friendship network that we discussed in Chapter 2, showing who is friends with whom, and the *global name-recognition network*, in which there is a link from person A to person B if A has heard of B. This latter network is directed and in fact quite

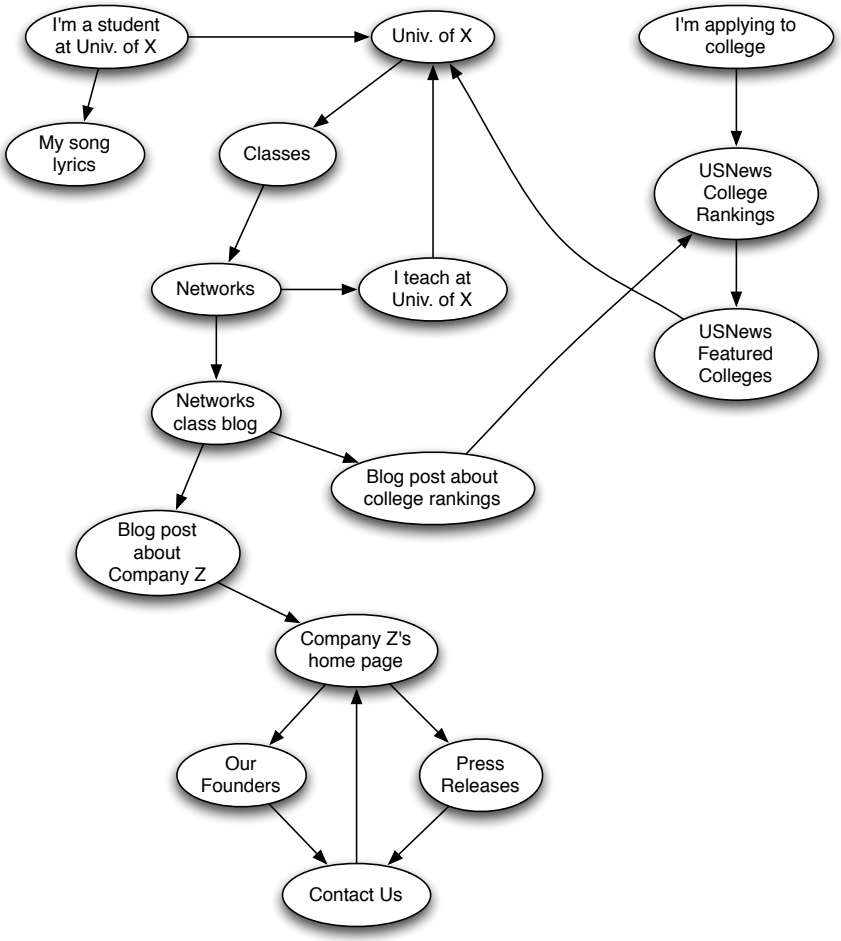

Figure 13.5: A directed graph formed by the links among a small set of Web pages.

asymmetric — famous celebrities are recognizable to millions of people, and in fact millions closely track the goings-on in their lives, but one doesn't expect that such celebrities are in any sense aware of the names or identities of all these fans. In other words, the global name-recognition network is structurally more similar to an information network like the Web than it is to a traditional social network defined by friendship.

Paths and Strong Connectivity. The connectivity of undirected graphs was defined in terms of paths: two nodes are linked by a *path* if we can follow a sequence of edges from one to the other; a graph is *connected* if every pair of nodes is linked by a path; and we can break up a disconnected graph into its connected *components*. Now that we're dealing with a directed graph, we're going to try following the same general strategy for talking about connectivity; but to do this, we first need to rework the definition of a path to take directions

into account, and this will necessarily make the subsequent definitions more subtle.

First, a *path* from a node A to a node B in a directed graph is a sequence of nodes, beginning with A and ending with B, with the property that each consecutive pair of nodes in the sequence is connected by an edge pointing in the forward direction. This "pointing in the forward direction" condition makes the definition of a path in a directed graph different from the corresponding definition for undirected graphs, where edges have no direction. On the Web, this notion of following links only in the forward direction corresponds naturally to the notion of viewing Web pages with a browser: we can follow a link when it's emanating from the page we're on, but we aren't in general aware of all the links that point *to* the page we're currently visiting.

We can try out this definition on the example in Figure 13.5, which shows the directed graph formed by the links among a small set of Web pages; it depicts some of the people and classes associated with the hypothetical University of X, which we imagine to have once been a Featured College in a national magazine. By following a sequence of links in this example (all in the forward direction), we can discover that there's a path from the node labeled *Univ. of X* to the node labeled *US News College Rankings*: we can follow a link from *Univ. of X* to its *Classes* page, then to the home page of its class entitled *Networks*, then to the *Networks class blog*, then to a class blog post about college rankings, and finally via a link from this blog post to the page *US News College Rankings*. On the other hand, there's no path from the node labeled *Company Z's home page* to the node labeled *US News College Rankings* — there would be if we were allowed to follow directed edges in the reverse direction, but following edges forward from *Company Z's home page*, we can only reach *Our Founders*, *Press Releases*, and *Contact Us*.

With the definition of a path in hand, we can adapt the notion of connectivity to the setting of directed graphs. We say that a directed graph is *strongly connected* if there is a path from every node to every other node. So for example, the directed graph of Web pages in Figure 13.5 is not strongly connected, since as we've just observed, there are certain pairs of nodes for which there's no path from the first to the second.

Strongly Connected Components. When a directed graph is not strongly connected, it's important to be able to describe its *reachability* properties: identifying which nodes are "reachable" from which others using paths. To define this notion precisely, it's again useful to draw an analogy to the simpler case of undirected graphs, and try to start from there. For an undirected graph, its connected components serve as a very effective summary of reachability: if two nodes belong to the same component, then they can reach each other by paths; and if two nodes belong to different components then they can't.

But reachability in a directed graph is a harder thing to summarize. In a directed graph, we can have pairs of nodes for which each can reach the other (like *Univ. of X* and *US*

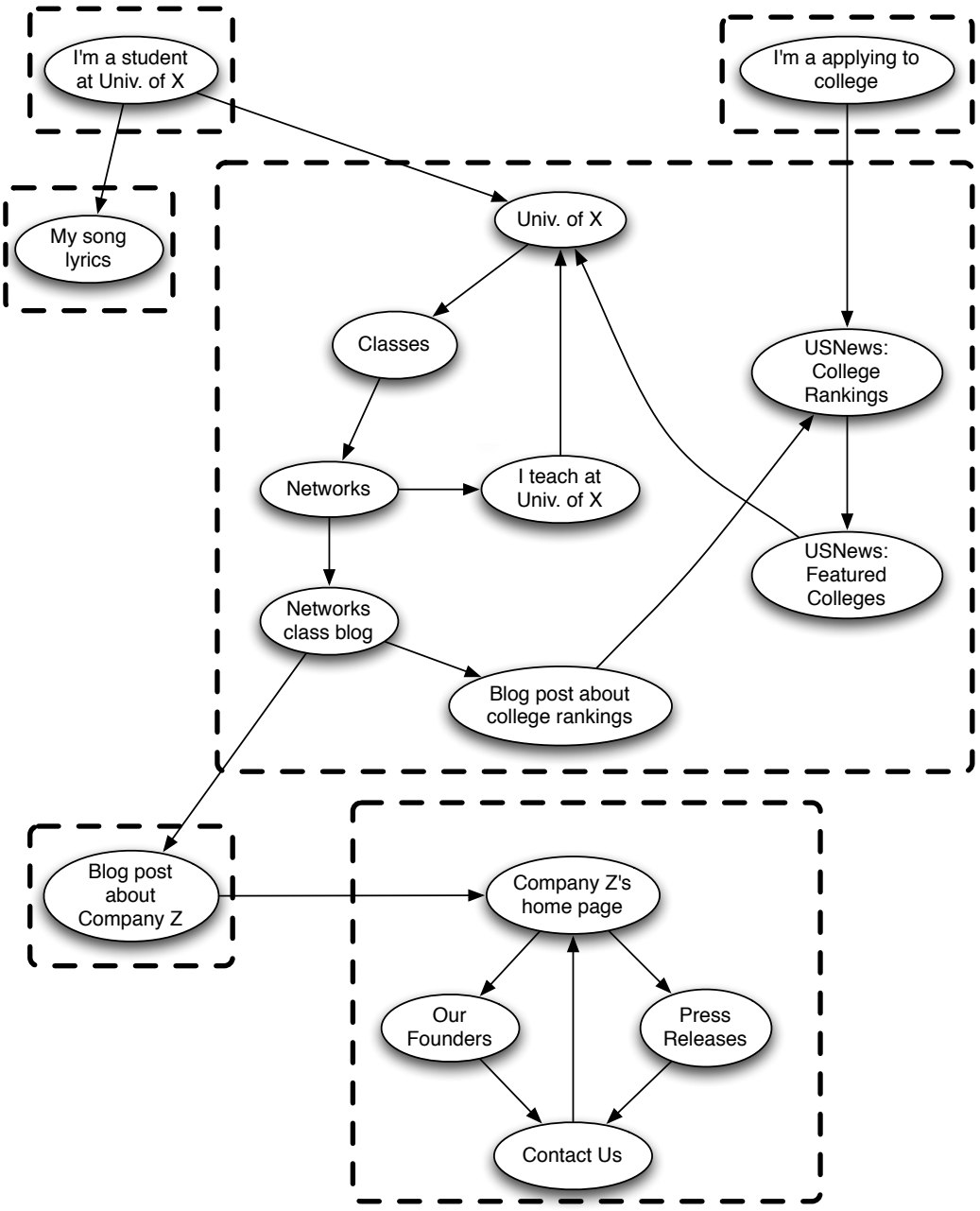

Figure 13.6: A directed graph with its strongly connected components identified.

News College Rankings), pairs for which one can reach the other but not vice versa (like *US News College Rankings* and *Company Z's home page*), and pairs for which neither can reach the other (like *I'm a student at Univ. of X* and *I'm applying to college*). Moreover, the conceptual complexity of reachability in a directed graph corresponds to a kind of "visual" complexity as well: whereas the components of an undirected graph naturally correspond to separate chunks of the graph with no edges between them, a directed graph that is not strongly connected does not break equally obviously into non-interacting pieces. How then should we describe its reachability properties?

The key is to find the right notion of a "component" for directed graphs, and in fact one can do this with a definition that strictly mirrors the formal definition of a component in an undirected graph.

> *We say that a* strongly connected component *(SCC) in a directed graph is a subset of the nodes such that: (i) every node in the subset has a path to every other; and (ii) the subset is not part of some larger set with the property that every node can reach every other.*

As in the undirected case, part (i) of this definition says that all nodes within a strongly connected component can reach each other, and part (ii) of this definition says that the strongly connected components correspond as much as possible to separate "pieces," not smaller portions of larger pieces.

It helps to consider an example: in Figure 13.6 we show the strongly connected components for the directed graph from Figure 13.5. Notice the role that part (ii) of the definition plays in producing the separate pieces of the graph in this picture: the set of four nodes consisting of *Univ. of X*, *Classes*, *Networks*, and *I teach at Univ. of X* collectively satisfy part (i) of the definition, but they do not form a strongly connected component because they belong to a larger set that also satisfies (i).

Looking at this picture, one can see how the SCCs serve as a compact summary of the reachability properties of the directed graph. Given two nodes A and B, we can tell if there is a path from A to B as follows. First, we find the SCCs containing A and B respectively. If A and B belong to the same SCC, then they can each reach each other by paths. Otherwise, viewing the SCCs themselves as larger "super-nodes", we see if there is a way to walk from the SCC of A to the SCC of B, following edges between SCCs in the forward direction. If there is a way to do this, then this walk can be opened up into a path from A to B in the graph; if there is no way to do this, then there is no path from A to B.

13.4 The Bow-Tie Structure of the Web

In 1999, after the Web had been growing for the better part of a decade, Andrei Broder and his colleagues set out to build a global map of the Web, using strongly connected components

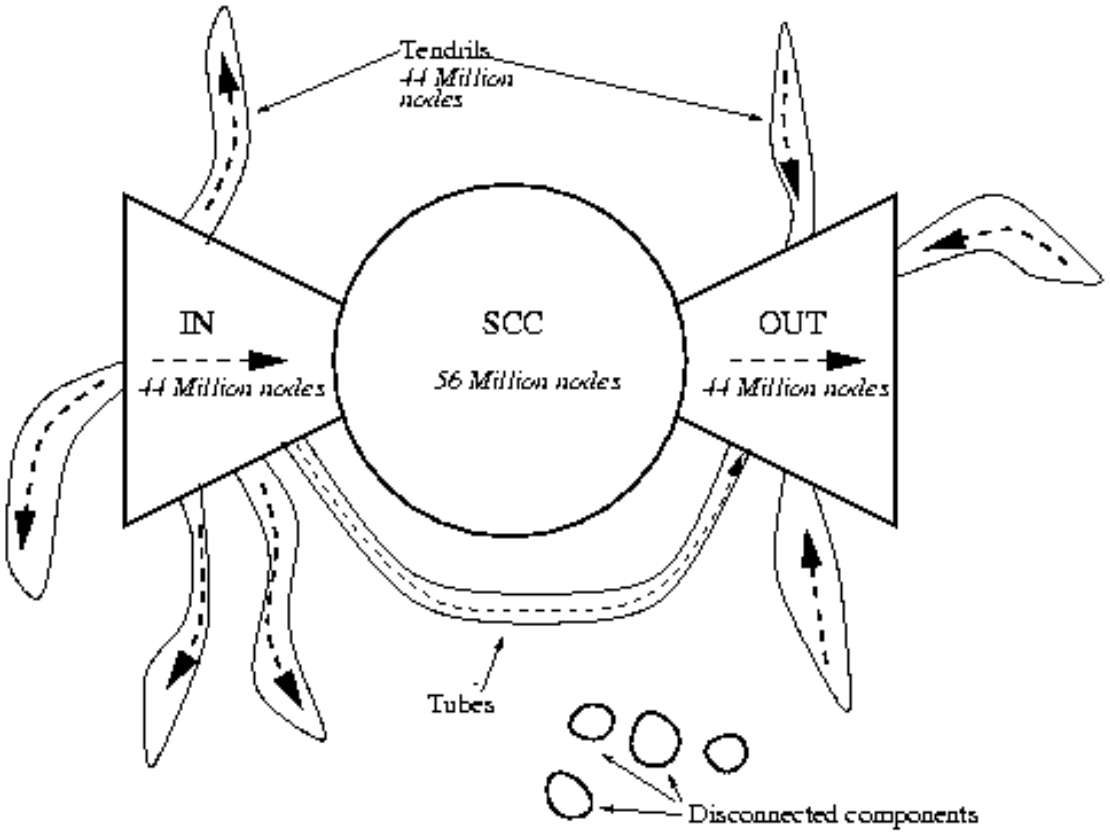

Figure 13.7: A schematic picture of the bow-structure of the Web (image from [80]). Although the numbers are now outdated, the structure has persisted.

as the basic building blocks [80]. For their raw data, they used the index of pages and links from one of the largest commercial search engines at the time, AltaVista. Their influential study has since been replicated on other, even larger snapshots of the Web, including an early index of Google's search engine [56] and large research collections of Web pages [133]. Similar analyses have been carried out for particular well-defined pieces of the Web, including the links among articles on Wikipedia [83], and even for complex directed graph structures arising in other domains, such as the network of interbank loans depicted in Figure 1.3 from Chapter 1 [50]. In this way, although the actual snapshot of the Web used by Broder et al. in their original study comes from an earlier time in the Web's history, the mapping paradigm they proposed continues to be a useful way of thinking about giant directed graphs in the context of the Web and more generally.

A Giant Strongly Connected Component. A "map" of the Web clearly can't resemble a map of the physical world in any serious sense, given the scale and complexity of the network being analyzed. Rather, what Broder et al. [80] wanted was something more conceptual — an abstract map dividing the Web into a few large pieces, and showing in a stylized way how these pieces fit together.

Their first finding was that the Web contains a giant strongly connected component. Recall from our discussions in Chapter 2 that many naturally occuring undirected graphs have a giant connected component — a single component containing a significant fraction of all the nodes. The fact that the directed analogue of this phenomenon holds for the Web is not hard to believe based on analogous thought experiments. Roughly, the point is that a number of major search engines and other "starting page" sites have links leading to directory-type pages from which you can, in turn, reach the home pages of major educational institutions, large companies, and governmental agencies. From here one can reach most of the pages within each of these large sites. Further, many of the pages within these sites link back to the search engines and starting pages themselves. (The path from *US News College Rankings* to a class blog and back in Figures 13.5 and 13.6 suggests a concrete example for how this happens.) Thus, all these pages can mutually reach one another, and hence all belong to the same strongly connected component. Given that this SCC contains (at least) the home pages of many of the major commercial, governmental, and non-profit organizations in the world, it is easy to believe that it is a giant SCC.

From here, we can invoke an argument — familiar from the undirected case as well — that there is almost surely at most one giant SCC. For if there were two giant SCCs — call them X and Y — all it would take is a single link from any node in X to any node Y, and another link from any node in Y to any node in X, and X and Y would merge to become part of a single SCC.

The Bow-Tie Structure. The second step in the analysis by Broder et al. [80] was to position all the remaining SCCs in relation to the giant one. This involves classfying nodes by their ability to reach and be reached from the giant SCC. The first two sets in this classification are the following.

(1) *IN*: nodes that can reach the giant SCC but cannot be reached from it — i.e., nodes that are "upstream" of it.

(2) *OUT*: nodes that can be reached from the giant SCC but cannot reach it — i.e., nodes are "downstream" of it.

Figure 13.6 forms a useful example for trying out these definitions. Although the network in Figure 13.6 is much too small for any of its SCCs to be considered "giant," we can imagine its largest SCC as the giant one and consider how the other nodes are positioned in relation to

it. In this case, the pages *I'm a student at Univ. of X* and *I'm applying to college* constitute *IN*, and the pages *Blog post about Company Z* and the whole SCC involving Company Z constitute *OUT*. And this is roughly what one intuitively expects to find in these sets: *IN* contains pages that have not been "discovered" by members of the giant SCC, while *OUT* contains pages that may receive links from the giant SCC, but which choose not to link back.

Figure 13.7 shows the original schematic image from Broder et al. [80], depicting the relation of *IN*, *OUT*, and the giant SCC. Because of the visual effect of *IN* and *OUT* as large lobes hanging off the central SCC, Broder et al. termed this the "bow-tie picture" of the Web, with the giant SCC as the "knot" in the middle. The actual sizes of the different pieces shown in the Figure come from the 1999 AltaVista data, and are long since obsolete — the main point is that all three of these pieces are very large.

As Figure 13.7 also makes clear, there are pages that belong to none of *IN*, *OUT*, or the giant SCC — that is, they can neither reach the giant SCC nor be reached from it. These can be further classified as

(3) *Tendrils:* The "tendrils" of the bow-tie consist of (a) the nodes reachable from *IN* that cannot reach the giant SCC, and (b) the nodes that can reach *OUT* but cannot be reached from the giant SCC. For example, the page *My song lyrics* in Figure 13.6 is an example of a tendril page, since it's reachable from *IN* but has no path to the giant SCC. It's possible for a tendril node to satisfy both (a) and (b), in which case it's part of a "tube" that travels from *IN* to *OUT* without touching the giant SCC. (For example, if the page *My song lyrics* happened to link to *Blog post about Company Z* in Figure 13.6, it would be part of a tube.)

(4) *Disconnected:* Finally, there are nodes that would not have a path to the giant SCC even if we completely ignored the directions of the edges. These belong to none of the preceding categories.

Taken as a whole, then, the bow-tie picture of the Web provides a high-level view of the Web's structure, based on its reachability properties and how its strongly connected components fit together. From it, we see that the Web contains a central "core" containing many of its most prominent pages, with many other nodes that lie upstream, downstream, or "off to the side" relative to this core. It is also a highly dynamic picture: as people create pages and links, the constituent pieces of the bow-tie are constantly shifting their boundaries, with nodes entering (and also leaving) the giant SCC over time. But subsequent studies suggest that the aggregate picture remains relatively stable over time, even as the detailed structure changes continuously.

While the bow-tie picture gives us a global view of the Web, it doesn't give us insight into the more fine-grained patterns of connections within the constituent parts — connections which could serve to highlight important Web pages or communities of thematically related

pages. Addressing these latter issues will require more detailed network analysis, which we undertake in Chapter 14; as we will see, this requires us to think about what it means for a Web page to occupy a "powerful" position, and it leads to methods that bear directly on the design of Web search engines. More generally, network analysis of the Web forms one ingredient in a broader emerging research agenda that aims to understand the structure, behavior, and evolution of the Web as a phenomenon in itself [220].

13.5 The Emergence of Web 2.0

The increasing richness of Web content, which we've encountered through the distinction between navigational and transactional links, fueled a series of further significant changes in the Web during its second decade of existence, between 2000 and 2009. Three major forces behind these changes were

(i) the growth of Web authoring styles that enabled many people to collectively create and maintain shared content;

(ii) the movement of people's personal on-line data (including e-mail, calendars, photos, and videos) from their own computers to services offered and hosted by large companies; and

(iii) the growth of linking styles that emphasize on-line connections between people, not just between documents.

Taken together, this set of changes altered user experience on the Web sufficiently that technologists led by Tim O'Reilly and others began speaking in 2004 and 2005 about the emergence of *Web 2.0* [335]. While the term evokes images of a new software release, there is agreement that Web 2.0 is principally "an attitude, not a technology" [125]. There has never been perfect consensus on the meaning of the term, but it has generally connoted a major next step in the evolution of the Web, driven by versions of principles (i), (ii), and (iii) above (as well as others), and arising from a confluence of factors rather than any one organization's centralized decisions.

Indeed, there was an explosion of prominent new sites during the period 2004–2006 that exemplified these three principles (i), (ii), and (iii), sometimes in combination. To name just a few examples: Wikipedia grew rapidly during this period, as people embraced the idea of collectively editing articles to create an open encyclopedia on the Web (principle (i)); Gmail and other on-line e-mail services encouraged individuals to let companies like Google host their archives of e-mail (principle (ii)); MySpace and Facebook achieved widespread adoption with a set of features that primarily emphasized the creation of on-line social networks (principle (iii)).

Many sites during this period combined versions of all three principles. For example, the photo-sharing site Flickr and subsequently the video-sharing site YouTube provided users with a centralized place to store their own photos and videos (principle (ii)), simultaneously enriched this content by allowing a large user community to tag and comment on it (principle (i)), and allowed users to form social connections to others whose content they followed (principle (iii)). The micro-blogging service Twitter extended principle (ii) further, by creating an on-line forum for personal data (in the form of short real-time descriptions of one's experiences, thoughts, and questions) that would otherwise never have been recorded at all. Because many people will all comment at roughly the same time on a current event in the news, Twitter also creates collective summaries of worldwide reactions to such events (principle (i)), and allows users to construct links by which they follow the writings of other users (principle (iii)).

Even if some (or many) of these specific sites are replaced by others in the coming years, the principles they embody have clearly brought about a lasting change in perspective on Web content. These principles have also led to a point that we discussed early in Chapter 1: designers of Web sites today need to think not just about organizing information, but also about the social feedback effects inherent in maintaining an audience of millions of users — users who are able to interact directly not just with the site itself but with one another.

This helps to explain why many of the central concepts in this book relate to phenomena that surround this current phase of the Web's evolution. For example, many of the key rallying cries that accompanied the emergence of Web 2.0 are in a sense shorthand for social phenomena that we discuss in other chapters:

- *"Software that gets better the more people use it."* A core principle of Web 2.0 is that on-line Web sites and services can become more appealing to users — and in fact, often genuinely more valuable to them — as their audiences grow larger. When and how this process takes place forms a central focus in chapters from the next two parts of the book, particularly Chapters 16, 17, and 19.

- *"The wisdom of crowds."* The collaborative authoring of an encyclopedia by millions on Wikipedia, the elevation of news content by group evaluation on Digg, the fact that photos of breaking news now often appear on Flickr before they do in the mainstream news, and many similar developments highlighted the ways in which the audience of a Web 2.0 site — each contributing specific expertise and sometimes misinformation — can produce a collective artifact of significant value. But the "wisdom of the crowds," as this process is now often called, is a subtle phenomenon that can fail as easily as it can succeed. In Chapter 22 we discuss some of the basic work in the theory of markets that helps explain how collective information residing in a large group can be synthesized successfully; and in Chapter 16 we describe ways in which this process can also lead to unexpected and sometimes undesirable outcomes.

- *"The Long Tail."* With many people contributing content to a Web 2.0 site, the system will generally reach a balance between a small amount of hugely popular content and a "long tail" of content with various levels of niche appeal. Such distributions of popularity have important consequences, and will be the topic of Chapter 18.

In addition to the ideas suggested by such mantras, the premises underlying Web 2.0 appear in many other contexts in the book as well. The social-networking aspects of Web 2.0 sites provide rich data for large-studies of social network structure, as discussed in Chapter 2. They offer a basis for empirical studies of the ideas of triadic closure and group affiliation from Chapters 3 and 4, and have been used to evaluate the theories underlying the small-world phenomenon in Chapter 20.

Moreover, many of the features that are common to Web 2.0 sites are designed to explicitly steer some of the underlying social feedback mechanisms in desirable directions. For example, *reputation systems* and *trust systems* enable users to provide signals about the behavior — and misbehavior — of other users. We discussed such systems in the context of structural balance in Chapter 5, and will see their role in providing information essential to the functioning of on-line markets in Chapter 22. Web 2.0 sites also make use of *recommendation systems*, to guide users toward items that they may not know about. In addition to serving as helpful features for a site's users, such recommendation systems interact in complex but important ways with distributions of popularity and the long tail of niche content, as we will see in Chapter 18.

The development of the current generation of Web search engines, led by Google, is sometimes seen as a crucial step in the pivot from the early days of the Web to the era of Web 2.0. In the next two chapters we will discuss how thinking of the Web as a network helped form the foundation for these search engines, and how models based on matching markets helped turn search into a profitable business.

13.6 Exercises

1. Consider the set of 18 Web pages drawn in Figure 13.8, with links forming a directed graph. Which nodes constitute the largest strongly connected component (SCC) in this graph? Taking this as the *giant SCC*, which nodes then belong to the sets IN and OUT defined in Section 13.4? Which nodes belong to the *tendrils* of the graph?

2. As new links are created and old ones are removed among an existing set of Web pages, the pages move between different parts of the bow-tie structure.

   (a) Name an edge you could add or delete from the graph in Figure 13.8 so as to increase the size of the largest strongly connected component.

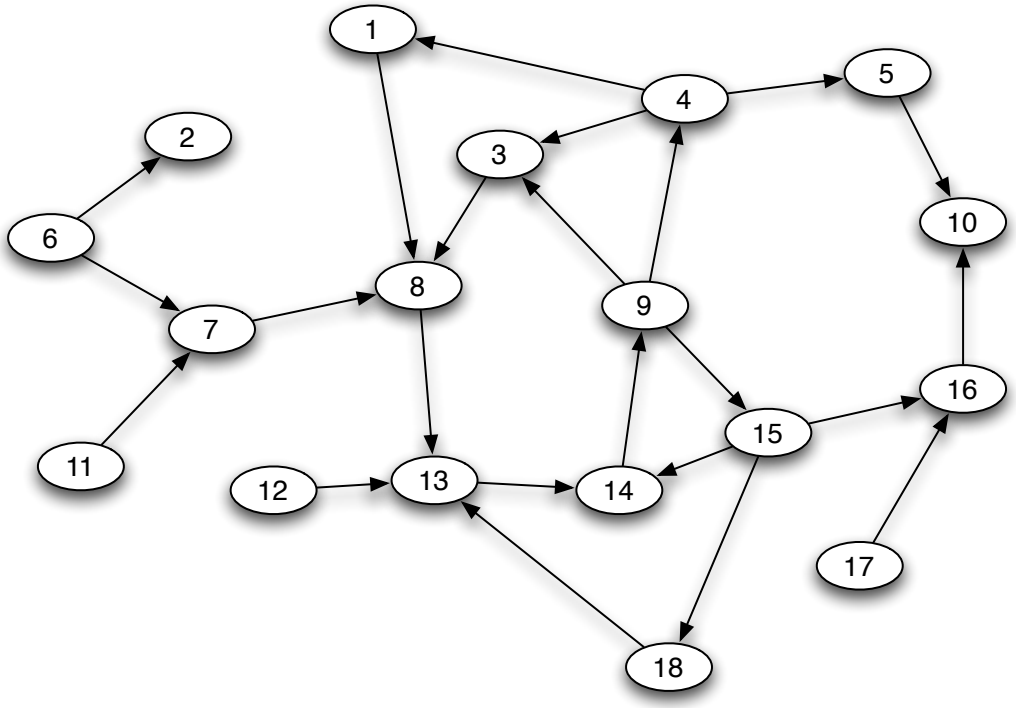

Figure 13.8: A directed graph of Web pages.

(b) Name an edge you could add or delete from the graph in Figure 13.8 so as to increase the size of the set *IN*.

(c) Name an edge you could add or delete from the graph in Figure 13.8 so as to increase the size of the set OUT.

3. In Exercise 2, we considered how the consistuent parts of the bow-tie structure change as edges are added to or removed from the graph. It's also interesting to ask about the magnitude of these changes.

(a) Describe an example of a graph where removing a single edge can reduce the size of the largest strongly connected component by at least 1000 nodes. (Clearly you shouldn't attempt to draw the full graph; rather, you can describe it in words, and also draw a schematic picture if it's useful.)

(b) Describe an example of a graph where adding a single edge can reduce the size of the set OUT by at least 1000 nodes. (Again, you should describe the graph rather than actually drawing it.)