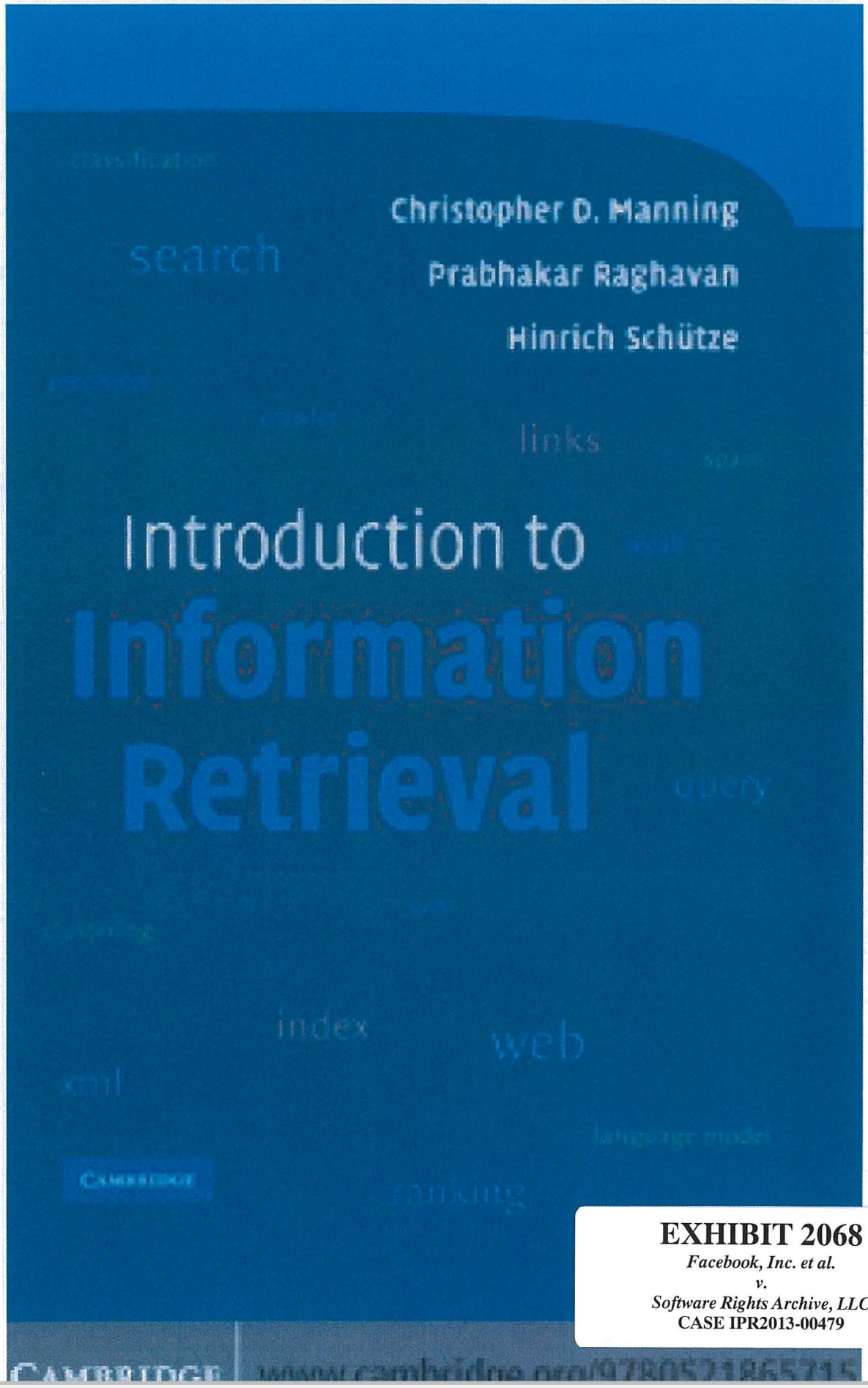

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK
Published in the United States of America by Cambridge University Press, New York
www.cambridge.org
Information on this title: www.cambridge.org/9780521865715

First published in print format 2008

ISBN-13 978-0-511-41405-3 eBook (EBL)

ISBN-13 978-0-521-86571-5 hardback

Figure 19.5 Cloaking as used by spammers.

property and their agents, therefore, have a strong incentive to create web pages that rank highly on this query. In a search engine whose scoring was based on term frequencies, a web page with numerous repetitions of maui golf real estate would rank highly. This led to the first generation of *spam*, which (in the context of web search) is the manipulation of web page content for the purpose of appearing high up in search results for selected keywords. To avoid irritating users with these repetitions, sophisticated *spammers* resorted to such tricks as rendering these repeated terms in the same color as the background. Despite these words being consequently invisible to the human user, a search engine indexer would parse the invisible words out of the HTML representation of the web page and index these words as being present in the page.

SPAM

At its root, spam stems from the heterogeneity of motives in content creation on the Web. In particular, many web content creators have commercial motives and therefore stand to gain from manipulating search engine results. You might argue that this is no different from a company that uses large fonts to list its phone numbers in the yellow pages; but this generally costs the company more and is thus a fairer mechanism. A more apt analogy, perhaps, is the use of company names beginning with a long string of As to be listed early in a yellow pages category. In fact, the yellow pages' model of companies paying for larger/darker fonts has been replicated in web search: In many search engines, it is possible to pay to have one's web page included in the search engine's index – a model known as *paid inclusion*. Different search engines have different policies on whether to allow paid inclusion, and whether such a payment has any effect on ranking in search results.

PAID INCLUSION

Search engines soon became sophisticated enough in their spam detection to screen out a large number of repetitions of particular keywords. Spammers responded with a richer set of spam techniques, the best known of which we now describe. The first of these techniques is *cloaking*, shown in Figure 19.5. Here, the spammer's web server returns different pages depending on whether the http request comes from a web search engine's

a web page that has altogether different content than that indexed by the search engine. Such deception of search indexers is unknown in the traditional world of IR; it stems from the fact that the relationship between page publishers and web search engines is not completely collaborative.

A *doorway page* contains text and metadata carefully chosen to rank highly on selected search keywords. When a browser requests the doorway page, it is redirected to a page containing content of a more commercial nature. More complex spamming techniques involve manipulation of the metadata related to a page including (for reasons we will see in Chapter 21) the links into a web page. Given that spamming is inherently an economically motivated activity, there has sprung around it an industry of *search engine optimizers*, or SEOs, to provide consultancy services for clients who seek to have their web pages rank highly on selected keywords. Web search engines frown on this business of attempting to decipher and adapt to their proprietary ranking techniques and indeed announce policies on forms of SEO behavior they do not tolerate (and have been known to shut down search requests from certain SEOs for violation of these). Inevitably, the parrying between such SEOs (who gradually infer features of each web search engine's ranking methods) and the web search engines (who adapt in response) is an unending struggle; indeed, the research subarea of *adversarial information retrieval* has sprung up around this battle. To combat spammers who manipulate the text of their web pages is the exploitation of the link structure of the Web – a technique known as *link analysis*. The first web search engine known to apply link analysis on a large scale (to be detailed in Chapter 21) was Google, although all web search engines currently make use of it (and correspondingly, spammers now invest considerable effort in subverting it – this is known as *link spam*).

SEARCH ENGINE OPTIMIZERS

ADVERSARIAL INFORMATION RETRIEVAL

LINK SPAM

? **Exercise 19.1** If the number of pages with in-degree i is proportional to $1/i^{2.1}$, what is the probability that a randomly chosen web page has in-degree 1?

Exercise 19.2 If the number of pages with in-degree i is proportional to $1/i^{2.1}$, what is the average in-degree of a web page?

Exercise 19.3 If the number of pages with in-degree i is proportional to $1/i^{2.1}$, then as the largest in-degree goes to infinity, does the fraction of pages with in-degree i grow, stay the same, or diminish? How would your answer change for values of the exponent other than 2.1?

Exercise 19.4 The average in-degree of all nodes in a snapshot of the web graph is 9. What can we say about the average out-degree of all nodes in this snapshot?

as MSN, America Online, Yahoo!, and CNN). The primary purpose of these advertisements was *branding*: to convey to the viewer a positive feeling about the brand of the company placing the advertisement. Typically these advertisements are priced on a *cost per mil* (*CPM*) basis: the cost to the company of having its banner advertisement displayed 1,000 times. Some websites struck contracts with their advertisers in which an advertisement was priced not by the number of times it is displayed (also known as *impressions*), but rather by the number of times it is *clicked on* by the user. This pricing model is known as the *cost per click* (*CPC*) model. In such cases, clicking on the advertisement leads the user to a web page set up by the advertiser, where the user is induced to make a purchase. Here, the goal of the advertisement is not so much brand promotion as to induce a transaction. This distinction between brand- and transaction-oriented advertising was already widely recognized in the context of conventional media such as broadcast and print. The interactivity of the web allowed the CPC billing model – clicks could be metered and monitored by the website and billed to the advertiser.

CPM

CPC

The pioneer in this direction was a company named Goto, which changed its name to Overture before eventual acquisition by Yahoo! Goto was not, in the traditional sense, a search engine; rather, for every query term q it accepted *bids* from companies who wanted their web page shown on the query q. In response to the query q, Goto would return the pages of all advertisers who bid for q, ordered by their bids. Furthermore, when the user clicked on one of the returned results, the corresponding advertiser would make a payment to Goto (in the initial implementation, this payment equaled the advertiser's bid for q).

Several aspects of Goto's model are worth highlighting. First, a user typing the query q into Goto's search interface was actively expressing an interest and intent related to the query q. For instance, a user typing golf clubs is more likely to be imminently purchasing a set than one who is simply browsing news on golf. Second, Goto only got compensated when a user actually expressed interest in an advertisement – as evinced by the user clicking the advertisement. Taken together, these created a powerful mechanism by which to connect advertisers to consumers, quickly raising the annual revenues of Goto/Overture into hundreds of millions of dollars. This style of search engine came to be known variously as *sponsored search* or *search advertising*.

SPONSORED SEARCH

SEARCH ADVERTISING

Given these two kinds of search engines – the "pure" search engines such as Google and Altavista versus the sponsored search engines – the logical next step was to combine them into a single user experience. Current search engines follow precisely this model: They provide pure search re-

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.