Privacy in Information Retrieval

Ian Soboroff
National Institute of Standards and Technology*

responding to comments from Omar, Jan 2, 2024

1 Introduction

Information retrieval (IR) is fundamentally the discipline of developing algorithms and systems to help people find information, whether stored in books, articles, newspapers, social media, and the web. Historically, privacy has not been a primary concern for IR, but as the technology reaches relative maturity and near complete penetration into everyday computer usage, that is beginning to change.

In 2014, a workshop held at the ACM Conference on Research and Development in Information Retrieval (SIGIR) [30] attempted to sketch a research agenda for privacy-preserving IR, and the ensuing discussion revealed that privacy is a complicated subject, representing different things to different people: What was meant to be private? How do we define "private"? Private to whom, and for how long, and why?

The Electronic Frontier Foundation (EFF), on its web page describing the need for online privacy, writes:[1]

> Your Web searches about sensitive medical information might seem a secret between you and your search engine, but companies like Google are creating a treasure trove of personal information by logging your online activities, and making it potentially available to any party wielding enough cash or a subpoena.

This aptly illustrates one type of privacy in information retrieval: the privacy of the user's search. Personalization is at the root of high-quality web search, and large web search engine companies make use of user queries, and indeed all sorts of user profile and behavior signals (including past searches, browsing history, click behavior, advertising interactions, location, IP address, and more) in learning to optimally answer your query.

*This work may mention certain products, companies, or services. Any such mentions do not serve to endorse, recommend, or review those services, companies, or products by NIST.

[1] `https://www.eff.org/issues/privacy`, visited on May 26, 2020

As user information is amassed by search providers, they themselves are faced with a second type of privacy problem. How much of that data should people in the organization be able to see? Can any of that information be released outside the organization? In 2006, America Online (AOL), a large internet service provider, made a log of searches available to researchers. AOL thought that the data was anonymized, but reporters at the New York Times were able to identify several individuals whose queries were in the log, and those people were surprised that their queries were exposed to the world.[2]

What parts of those logs are useful for training models and should be kept? If data is retained long-term, governments may attempt to search it, placing the organization in an uncomfortable gatekeeper role, especially considering that many search providers operate around the world and no two countries have the same legal perspectives on privacy.[3] Should the users who were full partners in the creation of that data be able to see it? Understand how it is used, both for their benefit and the benefit of others? May they expect that it would be removed if they asked?

A third type of privacy problem is embodied in the information being searched. Where was it collected from? Should all accessible information be presumed to be searchable? Perhaps different users should only be able to discover information that they are allowed to discover.[4] In some collections it is possible to define notions of privilege and protection that could drive access policies. In the world of social media where anyone may share anything about anyone else, perhaps not.

What do we mean by privacy, anyway? In this chapter I take a broad view that privacy means the ability of people to perceive how "their information" is known, understand how that exposure affects their information retrieval experience, and control what information is exposed and to whom. "People" here are everyday searchers, search providers, and people whose information is being searched. The notion of information ownership is also somewhat slippery, and so I will again take a broad perspective that entities have an "ownership" relation to information for which they are concerned for its privacy. Hence, my interest in the queries I issue to a web search engine is just as valid as the search engine's interest.

Clearly there are larger views on privacy, such as policies, laws, and social norms. This chapter tries to survey the technical issues and the IR research landscape, leaving these broader questions to other venues. For those who are interested in deeper theories of privacy and how they are situated within societies

[2] See `https://en.wikipedia.org/wiki/AOL_search_log_release` (visited Dec 12, 2023) for more details on the AOL query log release. The New York Times article, "A Face is Exposed for User No. 4417749", is at `https://www.nytimes.com/2006/08/09/technology/09aol.html` (visited Jan 2, 2024).

[3] The Internet Association of Privacy Professionals (IAPP) maintains a website on privacy legislation in US states, `https://iapp.org/resources/article/us-state-privacy-legislation-tracker/` (updated on Dec 22, 2023), which is helpful for seeing how different laws are evolving in different states in the *same* country.

[4] This could mean a right to be forgotten, or an absence of the right to information, depending on your perspective.

and countries, McDonald and Forte's recent survey (2020) presents an excellent starting point.

The study of privacy in information retrieval is just in its infancy, with few fundamental results to describe. Thus, my goal is to structure the nature of the problem, indicate critical work in each of those directions, and light a path to future steps.

2 Risks and Rewards of Privacy Leakage

Commercial search engines make heavy use of personalization to refine search results [20]. This stems directly from the abstractive process of querying for an information need, which makes queries hard to understand, combined with the fact that different users make different assessments of search relevance, which makes it hard to know what you really want. These two factors create a ceiling on the effectiveness that a search system can achieve based on content-query matching alone. Personalization is one path past that barrier, since it effectively expands the feature space for matching the user to search results.

Personalization is based on a range of searcher behaviors, from search queries issued and clicks following those search queries, to dwell-time on the search page, the time between leaving the search page to returning with a new search or to click on a different result, and good-abandonment scenarios where the user may not click or scroll at all because their information need was addressed by the results page itself. If the searcher is using a browser from a search company, that browser may be sending telemetry back to the search engine as well. Tracking techniques such as explicit ad trackers, decoy single-pixel images, and cookie harvesting that live on destination pages but report back to the search engine can fill in the gaps to form a rich surveillance of the web user. Personalization happens at a number of different levels of granularity: from a browser's IP address, the search engine can guess the searcher's geographic location and internet service provider; from the browser user-agent the search engine can know computer and operating system types, which can reinforce other demographic information. Note that some of this information is available to the search engine even from browsers using an "incognito" or "private" browsing mode, specific ad- and tracker-blocking add-ons, or a VPN.

Mobile devices add another level of behavior tracking. Mobile devices know your physical location, because that's part of how cellular telephone networks work. Phones often have a GPS receiver and can provide even more exact location information by triangulating from wireless access points. The camera app on a phone can store the GPS coordinates of the camera in a photo's metadata block, so you can tell where the photo was taken. You can use your phone to make purchases, and so the phone may store your credit card, above and beyond any e-commerce sites that use it. Because credit cards and phones can be linked, there is a history of where and when you've purchased anything. The information provided by a phone is not only available to web sites but also to apps that run natively on the phone. Some apps like running trackers actually

provide you an interface to see your location history in certain contexts. Phone operating systems provide APIs to gateway access to this data, but once you've given consent in one situation, that consent can propagate to other locations and services.

Phones are not the only mobile device. Many new cars have a number of wireless communication technologies integrated, including Wifi, Bluetooth, and cellular network connections.[5] For example, the car can call emergency and roadside assistance services, which means they can connect to a cellular network (usually an older network no longer in active use, such as 3G), which means they know your location, which they might share in order to provide you the service you need.

All of those observations represent privacy leakage. Some of that leakage could be explicit: sequences that match the format of credit card numbers or social security numbers appear in search query logs, and who among us hasn't accidentally typed our password into the username space on the login form. Browsing paths and history might not seem that revealing, but in aggregate the search engines know quite a lot about you just from the ads that are shown on pages that you visit, whether you click on them or not.[6] And because those observations happen all the time, they can easily cross the lines from very private to personal to proprietary to official.

There is a payoff. In return for your ~~soul~~ complete behavior history, the search engine can help you find pages you searched for previously, or browsed to previously; it can suggest queries you have used in the past; it can promote the ranking of pages it predicts will be relevant to your search; it can provide search results relevant to your physical location, professional interests, or hobbies; guess that you only want search results in certain languages; cache results you have used in the past; and more. All of these features improve the search experience, and many of them can contribute to improved search effectiveness. For example, my search history can help disambiguate the acronym "TREC", so the search engine knows I want information relating to the information retrieval conference and not the Texas (or Tennessee) Real Estate Commission.

Although the tradeoff presented above is from the perspective of web search engines, it is equally applicable to domain-specific search systems (such as the ACM Digital Library, or your local public library), search systems internal to your workplace, search engines you pay subscription fees towards (like Lexis Nexis), and internal site search on individual web sites. The degree of observation may be different, the search system may not have broad access to many users from whom to leverage personal data, and the search system may not return that to you in improvements to your search results, but the potential is there.

A critical issue with this tradeoff is that it's all-or-nothing: users don't know

[5] `https://www.synopsys.com/blogs/chip-design/connected-vehicle-cybersecurity-wireless-comm.html` (visited Jan 2, 2024)

[6] According to the Wikipedia page `en.wikipedia.org/wiki/AOL_search_log_release`, not only were individual searchers identified by journalists, a play and a documentary have been composed based on entries in the log.

what personal information is being collected, or how it is used, or how it directly improves the service they receive. There's no way for a user to provide only the information they wish to share in order to receive a specific level of service. Some undefined quantity of information is collected, and the benefit of that collection to the user is unquantifiable. The user should really be able to see exactly what data is collected, and how that data improves their user experience.

Search behavior, purchase histories, and mobile tracking data are immensely valuable to companies, and there are data brokers who aggregate and re-sell browsing, behavior, and location data. In August 2022, the US Federal Trade Commission (FTC) sued a data brokerage for "selling geolocation data from hundreds of millions of mobile devices that can be used to trace the movements of individuals to and from sensitive locations." The FTC accused the company of selling data that allowed customers to physically track individual phone users.[7] The data broker industry is valued at somewhere between $200bn and $400bn with a projected growth rate of 5-10% through 2030, based on figures from a number of web market analyses. It's not simple to calculate how much this industry values an individual's data contribution, since the value is in the aggregation, but with 5.7mn people online[8] it averages out to around $50 per person. Again, because users can't directly see how their data impacts their experience, users can't say if they see that level of value spent on them.

In a very interesting 2010 paper, Krause and Horvitz composed an information-theoretic model of the tradeoff of privacy for service [23]). They model the utility of accessing private data with information gain, and the cost of sharing private data with loss of entropy in the distribution of possible users. They then convert these formulations into loss functions and frame the tradeoff as an optimization problem. They computed utility and cost values based on a search log of users who had agreed to participate in a program about personal data use and enhanced search effectiveness. Lastly, they surveyed 1,451 individuals on their perceptions of sensitivity and privacy risk, and used these responses and the log data to calibrate the model. They found that significant personalization gains can be made with only a small amount of personal data, and that in many cases there are diminishing returns to using more data. Of couse, since that time, nearly every online services has increased the amount of personal data they collect, but this work shows that it is possible for search engines to optimize the trade-off between privacy and performance from personalization.

3 Privacy for searchers

In this first section I want to survey efforts to support the privacy of the searcher. Some of this work investigates the prevalence of tracking technologies, while other work tries to actively inform the user of impending privacy leakage. I

[7] The press release can be found at `https://www.ftc.gov/news-events/news/press-releases/2022/08/ftc-sues-kochava-selling-data-tracks-people-reproductive-health-clinics-places-worship-other`. As of January 2024 the case is still pending.

[8] `https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx`, Jan 2, 2024

mostly discuss web search, since it represents an active and adversarial marketplace for users' personal behavior data, and thus the most vibrant ecosystem for the evolution of technology in that space. Clearly, when users work with site search, vertical search, or other focused search sites hosted on the web, any and all behavior can be tracked, but the major web companies are also able to track a single user's behavior across the web. When searching on a non-web-based search system, or a specialized search intentionally detached from the tracking networks, say a local library catalog, privacy leakage is still possible but mostly constrained to that site.[9]

Princeton's Web Transparency and Accountability Project (WebTAP)[10] conducted a study of one million web pages, and a survey of users of ad blocker browser extensions, to understand both the reach of tracking technology and how users understand and react to it [15, 24]. These papers together make an excellent primer on tracking technology, at least as it existed at that time (2016 – 2018). They define the *first party* to be the site that the user visits and intends to interact with. In the process of that interaction, they may connect to several *third parties* linked by the first party, but not necessarily revealed to the user. Each third party adds cookies to the user's browser, and they can see their own cookie whenever a user visits a page with a tracker from that party. There are also non-cookie-based approaches to identifying the user, such as computing a fingerprint based on properties of the user's browser and network connection. There are techniques to try and steal tracking information between third parties. Mathur et al (2018) found through surveys that users frequently use browser plugins to block ads and trackers, but usually do so to improve their browsing experience rather than out of privacy concerns. Englehardt and Naranayan (2016) present the most recent published survey of web tracking technology [15].

Here is an example session. A user goes to a web search engine and types their query. The search engine knows the query, the results, and the user, either because they are directly logged into the search engine or through trackers or fingerprints. The search engine instruments the search result page so that they know how the user scrolls on the page, as well as any results they click on. When the user follows a search result, the destination page may link to trackers from the referring search engine as well as other third parties. Each of those parties now notes that this user has visited this page; depending on the third party, and the degree to which they have "tagged" this user, they have a more or less complete trace of the user's site visits. The search engine is the only party that knows the query, unless they sell or unintentionally leak it through bugs in their tracking technology. It's conceivable that a tracker network with a cluster of page visits surrounding a search engine could deduce the query or at least a probable search intent. The provider of the user's internet connection knows the

[9] Library catalogs are my standard contrastive example to large-scale web search, since librarians in the United States have a strong culture of supporting privacy; see, for example, the American Libraries Association's page on privacy advocacy, `https://www.ala.org/advocacy/privacy` (visited Dec 12, 2023) However, libraries may obtain online catalog access services from companies that may not share that perspective.

[10] `https://webtap.princeton.edu/`, visited on Feb 28, 2021

destination of all packets in or out, and for unencrypted connections (HTTP and not HTTPS) can know the full details of all information transmitted. Lastly, the user's browser has the ability to log all the behavior of the user, including passwords; this logging is used to support automatic completion of various kinds (URLs, search queries, addresses, credit card numbers, etc.) and may be subject to attack by first or third parties.

Biega et al. (2014) advocate for a user-centric conception of privacy in IR. They formulate a probabilistic model that estimates the chance that a searcher might use a sensitive keyword, based on other words they have queried for. This simple model was fairly limited, but drove a vision of tools that could watch over the shoulder of a searcher and guide them if they were conducting privacy-revealing searches [9].

Zimmerman et al (2019) investigated enhancing a search engine result page with "nudges" that warn searchers away from websites that might risk privacy leakage. Given search results for health-related queries, they counted trackers on result sites, and found that users warned away from sites with heavy tracker presence were somewhat less likely to encounter harmful information, and at any rate encountered fewer trackers [40]. However, in follow-on work, they did not find that their most effective nudge technique reduced privacy impacts for studied users [39]. In 2020, Zimmerman et al found that if sites had a top-level domain of .org or .gov, they were likely to employ fewer third-party trackers, and proposed that such page cues could help users to assess privacy risks when using search data [41].

Query suggestions are an interesting privacy risk surface, since they involve sequences of partial queries which themselves may reveal more than a fully-formed query. Suggestions happen at multiple levels. At the search provider, the system may suggest alternative queries to the user along with the search results, or suggestions may be offered in real time, as the user types into the search box, or a combination of the two. Additionally, the browser may offer suggestions of URLs or search queries in the location bar. Those suggestions are driven by local history as well as APIs from search engines. In 2013, Facebook published an exploratory analysis of self-censorship on their platform, that is, when users would be typing a status update and delete it before posting [12]. However, social media posts have a notional audience, whereas users may not perceive any audience for their search queries.

3.1 Privacy policies

Privacy policies are required by law in many countries. For example, Europe's General Data Protection Regulation (GDPR) not only requires a privacy policy, but it must be "concise, transparent, intelligible, easily accessible, and must use clear and plain language."[11] In the United States, there are a number of overlapping federal and state laws intended to protect user privacy.

[11] `https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/individual-rights/right-to-be-informed/`, visited Feb 10, 2024

Pew Research's 2023 American Trends Panel found that people overall have low expectations of and trust in companies and the government when it comes to online privacy. 71% of Americans say they are worried about government use of people's data, 67% say they understand little to nothing about what companies do with their personal data, and 73% feel thy have little to no control over what companies do with their data. 72% of Americans say there should be more government regulation covering what companies do with users' data, but 71% say they have little to no trust that tech companies will be held responsible for data misuse. Privacy policies don't help: 56% of people say they frequently click "agree" without actually reading the policy and 61% think they don't effectively describe what companies do with user data.[12]

In 2002, the World Wide Web Consortium issued the Platform for Privacy Preferences (P3P) as an official recommendation. P3P is a structured format for expressing privacy policies; web sites could link to such policies and browsers could conceptually implement them.[13] However, the standard was last revised in 2006, and officially retired in 2018. Even during its development, the Electronic Privacy Information Center found the then-under-development P3P "a complex and confusing that will make it more difficult for Internet users to protect their privacy."[14]. The problem was complicated, the companies involved had little incentive for successful solutions, and non-technical users found them unusable anyway.

It's clear that technologies are moving to support the user's understanding of their possible privacy exposure. Mainstream browsers support the Do-Not-Track HTTP header, which allows users to express to sites they visit that they wish to opt out of being tracked. Online tools such as EFF's Cover Your Tracks[15] conduct an analysis on your browser to reveal sources of privacy leakage, and it seems reasonable to suppose that such information could be presented as simply as a padlock icon indicates an HTTPS connection. Users typically have several ways of issuing a search besides going to a search engine's web page, including typing the search in a URL bar or in a special search box in the browser, so the browser can watch for some privacy-revealing search queries. Browsers are starting to mediate search queries from the URL bar by default, proxy connections to obscure source addresses, and support generating temporary email addresses for registration.

In summary, searcher privacy risks include inadvertently providing personal information to search engines, through individual queries, query sessions, and contextual information from cookies and tracking information. Challenges to overcoming these risks include the enormous profit potential for companies in collecting personal information and the opaque and complex ways in which collectible information is created and made available. Opportunities include browsers that put their users ahead of the companies that develop them, regu-

[12] `https://www.pewresearch.org/internet/2023/10/18/how-americans-view-data-privacy/`

[13] `https://www.w3.org/TR/P3P11/`

[14] `https://archive.epic.org/reports/prettypoorprivacy.html`

[15] `https://coveryourtracks.eff.org/`

lations on data privacy that have penalties in alignment with profit gain, and making information exposure more obvious to users.

4 Privacy for search engines

Web search engines collect data for personalization, so how should they limit the risk that this data will be misused? Will search engine company employees stalk, harrass, or doxx users for personal or ideological reasons? Are users who are famous at risk of having their search history sold to a newspaper? Government law-enforcement and intelligence agencies are quite interested in users' search and browsing history, and companies need to decide how much to cooperate and how. Since personalization can improve search results, we might also reasonably expect site search, verticals, and smaller search portals to also offer personalization. Perhaps the big web companies will devise a way to sell personalization as a service.

In 2020, NIST created a Privacy Framework [26] to help organizations create or improve privacy programs. A risk-management framework, the Privacy Framework guides the reader through identifying data, conducting a privacy risk assessment, developing privacy policies, understanding legal obligations, building appropriate controls and protections, and maintaining the risk profile over time. For organizations deploying search or companies providing search services, this framework can support responsible use of data collected for personalization, especially if incorporated from the start.

Appropriate controls might include suitably-constructed access control mechanisms like role-based and rule-based access controls. Access control in computer systems achieved prominence through Bell and LaPadula's work in the context of development of the Multics operating system and Denning's lattice-based access control model.[3, 4, 2, 13] This and other proposals were formal models of computer security which theoretically could provide provable guarantees on information access, integrity, and/or security. A few government systems such as Multics and Trusted Xenix were built based on these models. These concepts were delineated as Mandatory Access Controls (MAC) and Discretionary Access Controls (DAC). Subsequent concepts like role-based access control, rule-based access control, and organization-based access control developed from these, trying to make these systems easier to understand, develop, and maintain.

Biega et al (2017) proposed a tiered approach, "privacy through solidarity," where proxies called *mediator accounts* scramble the usage profiles from many users before passing along a request to a search engine. The setup is similar to a metasearch architecture. User queries are proxied to a matching mediator, which aggregates requests from many users. Beyond requests, responses, ratings, and other behavioral data may be aggregated as well. The mediator acts as an agent for all the users querying through it, presenting a single face to the search engine made up of aggregated behavioral information. After the interaction is complete, the mediator drops any attribution to the querying user. Such an architecture requires that the user trust the mediator proxy, and it also provides

a deniability shield for the search engine. Biega and colleagues investigated different methods of assigning queries to mediators, and the resulting privacy-effectiveness trade-offs using data from StackExchange [7]. If future internet regulation forces search engines and social media companies to decouple from advertising and tracking networks, a mediator architecture may offer a technical solution.

4.1 Approaches to data retention policies

The European Union's GDPR requirements state that personal information collection be limited to what is relevant and necessary. Biega et al (2020) explore this ill-defined standard in the context of recommender systems, to see if data minimization is feasible, and how it would impact different common recommendation algorithms [8]. They suggest tying the minimization standard to effectiveness metrics, and then explore the effect of global and per-user minimization on nearest-neighbor and matrix-factorization approaches using the familiar MovieLens and Google Locations recommender systems datasets.

Many approaches to collecting private data assume that some level of permission is obtained which does not then decrease either by users changing their minds or governments changing their policies. Ginart et al (2019) observed that for many machine learning algorithms, the only method to reliably delete information is to retrain from scratch, but they developed deletion algorithms for k-means clustering that were more efficient than retraining [18]. Clearly, there is significant cost to companies to getting rid of personal data.

4.2 Anonymization and Differential Privacy

Sweeney's k-Anonymity paper [32] presents the earliest work in anonymity in databases, and this is the work that the original differential privacy approaches compare to. She also identifies some of the most straightforward privacy attacks, such as linking across databases. But most importantly to the IR research community, she asks the question of whether a data holder can release a version of their data to the scientific community such that the individuals in the data cannot be re-identified. Beyond release of data, it's interesting to consider queries against a search engine as probes for personalization data leaks, and wonder if there is a formal framework within which this could be studied.

Differential privacy is a cryptographic approach to privacy-preserving data analysis. Dwork and Roth's monograph (2014) provides an excellent comprehensive introduction, although of course considerable work has been done in the area since then [14]. A differentially-private database allows the analyst to study the population represented in the database without exposing individual records. The differentially-private database is produced by the data owner from the original database and random information. The resulting database offers a privacy guarantee: an ϵ-differentially private database differing in one row from the original cannot be distinguished from the original with a probability related to $exp(\epsilon)$. Such a database can be created by injecting noisy rows. Proving the

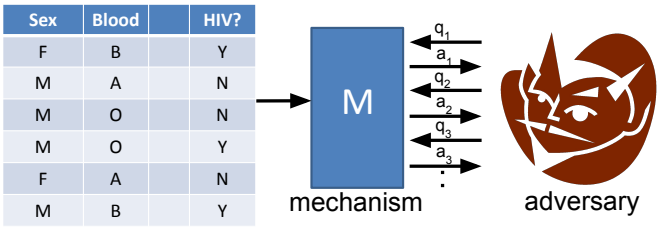

Figure 1: A simple diagram illustrating differential privacy (DP). A DP mechanism acts as an intermediary between the data and someone querying the data, who may be an adversary. The mechanism may be injecting random noise into the data. DP promises that individuals are not identifiable over a series of queries. (Illustration from Salil Vadhan, James Honaker, and Wanrong Zhang's course Applied Privacy for Data Science, slides at `https://opendp.github.io/cs208/spring2022/`, lecture on 2/08 introducing DP.)

privacy guarantee given a particular method of creating the private database, and methods which incorporate the kinds of queries permitted on the database, are the devils within the details. Applying differential privacy to numerical user behavior data is relatively straightforward, whereas applying it to user queries or other linguistic data is more complex.

Singla et al (2014) take a different approach where they stay within the online service and consider privacy risk to be the probability that a particular user is drawn in a sample that will be used to improve the quality of the service. (Note that the online service already *has* the data, so the issue is not whether data is collected or not, just whether that data will be subsequently used.) In their "stochastic privacy" framework, users are asked to state their "privacy risk" which is actually the rate at which they are willing to be sampled in any given optimization, and are offered incentives if the data is in fact sampled and used. Generally, sampling under those constraints is an NP-hard problem, but being careful about how the utility to the online service is formulated allows tractable sampling procedures that honor the users' stated privacy risks. The problem of how to elicit privacy risk, or in fact whether users can usefully estimate their risk in this framework at all, is left as a future direction [31].

4.3 Query logs

Adar (2007) represents early work on query log anonymization, which became a hot topic after the AOL query log release in 2006 [1]. AOL queries were easy to track, because users would log into AOL prior to executing any searches. Adar observed that only 87k of 22M clicks were to HTTPS sites. At any rate, Adar proposed eliminating rare or unique queries, or masking them using cryptographic methods. Biega et al. (2014) observed that Adar's method would still reveal many low-frequency queries that might expose private information [9].

Jones et al (2007) showed how query log attacks can be staged given a separate log of queries and stock of user profile information. They trained simple classifiers to predict demographic information from known queries, and showed that by combining these classifiers the "search space" of possible people to identify can be greatly reduced [21].

Korolova et al (2009) published perhaps the first differentially-private algorithm for releasing a set of queries and clicks. By adopting the differential privacy framework, the authors took a serious step forward from intuitive notions of security to cryptographic ones. This approach results in releasing a fraction of a percent of queries in a log, but it is still vulnerable to attacks with an external log such as those by Jones et al [21, 22].

Zhang et al (2016) presented another differentially-private query log anonymization approach, but in contrast to Korolova et al, they expressly consider the impact on search effectiveness of the resulting log [38]. The steps are

1. first remove queries appearing fewer than five times in the log, and filtering out sensitive information like social security numbers;
2. keep only the first n queries and clicks from a particular session;
3. expand the log with queries from some other source;
4. (DP step) apply Laplacian noise to the query counts, and select the queries whose counts exceed a fixed threshold;
5. (DP step) release log statistics based on the noised data;
6. clean up counts for sequences of queries.

Steps (1) and (2) obscure a lot of potentially private information, but the DP steps add a mathematically-supported bound on privacy loss.

4.4 Personalization leaks

Englehardt et al (2015) is primarily about surveillance of web traffic using tracking cookies that other sites place there. This is really not about IR per se except that searches are often followed by site visits [16].

Bi et al (2013) combined search query logs, Facebook "likes", the myPersonality dataset which links "likes" to demographics, and Open Directory Project categories in order to predict users' age, gender, politicial views, and religion based on search queries. This academic work is a microcosm of what is done to implement targeted advertising; the key difference is that this paper tells you what data they used, where it came from, and how effective they found it to be [6]. By this writing, "demographic prediction" is perceived as a task for researchers; see, for example, Wu et al., where demographics are predicted from search queries, with no occurrence of "privacy" in the text of the paper.

Yom-Tov (2019) analyzed demographic features associated with search queries in the context of identifying patient populations in internet-based health research. In this domain, for example, researchers may wish to track the incidence of a disease based on search engine query patterns [36].

Weber and Castillo (2010) investigated predicting demographic information for searchers based on query logs intersected with profile information and census data. They found that taking demographics into account (i.e., light personalization) improved web search [33]. In 2012, Weber et al. similarly mashed up a collection of political blogs and Wikipedia pages with a search log to predict political views [34].

Search engines clearly have significant exposure to privacy risks, from within the organization and without. The costs of privacy breaches from outside, including hacks, data exfiltrations, and third-party leaks of licensed data, are measured in negative publicity which may drive some users away. The costs of loose internal privacy controls are harder to measure, but again, cases that appear in the news can have an effect. This area is ripe for research, both within companies and in the academic community, as regulation becomes more and more of a possibility.

In summary, searcher privacy risks on the side of the search engine arise from the immense value of people's behavioral data. Challenges include IT security, employee curiosity (or more sinister motivations), government surveillance, legal intrusions, and inadvertent privacy leakage through cookies, trackers, and referrers. Opportunities include technologies like differential privacy (eventually) and policies and practices for data retention, security, and access that limit the amount of collected user data kept over time.

5 Privacy for document owners

The creators and owners of documents have privacy risks as well. This concept has been explored in the computer security domain (although referred to as access control and not privacy control). While much of that work was done in the context of military information systems, today's knowledge enterprises might find themselves of the same need. And last but not least, the emergence of large language models that produce text, images, and video based on training from massive amounts of online content has brought issues of content ownership squarely into the public eye.

Zerr et al (2012) investigated whether image classifiers could learn to distinguish private images, where in this case "private" is defined to be an image judged by a third party to contain information "outside the sphere of the photographer." The authors collected images from Flickr that indicated unrestricted usage rights, and crowdsourced labels using a gamification setup. The resulting classifier obtains 0.88 precision at 0.6 recall, and 0.93 precision at 0.4, which seems low for a privacy preserving approach. However, this work is in classical image classification, predating the deep learning revolution, and so performance might improve [37].

How high should that performance be? It depends on how the classifier is applied. If it is used to give nudges at the time that the image is posted, a lower threshold might be acceptable, versus trying to weed out private images at search time.

A related concept is the "right to be forgotten," the idea that people might have inherent rights to not be discovered in data in certain ways. The classic example is revenge porn: contrary to conventional content policies, platforms may have a responsibility to prevent posting illicit images of a person with intent to damage their reputation because of the rights of the person imaged. The GDPR enshrines this concept in law.[16] The paper cited above by Ginart [18] provides some instructive bounds on the technical barriers to supporting ad hoc data removal from search indices.

The 2019 Fairness, Accountability, Confidentiality, Transparency and Safety in IR (FACTS-IR) workshop [27] noted that identifying confidential information in document collections, and protecting that content from unauthorized use, should be an important research focus for information retrieval. Entire documents, parts of documents, or the acknowledgement of the existence of documents may need to be restricted. End-user search should make access to permitted material easy while controlling access to restricted material, without placing a burden on the searcher.

Restricted access to databases has been studied since there were databases, and the work in the 1970s on the Multics system has already been mentioned above. Another example was IBM's System R, the first SQL system [19, 17]. Bertino et al (2011) provides an excellent survey [5]. Early work distinguished between discretionary access control mechanisms, such as role-based or temporal access, and mandatory access controls based on subject and data classification. Modern questions have included handling insider threats, making databases available to third parties, and regulatory requirements. The critical difference in the IR perspective is that while the database (document collection) exists, the objects in the database have not been classified a priori to support a multilevel database, and the types of accesses that might be permitted on a discretionary basis may not be fully decided when the database is built.

As an example, consider a collection of personal papers, donated by a well-known individual to an archives upon their death. Traditionally, such collections go through a manual accession process through which access controls and considerations emerge. In very large collections, that process may be done at access time rather than when the data is accessioned. A related concept is freedom-of-information requests, where the "database" of documents consists of all documents produced by a government agency and only exists as an abstract concept covering dozens or thousands of data storage systems. Upon request, applicable records are collected and then reviewed according to statutory categories of exempt information prior to release. The information retrieval perspective asks if such processes might be automated, and given that even manual processes must make two kinds of errors, false positives and false negatives, what are the

[16] `https://gdpr.eu/right-to-be-forgotten/`

performance characteristics of manual and automatic processes.

In the early 2000s, James Hendler and others were involved in developing what was called the "Policy-Aware Web" (PAW). PAW provided an access control model at the level of the URI/URL, and proposed that the model would be mediated during an HTTP request featuring authentication and proofs of accessibility.[17]. While PAW did not graduate into a W3C standard or recommendation, companies have developed the idea into commercial web server products, and the spreading use of single-sign-on protocols could possibly provide the authentication infrastructure to support access control at the scale of the web.

Büttcher and Clarke (2005) presented an implementation of a security model for and IR system for file system search on multi-user systems [11]. The model is based on the UNIX security model, and can be implemented either by post-processing the result candidate list, or by filtering the postings list during query processing. They show that the post-processing implementation is exploitable if the system returns scores and if the ranker is BM25;[18] in the exploit, an attacker can deduce the existence of files they don't have permission to access. Their work goes on to examine performance characteristics of the filtering implementation and how to improve them. This paper is the most direct descendant of the foundational database systems research on access controls ([19, 17]) as applied to IR.

Sayed and Oard (2019) proposed a learning-to-rank approach that jointly considered both traditional retrieval features as well as sensitivity features which would indicate that access should be restricted. They developed a simple retrieve-then-filter approach, and then developed a variant of the nDCG metric[19] that incorporated sensitivity. This metric was then optimized by several standard learning-to-rank algorithms [28]. In subsequent work, Sayed et al developed a test collection for sensitivity-aware search, including search topics, relevance judgments, and sensitivity judgments [29].

Searchable encryption (SE) is the notion of allowing searching directly on encrypted data. SE allows the data owner to encrypt the data and place it on an untrusted server. The owner can search the data, and other authorized users can also search. The host holding the data can observe the query stream, but does not know what is being queried or (depending on the scheme) what is being retrieved. Thus, SE provides security for the document owner at the granularity of the entire collection, and to searchers within the context of a specific search. SE schemes are provably robust to leaks of index information, but may leak information about the query stream (called the "search pattern") or the retrieved documents (the "access pattern"). Following the initial conceptualization of this idea in 2000, the notion of security in searchable encryption was formalized and a number of schemes have been proposed, differing in the amount of security, the efficiency of the scheme, and the expressiveness of queries. Bösch et al

[17] `http://www.policyawareweb.org/`, see also `https://www.nsf.gov/awardsearch/showAward?AWD_ID=0427275`

[18] `https://en.wikipedia.org/wiki/Okapi_BM25` (visited Dec 26, 2023)

[19] `https://en.wikipedia.org/wiki/Discounted_cumulative_gain`, visited Dec 26, 2023.

(2015) provide an excellent survey [10].

Overall, the area of privacy for document owners is under-studied, possibly because the dominant paradigm in web search is centralized search engines that crawl vast amounts of the web, and that web search is itself the dominant search paradigm. Without some way for a web site to express access controls to the crawler, it isn't possible for a search engine to support them; moreover, since the web site is unaware of (most of) the search traffic, it might be impossible to express useful access controls beyond manually blocking access to privileged documents. One might imagine a default access control policy at the search engine that searchers from the domain of the content may access the content, unless the content (or parts of it) are otherwise made more open. I don't think we would recognize that web if we saw it today. Lastly, most policy discussions in this area focus on intellectual property ownership and compensating document sources, without much thought to privacy.

Moving away from the web paradigm, document owners can provide search access to their private documents themselves, and this can support whatever level of information access the owner has the capability to apply, for example access controls at the document and searcher level. Such approaches may be vulnerable to probing attacks where a searcher can learn the presence/absence or distribution of query terms in the collection, and the presence or absence of known documents.

In summary, content owner privacy risks include both intentional and inadvertent publishing of information online; as people say, once something is posted online, it is there forever. Challenges are both technical, in that people have few mechanisms to express privacy and access controls for their content, and societal, in that there is an unavoidable collision between someone's expectation that their own data will be private and their expectation that other people's data will be discoverable.[20] Opportunities in the current web environment are hard to imagine, but recent technical and governmental interest in content tracking and watermarking for AI-generated content could perhaps expand to cover real content as well.

6 Conclusion

Privacy is a recent area of interest for IR. While it is possible to define privacy differently depending on whether one takes a social, technical, or cryptographic approach to the question, this chapter takes its cue from the communities engaged in a search process: the document owner, the searcher, and the service host or search engine.

Although the work cited here gets an earnest start more than twenty years ago, there is quite a ways to go before privacy can be reliably supported for all of those users. Privacy engineers have put considerable effort into increasing users' privacy during web search. Document owners have some theoretical support

[20] We might claim to believe in others' privacy as well as our own, but profits from tabloids, data brokerages, and Facebook seem to indicate otherwise.

from the cryptography community, as do search providers. Within the paradigm of web search, which is where most attention has been focused, a start has been made.

Within other search paradigms, such as desktop search or search within private, subscription, or secure document environments, the weight of effort is on the other side. There is considerable work on maintaining the privacy of the document collection on the whole. Access policies and cryptographic schemes can support a great deal of use cases, but they assume considerable effort on the part of the owner or host system. If the document collection is too large or complex for manually-applied access controls, there has been some work in trying to predict which information should be controlled. Little or no work has focused on the privacy of searchers or the host of the search service who may wish to personalize searches to maximize success.

Some of the ideas proposed for private information retrieval would involve a proxy architecture where information about documents and searches is obscured. That kind of "privacy meta-search" architecture seemed completely out of touch until recently, however if governments decide to regulate privacy and tracking, it may be the best class of solutions.[21]

Overall, the IR community has not yet paid a lot of attention to the privacy of the people who sit just outside the searches we think about: the searcher, the search service provider, the document owner. In our pervasively networked, monitored, and advertised world, that may be changing. There is a lot of very good research to be done.

References

[1] Eytan Adar. "User 4xxxxx9: Anonymizing query logs". In: *Proc of Query Log Analysis Workshop, International Conference on World Wide Web*. 2007.

[2] David Elliott Bell. "Looking back at the Bell-La Padula model". In: *21st Annual Computer Security Applications Conference (ACSAC'05)*. 2005, 15 pp.–351. DOI: 10.1109/CSAC.2005.37.

[3] David Elliott Bell and Leonard J. LaPadula. *Secure Computer Sytems: Mathematical Foundations*. Tech. rep. MITRE Corporation, 1973.

[4] David Elliott Bell and Leonard J. LaPadula. *Secure Computer Sytems: Unified Exposition and Multics Interpretation*. Tech. rep. MITRE Corporation, 1976.

[5] Elisa Bertino, Gabriel Ghinita, and Ashish Kamra. "Access Control for Databases: Concepts and Systems". In: *Foundations and Trends in Databases* 3.1 (2010), pp. 1–148. ISSN: 1931-7883, 1931-7891. DOI: 10.1561/1900000014. URL: http://www.nowpublishers.com/article/Details/DBS-014 (visited on 02/28/2021).

[21] Between drafts of this chapter, privacy-enhanced search became a standard feature in the Safari and Brave web browsers.

[6] Bin Bi et al. “Inferring the demographics of search users: social data meets search queries”. In: *Proceedings of the 22nd international conference on World Wide Web*. WWW ’13. Rio de Janeiro, Brazil: Association for Computing Machinery, May 13, 2013, pp. 131–140. ISBN: 978-1-4503-2035-1. DOI: 10.1145/2488388.2488401. URL: https://doi.org/10.1145/2488388.2488401 (visited on 07/17/2020).

[7] Asia J. Biega, Rishiraj Saha Roy, and Gerhard Weikum. “Privacy through Solidarity: A User-Utility-Preserving Framework to Counter Profiling”. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’17. Shinjuku, Tokyo, Japan: Association for Computing Machinery, Aug. 7, 2017, pp. 675–684. ISBN: 978-1-4503-5022-8. DOI: 10.1145/3077136.3080830. URL: https://doi.org/10.1145/3077136.3080830 (visited on 05/06/2020).

[8] Asia J. Biega et al. “Operationalizing the Legal Principle of Data Minimization for Personalization”. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’20: The 43rd International ACM SIGIR conference on research and development in Information Retrieval. Virtual Event China: ACM, July 25, 2020, pp. 399–408. ISBN: 978-1-4503-8016-4. DOI: 10.1145/3397271.3401034. URL: https://dl.acm.org/doi/10.1145/3397271.3401034 (visited on 02/25/2021).

[9] Joanna Biega, Ida Mele, and Gerhard Weikum. “Probabilistic Prediction of Privacy Risks in User Search Histories”. In: *Proceedings of the First International Workshop on Privacy and Secuirty of Big Data*. PSBD ’14. Shanghai, China: Association for Computing Machinery, Nov. 7, 2014, pp. 29–36. ISBN: 978-1-4503-1583-8. DOI: 10.1145/2663715.2669609. URL: https://doi.org/10.1145/2663715.2669609 (visited on 05/06/2020).

[10] Christoph Bösch et al. “A Survey of Provably Secure Searchable Encryption”. In: *ACM Computing Surveys* 47.2 (2015), pp. 1–51.

[11] Stefan Büttcher and Charles L. A. Clarke. “A Security Model for Full-Text File System Search in Multi-User Environments”. In: *Proceedings of the 4th USENIX Conference on File and Storage Technologies (FAST 2005)*. San Francisco, CA, 2005. URL: https://www.usenix.org/legacy/event/fast05/tech/full_papers/buettcher/buettcher_html/.

[12] Sauvik Das and Adam Kramer. “Self-Censorship on Facebook”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 7.1 (June 28, 2013). Number: 1. ISSN: 2334-0770. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14412 (visited on 02/28/2021).

[13] Dorothy E. Denning. “A lattice model of secure information flow”. In: *Commun. ACM* 19.5 (May 1976), pp. 236–243. ISSN: 0001-0782. DOI: 10.1145/360051.360056. URL: https://doi.org/10.1145/360051.360056.

[14] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". In: *Foundations and Trends® in Theoretical Computer Science* 9.3 (Aug. 10, 2014). Publisher: Now Publishers, Inc., pp. 211–407. ISSN: 1551-305X, 1551-3068. DOI: 10.1561/0400000042. URL: https://www.nowpublishers.com/article/Details/TCS-042 (visited on 02/28/2021).

[15] Steven Englehardt and Arvind Narayanan. "Online Tracking: A 1-million-site Measurement and Analysis". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS'16: 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna Austria: ACM, Oct. 24, 2016, pp. 1388–1401. ISBN: 978-1-4503-4139-4. DOI: 10.1145/2976749.2978313. URL: https://dl.acm.org/doi/10.1145/2976749.2978313 (visited on 02/28/2021).

[16] Steven Englehardt et al. "Cookies That Give You Away: The Surveillance Implications of Web Tracking". In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. Florence, Italy: International World Wide Web Conferences Steering Committee, May 18, 2015, pp. 289–299. ISBN: 978-1-4503-3469-3. DOI: 10.1145/2736277.2741679. URL: https://doi.org/10.1145/2736277.2741679 (visited on 07/15/2020).

[17] Ronald Fagin. "On an Authorization Mechanism". In: *ACM Transactions on Database Systems* 3.3 (1978), pp. 310–319.

[18] Antonio Ginart et al. "Making AI Forget You: Data Deletion in Machine Learning". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/cb79f8fa58b91d3af6c9c991f63962d3-Paper.pdf.

[19] Patricia P. Griffiths and Bradforf W. Wade. "An Authorization Mechanism for a Relational Database System". In: *ACM Transactions on Database Systems* 1.3 (1976), pp. 242–255.

[20] Alon Halevy, Peter Norvig, and Fernando Pereira. "The Unreasonable Effectiveness of Data". In: *IEEE Intelligent Systems* 24 (2009), pp. 8–12. URL: http://www.computer.org/portal/cms_docs_intelligent/intelligent/homepage/2009/x2exp.pdf (visited on 02/26/2021).

[21] Rosie Jones et al. ""I know what you did last summer": query logs and user privacy". In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. CIKM '07. Lisbon, Portugal: Association for Computing Machinery, Nov. 6, 2007, pp. 909–914. ISBN: 978-1-59593-803-9. DOI: 10.1145/1321440.1321573. URL: https://doi.org/10.1145/1321440.1321573 (visited on 07/14/2020).

[22] Aleksandra Korolova et al. "Releasing search queries and clicks privately". In: *Proceedings of the 18th international conference on World wide web*. WWW '09. Madrid, Spain: Association for Computing Machinery, Apr. 20, 2009, pp. 171–180. ISBN: 978-1-60558-487-4. DOI: 10.1145/1526709.15

26733. URL: https://doi.org/10.1145/1526709.1526733 (visited on 07/14/2020).

[23] Andreas Krause and Eric Horvitz. "A utility-theoretic approach to privacy in online services". In: *Journal of Artificial Intelligence Research* 39.1 (Sept. 1, 2010), pp. 633–662. ISSN: 1076-9757.

[24] Arunesh Mathur et al. "Characterizing the Use of Browser-Based Blocking Extensions To Prevent Online Tracking". In: Fourteenth Symposium on Usable Privacy and Security ({SOUPS} 2018). 2018, pp. 103–116. ISBN: 978-1-939133-10-6. URL: https://www.usenix.org/conference/soups2018/presentation/mathur (visited on 02/28/2021).

[25] Nora McDonald and Andrea Forte. "The Politics of Privacy Theories: Moving from Norms to Vulnerabilities". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–14. URL: https://doi.org/10.1145/3313831.3376167.

[26] National Institute of Standards and Technology. *NIST PRIVACY FRAMEWORK:: A TOOL FOR IMPROVING PRIVACY THROUGH ENTERPRISE RISK MANAGEMENT, VERSION 1.0*. NIST CSWP 01162020. Gaithersburg, MD: National Institute of Standards and Technology, Jan. 16, 2020, NIST CSWP 01162020. DOI: 10.6028/NIST.CSWP.01162020. URL: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.01162020.pdf (visited on 02/28/2021).

[27] Alexandra Olteanu et al. "FACTS-IR: Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval". In: *ACM SIGIR Forum* 53.2 (2019), p. 24.

[28] Mahmoud F. Sayed and Douglas W. Oard. "Jointly Modeling Relevance and Sensitivity for Search Among Sensitive Content". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '19: The 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Paris France: ACM, July 18, 2019, pp. 615–624. ISBN: 978-1-4503-6172-9. DOI: 10.1145/3331184.3331256. URL: https://dl.acm.org/doi/10.1145/3331184.3331256 (visited on 02/28/2021).

[29] Mahmoud F. Sayed et al. "A Test Collection for Relevance and Sensitivity". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '20: The 43rd International ACM SIGIR conference on research and development in Information Retrieval. Virtual Event China: ACM, July 25, 2020, pp. 1605–1608. ISBN: 978-1-4503-8016-4. DOI: 10.1145/3397271.3401284. URL: https://dl.acm.org/doi/10.1145/3397271.3401284 (visited on 02/28/2021).

[30] Luo Si and Hui Yang. "Privacy-preserving IR: when information retrieval meets privacy and security". In: *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. SIGIR '14. Gold Coast, Queensland, Australia: Association for Computing Machinery, July 3, 2014, p. 1295. ISBN: 978-1-4503-2257-7. DOI: 10.1145/2600428.2600737. URL: https://doi.org/10.1145/2600428.2600737 (visited on 05/06/2020).

[31] Adish Singla et al. "Stochastic privacy". In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. AAAI'14. Québec City, Québec, Canada: AAAI Press, July 27, 2014, pp. 152–158. (Visited on 02/26/2021).

[32] Latanya Sweeney. "k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.5 (Oct. 1, 2002). Publisher: World Scientific Publishing Co., pp. 557–570. ISSN: 0218-4885. DOI: 10.1142/S0218488502001648. URL: https://www.worldscientific.com/doi/abs/10.1142/S0218488502001648 (visited on 02/26/2021).

[33] Ingmar Weber and Carlos Castillo. "The demographics of web search". In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '10. Geneva, Switzerland: Association for Computing Machinery, July 19, 2010, pp. 523–530. ISBN: 978-1-4503-0153-4. DOI: 10.1145/1835449.1835537. URL: https://doi.org/10.1145/1835449.1835537 (visited on 07/17/2020).

[34] Ingmar Weber, Venkata Rama Kiran Garimella, and Erik Borra. "Mining web query logs to analyze political issues". In: *Proceedings of the 4th Annual ACM Web Science Conference*. WebSci '12. Evanston, Illinois: Association for Computing Machinery, June 22, 2012, pp. 330–334. ISBN: 978-1-4503-1228-8. DOI: 10.1145/2380718.2380761. URL: https://doi.org/10.1145/2380718.2380761 (visited on 07/17/2020).

[35] Chuhan Wu et al. "Neural Demographic Prediction using Search Query". In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. WSDM '19. Melbourne VIC, Australia: Association for Computing Machinery, Jan. 30, 2019, pp. 654–662. ISBN: 978-1-4503-5940-5. DOI: 10.1145/3289600.3291034. URL: https://doi.org/10.1145/3289600.3291034 (visited on 07/17/2020).

[36] Elad Yom-Tov. "Demographic differences in search engine use with implications for cohort selection". In: *Information Retrieval Journal* 22.6 (Dec. 1, 2019), pp. 570–580. ISSN: 1573-7659. DOI: 10.1007/s10791-018-09349-2. URL: https://doi.org/10.1007/s10791-018-09349-2 (visited on 07/17/2020).

[37] Sergej Zerr et al. "Privacy-aware image classification and search". In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '12. New York, NY, USA: Association for Computing Machinery, Aug. 12, 2012, pp. 35–44. ISBN:

978-1-4503-1472-5. DOI: 10.1145/2348283.2348292. URL: https://doi.org/10.1145/2348283.2348292 (visited on 08/20/2020).

[38] Sicong Zhang, Hui Yang, and Lisa Singh. "Anonymizing Query Logs by Differential Privacy". In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. SIGIR '16. Pisa, Italy: Association for Computing Machinery, July 7, 2016, pp. 753–756. ISBN: 978-1-4503-4069-4. DOI: 10.1145/2911451.2914732. URL: https://doi.org/10.1145/2911451.2914732 (visited on 05/06/2020).

[39] Steven Zimmerman et al. "Investigating the Interplay Between Searchers' Privacy Concerns and Their Search Behavior". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'19. Paris, France: Association for Computing Machinery, July 18, 2019, pp. 953–956. ISBN: 978-1-4503-6172-9. DOI: 10.1145/3331184.3331280. URL: https://doi.org/10.1145/3331184.3331280 (visited on 07/17/2020).

[40] Steven Zimmerman et al. "Privacy Nudging in Search: Investigating Potential Impacts". In: *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. CHIIR '19. Glasgow, Scotland UK: Association for Computing Machinery, Mar. 8, 2019, pp. 283–287. ISBN: 978-1-4503-6025-8. DOI: 10.1145/3295750.3298952. URL: https://doi.org/10.1145/3295750.3298952 (visited on 05/06/2020).

[41] Steven Zimmerman et al. "Towards Search Strategies for Better Privacy and Information". In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. CHIIR '20. Vancouver BC, Canada: Association for Computing Machinery, Mar. 14, 2020, pp. 124–134. ISBN: 978-1-4503-6892-6. DOI: 10.1145/3343413.3377958. URL: https://doi.org/10.1145/3343413.3377958 (visited on 07/17/2020).