Chapter 19 Web search basics

- The distribution of web page is widely reported to be a power law, in which the total number of web pages with in-degree i is proportional to $1/i^\alpha$; the value of α of typically reported by studies is 2.1.
- 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.
- Notably, in several studies IN and OUT are roughly equal in size, whereas SCC is somewhat larger; most web pages fall into one of these three sets.
- A range of studies has concluded that the average number of keywords in a web search is somewhere between two and three.
- Google identified two principles that helped it to grow at the expense of its competitors: (i) A focus on relevance, specifically precision rather than recall in the first few results; and (ii) a user experience that is lightweight, meaning that both the search query page and the search results page are uncluttered and almost entirely textual, with very few graphical elements.
- The Web contains multiple copies of the same content. By some estimates, as many as 40% of the pages on the Web are duplicates of other pages.